BOOK OF ABSTRACTS

XI COLLOQUIUM CHEMOMETRICUM MEDITERRANEUM

27-30 JUNE 2023
PADOVA - ITALY

SPONSORS

Università degli Studi di Padova

DII Dipartimento di Ingegneria Industriale

CAPE >_lab

VIAVI

BRUKER

OPTOPRIM
BRILLIANT IN LASER SOLUTIONS

EIGENVECTOR RESEARCH INCORPORATED

UNIMORE
Università degli Studi di Modena e Reggio Emilia

Università degli Studi dell'Insubria

DISAT

Società Italiana di Chemiometria

SEQA
Sociedad Española de Química Analítica

SARTORIUS

Società Chimica Italiana

DataBloom
let your data flourish

TOPICS

Dear Colleagues, dear Friends

We are pleased to announce that the 11th Colloquium Chemometricum Mediterraneum (CCM XI 2023) will take place in Padova (Italy) from 27 till 30 June 2023.

In 1986, during an international conference, several French, Spanish and Italian chemometricians, each one speaking his own language, gathered and were able to share their own experiences on this topic. These facts were the seeds of the Colloquium Chemometricum Mediterraneum. The official languages of the conference are neo latin languages and English.

These Conferences have had a continuity over the years, starting in Barcelona (Spain, 1987), and following in San Miniato (Italy, 1991), Bastia (France, 1994), Burgos (Spain, 1998), Ustica (Italy, 2003), Saint Maximin La Sainte-Baume (France, 2007), Granada (Spain, 2010), Bevagna (Italy, 2013), Arles (France, 2017), Menorca (Spain, 2019).

The conference brings an outstanding and unique opportunity for exchanging knowledge on leading edge developments. The 11th Colloquium Chemometricum Mediterraneum welcomes all contributions on classical tools and new tools in Chemometrics. A non-limitative list of topics includes:

- Pattern Recognition and Calibration
- Experimental Design (DOE) and System Optimization
- Qualimetrics and Chemical Metrology
- QSAR/QSPR
- Image Analysis and hyperspectral imaging
- Process Analytical Technology (PAT)
- Multiway and Multiset analysis
- Teaching Chemometrics
- Recent Chemometrics developments
- Applications of Chemometrics in different domains: Cultural heritage, Environment, Food control, Omics sciences, Pharmaceuticals...

Following the tradition of previous editions, a Virtual Special Issue (VSI) on Chemometrics and Intelligent Laboratory Systems (Elsevier) will be published. (Further info here).

The Organising Committee

COMMITTEES

SCIENTIFIC COMMITTEE

Prof. Marina Cocchi, Università di Modena e Reggio Emilia
Prof. Luis Cuadros, Universidad de Granada
Prof. Cyril Ruckebusch, Université de Lille
Prof. Riccardo Leardi, Università degli Studi di Genova
Prof. Federico Marini, Sapienza Università di Roma
Prof. Jean-Michel Roger, IRSTEA
Prof. Luis Sarabia, Universidad de Burgos
Prof. Michelle Sergent, Aix-Marseille University
Prof. Romà Tauler, IDAEA-CSIC
Prof. Roberto Todeschini, Università Milano - Bicocca

ORGANIZING COMMITTEE

Prof. Pierantonio Facco, Università di Padova
Prof. Davide Ballabio, Università di Milano – Bicocca
Dr. Eugenio Alladio, Università di Torino
Dr. Caterina Durante, Università di Modena e Reggio Emilia
Prof. Barbara Giussani, Università dell'Insubria
Prof. Paolo Oliveri, Università di Genova
Dr. Gianmarco Barberi, Università di Padova
Ing. Francesco Sartori, Università di Padova

INDEX

PLENARIES

Toward the systematic development of hybrid models

Marco S. Reis

Univ Coimbra, CIEPQPF, Department of Chemical Engineering, Rua Sílvio Lima, Pólo II –
Pinhal de Marrocos, 3030-790 Coimbra, Portugal,
E-mail: marco@eq.uc.pt, phone: +351 239 798 700, FAX: +351 239 798 703

As industrial data is becoming easier to assess, distribute, process, and store, more and more research efforts have been devoted to taking advantage of this resource for different purposes. In Industrial Data Science/Process Analytics, the stream of new statistical, machine learning (ML), and artificial intelligence methods has been growing (AI), to adapt to the increasing demand for data analysis and to turn data into value. The high visibility of deep AI methods for large-scale image and video recognition (with deep convolutional neural networks, such as the VGG-16, achieving remarkable accuracies, namely in the ImageNet challenge), and, more recently, for natural language processing (NLP) (with the generative pre-trained transformers, such as GPT-n, currently at n=4, enabling close-to-human expert interactions in different domains), has been attracting the attention of other scientific communities to test and adapt these large-scale AI models to their own domains.

A common feature of deep AI methods is their strong dependence on rich data sets where they can be trained and where the relevant features to be learned must be abundantly present. This is not a problem in image/video recognition, where more data can usually be collected or generated, and is certainly not a problem for NLP, given the huge resources of text data. But when addressing the analysis of industrial processes and their problems, the scenario is much different. And it is not just a matter of quantity, which could be the first obvious obstacle (further aggravated by the omnipresence of non-stationarity, that makes stored past data obsolete and useless for most current purposes). The main issue is the amount of information contained in industrial data sets, which may be too limited for robustly training a medium-scale statistical/ML model, let alone to meet the eagerness of the data-thirsty deep AI methods. In fact, most industrial processes operate around tight operational bounds, with as low variation as possible, which for the purposes of training data-driven models, means low information.

However, data is not the only source of knowledge available for building models for industrial processes. Phenomenological knowledge and the application of first principles offer further channels of information that can be fed and used during model building and development, and that can help overcome the low information density of industrial databases or to make better use of the data collected in experiments conducted at any scale: laboratory, pilot or industrial. This is the goal of the so-called hybrid models, first developed in the 2010's (Georgieva et al., 2003; von Stosch et al., 2014), and more recently referred to by other designations, such as physics-inspired/guided/informed, grey-box, knowledge based, etc. (Sansana et al., 2021). The original developments usually have in common the need to fill in for some specific knowledge gap in the mechanistic description of the system, such as the kinetic model for a reaction. They were developed in a knowledge-intensive environment, and data were used to induce the behavior of one of its components. Currently, even though such situations remain valid and relevant in practice, one often must deal with data-intensive environments, that may be however information poor, for the reasons described above. This new context calls for new hybrid methods, and a new wave of interest has been emerging recently, leading to novel approaches such as Physical Inspired Neural Networks (PINN's) (Raissi et al., 2019), among others.

In this talk, different scenarios and applications are referred to, where the development of hybrid models is opportune to complement the limitations of data-driven approaches and/or make the results more informative and interpretable, including the analysis of batch/continuous processes, multimode operation, and degradation phenomena. We will also discuss some principles that structure the development of hybrid models, to contribute to a more systematic and structured activity for addressing this field under fast expansion.

References

Georgieva, P., Meireles, M.J., and Feyo de Azevedo, S. (2003). Knowledge-Based Hybrid Modelling of Batch Crystallisation when Accounting for Nucleation, Growth and Agglomeration Phenomena. Chemical Engineering Science *58*, 3699-3713.

Raissi, M., Perdikaris, P., and Karniadakis, G.E. (2019). Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. Journal of Computational Physics *378*, 686-707.

Sansana, J., Joswiak, M.N., Castillo, I., Wang, Z., Rendall, R., Chiang, L.H., and Reis, M.S. (2021). Recent trends on hybrid modeling for Industry 4.0. Computers & Chemical Engineering *151*, 107365.

von Stosch, M., Oliveira, R., Peres, J., and Feyo de Azevedo, S. (2014). Hybrid semi-parametric modeling in process systems engineering: Past, present and future. Computers & Chemical Engineering *60*, 86-101.

Application of DoE in pharmaceutical formulation development

Pierre Lebrun

Addresses: Pharmalex, Global Statistics and Data Sciences, Rue E Belin, 5, 1435 Mont-Saint-Guibert, Belgium
pierre.lebrun@pharmalex.com (corresponding author)

Formulation development of biologics is a complex process as the number of excipients to be added to the drug substance and buffer can be large, to ensure the bioavailability, compatibility, safety, potency, and most importantly the guarantees of an appropriate shelf life of the drug product.

The experimental studies hence come with a potentially elevated cost of material and the need to produce small drug product batches that are more difficult to control. Finally, a large cost is associated to the attributes that are verified on the drug product. Classical release attributes must be analyzed, which can be numerous for biologics. Furthermore, the product is often followed in stability studies, including long term, accelerated, and stress testing, such as high temperature, agitation, or freeze-thaw studies.

The need to create minimal designs is therefore prominent, as the costs of all subsequent studies is, if not multiplied, directly related to the number of formulations to create.

In a designed experiment, each formulation would count for a line of a design matrix. Strong knowledge on potential interaction may help to avoid including them all, or bracketing/screening strategies may help to decrease the costs. Often, design augmentation to uncover better interactions and quadratic effects is not a preferred strategy, for example because starting from scratch a new 2-year long stability is not an option. Hence, careful risk analysis of the formulation parameters (design factors) is paramount.

After collecting the data, it's important to consider a statistical analysis encompassing a proper uncertainty analysis in order to make relevant inference. An optimization metric in such studies is the so-called probability of success (P(success)), which is the probability to have all attributes, including longitudinal stability profiles, to be satisfactory enough to ensure the future drug product will be of appropriate quality, from release to end of shelf-life.

In this presentation, we will review several designs that can be applied in such context, from very sparse (e.g. super-saturated design) to more complete that allow for a predictive ability. We will see some common methodologies that allow creating these DoEs, and how they are generally analyzed to ensure nothing is left aside.

Then, we will see how Bayesian statistics is the answer to the somewhat complicated P(success) computation using multivariate multiple regression, while some software provides solutions that may be appropriate or not depending on the uncertainty structure (e.g. variance profiles)

Finally, we will oversee future developments of similar studies that could beneficiate from kinetic modelling of the degradation using ordinary differential equations. For such non-linear models, the question of the design and analysis is generally much more complex, while the final result is similar: guaranteeing long term quality of the drug product. Using kinetic modeling, the target is to have models that are better informed that first order model, so the number of experiments may be further decreased.

Harnessing deep learning for drug discovery – challenges and opportunities

Francesca Grisoni[1,2]

[1]Institute for Complex Molecular Systems, Dept. Biomedical Engineering, Eindhoven University of Technology, 5612AZ Eindhoven, The Netherlands;
[2]Centre for Living Technologies, Alliance TU/e, WUR, UU, UMC Utrecht, 3584CB Utrecht, The Netherlands.
f.grisoni@tue.nl

Deep learning has the potential to revolutionize drug discovery by accelerating the drug development process, reducing costs, and improving the success rate of drug candidates. This talk will focus on recent studies that harnessed deep learning for drug discovery, with a particular emphasis on bioactivity prediction and de novo design. These studies will be used to illustrate the potential of deep learning in the chemical sciences, but also to underscore current limitations. Finally, I will provide a personal perspective on future opportunities for deep learning in medicinal and organic chemistry, to accelerate molecule discovery and chemical space exploration.

UNBLIND CROSS-VALIDATION

Sergey Kucheryavskiy[1], **Oxana Rodionova**[2], **Alexey Pomerantsev**[2]

[1]Aalborg University, Department of Chemistry and Bioscience, Niels Bohrs vej 8, Esbjerg, 6700, Denmark

[2]Semenov Federal Research Center for Chemical Physics RAS, Moscow, Russia

E-mail (svk@bio.aau.dk)

Validation is one of the most important steps in chemometric modelling, which can drastically change the final results and the performance of a model, especially when it is being applied for the prediction of new data. The best solution for validation of a final model with optimized hyperparameters is to use a validation set – an independent set of measurements comprising the same population as the calibration set.

However, validation is also required for the optimization of the model, for selection of the best set of preprocessing methods and for the selection of the most important variables. In this case, the well-established solution is to use cross-validation, which is based on iterative resampling. However, cross-validation has several drawbacks. It does not actually validate the model of interest. The cross-validation outcomes (e.g. predicted values) are computed for a set of local models, which are built using initial calibration subsets. Another issue is how to split the objects from the calibration set into segments. Many different techniques exist, however, it is difficult to prove which one is correct in a particular case, as the only outcome cross-validation provides are predictions. Finally, conventional cross-validation can be time-consuming, especially in the case of large datasets, as the resampling procedure must be repeated over and over until the optimization is done.

These drawbacks can be overcome by using a recently proposed Procrustes Cross-Validation (PCV), which takes the best of the two methods – validation with an independent set and cross-validation [1, 2]. The main idea of PCV is to use cross-validation for estimation of sampling error and then introduce this error into the calibration set, which results in a new set of measurements – PV-set. PV-set can then be used for validation of the global models in the same way as the independent validation set. It also provides several additional tools that help to assess the PV-set quality and find the best splitting strategy. One could say that PCV is the way to "unblind" conventional cross-validation.

In this presentation, we will show the main PCV principles and how it can be used for validation of a wide range of models, including PCA/SIMCA, PCR, and PLS regression. We will also present various diagnostic tools that can be used to assess the quality of the PV-set and optimize the splitting strategy of the calibration set. Finally, we will propose a new splitting strategy that was developed with the help of these tools.

References

1. Kucheryavskiy S, Zhilin S, Rodionova O, Pomerantsev O, **(2020)** Procrustes Cross-Validation—A Bridge between Cross-Validation and Independent Validation Sets. *Analytical Chemistry*, **92** 11842–11850.
2. Kucheryavskiy S, Rodionova O, Pomerantsev O, **(2023)** Procrustes cross-validation of multivariate regression models. Analytica Chimica Acta, **1255**

Last advances in the development of imaging biomarkers for breast cancer detection

Aguado-Sarrió, E.[a)]; Prats-Montalbán, J.M.[b)]; Ortiz, C.[a)]; Camps-Herrero, J.[c)]; and Ferrer, A.[b)]

a) KENKO Imalytics S.L. Calle Pedro Duque, Edicifio 9B. Universitat Politènica de València. Cno de Vera s/n. 26022 Valencia, Spain

b) Multivariate Statistical Engineering Group. Applied Statistics Department. Universitat Politènica de València. Cno de Vera s/n. 26022 Valencia, Spain

c) Grupo Ribera Salud. Avd. Cortes Valencianas, 58. 46015, Valencia, Spain

---

Currently, breast cancer is one of the types of cancer with highest incidence in the world. Approximately 2,261,419 cases of breast cancer were diagnosed worldwide in 2020, representing 12.5% of all malignant tumors. Due to its prevalence and importance in women, worldwide technological and scientific efforts are oriented towards the search for a more efficient and earlier detection. For the identification of cancer, currently, in most hospital radiology services, the study of MRI images is done qualitatively, where the expert's criterion is what determines the result of the radiological reports. This way, there is a need in hospitals to support this methodology with objective (quantitative) parameters calculated from the images that report on certain characteristics of the tissues, the so-called "imaging biomarkers (IB)". The IB are location and quantification maps related to the presence and development of certain biological processes (for example, cancerous tumors) that provide, for each pixel of the images, quantitative and objective information that can be interpreted by doctors, thus improving diagnosis and providing information on how tumors develop. These IB are derived from perfusion (DCE-MRI) and from diffusion (DW-MRI) sequences.

In the case of diffusion (DW-MRI), although in certain tissues (for example, the prostate) the restrictions can be considered similar for all directions and, consequently, isotropy can be assumed, in the case of other organs (for example, the breast or the brain) there are preferential directions for the water diffusion. Therefore, it is a case of anisotropic diffusion where it is not possible to characterize diffusion by means of a single apparent diffusion coefficient, ADC. Therefore, in the presence of anisotropy, the diffusion tensor technique (Diffusion Tensor Imaging, DTI) should be used instead, obtaining the IB from the calculation of the eigenvalues of the tensor.

Despite the enormous progress in the development of these imaging biomarkers, they still present a series of drawbacks:

- Biomarkers are calculated pixel by pixel through mathematical models (complex on many occasions), without taking advantage of the internal correlation structures between pixels, and without offering information on the quality of their estimation.
- The types of biomarkers to analyze depend on the organ studied.
- Simultaneous evaluation of several biomarkers is difficult, even more so when considering the different combinations that may be related to a tumor. This can be complicated and requires a priori knowledge.
- Some of the parameters obtained (e.g. K^{trans}) have a complex physiological interpretation. Therefore, it seems interesting to obtain new biomarkers with easy clinical interpretability.

One way to address these problems and challenges is to analyze MRI sequences by multivariate statistical projection models, such as MCR, PCA, or PLS ; obtaining simplified structures that help to understand the relationships between the variables studied and the underlying physiological phenomena.

The application of these techniques to images is called Multivariate Image Analysis (MIA). In the case of medical images, the observations are formed by each of the pixels of the images, while the columns contain the signal of each pixel at each instant of time in the case of pharmacokinetic perfusion models, or b values in diffusion models.

Thus, the application MIA is presented as an appropriate way to deal with this natural variation and develop new biomarkers, as well as to improve existing ones, such as the calculation of the DTI from PLS models.

Moreover, through the application of MCR, new easy-to-interpret imaging biomarkers are obtained, so radiologists can understand them, regardless of their training and experience. In addition, error images are provided, locating those pixels of non-reliable values.

Finally, it is possible to integrate all relevant biomarkers for diagnosis or prognosis depending on each organ, in a single image or probabilistic map (Virtual Biopsy) that quantifies in each pixel the probability of existence of a lesion, making the task of the radiologist much easier and, again, allowing its use by inexperienced personnel.

ORAL CONTRIBUTIONS

	Tuesday June 27 *Centro Culturale Altinate San Gaetano*	Wednesday June 28 *Centro Culturale Altinate San Gaetano*	Thursday June 29 *Centro Culturale Altinate San Gaetano*	Friday June 30 *Aula Magna - Palazzo Bo*
09:00 - 09:20		PL02	PL04	PL05
09:20 - 09:40		Lebrun	Kucheryavskiy	Montalban
09:40 - 10:00		Fornasaro	Amigo Rubio	Olarini
10:00 - 10:20		Benedetti	Beese	Rocha de Oliveira
10:20 - 10:40		Cenci	Calvini	Vitale
10:40 - 11:00		Garcia Carrion	Caponigro	Macchia
11:00 - 11:30		break + posters	break + posters	break
11:30 - 11:50		Sergent	Ruckebusch	Marini
11:50 - 12:10		Magnaghi	Cruz Muñoz	Gomez Sanchez
12:10 - 12:30		Farinini	Guerrini	Brandolini-Bunlon
12:30 - 12:50		Rovira	Jaumot	Iaccarino
12:50 - 13:10		Camacho	Menozzi	Biancolillo
13:10 - 14:30	opening 14-14:30	lunch + posters	lunch + posters	lunch
14:30 - 14:50	PL01	PL03	awards	Sozzi
14:50 - 15:10	Reis	Grisoni	Tauler	Kassouf
15:10 - 15:30	Borras-Ferris	Boldini	Qannari	Cazzaniga
15:30 - 15:50	Cairoli	Gambacorta	Arroyo Cerezo	Durante
15:50 - 16:10	Giussani	Birolo	Bernard	closure
16:10 - 16:30	Sanchez	social event	Sarabia	
16:30 - 17:00	break + posters		break + posters	
17:00 - 17:20	de Juan		Roger	
17:20 - 17:40	Sartori		Gastone	
17:40 - 18:00	Mas Garcia		Lesnoff	
18:00 - 18:20	Barra		Sandrucci	
18:30 - 20:00	welcome cocktail			
20:00 - 24:00			social dinner	

Multivariate Statistical Process Control via SMB-PLS

J. Borràs-Ferrís[1], **C. Duchesne**[2], **A. Ferrer**[1]

[1] Universitat Politècnica de València, Department of Applied Statistics and Operational Research and Quality, Camí de Vera, València, 46022, Spain
2 Laval University, Chemical Engineering Department, Rue de l'Université, Quebec, 2325, Canada
joaborfe@eio.upv.es

In modern manufacturing processes, massive amounts of multivariate data are routinely collected through automated in-process-sensing. These data often exhibit high correlation, rank deficiency, low signal-to-noise ratio and missing values. Conventional univariate and multivariate statistical process control techniques are not suitable to be used in these environments and, hence, several authors advocate the use of multivariate statistical process control based on latent variable modes as an efficient statistical tool for process understanding (Ferrer, 2007). In this sense, Partial Least Squares (PLS) regression is of interest because it models the variation in the critical input parameters space as well as in the critical quality attributes (CQAs) space of the final product (Kourti, 2005).

Besides, very frequently in the industry one encounters processes composed of two critical input parameter blocks having an impact on CQAs (Y): the raw materials block (Z) and the process conditions block (X). In such a case, rather than building a model for both blocks, one could build a model for the full process that will take into account the interactions between these blocks by means of the Multi-block PLS (MB-PLS) (Kourti, 2005). However, raw material variations are often intermittent and unpredictable (i.e., special sources of variations). In such a case, process conditions are usually modified by control actions to compensate for these variations and, hence, they are propagated from raw materials to process conditions. The latter hinders the process monitoring by a conventional MB-PLS. Besides, forcing the supplier to be in statistical control either would result in higher raw materials cost or may not be feasible, for instance, when raw materials come from natural resources.

Sequential Multi-block PLS (SMB-PLS) algorithm has very interesting advantages to address such problem, since it allows considering the sequential nature of the blocks (Z and X) when modeling the relationship with the quality (Y) (Azari et al., 2015). Indeed, the algorithm imposes a sequential pathway between the regressor blocks according to the process flowsheet, and then uses orthogonalization to separate correlated information between the blocks from orthogonal variations. Hence, variations in process conditions that are correlated with raw material properties (e.g., control actions) are extracted in a first block of latent variables, and orthogonal variations in process conditions (i.e., uncorrelated with raw material properties) are then extracted in a second block (e.g., operator actions, known disturbances, etc.). For that reason, the SMB-PLS is more efficient to establish the multivariate statistical process control when raw material properties and process conditions are correlated as better sort the contribution of both on CQA variations.

The methodology was applied to a real case study from a food manufacturing process. In this case, raw material properties are subject to special sources of variations affecting process conditions due to control actions in the manufacturing process. The SMB-PLS ends up building two blocks of latent variables. The first one, related to raw material properties and correlated process conditions, can be used not only to predict the expected food CQAs before manufacturing, but also to improve control actions if it is required. Note that, the raw material monitoring would not be suitable in this case due to the special sources of variation. By contrast, the second block, related to the process conditions orthogonal to the food properties, can be used for process monitoring in order to detect significant variations with respect to the expected control actions.

References

Azari, K., Lauzon-Gauthier, J., Tessier, J., & Duchesne, C. (**2015**). Establishing multivariate specification regions for raw materials using SMB-PLS. *IFAC-PapersOnLine*, ***48***, 1132–1137.

Ferrer, A. (**2007**). Multivariate Statistical Process Control Based on Principal Component Analysis (MSPC-PCA): Some Reflections and a Case Study in an Autobody Assembly Process. Quality Engineering, **19**, 311–325.

Kourti, T. (**2005**). Application of latent variable methods to process control and multivariate statistical process control in industry. *International Journal of Adaptive Control and Signal Processing*, ***19***, 213–246.

Fingerprint-2-Footprint: dynamically reformulating animal feed production via NIR spectroscopy and life-cycle assessment

Maria Cairoli[1], Anne Ottenbros[2], Sin Yong Teng[1], Mark Schoot[3], Christiaan Kapper[3], Steef Hanssen[2], Rosalie van Zelm[2], Lutgarde Buydens[1], Mark Huijbregts[2], Jeroen Jansen[1]

Addresses: [1]Radboud University, Department of Analytical Chemistry & Chemometrics, Heyendaalseweg 135, 6525 AJ Nijmegen, The Netherlands.
[2]Radboud University, Department of Environmental Science, Heyendaalseweg 135, 6525 AJ Nijmegen, The Netherlands.
[3]Nutricontrol, N. C. B. Laan 52, 5462GE Veghel, The Netherlands.
E-mail: maria.cairoli@ru.nl

The necessity of incorporating environmental criteria in decision making is increasingly recognized by stakeholders, that are urgently asked by international policies and agreements to adopt sustainable production pathways[1,2]. Environmental sustainability and economic profit, however, are not always in synergy. A better understanding of potential synergies between reducing environmental and economic costs may lead to more sustainable production strategies and win-win situations. In this study, we aim to explore whether new combinations of existing digital technologies can provide more efficient production strategies to achieve such synergy. Considering a standard animal feed production process, we demonstrate how NIR spectroscopy, Chemometrics, and life-cycle assessment can be integrated and employed as a complementary tool to optimize recipe formulations while enhancing the economic profit, lowering the environmental footprint and maintaining the quality standards.

Animal feeds are mixtures of several feedstuffs (*e.g.*, wheat, soybean) that are grown in different origins, shipped globally and combined in nutritionally optimal mixing ratios to satisfy market requirements. NIR spectroscopy allows a non-destructive, time- and cost-effective, and accurate characterization of feedstuff properties, including nutritional content and origin[3]. LCA allows estimating the environmental impacts of feedstuffs of various origins over their life cycle. Chemometric predictive models were employed to link the chemical PAT spectra to the environmental footprint of the corresponding samples, predicting not only feedstuffs' nutritional content, but also the respective environmental impact. This predicted information was combined with feedstuffs' price in a stochastic multi-objective optimization to obtain a set of candidate optimal formulations (*i.e.,* Pareto front) with reduced costs. Multi-criteria decision making was ultimately applied to select the optimal feedstuffs' mixing ratio that better approaches the synergy between environmental and economic costs (**Figure 1**).

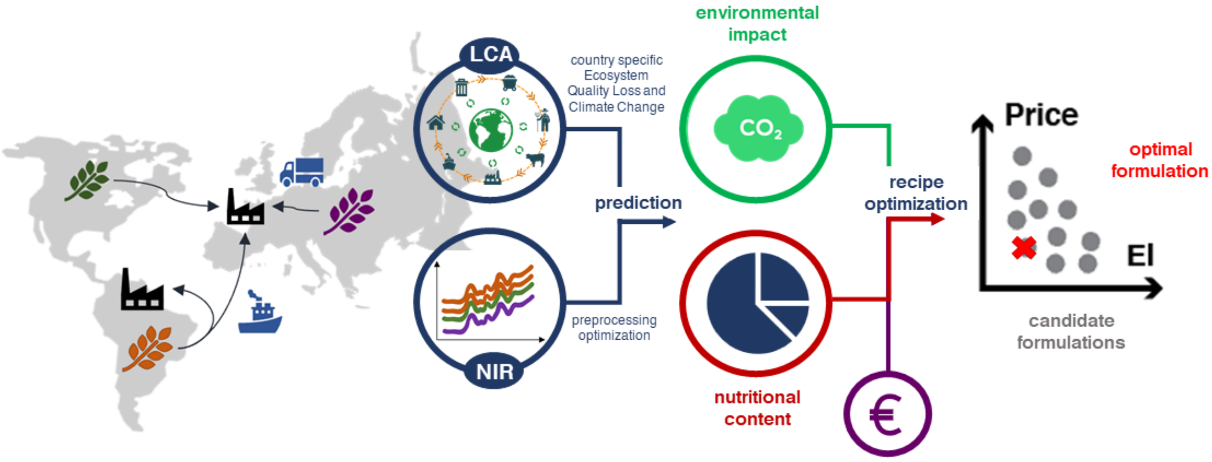


Figure 1 Workflow for the animal feed recipe optimization framework integrating NIR spectroscopy, Chemometrics and LCA.

This interdisciplinary approach enabled deriving environmentally and economically optimal feedstuffs mix ratios (including feedstuff origin), while meeting product market requirements, from spectral fingerprints alone. As we include the variability in environmental and economic costs in our optimization, and use NIR spectroscopy, which is already routinely employed to assess quality standards, our approach constitutes both a pragmatic and efficient digital tool solution for companies to make sensible decisions to achieve environmental and quality targets without additional costs.

Acknowledgements

The authors would like to thank the members of the ISPT "Management 4 Measurement" consortium for their financial and in kind contribution. This consortium consists of the following organizations: DSM, ISPT, Kraft Heinz, Magion, Nexperia, Nouryon, Nutricontrol, Radboud University, RIWA Rijn, Unilever. This project received funding from TKI E&I with the supplementary grant 'TKI-Toeslag' for Topconsortia for Knowledge and Innovation (TKI's) of the Ministry of Economic Affairs and Climate Policy.

References

1. Fuso Nerini, F. *et al.* Connecting climate action with other Sustainable Development Goals. *Nat. Sustain.* **2**, 674–680 (2019).

2. Dantas, T. E. T. *et al.* How the combination of Circular Economy and Industry 4.0 can contribute towards achieving the Sustainable Development Goals. *Sustain. Prod. Consum.* **26**, 213–227 (2021).
3. Karoui, R. & De Baerdemaeker, J. A review of the analytical methods coupled with chemometric tools for the determination of the quality and identity of dairy products. *Food Chem.* **102**, 621–640 (2007).

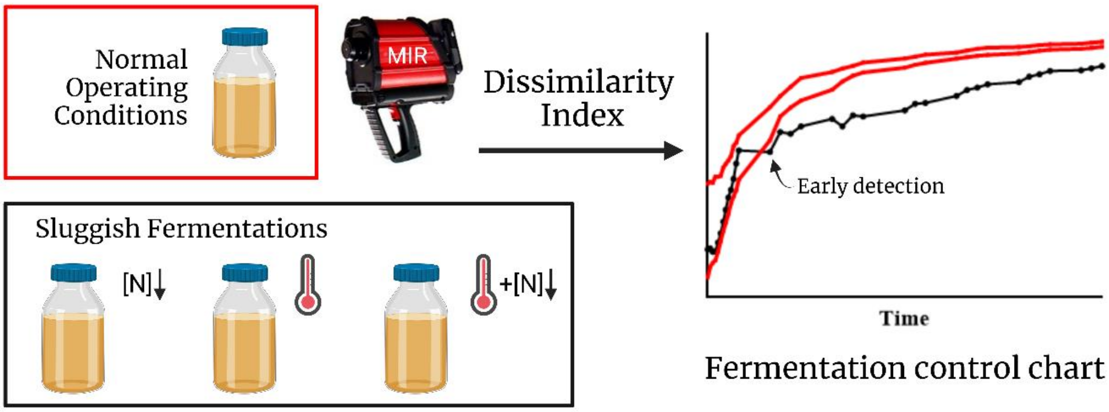
A new approach to the Dissimilarity Index as a process control tool to detect deviations in alcoholic fermentation

Daniel Schorn-García[a], Jokin Ezenarro[a], Olga Busto[a], Laura Aceña[a], Ricard Boqué[a], Montserrat Mestres[a], Barbara Giussani[b]

[a]QAQO Dpt, Universitat Rovira i Virgili, Tarragona, Spain

[b]DISAT, Università Degli Studi Dell'Insubria, Como, Italy

barbara.giussani@uninsubria.it



Food processing is highly complex due to the presence of many species, including microorganisms, which may participate in the biochemical transformations that take place. Quality control is therefore a fundamental step in all processing and production phases. Traditionally, quality control has relied on daily univariate measurements and final product characterization. However, Process Analytical Technologies (PAT), which have become popular in the pharmaceutical process, are also gaining implementation in the food and beverage sector. The concept behind PAT is to guarantee the quality throughout the entire production process [1]. Spectroscopic techniques, such as infrared, play a critical role in the implementation of PAT as they offer fast, non-destructive and in-line or on-line measurement possibilities that provide information on the molecular composition of the process [2].

Spectroscopic techniques generate multivariate data, and for this reason, Multivariate Statistical Process Control (MSPC) techniques are used to monitor and control the process. The Dissimilarity Index (DI), firstly proposed by Muncan et al. for monitoring yogurt fermentation using near-infrared spectroscopy, is based on the use of a principal component (PC) loading to describe the progress of the process [3]. Using a moving-window approach and the first loading of the fermentation, changes in the process can be identified. The index is described as follows:

$$A_i=1-|\mathbf{p}(i)^T\mathbf{p}(0)|$$

where A_i is the dissimilarity index to detect changes between angles of the principal component loading ($\mathbf{p}(i)^T$) of the i^th^ sampling point and a reference loading $\mathbf{p}(0)$ of a preceding time.

In this research, a modified version of the DI is proposed to monitor wine alcoholic fermentation. This modified DI uses an evolving window approach and calculates de reference loading from the first hours' spectra working under Normal Operating Conditions (NOC). Then, spectra from the beginning of the fermentation to a given sampling point are considered and their first PC loading calculated and referred to the reference loading. The proposed approach was able to monitor alcoholic fermentation and, using an MSPC control chart, to detect deviations of the process. For this purpose, experiments were planned and performed by simulating temperature control problems and an initial deficit in assimilable nitrogen compounds that affect yeast metabolism. In both cases, a satisfactory detection was achieved at early stages. In addition, the effect of the deviations was studied with ANOVA–simultaneous component analysis (ASCA), showing a greater impact of the initial nitrogen concentration on the evolution of the process compared to sudden changes in temperature.

References

[1] FDA Official Document (**2004**) Guidance for Industry PAT — A Framework for Innovative Pharmaceutical Development, Manufacuring, and Quality Assurance 16.

[2] Cozzolino D (**2022**) Advantages, Opportunities, and Challenges of Vibrational Spectroscopy as Tool to Monitor Sustainable Food Systems, Food Anal Methods. **15** 1390–1396.

[3] Muncan J, Tei K, Tsenkova R (**2021**) Real-time monitoring of yogurt fermentation process by aquaphotomics near-infrared spectroscopy, *Sensors (Switzerland)*. **21** 1–18.

Partial Least Squares model inversion for intervals

M.C. Ortiz, L.A. Sarabia, M.S. Sánchez

Universidad de Burgos, Faculty of Sciences, Plaza Misael Bañuelos s/n, 09001 Burgos, Spain
E-mail (corresponding author): ssanchez@ubu.es

The work addresses the inversion of latent variables models (Partial Least Squares, PLS, models in this case) for intervals in the response space. The procedure applies in the general context of quality control under the Quality by Design (QbD) paradigm when the quality characteristic of interest has specification limits. In that case, the objects resulting from the inversion will be in the design space for the given process.

Precisely in the present work, there is a single response to be fitted with a PLS model and the interest focusses on an interval of response values, i.e., in the output space. The goal is to determine, if any, a region that contains values of the predictor variables (i.e., in the input space) whose predictions with the PLS model lie inside the given interval.

This interval of interest can define not only specification limits for a quality characteristic, it can be just a confidence interval on the prediction or a tolerance interval built for some predicted values of interest. Irrespective of its meaning, the situation is the same, the inversion of an interval of response values (y_1, y_2).

Contrary to other approaches, with confidence intervals [1], and given the linear nature of PLS models, in the present work the inversion of the PLS model is only made, separately, for y_1 and y_2 to obtain parallel hyperplanes in the input space that account for the null spaces [2] related to both the projection and regression steps.

The situation is illustrated in Figure 1 with a simulated three-dimensional data set **X** (emulating three parameters on which the process under study depend). The two extremes of a given interval in the output space, via the PLS model inversion, give two parallel planes inside the experimental domain, depicted in Figure 1 in blue and green. For the graphical representation in Figure 1, the experimental domain is considered to be the "cube" defined with the points in **X** (black diamonds).

For practical purposes, these estimates preliminarily delimit the design space which is a set that will be inside the experimental domain and inside the region "between" these hyperplanes.

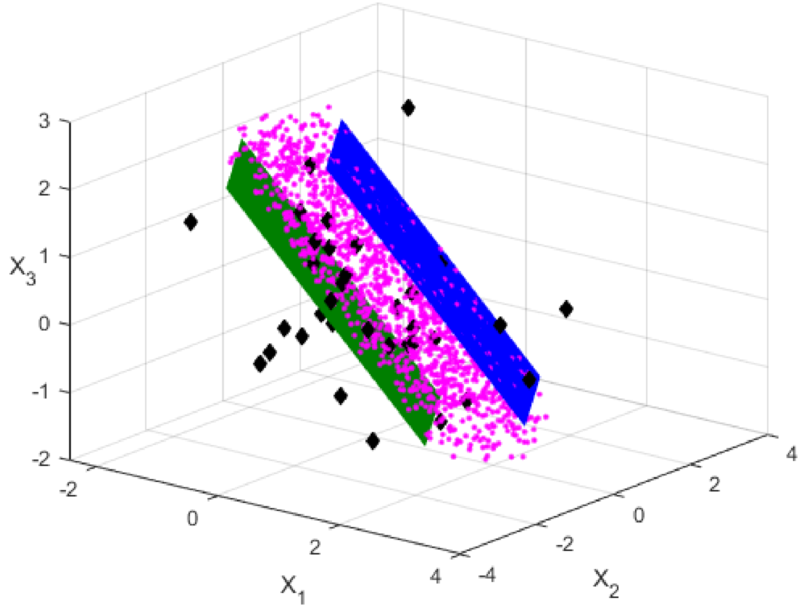


Figure 1. Three-dimensional input space. The objects in **X** are in black diamonds, the green and blue planes resulted from the inversion of a PLS model for an interval and define boundaries for the settings in the design space (magenta dots).

To explore and more precisely delimit the boundaries of the design space, this region should be adequately filled to select points for the experimental validation. The magenta points in Fig. 1 mark some solutions, obtained with a Monte Carlo method, all of them with predictions that will lie inside the given interval.

The method will be illustrated for the inversion of tolerance intervals, which can be relevant in some classification studies where it is more important to reduce false negatives than false positives.

Acknowledgements

The work is part of the research project ref. BU052P20 co-financed by the Spanish Junta de Castilla y León, and European FEDER and *«Recovery plan for Europe – NextGenerationEU»* funds.

References

[1] P. Facco, F. Dal Pastro, N. Meneghetti, F. Bezzo, M. Barolo (**2015**) Bracketing the Design Space within the Knowledge Space in Pharmaceutical Product Development, *Ind. Eng. Chem. Res.,* **54,** 5128–5138. DOI: http://doi.org/10.1021/acs.iecr.5b00863

[2] S. Ruiz, L.A. Sarabia, M.C. Ortiz, M.S. Sánchez (**2020**), Residual spaces in latent variables model inversion and their impact in the design space for given quality characteristics, *Chemometr. Intell. Lab. Systems,* **203,** 104040. DOI: https://doi.org/10.1016/j.chemolab.2020.104040

Multivariate Curve Resolution for incomplete multisets. Addressing the problem with a single factorization model.

Anna de Juan[1], Adrián Gómez-Sánchez[1,2], Romà Tauler[3]

Chemometrics group. Universitat de Barcelona. Martí i Franquès, 1. 08028 Barcelona.
LASIRE. Laboratory of Advanced Spectroscopy, Interactions, Reactivity and Environment
Université Lille, CNRS, UMR 8516, Cité Scientifique, Bâtiment C5, 59000, Lille, France.
IDAEA-CSIC. Jordi Girona 18, 08034 Barcelona.
E-mail: anna.dejuan@ub.edu

The use of Multivariate Curve Resolution (MCR) for multiset analysis was a breakthrough for this kind of methodology[1]. Indeed, multisets allowed for the combination of data blocks containing very different information into a single augmented matrix structure that could be subsequently decomposed into a bilinear model of meaningful contributions.

When a multiset lacks missing values, algorithms such as MCR-ALS (Multivariate Curve Resolution-Alternating Least Squares) work iteratively using a single factorization to obtain the basic bilinear model in equation 1:

$$\mathbf{D} = \mathbf{C}\mathbf{S}^{\mathrm{T}} + \mathbf{E} \qquad (1)$$

where **D** is the initial multiset data, **C** and $\mathbf{S}^{\mathrm{T}}$ are the single (or augmented) matrices, often related to concentration profiles and pure instrumental responses of the components of interest, respectively, and **E** is the matrix of model residuals.

However, there are situations where a multiset may contain full missing blocks of data, e.g., when some experiments could not be monitored by all available techniques, when images of different platforms do not cover identical sample areas or when in environmental campaigns, the values of some sets of parameters are systematically missing at a particular time period. In these scenarios, an adapted MCR-ALS variant was proposed to obtain the sought bilinear model, based on the combination of the results of multiple bilinear factorizations performed on complete multisets, which were defined as subsets of the total incomplete multiset structure[2-4]. Such a solution avoided the challenging task of data imputation when systematic patterns of missing values were present, but the strategy to combine the results of the multiple factorizations into a single bilinear model was not trivial.

The new approach proposed allows analyzing incomplete multisets by MCR-ALS with a single factorization and without the need of performing data imputation. Whereas the usual MCR-ALS optimization obtains **C** and $\mathbf{S}^T$ matrices in a single matrix operation step using **D** and the bilinear counterpart matrix, the new methodology is based on the fact that the **C** and $\mathbf{S}^T$ matrices can be also calculated in a least-squares manner one row at a time (eq. 2) and one column at a time (eq. 3), obeying the following formulations of the bilinear model:

$$\boldsymbol{d}(i,:) = \boldsymbol{c}(i,:)\mathbf{S}^T + \boldsymbol{e}(i,:) \qquad (2)$$

$$\boldsymbol{d}(:,j) = \mathbf{C}\boldsymbol{s}(:,j) + \boldsymbol{e}(:,j) \qquad (3)$$

Following these formulations and using only the non-missing values of every **D** row and the matching columns of $\mathbf{S}^T$, the **C** matrix can be obtained one row at a time (eq. 2). Likewise, with the non-missing values of every column of **D** and the matching rows of **C**, the full $\mathbf{S}^T$ matrix can be obtained one column at a time (eq. 3). Working in this way, no data imputation is needed and a single factorization is only required.

The approach is shown in simulated and real data sets of different typologies to show the advantages and limitations of the algorithm.

References

1. Tauler R, Maeder M, de Juan, A (**2020**). Multiset data analysis: Extended Multivariate Curve Resolution, in *Comprehensive Chemometrics*, S. Brown, R. Tauler, and R. Walczak, Eds., Elsevier, 2020, 2, pp 305-336.
2. Alier M, Tauler R (**2013**). Multivariate curve resolution of incomplete data multisets. *Chemom. Intell. Lab. Sys.*, **127**, 2013, 17-28.
3. Piqueras S, Bedia C, Beleites C, Krafft C, Popp J, Maeder M, Tauler R, de Juan A (**2018**). Handling different spatial resolutions in image fusion by multivariate curve resolution-alternating least squares for incomplete image multisets. *Anal. Chem.*, **90**, 6757-6765.
4. de Luca M, Ragno G, Ioele G, Tauler R (**2014**). Multivariate curve resolution of incomplete fused multiset data from chromatographic and spectrophotometric analyses for drug photostability studies. *Anal. Chim. Acta,* 837, 31-37.

An automated pre-processing framework for uneven-length multiphase batch processes

Francesco Sartori[1], Federico Zuecco[2], Pierantonio Facco[1], Fabrizio Bezzo[1], Massimiliano Barolo[1]

[1] CAPE-Lab – Computer Aided Process Laboratory, Department of Industrial Engineering, University of Padova, Via Marzolo 9, 35131 Padova (Italy)
[2] BASF Italia S.p.A., Via Pila 6/3, 40037 Pontecchio Marconi, BO (Italy)

Batch processes are ubiquitous in many industries producing low volumes of high added-value goods. They consist of cyclic repetitions of finite-length operations (framed in a recipe) performed in specialized process equipment, and are typically adopted when flexibility is required. Both common-cause and abnormal variability can affect these processes, ending up in batches having uneven length. Furthermore, some processes (called multiphase processes) consist in more than one operating stage. The availability of process data due to the Industry 4.0 framework boosted the development of data-driven models for tasks such as product quality prediction and product classification. These models are faster and less expensive to develop than first-principles ones. Among the empirical modelling methodologies, time-resolved methods are frequently used because they preserve and incorporate time information in the modelling step (Rendall *et al.*, 2019). However, in order to use this modelling approach with uneven-length batch processes, a synchronization step is required in order for all the batch data to be compared at the same degree of completion (García-Muñoz *et al.*, 2011). Several approaches have been proposed for batch synchronization, all of them requiring a certain degree of process knowledge to be applied. The indicator variable technique is one such methodology, which is widely adopted due to its effectiveness and simplicity. In multiphase batch processes, this technique can be applied for each phase separately, because a single variable indicating the degree of completion of the process across the whole batch is usually difficult to find (García-Muñoz et al., 2003).

In order to apply the indicator variable methodology, a phase partition must be carried out in advance for all batches. Data-driven techniques can be used for identifying the correct phase partition, for example, by detecting correlation changes among the measured variables along the batch profile (Guo & Jin, 2019). Both phase partition and batch synchronization have a strong impact on the performance of the final model.

However, both batch synchronization and phase partition are usually applied by adjusting their parameters manually.

In this study, an optimization framework is proposed, based on surrogate optimization, to identify the optimal parameters for phase partition and batch synchronization that guarantee that the phase partition and synchronization are applied in order to obtain the smallest prediction error in regression (or the maximum accuracy in classification). In order to implement this framework, a novel methodology for automatic identification of an appropriate indicator variable is proposed together with a novel phase partition methodology, which is robust to measurement noise. Two case studies, a simulated Penicillin fermentation process and an industrial batch process for the manufacturing of a specialty chemical, are used to illustrate the capabilities of the proposed approach with multiphase, uneven-length batch processes. In the first case study the proposed framework resulted in performance comparable to the state-of-the-art, but with a 94% reduction in computational time, while in the second case study the performance of the proposed framework was better than the state-of-the-art performance with a comparable computational time required.

References

García-Muñoz, S., Kourti, T., MacGregor, J. F., Mateos, A. G., & Murphy, G. (2003). Troubleshooting of an industrial batch process using multivariate methods. *Industrial and Engineering Chemistry Research*, *42*(15), 3592–3601.

García-Muñoz, S., Polizzi, M., Prpich, A., Strain, C., Lalonde, A., & Negron, V. (2011). Experiences in batch trajectory alignment for pharmaceutical process improvement through multivariate latent variable modelling. *Journal of Process Control*, *21*(10), 1370–1377.

Guo, R., & Jin, Y. (2019). Phase Identification and Online Monitoring for the Uneven Batch Processes. *IEEE Access*, *7*, 81351–81363.

Rendall, R., Chiang, L. H., & Reis, M. S. (2019). Data-driven methods for batch data analysis – A critical overview and mapping on the complexity scale. *Computers and Chemical Engineering*, *124*, 1–13.

BENEFITS OF MULTIVARIATE CURVE RESOLUTION METHODS TO ANALYZE LARGE-SCALE RAW TIME-DOMAIN NMR DATA

Silvia Mas Garcia[1,2], Jean-Michel Roger [1,2], Mireille Cambert [3], Ruzica Ferbus[3], Denis Lourdin[4], Corinne Rondeau-Mouro [3]

[1] ITAP, INRAE, Institut Agro, University Montpellier, 34196 Montpellier, France
[2]ChemHouse Research Group, 34196 Montpellier, France
[3]INRAE, UR1466 OPAALE, 17 Avenue de Cucillé, CS 64427, F-35044 Rennes, France
[4]INRAE, UR1268 Biopolymères Interactions Assemblages, 44300 Nantes, France
silvia.mas-garcia@inrae.fr

Time-domain nuclear magnetic resonance (TD-NMR) provides new opportunities for large-scale and non-destructive studies of water distribution and transfers in many biological systems. TD-NMR is based on longitudinal (T_1) and transverse (T_2) relaxation time measurements [1]. These parameters that measure the molecular dynamics through the magnetic properties of protons (hydrogens), give access to molecular information, and to their physico-chemical properties in the investigated systems. Due to relaxation time phenomena, the measured NMR signal consists of a sum of decreasing exponentials $e^{-t/\tau}$, where the distribution of τ values produces T_1 or T_2 spectra. Thus, the classical processing of such data consists of a Non-Negative Least Squares (NNLS) fitting procedure [2] or uses a numerical inversion of the Laplace transform [3]. Such signal processing task is known to be an ill-conditioned and ill-posed problem, resulting in a large number of solutions that small noise in the data can easily affect.

To deal with this limitation, this work aims at proposing the use of chemometrics methods as an alternative, in particular multi-curve resolution approaches [3]. When MCR is applied to TD-NMR data, the pure spectra are NMR signals that can be processed by an inverse Laplace transform without numerical instability problems. This provides reliable information on the compounds present in the system. Moreover, concentration profiles estimated by MCR allow highlighting temporal phenomena, if the system monitored is process-related, and/or increases the knowledge of spatial structures if the raw data are organised in images. The advantages of using chemometrics instead of classical processing methods will be shown through the study of changes in hydration properties of nineteen genotypes of Arabidopsis seeds during water imbibition [4] and the monitoring by Magnetic Resonance microimaging of a potato starch blend swelling process [5].

References

[1] Modern Magnetic Resonance, Graham A. Webb (Eds.), Springer International Publishing AG, London, 2017

[2] C.L. Lawson, R.J. Hanson, Solving Least Squares Problems, Prentice Hall, Englewood Cliffs, NJ, 1974.

[3] R.L. Parker, Y.Q. Song, Assigning uncertainties in the inversion of NMR relaxation data, J. Magn. Reson. 174 (2005) 314–324.

[4] S. Mas Garcia, J.-M. Roger, M. Cambert, C. Rondeau-Mouro, Untargeted analysis of TD-NMR signals using a multivariate curve resolution approach: Application to the water-imbibition kinetics of Arabidopsis seeds, Talanta. 233 (2021) 122525

[5] S. Mas Garcia, J.-M. Roger, R. Ferbus, D. Lourdin, C. Rondeau-Mouro. Monitoring of water sorption and swelling of potato starch-glycerol extruded blend by magnetic resonance imaging and multivariate curve resolution. Talanta *("in revision")*

SOIL SPECTROSCOPY: Use of the moment distance index for fine-tuning spectra acquisition and chemometric models improvement

I. Barra[1], L. Khiari[2], S. Haefele[3], R. Sakrabani[4], F. Kebede[1]

[1]*Center of Excellence in Soil and Fertilizer Research in Africa (CESFRA) Mohammed VI Polytechnic University (UM6P), Benguerir, Morocco*
[2]*Department of Soils and Food Engineering, Faculty of Agriculture and Food Sciences, Laval University, Quebec, Canada*
[3]*Department of Sustainable Agriculture Sciences, Rothamsted Research, Harpenden, UK*
[4]*School of Water, Energy and Environment, Cranfield University, UK*
E-mail: *Issam.barra@um6p.ma*

Vibrational spectroscopy such as Fourier-transform infrared (FTIR), has been used successfully for soil diagnosis owing to its low cost, minimal sample preparation, non-destructive nature. The present study aimed at facilitating the work of soil spectroscopist by optimizing one of the essential settings during the acquisition of FTIR spectra (viz. Scans number) using the standardized moment distance index (SMDI) as a metric that could trap the fine points of the curve and extract optimal spectral fingerprints of the sample. Furthermore, it can be used successfully to assess the spectra resemblance. The study revealed that beyond 50 scans the similarity of the acquisitions has been remarkably improved [1].

Subsequently, the effect of the number of scans on the predictive ability of the partial least squares regression models for the estimation of five selected soil properties (i.e., water pH, soil organic carbon, total nitrogen, cation exchange capacity and Olsen phosphorus) was assessed, and the results showed a general tendency in improving the correlation coefficient (R^2) as the number of scans increased from 10 to 80. In contrast, the cross-validation error RMSECV decreased with increasing scan number, reflecting an improvement of the predictive quality of the calibrated models with an increasing number of scans.

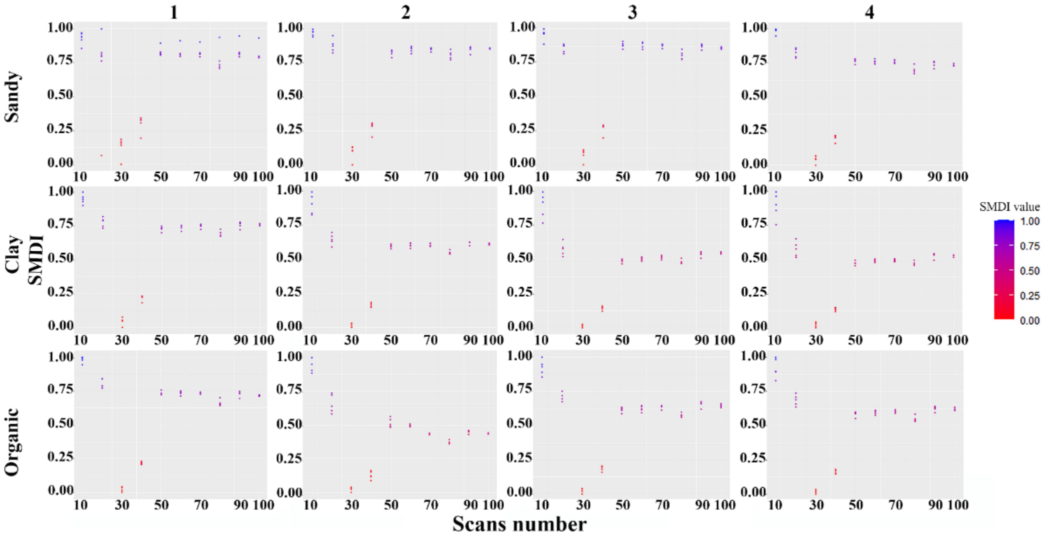


Figure 1 – Calculated standardized moment distance index (SMDI) for all spectra of the twelve soils samples resulting from different scan numbers. The points' (SMDI) rapprochement indicates the improvement of the spectra's similarity

The final finding of the present study showed that the number of scans has a remarkable effect on spectral stability and represents an important parameter to be taken into consideration when recording FTIR spectra of soil samples for the set-up of partial least squares models in soil spectroscopy. On the other, the chemometric methodology used as part of this study showed that the recorded spectra's quality (stability and repeatability) was improved by increasing the number of scans.

References

[1] I. Barra, L. Khiari, S.M. Haefele, R. Sakrabani, F. Kebede, optimizing setup of scan number in FTIR spectroscopy using the moment distance index and PLS regression: application to soil spectroscopy, Scientific Reports, 11 (2021) 13358.

Characterization and optimization of a novel UV-C LED aerodynamic device for airborne microbe viability abatement

Stefano Fornasaro[1], Sabrina Semeraro[2], Anastasia Serena Gaetano[1], Sabina Licen[1], Rita De Zorzi[1], Rosario Russo[4], Livio Tenze[4], Giorgio Vinciguerra[4], Carlo Poloni[3], Pierluigi Barbieri[1,2]

[1]University of Trieste, Department of Chemical and Pharmaceutical Sciences, via L. Giorgieri 1, 34127 Trieste;
[2]INSTM National Interuniversity Consortium of Materials Science and Technology, Via G. Giusti 9, 50121 Firenze, Italy;
[3]University of Trieste, Department of engineering and architecture, via A. Valerio 6 34127 Trieste;
[4]Esteco spa, Galleria Padriciano, 99/building b, 34149 Trieste
sfornasaro@units.it (corresponding author)

The spread of Covid-19 has intensified studies related to airborne infectious disease transmission in indoor environments, highlighting the critical need for cost-efficient and effective air disinfection solutions. The aim of the present work is the development of a complete framework based on designed experiments for exploring and optimizing the bioaerosol removal and inactivation efficiency of a novel air disinfection device that combines the aerodynamic effect of a three-dimensional vortex structure with UV-C radiation provided by commercially-available UV-C light-emitting diodes (UV-C-LEDs). The system was designed and tested to locally maintain a high radiation intensity suitable for bioaerosol disinfection. A controlled experimental laboratory model of bioaerosol aerosolization was set up using an impinger medical vibrating nebulizer, a cylindrical chamber for bioaerosol travel, and an SKC biosampler[1]. A non-pathogenic strain of *E. coli* (BL21-DE3), was used as a model of airborne bacteria, in accordance with the ISO 15714:2019 standard. The inactivation efficiency was assessed based on the enumeration of the colonies originating from viable *E. coli*. Interactions between analytical factors and their optimal levels were investigated using sequential D-optimal designs adapted to domain constraints and previous simulations of the aerodynamic performance of the device. Six variables (concentration of aerosolized bacteria; aerosol particle size; volumetric airflow; LED power; presence of reflective material; single or double filtration unit) were considered as factors in the optimization process. According to the results, the effects of particle size and concentration of aerosolized bacteria were negligible. Moreover, response surfaces allowed for the identification of the ideal working conditions to maximize the efficiency of the device even under high airflow, an essential requirement for the device's future exploitation in real-world settings.

References

Zupin L, Licen S, Milani M, Clemente L, Martello L, Semeraro S, Fontana F, Ruscio M, Miani A, Crovella S, Barbieri P (**2021**) Evaluation of Residual Infectivity after SARS-CoV-2 Aerosol Transmission in a Controlled Laboratory Setting. *Int. J. Environ. Res. Public Health*, **18**, 11172.

The power of DoE: how to solve complex analytical problems by multivariate optimization strategies

Barbara Benedetti

University of Genoa, Department of Chemistry and Industrial Chemistry, Via Dodecaneso 31, Genoa, 16146, Italy
barbara.benedetti@unige.it

Within the world of chemometrics, design of experiment (DoE) is a multivariate technique whose application should be desirable in any research study. It aims at rationally planning the experiments to gain the maximum information of a chemical system with the minimum experimental effort [1,2]. Despite being introduced almost a century ago (1935), it is still not widely employed by chemists and its usefulness continues to be underestimated. The application of one or more suitable experimental designs allows to perform a reasonable number of experiments to optimize even very complex analytical problems, avoiding the costly and tedious OVAT strategy. Indeed, through the application of response surface methodology, simple mathematical models can be built, relating the independent variables to the considered response(s), thus estimating effects and interactions. Depending on the research stage, as well as available time and costs, a "fit-for-purpose" experimental design can be chosen among screening designs (full factorial, Plackett-Burman) and response surface designs (central composite, Box-Behnken…).

Herein, several applications of DoE in both sample preparation and instrumental analysis are presented, also highlighting the importance of combining DoE with other simple chemometric tools, such as principal component analysis.

Generally, when several variables hypothetically influence more responses, a first screening followed by a response surface design is necessary. This approach was applied to maximize the QuEChERS-LC-MS/MS method accuracy in the determination of phytoestrogens in soy-based food. A Plackett-Burman design was used to explore the significance of 7 variables, 3 of which were then studied by a Box-Behnken design. Satisfactory mathematical models were obtained for the 10 responses involved, and a pareto-front strategy was employed to find the compromise for the optimal conditions.

The power of DoE is particularly evident when dealing with chromatographic separations. A central composite design allowed to understand the effect of flow, temperature and gradient velocity on peak resolution of 21 bioactive compounds, analysed by non-aqueous reverse phase chromatography. In less than one day of analysis (just seventeen runs) the data were collected and the best conditions to separate coeluted peaks were defined.

The extraction of 16 polycyclic aromatic hydrocarbons (PAHs) from complex matrices was optimized in three different applications, by employing suitable experimental designs and by combining DoE with PCA. Two studies regarded the magnetic molecular imprinted polymer dispersive extraction from seawater and plant extracts, respectively. In the first, 9 variables were studied by a screening design of 12 experiments and the correlation among responses was investigated by PCA. Two models (instead of 16) were built and revealed that only one variable significantly affected the recovery, with a consequent rapid optimization. In the second study, the D-optimal design was employed, defining the most informative experiments to explore the effect of 4 variables on the PAHs recovery (with 18 experiments). Finally, a recent work regarded the optimization of a SPME method for the PAHs analysis in plant material, for which two sequential central composite designs were employed, since the first gave indications on a better region of the experimental domain to explore. In both cases, the models were built after an exploratory data analysis by PCA and the optimal conditions were rapidly identified.

All these case studies demonstrate that DoE represents the winning strategy to face complex chemical problems and solve them in a costly- and timely-effective manner.

References

1. Leardi, R. Experimental design in chemistry: A tutorial. *Anal. Chim. Acta* **652**, 161–172 (2009).
2. Benedetti, B., Caponigro, V. & Ardini, F. Experimental Design Step by Step: A Practical Guide for Beginners. *Crit. Rev. Anal. Chem.* **52**, 1015–1028 (2022).

Trade off between space exploration and information maximization in experimental design

Francesca Cenci,[a] Arun Pankajakshan,[b] Federico Galvanin,[b]
Pierantonio Facco[a]

[a] CAPE-Lab – Computer-Aided Process Engineering Laboratory, Department of Industrial Engineering, University of Padova, via Marzolo 9, 35131 Padova, Italy
[b] University College London, Department of Chemical Engineering, Torrington Place, London WC1E 7JE, United Kingdom
E-mail of corresponding author: pierantonio.facco@unipd.it

Identification of a system from observed data relies on the quality of the data generated by the system itself. It is therefore crucial to design experiments that are informative for the purpose of system identification. A well-established method to design experiments is factorial Design of Experiments (DoE; Montgomery, 2013), which requires relatively small information on the system beforehand and is effective in exploring the entire design space. Moreover, factorial DoE allows to identify factors with significant impact on the response variable and to build response surfaces suitable for optimization (Politis et al., 2017). However, factorial DoE relates response variables to input factors through linear regression models that may become inadequate for highly nonlinear processes (Barz at al., 2022). Moreover, detailed first-principles models to describe chemical, physical and biological phenomena occurring in the system are more robust particularly in unexplored regions of the design space. A method that exploits process understanding to design experiments is the so-called Model-Based Design of Experiments (MBDoE; Espie and Macchietto, 1989). It uses model equations and parameters guesses to calculate the optimal experimental conditions for a specific purpose in system identification. For instance, , MBDoE methods can be employed to design experiments to maximize parameters precision of the specific model structure (Franceschini and Macchietto, 2008). However, the resulting model is not necessarily characterized by a minimum prediction variance throughout the design space. Moreover, space exploration may be insufficient if too many replicates of optimal experiments are selected, which may worsen model prediction variance in the unexplored regions. We tackle these issues by proposing a novel MBDoE method, *G-map eMBDoE* (G-map explorative MBDoE), that aims at increasing space exploration, while guaranteeing a significant reduction of model prediction variance. The exploration is enhanced by the use of mapping of G-optimality (Kiefer and Wolfowitz, 1960), which estimates model prediction variance at a given experimental condition. We have tested *G-map eMBDoE* on two simulated processes: (i) a system described by nonlinear algebraic model and (ii) a system described by differential and algebraic (DAE) model. The latter simulates the fermentation of Baker's yeast through a Monod-type kinetic model for biomass growth and substrate consumption (Chen and Asprey, 2003). The results from *G-map eMBDoE* is compared to that of conventional MBDoE and full factorial DoE: *G-map eMBDoE* explores regions of the design space that are neglected by MBDoE and samples highly informative experiments that are not considered by the factorial DoE. Overall, the novel method has the best performance in terms of precisely estimating model parameters within the experimental budget and minimizing the model prediction variance in the whole design space. This suggests that *G-map eMBDoE* is highly efficient as the best trade-off approach for space exploration and information maximization with the minimum number of experiments required.

References

Barz, T., Kager, J., Herwig, C., Neubauer, P., Mariano Bournazou, M. N. C., Galvanin, F., 2022. Chapter 11 - Characterization of reactions and growth in automated continuous flow and bioreactor platforms, From linear DoE to model-based approaches. In book: Simulation and Optimization in Process Engineering, Elsevier, pages 273-319.

Chen, B. H., Asprey, S. P. (2003). On the design of optimally informative dynamic experiments for model discrimination in multiresponse nonlinear situations. *Ind. Eng. Chem. Res.*, ***42*(7),** 1379–1390

Espie, D., Macchietto, S. (1989). The Optimal Design of Dynamic Experiments. *AIChE J.*, **35**, 223-229

Franceschini, G., Macchietto, S. (2008). Model-based design of experiments for parameter precision: State of the art. *Chem. Eng. Sci.,* **63(19)**, 4846–4872.

J. Kiefer, Wolfowitz, J. (1960), The equivalence of two extremum problems, *Canad. J. Math.*, **12**, 363-366.

Montgomery DC. (2013). Design and analysis of experiments. 8th edition. New York: John Wiley Sons, Inc.

Politis, S. N., Colombo, P., Colombo, G., Rekkas, D. M. (2017). Design of experiments (DoE) in pharmaceutical development. *Drug Dev. Ind. Pharm.,* **43(6)**, 889–901.

On the use of retrospective DOE for process optimization from historical data in Industry 4.0

García-Carrión, Sergio; Borràs-Ferrís, Joan; Ferrer, Alberto

Multivariate Statistical Engineering Group, Department of Applied Statistics and Operational Research, and Quality, Universitat Politècnica de València (UPV), Camino de Vera s/n, Valencia, 46022, Spain
sergarc6@doctor.upv.es

Model-based process optimization requires a causal model that relates changes in input variables (X) to those in the process outputs (Y). Data-driven models are commonly used to obtain causal models since deterministic models based on first principles, although always desirable, are unapproachable in most cases in practice (Liu & MacGregor, 2005). However, independent variation in the input variables is required to ensure causality when using data-driven approaches (Box, Hunter & Hunter, 2005).

Design of Experiments (DoE) is the key tool for obtaining empirical causal models, but its use is usually not feasible in practice in the Industry 4.0 era due to the impractically extensive experimentation that would be required as a consequence of the generally high number of potential factors, and the complex aliasing structure among them, which may prevent some independent variations (Ferrer, 2021).

Nevertheless, most companies in Industry 4.0 have huge amounts of historical data that may present some unplanned excitations. The problem with these data is that variations in the input variables (X) are usually not independent, and therefore they do not meet the requirement stated above. Consequently, predictive models such as linear regression (LR) (Draper & Smith, 1998) and machine learning (ML) techniques (Hastie, Tibshirani & Friedman, 2009; Provost & Fawcett, 2013) cannot be used for extracting causal models from raw historical data for process optimization (MacGregor, 2018).

In this work, a retrospective DoE methodology is proposed to exploit this type of data (i.e. data from daily production) for optimization purposes. It consists of retrospectively fitting one (or more) experimental designs to available data and later carrying out the analysis in the usual way (Loy et al., 2002; Wold et al., 2004). The possibility of eventual not available (i.e. missing) treatment conditions was also addressed. Moreover, a metaanalysis can be performed in those cases where more than one experimental design was retrospectively fitted, which allows to check the consistency of the conclusions.

Several datasets from real industrial processes were used to test the proposed methodology, identifying its advantages and disadvantages. Our approach is expected to be useful in extracting useful information from historical databases, gaining knowledge about the process and guiding subsequent real experimentation, all with significant cost savings.

References

Box, G. E.; Hunter, W. H.; & Hunter, S. (**2005**) Statistics for experimenters: Design, Innovation, and Discovery. *John Wiley and sons*.

Draper, N. R.; Smith, H. (**1998**) Applied regression analysis. *John Wiley & Sons*.

Ferrer, A. (**2021**) Multivariate six sigma: A key improvement strategy in industry 4.0. *Quality Engineering*, **33(4)**, 758-763.

Hastie, T.; Tibshirani, R.; Friedman, J. (**2009**). The elements of statistical learning: data mining, inference, and prediction. *Springer Science & Business Media.*

Liu, J. J.; MacGregor, J. (2005). Modeling and optimization of product appearance: application to injection-molded plastic panels. *Industrial & engineering chemistry research*, **44(13)**, 4687-4696.

Loy, C.; Goh, T. N.; Xie, M. (**2002**) Retrospective factorial fitting and reverse design of experiments. *Total Quality Management*, **13(5)**, 589-602.

MacGregor, J. (**2018**). Empirical Models for Analyzing BIG Data. What's the Difference?. *In 2018 Spring Meeting and 14th Global Congress on Process Safety*. AIChE.

Provost, F.; Fawcett, T. (**2013**) Data Science for Business: What you need to know about data mining and data-analytic thinking. *O'Reilly Media, Inc*.

Wold, S.; Josefson, M.; Gottfries, J.; Linusson, A (**2004**). The utility of multivariate design in PLS modeling. *Journal of chemometrics*, **18(3-4)**, 156-165.

QUALITY BY DESIGN: PRINCIPLE AND CONSTRUCTION OF PROVEN ACCEPTABLE INDEPENDANT RANGE

Michelle Sergent[1], Magalie Claeys-Bruno[2], Sacha Guedj[3]

[1] Institut Méditerranéen de Biodiversité et d'Ecologie marine et continentale, Aix-Marseille Université, UMR CNRS IRD Avignon Université, Site de l'Etoile, Marseille, France
E-mail : michelle.sergent@univ-amu.fr
[2] Institut Méditerranéen de Biodiversité et d'Ecologie marine et continentale, Aix-Marseille Université, UMR CNRS IRD Avignon Université, Site de l'Etoile, Marseille, France
E-mail : m.claeys-bruno@univ-amu.fr
[3] Azurad, 33 Place des Héros, 13013 Marseille
E-mail : contact@azurad.fr

In pharmaceutical studies, the Quality by Design (QbD) [1-6] approach is increasingly being implemented to improve product development. Product quality is tested at each step of the manufacturing process, allowing a better process understanding and a better risk management, thus avoiding manufacturing defects. A key element of QbD is the construction of a Design Space (DS), i.e., a region in which the probability that the specifications on the output parameters will be met, is imposed. Among the various possible construction methods, Designs of Experiments (DoE), and more precisely Response Surface Methodology [7], represent a perfectly adapted tool [8]. Nevertheless, the DS obtained may have any geometrical shape; consequently, the acceptable variation range of an input may depend on the value of other inputs (green arrows – Fig.1). However, the experimenters would like to directly know the variation range of each input so that their variation domains are independent. In this context, we developed a method to determine the "Proven Acceptable **Independent** Range" (PAIR) (purple rectangle – Fig.1). It consists of looking for all the hyper polyhedra included in the multidimensional DS and selecting a hyper polyhedron according to various strategies.

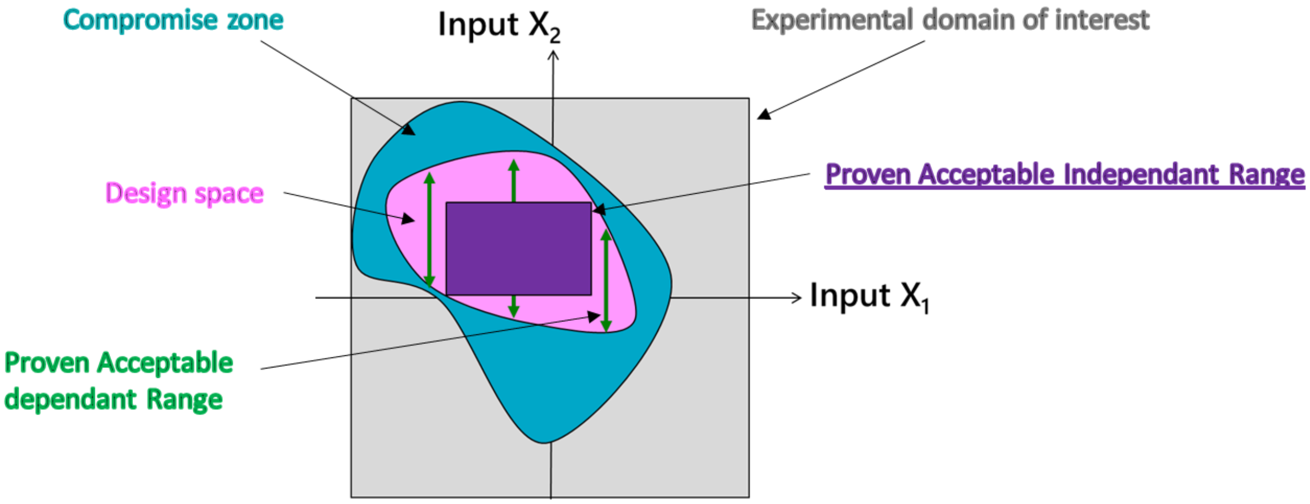


Fig.1 : QbD concepts in cubic experimental domain

We present an algorithm which looks for all possible hyper polyhedra within the DS, and determines the PAIR using various strategies and considering different selection criteria [9].

We apply this method on different DoE cases to compare these different strategies and to show the major benefit of this approach for pharmaceutical and chemical studies.

References

1. Woodcock J, (**2004**) The concept of pharmaceutical quality. *American Pharmaceutical Review*, 1–3.
2. Sangshetti JN, Deshpande M, Zaheer Z, Shinde D B, Arote R, (**2017**) Quality by design approach: Regulatory need. *Arabian Journal of Chemistry*, **10**, S3412-S3425.
3. Bohlin M, Jones H, Black S, (**2009**) New developments in scale-up and QbD to ensure control over product quality. *American Pharmaceutical Review*, **12**, 52–57.
4. Juran JM, (**1992**) Juran on quality by design: the new steps for planning quality into goods and services. *New York: The Free Press.*
5. Mhatre R, Rathore AS, (**2009**) Quality by design: An overview of the basic concepts. in: Quality by design for biopharmaceuticals. *Wiley, New York*, 1–8.
6. International Conference on harmonization of technical requirements for registration of pharmaceuticals for human use, ICH harmonized tripartite guideline. **(2009)** *Pharmaceutical Quality System.*: Q8(R2).
7. Sarabia LA, Ortiz MC, (**2009**) Response Surface Methodology. *In Comprehensive Chemometrics T.R. Brown S, Walczak R Editor. Elsevier: Oxford.*, 251-300.
8. Manzon D, Claeys-Bruno M, Declomesnil S, Carite C , Sergent M, **(2020)** Quality by design: comparison of Design Space construction methods in the case of design of experiments, *Chemometrics and Intelligent Laboratory Systems*, **200,** 104002.
9. Manzon D, Ghattas B, Claeys-Bruno M, Declomesnil S, Carite C, Sergent M, (**2023**) Looking for a hyper polyhedron within the multidimensional space of Design Space from the results of Designs of Experiments, *Chemometrics and Intelligent Laboratory Systems,* **232**, 104712

DOE for sustainability: bioplastic films composition and preparation optimization for lab-scale production

Lisa Rita Magnaghi[1,2], Marta Guembe Garcia[1,3], Camilla Zanoni[1], Giorgia Condrò[4], Paola Perugini[4], Giancarla Alberti[1], Raffaela Biesuz[1,2]

(1)Università degli Studi di Pavia, Dipartimento di Chimica, Viale Taramelli 12, Pavia, 27100, Italy
(2) INSTM, Unità di Ricerca di Pavia, Via G. Giusti 9, Firenze, 50121, Italy
(3) Universidad de Burgos, Facultad de Ciencias, Departamento de Química, Plaza de Misael Bañuelos s/n, Burgos, 09001, Spain
(4) Università degli Studi di Pavia, Dipartimento di Farmacia, Viale Taramelli 12, Pavia, 27100, Italy
lisarita.magnaghi@unipv.it

The development of biodegradable or biocompatible materials has represented an emergent trend in recent years, and among all the field of food packaging is one of the most involved in this trend. In fact, contaminations associated with synthetic food packaging and current increasing concerns related to the negative environmental impact of plastic packaging materials derived from petroleum have driven significant interest from both academia and industry in natural and biodegradable materials [1]. Most of the papers recently published on this topic have in common the choice of starch, as the main component, glycerol and citric acid, as additives, and casting method, as deposition strategy. Different experimental procedures are proposed, but a rational investigation of the influent factors and an optimization of reactive amounts and ratios have never been performed [2].

As far as bioplastics composition is concerned, starting from a deep literature survey, we identified starch, glycerol and carboxymethyl cellulose (CMC) as the main components, and we further expanded our investigation by testing starches from four different sources and CMC with different molecular weight and degree of functionalization.

The very preliminary experiments were focused on defining the pseudo-components domain, expressed as the fraction of starch, glycerol and CMC, in which a suitable material was always obtained, applying the most common casting deposition procedure. Then Mixture and Process Design was exploited to optimize the mixture composition, together with the type of starch and CMC. Once identified model equation and candidate points, D-Optimal Design was employed to select the most informative submatrices of training experiments. Bioplastic films were then tested in terms of tensile properties, hydrophilicity and water solubility, and the results of these characterizations were exploited as DOE responses.

Moving to preparation optimization, thousands of procedures can be found in the literature that slightly differ from each other in terms of order of addition of the reagents, time and temperature required for starch gelatinization and film formation, stirring rate, type of mold and other procedural aspects. Given that none of these procedures derives from a suitable multivariate optimization, we first applied Plackett-Burman Design to identify the influent factors, with a particular focus on the final films' homogeneity and absence of bubbles, and finally, we defined the optimized preparation procedure.

In conclusion, by combining different DOE techniques, we could define a reliable methodology to produce a library of bioplastic films with different properties, among which the best candidate can be selected, depending on the final application of interest. Among all, an interesting option could be the possibility of exploiting these materials for food packaging, either as inert packaging material or as support for active or intelligent packaging devices. [4]

References

[1] Medina-Jaramillo C, Ochoa-Yepes O, Bernal C, Famá L (**2017**) Active and smart biodegradable packaging based on starch and natural extracts. *Carbohydrate Polymers,* **176**, 187-194

[2] La Fuente Arias CI, Ka-neiwa Kubo MT, Tadini CC, Duarte Augusto PE (**2021**) Bio-based multilayer films: A review of the principal methods of production and challenges. *Critical Reviews in Food Science and Nutrition*

[3] Lundstedt T, Seifert E, Abramo L, Thelin B, Nystrom A, Pettersen J, Bergmanä Bergman R (**1998**) Experimental Design and Optimization. *Chemometrics and Intelligent Laboratory Systems*, **42**, 3–40

[4] Dodero A, Escher A, Bertucci S, Castellano M, Lova P. Intelligent Packaging for Real-Time Monitoring of Food-Quality: Current and Future Developments (**2021**) *Applied Science*, **11**, 3532

Development of eco-efficient cements by a multi-step Experimental Design

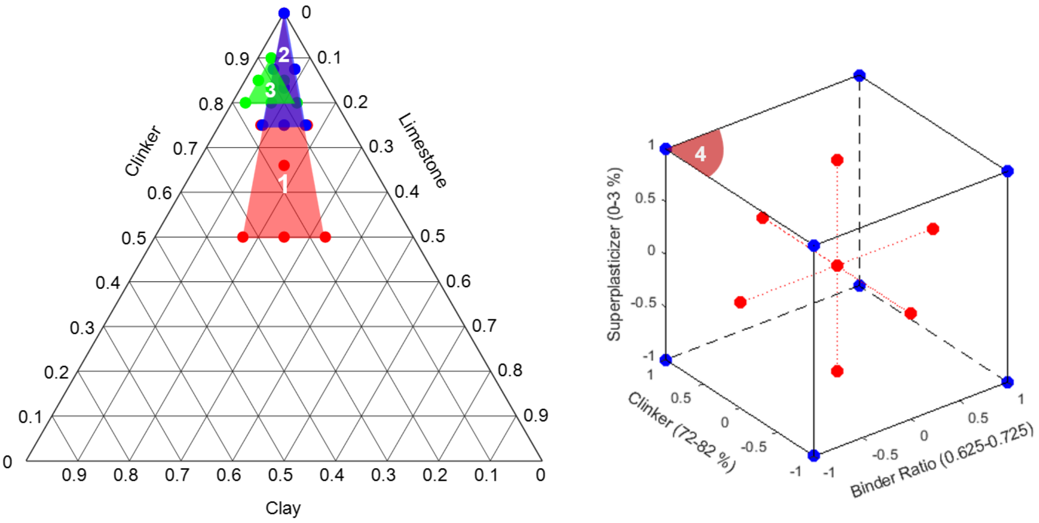
Emanuele Farinini[a], **Riccardo Leardi**[a], **Guilherme Ascensão**[b]

Addresses: [a] Department of Pharmacy, University of Genova, Viale Cembrano 4, I-16148 Genova, Italy.
[b] RISCO, Department of Civil Engineering, University of Aveiro, 3810-193 Aveiro, Portugal
E-mail: emanuele.farinini@edu.unige.it

The global production of cement has grown very rapidly in recent years, and after fossil fuels and land-use change, it is the third-largest source of anthropogenic emissions of carbon dioxide. The main source of pollution is limestone decarbonation during the production of clinker, the main component of cement [1]. Supplementary cementitious materials (SCMs) are used to partially replace clinker in cement production. However, conventional SCMs (slags and coal fly ashes) have limited incorporation thresholds and cannot meet the increasing global demand for cementitious materials. Limestone Calcined Clay Cements (LC3) are being explored as a sustainable alternative. In LC3, clinker is partially replaced by calcined clay and finely milled limestone. The pozzolanic reaction delivered by calcined clay and the filler effect of limestone are enhanced by secondary reactions between these SCMs, enabling higher clinker replacement levels without sacrificing performance. Early studies indicate that the process is robust and tolerant to variations in the composition of the feedstock materials. The development of tailor-made LC3 cements and building materials with *ad hoc* characteristics for rehabilitation purposes is a novelty of this work [2].

This study reports a multi-step mixture-process (D-optimal) design with relational constraints (Aggregate/Binder=3 and Water/Binder=0.50 according the EN 196-1 standards [3]; 0.5 ≤ Clay/Limestone ≤ 2) studied in pseudo-components of a LC3 formulation focusing on the reduction of clinker without compromising technical characteristics to reduce associated environmental impacts and production costs. Alongside, for the first three steps gypsum was considered as process variable evaluated at three levels. To assess the different products seven responses were considered, mostly physico-mechanical properties tests after 28 days: 1-2) Apparent density at 7 and 28 days; 3) Modulus of elasticity; 4) Water absorption; 5) Open porosity; 6) Flexural strength; 7) Compressive strength; 3,6, 7 to be maximized; 4,5 to be minimized. After a three-step experimental design, to find the best compromise a Pareto Front approach has been followed which allowed to find the best formulation. Comparing the predicted results to the “reference” experiments (conditions: 100% clinker, 5% gypsum), a product characterized by improved quality especially for what regards the compressive strength (+22 %) can be obtained together with a reduction of 20% in clinker. A central composite design is planned with the aim of further reducing the amount of clinker by adding the superplasticizer as a new factor.

Figure 1. Multi-step mixture-process design within relational constraints followed by CCD optimization.



Acknowledgment

G.A work was supported by the Foundation for Science and Technology (FCT)—Aveiro Research Center for Risks and Sustainability in Construction (RISCO), Universidade de Aveiro, Portugal [FCT/UIDB/ECI/04450/2020]. This publication reflects only the authors’ views, exempting the funding agency from any liability.

References

[1] Boden T A, Marland, G, & Andres R J (2009). Global, regional, and national fossil-fuel CO2 emissions. Carbon dioxide information analysis center, Oak ridge national laboratory, *US department of energy*, Oak Ridge, Tenn., USA doi, **10**.

[2] Sivakrishna A, Adesina A, Awoyera P O, Kumar K R (2020). Green concrete: A review of recent developments. *Materials Today: Proceedings*, **27**, 54-58.

[3] EN 196 - 1, Test Methods for Cement, CEN TC 51.

CORRECTION OF THE EFFECT OF SEASONALITY THROUGH A DATA STANDARDIZATION STRATEGY IN THE AUTHENTICATION OF EXTRA VIRGIN OLIVE OIL

Glòria Rovira, **M. Pilar Callao, Itziar Ruisánchez**

Rovira i Virgili University, Chemometrics, Qualimetric and Nanosensors Group,
Department of Analytical and Organic Chemistry, Marcel·lí Domingo s/n, Tarragona,
43007, Spain
gloria.rovira@urv.cat

Extra Virgin Olive Oil (EVOO) is one of the principal ingredients of the Mediterranean diet. Because of its nutritional and sensory properties and the multiple health benefits, its use has increased significantly worldwide. As a consequence of this use, fraud also increased. The most studied types of fraud are framed, mainly, in two large blocks: adulteration with other cheaper oils and lack of authenticity due to fraudulent labeling [1,2].

To authenticate Arbequina EVOO from two Protected Catalan Designations of Origin (Les Garrigues and Siurana, Catalonia, Spain) we have developed and validated a two-class PLS-DA model from fluorescence spectra of the 2019 harvest samples.

In the case of the authentication of seasonal products (or their derivatives), the future samples may contain sources of variability not considered when the model was developed. Therefore, maintaining the performance parameters and making the model applicable to seasonal variability remains a challenge in the authentication field.

The correction of the differences between the measurements (spectra) performed under different conditions is done by chemometric methods known as standardization or transfer methods. Transfer methods calculate a transform function that corrects the measurements done in the new conditions according to the training set used to build the multivariate model (primary conditions). As a result, the model can be used to predict samples measured in the new conditions. This strategy aims to correct the difference between measurements performed under different conditions to make a model transferable and useful at lower experimental costs.

In this work, a strategy to account for the season variability is proposed by adapting the Piecewise Direct Standardization (PDS) method. PSD transforms the response from second/new conditions to correspond to the response from the primary conditions from a subset of samples measured under both conditions [3]. Alternatives for the determination of the transfer function have been proposed when it is not possible to have a set of samples measured in both conditions [4]. We propose to obtain it from the average of the spectra in the first conditions concerning the mean of a set of spectra in the second conditions.

Samples from 2020, 2021, and 2022 harvests have been measured with the same instrument. The corresponding transfer functions have been established and the spectra have been corrected so that they can be predicted using the PLS-DA model established from 2019 harvest samples.

References

[1] Yan J, Erasmus SW, Toro MA, Huang H, van Ruth SM (**2020**) Food fraud: Assessing fraud vulnerability in the extra virgin olive oil supply chain. *Food Control,* **111**, 107081.

[2] Callao MP, Ruisánchez I (**2018**) An overview of multivariate qualitative methods for food fraud detection. *Food Control*, **86**, 283-293.

[3] Wang Y, Veltkamp DJ, Kowalski BR **(1991)** Multivariate instrument standardization. *Analytical Chemistry,* **63**, 2750-2756.

[4] Mishra P, Nikzad-Langerodi R, Marini F, Roger JM, Biancolillo A, Rutledge DN, Lohumi S (**2021**) Are standard sample measurements still needed to transfer multivariate calibration models between near-infrared spectrometers? The answer is not always. *TrAC Trends in Analytical Chemistry,* **143**, 116331.

Comparison of VASCA, GASCA and its combined version G-VASCA

José Camacho, **Edoardo Saccenti**, **Age Smilde**

Addresses: University or Company, Department, Street, City, Postcode, Country
E-mail (corresponding author)

Variable-selection ANOVA Simultaneous Component Analysis (VASCA, Camacho et al., 2023) is an extension of the popular ANOVA Simultaneous Component Analysis (ASCA) that applies variable-selection using a step-up procedure. VASCA has proved to be more statistically powerful than both ASCA and the False Discovery Rate (FDR, Benjamini y Hochberg, 1995) for the analysis of highly multivariate responses (e.g., omics data) in designed experiments. The Group-wise ASCA (Saccenti et al., 2018) is a sparse ASCA implementation with the aim of obtaining models that are easier to interpret. GASCA is based on the concept of group-wise sparsity introduced in the Group-wise Principal Components Analysis (GPCA, Camacho et al., 2017) where structure to impose sparsity is defined in terms of groups of correlated variables found in the correlation matrices calculated from the effect matrices. Both VASCA and GASCA represent complementary approaches to identify relevant biomarkers in omics data. While VASCA is a more solid inferential tool, it may select more biomarkers than the relevant ones, and GASCA provides an elegant solution to this problem. We investigate this by comparing VASCA, GASCA and its combination (G-VASCA).

References

Camacho J, Vitale R, Morales-Jiménez D, Gómez-Llorente C. Variable-selection ANOVA Simultaneous Component Analysis (VASCA). Bioinformatics. 2023 Jan 1;39(1):btac795. doi: 10.1093/bioinformatics/btac795. PMID: 36495189; PMCID: PMC9825241.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. B (Methodological), 57, 289–300.

Saccenti E, Smilde AK, Camacho J. Group-wise ANOVA simultaneous component analysis for designed omics experiments. Metabolomics. 2018;14(6):73. doi: 10.1007/s11306-018-1369-1. Epub 2018 May 21. PMID: 29861703; PMCID: PMC5962647.

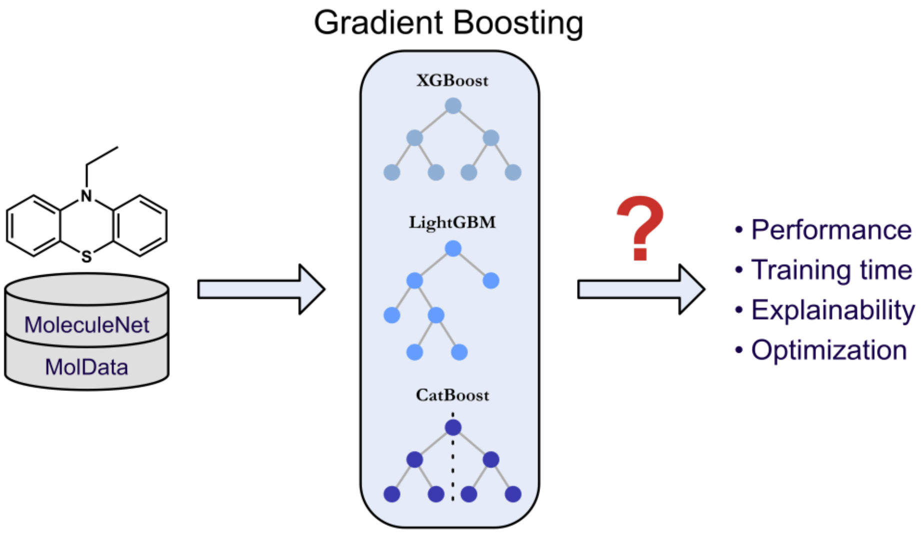
Camacho, J., Rodríguez-Gómez, R., Saccenti, E. Group-wise Principal Component Analysis for Exploratory Data Analysis. Journal of Computational and Graphical Statistics, 2017, 26 (3): 501-512.

Practical guidelines for the use of gradient boosting for molecular property prediction

Davide Boldini,[†] Francesca Grisoni,[‡] Lukas Friedrich[§], Daniel Kuhn[§]
and Stephan A. Sieber[†]

[†]Technical University Munich, Department of Bioscience, Ernst-Otto-Fischer Straße 8, Garching bei München, 85748, Germany
‡Eindhoven University of Technology, Department of Biomedical Engineering, Groene Loper 15, Eindhoven, 5600, Netherlands
§Merck Healthcare KGaA, Frankfurter Straße 250, Darmstadt, 64293, Germany
stephan.sieber@tum.de

Decision tree ensembles are among the most robust, powerful, and computationally efficient machine learning approaches for modeling quantitative structure-activity relationships (QSAR). Among them, gradient boosting (GBM) has recently drawn significant attention due to its performance in data science competitions,[1] virtual screening campaigns[2] and bioactivity prediction benchmarks.[3,4] However, there are several variations of the GBM algorithm, of which XGBoost,[5] LightGBM,[6] and CatBoost[7] are the most popular. Our study provides the first comprehensive comparison of these approaches for QSAR. To this end, we trained 157,590 gradient boosting models and we evaluated them on 11 datasets with 89 different endpoints, totaling 1.4 million compounds.[8,9] Our results show that XGBoost generally achieves the best classification performance, outperforming the other alternatives in terms of ROC-AUC and PR-AUC by 5% on average. LightGBM requires the least training time, especially on larger datasets, where it can be trained up to 100 times faster than XGBoost and CatBoost. In terms of explainability, the GBM models surprisingly rank molecular features differently in terms of importance for predictive performance, highlighting the need for expert knowledge to evaluate data-driven explanations of bioactivity. Finally, we also assessed the importance of tuning each hyperparameter when employing GBM models in order to maximize their predictive performance while reducing the number of optimization iterations. Our analysis shows that the impact of each hyperparameter varies greatly across datasets and that it is critical to optimize as many hyperparameters as possible to maximize predictive performance. In summary, our study provides the first set of guidelines for cheminformaticians to select, train, and optimize gradient-boosting models for virtual screening and QSAR.



References

1. Bentéjac, C.; Csörgő, A.; Martínez-Muñoz, G. A Comparative Analysis of Gradient Boosting Algorithms. *Artif. Intell. Rev.* **2021**, *54* (3), 1937–1967.
2. Zhu, Y.; Brettin, T.; Evrard, Y. A.; Partin, A.; Xia, F.; Shukla, M.; Yoo, H.; Doroshow, J. H.; Stevens, R. L. Ensemble Transfer Learning for the Prediction of Anti-Cancer Drug Response. *Sci. Rep.* **2020**, *10* (1), 18040.
3. Jiang, D.; Wu, Z.; Hsieh, C.-Y.; Chen, G.; Liao, B.; Wang, Z.; Shen, C.; Cao, D.; Wu, J.; Hou, T. Could Graph Neural Networks Learn Better Molecular Representation for Drug Discovery? A Comparison Study of Descriptor-Based and Graph-Based Models. *J. Cheminformatics* **2021**, *13* (1), 12
4. van Tilborg, D.; Alenicheva, A.; Grisoni, F. Exposing the Limitations of Molecular Machine Learning with Activity Cliffs. *J. Chem. Inf. Model.* **2022**, *62* (23), 5938–5951.
5. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; ACM: San Francisco California USA, 2016; pp 785–794.
6. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.-Y. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc., 2017; Vol. 30.
7. Prokhorenkova, L.; Gusev, G.; Vorobev, A.; Dorogush, A. V.; Gulin, A. CatBoost: Unbiased Boosting with Categorical Features, arXiv:1706.09516.
8. Wu, Z.; Ramsundar, B.; Feinberg, E. N.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; Pande, V. MoleculeNet: A Benchmark for Molecular Machine Learning. *Chem. Sci.* **2018**, *9* (2), 513–530.
9. Keshavarzi Arshadi, A.; Salem, M.; Firouzbakht, A.; Yuan, J. S. MolData, a Molecular Benchmark for Disease and Target Based Machine Learning. *J. Cheminformatics* **2022**, *14* (1), 10.

A multiple explainable machine learning framework for discriminating selective and unselective cannabinoid binders.

Gambacorta,N. [a]; **Ciriaco,F.** [b]; **Bajorath, J.** [c] ; **Nicolotti,O.** [a]

[a] *Università degli Studi di Bari "Aldo Moro", Dipartimento di Farmacia Scienze del Farmaco, Via E.Orabona, 4,I-70125 Bari, Italy;*
[b] *Università degli Studi di Bari "Aldo Moro", Dipartimento di Chimica, Università degli Studi di Bari "Aldo Moro", Via E. Orabona, 4, I-70125 Bari,Italy;*
c Rheinische Friedrich-Wilhelms-Universität , Department of Life Science Informatics and Data Science, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, , Friedrich-Hirzebruch-Allee 5/6, D-53115 Bonn, Germany;
Nicola.gambacorta1@uniba.it

The endocannabinoid system, which includes CB1 and CB2 receptor isoforms is involved in the onset of several pathologies including neurodegeneration, cancer, neuropathic and inflammatory pain, obesity, and inflammatory bowel disease. Given the high similarity of CB1 and CB2 receptor isoforms, understanding what are the molecular determinants for finding selective ligands is very difficult. In the study presented herein, a machine learning approach has been employed to predict compounds with dual or single target CB1/CB2 activity[1]. For model derivation, CB1/CB2 compound activity data was extracted from the latest release of the ChEMBL database[2] and the core-substituent molecular fingerprint was used as descriptor[3]. Finally, Shapley values[4] were computed to identify features determining correct predictions and explain machine learning models. The newly developed models are expected to better support the identification of new active compounds for CB1 and/or CB2 or the repurposing of known drugs.

References

1. Feldmann C, Bajorath J. Differentiating Inhibitors of Closely Related Protein Kinases with Single- or Multi-Target Activity via Explainable Machine Learning and Feature Analysis. Biomolecules 2022;12:557
2. Mendez D, Gaulton A, Bento AP, Chambers J, De Veij M, Félix E, et al. ChEMBL: towards direct deposition of bioassay data. Nucleic Acids Res 2019 ;47:D930–40.

3. Janela T, Takeuchi K, Bajorath J. Introducing a Chemically Intuitive Core-Substituent Fingerprint Designed to Explore Structural Requirements for Effective Similarity Searching and Machine Learning. Molecules 2022;27:2331.
4. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell 2020;2:56–67.

Machine Learning Models for Making Cocrystallization More Sustainable

R. Birolo[1], F. Bravetti[1], E. Alladio[1], A. Aramini[2], G. Bianchini[2], R. Gobetto[1], M. R. Chierotti[1]

[1]Department of Chemistry, University of Torino, Via P. Giuria 7, 10125 Torino, Italy

[2]Research and Early Development, Dompé Farmaceutici S.p.A., 67100 L'Aquila, Italy

E-mail: rebecca.birolo@unito.it

Due to their promising performance for data-driven prediction, data-driven machine learning (ML) methods have become increasingly popular in the chemical and pharmaceutical fields. Among the several application areas, algorithms such as Random Forest (RF), Deep Neural Networks (NN) and Support Vector Machines exhibit positive results in the context of cocrystal prediction.[1] Pharmaceutical cocrystals are crystalline materials composed of at least two molecules, *i.e.,* an active pharmaceutical ingredient (API) and a coformer, assembled by non-covalent forces. Cocrystallization is successfully applied to improve the physicochemical properties[2] of APIs, such as solubility, dissolution profile, pharmacokinetics, and stability.[3] However, choosing the ideal coformers is a challenging task in terms of time, effort and laboratory resources. Several computational tools and ML models were proposed to mitigate this problem. However, the challenge of achieving a robust and generalizable predictive method is still open.[4,5] In this study, we propose a new approach to quickly predict the formation of cocrystals, employing the PLS-DA, RF and NN algorithms. The models were based on a training set consisting of 543 cases of cocrystallization with both positive and negative experimental outcomes. At the same time, the features were specially selected from a variety of molecular descriptors to explain the phenomenon of cocrystallization. All the proposed ML models showed cross-validation accuracy higher than 80%. Furthermore, this approach was successfully applied to drive the cocrystallization experimental tests of 2-phenylpropionic acid, showcasing the high potential of the ML models in practice.

References

(1) Mswahili, M. E.; Lee, M.-J.; Martin, G. L.; Kim, J.; Kim, P.; Choi, G. J.; Jeong, Y.-S. Cocrystal Prediction Using Machine Learning Models and Descriptors. **2021**, 12.

(2) Cerreia Vioglio, P.; Chierotti, M. R.; Gobetto, R. Pharmaceutical Aspects of Salt and Cocrystal Forms of APIs and Characterization Challenges. *Adv. Drug Deliv. Rev.* **2017**, *117*, 86–110. https://doi.org/10.1016/j.addr.2017.07.001.

(3) Dai, X.-L.; Chen, J.-M.; Lu, T.-B. Pharmaceutical Cocrystallization: An Effective Approach to Modulate the Physicochemical Properties of Solid-State Drugs. *CrystEngComm* **2018**, *20* (36), 5292–5316. https://doi.org/10.1039/C8CE00707A.

(4) Hao, Y.; Hung, Y. C.; Shimoyama, Y. Investigating Spatial Charge Descriptors for Prediction of Cocrystal Formation Using Machine Learning Algorithms. *Cryst. Growth Des.* **2022**, *22* (11), 6608–6615. https://doi.org/10.1021/acs.cgd.2c00812.

(5) Devogelaer, J.; Meekes, H.; Tinnemans, P.; Vlieg, E.; Gelder, R. Co-crystal Prediction by Artificial Neural Networks**. *Angew. Chem. Int. Ed.* **2020**, *59* (48), 21711–21718. https://doi.org/10.1002/anie.202009467.

First evidence of Microplastics translocation in mussel tissues using Raman Imaging and Chemometrics

Julene Aramendia[1], Nerea García-Velasco[2], Jose Manuel Amigo[1,3], Urtzi Izagirre[2], Andreas Seifert[3,4], Laura Arévalo[4], Manu Soto[2] and Kepa Castro[1]

[1] IBeA Research Group, Analytical Chemistry department, Faculty of Science and Technology, University of the Basque Country UPV/EHU, E-48080, Bilbao, Basque Country, Spain

[2] Cell Biology in Environmental Toxicology (CBET) Research Group, Dept. Zoology and Animal Cell Biology, Faculty of Science and Technology and Research Centre for Experimental Marine Biology and Biotechnology PIE-UPV/EHU, University of the Basque Country UPV/EHU, E-48080, Bilbao, Basque Country, Spain

[3] IKERBASQUE, Basque Foundation for Science, Euskadi Plaza 5, 48009 Bilbao, Spain

[4] CIC nanoGUNE BRTA, Tolosa Hiribidea 76, 20018 San Sebastian, Spain

Presenting author: josemanuel.amigo@ehu.eus

Even though the importance of microplastics (MPs) in the environment, fauna and biota have been widely studied, by today it has not been clearly demonstrated that MPs of a certain size can be translocated via ingestion from the lumen of the digestive tract to tissues through epithelial cells of mussels. This fact is mostly due to inherent difficulties of sampling, measuring and analyzing; and the plausible spectral and matrix interferences during the analysis that hampers the reliability of any methodology.

This presentation will put forward an analytical methodology composed of a meticulous sample processing, accurate spectral imaging and a precise chemometric strategy to demonstrate unequivocally that MPs of a certain size can actually trespass the biological barriers (i.e. digestive epithelium) of mussels and being internalized by cells and tissues.

For this purpose, mussels under controlled laboratory conditions were exposed to 1 µm size polystyrene (PS) MPs. Following, confocal volumetric Raman imaging measurements were done (using a 532 nm excitation laser) using a 50x objective, and a scanning lateral resolution of 1 µm/pixel in the XYZ axes. The spectra were pre-processed, and a strategy combining Principal Component Analysis (PCA) and Multivariate Curve Resolution–Alternating Least Squares (MCR-ALS) was implemented.

Acknowledgements

This work was funded by Basque Government (KK 2021/00001 ELKARTEK 2021/2022); further financial support by grant CEX2020-001038-M funded by MCIN/AEI/ 10.13039/501100011033. J. Aramendia thanks the post-doctoral contract from the UPV/EHU

On determining the factor ambiguity for incomplete data sets

Martina Beese[1,2], Tomass Andersons[1], Mathias Sawall[1], Klaus Neymeyr[1,2]

[1]University of Rostock, Mathematical Institute, Ulmenstraße 69, Rostock, 18057, Germany
[2]Leibniz Institute for Catalysis, Numerical Analysis, Albert-Einstein-Straße 29A, Rostock, 18059, Germany
martina.beese@uni-rostock.de

When being faced with incomplete data sets in MCR analyses one may ignore the problematic parts or can estimate the missing data. However, ignoring some data typically increases the ambiguity of the possible pure components. A method for dealing with this problem has been proposed by Alier and Tauler [1]. Their method results in a single set of pure component profiles. Here, the goal is to obtain all feasible solutions without ignoring any data, but reconstructing data – thus keeping the ambiguity to a minimum.

The proposed method is applied to image data where a block of pixels suffers from (sensor) saturation and is therefore assumed to have missing values. For instance, a noise free data set is shown in Figure 1. Our method requires a good signal-to-noise ratio. We compare the resulting ambiguity for either the case where the affected pixels are completely ignored with a new local reconstruction approach. This uses bilinearity to predict missing parts of the data from bilinearity constraints, see Figure 1 (right) for one component of the spectral factor. As shown here, the use of this method can considerably improve the results for saturated data.

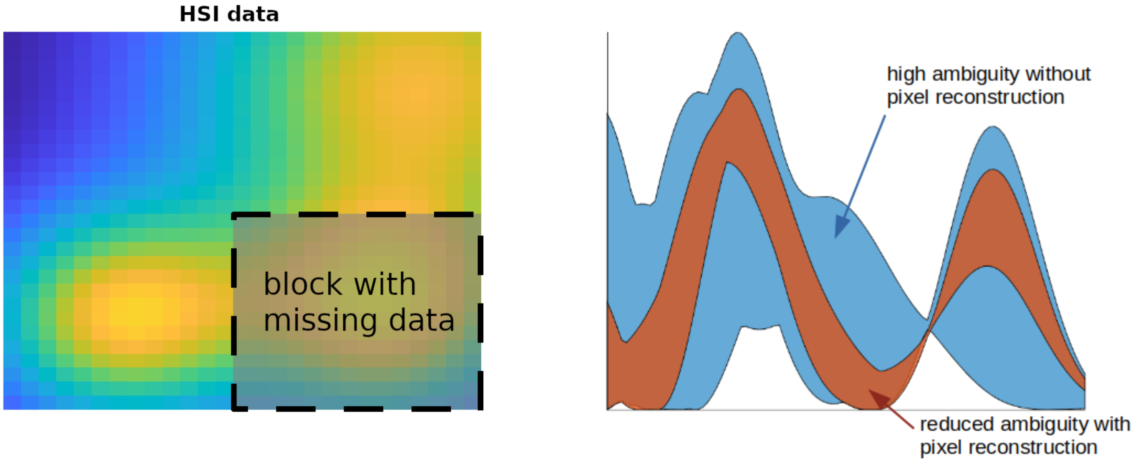


Figure 1: Benefits of pixel reconstruction

References

[1] Alier, M., Tauler, R. (2013). Multivariate curve resolution of incomplete data multisets. *Chemometrics and Intelligent Laboratory Systems*, *127*, 17-28.

[2] Beese, M., Andersons, T., Sawall, M., Neymeyr, K. (2023). Calculating the area of feasible solutions for incomplete data in MCR problems. *In preparation.*

Exploratory analysis of large hyperspectral datasets: coupling randomness and sparsity

Rosalba Calvini[1], Alessandro Ulrici[1], José Manuel Amigo[2,3]

[1] *Department of Life Sciences, University of Modena and Reggio Emilia, Pad. Besta, Via Amendola, 2, 42122, Reggio Emilia, Italy*
[2] *Ikerbasque, Basque Foundation for Sciences, María Díaz de Haro, 3, Bilbao, 48013, Spain*
[3] *Department of Analytical Chemistry, University of the Basque Country, Barrio Sarriena S/N, Leioa, 48940, Spain*
E-mail rosalba.calvini@unimore.it

Hyperspectral imaging (HSI) combines classical spectroscopic methods with imaging systems in order to obtain both spectral and spatial information from a sample. HSI images contain a wide amount of data since each image can contain up to thousands of spectra and, depending on the device, each spectrum can be composed of hundreds of wavelengths. On the one hand, this huge amount of data represents an advantage of HSI systems thanks to the possibility of obtaining a detailed chemical mapping of the analysed samples. On the other hand, this aspect is also known as the *curse* of dimensionality since the presence of a large amount of information in big data matrices generally makes their analysis time consuming and requiring a lot of random-access memory (RAM) and high computing power (Burger and Gowen, 2011).

Furthermore, in practical applications of HSI systems it is often necessary to simultaneously analyse many images altogether (e.g., in time series analysis) and in this scenario the problem of effectively extracting useful information from such data becomes even more relevant.

In order to overcome this issue, data reduction strategies can be applied allowing to analyse big data arrays with a reduced computation time without losing relevant analytical information.

Among the different data reduction strategies proposed, the present study focuses on the use of a robust randomization method based on local-rank approximations, which allows to construct an approximate matrix factorization that, using a small fraction of the information contained in the hyperspectral data matrix, can provide the same results (Halko et al., 2011).

The prosed approach was recently tested in the frame of exploratory image analysis using Principal Component Analysis with successful results (Cruz-Tirado et al., 2022). In this study, the same approach has been evaluated in combination with sparse PCA (sPCA), an extension of PCA with sparsity induced on the loading vectors to force uninformative variables to be equal to zero and increase model interpretability (Calvini et al., 2015).

More in detail, we evaluated the impact of different model parameters such as sPCA algorithm (Alternating Shrunken Least Squares or deflation-based), sparsity level, number of principal components, number of iterations and oversampling on the consistency between sPCA and reduced sPCA models as well as on computation time. The simulations were performed considering two datasets of hyperspectral images: a benchmark Raman hyperspectral of an oil-in-water emulsion and a large time series dataset of Near Infrared (NIR) hyperspectral images of bread samples.

References

Burger, J., Gowen, A. (**2011**). Data handling in hyperspectral image analysis. *Chemometrics and Intelligent Laboratory Systems*, **108(1)**, 13-22.

Calvini, R., Ulrici, A., Amigo, J. M. (**2015**). Practical comparison of sparse methods for classification of Arabica and Robusta coffee species using near infrared hyperspectral imaging. *Chemometrics and Intelligent Laboratory Systems*, **146**, 503-511.

Cruz-Tirado, J. P., Amigo, J. M., Barbin, D. F., Kucheryavskiy, S. (**2022**). Data reduction by randomization subsampling for the study of large hyperspectral datasets. *Analytica Chimica Acta*, **1209**, 339793.

Halko, N., Martinsson, P. G., Tropp, J. A. (**2011**). Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, **53(2)**, 217-288.

CHEMOMETRIC STRATEGIES FOR SPATIAL AND CHEMICAL ANALYSIS OF MALDI MASS SPECTROMETRY IMAGING (MSI) OF HETEROGENEOUS TISSUES. CASE STUDY: PAROTID TUMOUR

V. Caponigro[1], E. Salviati1, F. Marini[2], M. Grimaldi[1], A. M. D'Ursi[1], E. Sommella[1], P. Campiglia[1]

1 Department of Pharmacy, University of Salerno, I-84084 Fisciano, (SA), Italy

2 Dipartimento di Chimica, Sapienza Università di Roma, I-00185 Rome, Italy.

vcaponigro@unisa.it

MALDI mass spectrometry imaging technology (MALDI-MSI) combines chemical and spatial information. Indeed, the potential of MALDI-MSI lies in the spatial distribution of numerous biomolecules that can be used for both diagnostic and prognostic purposes, leading to its increasing use with effective benefits for early diagnosis (Mas et al.,2020; Meding et al. (2012); Ogrinc et al. (2022)). In addition to the benefits, the complexity of the data acquired with MALDI-MSI imaging can make data analysis challenging. Indeed, application to tissues with heterogeneous cellular composition, such as parotid gland tissue, can lead to two levels of complications. First, diagnosing parotid gland tumours and identifying malignant lesions, as well as distinguishing between different tumour types, can be challenging because tumour lesions can have a variety of histological appearances. Therefore, it is difficult to accurately identify and label each pixel to avoid errors in model training and analysis of results. Second, the enormous amount of data generated by this technique must be efficiently reduced to select the correct number of pixels and chemical information representing the tissue to generalise the changes associated with cancer cell signalling, mechanisms of disease development and tumour progression. In addition, to reduce the influence of the individual patient signature, an appropriate method for pre-processing the data must be used.

In this work, different chemometrics strategies are compared to define the optimal workflow for heterogeneous tissue analysis using a case study of cancer in parotid gland tissue. MATLAB R2022B was used to import and analyse the raw data set. Principal

component analysis (PCA) was first used for an exploratory study. Then, to find out certain differences between the pathological and healthy groups and compare the different strategies, Partial Least Squares Discriminant Analysis (PLS-DA) was performed. Each approach was examined at both the pixel and object level to facilitate interpretation. In both cases, random selection and the use of a mask based on histological information (after image registration) were compared. To maximise information related to cancer cell signalling, mechanisms of disease development and tumour progression, different pre-processing and normalisation methods were compared, including full range, PQN, MSROI (Pérez-Cova et al.,2021) and co-localisation approaches.

Acknowledgment

This work was funded by project PIR01_00032 BIO OPEN LAB BOL "CUP" J37E19000050007 to P. Campiglia.

References

Mas, S., Torro, A., Fernández, L., Bec, N., Gongora, C., Larroque, C., Martineau, P., de Juan, A., & Marco, S. (2020). MALDI imaging mass spectrometry and chemometric tools to discriminate highly similar colorectal cancer tissues. *Talanta*, 208, 120455.

Meding, S., Nitsche, U., Balluff, B., Elsner, M., Rauser, S., Schöne, C., Nipp, M., Maak, M., Feith, M., Ebert, M. P., Friess, H., Langer, R., Höfler, H., Zitzelsberger, H., Rosenberg, R., & Walch, A. (2012). Tumor Classification of Six Common Cancer Types Based on Proteomic Profiling by MALDI Imaging. *Journal of Proteome Research*, 11(3), 1996–2003.

Ogrinc, N., Attencourt, C., Colin, E., Boudahi, A., Tebbakha, R., Salzet, M., Testelin, S., Dakpé, S., & Fournier, I. (2022). Mass Spectrometry-Based Differentiation of Oral Tongue Squamous Cell Carcinoma and Nontumor Regions with the SpiderMass Technology. *Frontiers in Oral Health*, 3, 1–11.

Pérez-Cova, M., Bedia, C., Stoll, D. R., Tauler, R., & Jaumot, J. (2021). MSroi: A pre-processing tool for mass spectrometry-based studies. *Chemometrics and Intelligent Laboratory Systems*, 215, 104333.

CALCULATING ESSENTIAL INFORMATION IN THE FOURIER DOMAIN TO ACCELERATE HYPERSPECTRAL IMAGING

Cyril Ruckebusch, **Laureen Coic, Raffaele Vitale**

Address: U. Lille, CNRS, LASIRE, Laboratoire Avancé de Spectroscopie pour les Interactions, la Réactivité et l'Environnement, Cité Scientifique, F-59000 Lille, France
Cyril.ruckebusch@univ-lille.fr

In the domain of multivariate curve resolution (MCR) of spectroscopic data, the selection of essential information is driven by the properties of linear spectral mixtures. In this regard, a very useful practical tool to select the most relevant spectral information before MCR analysis has recently been proposed, both yielding a significant data compression rate and boosting the overall data processing speed. However, this approach relies on the use of principal component analysis (PCA) to determine the convex hull that encapsulates the entire set of investigated data points in the space of its normalized principal component scores. Since PCA is a data-driven transform, the evaluation of the essential information in this case can, therefore, only be achieved after all the data have been acquired. In other words, all spectra must have been registered before one can decide on the essentiality of each of those spectra. This obviously hampers the evaluation of essential information on-the-fly, *i.e.*, while spectral data are being acquired. Nonetheless, not only factorial representations such as PCA but any linear transformation preserving the convex geometry of a dataset can be used to retrieve essential information. We thus propose to use Fast Fourier Transform to provide representations of spectroscopic signals characterized by sine and cosine functions of varying frequencies. These representations are called phasor and allow convex hull calculations to be performed directly in the space defined by the Fourier coefficients of a given harmonic. This constitutes an alternative approach to identify and recover essential information, and this approach has striking advantages for the sake of curve resolution. To illustrate them, we show the efficiency and robustness of this novel strategy through the analysis of Raman hyperspectral imaging data. We also propose a new approach for data acquisition based on selective sampling which is capable of guaranteeing a fifty-fold acceleration of the whole Raman imaging procedure.

References

Ghaffari, M.; Omidikia, N.; Ruckebusch, C. Essential Spectral Pixels for Multivariate Curve Resolution of Chemical Images. *Anal. Chem.* **2019**, *91* (17), 10943–10948.

Coic, L.; Sacré, P. Y.; Dispas, A.; De Bleye, C.; Fillet, M.; Ruckebusch, C.; Hubert, P.; Ziemons, É. Selection of Essential Spectra to Improve the Multivariate Curve Resolution of Minor Compounds in Complex Pharmaceutical Formulations. *Anal. Chim. Acta* **2022**, *1198*.

Ruckebusch, C.; Vitale, R.; Ghaffari, M.; Hugelier, S.; Omidikia, N. Perspective on Essential Information in Multivariate Curve Resolution. *TrAC - Trends in Analytical Chemistry*. Elsevier B.V. November 1, **2020**.

Characterization of pyrite weathering products by Raman hyperspectral imaging and chemometrics techniques

Enmanuel Cruz Muñoz[a*], Fabio Gosetti[a], Davide Ballabio[a], Sergio Andò[b], Olivia Gómez-Laserna[c], Jose Manuel Amigo[c,d], Eduardo Garzanti[b]

[a]Milano Chemometrics and QSAR Research Group. Department of Earth and Environmental Sciences, University of Milano-Bicocca, Piazza della Scienza 20126, Milano, Italy
[b]Laboratory for Provenance Studies. Department of Earth and Environmental Sciences, University of Milano-Bicocca, Piazza della Scienza 20126, Milano, Italy
[c]Department of Analytical Chemistry, University of the Basque Country, Barrio Sarriena S/N, 48940, Leioa, Spain
[d]IKERBASQUE, Basque Society for the Promotion of Science, Plaza Euskadi 5, 48009, Bilbao, Spain
[*]E-mail (e.cruzmunoz@campus.unimib.it)

The evaluation of chemical weathering processes along with the surface alteration sulfide minerals undergo after oxidation [1] is fundamental to assessing the effect of acid drainage problematic that causes low-pH effluents and the release of potentially toxic metals to the environment [2]. In this study we propose and apply a novel approach based on chemometrics and Hyperspectral Raman imaging for the analysis of chemically heterogeneous surfaces of weathered minerals. The approach was applied to a pyrite crystal from *Lugagnano Val d'Arda*, northern Italy.

This methodology includes data pre-processing and subsequent multivariate modelling with different approaches [3],[4]. Principal Component Analysis (PCA) has been initially used to evaluating data structure and to identifying Raman phases and vibrational modes related to

minor and major weathering features. Afterwards, Multivariate Curve Resolution-alternating least squares (MCR-ALS) and K-means clustering were used to identify the specific chemical components and end-members spectra, as well as to group similar pixels of low percentage of variability in the hyperspectral image, respectively. Particularly, MCR-ALS was applied also considering the availability of a Raman library, proposing a procedure for the selection of the best candidates to be used as reference spectra.

In the analyzed sample (Figure 1), three major components were found corresponding with pyrite vibrational modes and alteration products, whereas four additional spectral signatures associated with the presence of sulfates were identified after the analysis of the minor components fraction.

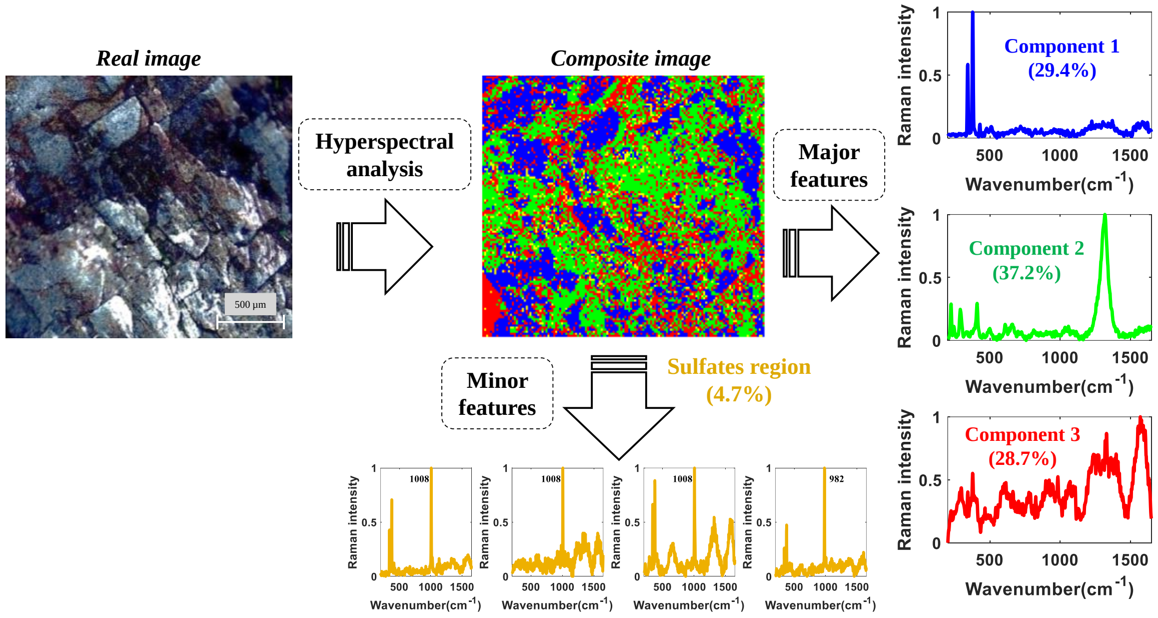


Figure 1.- Graphical abstract of the approach accomplished. Pure spectral signatures of components 1 and 2 correspond to pyrite and hematite, respectively, while an undefined spectral signature related to the presence of an amorphous patina was identified for component 3.

References

[1] A.P. Chandra; A.R. Gerson, (**2010**). The mechanisms of pyrite oxidation and leaching: A fundamental perspective, Surface Science Reports. **65**, 293–315.

[2] M.C. Moncur; J.L. Jambor et al., (**2009**). Mine drainage from the weathering of sulfide minerals and magnetite, *Applied Geochemistry*. **24**, 2362–2373.

[3] J.M. Amigo, (**2019**). Hyperspectral Imaging, *Data Handling in Science and Technology*. **32**,0-630

[4] A.de Juan, R.Tauler (**2021**). Multivariate Curve Resolution: 50 years addressing the mixture analysis problem – A review, *Analytica Chimica Acta*. **1145**, 59-78

Towards automatic clustering in MALDI imaging for a real unsupervised exploration of complex biological tissues

Ruggero Guerrini[1], **Nina Ogrinc**[2], **Ludovic Duponchel**[1]

[1]Univ. Lille, CNRS, UMR 8516 – LASIRE – Laboratoire de Spectroscopie pour Les Interactions, La Réactivité et L'Environnement, Lille, 59000, France
[2]University of Lille, Inserm, CHU Lille, U1192 - Protéomique Réponse Inflammatoire Spectrométrie de Masse – PRISM, Lille, France
ruggero.guerrini.etu@univ-lille.fr
ludovic.duponchel@univ-lille.fr
nina.ogrinc@univ-lille.fr

MALDI imaging has a high potential for the exploration of biological samples, such as tissue sections, as it provides a complete information on molecular composition and distribution. It is then possible to observe the spatial distribution of proteins, peptides, lipids, etc. Due to the huge amount of information contained in such datasets, the use of chemometric methods is important in order to be able to proceed with a deeper analysis.

The most used chemometric tool in this domain is certainly clustering. Clustering, applied to an image, allows the visualisation of *k* regions with a similar molecular composition, thus potentially allowing for different region of the tissue to be delineated.

Out of the many segmentation algorithms, many researchers in in the MALDI imaging field use the the so-called bisecting *k-means*[1,2], defined as a hierarchical application of *k-means* clustering. For each step of the dataset partioning, a cluster is systematically divided in two sub-clusters, with the *k-means* algorithm, until the desired number of clusters is reached. It is above all the observation of the clustering maps that guides the researcher's choice in this segmentation, which is not an optimal solution and potentially proposes a biased exploration of the data set. Firstly, from our point of view, this systematic division of clusters into two subsets is not justified. Secondly, the selection of the number of clusters should be done automatically without any intervention from the researcher who will always tend to look for a partitioning he would like to observe. This inevitably leads to a biased exploration of the dataset. The work we propose will be based on two aspects. Firstly, we will keep a hierarchical approach in the search for clusters as it is a way to handle two critical situations, namely unbalanced clusters but also non-spherical cluster structures. Second, we will focus on the use of statistical indices to propose an optimal number of clusters without human intervention[1]. This presentation will be an opportunity to present an algorithm that we have developed called *embedded k-means*[3] which takes up the two previous points. Thus in a first step, a *k-means* analysis is performed on the MALDI dataset considering a variable number of clusters. The use of statistical indices then allows in this step to find an optimal number of clusters. In a second step, a new *k-means* analysis is performed on each cluster found previously. The same statistical indices are of course used to find an optimal number of sub-clusters for each of them.

Our approach and bisecting *k-means*[1,2] will compared in the context of clustering a MALDI dataset acquired from a rat brain tissue section. It is indeed an ideal sample (often considered as a gold standard in MALDI imaging) as it has a well-defined structure very detailed in the atlases (ex: http://atlas.brain-map.org/).

References

(1) Steinbach, M.; Karypis, G.; Kumar, V. A Comparison of Document Clustering Techniques. **2000**.

(2) Trede, D.; Schiffler, S.; Becker, M.; Wirtz, S.; Steinhorst, K.; Strehlow, J.; Aichler, M.; Kobarg, J. H.; Oetjen, J.; Dyatlov, A.; Heldmann, S.; Walch, A.; Thiele, H.; Maass, P.; Alexandrov, T. Exploring Three-Dimensional Matrix-Assisted Laser Desorption/Ionization Imaging Mass Spectrometry Data: Three-Dimensional Spatial Segmentation of Mouse Kidney. *Anal. Chem.* **2012**, *84* (14), 6079–6087. https://doi.org/10.1021/ac300673y.

(3) Nardecchia, A.; Fabre, C.; Cauzid, J.; Pelascini, F.; Motto-Ros, V.; Duponchel, L. Detection of Minor Compounds in Complex Mineral Samples from Millions of Spectra: A New Data Analysis Strategy in LIBS Imaging. *Anal. Chim. Acta* **2020**, *1114*, 66–73. https://doi.org/10.1016/j.aca.2020.04.005.

(4) Aichler, M. MALDI Imaging Mass Spectrometry: Current Frontiers and Perspectives in Pathology Research and Practice. *Lab. Invest.* **2015**, *95*, 10.

EXPLORING SEA BASS MULTIOMIC IMAGES FOR SPATIALLY RESOLVED TRANSCRIPTOMICS AND METABOLOMICS

Albert Menéndez-Pedriza[1], Mercedes Blázquez[2], Eduardo Chicano-Gálvez[3], Laia Navarro-Martín[1] and Joaquim Jaumot[1]

Addresses:
1. Institute of Environmental Assessment and Water Research – Spanish National Research Council (IDAEA-CSIC), Department of Environmental Chemistry, Jordi Girona 18-26, E08034. Barcelona, Spain.
2. Institute of Marine Sciences – Spanish National Research Council (ICM-CSIC), Passeig Maritim 37-49, E08003 Barcelona, Spain.
3. Maimonides Institute of Biomedical Research of Córdoba (IMIBIC), Mass Spectrometry and Molecular Imaging Unit (IMSMI), Avda Menéndez Pidal s/n, E14004 Córdoba, Spain
E-mail: joaquim.jaumot@idaea.csic.es

Omic studies traditionally focused on analyzing metabolites from homogenized samples, resulting in the loss of morphological information. However, new techniques like mass spectrometry imaging (MSI)[1], useful for proteomics and metabolomics studies, and spatial transcriptomics[2] (ST) have emerged, allowing for the acquisition of images while retaining morphological information. MSI generates an image containing a mass spectrum in each pixel whereas ST allows for the identification of gene expression information and its location at the image. These techniques provide both structural information that characterizes and identifies the compounds or genes found on the surface of the samples, and morphological information that represents where these compounds are located.

The latest developments involve the fusion of multiomic images, which can be achieved by generating transcriptomic, proteomic, and/or metabolomic images from consecutive tissue sections[3]. Targeted studies analyze pre-defined molecules, while untargeted methods consider the entire dataset, using multivariate analysis to reveal most of the biological information and generate new hypotheses.

In this study, we present chemometric strategies for analyzing spatial transcriptomics and metabolomics (lipidomics) images, with the goal of identifying spatial regions with similar behavior and determining the genes or chemical compounds responsible for the observed differentiation. To illustrate our approach, we consider several testicular sections from European sea bass testicular sections in which different developmental stages, according to the maturation level, could be identified.

Our chemometric workflow includes several key stages, such as data import, quality control, filtering, and dimensionality reduction. We apply chemometrics tools for resolution and clustering to identify regions of interest and their potential markers. Figure 1 shows distribution maps for spatial transcriptomics and lipidomics for two sea bass testicular sections, revealing distinct clusters characterized by specific molecular signatures.

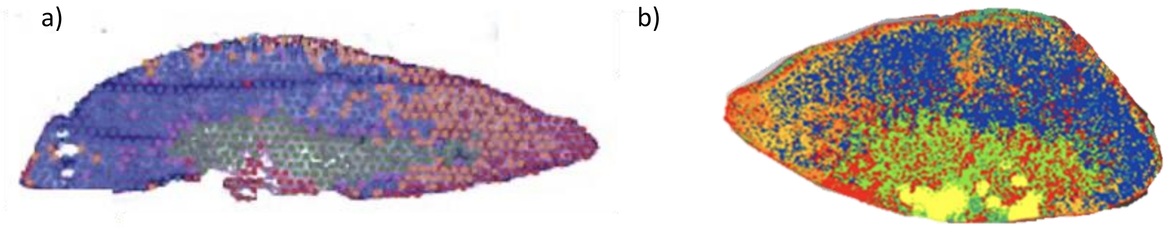


Figure 1. Clustering images obtained for a) spatial transcriptomics and b) spatial lipidomics

Finally, we aim to obtain a holistic picture by fusing the spatial multiomics images. However, the fusion process presents certain drawbacks from a chemometric perspective, such as ensuring correspondence between images (i.e., similarity between consecutive tissue slices) and addressing differences in spatial resolution used by the different spatial omics approaches. Therefore, an additional step of pixel size homogenization is required before the final chemometric analysis and the different considered approaches will be discussed.

Acknowledgements

Grant PID2021-122929OB-C33 funded by MCIN/AEI/10.13039/501100011033 and, as appropriate, by "ERDF A way of making Europe.

References

1. McDonnell, L.A., Heeren, R.M.A. Imaging mass spectrometry (**2007**) Mass Spectrometry Reviews, **26** (4), pp. 606-643.
2. Ståhl, P.L., Salmén, F., et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics (**2016**) Science, **353** (6294), pp. 78-82.
3. Porta Siegel, T., Hamm, G., et al. Mass Spectrometry Imaging and Integration with Other Imaging Modalities for Greater Molecular Understanding of Biological Tissues (**2018**) Molecular Imaging and Biology, **20** (6), pp. 888-901.

A SMARTPHONE-BASED DEVICE FOR IN-VINEYARD MONITORING OF RED GRAPES BY MULTIVARIATE ANALYSIS OF RGB IMAGES

Camilla Menozzi [1], **Rosalba Calvini** [1,2], **Veronica Ferrari** [1], **Giorgia Foca** [1,2], **Alessandro Ulrici** [1,2]

[1] *University of Modena and Reggio Emilia, Department of Life Sciences, Via Amendola, 2, Reggio Emilia, 42122, Italy*
[2] *University of Modena and Reggio Emilia, Interdepartmental Research Centre BIOGEST-SITEIA, Piazzale Europa 1, Reggio Emilia, 42122, Italy*
camilla.menozzi@unimore.it

Fast determination of food chemical composition is ever more performed by means of non-destructive analytical techniques, which offer to the opportunity to extract the information of interest by proper chemometric elaboration of signals/images that can be acquired using objects of common use, such as smartphones. Within this framework, we have recently developed a device in order to provide a real-time support in the monitoring of the anthocyanins content of red grapes over time, by multivariate analysis of RGB images of grape berries acquired using a smartphone camera. The device has been designed to be used directly in the vineyard thanks to its small size and to the possibility of acquiring geolocated images of the berries under controlled lighting conditions. The system essentially consists of a hardware part (3D-printed controlled lighting acquisition camera equipped with LED lights, colour reference, grape sample and smartphone housings) and a software interface (smartphone app coupled to a web interface). The core idea consists in relating the color information of the grape samples contained in RGB images with the chemical parameters of interest, i.e., total and extractable anthocyanins, using a data reduction algorithm and multivariate calibration. To this aim, we basically followed three steps: i) images standardization to account for possible variations of the lighting conditions; ii) conversion of the images into one-dimensional signals, named colourgrams (Calvini, 2020; Orlandi, 2018), which codify the colour-related properties within the image; iii) development and external validation of PLS regression models using the colourgrams to estimate the anthocyanins content. The device was tested on grape samples of three different varieties (*Ancellotta*, *Lambrusco Salamino* and *Sangiovese*), in the context of the research project "Technological innovation for an efficient harvest forecast - VITEVEN" funded by the Emilia-Romagna Region as part of the Rural Development Program 2014-2020. Samples were collected at different harvest times from veraison to maturity during vintages 2020 and 2021, and considering nine different vineyards on the whole.

References

Calvini R, Orlandi G, Foca G, Ulrici A (**2020**) Colourgrams GUI: A graphical user-friendly interface for the analysis of large datasets of RGB images. *Chemometrics and Intelligent Laboratory Systems,* **196**, 103915.

Orlandi G, Calvini R, Pigani L, Foca G, Vasile Simone G, Antonelli A, Ulrici A (**2018**) Electronic eye for the prediction of parameters related to grape ripening. *Talanta,* **186**, 381–388.

Flexible implementation of multilinearity constraints in Multivariate Curve Resolution Alternating Least Squares (MCR-ALS)

Xin Zhang[1], **Romà Tauler**[2*]

[1]Department of Chemistry, Capital Normal University, Beijing, 100048, China
[2]Institute of Environmental Assessment and Water Research (IDAEA-CSIC), Barcelona, 08043, Spain
[*]Corresponding author: Roma.Tauler@idaea.csic.es

The Multivariate Curve Resolution Alternating Least Squares (MCR-ALS) method[1-4] is used to analyze multiway data by imposing multilinearity constraints. However, strict adherence to these constraints can lead to unrealistic results when the structure of the multiway data does not match the requirements of the multilinear models. To address this issue, various methods have been proposed to relax the trilinear model constraints, such as the PARAFAC2[5], PARASIAS[6], and Direct Non-Trilinear Decomposition (DNTD)[7] methods.

This work demonstrates how the multilinearity constraint in MCR-ALS can be adapted to different data scenarios where some of the component profiles are shifted or have changed shapes in one of the data modes. This adaptation is particularly useful in applications such as GC and LC, and FIA. Instead of preprocessing the raw data to correct the shifts, the shift correction constraint can be incorporated into the ALS iterations.

In more complex scenarios where the profiles of different components are linearly dependent, or when the shapes of the profiles change, the application of the multilinear model becomes too demanding, and the recovery of the chemical profiles is not realistic. In such cases, multiway data analysis can still be performed effectively using mixed bilinear multilinear or simpler bilinear model factor decomposition methods, by augmenting the matrix data along one of the data modes (usually the spectral mode). Although the factor decomposition is no longer unique, the application of MCR-ALS with mixed bilinear multilinear or only bilinear is still possible, yielding realistic and meaningful results, especially if additional constraints such as local rank, selectivity, correlation, or unimodality are imposed. Results from different data sets will be presented to demonstrate the efficiency and flexibility of implementing the multilinearity constraint in MCR-ALS in these different data scenarios.

References

1. Multivariate curve resolution applied to second order data, Tauler R. Chemomet. and Intell. Lab. Syst., 1995, 30, 133-146; 18. https://doi.org/10.1016/0169-7439(95)00047-X
2. Multivariate curve resolution of multiway data using the multilinearity constraint Tauler R., **J.Chemometrics**, 2020, e3279, https://doi.org/10.1002/cem.3279
3. Flexible Implementation of the trilinearity Constraint in Multivariate Cureve resolution Alternating least squares (MCR-ALS) of Chromatographic and Other type of Data. Zhang X.and Tauler R.. **Molecules**, 2022, 27, 2338, https://doi.org/10.3390/molecules27072338
4. A model for simultaneous evaluation of NO2, O3 and PM10 pollution in urban and rural areas: handling incomplete data sets with multivariate curve resolution analysis Gorrochategui E.,. Hernandez I. and Tauler R.. **Atmospheric Chemistry and Physics**, egusphere-2022-117 https://doi.org/10.5194/egusphere-2022-117
5. Parafac2—Part, A direct fitting algorithm for the PARAFAC2 model, H.A. Kiers, J.M. Ten Berge, R. Bro, **J. Chemometr**. 1999, 13 (3-4) 275–294. https://doi.org/10.1002/(SICI)1099-128X(199905/08)13:3/4<275::AID-CEM543>3.0.CO;2-B
6. PARASIAS: A new method for analyzing higher-order tensors with shifting profiles
Yu H. Bro R., Gallagher N.B. **Analytica Chimica Acta** 2023, 1238, 339848, https://doi.org/10.1016/j.aca.2022.339848
7.Direct non-trilinear decomposition for analyzing high-dimensional data with imperfect trilinearity, Zhang J., Guo C., Cai W., Shao X. Chemometrics and Intelligent Laboratory Systems, 210 (2021) 104244. https://doi.org/10.1016/j.chemolab.2021.104244

A new strategy of analysis for relating datasets

Benoît Jaillais, El Mostafa Qannari, Benjamin Mahieu

Oniris, INRAE, StatSC,44300, Nantes, France
elmostafa.qannari@oniris-nantes.fr

We introduce a new strategy of analysis for relating data sets. This encompasses the case of a response dataset to be predicted by one or several datasets X and the case of path modeling. We discuss the advantages of this strategy of analysis over alternative methods, particularly those methods pertaining to PLS regression. Illustrations and comparison of methods based on real datasets are presented.

All the datasets considered throughout this presentation are supposed to be centered. Standardization and pre-scaling may also be applied.

It is of paramount importance to bear in mind two pre-requisites. Firstly, given a matrix, $\boldsymbol{X}$, the matrix $\boldsymbol{XX}^T$ may be seen as a pseudo-projection upon the space spanned by the columns of $\boldsymbol{X}$. Secondly, given a variable, ξ, that we shall refer to as auxiliary variable, we define a latent variable in the X-space as $\boldsymbol{XX}^T\xi$. Throughout this presentation, we shall assume that ξ is of unit length. Obviously, the vector of loadings-weights associated with the latent variable $\boldsymbol{XX}^T\xi$ is $\boldsymbol{X}^T\xi$. This vector may be standardized to unit length for an interpretation purpose.

Relationships between two datasets X and Y.

Let us consider two dataset $\boldsymbol{X}$ and $\boldsymbol{Y}$. We aim at predicting $\boldsymbol{Y}$ from $\boldsymbol{X}$. We consider a latent variable, $\boldsymbol{YY}^T\xi$, in the Y-space. The (pseudo-) projection of this latent variable upon the X-space is given by $\boldsymbol{XX}^T\boldsymbol{YY}^T\xi$. We seek ξ so as to maximize $cov(\boldsymbol{YY}^T\xi, \boldsymbol{XX}^T\boldsymbol{YY}^T\xi)$, where $cov()$ stands for the covariance. The rationale behind this maximization problem is intuitively appealing since it bears resemblance to the rationale behind linear regression and more generally canonical correlation where we maximize the correlation between (latent) variables and their projection upon the X-space. The solution to the previous problem is given by setting ξ to the eigenvector of $\boldsymbol{YY}^T\boldsymbol{XX}^T\boldsymbol{YY}^T$ associated with the largest eigenvalue. Alternatively, the problem could be solved by means of an iterative algorithm that runs as follows: (*0*) select a random vector ξ and standardize it to unit length; (*i*) compute $\boldsymbol{YY}^T\xi$; (*ii*) 'project' this vector upon the X-space yielding $\boldsymbol{XX}^T\boldsymbol{YY}^T\xi$; (*iii*) 'project' back this vector upon $\boldsymbol{Y}$, yielding $\boldsymbol{YY^TXX^TYY^T}\xi$; (*iv*) This new vector is standardized and used as an update of ξ. (*v*) the process starting from step (*i*) is iterated until convergence.

Subsequent latent variables can be found after deflation. Several interesting properties of the strategies of analysis will be shown.

Multiblock setting.

The extension of the previous analysis to a multiblock setting where we aim at predicting a dataset $\boldsymbol{Y}$ from several datasets $\boldsymbol{X_1}$, $\boldsymbol{X_2}$,…, $\boldsymbol{X_m}$ is straightforward: *(i)* consider a latent variable $\boldsymbol{YY^T}\xi$; *(ii)* this variable is 'projected' upon the X-spaces, yielding $\boldsymbol{X_kX_k^TYY^T}\xi$ ($k=1 \ldots m$); (iii) the auxiliary variable, ξ, is sought so as to maximize the sum of the covariances between $\boldsymbol{YY^T}\xi$ and their (pseudo) projections $\boldsymbol{X_kX_k^TYY^T}\xi$.

Path modeling.

The case of path modeling is of particular interest and the strategy of analysis proposed herein presents advantages over well-established alternatives. For the sake of simplicity, consider the case of three datasets $\boldsymbol{X_1}$, $\boldsymbol{X_2}$, and $\boldsymbol{X_3}$. Assume that these datasets are linked by the path diagram depicted in figure 1. As customary in path modeling, we associate to each dataset a latent variable. A notable feature of the approach adopted herein is that all these latent variables are associated with the same auxiliary variable, ξ. Another feature is that the auxiliary variable is determined taking account of the direction of causality between the datasets.

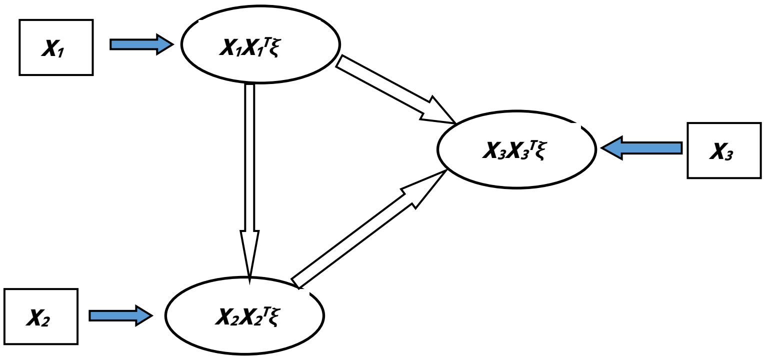


Figure 1. Path diagram showing the causal effects of $\boldsymbol{X_1}$ upon $\boldsymbol{X_2}$ and $\boldsymbol{X_3}$ and of $\boldsymbol{X_2}$ upon $\boldsymbol{X_3}$. It also shows the latent variables associated to $\boldsymbol{X_1}$, $\boldsymbol{X_2}$ and $\boldsymbol{X_3}$. These latent variables are defined by the

STUDY OF THE CORRECTION OF SPECTRAL DATA OBTAINED BY SORS USING CHEMOMETRIC BLIND SIGNAL SEPARATION (BSS)

A. Arroyo-Cerezo, A.M. Jiménez-Carvelo, M. Medina-García, L. Cuadros-Rodríguez.

Address: University of Granada, Department of Analytical Chemistry, Fuentenueva s/n,
Granada, 18071, Spain
arroyoc@ugr.es

Spatially offset Raman spectroscopy (SORS) is an analytical technique with characteristic of acquiring the Raman signal at a point shifted from the point of incidence of the excitation laser. Thus, it collects the signal from deeper layers of the analyzed material, avoiding the contribution of the surface layers. This technique therefore becomes very relevant for non-invasive control, for example, in the food industry, since food could be analyzed through packaging in a fast and non-destructive way.

Currently some portable instruments are available, such as the Vaya Raman model (Agilent). This equipment collects two Raman signals in a single SORS-type measuring run: (i) one acquired at the same point of incidence of the laser [ZERO] and (ii) another acquired at a point shifted with respect to the point of incidence of the laser [OFFSET] (Figure 1). The ZERO measurement consists of the Raman signal in a conventional way, and therefore has a higher contribution from the surface layers of the measured material. While the OFFSET measurement allows the laser to travel deeper and collect the Raman signal from deep layers of the measured material. However, OFFSET spectrum also contains some contribution from the surface layers, as the collected signal, although it travels in depth, then returns to pass through these layers (surface) and then the equipment collects the signal. This process is carried out by the equipment's software, which works as a black box and does not allow the user to control the conditions under which a 0-1 normalized FINAL spectrum is reached.

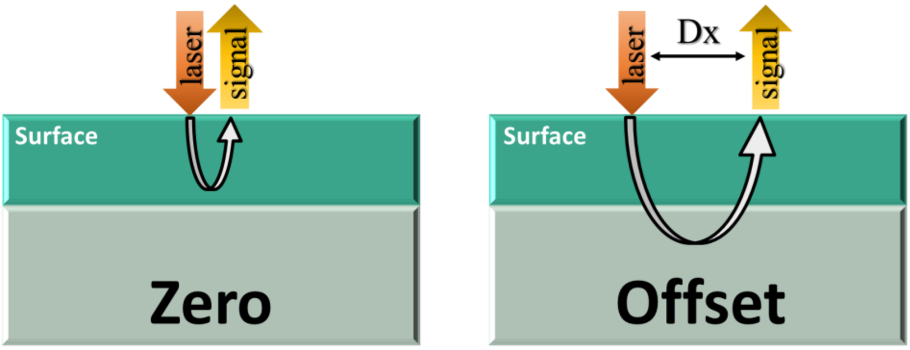


Figure 1. Raman signals collected in a single SORS-type measuring run: ZERO and OFFSET.

Therefore, to acquire the "pure" Raman target spectrum of the measured material inside of, for example, a food inside its original packaging, the correction of the OFFSET spectrum is a critical step, to remove any contribution from the surface layers that has been collected in this second measurement. This could not be performed by the preprocessing methods that are used as a preliminary step for the application of chemometric methods.

This communication presents the development of a methodology to correct the spectra obtained with Vaya Raman equipment (Agilent) of 4 materials analyzed through 2 types of packaging (polypropylene and polyethylene terephthalate). To achieve this objective, two types of analysis have been used: Independent Component Analysis (ICA) and Multivariate Curve Resolution (MCR). The resulting spectra for four chemical compounds (sucrose, sodium sulfate anhydrous, ethanol 96% and glycerol) after chemometric separation have been compared by similarity analysis with both the FINAL equipment spectra and the target "pure" Raman spectra, this last one retrieved from an available spectral database.

References

1. Mosca, S., Conti, C., Stone, N., Matousek, P. (**2021**) Spatially offset Raman spectroscopy. *Nature Reviews Methods Primers*, **1**, 21.
2. Arroyo-Cerezo, A., Jimenez-Carvelo, A. M., González-Casado, A., Koidis, A., Cuadros-Rodríguez, L. (**2021**) Deep (offset) non-invasive Raman spectroscopy for the evaluation of food and beverages–A review. *LWT*, **149**, 111822.
3. Liu, Z., Huang, M., Zhu, Q., Qin, J., Kim, M. S. (**2022**) A packaged food internal Raman signal separation method based on spatially offset Raman spectroscopy combined with FastICA. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, **275**, 121154.

Spectroscopic online and at-line monitoring of Anaerobic digestion at various process stages

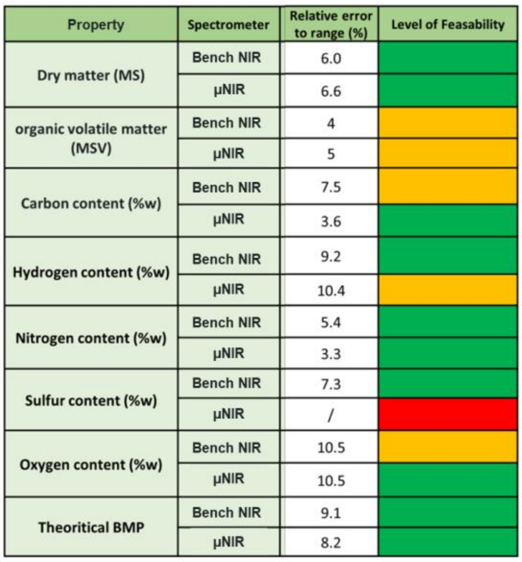
BERNARD Nicolas, BARBIER Nathalie, BRUN Nadège, SAMBUSITI Cécilia

Addresses: TotalEnergies, TRTG, Zone Industrielle Carrefour n° 4 - BP 27, 76700, Rogerville, France
nicolas.bernard@totalenergies.com

Anaerobic digestion can be described as driven oxygen-less biochemical digestion of organic waste that lead to the production of a substrate named digestate and a complex gas mix mainly composed of methane (CH4) and carbon dioxide (CO2). Organic feedstock of various nature and origin can be used as substrate for fermentation process, from organic waste to crops, slurry and manure but it has to pass first through a series of preprocessing steps such as filtration, thickening, dewatering, drying, hygienization and disintegration. Reaction of digestion triggered by chosen microorganism under anaerobic condition lead to a mean ratio of around 70 : 30 for CH4 over CO2 percent rate, depending both on feedstock nature and quality at the entrance of the fermentation unit and internal control parameters such as temperature, humidity, pH, oxygen rate… Gazes produced by fermentation are then post treated in a gas treatment unit until a minimum methane purity of around 97 percent is achieved. Bio-methane is then ready to be commercialized.

Monitoring of Anaerobic digestion with analytical technique applies both at gazes and substrate characterization but requirements and constraints are not the same for these two applications. For feedstock and digestate analysis, both a rapid, low-cost at-line characterization and an online monitoring inside the fermentation unit can provide useful information for process control. For biogas, online monitoring at different steps of product purification process is more adapted to ensure safety and specifications. On-line monitoring of bio-methane concentration at the output of digester is also the most reliable way to follow process performance in real-time, thus confirming Biochemical Methane Potential that are targeted in the feedstock quality control strategy. Even if biogas and feedstock characterization are part of different proof-of-concept project, objectives are joining in a common direction: improving performance and process control of anaerobic digestion. Today spectroscopic techniques such as NIR or Raman system show real advantages regarding these requirements compare to classical physico-chemical analytical technique, particularly since a growing offer of miniaturized low-cost solution are now available on the market. Here, focus is made on feedstock characterization. Main challenges to tackle are, according to literature: samples variability, heterogeneity and preparation.

Different technologies were tested: µNIR, benchtop NIR and Raman instruments. Proof-of-concept have highlighted the potential of NIR spectroscopy combined with chemometrics, and especially for rapid, low-cost and low-preparation measurement of dry matter content directly on site. Next step is to extend the PoC to biogas market.



Property	Spectrometer	Relative error to range (%)	Level of Feasability
Dry matter (MS)	Bench NIR	6.0	🟩
	µNIR	6.6	🟩
organic volatile matter (MSV)	Bench NIR	4	🟨
	µNIR	5	🟨
Carbon content (%w)	Bench NIR	7.5	🟨
	µNIR	3.6	🟩
Hydrogen content (%w)	Bench NIR	9.2	🟩
	µNIR	10.4	🟨
Nitrogen content (%w)	Bench NIR	5.4	🟩
	µNIR	3.3	🟩
Sulfur content (%w)	Bench NIR	7.3	🟩
	µNIR	/	🟥
Oxygen content (%w)	Bench NIR	10.5	🟨
	µNIR	10.5	🟩
Theoritical BMP	Bench NIR	9.1	🟩
	µNIR	8.2	🟩

Table 1: Summary of results for each properties tested

References

Mallet A., Charnier C., Latrille E., Bendoula R., Roger J.-M., and Steyer J.-P. (2022) Fast and robust NIRs-based characterization of raw organic waste: Using non-linear methods to handle water effects. Water Research, 227:119308.

Peng W., Beggio G., Pivato A., Zhang H., Lu F., and He P. (2022). Applications of near infrared spectroscopy and hyperspectral imaging techniques in anaerobic digestion of bio-wastes: A review. Renewable and Sustainable Energy Reviews, 165:112608.

PLS CLASS MODELING USING ERROR CORRECTION OUTPUT CODE MATRICES AND NIR SPECTROSCOPY TO DETECT DEFICIENCIES IN BAKERY DOUGHS

D. Castro[1,2], M.C. Ortiz[1], S. Sanllorente[1], I. García[2], L.A. Sarabia[3]

Universidad de Burgos, Dep. Chemistry[1], Dep. Mathematics and Computation[3],
Plaza Misael Bañuelos s/n, Burgos 09001, Spain
[2]AOTECH (Advanced Optical Technologies S.L.), Escuela Ing. De Bilbao,
Plaza Ingeniero Torres Quevedo, 1 2º, Bilbao 48013, Spain
E-mail (lsarabia@ubu.es)

Recently, it has been verified the usefulness of the error correction output code matrices application to codify the responses of a Partial Least Squares model with multiple classes (PLS2-ECOC). In the ref. [1], the authors found that the K-class-model obtained has shown an improvement over the model based on class indicator variables One- versus All (OVA).

In this work, the AONIR in-line measuring platform developed by AOTECH [2] is used in order to monitor the dough kneading and rolling processes in a bakery industry [3]. The versatility of the Near-infrared spectroscopy (NIR) in the food industry to assess the quality of raw materials and the final product is widely recognized for being a fast and non-destructive technique, and without sample preparation. What makes it increasingly demanded by the food industry. The purpose is to demonstrate the possibility to detect deficiencies in the kneading and rolling processes through NIR spectroscopy. For that, 579 and 481 samples have been monitored respectively in both processes, measuring the reflectance in a wavelength range between 900 and 1670 nm (125 wavelengths). The samples are grouped into four different classes, that reflect the operation of the kneading process: correct dough (C_1), dough with an excess of water (C_2), dough with lack of water (C_3) and dough with little kneading time (C_4). The proposed PLS2-ECOC entails a decomposition of a class-modelling problem into a series of binary learners [1], based on a family of code matrices with different code length, which are evaluated to obtain simultaneous compliant class-models with the best performance. The codification that provides the optimal model is obtained by comparing the sensitivity-specificity matrices in prediction of the built models with each one of the ECOC matrices. For that comparation, the DMECEN (Diagonal Modified Confusion Entropy) index is used. This index is more sensitive to changes in the sensitivities and specificities of the class models in difference with others like the total efficiency or accuracy.

In table 1 the number of samples in each class for each process is noted. Each class has been divided in training samples (70%) and prediction ones (30%) by using the Kennard-Stone procedure.

Table 1. Number of samples in each class for each process. TG y TT mean training and test.

	C_1		C_2		C_3		C_4	
	TG	TT	TG	TT	TG	TT	TG	TT
Rolling	88	38	88	38	77	33	83	36
Kneading	90	39	90	38	108	46	118	50

In table 2 the sensitivity and specificity matrix in training and in prediction for the kneading process, obtained using a code length of five, is noted.

Table 2. Kneading process. Sensitivity and specificity matrices

		Training					Prediction			
		Class model					Class model			
		C_1	C_2	C_3	C_4		C_1	C_2	C_3	C_4
True class	C_1	0.97	1.00	1.00	1.00	C_1	1.00	1.00	1.00	1.00
	C_2	0.97	1.00	0.34	0.42	C_2	1.00	1.00	0.26	0.29
	C_3	0.97	0.78	0.96	0.06	C_3	0.93	0.74	0.96	0.13
	C_4	1.00	0.92	0.20	0.97	C_4	1.00	0.98	0.14	0.98

For the rolling process, a code length of four was needed. In this case, both for training and prediction, sensitivities greater than 0.98 were obtained for every class, while specificities were equal to one. Furthermore, in prediction both parameters are equal to 1 for all classes.

Acknowledgements

This work has been funded by the Ministerio de Industria Comercio y Turismo under project PHOTONICS4BAKERY (AEI-010500-2021b-111) and Consejería de Educación de la Junta Castilla y León (BU052P20) cofinancied with FEDER funds.

References

[1] Valencia O, Ortiz MC, Sánchez MS, Sarabia LA (2022) Simultaneous class-modelling in chemometrics: A generalization of Partial Least Squares class modelling for more than two classes by using error correcting output code matrices. *Chemometr. Intell. Lab. Syst.* **227**, 104614.

[2] AOTECH. Advanced Optical Technologies. https://www.aotech.es/

[3] Daveiga. Galletas Mariñeiras. https://daveiga.es/web/es/

g-CovSel: covariance oriented variable clustering

Jean-Michel ROGER[a,b], Alessandra BIANCOLILLOc, Federico MARINI[d]

[a]ITAP-INRAE, Institut Agro, University Montpellier, 34196 Montpellier, France
[b]ChemHouse Research Group, 34196 Montpellier, France
[c]Department of Physical and Chemical Sciences, University of L'Aquila, 67100, Coppito, L'Aquila, Italy
[d]Department of Chemistry, University of Rome "La Sapienza", 00185, Rome, Italy
jean-michel.roger@inrae.fr

Variable selection is a highly studied topic in machine learning. In chemometrics, variable selection can be used as preprocessing to make calibration easier. In this case, it is a matter of eliminating uninformative variables, e.g. by means of a filter, such as VIP 1. Variable selection can also be used to simulate a simpler measurement device or to identify the chemical compounds of interest. It is then a matter of finding a subset of variables related to the phenomena of interest. This can be done by wrapper methods, such as genetic algorithms, or with specific methods, such as sparse-PLS 2 or CovSel 3. CovSel selects the columns of a matrix $\mathbf{X}$ in a stepwise process, maximising the squared covariance with a response matrix $\mathbf{Y}$. A deflation between each step ensures a parsimonious selection of poorly correlated variables. A multiblock (SO-CovSel, 4) and multiway (N-CovSel, 5) versions have been proposed. They too provide a parsimonious set of poorly correlated features. CovSel-based methods are therefore particularly suitable for the identification of compounds of interest (6, 7, 8). However, in some cases the identification of a small number of compounds of interest may not be sufficient, as for example in the case of mass spectrometry, when the molecules are fragmented. Indeed, in this case, clusters of fragments should be selected. The g-CovSel method proposes to generalise the x-CovSel methods (CovSel, SO-CovSel, N-CovSel). But below let us consider only the case of CovSel.

Let $\mathbf{X}(N,P)$ and $\mathbf{Y}(N,Q)$ be the descriptor and response matrices, respectively. Let $L = \{I_1, I_2, \dots, I_K\}$ be a list of indices of K variables of $\mathbf{X}$. Let $\mathbf{G}(P,L)$ be a matrix containing for each variable of $\mathbf{X}$ the degrees of membership in K groups corresponding to the K variables of L. The result of CovSel is a "hollow" matrix $\mathbf{G_0}$, i.e. such that $g_{ii} = 1, i \in L$ and $g_{ij} = 0,$ otherwise. The aim of g-CovSel is to produce a non-hollow matrix defining groups centred on the variables selected by CovSel. To do this, two solutions are possible. Either the selection $\mathbf{G}_0$ is modified, or the CovSel algorithm is adapted to generate a non-hollow matrix $\mathbf{G}$.

The first solution can be implemented using one of the many variable clustering algorithms. We propose, to remain in line with the CovSel philosophy, based on orthogonal projections, to adopt the following algorithm:

$$\mathbf{Z} = \mathbf{X}(:,L)$$

$$\mathbf{G} = \mathbf{X}^{\mathrm{T}}\mathbf{Z}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}$$

The second solution consists in calculating, at each step of the Covsel algorithm, a column of the matrix $\mathbf{G}$ by:

$$\mathbf{g}_j = \frac{diag(\mathbf{X}^{\mathrm{T}}\mathbf{Y}\mathbf{Y}^{\mathrm{T}}\mathbf{X})}{\sum diag(\mathbf{X}^{\mathrm{T}}\mathbf{Y}\mathbf{Y}^{\mathrm{T}}\mathbf{X})}$$

First solution produces a matrix of scores, from which membership degrees can be derived and can be applied to any variable selection method. Second solution produces directly membership degrees and is intricated in the CovSel algorithm. Both solutions will be presented, discussed and illustrated with real applications, and the extensions to multiblocks and multiways will be discussed.

References

1. Wold, Svante, Erik Johansson, and Marina Cocchi. "PLS: partial least squares projections to latent structures." *3D QSAR in Drug Design: Theory, Methods and Applications.*. Kluwer ESCOM Science Publisher, 1993. 523-550.
2. Lê Cao, Kim-Anh, Simon Boitard, and Philippe Besse. "Sparse PLS discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems." *BMC bioinformatics* 12.1 (2011): 1-17.
3. Roger, J. M., et al. "CovSel: variable selection for highly multivariate and multi-response calibration: application to IR spectroscopy." *Chemometrics and Intelligent Laboratory Systems* 106.2 (2011): 216-223.
4. Biancolillo, Alessandra, Federico Marini, and Jean-Michel Roger. "SO-CovSel: A novel method for variable selection in a multiblock framework." *Journal of Chemometrics* 34.2 (2020): e3120.
5. Biancolillo, Alessandra, Jean-Michel Roger, and Federico Marini. "N-CovSel, a new strategy for feature selection in N-way data." *Analytica Chimica Acta* 1231 (2022): 340433.
6. Calvani, Riccardo, et al. "Identification of biomarkers for physical frailty and sarcopenia through a new multi-marker approach: results from the BIOSPHERE study." *GeroScience* 43.2 (2021): 727-740.
7. Picca, Anna, et al. "Gut microbial, inflammatory and metabolic signatures in older people with physical frailty and sarcopenia: results from the BIOSPHERE study." *Nutrients* 12.1 (2019): 65.
8. Deuscher, Zoé, et al. "Volatile compounds profiling by using proton transfer reaction-time of flight-mass spectrometry (PTR-ToF-MS). The case study of dark chocolates organoleptic differences." *Journal of Mass Spectrometry* 54.1 (2019): 92-119.

Self-Optimizing High Shear Wet Granulation with DeepMPC

Carolina Gastone[1,*], **Erica Franceschinis**[2], **Marino Sergi**[1] **and Stefano Zanella**[1]

[1] Eigengran srl – Via Giuseppe Galliano 15 - 10129 Torino - Italy
[2] Department of Pharmaceutical and Pharmacological Sciences - University of Padova - 35131 Padova - Italy
[*] Correspondence: carolina@eigengran.it

In this paper, we present the development of a self-optimizing high shear wet granulation process using **DeepMPC**, a deep reinforcement learning-based Model Predictive Control (MPC) technique. DeepMPC is a control approach that combines the benefits of **Deep** Learning and **M**odel **P**redictive **C**ontrol.

The development of self-optimizing high shear wet granulation processes is a key area of research in the pharmaceutical industry, particularly in the context of Pharma 4.0 and the use of advanced technologies such as Artificial Intelligence (AI) and the Internet of Things (IoT) to improve drug development and manufacturing processes. One promising approach to self-optimization in this context is the use of DeepMPC, which involves using machine learning algorithms to build models of the granulation process, which can then be used to predict the outcome of different process conditions and optimize the process in real-time.

However, a key challenge in the development of self-optimizing high shear wet granulation processes is the limited availability of sensor data. To overcome this challenge, this paper proposes the use of sensor data fusion and online learning to enhance the performance of DeepMPC. Sensor data fusion involves combining data from multiple sensors to improve the accuracy and reliability of the measurement, and online learning algorithms can continuously update the models developed using DeepMPC based on new sensor data as it becomes available. The proposed approach involves the integration of Process Analytical Technology (PAT) sensors, such as Near-Infrared Spectroscopy (NIRS) and (first-order) torque.

In addition to sensor data fusion and online learning, this paper also discusses the importance of nonlinear dynamics modeling in the development of self-optimizing high shear wet granulation processes. Nonlinear dynamics refer to the complex, non-linear relationships between different process variables, which can make it difficult to predict the outcome of different process conditions. To address this challenge, deep learning algorithms can be used to model the nonlinear dynamics of the granulation process and improve the accuracy of the predictions. One approach that has shown promise in this context is the use of **R**ecurrent **N**eural **N**etwork (**RNN**), which are capable of modeling complex, high-dimensional data and can be trained using reinforcement learning techniques to optimize the control policy based on real-time process data. By using deep learning algorithms to model the nonlinear dynamics of the granulation process, it is possible to improve the accuracy of the predictions and optimize the process more effectively.

Finally, this paper discusses the potential use of **D**eep **R**einforcement **L**earning (**DRL**) to optimize the granulation process in real-time based on the sensor data and models developed using DeepMPC.

DRL involves the use of deep learning algorithms to enable an AI system to learn from the interactions between an agent and its environment. In this case, the agent is the DeepMPC controller and the environment is the high shear wet granulation process. We demonstrate the use of DRL to enable the DeepMPC controller to adapt to changes in the process and improve its control performance over time. This represents a key aspect of cognitive control, as the controller is able to learn and adapt to the dynamic nature of the process.

Overall, this paper presents a comprehensive approach to the development of self-optimizing high shear wet granulation processes that is in line with PAT guidelines and incorporates a range of engineering advances, including sensor data fusion, online learning, nonlinear dynamics modeling, and deep reinforcement learning. These approaches can help to overcome the challenges posed by limited sensor data and enable real-time optimization of the granulation process, leading to improved efficiency and product quality in the pharmaceutical industry.

Averaging local PLSR models to predict chemical compositions and nutritive values of forages from spectral near infrared data

Matthieu Lesnoff[a,b, c]

Addresses:
a. SELMET, Univ Montpellier, CIRAD, INRAE, Institut Agro, Montpellier, France
b. CIRAD, UMR SELMET, Montpellier, France
c. ChemHouse Research Group, Montpellier, France
email: matthieu.lesnoff@cirad.fr

In the past, several studies (e.g. Lesnoff et al 2022) demonstrated the efficiency of averaging the predictions of a set of partial least squares regression (PLSR) models with different numbers of latent variables (LVs). Such an approach belongs to the ensemble learning (or model averaging methods) whose advantageous properties have been described (Burnahm & Anderson 1998; Hastie & Tibishirani 2009; Ballabio et al. 2019), in particular the reduction of the variances of the predictions. The principle is as follows. Assume that $\boldsymbol{x}_{new}$ is a new observation to predict, and that $\hat{y}_{new,r}$ is the prediction returned by the PLSR model having a number r of LVs. The averaging model prediction with A LVs is given by:

$$\hat{y}_{new,avg[A]} = w_0\hat{y}_{new,0} + w_1\hat{y}_{new,1} + \cdots + w_A\hat{y}_{new,A}$$

where w_r (r = 0, …, A) is the weight (bounded between 0 and 1) of the model with r LVs, with the constraint: $\sum_{r=0}^{A} w_r = 1$. Vector $\boldsymbol{w}$ = {w_0, w_1, …, w_A} represents a pattern of weights and the shape of this pattern is specific to a given averaging method.

LVs-averaging can be extended to any method using PLSR, in particular the local PLSR (also referred to as kNN-PLSR) where a given number of k neighbors to each new observation to predict is selected, and then a PLSR model is fitted on these neighbors. Local PLSR averaging was for instance implemented in the Winisi © software for the pipeline described by Shenk et al (1998).

In this communication, I present the extension to another local pipeline, referred to as kNN-locally weighted PLSR (kNN-LWPLSR) (Lesnoff et al. 2020) where, in addition to the neighbors' selection, training observations that are closer to the new observation to predict receive higher weights in the PLSR. I compare the efficiency (predictive errors) between the usual kNN-LWPLSR and various strategies of averaging. The method is applied to an extensive NIR database available at Cirad (Selmet research unit), built to predict the chemical composition of European and tropical forages and feed (twelve variables; Table 1). The study showed an overall superiority of the averaging over the usual approach where the number of LVs is set to a given value (partial results in Fig. 1).

Table 1: The twelve variables of forage and feed chemical composition to be predicted (*N* = total number of observations, n_{test} = size of the test set; for each variable, the test set was replicated five times).

Abbreviation	Unit	Variable	*N*	n_{test} (20% *N*)	Mean	Std
DM	%	Dry matter	17714	3543	92.2	2.5
ASH	%DM	Mineral matter	17299	3460	9.7	8.3
CP	%DM	Crude protein	16295	3259	14.9	12.3
EE	%DM	Crude fat	7004	1401	5.7	7.2
CF	%DM	Crude fiber	13006	2601	26.0	11.8
NDF	%DM	Neutral detergent fiber	10414	2083	52.1	18.0
ADF	%DM	Acid detergent fiber	11097	2219	33.3	13.4
ADL	%DM	Acid detergent lignin	10809	2162	8.2	6.5
DMDCELL	%DM	DM enzymatic digestibility	9043	1809	52.5	17.6
OMDCELL	%DM	OM enzymatic digestibility	8713	1743	50.4	17.9
STARCH	%DM	Starch	1919	384	31.3	20.8
SUGARS	%DM	Total sugars	2014	403	8.0	9.5

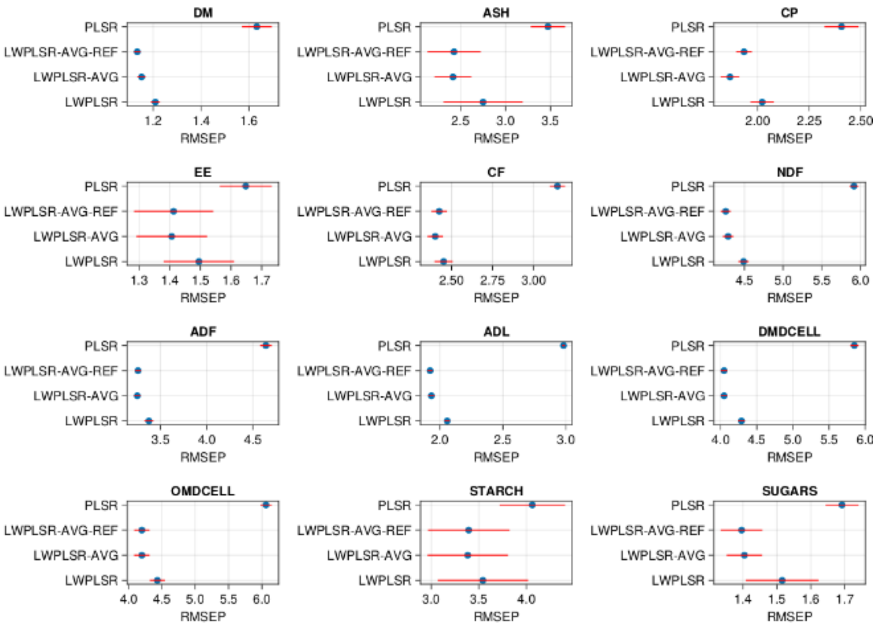


Figure 1: RMSEP$_{test}$ for kNN-LWPLSR: usual and various strategies of averaging.

Refernces

Ballabio D, Todeschini, R, Consonni, V (**2019**). Chapter 5 - Recent Advances in High-Level Fusion Methods to Classify Multiple Analytical Chemical Data, in: Cocchi M (Ed.), Data Handling in

Science and Technology, Data Fusion Methodology and Applications. Elsevier, pp. 129–155. https://doi.org/10.1016/B978-0-444-63984-4.00005-3

Burnham KP, Anderson DR (**1998**). Model selection and inference. A practical information-theoretic approach. Springer, New York.

Hastie T, Tibshirani R, Friedman J (**2009**). The elements of statistical learning: data mining, inference, and prediction, 2nd ed. Springer, New York.

Lesnoff M, Metz M, Roger J-M (**2020**). Comparison of locally weighted PLS strategies for regression and discrimination on agronomic NIR data. *Journal of Chemometrics* n/a, **e3209**. https://doi.org/10.1002/cem.3209.

Lesnoff M, Andueza D, Barotin C, Barre P, Bonnal L, Fernández Pierna JA, Picard F, Vermeulen P, Roger J-M (**2022**). Averaging and Stacking Partial Least Squares Regression Models to Predict the Chemical Compositions and the Nutritive Values of Forages from Spectral Near Infrared Data. *Applied Sciences* **12**, 7850. https://doi.org/10.3390/app12157850.

Monitoring the State of Health (SOH) of green batteries (GreenBat)

<u>Eugenio Sandrucci</u>, Sergio Brutti, Federico Marini

University of Rome "La Sapienza", Department of Chemistry, P.zzle A. Moro 05, Rome, 00185, Italy
eugenio.sandrucci@uniroma1.it

Here we focus on the study of state of health monitoring of secondary batteries with advanced "green" formulations. European recycle map on battery energy storage plans to replace positive electrodes consisting of cobalt (Co) and fluorinated (F) graphite-based negative electrodes constituents with environmentally sustainable alternatives. Recently, "green formulations" was validated; in particular, recent trends suggest the use of electrodes based on silicon (Si), Co-free manganites (LiMnO2) aqueous soluble binders and fluorine-free electrolytes [1]. Our aim is to describe the chemical and morphological changes occurring during the battery cycling, through spectroscopies (Raman and IR) operating in parallel with impedance analysis. These experimental data will be integrated with the electrochemical performance in a multiblock dataset, which will constitute the basis for chemometric processing. The purpose of this chemometric modelling is to correlate SOH evolution to the chemical and physical state of the battery [1]. Here we illustrate the analysis of SOH of two different pouch cells formulation constituted by LTO (Lithium-Titanate) as anode, LFP (Lithium-Iron-Phosphate) as cathode and the LP30 electrolyte (EC:DMC 1:1 and LiPF6 1M); the second formulation is constituted by a thin film layer of commercial Li metal (purity of 99.7%) as anode, LFP as cathode and an ethers based formulation as electrolyte. After formation cycles (CC-CV), the cells have been submitted to aging tests. The different operating conditions are chosen with an Experimental Design (ED). Turning to the chemometric modelling, as a starting point we verified the ability of PLS regression applied to the voltage profiles during charge cycles to estimate the SOH of a benchmark dataset by Lin et al. [1] (Figure 1c).

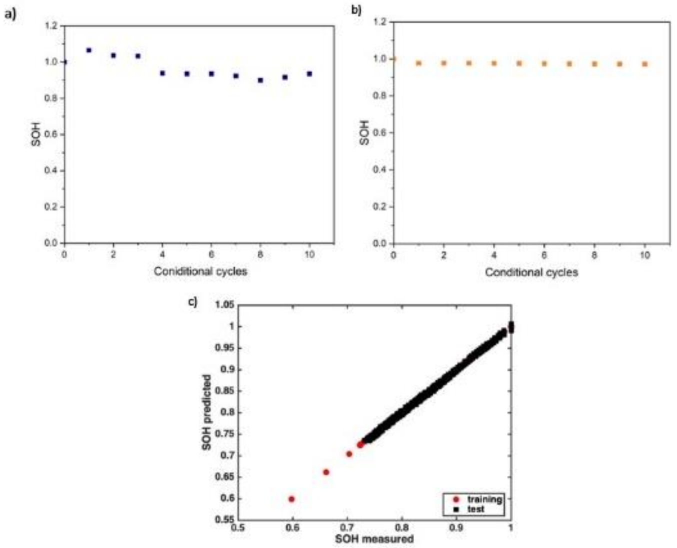


Figure 1: a) SOH trend of Li-ion battery during the aging test; b) SOH trend of Li metal battery during the aging test; c) Results of PLS modelling of the benchmark data in [1]: the plot displays the comparison between predicted and measured SOH for the training (data from 5 cells) and the test (data from the remaining three cells) sets. For the test set, the value of R2 is 0.9977 and RMSEP is 0.0034.

References

[1] M. Lin, D. Wu (**2022**), *Journal of Power Sources*, **518**, 230774.

USING DATA GEOMETRY TO HIGHLIGHT THE NECESSITY OF BRIDGING THE GAP BETWEEN CLUSTERING AND SPECTRAL UNMIXING IN COMPLEX SAMPLES

Alessandra Olarini [1,2], **Marina Cocchi** [2], **Ludovic Duponchel** [1], **Cyril Ruckebusch** [1]

[1]Université de Lille, LASIRE, Cité Scientifique, Villeneuve d'Ascq, 59650, France
[2]Università di Modena e Reggio Emilia, DSCG, via Campi 103, Modena, 41125, Italia
alessandra.olarini@unimore.it

Visualizing and understanding the structure of the spectral data at hand is a key step for selecting the appropriate analysis methodology. This becomes even more important when no a priori knowledge or ground truth information is available. In this work, a geometric interpretation based on convexity of normalized scores and loadings data points, resulting from a Singular Value Decomposition analysis (SVD), is proposed to extract the most useful (essential) information, in terms of linear mixture analysis, from a complex Laser Induced Breakdown Spectroscopy (LIBS) hyperspectral image of a turquoise rock sample [1,2,3]. Figure 1 illustrate the results obtained. In the normalized score plot (Fig.1a) visualized in 2D, where each data point represents a spectrum, purest spectra can be identified (highlighted with red circles at the vertices of a convex geometry).

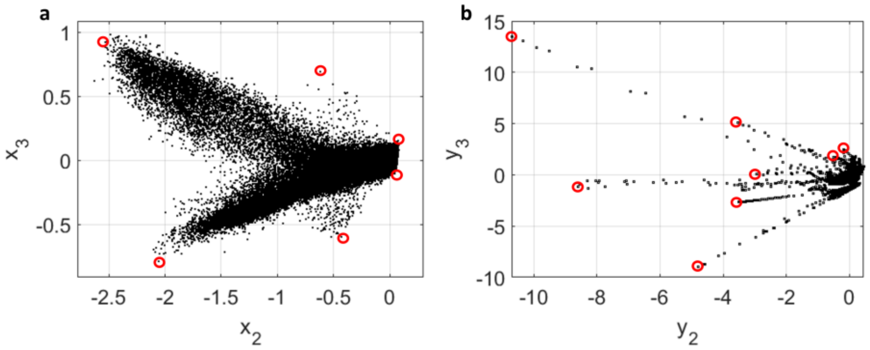


Fig.1: Plots of: a) normalized scores and b) normalized loadings of LIBS hyperspectral imaging data of a turquoise sample.

In the normalized loading plot (Fig.1b), where each point now denotes an image for a specific wavelength, the purest images can be identified in the same way, and more contributions may be found. The identified pure spectra and images are then used to initialize clustering and spectral unmixing approaches (initial estimation of number of clusters, components, components profiles). While clustering aims to find similar groups of pixels by looking at their spectral signatures, considering that the spectrum of one pixel belongs to one cluster only, linear spectral unmixing aims to find pure contributions under the assumption that a pixel is a linear combination of pure spectral profiles. However, in cases where the initial estimation of the number of clusters or components is tricky, or when spectra profiles are very similar, none of the approaches provides satisfactory results. We illustrate this issue by comparing the results obtained applying Kmeans for clustering and Multivariate Curve Resolution-Alternating Least Squares (MCR-ALS) for spectral unmixing analyses to the LIBS data introduced previously. The results show that none of the two individual methods can extract as comprehensive information as the one obtained from the extraction of essential information in both the scores and loadings subspaces. This translates that the structure of the investigated data deviates from the ideal one for K-means and MCR-ALS analyses. This can be explained by the fact that very similar distributions are associated with different spectra, reflecting the fact that some elements in the samples may exchange between different phases during the mineral crystallization. This preliminary work emphasizes the need to bridge the gap between clustering and spectral unmixing, opening the door to the development of new methods, in situations where the data structure deviates from the ideal one for clustering (pixel not characteristic of a unique cluster) and spectral unmixing (deviation from the ideal bilinearity).

References

[1] Grande BV, Manne R (**2000**). Use of convexity for finding pure variables in two-way data from mixtures. *Chemometrics and intelligent laboratory systems*, **50(1)**, 19-33.

[2] Ruckebusch C, Vitale R, Ghaffari M, Hugelier S, Omidikia N (**2020**). Perspective on essential information in multivariate curve resolution. *TrAC Trends in Analytical Chemistry*, **132**, 116044.

[3] Moncayo S, Duponchel L, Mousavipak N, Panczer G, Trichard F, Bousquet B, Frédéric P, Motto-Ros V (**2018**). Exploration of megapixel hyperspectral LIBS images using principal component analysis. *Journal of Analytical Atomic Spectrometry*, **33(2)**, 210-220.

NIR hyperspectral imaging combined with chemometrics for the monitoring of water patterns during dehydration of nonvascular epiphytic communities

Cristina Malegori[1], Giulia Canali[2], Rodrigo Rocha de Oliveira[1,3], Sara Gariglio[1,4], Paola Malaspina[2], Riccardo Voccio[1,4], Monica Casale[1], Paolo Oliveri[1], Paolo Giordani[1]

[1] *DIFAR – Department of Pharmacy, University of Genova, Viale Cembrano 4, 16148 – Genova (Italy)*

[2] *DISTAV – Department of Earth, Environment and Life Sciences, University of Genova, Corso Europa, 26, 16132 – Genova (Italy)*

[3] *Department of Chemical Engineering and Analytical Chemistry, Universitat de Barcelona, Martí i Franquès 1, 08028 – Barcelona, (Spain)*

[4] *DCCI – Department of Chemistry and Industrial Chemistry, University of Genova, Via Dodecaneso 31, 16146 - Genova (Italy)*

cristina.malegori@unige.it

The present work is aimed at developing water maps for understanding how a micro-ecosystem composed by nonvascular epiphytic communities (mainly, lichens and bryophytes), manages water and, consequently, how the thalli can prolong water availability depending on the morphology of the species and the composition of the community itself. For the first time, in fact, the nonvascular epiphytic community has been studied during the dehydration process with the objective of understanding the relationship between bryophytes and lichens and their competition or facilitation regarding water management.

To achieve this aim, the water status of bark, lichens (crustose and foliose) and bryophytes was monitored by means of near-infrared (NIR) hyperspectral imaging (HSI), working in the spectral range from 1000 to 2500 nm (SWIR3 – Specim, Finland). This spectroscopic technique appears to be the most suitable from both a spectral and a spatial perspective. From the spectral perspective, in fact, NIR spectroscopy presents strong absorption bands ascribable to the O-H bond and, so, water shows an unmistakable signature in this spectral region. Moreover, the application of an imaging technique allows to identify spatial patterns, a key aspect when a non-homogeneous sample is analysed. Furthermore, the non-destructive and non-contact nature of the technique offers the advantage of enabling the monitoring of the same sample for the entire duration of the process without interferences.

Regarding the sampling protocol, it was possible to analyse pure communities (composed mainly by a single organism), binary and ternary mixtures, for a total of 32 samples. All of the micro-ecosystem were naturally supported by chestnut bark. Samples of non-colonised (or as less colonised as possible) bark were also analysed.

The samples were first submitted to a full hydration protocol and, then, let dry at room temperature (19°C) and humidity (41%) for about 12 h. During the dehydration period, NIR-HSI acquisition was carried out every 45 minutes for 15 sampling points, obtaining 480 images. Before image acquisition at each sampling point, the samples were weighed to quantify their dehydration curve, relative to their weight, prior to the hydration phase. The system, in fact, over the period of time relative to the dehydration cycle, can be considered constant regarding its chemical composition and so the variations along NIR spectra can be ascribable mainly to water content variation.

On the images, an exploratory data analysis was performed, by means of principal component analysis (PCA), with the aim of comparing the drying trajectory in the orthogonal space defined by the lowest-order components with the dehydration curve. The study of these correspondences on pure communities opens the possibility to understand the water management of each organism and then to use such information in the study of more complex microclimates (binary and ternary mixtures). The projection of the microclimate images in the space defined by the pure samples allows to map changes in the water status of each complex sample, focusing on lichens and bryophytes and on their interaction. The role of the bark support was also considered.

After this exploratory step, a supervised strategy is proposed, with the aim of extracting not only the water pattern inside the microclimate but also a numerical parameter expressing the water content in the different regions of interest inside a single map.

The study of these communities and their relative behaviour in water stress condition plays a crucial role in understanding how these organisms are facing the climate change and the subsequent water scarcity or overabundance.

The *black hole effect* in multivariate curve resolution based on alternating least squares: a comprehensive overview and two possible solutions to overcome it

Raffaele Vitale, **Mohamad Ahmad**, **Cyril Ruckebusch**

Address: Dynamics, Nanoscopy and Chemometrics (DyNaChem) Group, Laboratoire
Avancé de Spectroscopie pour les Interactions, la Réactivité et l'Environnement (LASIRE
CNRS – UMR 8516), Université de Lille, F-59000 Lille, France
raffaele.vitale@univ-lille.fr

Least squares-based estimations lay behind most chemometric methodologies. In the domain of Multivariate Curve Resolution (MCR), for instance, the use of Alternating Least Squares (ALS) has been ubiquitous since the very first development of one of the workhorse algorithms proposed for such a purpose, MCR-ALS, in the 1980s. All over these years, the applications of MCR-ALS have significantly increased, which has nowadays rendered this technique one of the most commonly utilised by operators and practitioners.

Conversely, though, scarcer attention has been paid to the limitations MCR-ALS can suffer from in case the handled data structures exhibit certain characteristics violating specific distributional assumptions that should be ideally fulfilled when ALS-based approaches are concerned. As an example, we have recently highlighted how, in the presence of minor components underlying the datasets at hand and if the number of analysed data points is relatively large, the weight or relevance of those that may be essential for a MCR-ALS resolution might become too low for guaranteeing its correctness. In such a regard, this work aims at offering a comprehensive perspective on this particular deleterious phenomenon which has been lately denominated the *black hole effect*. A distinctive focus will be put on the analogy it has with the biased calibration of univariate and multivariate regression models that may occur when high-leverage outlying observations are collected and processed. Most importantly, two solutions to successfully overcome this effect, namely data pruning and object weighting based on a measure of essentiality for the sake of curve resolution, will be discussed and tested in both simulated and real case-studies.

References

Gampp H, Maeder M, Meyer CJ, Zuberbühler AD (1985) Calculation of equilibrium constants from multiwavelength spectroscopic data. III. Model-free analysis of spectrophotometric and ESR titrations. *Talanta*, **32**, 1133-1139.

Tauler R, Casassas E (1988) Principal component analysis applied to the study of successive complex formation data in the Cu(II) ethanolamine systems. *Journal of Chemometrics*, **3**, 151-161.

Ghaffari M, Omidikia N, Ruckebusch C (2019) Essential spectral pixels for multivariate curve resolution of chemical images. *Analytical Chemistry*, **91**, 10943-10948.

Vitale R, Ruckebusch C (2023) On a *black hole* effect in bilinear curve resolution based on least squares. *Journal of Chemometrics*, **37**, e3442.

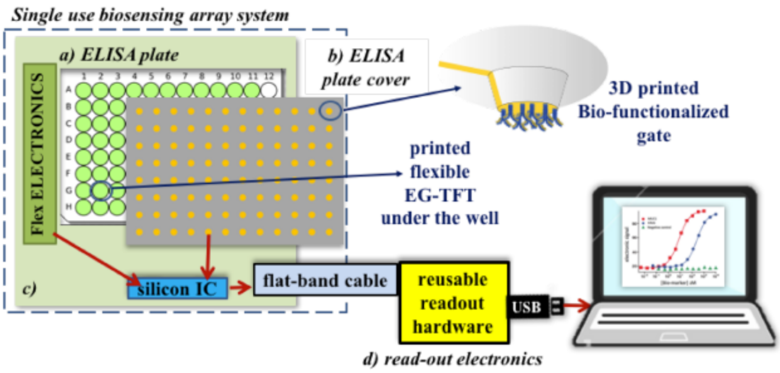
Ahmad M, Vitale R, Cocchi M, Ruckebusch C (submitted) Weighted Multivariate Curve Resolution-Alternating Least Squares based on sample relevance.

Single-molecule bioelectronic sensor: improving reliability with chemometric approaches

Eleonora Macchia * and Luisa Torsi

Addresses: University of Bari, Department of Pharmacy-Pharmaceutical Sciences, Via Orabona 4, Bari, 70124, Italy
Eleonora.macchia@uniba.it

Digitizing biomarkers analysis by quantifying them at the single-molecule level is the new frontier for advancing the science of precision health. The enhancement of the technical capabilities of bioelectronics systems, by giving clinicians the possibility to rely on biomarkers quantifications down to the single-molecule, holds the potential to revolutionize the way healthcare is provided. Such an analytical tool will indeed enable clinicians to associate a biomarker tiniest increase to the progression of a disease, particularly at its early stage.[1] Eventually, physicians will be able to identify the very moment in which the illness state begins. Such an occurrence will enormously enhance their ability of curing diseases by supporting better prognosis and permitting the application of precise treatment methods. The single molecule bio-electronic smart system array for clinical testing - SiMBiT - technology has been developed within the blooming field of precision medicine, leveraging on the single molecule with large transistor (SiMoT)[2] lab-based technology that can perform single-molecule detection of both proteins and DNA bio-markers.[3,4] Specifically, the SiMBiT technology has lately developed the SiMoT lab-based device into a cost-effective portable prototype multiplexing array that integrates, with a modular approach, standard components and interfaces with novel materials and exhibits enhanced sensing capabilities. The SiMBiT prototype has proven its potency in early detection of pancreatic cancer, being capable to discriminate among low-grade and high-grade mucinous cyst's lesions in peripheral biofluids, such as plasma samples. In this perspective, machine learning approaches play a pivotal role in developing classifiers for a fast, reliable multiparametric biosensors output. Supervised model based on multivariate data processing has been undertaken to enable multiplexing, *i.e.* the simultaneous quantification of three biomarkers, namely MUC1 and CD55 proteins and KRAS DNA mutated sequence, in plasma and cysts' fluid samples. The main technological aspect of the SiMBiT device, with particular emphasis on the potency of machine learning approaches, will be discussed.



SiMBiT portable prototype.

References

1. Macchia, E. *et al.*, *Chem. Soc. Rev.* 122, 4, 4636–4699 (2022).
2. E. Macchia, et al., *Nat. Commun.* 9 (2018).
3. L. Sarcina, et al., *Anal. Bioanal. Chem.* 414, 5657 (2022).
4. E. Macchia, et al., *Adv. Electron. Mater. 7 (2021)*.

REVISTING THE ROSA ALGORITHM

A. Biancolillo[1], C. Fiore[2], M. Ryckewaert[3], S. Mas Garcia[3], J.M. Roger[3], F. Marini[2]

[1]Dipartimento di Scienze Fisiche e Chimiche, Università degli Studi dell'Aquila, Via Vetoio – 67100 Coppito
[2]Dipartimento di Chimica, Università di Roma La Sapienza, P.le Aldo Moro, 5 – 00185 Roma
[3]ITAP, INRAE, Rue Breton 361, F-34196 Montpellier, Francia
E-mail federico.marini@uniroma1.it (corresponding author)

Recently, response-oriented sequential alternation (ROSA) was presented as a novel method for multiblock data analysis (Liland et al., 2006). With respect to competitive approaches for multi-block data analysis, ROSA exploits a "winner takes all" approach, where each component is calculated from the block of predictors that leads to the minimum residual error, to produce accurate model in exceptionally rapid computational times.

This new method, in fact, has numerous advantages as it is independent of the scaling and order of the blocks; it is computationally faster than other multi-block methods; each component of the model is calculated from the block of predictors that most reduce the current residual error.

However, in addition to these strengths, this method also has a major limitation: it has been noted, in fact, that the choice of the various blocks and, in particular, of the first block, considerably influences the choice of all the others and, hence, the resulting performance of the model.

The present work therefore focused on the criteria for choosing the various blocks during the regression, implementing two alternative criteria to the standard one, which provides for the default choice of the block which minimizes the overall residual error during the cross-validation (RMSECV). Compared to the standard criterion, the two alternative criteria investigated in this work are as follows. The first ("diff") imposes that the chosen block must necessarily be different from the one chosen in the previous iteration. With the second criterion ("Thres"), however, the method tries to take the second block, in the event that the second block chosen during a certain iteration is the same as the one chosen in the previous iteration, we will choose to select a different block (the second in order of criterion), only if the difference between the residuals is lower than a certain threshold, for example 5%, %; otherwise it continues with the one initially chosen.

Moreover, starting from the original formulation, which was meant for the prediction of a univariate response, we developed a multi-response version which was then extended to cope also with discriminant classification problems. In the present communication, these extensions of the original ROSA algorithm will be discussed together with their application to real and simulated data sets.

References

Liland KH, Næs T, Indahl UG. (**2016**) ROSA—a fast extension of partial least squares regression for multiblock data analysis. *J. Chemometri*cs, **30,** 651-662.

Revisiting PCA for data sets with missing values

Adrián Gómez-Sánchez[1,2], **Cyril Ruckebusch**[2], **Anna de Juan**[1]

1. *Chemometrics group. Universitat de Barcelona. Martí i Franquès, 1. 08028. Barcelona.*
2. *LASIRE. Laboratory of Advanced Spectroscopy, Interactions, Reactivity and Environment Université Lille, CNRS, UMR 8516, Cité Scientifique, Bâtiment C5, 59000, Lille, France*

E-mail: aderegomez@gmail.com

Principal Component Analysis (PCA)[1] is a widely used technique to reduce the dimensionality of multivariate data sets while retaining the most relevant information and to estimate the number of underlying sources of variation needed to reproduce the original data. Additionally, the reconstruction of the data using reduced number of principal components can be used as a noise filter to improve the convergence and the stability of iterative unmixing algorithms, such as Multivariate Curve Resolution-Alternating Least Squares (MCR-ALS)[2].

There are several alternatives to deal with missing values in PCA. Among those, the Single Imputation algorithm uses an iterative process to update the estimated values for the missing data at each iteration based on the prediction of the model itself[3,4]. However, the method can be computationally intensive and may not always be the most effective approach to estimate missing values in large datasets or in sets with systematic patterns of missing data[4].

Another good alternative to deal with missing values in PCA is the Nonlinear Iterative Partial Least Squares (NIPALS)[5] algorithm. In this context, NIPALS can be modified to deal with missing values by simply excluding those missing values from the row-by-row and column-by-column calculations of scores and loadings, without significantly impacting the accuracy of the PCA results, and being faster than other algorithms based on imputation approaches[3]. However, a limitation of NIPALS in the presence of missing values is that the calculated components are not perfectly orthogonal and, then, a small bias in the reproduction of the data is induced[3].

In this presentation, we explore a new alternative to both approaches, based on the implementation of the orthogonality constraint on the scores and loadings during the ALS, which could be called Orthogonalized-Alternating Least Squares (O-ALS). Starting from a preselected number of components, the O-ALS approach decomposes the data using a row-by-row and column-by-column ALS procedure. When a missing value is detected during the calculation, it is just skipped, as in the NIPALS case. As no imputation is required with O-ALS, the analysis is significantly speeded up while, at the same time, the calculated components are perfectly orthogonal resulting in unbiased reproduction of the data.

Here, we present the application of NIPALS, Single Imputation and O-ALS algorithms when missing values are present, showing the benefits and limitations of each method and the possibility to be applied as a noise filters and rank estimators.

References

1. Jollife, I. T., Cadima, J. (**2016**). Principal component analysis: A review and recent developments. *Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences*, 374 (2065).
2. Dadashi, M., Abdollahi, H., Tauler, R. (**2012**). Maximum likelihood principal component analysis as initial projection step in multivariate curve resolution analysis of noisy data. *Chemometrics and Intelligent Laboratory Systems*, 118, 33-40.
3. Grung, B., Manne, R. (**1998**). Missing values in principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 42(1-2), 125-139.
4. Walczak, B., Massart, D. L. (**2001**). Dealing with missing data: Part I. *Chemometrics and Intelligent Laboratory Systems*, 58(1), 15-27.
5. Wold, H. (**1975**). Soft modelling by latent variables: the non-linear iterative partial least squares (NIPALS) approach. *Journal of Applied Probability*, 12(S1), 117-142.

PLSDA VERSUS PCA ON BARYCENTERS APPLIED TO METABOLOMICS IN A CONTEXT OF DISCRIMINATION

Marion Brandolini-Bunlon[1], **Benoît Jaillais**[2], **Mohamed Hanafi**[2]

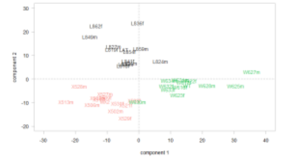
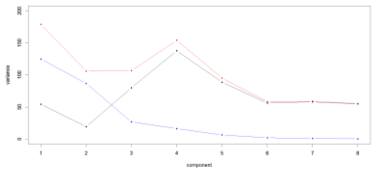
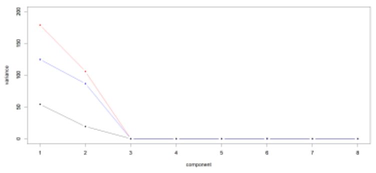
Addresses: [1]Université Clermont Auvergne, INRAE, UNH, Plateforme d'Exploration du
Métabolisme, MetaboHUB Clermont, 63000 Clermont-Ferrand, France
[2]Oniris, INRAE, StatSC, 44300 Nantes, France
E-mail: marion.brandolini-bunlon@inrae.fr

Introduction. Untargeted metabolomics is a powerful phenotyping tool to better understand the biological mechanisms involved in the physiopathological processes, and identify biomarkers of metabolic status. The complex data need dedicated preparation and treatments to extract meaningful information. The major specificity of metabolomics data is the large number of variables compared to the number of samples, as well as their high degree of correlation. The common analysis strategy consists in performing univariate and multivariate statistics to highlight variables of interest. In a discriminant context, partial least squares-discriminant analysis (PLSDA) is one of the most effective multivariate tools currently used, because of its ability to analyze collinear and noisy data. Another multivariate method that could be used is the Principal Component Analysis (PCA) of the matrix of barycenters of the observation groups (called here "PCA$_C$"). The objective of our study is to compare these approaches in terms of explained variances and important variables.

Material and methods. Published data from a project on the impact of genetic mutations in mice (ProMetIS) were used as a case study (Imbert, 2021). Mice (n=42), males and females, belonged to one of the three genotype groups (wild type, lacking the linker for activation of T cells gene, or lacking the MX dynamin-like GTPase 2 gene). The metabolomics dataset we used, was obtained from the analysis of plasma samples using a mass spectrometry-based untargeted approach (LC-MS), and contained 6104 variables after preparation. In the present work, data analysis was performed with the R-package "rchemo". Six atypical mice were removed to have a balanced experiment design, before Pareto scaling. Due to the sex effect, without interaction with the genotype, the data were centered by sex before applying PLSDA, and PCA$_C$, to discriminate genotype groups. On one hand, the optimal number of PLS components was determined according to the error in repeated cross-validation (30 repetitions of 10-

fold cross-validation) and application of the one-standard-error-rule. On the other hand, using PCA$_C$ model, subjects were projected onto the components. For each PLSDA or PCA$_C$ component, the total, inter- and intra-group variances were then calculated, and the group effect was assessed by ANOVA. ANOVA were also performed for metabolomics variables becoming important in the discrimination in the PLSDA model with the optimal number of components (called here "PLSDA$_{opt}$"), compared to the one with 2 components.

Results. The optimal number of PLS components was 3, and the number of PCA$_C$ components was 2. Only these components had a significant p-value in the ANOVA. As expected, the 1st component of both methods were the same, and, as shown below, in our study, the 2nd components were closely similar. The 3rd component of the PLSDA model was also of interest because it still significantly explained intergroup variability and highlighted other important discriminant variables.

	PLSDA		PCA$_C$
Score scatter plots			
Component variances ● total ● between ● within			

Discussion and conclusion. The PLSDA and the PCA$_C$ components maximize the inter-group variability. When 3 groups are to be discriminated, the PCA$_C$ finds 2 components while the PLSDA can find more. Presumably, each component of the PLSDA$_{opt}$ can discriminate one group from one or both others, and would thus allow a better discrimination.

References

Imbert A, *et al.* (**2021**) ProMetIS, deep phenotyping of mouse models by combined proteomics and metabolomics analysis. *Sci Data*, 8(1), 311.

DATA FUSION AND MULTIVARIATE CURVE RESOLUTION FOR EVALUATING THE METABOLIC IMPACT OF DIFFERENT GADOLINIUM-BASED CONTRAST AGENTS IN MICE

Nunzia Iaccarino[1], Francesca Romano[1], Gelsomina Riccardi[1], Antonio Randazzo[1], Nicola Cavallini[2], Francesco Savorani[2], Chiara Furlan[3], Enza Di Gregorio[3], Eliana Gianolio

[1] University of Naples, Department of Pharmacy, Via D. Montesano, 49, Naples, 80131, Italy
[2] Politecnico di Torino, Department of Applied Science and Technology, Corso Duca degli Abruzzi 24, Turin, 10129, Italy
[3] University of Turin, Department of Molecular Biotechnologies and Health Science, Via Nizza 52, Turin, 10126, Italy
nunzia.iaccarino@unina.it

Gadolinium-based contrast agents (GBCAs) are massively employed in radiology to increase the sensibility of magnetic resonance imaging. During the past decades, warnings about potential harmful effects from the use of linear GBCAs ascribable to the release of free gadolinium cations have been raised.[1] However, investigations aiming at detecting possible metabolic perturbations and/or potential adverse health effects due to gadolinium deposition are still lacking. Thus, the aim of the present work was to exploit a multi-organ (liver, kidney, spleen, cerebrum, and cerebellum) and plasma metabolomics approach to investigate the effects of multiple administrations on the main metabolic pathways in healthy mice of one linear (Omniscan, gadiodiamide), and one macrocyclic GBCA (ProHance, gadoteridol).

In this work, we have analysed aqueous extracts of five organs (brain, cerebellum, kidney, liver, and spleen) and plasma by employing, GC-MS and ^{1}H-NMR, from thirty mice divided in three study groups: control group, Omniscan-treated group, ProHance-treated group. Mice received twenty consecutive injections of each GBCA during a 5-week period, then organs and plasma were retrieved 1 month after the end of the treatments.

After GC-MS analysis, a low-level data fusion was performed by concatenating the data matrices of the five different organs to generate an individual data matrix which was submitted to Principal Component Analysis (PCA) allowing to draw interesting observations about the potential toxic effects of Omniscan and to identify the organs most affected by its administration. On the other hand, Multivariate Curve Resolution (MCR) was crucial to reliably assign the nature and the relative concentration of both metabolites and lipids from plasma samples analysed by ^{1}H-NMR. Also in this case, PCA was employed to explore the data, revealing a clear lipid dysregulation in the GBCAs-treated animals.

In conclusion, this study demonstrated the importance of using chemometric methods to extract relevant biological information from a complex multi-organ and plasma metabolomics dataset allowing, for the first time, a detailed identification of the perturbed biochemical pathways in mice exposed to two GBCAs having different chemical structure and thus, different capability to retain the gadolinium ion.

References

Do C, DeAguero J, Brearley A, Trejo X, Howard T, Escobar GP, Wagner B (**2020**) Gadolinium-based contrast agent use, their safety, and practice evolution. *Kidney360*, **1** (6), 561–568.

Authentication of PDO saffron of L'Aquila (Crocus sativus L.) by HPLC-DAD coupled with a discriminant multi-way approach

Alessandra Biancolillo[1], Maria Anna Maggi[1], Alessandra De Martino[2], Federico Marini[2], Fabrizio Ruggieri[1], Angelo Antonio D'Archivio[1]

[1]University of L'Aquila, Department of Physical and Chemical Sciences, Via Vetoio, L'Aquila, 67100, Italy

[2]University of Rome "La Sapienza", Department of Chemistry, Piazzale Aldo Moro 5, Rome, 00185, Italy

alessandra.biancolillo@univaq.it

Very often, arrays of data that would be naturally multi-way are re-elaborated to make them suitable for traditional approaches tailored for data matrices. This can be done by unfolding the arrays, or by feature selection. Among the other contexts, this happens with data provided by HPLC-DAD instruments. In fact, these instruments provide a three-way array of data, but, customarily, chromatograms are only taken at specific wavelengths. Regardless of the way dimensionality reduction occurs, unfolding or variable selection can lead to a loss of useful information. Given this premise, the present work proposes a multi-way strategy for the analysis of the entire spectro-chromatograms collected by HPLC-DAD. The outcome of this strategy will be shown on a classification problem concerning the discrimination of PDO saffron from L'Aquila from other samples of the same spice harvested in different areas [1]. The multi-way strategy has been compared to the analysis of the individual data matrices. The entire spectro-chromatograms were handled by N-Partial Least Squares (N-PLS) [2] and then classified by Linear Discriminant Analysis (LDA) [3], whereas individual matrices were classified using PLS-DA [4]. Both approaches provided satisfactory accuracies; the best results from the prediction point of view (estimated on an external set of samples) were achieved by the multi-way methodology. This achievement indicates that three-way spectro-chromatograms contain relevant information for the characterization of the investigated samples, which is lost restricting the analysis to only a few specific wavelengths.

References

1. Biancolillo A, Maggi MA, De Martino A, Marini F, Ruggieri F, D'Archivio AA (**2020**) Authentication of PDO saffron of L'Aquila (Crocus sativus L.) by HPLC-DAD coupled with a discriminant multi-way approach. *Food Control*, **110**, 107022.
2. Bro R (**1996**) Multiway calibration. Multilinear PLS. *Journal of Chemomometrics*, **10**, 47–61.
3. Fisher RA (**1936**) The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, **7**, 179–188.
4. Sjöström M, Wold S, Söderström B (**1986**) PLS discriminant plots. In *Pattern Recognition in Practice*, 461–470.

Nuclear Magnetic Resonance and Chromatography Data Fusion approach for authentication and traceability of Italian hazelnuts

M. Sozzi[1], M. R. Chierotti[2], L. Lucini[3], B. Senizza[3], F. Savorani[1], R. Gobetto[2]

[1] Polytechnic of Turin, Department of Applied Science and Technology,
Corso Duca degli Abruzzi, 24 – 10129 Turin, Italy
[2] University of Turin, Department of Chemistry,
Via Pietro Giuria, 7 – 10125 Turin, Italy
[3] Sacred Heart Catholic University, Department for Sustainable Food Process,
Via Emilia Parmense, 84 – 29122 Piacenza, Italy
E-mail: mattia.sozzi@polito.it

In the past years, the interest towards metabolomics has been considerably gaining momentum. Its application ranges from medicine to environmental science, also including life science, agriculture, and food science. In the latter, robust and sensitive methodologies need to be developed and continuously updated to assess traceability and to guarantee quality and safety of a huge variety of food products [1].

Accordingly, NMR-based and LC/MS-based approaches have been established as useful tools in Foodomics studies. The high reproducibility of the NMR technique, combined with its low difficulty in the automatization of the entire analytical process, even with very large numbers of samples, makes it suitable for high-throughput analysis [2]. On the other hand, LC/MS combined techniques show a remarkable sensitivity and excellent identification capabilities [3].

In this perspective, the present study proposes a promising NMR–LC/MS combined approach applied to traceability and authenticity of an Italian food excellence, the "Tonda Gentile Trilobata" (TGT) hazelnut variety, which in 1996 obtained the Protected Geographical Indication (PGI) collective label from the European Union [4].

Using ^{1}H solution nuclear magnetic resonance (^{1}H-NMR) spectroscopy and ultra-high-performance liquid chromatography coupled with electrospray ionization quadrupole time-of-flight mass spectrometry (UHPLC-ESI/QTOF-MS) we analyzed 54 hazelnuts samples of different origin, variety, and production year (2020 and 2021). All samples were finely ground and treated with a methanol extraction process. Then, different volumes of the extracted liquid were submitted to the two analytical techniques.

Data originating from both techniques were imported into MATLAB environment to be analysed with multivariate statistical tools. The two datasets, i.e. the spectral data from NMR analysis and the quantified compounds from LC/MS, were firstly treated separately with different pre-processing methods, mean centering and autoscaling respectively, and then investigated with principal component analysis (PCA). The PCA models of the two separate datasets revealed interesting clusters related to production year, but also groupings referable to Piedmont origin and to the TGT variety. For these reasons, we decided to apply a data fusion approach by creating a new joined dataset containing data from both techniques. Because of the diversity of the PCA results obtained from NMR and LC/MS, we chose a mid-level data fusion approach. The new dataset was therefore obtained combining the scores of the first 10 principal components of each PCA model, i.e., the individual model of each characterization technique.

The new fused dataset was explored with PCA and then used to build three different classification models using partial least square discriminant analysis (PLS-DA). The obtained cross-validated models, with different performances and robustness, were able to distinguish year of harvest, hazelnuts of Piedmont from other origins and TGT variety from other varieties.

Acknowledgments

Gruppo Italiano Discussione Risonanze Magnetiche (GIDRM)

References

Cevallos J. M., Reyes-De-Corcuera J. *et al* (**2009**) Metabolomic analysis in food science: a review. *Trends in Food Science and Technology*, **20**, 557-566.

Emwas A., Wishart D. S. *et al* (**2019**) NMR Spectroscopy for Metabolomics Research. *Metabolites*, **9**, 123.

Herrero M. *et al* (**2013**) Metabolomics approaches based on mass spectrometry for food safety, quality and traceability. *Trends in Analytical Chemistry*, **52**, 74-87.

Caligiani A., Arlorio M. *et al* (**2014**) Application of ^{1}H NMR for the characterization and authentication of "Tonda Gentile Trilobata" hazelnuts from Piedmont (Italy). *Food Chemistry*, **148**, 77-85.

Quantification of recycled PET in commercial bottles by IR spectroscopy and chemometrics

Nicholas Kassouf[1], **Alessandro Zappi**[1], **Alessandra Biancolillo**[2],
Pietro Morozzi[3], **Valentina Marassi**[1], **Dora Melucci**[1]

[1]University of Bologna, Department of Chemistry "G.Ciamician", Bologna, Italy
[2]University of L'Aquila, Department of Physical and Chemical Sciences, L'Aquila, Italy
[3]COOP ITALIA Soc. Cooperativa, Bologna, Italy

This work presents a novel approach for the quantification of recycled polyethylene-terephthalate (r-PET) in commercial bottles [1]. Several bottle samples of different brands, producers, and with different percentages of r-PET were purchased from the market or directly supplied from the producers.

Samples were analyzed by IR spectroscopy in two different settings: near-infrared (NIR) spectroscopy and attenuated total reflection (ATR) spectroscopy in mid-IR (MIR). The spectra were then used to compute partial-least squares (PLS) regressions, with the aim of create quantitative models for the prediction of the percentage of r-PET. A multi-block regression was also applied to join the two datasets and to strengthen the quantitative analysis.

All models were validated by cross-validation and by excluding and projecting onto the model the replicated spectra of one brand at a time. Results are promising, also considering the variability that commercial samples can have in terms of additives, shape, or thickness of the bottles: for samples close to the centroids of the models (i.e.: from 10 to 50% r-PET) the predictions seldom departed from the expected values of ±10%. Moreover, no chemical pre-treatment was applied before analyses, making them fast, simple, and cheap.

References

[1] Peñalver R, Marín C, Arroyo-Manzanares N, Campillo N, Viñas P (**2022**). Authentication of recycled plastic content in water bottles using volatile fingerprint and chemometrics, *Chemosphere*, **297**, 134156

NIR spectroscopy and chemometrics against food fraud: spotting Mechanically Separated Meat (MSM) in processed meat products

Elena Cazzaniga[1*], Nicola Cavallini[1], Alessandro Giraudo[1], Marzia Pezzolato[2], Francesco Geobaldo[1], Francesco Savorani[1]

1 Polytechnic of Turin, Department of Applied Science and Technology, Corso Duca degli Abruzzi 24, 10129 Torino, Italy
2 Istituto Zooprofilattico Sperimentale del Piemonte, Liguria e Valle d'Aosta, via Bologna 148, 10154 Torino, Italy
E-mail elena.cazzaniga@polito.it

Food fraud has been identified as an emerging risk to the food industry and as a significant concern to consumers [1]. Any food chain can be affected by food frauds, with no exception for the meat production chain: one of the most significant cases is represented by processed meat, where it becomes very difficult to distinguish the different meat source and it is therefore very easy to incur fraud. One product that shows evidence of such fraud are sausages produced from minced meat that can be mixed or replaced with Mechanically Separated Meat (MSM) [2]. The latter process works by exploiting high pressure to separate the still edible meat tissue, remaining after slaughter, from the bone, obtaining mashed meat which is then mixed with other ingredients, shaped in sausages and cooked. MSM products allow a reduction of the final price of the meat products, as it is of lesser food quality and safety if compared to selected meat cuts. The use of undeclared MSM in the production of sausages and bratwursts is therefore a fraud and may pose a safety risk.

Several analytical techniques have been tested with the aim of distinguishing products containing MSM from those only containing minced meat, but the main problem is the fact that those that proved successful are expensive and time-consuming [3]. In the present work, we inspected the use of NIR spectroscopy as a rapid and non-destructive tool for the investigation of MSM products. Three different NIR instruments were used for this purpose: a benchtop full range NIR spectrometer – MPA (Bruker), a pocket-size spectrometer – SCiO (Consumer Physics), and a portable one – MicroNIR (Viavi). The acquired data were processed with chemometric tools to elaborate explorative and classification models to enable the distinction between MSM and non-MSM products.

These models were built using Principal Component Analysis (PCA) and Partial Least Square–Discriminant Analysis (PLS-DA), respectively. Both explorative and classification models provided good results in distinguishing between MSM and non-MSM, with both classification and prediction accuracies in cross-validation (CV) higher than 95% for all the three analytical techniques.

References

[1] Barrere V, Everstine K, Théolier J, Godefroy S (2020) Food fraud vulnerability assessment: Towards a global consensus on procedures to manage and mitigate food fraud, Trends Food Science and Technology, **100**, 131–137

[2] Pospiech M, Zikmund T, Javůrková Z, Kaiser J, Tremlová B (2019) An Innovative Detection of Mechanically Separated Meat in Meat Products, *Food Analytical Methods*, **12**, 652–657.

[3] Wilhelm C, Hofsommer M, Wittke S (2022) Detection of Mechanically Separated Meat from Chicken in Sausages and Cold Meat by Targeted LC–MS/MS Analysis, *Food Analytical Methods*, 15, 1899–1908.

Comparative analysis of ROI-MCR and Compounds Discover protocols for preprocessing LC-MS signals in untargeted metabolomic studies of Parmigiano Reggiano cheese

Caterina Durante, Samuele Pellacani, Marina Cocchi, Lorenzo Strani

University of Modena and Reggio Emilia, Dept. Chemical and Geological Sciences, via Campi 103, Modena, 41125, Italy
caterina.durante@unimore.it

In recent years, the advancement of high-resolution mass spectrometry technologies has enabled the development of untargeted approaches for food product analysis [1,2], which do not require prior knowledge of the compounds being investigated, leading to the discovery of unexpected or unidentified compounds. However, in the use of liquid chromatography-mass spectrometry (LC-MS) or hyphenate techniques in general, the produced signals can reach sizes up to gigabytes and for their analysis, it is necessary to reduce the file size and compress the information to avoid memory issues on limited computers. Compression must be carefully studied to prevent the loss of useful information, eliminating possible sources of error or spurious variability. Furthermore, it is of utmost importance to use methodologies able to resolve overlapped signals and depict 'identitation' features.

In this study, based on an untargeted analysis, LC-MS signals, obtained from the metabolic profile analysis of Parmigiano Reggiano DOP cheese samples, were preprocessed using two different methodologies, namely the ROIMCR protocol [3] and the Compounds discover software, highlighting the strengths and weaknesses of both procedures. Finally, the characteristic features extracted from both procedures were subjected to Principal Component Analysis to differentiate Parmigiano Reggiano samples based on their geographic origin.

References

[1] L. Laghi et al. *TrAC* 59 (2014) 93–102
[2] M. Herrero et al. *Mass Spectrometry Reviews*, 2012, 31, 49–69
[4] Gorrochategui et al. *BMC Bioinformatics* (2019) 20:256

POSTERS

P01	Adrian	Gomez Sanchez	P46	Nicola	Caporaso
P02	Aina	Queral-Beltran	P47	Nicola	Cavallini
P03	Alessandro	D'Alessandro	P48	Oliver Aleksandrei	Polushkina
P04	Alessandro	Monari	P49	Paolo	Berzaghi
P05	Alessandro	Zappi	P50	Raimundo	Gargallo
P06	Ana Maria	Jimenez Carvelo	P51	Sabina	Licen
P07	Andrea	Balivo	P52	Samuele	Pellacani
P08	Andrea	Botton	P53	Sarah	Curro
P09	Andres	Cruz Conesa	P54	Stefano	Fornasaro
P10	Ane	Bordagaray	P55	Stiufiuc	Rares Ionut
P11	Beatriz	Quintanilla-Casas	P56	Tiziana	Forleo
P12	Berta	Torres Cobos	P57	Victor Hugo	Cavalcanti Ferreira
P13	Blanche	Krieguer	P58	Luis Antonio	Sarabia
P14	Claudia	Scappaticci	P59	Magalie	Claeys-Bruno
P15	Consuelo	Pizarro	P60	Magalie	Claeys-Bruno
P16	Davide	Ballabio	P61	Maria Cruz	Ortiz
P17	Elia	Arnese Feffin	P62	Raffaele Emanuele	Russo
P18	Elisa	Robotti	P63	Tommaso	Paravano
P19	Eloisa	Bagnulo	P64	Adrian	Venegas-Reynoso
P20	Eneko	Lopez	P65	Fabrizio	Mastrolorito
P21	Fabiola	Pereira Verbi	P66	Benoit	Jaillais
P22	Flavia	Bisozzi	P67	Daniele	Tanzilli
P23	Flavio	Montenegro de Andrade	P68	Giorgia	Sciutto
P24	Giacomo	Squeo	P69	Paolo	Oliveri
P25	Gianmarco	Barberi	P70	Pier Lorenzo	Rolando
P26	Giulia	Gorla	P71	Pietro	Bertani
P27	Gorka	Albizu Martinez	P72	Veronica	Ferrari
P28	Harun	Hano	P73	Arsenio	Muñoz de la Pena
P29	Henry	Mac Keown	P74	Edenir Rodrigues	Pereira Filho
P30	Isabel	Duran Meras	P75	Eduardo	Caballero Saldivar
P31	Jokin	Ezenarro	P76	Giorgia	Foca
P32	Jordi	Cruz	P77	Jokin	Ezenarro
P33	Jordi	Riu	P78	Jokin	Ezenarro
P34	Jordi	Riu	P79	José Antonio	Cayuela-Sànchez
P35	Laurence	Dujourdy	P80	Josè Luis	Gonzalez Solis
P36	Lenka	Halamkova	P81	Jose Maria	Gonzalez-Saiz
P37	Leonardo	Sibono	P82	Kateryna	Tkachenko
P38	Lorenzo	Castellino	P83	Kurt	Varmuza
P39	Lorenzo	Strani	P84	Maria Suliany	Rodriguez Barrios
P40	Lorenzo	Strani	P85	Marina	Cocchi
P41	Manuela	Mancini	P86	Roberto	Todeschini
P42	Martina	Fattobene	P87	Gaelle	Bordes
P43	Martina	Foschi	P88	Lorenza	Broccardo
P44	Miriam	Medina Garcia	P89	Qiang	Zhu
P45	Monica	Casale	P90	Riccardo	Voccio

Trilinearity constraint for incomplete data sets

Adrián Gómez-Sánchez[1,2], **Cyril Ruckebusch[2]**, **Anna de Juan[1]**

3. *Chemometrics group. Universitat de Barcelona. Martí i Franquès, 1. 08028. Barcelona.*
4. *LASIRE. Laboratory of Advanced Spectroscopy, Interactions, Reactivity and Environment Université Lille, CNRS, UMR 8516, Cité Scientifique, Bâtiment C5, 59000, Lille, France*

E-mail: aderegomez@gmail.com

Trilinear decomposition of multi-way data sets is a powerful chemometric technique that offers a clear advantage over other methods by providing unique solutions[1]. In the analytical chemistry field, there are several techniques, such as the Excitation-emission fluorescence matrices (EEM), which are widely used in chemical analysis and provide a trilinear structure based on the data configuration of excitation, emission, and sample modes[2]. EEM are regarded as one of the best-known paradigms of chemical measurements and their trilinear structure enables the identification of underlying factors and the exploration of the latent components. By applying decomposition techniques such as Parallel Factor Analysis (PARAFAC)[3] or Multivariate Curve Resolution-Alternating Least Squares (MCR-ALS)[4] with trilinear constraint, the information contained in EEM can be effectively extracted and utilized for a range of analytical and predictive purposes. However, in the case of EEM measurements, no fluorescence emission signal is recorded below the current excitation wavelength and such a fact results in a systematic pattern of missing data values. This characteristic poses a significant problem to the classical implementations of trilinear models in MCR-ALS or PARAFAC, since those rely on complete data sets. Dealing with the challenge of missing values in EEM data is a topic of active research, and several approaches have been proposed to address this issue. One approach involved the use of data imputation in PARAFAC[5], which has demonstrated a good performance in handling missing values. However, it should be noted that imputation methods can sometimes suffer from poor convergence because of the portion of missing values on the data and the systematic pattern, which can lead to suboptimal results[5].

On the other hand, a recent methodology based on implementing the trilinear constraint sequentially on complete subareas of the EEM landscape by Singular Value

Decomposition (SVD) in MCR-ALS has been published with good results[6]. This methodology avoids the imputation step, but requires several optimized factorizations on the EEM landscape of the pure components.

In this work, we propose a novel alternative to the traditional sequential SVD approach used to implement the trilinearity constraint in MCR-ALS, which is based on the use of the Nonlinear Iterative Partial Least Squares (NIPALS)[7] algorithm. NIPALS is applied calculating the first principal component linked to the pure EEM landscape to be constrained extracting the score and loading values doing a row-by-row and column-by-column calculation that plays using only the non-missing values and avoids the multiple SVD analyses used in the previously proposed approach. In this way, NIPALS speeds up the computation time and provides a very simple and scalable constraint to be applied on higher order tensors. With these advantages, MCR-ALS with the NIPALS-based trilinear constraint has the potential to become a promising tool for analyzing N-dimensional data sets with missing values. The performance of this new constraint has been tested on simulated data and experimental data sets.

References

1. Kruskal, J. B. (**1977**). Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear algebra and its applications*, 18(2), 95-138.
2. Marín-García, M., Tauler, R. (**2020**). Chemometrics characterization of the Llobregat river dissolved organic matter. *Chemometrics and Intelligent Laboratory Systems*, 201, 104018.
3. Bro, R. (1997). PARAFAC. Tutorial and applications. *Chemometrics and intelligent laboratory systems*, 38(2), 149-171.
4. De Juan, A., Tauler, R. Multivariate Curve Resolution: 50 years addressing the mixture analysis problem–A review. Analytica Chimica Acta, 2021, vol. 1145, p. 59-78.
5. Tomasi, G., Bro, R. (**2005**). PARAFAC and missing values. *Chemometrics and Intelligent Laboratory Systems,* 75(2), 163-180.
6. Gómez-Sánchez, A., Alburquerque, I., Loza-Álvarez, P., Ruckebusch, C., de Juan, A. (**2022**). The trilinear constraint adapted to solve data with strong patterns of outlying observations or missing values. *Chemometrics and Intelligent Laboratory Systems*, 231, 104692.
7. Wold, H. (**1975**). Soft modelling by latent variables: the non-linear iterative partial least squares (NIPALS) approach*. Journal of Applied Probability*, 12(S1), 117-142.

MCR-ALS multiset incomplete data block analysis of the sulfamethoxazole degradation by chlorination, photodegradation, Cl/UV, and pyrolysis.

Aina Queral-Beltran, Sílvia Lacorte, and Romà Tauler

Addresses: Institute of Environmental Assessment and Water Research (IDAEA-CSIC),
Jordi Girona 18-26, 08034, Barcelona, Spain
E-mail: aina.queral@idaea.csic.es

Sulfamethoxazole (SMX) has been used for many years as an antibiotic and has become a water contaminant worldwide. It is a bioactive contaminant with elevated risk due to its long-term persistence and difficulty to eliminate in conventional water treatment technologies. It is of paramount importance to develop methodologies for its removal and understanding the mechanism by which SMX degrades, and the transformation products generated.

The objectives of the present study were to evaluate the degradation of SMX in four different ways: a) by chlorination (Cl), adding a solution of hypochlorite solution which contains free chlorine; b) by photodegradation (UV), using a Suntest chamber which emits UV radiation; c) by the combination of both chlorine and photodegradation at the same time (Cl/UV), and finally d) by pyrolysis (Py). The reactions in which chlorine or UV light was employed were monitored spectrophotometrically. In parallel, aliquots were taken at different reaction times and analysed by liquid chromatography with a diode array detector and mass spectrometry in positive and negative ionization modes (LC-DAD-MS(+)-MS(-)). On the other hand, the degradation by pyrolysis was performed *in-situ* by Py-gas chromatography coupled to an Orbitrap mass spectrometer (GC-Orbitrap-MS).

The high amount of spectrophotometric and chromatographic data obtained in the different experiments were fused and analysed using Multivariate Curve Resolution-Alternating Least Squares (MCR-ALS)[1], in which the different blocks of data were fused and integrated. This arrangement allowed the comparison of the four different types of degradation processes (chlorination, light, their combination, and pyrolysis), to resolve the different chemical species produced and, for each of them, their kinetic profiles, elution profiles, UV-Vis spectra, MS(+) and MS(-) spectra and Orbitrap-MS. This information was compared with previous works[2,3] and allowed the identification of the compounds generated.

References

1. de Juan A, Tauler R. Multivariate Curve Resolution-Alternating Least Squares for Spectroscopic Data. *Data Handling in Science and Technology*. 2016;**30**:5-51. doi:10.1016/B978-0-444-63638-6.00002-4.
2. Marín-García M, de Luca M, Ragno G, Tauler R. Coupling of spectrometric, chromatographic, and chemometric analysis in the investigation of the photodegradation of sulfamethoxazole. *Talanta*. 2022;**239**. doi:10.1016/j.talanta.2021.122953.
3. Queral-Beltran A, Lacorte S, Tauler R. UV absorption spectrophotometry and LC-DAD-MS coupled to Chemometrics analysis of the monitoring of the degradation process of Sulfamethoxazole drug by Chlorination, Photodegradation, and UV/Chlorination Advanced Oxidation Processes. *(to be submitted)*.

Multivariate Curve Resolution (MCR) applied to hyperspectral images: case study on Pesto alla Genovese sauce

Alessandro D'Alessandro[1,2], **Marina Cocchi**[1], **Daniele Tanzilli**[1], **Lorenzo Strani**[1], **Jean Michel Roger**[3], **Silvia Mas-Garcia**[3] , **Maxime Ryckewaert**[3]

Addresses:
[1] *Department of Chemical and Geological Sciences*
University of Modena and Reggio Emilia, Via Campi 103, 41125 Modena, Italy
[2] *Research manager at Barilla G. e R. Fratelli, Via Mantova 166, 43122 Parma, Italy*
[3] *INRAe ChemHouse, 361 rue Jean-François Breton, 34196 Montpellier, France*
E-mail alessandro.dalessandro@unimore.it
alessandro.dalessandro@barilla.com

In the food industry, to evaluate the characteristics of products, there is an increasing need of fast and non-destructive analytical methods, especially for on-line monitoring.

Hyperspectral imaging is a powerful methodology joining the possibility to describe the morphological characteristics of the sample (the image) to acquiring detailed chemical information (the spectra available for each single pixel). In fact, with respect to digital images where the three red, green, and blue channels are acquired (RGB images), in hyperspectral imaging for each pixel a whole spectrum in visible and or NIR range is acquired.

In the present work, a preliminary study has been done on the evaluation of Italian "Pesto alla Genovese" sauces. "Pesto alla Genovese" is a typical Italian sauce containing mainly basil, olive oil, cheese, and other minor ingredients. It has a complex structure including several phases: water, oil in water emulsion, solid suspension, and a protein matrix. The evaluation of the structure became relevant in the cases where pesto loses the stability, where typically the emulsion breaks releasing oil with an unwanted evident phase separation.

Hyperspectral imaging in visible and NIR range coupled with proper data elaboration is potentially able to give information about the chemical components and their distribution on the imaged surface. In present study were used two hyperspectral cameras covering the spectral ranges from 409 to 987 nm and from 964 to 2494 nm, respectively. In each hyperspectral image was then collected a region of interest (ROI) and images spectra were pre-processed. Images were then unfolded to reduce dimensions from three to two. Multivariate Curve Resolution (MCR) where the applied to individuate the different chemical (matrices) classes present.

Reconstructing the 3D images is possible to individuate the location and the concentration of each component.

References

Jaumot J, de Juan A, Tauler R (**2015**) MCR-ALS GUI 2.0: new features and applications. *Chemometrics and Intelligent Laboratory Systems*, **140**,1-12

Amigo JM, Martí I, Gowen A (2013) Chapter 9 - Hyperspectral Imaging and Chemometrics: A Perfect Combination for the Analysis of Food Structure, Composition and Quality. *Data Handling in Science and Technology,* **28,** 343-370.

Lleó L, Roger JM, Herrero-Langreo A, Diezma-Iglesias B, Barreiro P (2011) Comparison of multispectral indexes extracted from hyperspectral images for the assessment of fruit ripening. *Journal of Food Engineering*, **104**, 612-620.

NOVEL ELECTRO-ANALYTICAL APPROACH FOR THE DETECTION AND DISCRIMINATION OF CANNABINOIDS IN CANNABIS EXTRACTS

Alessandro Monari[1], Barbara Zanfrognini[1], Giorgia Foca[2], Alessandro Ulrici[2], Chiara Zanardi[3], Laura Pigani[1]

[1]University of Modena and Reggio Emilia, Department of Chemical and Geological Sciences, via G. Campi 103, 41125 Modena, Italy
[2]University of Modena and Reggio Emilia, Department of Life Sciences, via Amendola 2, 42122 Reggio Emilia, Italy
[3]Ca' Foscari University of Venice, Department of Molecular Sciences and Nanosystems, Via Torino 155, 30170 Venezia Mestre, Italy
E-mail: alessandro.monari@unimore.it

CRYPTED FOR CONFIDENTIALITY REASON

Chemometrics in environmental chemistry: application of self-organizing maps for the study of Saharan dust events

Alessandro Zappi[1], **Sara Padoan**[2], **Laura Tositti**[1], **Thomas Adam**[2]

[1]University of Bologna, Department of Chemistry "G. Ciamician", Bologna, Italy
[2]University of the Bundeswehr Munich, Faculty of Mechanical Engineering, Institute of Chemistry and Environmental Engineering, Neubiberg, Germany
alessandro.zappi4@unibo.it

The association between environmental chemistry and chemometrics has deep and consolidated roots. Environment indeed, is characterized by an extreme degree of chemodiversity, with highly expressed time-variability (especially in the fluid phases), huge inhomogeneity over a massive size as compared to the laboratory scale, always claiming large datasets. Therefore, chemometric modeling is a necessary step for both understanding environmental complexity and diagnostic purposes.

The present work proposes the use Self-Organizing Maps (SOM)[1] for diagnostic purposes, i.e. to detect the influence of Saharan dust (SD) events in March 2022 in Munich, Germany, an unusual occurrence for this type of transport at these latitudes. Munich, indeed, besides being a heavily man-impacted city in southern Germany, is located beyond the Alps and is therefore rarely reached by Northern African air masses loaded in Saharan mineral dust, differently from the Mediterranean basin.

These events, however, are increasing in frequency and intensity, sometimes reaching latitudes as high as the UK, as a possible result of climate change. During these events, huge alterations of PM_{10} concentrations, often exceeding the EU air quality standards as well as of composition are observed leading to increased environmental and health hazard[2].

In this study, particulate matter (PM) was collected daily on quartz fiber filters from March to May 2022, and its metal composition was evaluated by Inductively Coupled Plasma Mass-Spectrometry (ICP-MS). After basic data pre-processing, data was then subjected to SOM with the aim of evaluating the enrichment of metal concentrations due to the presence of SD over Munich. Although Positive Matrix Factorization would have been more appropriate to achieve source apportioning of airborne particulate matter, it requires a huge amount of data to compute reliable models. Though many other multivariate models are usually applied in the case of limited matrices of data, SOM has been proved as a valid and reliable alternative to such methods, due to its simpler and faster computational procedure that can be carried out with any number of samples.

SOM results, also assisted by meteorological data and physical transport-based models, as backtrajectory analysis, successfully demonstrated not only the differences in trace element composition of PM in Munich due to the advent of SD, but even some differences in SD composition according to different source locations as a result of evolving atmospheric dynamics.

Acknowledgements

This research is funded by dtec.bw – Digitalization and Technology Research Center of the Bundeswehr [project MORE and LUKAS]. dtec.bw is funded by the European Union – NextGenerationEU.

References

Licen S, Franzon M, Rodani T, Barbieri P (**2021**) SOMEnv: An R package for mining environmental monitoring datasets by Self-Organizing Map and k-means algorithms with a graphical user interface, *Microchemical Journal*, **165**, 106181

Tositti L, Brattich E, Cassardo C, Morozzi P, Bracci A, Marinoni A, Di Sabatino S., Porcù F, Zappi A (**2022**) Development and evolution of an anomalous Asian dust event across Europe in March 2020, *Atmospheric Chemistry and Physics*, **22**, 4047–4073

TOWARDS THE CREATION OF A GLOBAL HARMONISED DATABASE AND A SINGLE ANALYTICAL MULTIVARIATE METHOD - TEQUILA AUTHENTICATION AS A CASE STUDY

Ana M. Jiménez-Carvelo[a], **Christian H. Pérez-Beltrán**[a], **José J. Olmos-Espejel**[b], **Guadalupe Pérez-Caballero**[b], **Luis Cuadros-Rodríguez**[a]

[a] *University of Granada, Department of Analytical Chemistry, Fuentenueva St, Granada, E-18071, Spain*

[b] *Universidad Nacional Autónoma de México, Unidad de Investigación Multidisciplinaria, campus 4, 54714, México*

amariajc@ugr.es

The authenticity assurance of foodstuff around the world is of great concern, since food fraud continues to be present, nationally or internationally, along the supply chain of food products [**1**]. Despite the availability of official analytical methods and quality assessment bodies, food fraud cases are constantly arising, such as adulteration or counterfeiting of olive oil [**2**], tequila [**3**] and more [**1**]. Thus, new analytical methods should be adopted to increase the food fraud detection and to strengthen the current quality control process of foodstuff among different routine analytical laboratories.

In this regard, the current study demonstrates the potential of the tandem *instrument-agnostizing* [**4,5**] and non-target fingerprinting methodologies [**6**] to transfer chromatographic signals (fingerprints) among laboratories to verify the stated quality of the food product in concern. In this framework, the reliability of the chromatographic signals of White Tequila samples, acquired from two different analytical instruments from two laboratories, is verified: first, applying the *ad hoc* developed 'Equity Function'; and, then, *agnostizing* the chromatographic signals (i.e., standardizing both intensities and retention times) of White Tequila samples. Afterwards, the chromatographic signals of each sample, before and after *agnostizing*, are compared using the nearness similarity index (NEAR) and cosine angle (COS). Finally, the *instrument-agnostizing* chromatographic signals are used to create a global database and a single PLS-DA or SVM multivariate mathematic model to authenticate the '100% agave' and 'mixed' categories of White Tequila.

Results showed that the critical evaluation of the chromatographic signals using the 'Runtime Ratio' and 'Starting-time Lag' (StL) parameters, as part of the Equity Function, is crucial in order to ensure that they are free of distortion effects. The suitability of the chromatographic signals was demonstrated with the increment of the NEAR and COS values, which was originally lower than 0.4 and higher than 0.7 after correction and *agnostizing*. The performance of the single mathematic model, built from the global *instrument-agnostic* fingerprints database, was better using the SVM model than the PLS-DA model, which allowed to obtain a saving index of 60%, indicating that the correctly classified samples coming from a different laboratory do not require to undergo potential confirmatory analyses.

This study describes the way of creating a global database of *instrument-agnostic* chromatographic signals from two different analytical laboratories, which was used to perform the first single multivariate mathematic model capable to distinguish the quality of two categories of White Tequila. This suggests new horizons for the food quality control activities, which will allow to increase the detection pace of food frauds worldwide.

References

[**1**] Knowledge Centre for Food Fraud and Quality, European Commission (**2021**) Monthly summary of articles on food fraud and adulteration. Access: Feb 7th 2023 https://knowledge4policy.ec.europa.eu/sites/default/files/food_fraud_newsletter_12-2022.pdf

[**2**] Lozano-Castellón J, López-Yerena A, Domínguez-López I, Siscart-Serra A, Fraga N, Sámano S, López-Sabater C, Lamuela-Raventós R.M, Vallverdú-Queralt A, Pérez M (**2022**) Extra virgin olive oil: A comprehensive review of efforts to ensure its authenticity, traceability, and safety, *Comprehensive reviews in food science and food safety*, **21**, 2639–2664.

[**3**] Warren-Vega WM, Fonseca-Aguiñaga R, González-Gutiérrez LV, Romero-Cano L (**2023**) A critical review on the assessment of the quality and authenticity of tequila by different analytical techniques: recent advances and perspectives, *Food Chemistry,* **408**, 135223.

[**4**] Cuadros-Rodríguez L, Ortega-Gavilán F, Martín-Torres S, Medina-Rodríguez S, Jiménez-Carvelo AM, González-Casado A, Bagur-González MG (**2021**) Standardization of chromatographic signals – Part I: Towards obtaining instrument-agnostic fingerprints in gas chromatography, *Journal of Chromatography A*, **1641**, 461983.

[**5**] Cuadros-Rodríguez L, Martín-Torres S, Ortega-Gavilán F, Jiménez-Carvelo AM, López-Ruíz R, Garrido-Frenich A, Bagur-González MG, González-Casado A (**2021**) Standardization of chromatographic signals – Part II: Expanding instrument-agnostic fingerprints to reverse phase liquid chromatography, *Journal of Chromatography A*, ***1641***, 461973.

[**6**] Jiménez-Carvelo AM, Martin-Torres S, Cuadros-Rodríguez L, González-Casado A (**2021**) Non-targeted fingerprinting approaches, in: C.M. Galanakis (Ed.), Food Authentication and Traceability, Academic Press /Elsevier, pp. 163–193.

Authentication of milk from buffaloes fed with hydroponic barley forage using E-nose

Andrea Balivo, Raffaele Sacchi, Felicia Masucci and Alessandro Genovese

Address: University of Naples Federico II, Department of Agricultural Sciences, Via Università 100, Portici (NA,), 80055, Italy
andrea.balivo@unina.it (corresponding author)

The use of fresh forage in the animals' diet improves the nutritional composition of milk and, hence, of dairy products [1]. Grazing or fresh forage cultivation may become less feasible as Mediterranean agriculture is threatened by climate change, drought and scarcity of natural resources such as water and arable land [2]. Recently, hydroponic cultivation system has emerged as a new soilless cultivation method for growing quality-constant fresh forage, offering advantages such as better environmental sustainability, reduced labour costs, lower resource and space requirements [3]. The evaluation of the best nutritional composition of the milk, to avoid possible frauds, has focused on time consuming and laborious analyses such as chromatographic ones. This study aimed to use an E-nose to differentiate raw milk samples from buffaloes fed on maize silage or hydroponic barley forage. 108 samples of raw milk were taken on three different days, from three groups of twelve buffaloes (Italian Mediterranean breed). The control group (C) was fed maize silage and concentrate (60:40). In the first experimental group maize silage was replaced at 50% (LH) by hydroponic barley forage, while in the second group at 100% (HH). The E-nose, equipped with ten metal oxide semiconductor sensors (MOS), was used in combination with statistical data analysis (LDA and PCA) to classify the milk samples. In order to obtain information on the volatile compounds that characterised the milk samples, a SPME-GC/MS analysis was also performed. The confusion matrix for the cross-validation results, obtained from LDA performed on the E-nose data vectors, showed a correct classification rate of 84.3%. Furthermore, the PCA observation plot of the resulting 10-dimensional patterns corroborates the LDA findings, showing good discrimination of the three types of milk samples (Figure 1).

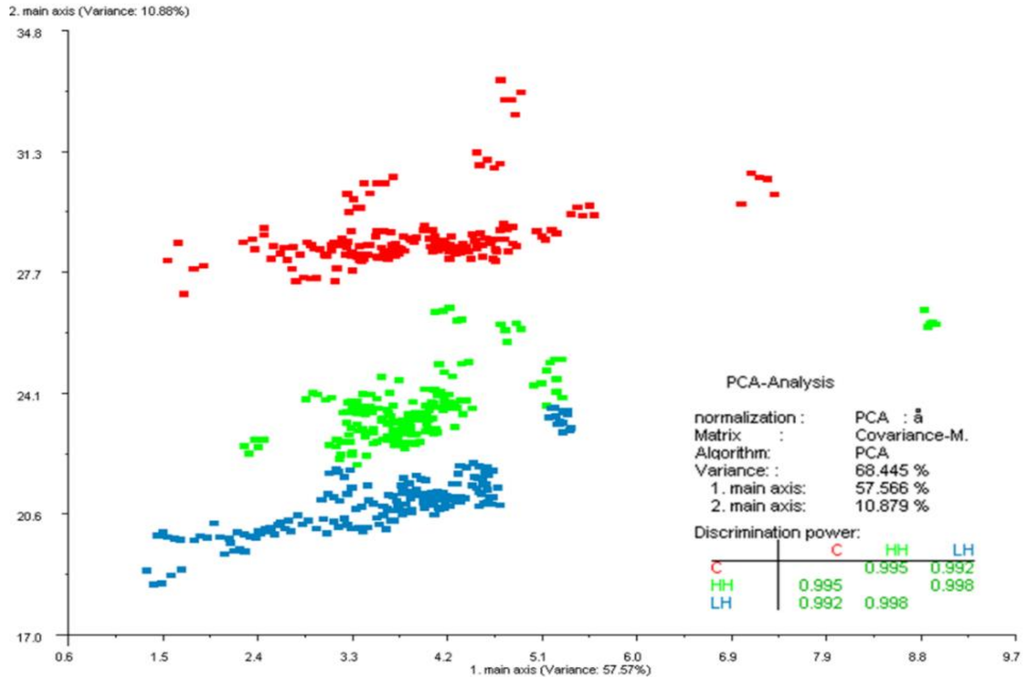


Figure 2 PCA results of electronic nose data vectors of 108 milk samples obtained from maize silage-fed buffaloes (control group, C, in red) and hydroponic forage-fed buffaloes, with 50% (LH, in blue) and 100% (HH, in green) silage replacement percentage. Data were extrapolated in the time range 55-59 s and processed with Winmuster v.1.6 software (Airsense Analytics GmbH, Schwerin, Germany).

This different response of MOS sensors to samples is due to quantitative differences of some volatile compounds. In fact, C milk samples had a higher overall abundance of volatile compounds, particularly for esters and ketones. In conclusion, the evidence from this study implies that E-nose can be a portable, solvent free and relatively inexpensive device for the rapid control of the authenticity of raw milk to protect producers and consumers of dairy products from potential fraud.

References

[1] Balivo A, Sacchi R, Genovese A (**2023**) The Noble Method in the dairy sector as a sustainable production system to improve the nutritional composition of dairy products: A review. *International Journal of Dairy Technology*, **in Press**, DOI:10.1111.1471-0307.12941.

[2] Aguilera E, Diaz-Gaona C, Garcia-Laureano R, Reyes-Palomo C, Guzmán G I, Ortolani L, Sánchez-Rodríguez M, Rodriguez-Estevez V (**2020**) Agroecology for adaptation to climate change and resource depletion in the Mediterranean region. A review. *Agricultural Systems*, **181**, 102809.

[3] Agius A, Pastorelli G, Attard E (**2019**). Cows fed hydroponic fodder and conventional diet: effects on milk quality. *Archives Animal Breeding*, **62**, 517-525.

Virtual Sensor to Forecast the Mooney Viscosity in an Industrial Rubber Production Process

Andrea Botton[1], Pierantonio Facco[1], Angelo Ferrando[2], Francesco Bonacini[2], Erik Mantovani[2], Gianni Marchetti[2]

Addresses: [1] CAPE-Lab, Computer Aided Process Engineering Laboratory, University of Padova, Department of Industrial Engineering, Via Marzolo 9, Padova, 35131, Italy
[2] Versalis s.p.a., Piazza Boldrini 1, San Donato Milanese (MI), 20097, Italy
pierantonio.facco@unipd.it

Objective

In the rubber production process, one of the most important parameters for quality assessment is Mooney viscosity. Appropriate instrumentation and dedicated personnel are usually employed to measure the Mooney viscosity through long-lasting and expensive laboratory analyses [1]. However, in case of anomalies the quality monitoring and the corrective intervention can be significantly delayed by measurements taken off line, leading to unacceptable levels of product quality and downgraded rubber.

Virtual sensors have been developed to estimate in real time the Mooney viscosity from online process measurements, accordingly speeding up the quality monitoring. For example, deep learning methodologies were employed to achieve an acceptable estimation accuracy [2]. These models are very complex to deal with the typical high non-linearities of rubber manufactory. However, they suffer from low interpretability and require huge amounts of data to be built. In this presentation we propose to use a moving window [3] dynamic [4] partial least squares [5] regression model to estimate the product quality of the rubber after the stripping section of the production plant. Not only the proposed virtual sensor is easily interpretable, but it requires also a relatively limited amount of data.

Results and conclusions

The accuracy of the estimation is quite satisfactory (*Mean Relative Error* = 2.3%) and the model typically identifies the top-quality products (i.e., products which meet specification requirements for Mooney) correctly (*Top-quality Identification Rate* = 96.2%). The identification of products to be downgraded (i.e., products which do not meet specification requirements for Mooney) is good (*Downgraded Identification Rate* = 63.2%), but it requires improvement to guarantee a fast and efficient intervention when the quality drifts out of specification. Hence, the virtual sensor increases the efficiency of quality monitoring and ensures a quick identification of out-of-spec products, which allows timely corrective interventions to face faults and malfunctions. As future work, we aim at improving the prediction accuracy of the virtual sensor through model hybridization by means of a process simulator.

References

1. Mark J.E., Erman B., Eirich F.R. (**2005**). Science and Technology of Rubber, 3rd ed. *Elsevier Academic Press*, San Diego, CA, USA.
2. Yang, Y., Lu J. (**2022**). A Fusion Transformer for Multivariable Time Series Forecasting: The Mooney Viscosity Prediction Case, *Entropy*, **24**(4), 528.
3. Jiang J., Barry J., Siezler H. W., Ozaki Y. (**2002**). Wavelength Interval Selection in Multicomponent Spectral Analysis by Moving Window Partial Least-Squares Regression with Applications to Mid-Infrared and Near-Infrared Spectroscopic Data, *Analytical Chemestry*, **74**, 3555-3565.
4. Ku W., Storer R. H., Georgakis C. (**1995**). Disturbance Detection and Isolation by Dynamic Principal Component Analysis, *Chemometrics and Intelligent Laboratory Systems*, **30**, 179-196.
5. Wold S., Sjöström M., Eriksson L. (**2001**). PLS-regression: a basic tool of chemometrics, *Chemometrics and Intelligent Laboratory System*, **58**, 109–130.

Near-infrared spectroscopy (NIRS) to determine apparent metabolizable energy (AME) and apparent ileal phosphorus digestibility (AIDP) in broiler diets

Andres Cruz-Conesa[1,2], Joan Ferré[1], Anna M. Pérez-Vendrell[2], Itziar Ruisánchez[1]

[1]Universitat Rovira i Virgili, Faculty of Chemistry, Department of Analytical and Organic Chemistry, Tarragona, Spain

[2]Institute of Agrifood Research and Technology, Animal Nutrition Más Bové, Constantí, Spain

andres.cruz@urv.cat

Precise feed formulation improves farm animal productivity without causing nutritional excesses or deficiencies. This is achieved by formulating based on parameters that reflect how the animal digests the feed. Two of the most important in broiler formulations are AME and AIDP. These values are obtained through bioassays in which the animal is fed a test diet and its excreta and/or ileal contents are collected. Bioassays are controversial due to the use of animals and expensive, mainly due to the large number of samples to be analysed (Zaefarian et al. 2021). In broiler diets, AME is determined by subtracting the gross energy (GE) of the excreta (obtained by calorimetry) from the GE of the feed and AIDP by subtracting the phosphorus (P) content (obtained by spectrophotometry) in the ileal digesta from the P of the feed. In this work, we evaluate the use of NIRS as an alternative methodology to determine AME and AIDP.

The diets and the excreta/ileal digesta were obtained from studies of the effect of different enzymes on AME and AIDP. Four modelling strategies using partial least squares regression (PLSR) were compared: a) using the feed spectra, b) using the excreta/ileal digesta spectra, c) combining feed and excreta/ileal digesta spectra (Campos et al. 2017) and d) including the phytase concentration as an additional predictor variable in AIDP strategies a, b and c. The standard error of leave one-out cross validation (SECV) was used to compare the models.

For AME, the best predictions were obtained by combining the feed and the excreta spectra (Fig. 1A). This shows that the spectra provide complementary information. Similar conclusions were found by Paternostre et al. 2021 in the prediction of net energy in pig diets. The success of this model means that the determination of GE in diets and excreta in order to find AME may no longer be needed. Although the model using only the feed spectra was less accurate, it could be useful on some occasions since the collection of the excreta and hence bioassays would not be necessary to make AME predictions. For AIDP the best predictions were obtained by combining the ileal digesta spectra with the phytase concentration (Fig. 1B). The reason is that most of the feeds had a very similar P concentration, so feed spectra did not correlate with AIDP variation, which is related to the differences in the digestive capacity of the broilers and to the phytase concentration in the diet. The success of this model means that the determination of P in diets and ileal digesta in order to find AIDP may no longer be necessary.

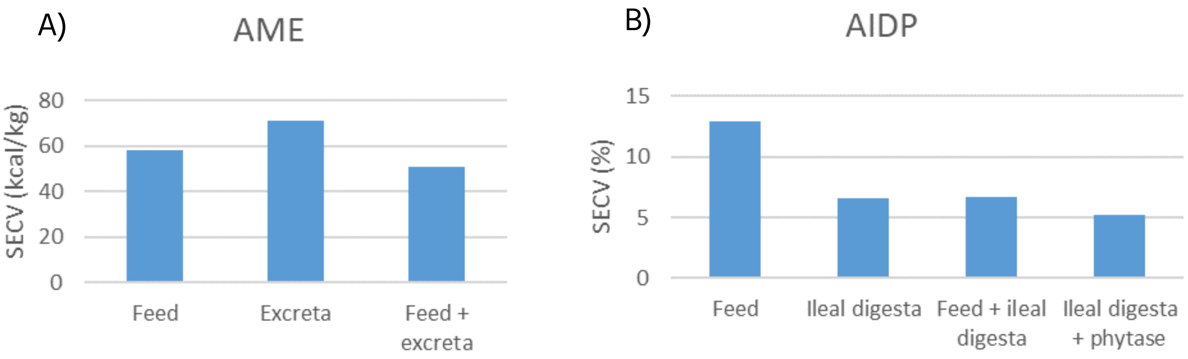


Figure 1. A) SECV of PLSR models for AME developed from a) using the feed spectra, b) using the excreta spectra and, c) combining feed and excreta spectra. B) SECV of PLSR models developed for AIDP from feed spectra, from a) using the feed spectra, b) using the ileal digesta spectra, c) combining feed and ileal digesta spectra and d) combining the ileal digesta spectra with the amount of phytase.

References

Zaefarian F, Cowieson AJ, Pontoppidan K, Abdollahi MR, Ravindran V (**2021**) Trends in feed evaluation for poultry with emphasis on in vitro techniques. *Animal Nutrition*, **7**, 268-281.

Campos MP, Sousa R, Pereira AC, Reis MS (**2017**) Advanced predictive methods for wine age prediction: Part II - A comparison study of multiblock regression approaches. *Talanta*, **171**, 132-142.

Paternostre L, Baeten V, Ampe B, Millet S, De Boever J (**2021**) the usefulness of NIRS calibrations based on feed and feces spectra to predict nutrient content, digestibility and net energy of pig feeds. *Animal Feed Science and Technology*, **281**, 115091.

DETERMINATION OF RIBOFLAVIN, FLAVIN ADENINE NUCLEOTIDE AND FLAVIN MONONUCLEOTIDE PRODUCED BY LACTIC ACID BACTERIA BY FLUORESCENCE SPECTROSCOPY

Irati Berasarte, Ane Bordagaray, Mª Teresa Dueñas, Rosa Garcia-Arrona, Irati Fradejas, Miren Ostra, Maider Vidal[1]

[1]Department of Applied Chemisty, Faculty of Chemistry, University of the Basque Country (UPV/EHU), 20018 Donostia/San Sebastian.
e-mail: ane.bordagaray@ehu.eus

B2 vitamin (riboflavin) and its cofactors are essential compounds required by all life forms. Lactic acid bacteria (LAB) have been widely used as starter cultures for fermentation in a wide variety of foods, since they can improve the nutritional content and quality of fermented foods. Among all lactic acid bacteria, there are certain strains capable of synthesizing B vitamins, including riboflavin. Riboflavin, or vitamin B2, is a precursor of two coenzymes responsible for the transport of electrons in certain redox reactions, so it is essential to take in adequate amounts through the diet. According to literature, different LAB strains have been isolated and their ability to fortify foods such as milk, whey and cereals has been demonstrated [1].

For this reason, nowadays the synthesis of RF and its cofactors has a growing importance in an industrial level, which means that the importance of rapid determinations of these micronutrients is also growing. According to this, the main objective of this work is to design and validate a method for a simultaneous and rapid determination of RF and the cofactors flavin adenine dinucleotide (FAD) and flavin mononucleotide (FMN), based on their most important property: fluorescence.

The three analytes show practically overlapping emission spectra, but their intensity varies as a function of pH (Figure 1).

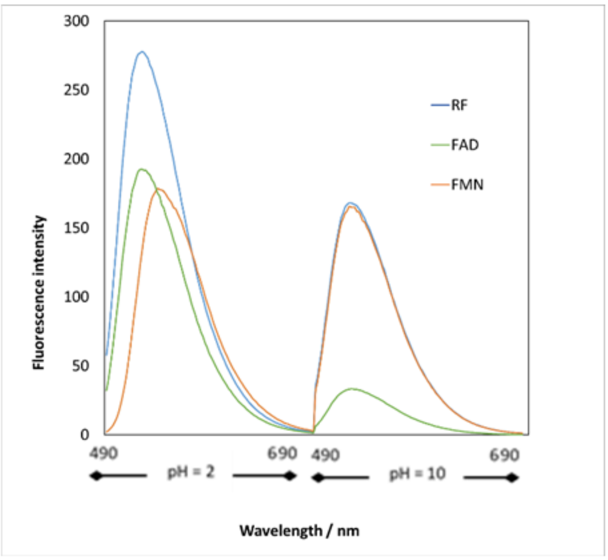


Figure 1. emission spectra at pH 2 and pH 10 for RF and cofactors RF and FAD.

The present work proposes the use of combined signals from mixtures of the three analytes recorded at pH 2 and pH 10 and their subsequent treatment with multivariate calibration algorithms. Excitation-emission fluorescence spectra have been obtained at excitation wavelength between 200 and 700nm and emission wavelength between 490 and 700 nm for mixtures containing RF, FMN and FAD. PLS models have been obtained for unfolded data as well as MCR-ALS models for three-way excitation-emission spectra. The obtained models have been applied for the determination of RF, FMN and FAD produced by LAB in several growing media and results have been compared to those obtained by HPLC.

Acknowledgements

Authors acknowledge the Basque Government (Group IT1662-22) and University of the Basque Country (PIF 19/131pre-doctoral scholarship) for the financial support.

References

[1] Levit, R.; Savoy de Giori, G.; de Moreno de LeBlanc, A.; LeBlanc, J. G. (**2021**) Recent Update on Lactic Acid Bacteria Producing Riboflavin and Folates: Application for Food Fortification and Treatment of Intestinal Inflammation. Journal of Applied Microbiology, **130**, 1412–1424. https://doi.org/10.1111/jam.14854

Handling saturation in untargeted GC-MS data analysis

Beatriz Quintanilla-Casas, Rasmus Bro

University of Copenhagen, Department of Food Science, Rolighedsvej 26
1958 Frederiksberg (Copenhagen), Denmark

Mass spectrometry (MS) is a widely used analytical technique applied in several research fields, such as food and environmental science. Despite of the many advantages this technique provides e.g., high sensitivity, some issues can appear when chemical compounds are present in high concentrations. Among them, saturation of the MS detector affects the quantification of chemical species because the area of the saturated spectrum does not reflect the actual concentration of the sample[1]. When gas chromatography precedes MS, extracted ion chromatograms of saturated masses (*m/z*) result in flat lines at the upper limit (Figure 1) which is the main reason for the deviation from linearity of the area versus concentration.

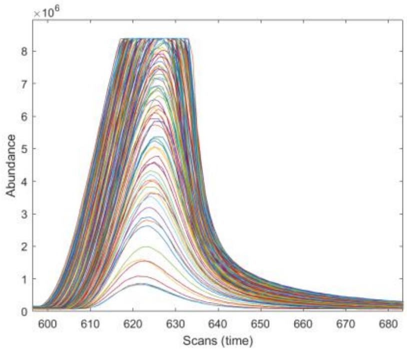


Figure 1. Extracted ion chromatogram of a given saturated *m/z*

We conjecture that such flat lines of saturation can be automatically detected and further that we can develop methods that optimally disregard these artefacts. This will then lead to a linear or near linear relation between imputed area and concentration. These alternative numerical methods will be assessed and incorporated to PARADISe, a PARAFAC2-based Deconvolution and Identification System for untargeted analysis of GC-MS data[2].

References

[1] Jung Wei AA, Joshi A, Chen Y, McIndoe JS. (**2020**) Strategies for avoiding saturation effects in ESI-MS. *International Journal of Mass Spectrometry*, 450, 116306.

[2] Quintanilla-Casas B, Bro R, Hinrich JL, Davie-Martin, CL (**2023**) Tutorial on PARADISe: PARAFAC2-based Deconvolution and Identification System for processing GC-MS data. *Available at Protocol Exchange*.

Nuttin' to hide: A showdown between Fingerprinting and PARADISe to uncover the secrets of hazelnut authenticity

B. Torres-Cobos[1,2], B. Quintanilla-Casas[1,2], M. Rovira[3], A. Romero[3], F. Guardiola[1,2], A. Tres[1,2], S. Vichi[1,2]

1 University of Barcelona, Department of Nutrition, Food Sciences and Gastronomy, Prat de la Riba 171, Santa Coloma de Gramenet, 08921, Spain.
2 University of Barcelona, Institute of Research on Food Nutrition and Safety (INSA-UB), Prat de la Riba 171, Santa Coloma de Gramenet, 08921, Spain.
3 Institute of Agrifood Research and Technology (IRTA), Ctra. de Reus – El Morell Km 4,5, Constantí, 43120, Spain.
bertatorres@ub.edu

Hazelnuts are a highly valued and popular ingredient in both sweet and savoury products, with their sensory and qualitative attributes greatly influenced by cultivar and geographical origin [1]. Their high value makes them susceptible to economically motivated fraud, further aggravated by the lack of effective fraud detection methods. Suitable tools for verifying the cultivar and origin of hazelnuts are critical to guarantee authenticity and protect consumers.

Metabolomics, focusing on the unsaponifiable fraction, has been proposed as a promising tool for both varietal and geographical authentication as several secondary metabolites from this fraction such as hydrocarbons, linear and terpene alcohols, and sterols, are known to be significantly influenced by the cultivar and environmental conditions [1,2]. Its analysis with/through gas chromatography mass spectrometry (GC-MS), provides comprehensive molecular-level information through 3-way data with high sensitivity.

The data treatment and chemometric analysis are crucial parts of the authentication strategy. Rather than using traditional targeted methods, untargeted approaches offer more information and are better suited for analyzing complex chromatograms as the ones of the unsaponifiable fraction of nuts, since they overcome the difficulties of identifying and quantifying analytical compounds [3].

In the present study two state-of-the art untargeted methods are compared to authenticate hazelnuts. One of them is fingerprinting, which consists on finding specific patterns, known as fingerprints, that are unique to a specific characteristic of the food sample, such as cultivar or geographical origin. Fingerprinting of three-way data, such as GC-MS data, typically entails complex multi-way chemometric algorithms, but the process can be simplified by transforming the data into a manageable two-way format, an unfolded matrix. An alternative is using advanced profiling techniques, such as powerful deconvolution tools to extract maximum information from the samples. The second method evaluated is a user-frienfly deconvolution and identification system for analysing high-dimensional data that has recently emerged, called PARADISe, which is based on PARAllel FACtor analysis 2 (PARAFAC2). PARADISe enables the analysis of three-way chromatographic data by providing pure mass spectra and determining the relative concentrations of each chemical compound in the sample peaks [4].

Both data processing techniques are compared to determine the most suitable method for developing efficient hazelnut authentication models. For this purpose, classification models based on GC-MS data of hazelnut unsaponifiable fraction using both fingerprinting and PARADISe methods are built. The performance of the models is assessed through the percentage of correct classification, sensitivity and specificity in an external validation.

References

[1] M. Özdemir, F. Ackurt, M. Kaplan, M. Yildiz, M. Löker, T. Gürcan, G. Biringen, A. Okay (**2001**) Evaluation of new Turkish hybrid hazelnut (Corylus avellana L.) varieties: fatty acid composition, α-tocopherol content, mineral composition and stability. *Food Chemistry*, **73**, 411-415

[2] S. V. Goriainov, C. A. Esparza, A. R. Borisova, S. V. Orlova, V. V. Vandyshev, Fadi Hajjar, E.A. Platonov, E. P. Chromchenkova, O. O. Novikov, R. S. Borisov, G. A. Kalabin (**2021**) Phytochemical Study of the Composition of the Unsaponifiable Fraction of Various Vegetable Oils by GC–MS. Journal of Analytical Chemistry, **76**, 1635-1644.

[3] B. Quintanilla-Casas, S. Bertin, K. Leik, J. Bustamante, F. Guardiola, E. Valli, A. Bendini, T. G. Toschi, A. Tres, S. Vichi (**2020**) Profiling versus fingerprinting analysis of sesquiterpene hydrocarbons for the geographical authentication of extra virgin olive oils. *Food Chemistry,* **307**, 125556.

[4] G. Baccolo, B. Quintanilla-Casas, S. Vichi, D. Augustijn, R. Bro (**2021**) From untargeted chemical profiling to peak tables – A fully automated AI driven approach to untargeted GC-MS. *TrAC Trends in Analytical Chemistry*, **145,** 116451.

ESR investigation of polymer materials irradiated with gamma and X-rays with SIMPLISMA treatment

Blanche Krieguer[1,2,3], Sylvain R.A. Marque[2], Samuel Dorey[1], Nathalie Dupuy[3], Fabien Girard[3], Nina Perier[1], Florent Kuntz[4], Nicolas Ludwig[4]

[1] Sartorius Stedim FMT S.A.S, Z.I. Les paluds, Avenue de Jouques CS91051, 13781, Aubagne Cedex, France ; blanche.krieguer@sartorius.com; samuel.dorey@sartorius.com; nina.perier@sartorius.com

[2] Aix Marseille Univ, CNRS, ICR, Case 551, 13397 Marseille, France ; sylvain.marque@univ-amu.fr

[3] Aix Marseille Univ, Avignon Univ, CNRS, IRD, IMBE, Marseille, France ; nathalie.dupuy@univ-amu.fr; fabien.girard@univ-amu.fr

[4] Aerial, 250 Rue Laurent Fries, 67400 Illkirch, France; florent.kuntz@aerial-crt.com; n.ludwig@aerial-crt.com

---

Gamma irradiation is one of the most widely used methods for sterilizing medical equipment and biopharmaceuticals. As a result of the existing sanitary environment (COVID) leading to a significant increase in the production of pharmaceutical goods, several doubts are raised about the current sterilizing capacity. Thus, to ensure the continued manufacture of these systems, alternate radiation sterilization techniques must be developed; X-ray sterilization is seen as a potential option.

In this research, the effects of gamma rays and X-rays on 22 medical device materials are examined. They consist of polypropylene (PP) and polyethylene (HDPE, LDPE, LLDPE). During a period of 380 days, EPR analyses were conducted on each substance. The purpose is to detect and quantify the radicals produced by gamma ray and X-ray irradiation of each polymer type.

The two examined classes of polymers produce distinctive radicals regardless of the irradiation technology. A comparable EPR signal was seen in all PP samples and ascribed to a peroxyl radical[1]. Using a curve resolution analysis (SIMPLISMA)[2], the radicals are identified. HDPE samples, unlike LDPE and LLDPE, exhibit a high EPR signal following gamma and X-ray irradiation, which is linked to their density and crystallinity. Alkyl, allyl, and polyenyl radicals are detected in all HDPE samples and distinguished[3] by their coupling constants (Figure 1). There was no substantial variation in EPR signals between gamma and X-ray irradiation among the 22 tested materials.

The comparability of EPR signals, kinetics, and radical concentration for the two examined classes of polymers allows for the identification of a radiation-matter interaction that is comparable.

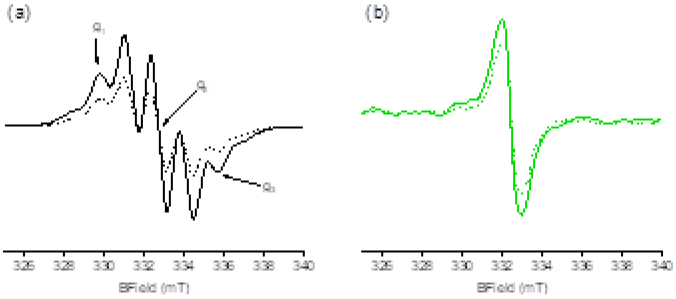


Figure 3: Signal of sample 8, 9 days after gamma (full line) and X-rays (dotted line) irradiation corresponding to a mixture of allyl, alkyl and polyenyl radicals. g1=2.021. g_2=2.003. g_3=1.985. (b) Signal of HDPE 60 days after gamma (full line) and X-rays (dotted line) irradiation corresponding to a polyenyl radicals

---

References

[1] Dogué L., Mermilliod N., Genoud F., *Journal of Polymer Science PartA: Polymer Chem*, **1994**, 32, 2193-2198
[2] Windig, W., Guilment J., *Journal of Analytica Chemistry*, **1991**, 63, 1425-1432
[3] Przybytniak G., Sadlo J., Walo M., *Express Polymer Lettres*, **2020**, 14, 556-565

Development of non-destructive tools for the authentication of the typical Italian vegetable species “Mugnoli di Pettorano sul Gizio”

Claudia Scappaticci, **Alessandra Biancolillo**, **Martina Foschi**, **Angelo Antonio D’Archivio**

Addresses: University of L’Aquila, Department of physical and chemical sciences, Via Vetoio, L’Aquila, 67100, Italy
E-mail: claudia.scappaticci@graduate.univaq.it

The aim of this work is to authenticate and characterize a particular type of Brassicaceae, namely the Mugnoli of Pettorano sul Gizio, a typical Italian product from Abruzzo (Central Italy).

Brassicaceae are commonly used vegetable species of particular scientific interest because they contain pharmacologically active substances.

The Mugnoli of Pettorano sul Gizio are one of the winter vegetables found in Italy. It resists to low temperatures and to the harsh winter climate that softens the leaves making them tastier.

The beginning of the cultivation of mugnoli dates back to the early 1900s and it is assumed these vegetables have been imported into Abruzzo by transhumant shepherds who returned from Apulia. With the passage of time, the Mugnoli have adapted to the climate of Pettorano sul Gizio becoming a unique typical product. The use of chemical fertilizers is not required for its cultivation as it is resistant to pests and this makes it a natural product.

The seed is not commercially available because it is handed down year by year, from cultivation to cultivation.

In order to characterize and authenticate this product, different cultivars of Brassicaceae (including mugnoli) have been harvested in the same experimental field of Pettorano sul Gizio (Abruzzo, Central Italy), and analyzed. In this way, all plants are subjected to the same pedoclimatic conditions, and it is possible to appreciate the different characteristics provided by their diverse varieties. Eventually, samples were analyzed by IR spectroscopy and Multivariate Image Analysis (MIA), with the aim of developing a green and rapid tool for the discrimination of this typical species with respect to other brassicaeae.

Eventually, the data obtained by means of the mentioned analytical techniques have been explored by Principal Component Analysis (PCA), to magnify differences and similarities among the various samples, and then analyzed by Partial Least Squares Discriminant Analysis (PLS-DA) [1,2] and Soft Independent Modelling of Class Analogies (SIMCA) [3], with the purpose of classifying mugnoli with respect to all the other individuals.

References

[1] Barker M, Rayens W (**2003**) Partial least squares for discrimination. *Journal of Chemometrics,* **17**, 166–173.

[2] Ståhle L, Wold S (**1987**) Partial least squares analysis with cross-validation for the two-class problem: A Monte Carlo study, *Journal of Chemometrics*, **1**, 185-196

[3] Wold M, Sjöström S (**1977**) SIMCA: a method for analyzing chemical data in terms of similarity and analogy. *Chemometrics: Theory and Application*, **52**, 243–282.

Rapid spectrochemical differentiation of amyotrophic lateral sclerosis and its progression based on ATR-FTIR coupled with chemometrics

Consuelo Pizarro, Kateryna Tkachenko, José María González Sáiz

Addresses: University of La Rioja, Department of chemistry, calle Madre de Dios 53,
Logroño, 26004, Spain
E-mail (consuelo.pizarro@unirioja.es)

Amyotrophic lateral sclerosis (ALS) disease still lacks reliable diagnostic and prognostic biomarkers to characterise the earliest phases of the disease. ALS is characterised by progressive muscle weakness, with rapid progression and fatal outcomes only a few years after diagnosis. In addition, the tardive appearance of ALS clinical symptoms delays the diagnosis and appropriate treatment. Here, we investigated the potential of attenuated total reflectance Fourier-transform infrared (ATR-FTIR) spectroscopy as a rapid tool for discriminating ALS patients from controls and ALS progression. In this preliminary study, the analysis based on LDA classification and SIMCA class modelling was performed on reduced spectra wavenumbers, providing excellent group separation. Therefore, both ALS at onset and after six months of follow-up and ALS at the disease onset from controls were discriminated with almost 100% accuracy by LDA and SIMCA. Several wavenumbers were identified as possible biomarkers towards tumour differentiation in the "fingerprint" region. These findings demonstrate the potential of ATR-FTIR spectroscopy towards ALS discrimination as a fast, low-cost, non-destructive and sensitive tool that could facilitate rapid disease management in clinical settings.

Chemometrics and molecular fingerprints to enhance LC-MS/MS spectral matching

Viviana Consonni, Fabio Gosetti, Veronica Termopoli, Roberto Todeschini, Davide Ballabio

Addresses: University of Milano - Bicocca, Department of Earth and Environmental Sciences, P.zza della Scienza, 1, 20126 Milano, Italy
E-mail davide.ballabio@unimib.it

Mass spectrometry (MS) is extensively used for the identification of compounds by matching the measured mass spectrum against a database of reference spectra. This approach can suffer from a partial coverage of the existing databases. Among the computational approaches for the identification of metabolite structures based on MS data, one possibility is to predict molecular fingerprints from the mass spectra with chemometric strategies and, in particular, multi-task neural networks [1,2]. Then, if models result to be accurate, predicted fingerprints can be used to screen compound libraries. However, it is necessary to train multi-task artificial neural networks from large datasets of mass spectra, used as inputs, while molecular fingerprints are used as outputs. We organized a large LC-MS/MS dataset from an on-line open repository and these data were used to train and test neural networks to predict molecular fingerprints. Predicted fingerprints of test compounds were then matched against an extended library including 14 millions of compounds to evaluate the capability to retrieve the correct compound. Effects of data sparseness and the effect of strategies of data curing and dimensionality reduction have been evaluated, as well as extensive diagnostics have been carried out to estimate modelling advantages and drawbacks as a function of the explored chemical space.

References

[1] Dührkop K (2022) Deep kernel learning improves molecular fingerprint prediction from tandem mass spectra, **Bioinformatics**, 38, i342–i349

[2] Consonni V, Gosetti F, Termopoli V, Todeschini R, Valsecchi C, Ballabio D (**2022**) Multi-task neural networks and molecular fingerprints to enhance compound identification from LC-MS/MS data, **Molecules**, 27, 5827

Estimation of null-space uncertainty in latent-variable model inversion: the case of correlated quality attributes

Elia Arnese-Feffin[a, *], Pierantonio Facco[a], Fabrizio Bezzo[a],
Massimiliano Barolo[a]

[a]CAPE-Lab – Computer-Aided Process Engineering Laboratory, Department of Industrial Engineering, University of Padova, via Marzolo 9, 35131 Padova, Italy
[*]elia.arnesefeffin@phd.unipd.it

Partial least-squares (PLS) modelling (Geladi *et al.*, 1986) is a data-driven modelling method widely used to establish regression models when predictor variables feature significant correlation, a common occurrence in data from (bio-)chemical processes. PLS models can be exploited to design process conditions that lead to a preset target quality by means of latent-variable model inversion (LVMI; Jaeckle *et al.*, 2000). A unique solution is obtained by LVMI if the dimension of the space of process conditions is less than, or equal to, the dimension of the quality space, while an infinite set of solutions is obtained otherwise: the so-called null-space. Theoretically, any combinations of process inputs falling on the null-space leads to the target quality. However, the PLS model is affected by uncertainty, which (upon model inversion) is back-propagated to the estimated null-space. If one wishes to move the LVMI solution along the null-space to optimize the LVMI solution (e.g., according to some economic objective), considering the null-space uncertainty is critical. Estimation of the null-space uncertainty is also the basis for the identification of the design space of a process, as commonly intended in pharmaceutical manufacturing (Facco *et al.*, 2015).

Different methods are available to estimate the null-space uncertainty in LVMI, or design space, based on resampling (Tomba *et al.*, 2012), simple analytical formulations (Zhang *et al.* 2009; Facco *et al.*, 2015), numerical optimization (Pantelides *et al.*, 2010), or rigorous analytical derivations (Palací-López *et al.*, 2019). However, none of the available methods simultaneously offers both computational simplicity and ability to deal with correlated quality attributes. Yet, output correlation is frequent in PLS modelling and LVMI, and can lead to numerical issues upon ill-conditioned matrix inversion.

Recently, Arnese-Feffin *et al.* (2022) proposed a generalization of LVMI to tackle the problem of correlated quality variables in LVMI. The proposed method is based on regularized matrix inversion and can efficiently solve the numerical issues discussed above. In this study, we extend such method to the estimation of the null-space uncertainty. We demonstrate our procedure on a simulated case study of a batch fermentation process. Our results show that the proposed method can efficiently solve conditioning issues with minimal information loss, allowing for a reliable estimation of the null-space uncertainty.

References

Arnese-Feffin E., Facco P., Bezzo F., Barolo M. (**2022**) Digital design of new products: Accounting for output correlation via a novel algebraic formulation of the latent-variable model inversion problem. *Chemometrics and Intelligent Laboratory Systems*, **227**, 104610.

Facco P., Dal-Pastro F. Meneghetti N., Bezzo F., Barolo M. (**2015**) Bracketing the Design Space within the Knowledge Space in Pharmaceutical Product Development. *Industrial and Engineering Chemistry Research*, **54**, 5128–5138.

Geladi P., Kowalski B. R. (**1986**) Partial Least-Squares Regression: A Tutorial. *Analytica Chimica Acta*, **185**, 1–17.

Jaeckle C. M., MacGregor J. F. (**2000**) Industrial applications of product design through the inversion of latent variable models. *Chemometrics and Intelligent Laboratory Systems*, **50**, 199–210.

Palací-López D., Facco P., Barolo M., Ferrer A. (**2019**) New tools for the design and manufacturing of new products based on Latent Variable Model Inversion. *Chemometrics and Intelligent Laboratory Systems*, **194**, 103848.

Pantelides C., Pinto M., Bermingham S. (**2010**) Optimization-based design space characterization using first-principles models. Proceedings AIChE Annual Meeting, 7–12.

Tomba E., Barolo M., García-Muñoz S. (**2012**) General framework for latent variable model inversion for the design and manufacturing of new products. *Industrial and Engineering Chemistry Research*, **51**, 12886–12900.

Zhang L., García-Muñoz S. (**2009**) A comparison of different methods to estimate prediction uncertainty using Partial Least Squares (PLS): a practitioner's perspective. *Chemometrics and Intelligent Laboratory Systems*, **97**, 152–158.

Identification of illicit storage treatment in fresh fish by a multi-omic approach coupled to chemometrics

Elisa Robotti[1], Masho Hilawie Belay[1], Elisa Calà[1], Simone Cerruti[1],
Elettra Barberis[2], Marcello Manfredi[2], Alessandro Benedetto[3], Paola Brizio[3], Marzia Pezzolato[3], Maria Cesarina Abete[3], Emilio Marengo[1]

Addresses: [1]University of Piemonte Orientale, Department of Sciences and Technological Innovation, Viale Michel 11, 15121, Alessandria, Italy; [2] CAAD, Center for Autoimmune and Allergic Disease, University of Eastern Piedmont, Novara, Italy; [3] Istituto Zooprofilattico Sperimentale del Piemonte, Liguria e Valle d'Aosta, Via Bologna 148, 10154 Torino, Italy
E-mail (elisa.robotti@uniupo.it)

In the field of food control of fresh products, the identification of foods subjected to illicit conservation treatments applied in order to extend the shelf life of the product plays a very important role. Fresh fish products are particularly subjected to this type of fraud, both because of the high commercial value of some fish species, and because they often have to be transported over a long distance from the fishing site, keeping their organoleptic characteristics unaltered. Treatments of this type involve for example the bleaching of the meat and / or the momentary abatement of the microbial load, however, the product continues the degradation process with the production of molecules that can also be dangerous to health.

It is therefore important to develop methods that allow the identification of illicit treatments based on sampling procedures that can be easily applied in the field during the control phase.

The study presented here was performed on sea bass samples divided into four groups: controls stored on ice in the fridge for 3 hours; controls stored under the same conditions for 24 hours; treated with a solution of hydrogen peroxide and citric acid for 3 hours; and treated with the same solution for 24 hours (these last two groups, while they are kept on ice in the fridge). Different parts of each fish (eye, muscle, skin, gills) were sampled by a sampling procedure based on the use of polymeric strips functionalized with adsorptive beads. The method, originally developed for the non-invasive sampling of cultural heritage, was exploited as a tool that can be easily applied in the field. The strips were then extracted with different media according to the final analysis: lipidomics (muscle, skin) by UHPLC-HRMS, metabolomics (muscle, skin, eye, gills) by GCxGC-MS, proteomics (muscle, skin) by UHPLC-HRMS.

The collected data were processed using multivariate statistical techniques of pattern recognition (PCA, cluster analysis) and classification (PLS-DA, SIMCA, Artificial neural networks) in order to correctly identify the samples treated illegally from the control samples and identify biomarkers of the illicit treatment. Bioinformatics was also applied to identify the pathways involved in the effect played by the applied treatment.

The work is financed by Ministero della salute (Ricerca Finalizzata 2016 - GR-2016-02363722 - The use of Led multi-spectral imaging for fast and innovative indexing of seafood preservation).

References

Manfredi M, Barberis E, Gosetti F, Conte E, Gatti G, Mattu C, Robotti E, Zilberstein G, Koman I, Zilberstein S, Marengo E, Righetti PG (**2017**) Method for Noninvasive Analysis of Proteins and Small Molecules from Ancient Objects. *Analytical Chemistry*, **89 (6)**, 3310-3317

CHEMOMETRICS AS A TOOL TO ANSWER TO THE CHALLENGES OF THE CLIMATE CHANGE AND POLITICAL SITUATIONS ON THE CONSISTENCY OF COCOA FLAVOUR QUALITY

Bagnulo, E.[1], Bortolini, C.[2], Liberto, E.[1]

[1] Dipartimento di Scienza e Tecnologia del Farmaco, Università degli Studi di Torino, Via Pietro Giuria 9, Turin, Italy
[2] Soremartec Italia S.r.l., P.le P. Ferrero 1, 12051 Alba (CN), Italy
E-mail: e.bagnulo@unito.it

Cocoa is one of the most economically important commodities in the world and currently the main producers are West and East Africa (three-quarters of world production) and Latin America and Asia [1-2]. Cocoa production is strongly influenced by socio-political balances in farm areas and especially by climate change which affects both the amount of cocoa production and the flavour quality of the final products [3-6]. Since the flavour of cocoa products is the main property that directs the consumer's choice, as well as acting as an indicator of the quality of the product itself, it is necessary to ensure a high level of raw material quality standards [5]. Therefore, tools to track sustainable cocoa production are of paramount importance, especially due to climate change and the introduction of foreign varieties to support the high global demand for cocoa-derived products, which heavily influence cocoa production, quality and purity. With this in mind, the study we report is part of a broader project aimed at analysing the flavour profiles of cocoa from different origins and defining their chemical-sensory identity card aimed at a constant quality standard supply by also considering cocoa from different origins to cope with climate change and political conflicts. The volatilome of one hundred sixty cocoa samples was analyzed by HS-SPME-GC-MS in combination with chemometrics for origin "identitation"[7-9]. To achieve the above-mentioned objectives, fingerprinting (non-targeted) and profiling (targeted) approaches were used, which were able to decipher the information contained in the complex volatilome dataset of both beans and liquor from different origins and validate the results, allowing the origins studied to be discriminated with successful classification models for their informative, discriminative and classification capacity. The results indicate a coherent, clear clustering of samples according to their origin with the two analytical strategies, both on raw beans and on cocoa liquors, albeit with differences at the molecular level. Predicting the classification of cocoa beans with the untargeted fingerprint on an external test set gave excellent results for beans with a classification rate of 100% and very good results for liquors (88%), despite the processing they underwent. Better results were obtained for targeted approaches with a classification rate above 92% (i.e. 92.86% for beans and 92.31% for liquors).

References

(1) Somarriba, E.; Peguero, F.; Cerda, R.; Orozco-Aguilar, L.; López-Sampson, A.; Leandro-Muñoz, M. E.; Jagoret, P.; Sinclair, F. L. **(2021)** Rehabilitation and Renovation of Cocoa (Theobroma Cacao L.) Agroforestry Systems. A Review. *Agron. Sustain. Dev.*, **41** (5), 1–19.

(2) ICCO. Statistics - International Cocoa Organization https://www.icco.org/ (accessed Sep 5, **2022**).

(3) Boeckx, P.; Bauters, M.; Dewettinck, K. **(2020)** Poverty and Climate Change Challenges for Sustainable Intensification of Cocoa Systems. *Curr. Opin. Environ. Sustain.*, **47**, 106–111.

(4) Aprotosoaie, A. C.; Vlad Luca, S.; Miron, A. (**2016)**Flavor Chemistry of Cocoa and Cocoa Products-An Overview. *Compr. Rev. Food Sci. Food Saf.*, **15** (1), 73–91.

(5) CABISCO/ECA/FCC. **(2015)** *Cocoa Beans : Chocolate & Cocoa Industry Quality Requirements*.

(6) Perotti, P.; Cordero, C.; Bortolini, C.; Rubiolo, P.; Bicchi, C.; Liberto, E. **(2020)** Cocoa Smoky Off-Flavor: Chemical Characterization and Objective Evaluation for Quality Control. *Food Chem.*, **309**, 125561.

(7) Magagna, F.; Liberto, E.; Reichenbach, S. E.; Tao, Q.; Carretta, A.; Cobelli, L.; Giardina, M.; Bicchi, C.; Cordero, C. (**2018)** Advanced Fingerprinting of High-Quality Cocoa: Challenges in Transferring Methods from Thermal to Differential-Flow Modulated Comprehensive Two Dimensional Gas Chromatography. *J. Chromatogr. A.*, **9**, 122-136.

(8) Medina, S.; Perestrelo, R.; Silva, P.; Pereira, J. A. M.; Câmara, J. S. Current Trends and Recent Advances on Food Authenticity Technologies and Chemometric Approaches (**2019)**. *Trends Food Sci. Technol.*, **85**, 163–176.

(9) Cuadros-Rodríguez, L.; Ortega-Gavilán, F.; Martín-Torres, S.; Arroyo-Cerezo, A.; Jiménez-Carvelo, A. M. **(2021)** Chromatographic Fingerprinting and Food Identity/Quality: Potentials and Challenges. *Cite This J. Agric. Food Chem*, **69**, 14428–14434.

Proper validation of classification models. A real-life example

Eneko Lopez[1], Jose Manuel Amigo[2,3], Jaione Etxebarria-Elezgarai[1], Andreas Seifert[1,2]

[1] *CIC nanoGUNE BRTA, Tolosa Hiribidea 76, San Sebastián, 20018, Spain*

[2] *IKERBASQUE, Basque Foundation for Science, Euskadi Plaza 5, Bilbao, 48009, Spain*

[3] *University of the Basque Country UPV/EHU, Leioa, 48940, Spain*

e.lopez@nanogune.eu

Chemometrics applied to spectroscopic data is becoming more and more popular in the field of medical diagnostics. By today, many user-friendly libraries with implemented chemometric algorithms exist to "facilitate" access to machine learning. However, one can observe trends in the scientific community towards over- or misuse of certain algorithms [1]. Software tools can easily and rapidly be used as a black box without proper evaluation (or knowledge) of whether the methods are appropriate regarding the three elements that make an algorithm successful: the data, the reference values, and the validation, as for example cross-validation and external validation.

This work presents a real-life example of misuse of Partial Least Squares Discriminant Analysis (PLS-DA) validation, resulting in overoptimistic or promising results. It shows how important a reliable validation procedure is when unbalanced classes are used or when the data has so much variance that it is difficult to create a good classification model [2].

The study is performed on data from dried human cerebrospinal fluid samples recorded by Fourier-transform infrared spectroscopy in attenuated total internal reflectance mode (ATR-FTIR). The dataset is small but balanced in terms of healthy and unhealthy groups. Each sample was measured three times to assure statistical significance. The samples were collected from 20 Alzheimer's-diagnosed and 19 healthy control subjects.

After preprocessing, data was split into training and validation sets, 80% and 20%, respectively [3]. **Figure 1** shows how the outcomes can be interpreted differently depending on the cross-validation method used. Leave-one-spectrum-out models follow the same trend as the training models, yielding high accuracy of nearly 100%. However, leave-one-subject-out cross-validation does not even exceed 70%. The next step of evaluating the model with an external validation set confirms the poor performance of the models, reflected by accuracies of around 65%.

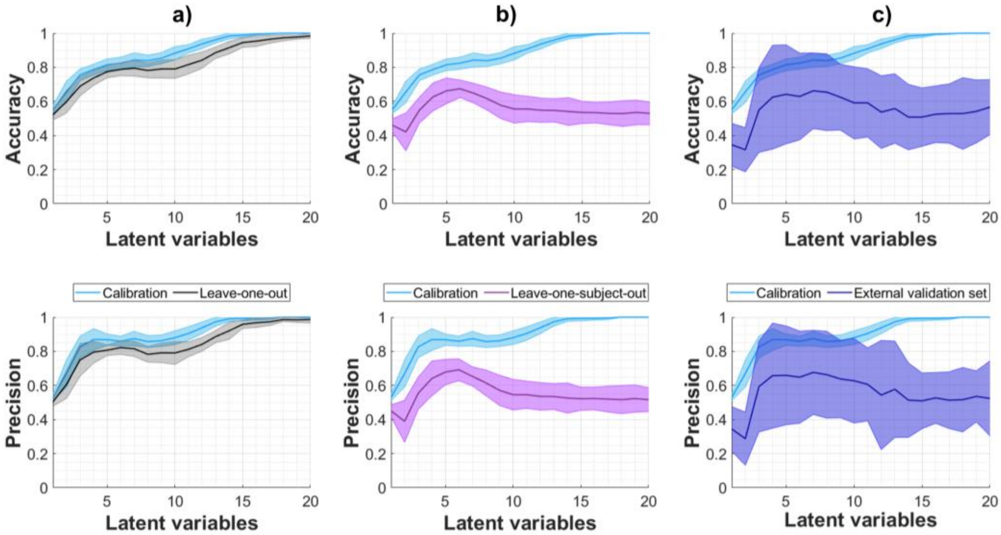


Figure 1: Accuracy and precision for different validation methods compared to the trained model. a) Leave-one-spectrum-out cross-validation. b) Leave-one-subject-out cross-validation. c) External validation set.

Our results indicate that the indiscriminate use of PLS-DA is inadvisable and that the utilization of intelligent criteria for data splitting can ensure the preservation of diversity within the dataset in both the training and testing subsets. Smart data selection is crucial for designing opportune validation strategies concerning specific modeling objectives. These findings have significant implications in chemometrics, particularly for predictive models in medical diagnostics. Moreover, we strongly have to emphasize the message that the best cross-validation (or validation) procedure is not the one giving the best figures of merit but the one giving the highest information on the data and the specific modeling objectives.

References

[1] Amigo JM (**2021**) Data mining, machine learning, deep learning, chemometrics: Definitions, common points and trends (Spoiler Alert: VALIDATE your models!). *Brazilian Journal of Analytical Chemistry*, **8**, 22–38.

[2] Westad F, Marini F (**2015**) Validation of chemometric models - A tutorial. *Analytica Chimica Acta*, **893**, 14–24.

[3] Morais CLM, Lima KMG, Singh M, Martin FL (**2020**) Tutorial: multivariate classification for vibrational spectroscopy in biological samples. *Nature Protocols*, **15**, 2143–2162.

Non-invasive analysis and chemometrics for macronutrients determination in non-conventional food plants (PANC)

Fabiola M. Verbi Pereira[1], Leticia da S. Rodrigues[1], Edenir R. Pereira-Filho[2]

Addresses: [1]Group of Alternative Analytical Approaches (GAAA), Bioenergy Research Institute (IPBEN), Institute of Chemistry, São Paulo State University (UNESP), Araraquara, São Paulo 14800-060, Brazil

[2]Group of Analytical Instrumental Analysis (GAIA), Chemistry Department, Federal University of São Carlos (UFSCar), São Carlos, São Paulo 13565-905, Brazil

E-mail fabiola.verbi@unesp.br

Plant analysis requires techniques with good sensitivity and minimal sample preparation using fewer chemicals in order to improve the analytical frequency [1]. For instance, X-ray fluorescence (XRF) is an effective technique for several elements' determination when in high concentration (1% or more). However, it has limitations for light elements ($Z < 13$) in bench-top instruments. Laser-induced breakdown spectroscopy (LIBS) makes chemical elements, such as P or S, difficult to detect depending on the instrumental configuration. Then, these two non-invasive techniques, energy-dispersive XRF (EDXRF) and LIBS were employed to predict the concentration levels of several macronutrients in non-conventional food plants (PANC). Partial least squares (PLS) models reveal remarkable results, using data from EDXRF fluorescence lines and LIBS atomic and ionic emission lines. The findings are that the EDXRF models can well-predict the macronutrients better than LIBS data, with standard error of calibration (SEC) between 0.57 and 0.70% (g/100g). Data fusion was also tested, and it did not reach improvements.

References

[1] Sultanbawa Y, Netzel ME (**2019**) Introduction to the Special Issue: Foods of Plant Origin. *Foods*, **8**, 555.

Application of PCA and classification method to prototype of fortified Piadina Romagnola

Flavia Bisozzi, Gabriella Tamasi, Claudio Rossi

Addresses: University of Siena, Department of Biotechnology, Chemistry and Pharmacy,
Via Aldo Moro 2, Siena, 53100, Italy
E-mail flavia.bisozzi@student.unisi.it

Nowadays a correct and balanced diet becomes an important topic for the human health. It is also interesting to evaluate the content of antioxidant compounds present in foods and the intake through the diet. In Mediterranean diet the mean content of polyphenols and flavonoids intake is reported to be around 664 – 1011 mg/day and 259 – 449 mg/day, respectively[1,2]. The assumption of grain derivates, contributes for about 3 % and 1% of the total polyphenols and flavonoids intakes [1,2]. For this reason, there is an interest on the enrichment of bread and other grain derivatives, using for instance oilseeds[3].

In this work a prototype of fortified Piadina Romagnola, made with refined (R) and whole grain (WG) flour, added with soy (S), flax seeds (FS), chia (CH) and sesame (SE) at different percentages were analysed, through spectrophotometric assay for the determination of the total content of polyphenols (TPC), flavonoids (TFC) and the Trolox equivalent antioxidant capacity (TEAC). The analyses were performed at the arrival of the product (fresh, F), after cooking at 180 °C (cook, C) and after the expiration date (shelf-life, SL). The fortified samples were compared with the corresponding commercial products with no addiction (CP). The results of the statistical analysis showed that there was a significant difference between the commercial product and the prototype additions (Student's t-test, $p < 0.05$) and from the two-way analysis of the variance (ANOVA) underlined a statistically significant difference ($p < 0.05$) between the fresh samples and the cooked and *shelf-life* ones. The principal component analysis (PCA) showed that the samples made with whole grain flour and with high percentage of addition had a higher value of TPC, TFC and TEAC than the samples made with refined flour, creating two different groups. On the other hand, any separation due to the treatments was revealed (C and SL).

A k-nearest neighbors (k-NN) classification method between three classes was also calculated: values of $\mathrm{TPC}_{\mathrm{total}} \geq 30$ mg GAE/100 g and $\mathrm{TFC}_{\mathrm{total}} \geq 30$ mg CE/100 g and

$\mathrm{TEAC}_{\mathrm{total}} \geq 100$ and an add ≥ 10 % (HH), values of $\mathrm{TPC}_{\mathrm{total}} \geq 30$ mg GAE/100 g and $\mathrm{TFC}_{\mathrm{total}} \geq 30$ mg CE/100 g and $\mathrm{TEAC}_{\mathrm{total}} \geq 100$ and an add < 10% (HL) and values of $\mathrm{TPC}_{\mathrm{total}} < 30$ mg GAE/100 g and $\mathrm{TFC}_{\mathrm{total}} < 30$ mg CE/100 g and $\mathrm{TEAC}_{\mathrm{total}} < 100$ and an add < 10% (LL). The division in training set (107 samples) and test set (54 samples) was done randomizing the rows. In cross validation for $k = 3$ the total percentage of correct prediction was 91% and in prediction the total percentage of correct prediction was 95%.

The results suggest the good performance of the model and the possibility to apply this chemometric model to other food matrices.

References

1 J. Godos, S. Marventano, A. Mistretta, F. Galvano and G. Grosso (**2017**), Dietary sources of polyphenols in the Mediterranean healthy Eating, Aging and Lifestyle (MEAL) study cohort, *Int. J. Food Sci. Nutr.*, **68**, 750–756.

2 R. Zamora-Ros, V. Knaze, J. A. Rothwell, B. Hémon, A. Moskal, K. Overvad, A. Tjønneland, C. Kyrø, G. Fagherazzi, M. C. Boutron-Ruault, M. Touillaud, V. Katzke, T. Kühn, H. Boeing, J. Förster, A. Trichopoulou, E. Valanou, E. Peppa, D. Palli, C. Agnoli, F. Ricceri, R. Tumino, M. S. de Magistris, P. H. M. Peeters, H. Bas Bueno-De-Mesquita, D. Engeset, G. Skeie, A. Hjartåker, V. Menéndez, A. Agudo, E. Molina-Montes, J. M. Huerta, A. Barricarte, P. Amiano, E. Sonestedt, L. M. Nilsson, R. Landberg, T. J. Key, K. T. Khaw, N. J. Wareham, Y. Lu, N. Slimani, I. Romieu, E. Riboli and A. Scalbert, (**2016**) Dietary polyphenol intake in europe: The european prospective investigation into cancer and nutrition (EPIC) study, *Eur. J. Nutr.*, **55**, 1359–1375.

3 B. De Lamo and M. Gómez, (**2018**)Bread enrichment with oilseeds. A review, *Foods*, **7**, 191-205.

Characterization of Bio-oil samples obtained from Thermal liquefaction

Flávio M. Andrade[1], Raquel S. Silva[3], José M. S. Rodrigues[3], Leandro Danielski[3], Diogo A. Simões[4], Marina Cocchi[2], Maria Fernanda Pimentel[3]

Addresses: [1]Federal University of Pernambuco, Department of Fundamental Chemistry, Av. Jorn. Aníbal Fernandes, s/n, Recife, 50740560, Brazil
[2]University of Modena and Reggio Emilia, Department of Chemical and Geological Sciences, Via Campi 103, Modena, 41125, Italy
[3]Federal University of Pernambuco, Department of Chemical Engineer, Av. dos Economistas, s/n, 50740590, Recife, Brazil
[4] Federal University of Pernambuco, Department of Biochemistry, Av. Prof. Moraes Rego, s/n, 50670420, Recife, Brazil
E-mail flavio.montenegro@ufpe.br

Aiming to supply the growing demand for energy by society, different methods of energy production have been developed, among them, the thermochemical conversion of biomass has been reported as one of the promising methods. While for pyrolysis or gasification processes, a preliminary drying step is mandatory, for thermal liquefaction (TL) processes, the biomass is used wet, being the main advantage of that approach. Several polar solvents can be employed in TL processes, among them water in near or supercritical conditions. The main product obtained from thermal liquefaction processes is the bio-oil, which has the potential to produce fuels with characteristics similar to fossil fuels, such as gasoline and diesel [1]. Especially for the hydrothermal process, as water is the major input component, it turns the process environmentally friendly. However, water is considered a contaminant to fuels. In this context, knowing the water content provides information about the efficiency of the phase separation in the production process as well as about the bio-oil quality. Some works have described the use of FTIR [2] [3] and NIR [4] data associated with chemometrics to predict the water content in bio-oils from pyrolysis or TL processes, though all of them are related to continuous process or use just one kind of biomass or even one solvent in the conversion process. In this context, this work aims to describe the use of FTIR data and chemometrics to build a predictive model to determine the water content of bio-oil samples from TL processes in batch, that employed water or ethanol as production solvent and different biomasses such as sugarcane bagasse, sugarcane straw, and eucalyptus wood. The reference water values were obtained from a volumetric Karl-fisher equipment.

References

[1] Katarzyna R. Arturi, Sergey Kucheryavskiy, Erik G. Søgaard (**2016**). Performance of hydrothermal liquefaction (HTL) of biomass by multivariate data analysis, *Fuel Processing Technology*, **150**, 94–103.

[2] A. Veses, J.M. López, T. García, M.S. Callén (**2018**). Prediction of elemental composition, water content and heating value of upgraded biofuel from the catalytic cracking of pyrolysis bio-oil vapors by infrared spectroscopy and partial least square regression models. *Journal of Analytical and Applied Pyrolysis,* **132,** 102–110.

[3] René B. Madsen, Konstantinos Anastasakis, Patrick Biller, and Marianne Glasius. (**2018**). Rapid Determination of Water, Total Acid Number, and Phenolic Content in Bio-Crude from Hydrothermal Liquefaction of Biomass using FT-IR. *Energy Fuels,* **32**, 7660−7669.

[4] Markandey M. Tripathi, El Barbary M. Hassan, Fang-Yu Yueh, Jagdish P. Singh, Philip H. Steele, Leonard L. Ingram Jr (**2009**). Reflection–absorption-based near infrared spectroscopy for predicting water content in bio-oil, *Sensors and Actuators B*, **136,** 20–25.

APPLICATION OF DOE AND MULTIVARIATE ANALYSIS FOR A TXRF METHOD DEVELOPMENT AND DATA ANALYSIS. A CASE-STUDY FROM THE AGRI-FOOD SECTOR

Giacomo Squeo[1], Ignazio Allegretta[2], Concetta E. Gattullo[1], Carlo Porfido[1], Francesco Caponio[1], Stefano Cesco[3], Carlo Nicoletto[4], Roberto Terzano[1]

[1]University of Bari Aldo Moro, Department of Soil, Plant and Food Sciences (DiSSPA), via Amendola 165/a, Bari, 70126, Italy
[2]University of Salento, Department of Biological and Environmental Sciences and Technologies, S.P. 6 Lecce-Monteroni, Lecce, 73100, Italy
[3]Free University of Bozen-Bolzano Faculty of Science and Technology, Piazza Università 5, Bolzano, 39100, Italy
[4]University of Padova, Department of Agronomy, Food, Natural Resources, Animals, and Environment (DAFNAE), Viale dell'Università 16, Legnaro, 35020, Italy
giacomo.squeo@uniba.it (corresponding author)

Total-reflection X-ray fluorescence (TXRF) spectroscopy is a suitable analytical method for the determination of the elemental composition of different kind of samples. In recent years, the technique has been widely used for agri-food products (Borgese et al., 2015). However, despite the growing application of TXRF, few studies have considered and applied multivariate strategies for improving method performance and usability. In particular, at least two critical steps could profit from the application of chemometric tools: i) sample preparation and ii) signal analysis. As for the first step, our goal was to find the best sample preparation for an organic material. In literature there is no accordance among the amounts of sample/suspender to be used. Thus, the design of experiments has been used as a rational approach to find suitable conditions of sample preparation (mass of the sample and dispersant volume). A 2^2 full factorial design was set up having as responses the recovery of twelve elements. The obtained response surfaces allowed to identify the region(s) of the experimental domain in which a suitable recovery (80-120%) was reached for most of the elements.

For what concern the second aspect, the output of TXRF is a continuous spectrum which is multivariate in its nature. Nonetheless, in literature these signals are seldom treated directly by multivariate methods, while, most commonly, a quantitation of the single elements is carried out. Thus, we aimed at evaluating the feasibility of TXRF coupled with multivariate data analysis for the discrimination of beans (twenty-four genotypes) from two growing sites comparing this analytical approach with the most common one, i.e., by using the quantified elemental composition for the multivariate analysis. The elemental dataset (144 × 12) and the spectral dataset (144 × 2312) were subjected to different preprocessing methods (according to the different nature of the data), explored by PCA and then different classification models were built and tested (Allegretta et al., 2023). The results showed that good discrimination between the growing sites could be achieved with both the datasets and approaches. In the case of the spectral dataset, the great variability associated to the bean genotypes masked that related to the growing sites which could be highlighted by using the GLSW filter. The practical advantage of the direct use of TXRF signals for classification purposes lied in the possibility to avoid the elemental quantification procedure (and related errors) thus speeding up the analysis and the classification assessment.

References

Borgese L, Bilo F, Dalipi R, Bontempi E, Depero LE (**2015**) Total reflection X-ray fluorescence as a tool for food screening. *Spectrochimica Acta Part B: Atomic Spectroscopy*, **113**, 1–15.

Allegretta I, Squeo G, Gattullo CE, Porfido C, Cicchetti A, Caponio F, Cesco S, Nicoletto C, Terzano R (**2023**) TXRF spectral information enhanced by multivariate analysis: A new strategy for food fingerprint. *Food Chemistry*, **401**, 134124.

ACCELERATING CELL LINES SELECTION IN BIOPHARMACEUTICAL PROCESS DEVELOPMENT THROUGH DATA-DRIVEN MODELING ON METABOLOMIC DYNAMIC INFORMATION

**Gianmarco Barberi[a], Antonio Benedetti[b], Paloma Diaz-Fernandez[c],
Daniel C. Sévin[d], Johanna Vappiani[d], Gary Finka[c], Fabrizio Bezzo[a],
Pierantonio Facco[a]**

a) CAPE-Lab – Computer-Aided Process Engineering Laboratory, Department of Industrial Engineering, University of Padova, via Marzolo 9, 35131 Padova PD, Italy
b) Process Engineering & Analytics, Product Development and Supply, GlaxoSmithKline R&D, Gunnels Wood Rd, SG1 2NY Stevenage, UK
c) Biopharm Process Research, Biopharm Product Development and Supply, GlaxoSmithKline R&D, Gunnels Wood Rd, SG1 2NY Stevenage, UK
d) Cellzome GmbH, GlaxoSmithKline R&D, Meyerhofstraße, 1, 69117 Heidelberg, Germany
E-mail: pierantonio.facco@unipd.it

The recent years have seen an exponentially growing production of monoclonal antibodies, which are clinically used for the treatment of immunological, oncological, and infectious diseases, such as Covid-19. The development of new monoclonal antibodies is an expensive and time-consuming process, typically lasting more than 10 years and costing more than 2 billion dollars (Epifa 2021; Barberi et al., 2022). Along monoclonal antibodies development, the identification of appropriate cell lines, that guarantee the desired product quality attributes (i.e., productivity, stability, product quality), requires an extensive experimentation and is a key step for all subsequent development stages. For this reason, pharmaceutical companies are looking at innovative modeling solutions to accelerate monoclonal antibodies process development, while preserving the main product quality attributes (Li et al., 2010).

In this work, we aim at accelerating the selection and scale-up of high productive cell lines exploiting the wealth of information contained in dynamic metabolomic data from AMBR15® scale through multivariate statistical techniques.

The use of multiway and evolving multivariate statistical techniques on dynamic metabolomic profiles identifies high productive cell lines already from the early stages of the experimentation at AMBR15® scale. In fact, a validation accuracy of 100% is achieved only exploiting the metabolomic profiles from the first week of culture. This early identification allows to make timely decision on process development, and eventually scale-up promising cell lines even before the end of the experimentation at AMBR15® scale.

Furthermore, the exploitation of dynamic metabolomic data and the in-depth inspection of the multivariate models allow to identify the metabolic trats of cell lines with industrially relevant phenotypes (i.e., high productivity). The biomarkers mostly associated with antibody productivity along the entire cell culture are identified, while the variation of the importance of metabolic pathways for the discrimination of cell productivity along the cultivation is studied. Specifically, metabolic pathways related to tricarboxylic acid cycle have a predominant importance in the early stage of the cultivation, while pathways related to amino and nucleotide sugars show larger importance in the late stages of the cultivation. This knowledge could be exploited to improve the confidence in the selection of promising cell lines, even mitigating the risk of progressing inadequate cell lines along the development stages. Furthermore, the identified cellular function provide insight for the further improvement of antibody productivity through genetic manipulation.

References

Epifa (2021). The pharmaceutical Industry in Figures: Key data 2021. https://www.efpia.eu/publications/downloads/efpia/the-pharmaceutical-industry-in-figures-2021/, last accessed 10/01/2023.

Barberi G., Bendetti A., Diaz-Fernandez P., Sévin D.C., Vappiani J., Finka G., Bezzo F., Barolo M., Facco P. (2022). Integrating metabolome dynamics and process data to guide cell line selection in biopharmaceutical process development. *Metabolic Engineering*, **72**, 353-364.

Li F., Vijayasankaran N., Shen A., Kiss R., Amanullah A. (2010). Cell culture process for monoclonal antibody production. *MAbs*, **2**, 466-479.

MULTIVARIATE STRATEGIES FOR UNLOCKING THE POTENTIAL OF INFRARED SPECTRA TO ENHANCE ADDITIVE EFFECTS IN HISTORICAL MORTARS

Giulia Gorla[1], Chiara Zannin[1], Cristina Corti[2], Ludovico Geminiani[1], Laura Rampazzi[2], Barbara Giussani[1]

1.University of Insubria, Department of Science and High Technology, Como, 22100, Italy

2. University of Insubria, Department of Human Sciences, Innovation, and Territory, Como, 22100, Italy

ggorla@uninsubria.it

Historical mortars are building materials mainly composed of lime, sand, and water. They have been used since antiquity to bond bricks and stones in constructions and to make plasters to protect them. Organic compounds have frequently been added to lime mortars for property modifications to satisfy various functional needs in building techniques [1]. Understanding the composition of historical mortars and the impact of various additives is of great importance for gaining insights into past technologies and improving modern conservation practices [2, 3].

Different groups of mortar samples were prepared with gum arabic and linseed oil added as additional ingredients in mortar samples. The samples were analysed over time using different infrared techniques: external reflection FTIR (ER-FT-IR) spectroscopy (7500 - 375 cm^{-1}), ATR-FT-MIR spectrometer (4000 - 600 cm^{-1}), and Miniaturized NIR spectrometers (7407- 4000 cm^{-1}). Spectra were organized in different arrays and multivariate strategies were investigated to understand how additives affect the production process of mortars. The carbonation kinetics were also monitored with X-ray diffraction spectroscopy and thermogravimetric analyses.

This research aimed to propose multivariate statistical strategies to deal with several datasets

composed of different samples acquired over time and with different instrumentations.

References

[1] Zhang, K., Grimoldi, A., Rampazzi, L., Sansonetti, A., & Corti, C. (**2019**). Contribution of thermal analysis in the characterization of lime-based mortars with oxblood addition. *Thermochimica Acta*, **678**, 178303.

[2] Rampazzi L, Colombini MP, Conti C, Corti C, Lluveras-tenorio A, Sansonetti A, Zanaboni M, (**2016**) Technology of medieval mortars: an investigation into the use of organic additives, *Archeometry*, **58**, 115-130.

[3] Brunello, V., Corti, C., Sansonetti, A., Tedeschi, C., & Rampazzi, L. (**2019**). Non-invasive FTIR study of mortar model samples: Comparison among innovative and traditional techniques, *The European Physical Journal Plus*, **134**(6), 270.

CHEMOMETRICS TO AVOID SYSTEMATIC ERROR IN THE CONTROL OF NICKEL IN ELECTROLESS NICKEL PLATING PROCESSES BY UV-VIS SPECTROPHOTOMETRY

Gorka Albizu, Ane Bordagaray, Rosa García-Arrona, Maider Vidal, Miren Ostra

University of the Basque Country (UPV/EHU), Department of Applied Chemistry, Manuel de Lardizabal, 3, Donostia-San Sebastián, 20018, Spain

gorka.albizu@ehu.eus

Electroless deposition is one of the most widely employed process in the coating industry, being nickel the most used metal. Nickel and its alloys represent 95 % of all the industrial electroless applications, with sodium hypophosphite being the most widely used reducing agent, leading to Ni-P coatings [1]. During the deposition process, the concentration of nickel and other compounds present in the electrolyte decreases, so they need to be replenished.

Titration with EDTA and spectrophotometric measurement at a single wavelength are the two procedures used in the industry to control nickel levels before replenishment. Due to the green color of nickel in an acidic environment, its concentration can be evaluated between 380-400 nm. However, while reagents are being added to the electrolyte to provide the optimum conditions, organic additives are also accumulating in the bath, altering the electrolyte spectrum in the UV region along time. As seen in Figure 1, this spectrum modification affects the wavelengths at which nickel is measured. Specifically, this organic compounds accumulation increases the absorbance, giving rise to a systematic error in the estimation of the metal even though multivariate calibration like Partial Least Square (PLS) is used, which increases as long as the bath ages, since the predicted values are higher than the actual concentrations.

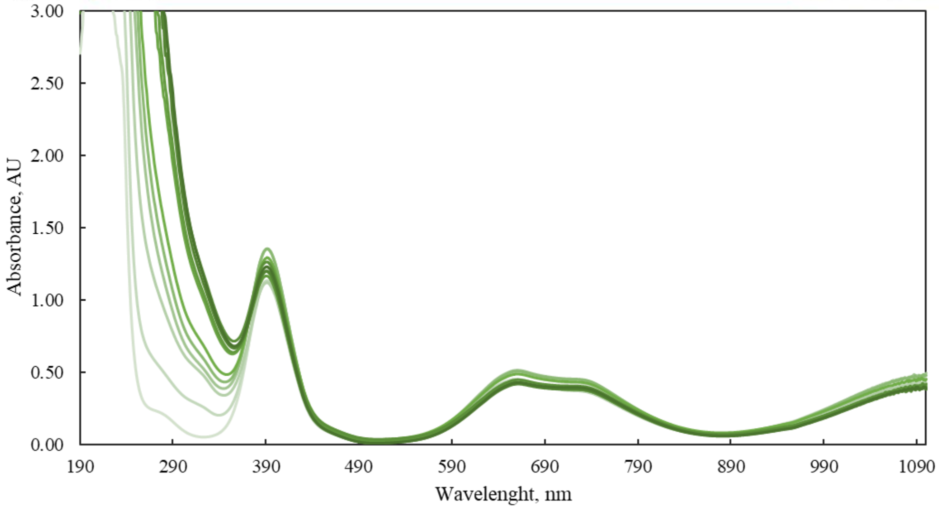


Figure 1. Obtained spectra of some of the analyzed samples, from light green (first coatings) to dark green (last coatings).

To avoid systematic error that can observed in the first two plots of the Figure 2, different strategies have been used. Instead of using exclusively standard solutions to build PLS models, titrated samples from another bath have also been included in the calibration set. In this work, the results obtained using different calibration sets, spectral ranges and data preprocessing are contrasted and presented.

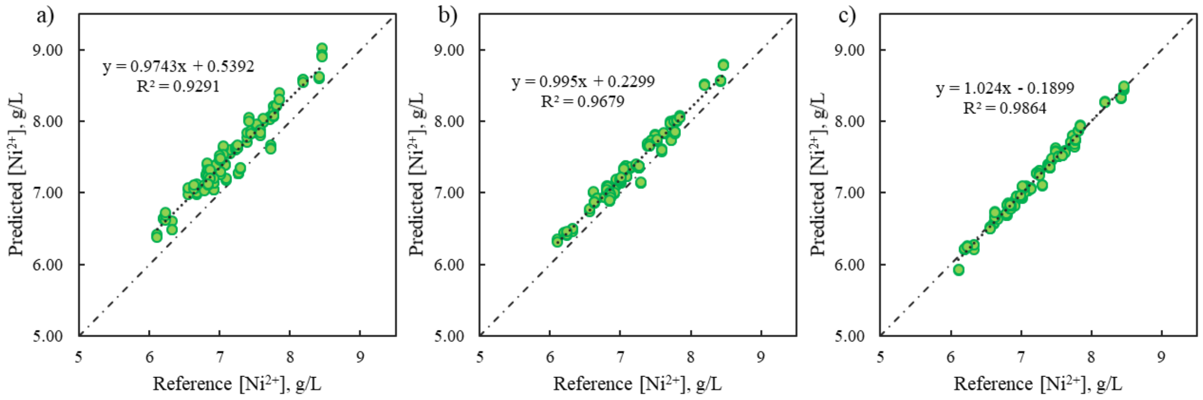


Figure 2. Comparison between the predicted Ni^{2+} concentration by UV-Vis using a) an univariate calibration at 392 nm; b) a PLS model of the spectrum between 300 and 1000 nm and as calibration set only standard solutions; c) a PLS model of the spectrum between 300 and 1000 nm and as calibration set standards solutions and samples from another bath; vs reference Ni^{2+} concentration.

References

[1] Sudagar J, Lian J, Sha W (2013) Electroless nickel, alloy, composite and nano coatings – A critical review. Journal of Alloys and Compounds, 571, 183-204.

COMPARISON OF CLASSIFICATION MODELS FOR LUNG CANCER DETECTION USING RAMAN SPECTROSCOPY

Harun Hano[1], Andreas Seifert[1,2]

[1] CIC nanoGUNE BRTA, Tolosa Hiribidea 76, San Sebastián, 20018, Spain
[2] IKERBASQUE, Basque Foundation for Science, Euskadi Plaza 5, Bilbao, 48009, Spain
h.hano@nanogune.eu

CRYPTED FOR CONFIDENTIALITY REASON

MULTIVARIATE INVESTIGATION ON THE INTERACTION OF TWO POLYMERIC MATERIALS WITH EMERGING CONTAMINANTS FOR SORPTIVE EXTRACTION APPLICATIONS

Henry MacKeown, Barbara Benedetti, Chiara Scapuzzi, Matteo Baglietto, Marina Di Carro, Emanuele Magi

University of Genoa, Department of Chemistry and Industrial Chemistry, Via Dodecaneso 31, Genoa, 16146, Italy
henry.mackeown@edu.unige.it

Sorptive extraction combines analyte extraction and preconcentration from water matrices with a simple sample preparation technique. In this work, the sorption ability of two polymeric materials was evaluated for the extraction of trace amounts of 40 contaminants with a wide range of polarities ($-3.8 < \log D < 6.88$) from water samples. These included different classes of emerging contaminants such as pharmaceuticals, UV-filters, estrogens, as well as a group of polar compounds of increasing interest, called "persistent mobile organic contaminants" (PMOCs).

Polyethersulfone (PES), in the form of a porous membrane, was chosen due to its explored capabilities for the extraction of a wide range of semi-polar and less polar compounds [1,2], while a starch-based biodegradable flexible film (SBF) was studied for the first time, as a greener option. Pieces of each polymer were placed in vials containing spiked water for 24h and then extracted with methanol. To find the best compromise for a list of compounds differing in so many properties, a design of experiments (DoE) was employed to simultaneously investigate several factors, possibly affecting the interaction mechanism. As a first approach, two identical Plackett-Burman screening designs were conducted for each material, to explore the effect of 5 variables on the extraction efficiency: pH, NaCl concentration, temperature, film dimension and % of organic modifier (ethanol). The LC-MS/MS signals in the eluate were used as responses to be modeled. Some of the studied variables have already been explored in PES sorptive extraction processes through DoE approaches, but for a smaller range of compound polarity [1,2].

Principal component analysis (PCA) was performed on the responses of the experiments, highlighting the correlation among some of them, which suggested a similar behavior for some groups of substances. However, in both cases PC1, PC2 and PC3 explained approximately 75% of the variance (approximately 35% and 25% for PC1 and PC2, respectively). Thus, single MLR models were computed, rather than using the scores' values as a unique response.

Some of the more polar analytes only slightly accumulated onto both films, behaving the same in all experiments. By contrast, UV filters and the anti-bacterial triclosan were completely extracted by SBF from water in all cases. Still, a positive effect of the temperature was observed for PES. For all non-steroidal anti-inflammatory drugs, uptake was significantly dependent on film dimension, but was positive for PES and surprisingly negative for SBF. The organic modifier effect was only statistically significant for four compounds using SBF. Instead, it had a negative effect on 9 semi-polar and less polar compounds for PES. In the studied pH range, some anionic compounds were negatively dependent on pH for SBF but not for PES.

Considering the outcomes of the screening designs, further experimental designs will be performed to study the quadratic models and understand potential interactions among variables. Finally, the sorptive extraction analytical methods will be optimized ad hoc for different classes of analytes.

References

[1] A. Prieto, R. Rodil, J.B. Quintana, I. Rodríguez, R. Cela, M. Möder, Evaluation of low-cost disposable polymeric materials for sorptive extraction of organic pollutants in water samples, Anal. Chim. Acta. 716 (2012) 119–127. https://doi.org/10.1016/j.aca.2011.12.023.

[2] A. Prieto, R. Rodil, J.B. Quintana, R. Cela, M. Möder, I. Rodríguez, Evaluation of polyethersulfone performance for the microextraction of polar chlorinated herbicides from environmental water samples, Talanta. 122 (2014) 264–271. https://doi.org/10.1016/j.talanta.2014.01.024.

3D FLUORESCENCE SPECTROSCOPY AS A TOOL FOR DISCRIMINATION OF OLIVE OILS FROM TRADITIONAL AND SUPERHIGH-DENSITY OLIVE GROVES

Isabel Durán-Merás[1], Elísabet Martín-Tornero[1], Daniel Martín-Vertedor[2], Rebeca Garrido González[1], Arsenio Muñoz de la Peña[1]

[1]Universidad de Extremadura, Departamento de Química Analítica, Av. de Elvas, Badajoz, 06006, Spain

[2]Junta de Extremadura, CICYTEX-INTAEX, Avenida Adolfo Suárez s/n, Badajoz, 06007, Spain.

iduran@unex.es

Olive oil is known to have important health benefits due to its particular composition of bioactive compounds. These compounds are strongly influenced by various agronomic factors, such as the type of crop, the maturation state of olives, the agroclimatic conditions, or the type of irrigation [1], giving rise to oils with different characteristics. On the other hand, superhigh-density olive groves and high-density olive groves are rapidly expanding in olive oil producer countries [2]. In this sense, consumers should have more information concerning how is produced the olive oil they buy. Consequently, having quick and simple tools that allow differentiating the type of oil is essential so that producers and consumers can choose.

Due to the native fluorescence of olive oil, in this work, the use of a non-destructive technique as fluorescence spectroscopy, combined with chemometric algorithms, is proposed to discriminate between olive oils obtained from olives belonging to rainfed crops and to super-intensive olive crops with different irrigation treatments: intensive adult dry, super intensive young irrigated, intensive young irrigated, super intensive adult irrigated and intensive adult irrigated. Olive oils were collected in two different maturation states (green and veraison) and for each production system and maturation stage, 6 oil samples (6 replicates) were extracted. Thus, a total of 60 samples were available (5 production systems x 2 maturation stages x 6 replicates).

The excitation-emission matrices (EEMs) of the samples were recorded in the two fluorescence regions that present the most relevant information. The first region was recorded with excitation wavelengths between 220 to 700 nm and emission from 300 to 600 m, corresponding to the presence of polyphenols and oxidation compounds. The second fluorescence region was recorded from 220 to 700 nm of excitation and from 650 to 800 nm of emission, corresponding to chlorophylls and pheophytins. The EEMs obtained were analyzed using parallel factor analysis (PARAFAC) as an exploratory analysis. The scores of PARAFAC were used to perform a classification model using linear discriminant analysis (LDA).

Each maturation state was analysed by separately. The two fluorescence regions and the combination of them were tested in both sets of samples, in order to select the one that allows better discrimination between the different groups. In all cases the optimum number of components was three and the excitation and emission loading profiles were similar for the two maturation stages. The best results were obtained when the two studied regions of the samples were fused. In this case, PARAFAC-LDA was able to discriminate between olive oils of the five production systems with an 80% of correct predictions in the green ripening state and a 100% of correct prediction in the veraison maturation state.

Acknowledgments

Financial support was provided by the Ministerio de Ciencia e Innovación de España por (Project PID2020-112996GB-I00 financed by MCIN/AEI/10.13039/501100011033) and Junta de Extremadura (Ayuda a Grupos GR21048-Research Group-FQM003 and Project IB20016) co-financed by Fondos Europeos de Desarrollo Regional.

References

[1] Dagdelen A, Tumen G, Ozcan MM, Dundar E (2013) Phenolics profiles of olive fruits (Olea europaea L.) and oils from Ayvalik, Domat and Gemlik varieties at different ripening stages. *Food Chem,* **136**, 41–45.

[2] Guerrero-Casado J, Carpio AJ, Tortosa FS, Villanueva AJ (2021) Environmental challenges of intensive woody crops: The case of super high-density olive groves. Science of the Total Environment **798**, 149212

Studying nectarine ripening with near infrared spectroscopy

Jokin Ezenarro, Daniel Schorn-García, Laura Aceña, Montserrat Mestres, Olga Busto, Ricard Boqué

Universitat Rovira i Virgili, Department of Analytical Chemistry and Organic Chemistry, Campus Sescelades, Edifici N4, C/Marcel·lí Domingo 1, Tarragona, 43007, Spain
jokin.ezenarro@urv.cat

Nectarines are a type of stone fruit widely consumed due to their sweet taste and various health benefits. As the ripening of nectarines plays a critical role in determining the final quality of the fruit and its shelf life, it is necessary to understand this process and the factors that affect it (Marini & Trout, 1984). The ripening of nectarines is influenced by various internal and external factors, including sunlight exposure, watering or temperature.

To gain a deeper understanding of the factors affecting ripening, a study was conducted using Fourier Transform-Near Infrared (FT-NIR) spectroscopy and chemometric techniques such as ANOVA-Simultaneous Component Analysis (ASCA) for quantifying the effects, Principal Component Analysis (PCA) for studying sample groupings and Partial Least Squares Regression (PLSR) for building calibration models for key quality parameters such as pH, penetromy, sugar content or total acidity.

In this study, the factors that could affect the most during the ripening time have been considered (Muleo et al., 1994). First, the evolution in time, that is, samples have been collected once a week during eleven weeks since stone hardening. Second, the position of the nectarines on the tree: nectarines in the top of the tree and middle height of the tree, nectarines at the tip of the branches and closer to the trunk and nectarines in the south-east and north-west sides of the tree have been collected and classified in these categories, four biological replicates for each combination and sampling day. Finally, the differences in the sampling point of the individual nectarines have been considered, systematically measuring the dark side and the light side, two replicates on each side. In addition, samples were measured with and without skin, in order to study its influence on the spectra and to conclude which measurement gives more reliable information.

The results show that all factors studied had a significant impact at some point on the ripening process, but the homogeneity between nectarines at the end of the ripening process is not strictly determined by the sunlight received. Different strategies such as spectra preprocessing and data unfolding were used to understand the differences between the ripening process of nectarine skin and nectarine flesh and to assess the influencing factors.

In conclusion, the combination of NIR spectroscopy and chemometrics shows to be a valuable tool for understanding the ripening process of nectarines. These findings could have important implications for the fruit industry and may lead to the development of new harvesting, storage and handling procedures to improve the quality of nectarines for consumers.

References

Marini, R. P., & Trout, J. R. (1984). Sampling Procedures for Minimizing Variation in Peach Fruit Quality. *Journal of the American Society for Horticultural Science*, *109*(3), 361–364. https://doi.org/10.21273/JASHS.109.3.361

Muleo, R., Masetti, C., Morini, S., Loreti, F., & Tellini, A. (1994). Modifications of Some Characteristics in Nectarine Fruit Induced by Light Deprival at Different Times of Fruit Growth. *Advances in Horticultural Science*, *8*, 75–1005. https://doi.org/10.1400/14296

Comparison of different spectroscopic data and chemometric tools for determining citalopram content in pharmaceutical tablets

J.Cruz[1], G.Gergov[2], E.Kirilova[2], S.Kasabov[3], O.Kostadinova[4]

[1]Escola Universitària Salesiana de Sarrià Passeig Sant Joan Bosco 74 08017 Barcelona, Catalonia.
[2]Bulgarian Academy of Sciences, Institute of Chemical Engineering, Acad. Georgi Bontchev Str., Bl.103, Sofia 1113, Bulgaria.
[3] Medical University of Sofia, Faculty of Pharmacy, Dunav 2, 1000 Sofia, Bulgaria.
[4]Institute of Electrochemistry and Energy Systems "Acad. E. Budevski", "Acad. Georgi Bonchev" Str. Bl.10, 1113 Sofia, Bulgaria.
E-mail (jcruz@euss.cat)

Citalopram hydrobromide salt is an antidepressant belonging to a group of drugs called selective serotonin reuptake inhibitors (SSRIs), and it is used to treat major depressive disorders. In the investigation of Durby et al. l [1], Near-Infrared (NIR) transmittance and Raman spectroscopy chemometric calibrations of the active substance content of a pharmaceutical tablet were developed using partial least squares regression (PLS) combined with suitable preprocessing and variable selection. In the present work, various dosage values for active substance were measured (120 tablets in total) and manufactured by pilot production scale.

The spectral acquisition range for NIR is 7400 –10 507 cm^{-1} (404 variables), and for Raman 3600 – 200 cm^{-1} (3401 variables) per sample. The citalopram content (ranging between 5.6–8.0 wt. %) was measured by high-performance liquid chromatography (HPLC). The original dataset was divided into a calibration set (80 samples) and a test set (40 samples) using Kennard–Stone (KS) algorithm.

The sensitivity of Raman scattering to detect citalopram in pharmaceutical tablets has been checked. The changes are easily tracked due to a distinct strongly active cyanide (CN) Raman mode, found at 2233 cm^{-1}. However, Raman analysis is problematic in some way, as the tablet spectra suffer from a strong fluorescence background. This background seems to originate from an unknown residual in the primary excipient, i.e. microcrystalline cellulose. Cellulose contains trace amounts of organic substances that can be highly fluorescent. Better results were obtained with NIR transmittance spectroscopy, relying on second aromatic stretching C-H overtones at 8830 cm^{-1} (corresponding to 1132 nm). This peak is partially overlapped with the peak at 8200 cm^{-1} (1220 nm) originating from the primary excipient, microcrystalline cellulose.

The present work aims to explore and compare the ability of Raman and NIR spectroscopy combined with PLS analysis to detect active citalopram substance, on the one hand and on the other, to assess the efficacy of a particular code for background fluorescence correction. Baseline correction was achieved by using adaptive iteratively reweighted penalized least squares (airPLS) [2]. It proved its usefulness as the baseline was successfully corrected without losing spectral information. Furthermore, two data fusion strategies were used (both low-level and mid-level fusion) for combining the NIR and Raman spectra to generate fusion spectra. For evaluating the response of different models to data fusion strategy, partial least squares (PLS) regression and radial basis function artificial neural network (RBF ANN) were applied to the modeling of four different groups of spectral data[3].

References

1. Dyrby, M., S. Balling Engelsen, L. Nørgaard, M. Bruhn, and L. Lundsberg-Nielsen. "Chemometric quantitation of the active substance (containing C≡ N) in a pharmaceutical tablet using near-infrared (NIR) transmittance and NIR FT-Raman spectra." *Applied spectroscopy* 56, no. 5 (2002): 579-585.
2. Zhang, Z. M., Chen, S., & Liang, Y. Z. (2010). Baseline correction using adaptive iteratively reweighted penalized least squares. *Analyst*, *135*(5), 1138-1146.
3. Azcarate, S. M., Ríos-Reina, R., Amigo, J. M., & Goicoechea, H. C. (2021). Data handling in data fusion: Methodologies and applications. *TrAC Trends in Analytical Chemistry*, *143*, 1163.

PAPER-BASED ORGANIC ELECTROCHEMICAL TRANSISTOR ARRAY FOR MULTI-ANALYTE DETECTION

Jordi Riu, **Ariadna Dasca, Pascal Blondeau, Francisco J. Andrade**

Universitat Rovira i Virgili, Department of Analytical Chemistry and Organic Chemistry, 43007 Tarragona, Spain

---

The increase in chronic diseases and the delay in their detection makes the development of point-of-care devices of vital importance. Traditionally, their prevention and control was centralized in laboratories because the analytical techniques used are complex, expensive, and require trained personnel. This caused that in extreme situations, such as the COVID-19 pandemic, many patients were diagnosed in late stages of their disease and others could not continue with their treatment due to congestion in hospitals. The development of artificial intelligence along with machine learning, together with the development of portable sensing devices, has paved the way for the creation of decentralized platforms capable of remote communication to facilitate decision making.

Electrochemical biosensors have emerged as devices capable of overcoming these limitations by being fast, simple of operation and low-cost. In this context, the use of chemometrics is a key factor in electrochemical analysis, providing a valuable resource for the precise and simultaneous quantification of analytes in complex matrices with multiple interferers [1].

Organic electrochemical transistors (OECTs) are presented as a promising solid-state devices due to the low power consumption, ion-to-electron transduction and high signal amplification. Combination of thick-film technologies with ion-selective membranes provides high analytical performance, low operational voltages and high switching speeds [2].

The present work proposes an array of biochemical transistors capable of predicting and monitoring K^+, Na^+ and NH_4^+ simultaneously thanks to the development of chemometric models. This compact device will open the doors to remote monitoring for people with cystic fibrosis.

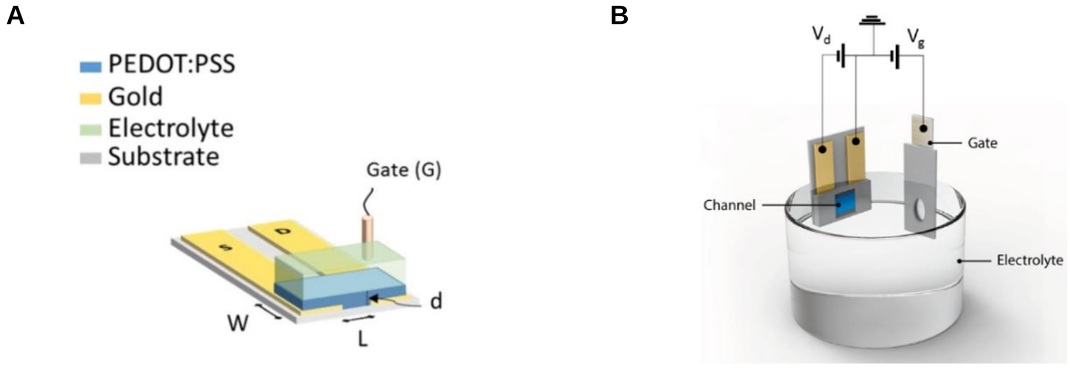


Figure 4. (A) Schematic of an OECT, where S is the source, D is the Drain, W is the width, L is the length, and d is the thickness of the channel. (B) Schematic of the measurement cell.

References

[1] Hadi Parastar, Dmitry Kirsanov **(2020)** Analytical figures of merit for multisensor arrays. *ACS Sensors*, **5**, 580-587

[2] Adil Ait Yazza, Pascal Blondeau, F. Javier Andrade (**2021**) Simple approach for Bublding high transconductance paper-based organic electrochemical transistor (OECT) for chemical sensing. *ACS Applied Electronic Materials*, **3**, 1886-1895

AN INSIGHT INTO THE PROBLEMS OF DETERMINING THE ADULTERATION OF ALMOND FLOUR WITH LOW-COST MINIATURISED NIR INSTRUMENTS

Jordi Riu[1], Somaya Ouzakar[1], Barbara Giussani[2]

[1]Universitat Rovira i Virgili, Department of Analytical Chemistry and Organic Chemistry,
43007 Tarragona, Spain
[2]Università degli Studi dell'Insubria, Science and High Technology Department,
22100 Como, Italy
jordi.riu@urv.cat

Food adulteration is a more common practice than it may seem, and that can have serious consequences for public health [1]. Adulterated food products can contain low-cost or harmful substances that may cause illnesses, allergies or even death. It is crucial to detect and prevent food adulteration to ensure food safety and to prevent frauds against consumers [2]. To achieve this, it is essential to have effective methods for detecting food adulteration. Traditional laboratory techniques may be time-consuming and expensive, leading to delayed detection and a higher risk of contaminated food entering the market.

Portable instruments offer a convenient and cost-effective alternative for detecting food adulteration on-site, allowing for quicker and more efficient monitoring of food quality. These instruments are easy to use, require minimal sample preparation, and provide rapid and reliable results. Near-infrared (NIR) spectroscopy is a well-established and mature technique, which equipment, as most of the analytical instrumentation, has evolved from large, benchtop-based instruments through on-line and in-line instruments for industry, to miniaturized portable devices. Miniaturized NIR instrumentation is gaining importance in the last years despite the plethora of different instruments with different technological solutions which require of suitable and structured analytical strategies to obtain reliable results [3]. It is also important to be aware that it is not usually possible to directly adapt analytical methods based on benchtop NIR instrumentation to miniaturized NIR instrumentation, because both types of instruments have their own sources of variability. Therefore, understanding the different sources of variability [4] and the influence of the experimental parameters, either from the sample to be measured or from the temporary distribution of the measurements, may help to obtain a better performance of the results obtained with these instruments and to have a clear vision of the features and limitations of miniaturized NIR instruments.

In this study we have used different NIR miniaturized instruments from two different companies, with different characteristics and modes of operation, for the measurement of sweet and bitter almond flour as a case study, with the aim of evaluating the performance of the constructed multivariate models for the prediction of the adulteration of sweet almonds flour with bitter almonds. Different experiment parameters such as the granulometry of the samples, the influence of the day of measurement or the number of replicates, have been varied during the measurements with the different instruments.

References

[1] Jose Marcelino Netto, Fernanda A. Honorato, Paulo Gustavo Celso, Maria Fernanda Pimentel, (**2023**) Authenticity of almond flour using handheld near infrared instruments and one class classifiers. *Journal of Food Composition and Analysis*, **115**, 104981

[2] Oksana Ye, Rodionova, Alexey L. Pomerantsev, (**2020**), Chemometric tools for food fraud detection: The role of target class in non-targeted analysis, *Chemometrics and Intelligent Laboratory Systems*, **317**, 126448

[3] Barbara Giussani, Giulia Gorla, Jordi Riu, (**2022**) Analytical Chemistry Strategies in the Use of Miniaturised NIR Instruments: An Overview. *Critical Reviews in Analytica Chemistry*. DOI: 10.1080/10408347.2022.2047607

[4] Giulia Gorla, Alessandro Taiana, Ricard Boqué, Paolo Bani, Olga Gachiuta, Barbara Giussani **(2022**) Unravelling error sources in miniaturized NIR spectroscopic measurements: The case study of forages. *Analytica Chimica Acta*, **1211**, 339900

Study of the resistance of *Bacillus subtilis* spores to UV light via Raman, FTIR and O-PTIR spectroscopy. Use of multiple co-inertia analysis (mCIA)

J. Dikec,[1] M. Pacheco,[1] L. Dujourdy,[3] C. Sandt,[4] P. Winckler,[1,4] JM Perrier-Cornet[1,4]

1 UMR Procédés Alimentaires et Microbiologiques, L'Institut Agro Dijon, Université de Bourgogne-Franche-Comté, 1, Esplanade Erasme, 21000 Dijon, France.

2 L'institut Agro Dijon – Direction Scientifique - Cellule d'appui à la recherche en science des données, F-21000 Dijon, France : laurence.dujourdy@agrosupdijon.fr

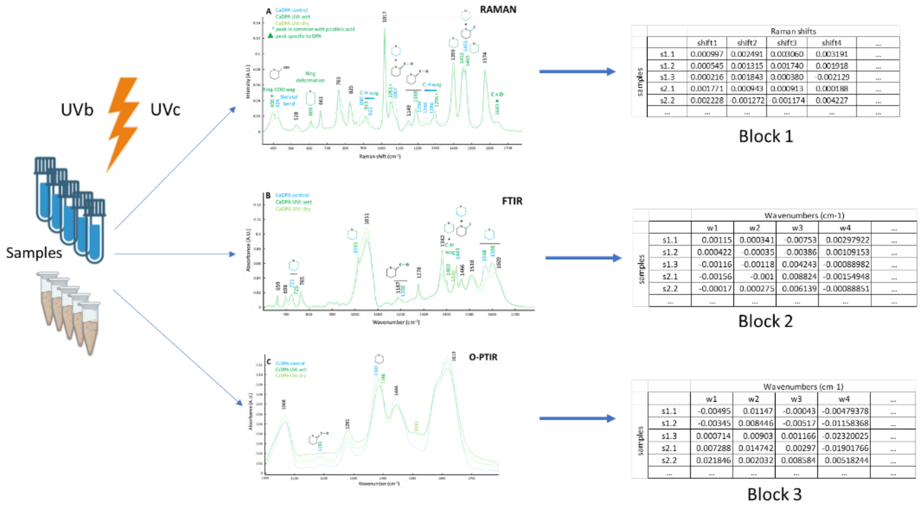
3 SMIS beamline, Synchrotron Soleil, L'Orme des Merisiers, BP 48 Saint Aubin, 91192 Cedex Gif-sur-Yvette, France

4 Dimacell Imaging Facility, L'Institut Agro Dijon, Université de Bourgogne Franche-Comté, 1 Esplanade Erasme, 21000 Dijon, France.

Bacterial spores can cause major problems, such as food poisoning (like neurotoxins or emetic toxins) or serious diseases (like anthrax or botulism). This dormant form of bacteria is made of several barriers offering an extreme resistance to many abiotic stresses (radiation, temperature, pressure, etc.) which makes its study in situ difficult. In order to better understand the biological and chemical mechanisms involved and specific to spore resistance, the acquisition of environmental parameters is necessary (Pacheco et al. 2022). Dipicolinic acid (DPA) is a specific molecule in bacterial spores that is essential for their resistance to various stresses such as exposure to ultraviolet (UV) light and for their germination. DPA has a particular photochemistry that is not fully understood. More than 90% of DPA is chelated to calcium ions forming calcium dipicolinate (CaDPA) (Dikec et al. 2022). CaDPA is known to be either a photoprotective or a photosensitiser agent in terms of UV radiation, depending on its hydration state. This molecule is also known to produce numerous photoproducts (DPAp) upon UV exposure, some of which influence spore resistance.

The purpose of this study is to better characterise the influence of hydration on the photochemistry of CaDPA using vibrational spectroscopy through three different techniques: Raman, FTIR and O-PTIR microspectroscopy. All analyses were performed at the SMIS beamline of the Synchrotron Soleil.

Synthetic CaDPA was exposed to UVc or UVb in dry conditions or in a water solution. Raman, FTIR and O-PTIR spectra of dry or wet CaDPA were compared before and after irradiation. Fig. 1 shows the schematic protocol.



		Raman shifts				
		shift1	shift2	shift3	shift4	...
samples	s1.1	0.000997	0.002491	0.003060	0.003191	...
	s1.2	0.000545	0.001315	0.001740	0.001918	...
	s1.3	0.000216	0.001843	0.000380	-0.002129	...
	s2.1	0.001771	0.000943	0.000913	0.000188	...
	s2.2	0.002228	-0.001272	-0.001174	0.004227	...
	...	...	...	...	...	...

Block 1

		Wavenumbers (cm-1)				
		w1	w2	w3	w4	...
samples	s1.1	0.00115	0.000341	-0.00753	0.00297922	...
	s1.2	0.000422	-0.00035	-0.00386	0.00109153	...
	s1.3	-0.00116	-0.00118	0.004243	-0.00088982	...
	s2.1	-0.00156	-0.001	0.008824	-0.00154948	...
	s2.2	-0.00017	0.000275	0.006139	-0.00088851	...
	...	...	...	...	...	...

Block 2

		Wavenumbers (cm-1)				
		w1	w2	w3	w4	...
samples	s1.1	-0.00495	0.01147	-0.00043	-0.00479378	...
	s1.2	-0.00345	0.008446	-0.00517	-0.01158368	...
	s1.3	0.000714	0.00903	0.001166	-0.02320025	...
	s2.1	0.007288	0.014742	0.00297	-0.01901766	...
	s2.2	0.021846	0.002032	0.008584	0.00518244	...
	...	...	...	...	...	...

Block 3

Figure 1: schematic representation of the analytical protocol.

Several data tables were obtained for each spectroscopic analysis. In this study, the differences in the spectra of UV-exposed CaDPA are investigated to understand the particularities of these spectra following exposure. Some of these differences are directly visible on the spectra but others are more difficult to detect and interpret. Sometimes minor differences may contain a large amount of information which can be critical to understanding the data completely. For this reason, direct interpretation of the spectral results can prove difficult. To overcome these difficulties and obtain deeper insights, we need to use data mining methods. These methods must be suitable for handling large multidimensional datasets and exploring all spectral information simultaneously.

Principal component analysis as an exploratory method was used initially after data pre-processing (which may differ for each type of spectral analysis), but we sought to identify common relationships and assess concordance among the multiple datasets. So, multiple co-inertia analysis (mCIA) was used because it is an ordination method based on a criterion for optimising the covariance between tables. It provides a representation of the stable and unstable parts of the variability of the assemblage structure with respect to a reference structure derived from each table. Recently it has been applied to integrative analysis of multiple -omics datasets (Cantini et al. 2021 and Subramanian et al. 2020). Similarities between the datasets or relevant features can be identified by the use of graphical representations. On the basis of this exploratory study, it may be said that whatever the platform considered (FTIR, Raman or O-PTIR), UVc wet and control samples stand out from the others. The FTIR domain is very discriminating: dry samples exposed to UVc are well separated along the dimension 1 from samples in solution exposed to UVc; dimension 2 separates dry samples exposed to UVb from samples in solution exposed to UVb. With Raman spectroscopy, dimension 1 separates the samples according to the nature of the UV: samples treated with UVb are positioned along the negative side of the axis and samples treated with UVc are on the positive side. In the case of O-PTIR spectroscopy, this separation between the two UV treatments is less clear.

While the information obtained via O-PTIR seems to be closer to that provided by Raman than to that provided by FTIR, the three techniques seem to be quite complimentary.

The results provide details on the influence of the hydration of CaDPA during both UVc and UVb exposure. The hydration state seems to have a slight influence on CaDPA photoreactions, also visible in the pyridine ring bands. These findings may help to unravel the molecule’s dual role in the UV resistance of spores (Dikec et al.,2023).

References

Pacheco M, Dikec J, Winckler P, Coelho C, Perrier-Cornet J-M (**2022**) Spectroscopic and microscopic characterization of dipicolinic acid and its salt photoproducts – A UVc effect study on DPA in solution and in bacterial spores, *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, **280**, 121502.

Dikec J, Pacheco M, Lavaud M, Winckler P, Perrier-Cornet J-M (**2022**) Uptake of UVc induced photoproducts of dipicolinic acid by Bacillus subtilis spores – Effects on the germination and UVc resistance of the spores, *Journal of Photochemistry and Photobiology B: Biology*, **236**, 112569.

Cantini, L., Zakeri, P., Hernandez, C. Naldi A, Thieffry D, Remy E, Baudot A (**2021**) Benchmarking joint multi-omics dimensionality reduction approaches for the study of cancer. *Nature Communications*, **12,** 124-136.

Subramanian I, Verma S, Kumar S, Jere A, Anamika K. (**2020**) Multi-omics Data Integration, Interpretation, and Its Application. *Bioinformatics and Biology Insights*, **14**, 1177932219899051.

Dikec J, Pacheco M, Dujourdy L, Sandt C, Winckler P, Perrier-Cornet J-M (**2023**), Influence of hydration on calcium dipicolinate (CaDPA) during UVb and UVc exposure studied via Raman, FTIR and O-PTIR spectroscopy, Journal of Photochemistry & Photobiology, A: Chemistry, **submitted**.

Leveraging Machine Learning Approaches for Forensic Applications using the Attenuated total reflectance Fourier transform infrared (ATR-FTIR) Spectroscopy

Lenka Halamkova, Bilkis Mitu, Migdalia Cerda

Address: Department of Environmental Toxicology
1207 Gilbert Drive, Box 41163, Lubbock, Texas
Texas Tech University
Email: Lenka.Halamkova@ttu.edu

Vibrational spectroscopy, including Raman scattering and Infrared absorption, is a very selective spectroscopic technique providing precise information on chemical/biochemical composition and molecular structure of complex samples. Attenuated total reflection Fourier transform infrared (ATR FT-IR) spectroscopy is a technique that has been widely used in many fields. This technique is fast, simple, non-destructive, and requires only small quantities of measured material with minimal-to no sample preparation. A significant advantage of ATR FT-IR is a specific spectral signature for different sample types based on their biochemical composition. The infrared spectrum displays unique vibrational characteristics of a sample based on the different absorption frequencies of the individual functional groups (Manheim et al., 2016). In addition to that, spectra from heterogeneous biological systems such as cells, body fluids, etc. can be associated with hundreds or thousands of biomolecules. Biological samples for forensic and medical analysis contain components such as proteins, nucleic acids, carbohydrates and lipids, etc. FT-IR spectra of these samples are very complex and must be interpreted with caution. Such spectra are very complex with underlying systematic variation that can make statistical analysis difficult.

To elucidate multiplex spectral information of sample composition and to enable the analysis of many spectra required for further application, multivariate statistical techniques are needed. Machine learning (ML) based computational framework has the potential to characterize such complex data sets and to enable understanding of very complex mathematical patterns and relationships hidden within the original spectral fingerprinting targeted in a particular project (Srivastava et al., 2020).

ML works on principles of pattern recognition in data and error minimization. ML algorithms seek regularities in spectral signatures of known examples, train a system to identify learned patterns and make subsequent predictions on unknown data. This is a contrary approach to traditional FTIR spectral library search approach, since it does not exclusively focus on spectral features originating from one specific molecule/compound itself. Specifically, we will demonstrate here the great potential of ML for ATR FT-IR spectroscopy in forensic applications to extract knowledge from ATR FT-IR data and produce outputs that could be used for individual outcome prediction analysis and forensic decision making (Takamura et al., 2019; Mistek et al. 2019).

References

Srivastava R, Kumar Mallick P, Rautaray SS, Pandey M (**2020**). Computational intelligence for machine learning and healthcare informatics, Berlin, Boston: DeGruyter, https://doi.org/10.1515/9783110648195

Manheim J, Doty KC, McLaughlin G, Lednev I K (**2016**). Forensic hair differentiation using attenuated total reflection Fourier transform infrared (ATR FT-IR) spectroscopy. *Applied Spectroscopy*, **70**(7), 1109-1117.

Takamura A, Halámková L, Ozawa T, Lednev IK (**2019**). Phenotype Profiling for Forensic Purposes: Determining Donor Sex Based on Fourier Transform Infrared Spectroscopy of Urine Traces. *Analytical Chemistry*, **91**(9), 6288-6295, DOI: 10.1021/acs.analchem.9b01058

Mistek E, Halámková L, Lednev IK (**2019**). Phenotype profiling for forensic purposes: Nondestructive potentially on scene attenuated total reflection Fourier transform-infrared (ATR FT-IR) spectroscopy of bloodstains. *Forensic Chem.*, **16**, 100176. 10.1016/j.forc.2019.100176.

Multivariate statistics approach to infer metabolite profile variations in Pecorino Romano Cheese

Leonardo Sibono[1], Massimiliano Grosso[1], Stefania Tronci[1],
Massimiliano Errico[2], Cristina Manis[3] , Pierluigi Caboni[3]

[1] Department of Mechanical, Chemical and Materials Engineering, University of Cagliari, Via Marengo 2, Cagliari, 09123, Italy; [2] Department of Green Technology, University of Southern Denmark, Campusvej 55, 5230 Odense, Denmark; [3] Department of Life and Environmental Sciences, University of Cagliari, 09124 Cagliari, Italy
Email: leonardo.sibono@unica.it

The production season of milk is a crucial factor that affects the fatty acid composition (Scano and Caboni, 2021) and mineral content (González et al., 2011) of cheese products, thus influencing their nutritional quality. Sheep's milk cheese, specifically, can have a positive impact on human health due to its bioactive unsaturated fatty acids, such as conjugated linoleic acids (CLA, e.g. C18:2 *cis* 9, *trans* 11) and vaccenic acid (C18:1 *trans* 11), which play a significant role as nutraceutical substances. The present study aimed to investigate the effect of cheese production season on the fatty acid profile, expressed as % Fatty Acids Methyl Esters (FAME). The mineral profile was also assessed in terms of calcium, magnesium, sodium, potassium, phosphorus, sulphur, zinc, and iron content. The study involved two tasks, namely classification and inference. First, 27 cheese samples were classified according to their production month (January, April, June). Three external samples were then taken from each period (totalling 9 external samples). These samples were used to infer the unknown %FAME profile and were classified for external validation purposes. Probabilistic Principal Component Analysis (PPCA) (Shah et al. 2017) was used to estimate missing data. Linear Discriminant Analysis (LDA) was then performed on the first two Principal Components (PCs) as a classification tool. The analysis of the model loadings allowed for the assessment of the correlation between the variables and the cheese production month. The PPCA scores relevant to the external samples were used to reconstruct missing data observations and were plotted in the PPCA+LDA score plot to evaluate the model's capability in classifying external samples whose fatty acid profile was unknown. 5-fold Venetian Blind Cross Validation (VBCV) was used for internal validation to determine the robustness of the model. The results indicated a positive correlation between short-chain fatty acids and the winter period (January), between polyunsaturated fatty acids (PUFA) and spring (April), and a prevalence of long-chain saturated fatty acids during the early summer (June). Such results are in line with those reported in other studies (Caboni et al., 2019). The model showed remarkable performance in autoprediction and cross-validation, with 100% of the samples being correctly classified in both cases. The external validation results indicated that the PPCA+LDA model was effective in correctly classifying all the samples, suggesting that the proposed approach may be useful in determining the fatty acid content and classifying dairy products based on their production season. Classification results are reported in Fig. 1.

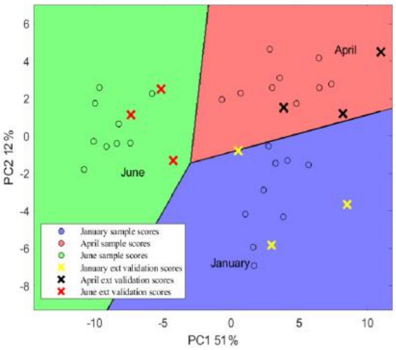


Fig. 1. Score plot for PPCA+LDA classification. Dot symbols: autoprediction samples. Cross symbols: external validation samples.

References

Caboni, P., Maxia, D., Scano, P., Addis, M., Dedola, A., Pes, M., Murgia, A., Casula, M., Profumo, A., & Pirisi, A. (2019). A gas chromatography-mass spectrometry untargeted metabolomics approach to discriminate Fiore Sardo cheese produced from raw or thermized ovine milk. Journal of Dairy Science, 102(6), 5005–5018. DOI: 10.3168/jds.2018-15885

González-Martín MI, Hernández-Hierro JM, Revilla I, et al. Differentiation of organic and non-organic ewe's cheeses using main mineral composition or near infrared spectroscopy coupled to chemometric tools: A comparative study. Talanta. 2011;85(4):1915-1919. doi:10.1016/j.talanta.2011.07.021

Shah, S. M. S., Batool, S., Khan, I., Ashraf, M. U., Abbas, S. H., & Hussain, S. A. (2017). Feature extraction through parallel Probabilistic Principal Component Analysis for heart disease diagnosis. Physica A: Statistical Mechanics and Its Applications, 482, 796–807. DOI:10.1016/j.physa.2017.04.113

Anomaly Detection in Production – Diagnostics and Insights from Data

Castellino L.[a], Alladio E.[a], Foglio P.[b]

[a]: Università degli Studi di Torino, Dipartimento di Chimica, Via P. Giuria 7, Torino, 10125, Italy
[b]: Iveco Group, Via Puglia 35, Torino, 10156, Italy
lorenzo.castellino@unito.it

Anomaly detection refers to the identification of data points that are unusual or deviate significantly from what is expected[1]. This technique has many applications, including fraud detection, cybersecurity, medical diagnosis, and production monitoring. Regarding production monitoring applications, the ability to identify the development of an anomaly in the behavior of the monitored machine with a reasonable margin of time allows anticipating the insurgence of a blocking issue, providing potentially great value from an economical and environmental standpoint.

The detection of anomalies can be achieved through statistical methods, machine learning algorithms, and deep learning techniques. Recent years have seen an increase in popularity of the latter, as it has proven to be highly effective in detecting anomalies even in complex data sets. The main issue with deep learning techniques resides in the fact that the obtained models are affected by "black-boxing", meaning it is usually difficult for the user to understand the relations learned by the model, making it difficult to understand how the model is performing its predictions.

With the present work, we want to share our experience on the topic that has matured working with Iveco Group in the context of the DOLPHINS Project[2] for the anomaly detection and applied predictive maintenance on machineries of the production lines in the Brescia(IT), Suzzara(IT) and Valladolid(ES) plants.

Three methodologies for signal reconstruction are presented: Auto Associative Kernel Regression (AAKR)[3], Principal Component Analysis (PCA)[4] and Artificial Neural Nets Autoencoders[5]. The whole project has been coded in Python, leveraging already existing libraries when possible and writing custom routines when needed. The whole data pipeline is handled with the Kedro library, facilitiating communication between the project partners while allowing code's reproducibility and portability.

The techniques presented could be applied in many contexts and at very different scales, providing tools to understand better and take better decisions in planning and maintaining a production facility.

References

1. Varun C., Arindam B., Vipin K. (**2009**) Anomaly detection: A survey. *ACM Comput. Surv.*, **41**, 3-61.
2. Alladio, E., Baricco, M., Leogrande, V., Pagliari, R., Pozzi, F., Foglio, P., Vincenti, M. (**2021**). The “DOLPHINS” Project: A Low-Cost Real-Time Multivariate Process Control From Large Sensor Arrays Providing Sparse Binary Data. *Frontiers in Chemistry*, **9**, 734132.
3. Baraldi P., Di Maio F., Turati P., Zio E. (**2015**) Robust signal reconstruction for condition monitoring of industrial components via a modified Auto Associative Kernel Regression method. *Mechanical Systems and Signal Processing*, **60–61**, 29-44.
4. Harrou, F., Kadri, F., Chaabane, S., Tahon, C., Sun, Y. (**2015**) Improved principal component analysis for anomaly detection: Application to an emergency department. *Computers & Industrial Engineering*, **88**, 63-77.
5. Bengio, Y., Courville, A., Vincent, P. (**2013**) Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, **35**, 1798-1828.

Chemometrics strategies to assess mountain products identity

Lorenzo Strani[1], Caterina Durante[1], Pierpaolo Becchi[1], Sara Michelini[2], Valentina Pizzamiglio[2], Marina Cocchi[1]

Addresses: [1]University Modena and Reggio Emilia, Department of Chemical and Geological Sciences, Via Campi 103, Modena, 41125, Italy
[2]Consorzio del Formaggio Parmigiano Reggiano, Via Kennedy 18, Reggio Emilia, 41214, Italy
E-mail lostrani@unimore.it

“Mountain Product” is a relatively new quality label defined in the European Community regulation [1] reserved for food products produced and processed in mountain areas, aiming to support the economy in disadvantaged areas. However, the “Mountain Product” denomination is still not enough spread and not well known by consumers, the present work is part of a larger research project whose objective is to develop an analytical methodology to conjugate identity, sustainability, and value, i.e., MOUNTAIN-ID (https://www.mountainid.unimore.it). Moreover, very few studies attempt to differentiate the same food product produced in a mountain area from the one obtained in the plane. Both dairy farmers and consortia want to protect the authenticity of their products from analogues and, at the same time, promote them. In this scenario, the present study aimed at assessing among different samples of Parmigiano Reggiano, a PDO (Protected Designation of Origin) product [2], the ones with the mountain denomination, namely the “Progetto Territorio Parmigiano Reggiano Prodotto di Montagna” (PdM). In order to obtain the identitary traits, different analytical approaches are employed, including spectroscopic techniques used as fingerprinting (untargeted) strategies. With the aim to differentiate PdM with respect non-PdM (as well PDO Parmigiano Reggiano, but produced in plain areas) products, chemometrics models are developed testing different strategies, i.e. class modeling and discrimination, and fusing information from different spectroscopies.

References

[1] regulation (EU) No 1151/2012
[2] https://www.parmigianoreggiano.com/product-guide-seals-and-marks#5

Assessment of iron nuclearity in mineral fibers by means of MCR-ALS

Lorenzo Strani, Marina Cocchi, Alessandro F. Gualtieri

Addresses: University Modena and Reggio Emilia, Department of Chemical and
Geological Sciences, Via Campi 103, Modena, 41125, Italy
E-mail lostrani@unimore.it

Iron nuclearity, defined as the number of iron atoms combined in a singular coordination object by bridging oxygen atoms, is one of the parameters used to assess the toxicity potential of mineral fibres. Isolated ferrous iron atoms have the highest potential to catalyse the formation of toxic reactive oxygen species (ROS) responsible for adverse effects *in vivo* while iron clusters seem less effective in producing ROS in the biological medium (Gualtieri, 2021). In this context, UV-Vis spectroscopy has proven promising in identifying iron structure in these complex natural samples (see for example, Martra et al., 1999 with crocidolite and Borghi et al., 2010 with chrysotile). However, these methods show low sensitivity, being unreliable if applied to natural systems with low iron contents. In the current study, the authors propose a novel approach, coupling UV-Vis spectroscopy with Multivariate Curve Resolution – Alternating Least Squares (MCR-ALS, De Juan & Tauler, 2006), aiming to obtaining information on the iron chemical environment, namely if the iron nuclearity can be ascribed to Fe^{2+} isolated, Fe^{3+} isolated or Fe cluster. Twenty-seven different samples of both synthetic and natural iron-containing mineral fibres have been collected and analysed by Solid UV Vis with integrating sphere accessory. MCR-ALS method has been employed to resolve specific spectral profiles for each type of nuclearity and assess the extent of each of them in the sample. Data has been partitioned in calibration (17 samples) and test (10 samples) sets, in order to validate the model. The results show that a model with 4 components provided the best results, being able to distinguish mineral fibres according to the three investigated iron nuclearity modes. These results will be of great help to clarify the role of iron in the existing models for the assessment of the toxicity potential of mineral fibres.

References

Borghi E, Occhiuzzi M, Foresti E, Lesci I G, Roveri N (**2010**) Spectroscopic characterization of Fe-doped synthetic chrysotile by EPR, DRS and magnetic susceptibility measurements. *Physical Chemistry Chemical Physics*, **12(1)**, 227-238.

De Juan A, Tauler R (**2006**) Multivariate curve resolution (MCR) from 2000: Progress in concepts and applications. *Critical Reviews in Analytical Chemistry*, **36**, 163-176.

Gualtieri AF (**2021**). Bridging the gap between toxicity and carcinogenicity of mineral fibres by connecting the fibre crystal-chemical and physical parameters to the key characteristics of cancer. *Current Research in Toxicology*, **2**, 42-52.

Martra G, Chiardola E, Coluccia S, Marchese L, Tomatis M, Fubini B (**1999**) Reactive sites at the surface of crocidolite asbestos. *Langmuir*, **15(18)**, 5742-5752.

Hyperspectral near infrared imaging to investigate waste wood material characteristics

M. Mancini[1,2], Å. Rinnan[1], M. Mäkelä[3], G. Toscano[2]

[1]Department of Food Science, Faculty of Science, University of Copenhagen, Rolighedsvej 30, DK-1958 Frederiksberg C, Denmark
[2]Agricultural, Food and Environmental Sciences, Università Politecnica delle Marche, via Brecce Bianche, IT-60131 Ancona, Italy
[3]VTT Technical Research Centre of Finland Ltd., PO Box 1000, FI-02044 Espoo, Finland
E-mail (manuela.mancini@staff.univpm.it)

Nowadays, there is an urgent need for a transition from fossil fuels to biomass-based economy and, in the last decades, European policies are moving in that direction. In this scenario, wood – and lignocellulosic material in general – are becoming more appealing. At the end of the life cycle, the utilized material ends up in the waste stream generating waste wood. Waste wood includes wood or wood-containing material from different sectors, which means it presents inherent high variability (Edo et al., 2016). Considering the high and continuously increasing demand in wood (Mantau et al., 2010) the big challenge is to reuse and recycle instead of dispose it. This is also in line with Waste Framework Directive of the European Union. Most waste wood can be reused or recycled and only wood with hazardous substances needs to be sent to the disposal with no benefits. Because of these contaminants the recycling potential of waste wood is still low (Faraca et al., 2019) and could increase if the material is properly sorted and handled based on its quality and characteristics. To this aim, waste wood material was divided in three categories: i) virgin wood that can be used both for panel board production and bioenergy applications, ii) treated wood that can be used for panel board production and iii) impregnated and painted wood (disposal wood) that should be sent to the disposal with no reuse. The material has been investigated using near-infrared (NIR) spectroscopy since such technique has the potential for the on-line sorting of the material and could represent a valid solution for the rapid screening.

The results of a previous study (under revision) proved that the spectra collected with a bench-top spectrophotometer can be used to develop Soft Independent Modeling of Class Analogy (SIMCA) classification models and discriminate between the three waste wood categories. But it showed also that some samples are more challenging to be classified than others because of their similarities in the chemical composition. For this reason, a subset of waste wood samples have been analyzed with NIR hyperspectral camera (Specim SWIR 3 - Specim, Spectral Imaging, Ltd, Oulu, Finland). In detail five waste wood types (virgin wood, treated wood, old wood, impregnated wood and painted wood) have been selected for this study for a total of 16 images of pure materials and 12 images of mixtures material. The idea is to investigate if NIR hyperspectral camera can facilitate the visualization of the distribution of different chemical components in a sample. In fact, since each spectrum is related to a pixel in the image, it is most suitable for analysis of heterogeneous samples. As a first step, Principal Component Analysis (PCA) has been computed to investigate the spectral differences and search for groupings among the five waste wood types. PCA is also a powerful tool for describing the variation in multivariate data sets and investigate spectral variability. We considered both pixel-based and object-based approaches and first conclusions were drawn on their advantages and limitations, also compared to the 'traditional' NIR technique. Next steps will aim in quantifying their respective effects (both positive and negative) on the classification results.

Acknowledgements

The project leading to this application has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No. 838560.

References

Edo, M., Björn, E., Persson, P.-E., & Jansson, S. (**2016**). Assessment of chemical and material contamination in waste wood fuels – A case study ranging over nine years. *Waste Management*, *49*, 311–319. https://doi.org/10.1016/J.WASMAN.2015.11.048

Faraca, G., Tonini, D., & Astrup, T. F. (**2019**). Dynamic accounting of greenhouse gas emissions from cascading utilisation of wood waste. *Science of The Total Environment*, *651*, 2689–2700. https://doi.org/https://doi.org/10.1016/j.scitotenv.2018.10.136

Mantau, U., Saal, U., Prins, K., Steierer, F., Lindner, M., Verkerk, H., Eggers, J., Leek, N., Oldenburger, J., & Asikainen, A. (**2010**). *EUwood–real potential for changes in growth and use of EU forests. Final report. Hamburg, Germany. 160 p.*

Evidence of the cultivation method of durum wheat by ICP-MS measurements of the elemental content in the seeds and flours.

Fattobene Martina[*], Berrettoni Mario[*], Conti Paolo[*], Governatori Catia[+], Liu Fuyong[*], Nardi Sandro[+], Zamponi Silvia[*]

[*] School of Science and Technology - Chemistry section "Chemistry Interdisciplinary Project" Building Via Madonna delle Carceri - 62032 Camerino (MC) ITALY

[+] ASSAM - Agenzia per i Servizi nel Settore Agroalimentare della regione Marche Via Industria, 1 - 60027 Osimo Stazione (AN) ITALY

martina.fattobene@unicam.it

Cereals are among the most important staple food crops; they are cheap source of calories, protein and elements for the inhabitants worldwide. Among the cereals, Durum wheat (Triticum durum Desf.) is the base of many largely used foods like pasta. These foods provide a large part of the essential elements, however, when polluted, they can contribute significantly to overexposure to some elements. Wheat plants exploit the elements present in the soil for their biological needs but there is a significant difference in the concentrations of these elements in the different soils. Another important source of some elements is both atmospheric and soil pollution, furthermore in agricultural practices there is extensive use of substances containing potentially toxic elements for humans. Wheat plants will therefore accumulate elements differently based on different exposures of the plants to the aforementioned sources. It is important to understand, since the cereal seeds are used for food purposes, how many and which elements are accumulated in the seeds.

Some studies investigated the mineral elements in cereals [1,2,3], but only few of them concerned the durum wheat[4], however, to the best of our knowledge, none investigate the distribution of the element in the grain kernel. The distribution in the kernel is important because going from the outer to the inner of the seed, during the grinding process, we can obtain: bran, semolina and flour that are used for different kind of food products. The knowledge of their elemental affinity can help to produce food of special characteristic or reduce the impact of environmental pollution on the final food products.

Here we analyzed, by ICP-OES, 13 variety of durum wheat each grown both with traditional and organic methods. Twentyeight elements were quantified in each sample. The data matrix was analyzed in order to highlight the differences between the kind of grinded material and of the growing methods. There is evidence that most of the elements concentrate in the bran. PCA easily distinguish two groups, one with whole seed and bran and the other with semolina and flour. The difference between organic and conventional growing is more evident in the group with semolina and flour. Classification with respect to agronomic method gives good result by PLS-DA or even better with LDA and QDA. Only few elements characterize the growing methods.

References

[1] Zhang, Y.; Song, Q.; Yan, J.; Tang, J.; Zhao, R.; Zhang, Y.; He, Z.; Zou, C.; Ortiz-Monasterio, I. Mineral Element Concentrations in Grains of Chinese Wheat Cultivars. *Euphytica* **2009**, *174* (3), 303–313.

[2] Brizio, P.; Benedetto, A.; Squadrone, S.; Curcio, A.; Pellegrino, M.; Ferrero, M.; Abete, M. C. Heavy Metals and Essential Elements in Italian Cereals. *FOOD ADDITIVES & CONTAMINANTS PART B-SURVEILLANCE* **2016**, *9* (4), 261–267.

[3] Baize, D.; Bellanger, L.; Tomassone, R. Relationships Between Concentrations of Trace Metals in Wheat Grains and Soil. *Agronomy For Sustainable Development* **2009**, *29* (2), 297–312.

[4] Harmankaya, M.; Özcan, M. M.; Gezgin, S. Variation of Heavy Metal and Micro and Macro Element Concentrations of Bread and Durum Wheats and Their Relationship in Grain of Turkish Wheat Cultivars. *Environ. Monit. Assess.* **2011**, *184* (9), 5511–5521.

ICP-MS AND CHEMOMETRICS FOR THE CHARACTERIZATION AND DISCRIMINATION OF THREE DIFFERENT SPECIES OF EDIBLE INSECTS

Martina Foschi, Alessandra Biancolillo, Claudia Scappaticci, Angelo Antonio D'Archivio

Address: University of L'Aquila, Department of Physical and chemical sciences, Via Vetoio, Coppito, L'Aquila, 67100, Italy
E-mail (martina.foschi@univaq.it)

In light of the recent inclusion of the yellow mealworm (*Tenebrio molitor* larva) on the European Union's list of authorized foods, a study was designed to differentiate and characterize different edible insects. The work focused on three specific types of edible insects, namely the house cricket (*Acheta domesticus* L.), the buffalo worm (*Alphitobius Diaperinus*), and the yellow mealworm (*Tenebrio molitor* larva), in accordance with the European Food Safety Authority's (EFSA) deliberations regarding the limited number of insects approved for consumption. There are several reasons why (RE)-introducing edible insects into the human diet could be valuable. Firstly, insects have been a significant part of the human diet since ancient times, which makes it important to emphasize the term "reintroduction". Secondly, this could be part of a series of actions to support the 2030 Agenda for Sustainable Development goals. Indeed, adopting insects as a food source could help ensure food security and provide environmental and socio-economic benefits [1]. In general, insects have proved to be a good and low-input source of protein, minerals, vitamins, polyphenols, and fatty acids. However, it is evident that the nutritional composition of insects varies depending on the species, life stages, environment, and diet. Therefore, the purpose of the study was to characterize the elemental profile of insect species that have been authorized for introduction into the European market. The samples were purchased from a European online shop; dried aliquots were mineralized using a microwave-assisted digester (Ethos-One, Milestone, Bergamo, Italy). They were then analyzed using an Inductively Coupled Plasma-Mass Spectrometer (iCAP-TQe, Thermo Fisher). After validating the analytical method for accuracy and precision, the multi-elemental profiles were processed using chemometric methods to distinguish species based on their mineral composition.

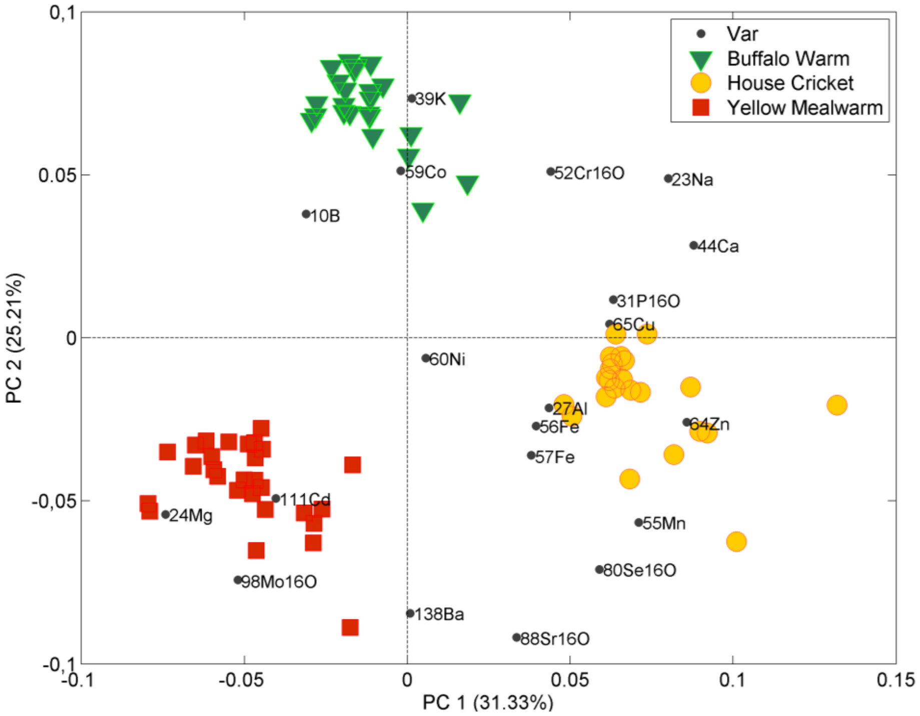


Fig.1: Biplot of the Principal Component Analysis obtained by the data matrix (75x20).

The Linear Discriminant Analysis produced satisfying results in predicting the external set (designated according to the Duplex algorithm). Several variable selection and variable reduction methods were tested and compared in this context [2].

References

Papastavropoulou K, Koupa A, Kritikou E, Kostakis M, Proestos C (**2022**) Edible insects: Benefits and potential risk for consumers and the food industry, *Biointerface Res. Appl. Chem.* **12**, 5131 – 5149.
Mishra P (**2022**) A brief note on a new faster covariate's selection (fCovSel) algorithm, *Journal of Chemometrics,* **36**, e3397.

MULTIVARIATE ANLYTICAL METHOD BASED ON LIQUID CROMATOGRAPHIC FINGERPRINTING AND CHEMOMETRICS TO AUTHENTIFICATE THE GEOGRAPHICAL ORIGIN OF TIGERNUTS

Miriam Medina-García*, **Antonio González-Casado, Ana M. Jiménez-Carvelo**

University of Granada, Dpto. Analytical Chemistry, Av. Fuentenueva, Granada, 18071, Spain.
miriammedina@ugr.es

Tigernuts is a typical foodstuff from a specific region of Valencia (Spain) known as "L'Horta Nord". Some tigernuts cultivated in this region are protected under a Protected Designation of Origin (PDO) "Chufa de Valencia" because they present unique characteristics associated with their particular production area. The high energetic and nutritive properties of tigernuts combined with its potential as an alternative foodstuff to other products of animal origin, have led the increasing in their demand and, therefore, the expansion of its cultivation around the world. For that reason, an exhaustive quality control to check the geographical origin is necessary in order to ensure the quality product for the consumer. In this line, the main goal of the research is the development of a new multivariate analytical method capable of authenticating the geographical origin of tigernut samples applying the tandem "fingerprinting" and "chemometrics".

For that purpose, the fat fraction of the tigernuts was extracted using the Pressurized Liquid Extraction technique (PLE). The optimal conditions of the extraction procedure were calculated using the design of experiments methodology (DoE). An ad-hoc analysis method was developed using liquid chromatography coupled with an ultraviolet-visible detector in a normal phase mode to acquire the instrumental fingerprint from fat fraction. Firstly, the fingerprint was examined by unsupervised chemometrics tools to observe the natural grouping of the samples using PCA method. After that, classification models (PLS-DA, SIMCA) were developed to discriminate between PDO tigernuts and Non-PDO tigernuts.

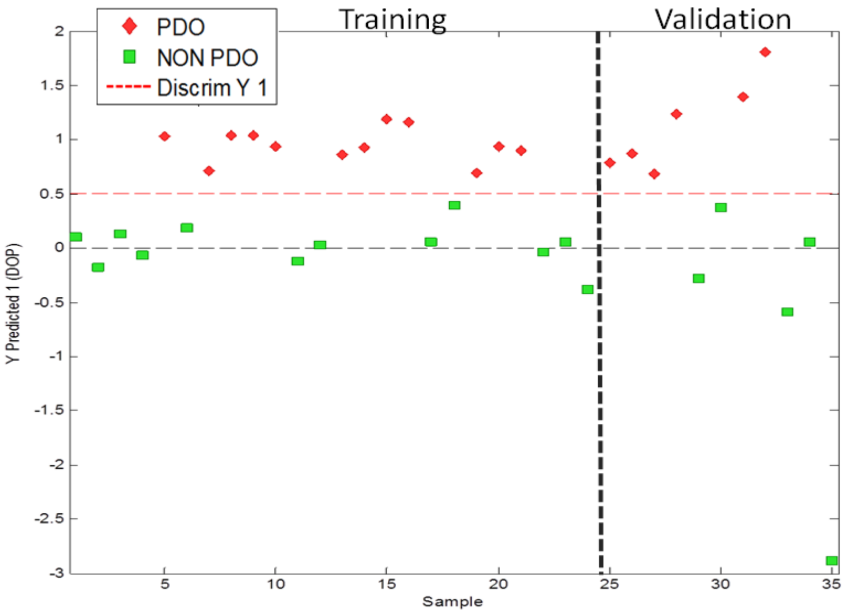


Figure 1: Sample assignment of tigernut samples from PLS-DA model developed with two input-classes: PDO and NON-DOP.

As a result, a methodology developed in this research made it possible the authentication of the tigernuts from different geographical origins, which makes it a tool with great potential to prevent food fraud.

References

1. Cuadros Rodríguez, L.; Pérez Castaño, E.; Ruiz Samblás, C. Quality performance metrics in multivariate classification methods for qualitative analysis (**2016**). *Trends Anal. Chem.*, **80**, 612-624.

A validated method for identifying mechanically separated meats by using multivariate analysis of 43 trace elements determined by iCP-MS

Monica Casale[a] , Oto Miedico[b], Valeria Nardelli[b], Teresa D'Amore[b], Paolo Oliveri[a], Cristina Malegori[a], Giuseppe Paglia[c], Marco Iammarino[b]

[a] Department of Pharmacy, University of Genoa, Viale Cembrano 4, 16148 Genoa, Italy
[b] Chemistry Department, Istituto Zooprofilattico Sperimentale della Puglia e della Basilicata, Via Manfredonia 20, 71121 Foggia, Italy
[c] School of Medicine and Surgery, University of Milano-Bicocca, Piazza dell'Ateneo Nuovo 1, 20126 Milan, Italy
E-mail (corresponding author): monica.casale@unige.it

The European Food Safety Authority asked for novel approaches for identifying mechanically separated meat (MSM) in meat products, due to food safety concern [1]. In this study, a novel approach based on multivariate analysis of 43 trace elements in meat products is described. Overall, 27 trace elements and 16 rare earth elements were determined by using ICP-MS after sample mineralization of 100 meat samples, composed of different percentages of MSM, obtained at low and high pressure, and without MSM. Chemometric classification models were developed in order to discriminate between meat products with or without MSM. Classification models were rigorously validated, both internally and externally: in details, a cross-validation strategy with six cancellation groups was applied for an internal validation achieving a total prediction rate of 97.2%. Then, the final validation of this approach was completed by analyzing and classifying 10 'blind' meat samples, obtaining a method accuracy equal to 90%, a significant improvement in field. Thus, the described method is applicable for the discrimination of meat products containing MSM, even if the processing was carried out at low and high pressure. This is the main strength of this approach. Indeed, most available approaches for the identification of MSM in meat products allow the clear distinction between MSM obtained by high pressure technology and meat not mechanically separated, while the differentiation between MSM obtained by low pressure technology and simple minced meat is not possible. This is due to the similarity between low-pressure MSM and fresh meat, since this type of mechanical deboning does not lead to significant release of bone fragments, bone marrow, etc., in meat. The analytical approach described in this study, on the contrary, can also be applied for identifying this type of MSM, since it exploits the potential of multi-elemental analysis (43 parameters) coupled to multivariate analysis.

This analytical method was also evaluated using the AGREE tool which investigates the 12 principles of green analytical chemistry. The final value obtained of 0.6, being higher than 0.5, assures the “GREEnness” of this approach [2-4].

Acknowledgment

This work was supported by Ministero della Salute (Rome, Italy), who financed the projects GR-2013-02358862 and IZS PB 01/17 RC. Mr. Rosario Accettulli, Mr. Michele Nicolini and Dr. Chiara Preite (Istituto Zooprofilattico Sperimentale della Puglia e della Basilicata, Foggia - Italy) are gratefully ackowledged for technical assistance.

References

[1] European Food Safety Authority. (2013). Scientific opinion on the public health risks related to mechanically separated meat (MSM) derived from poultry and swine. The EFSA Journal, 11(3), 3137.

[2] Pena-Pereira, F., Wojnowski, W., & Tobiszewski, M. (2020). AGREE—Analytical GREEnness Metric Approach and Software. Analytical Chemistry, 92(14), 10076–10082.

[3] Płotka-Wasylka, J. (2018). A new tool for the evaluation of the analytical procedure: Green Analytical Procedure Index. Talanta, 181, 204–209.

[4] Gałuszka, A., Migaszewski, Z., & Namie´snik, J. (2013). The 12 principles of green analytical chemistry and the SIGNIFICANCE mnemonic of green analytical practices. TrAC Trends in Analytical Chemistry, 50, 78–84.

ANALYSIS OF ELECTRONIC NOSE DATA FOR THE DISCRIMINATION OF VOLATILES FROM FOOD-SALIVA INTERACTION IN INDIVIDUALS WITH DIFFERENT BMI

Nicola Caporaso[1,2], Andrea Balivo[1], Raffaele Sacchi[1], Alessandro Genovese[1]

1. University of Naples "Federico II", Department of Agriculture, Portici (NA), 80055, Italy
2. Department of Research & Development, Buhler Sortex, London, E16 2BF, UK
nicola.caporaso3@unina.it; alessandro.genovese@unina.it (corresponding authors)

The lipid fraction of foods has a dramatic influence on their nutritional profile as well as their sensory impact. During the eating process, volatile compounds originate from the interaction between saliva and the food matrix, due to mechanical action and enzymatic reactions.

It is known that the activity of lipase and α-amylase is higher in subjects with larger body mass index (BMI). Recent research showed that the release of certain volatiles is linked to the variation of salivary composition, that is different in normal weight and obese subjects [1,2].

The electronic nose (EN) is an instrument using electronic chemical sensors that detect specific group of volatile molecules. An EN often contains multiple sensors, such that several groups of molecules are recorded.

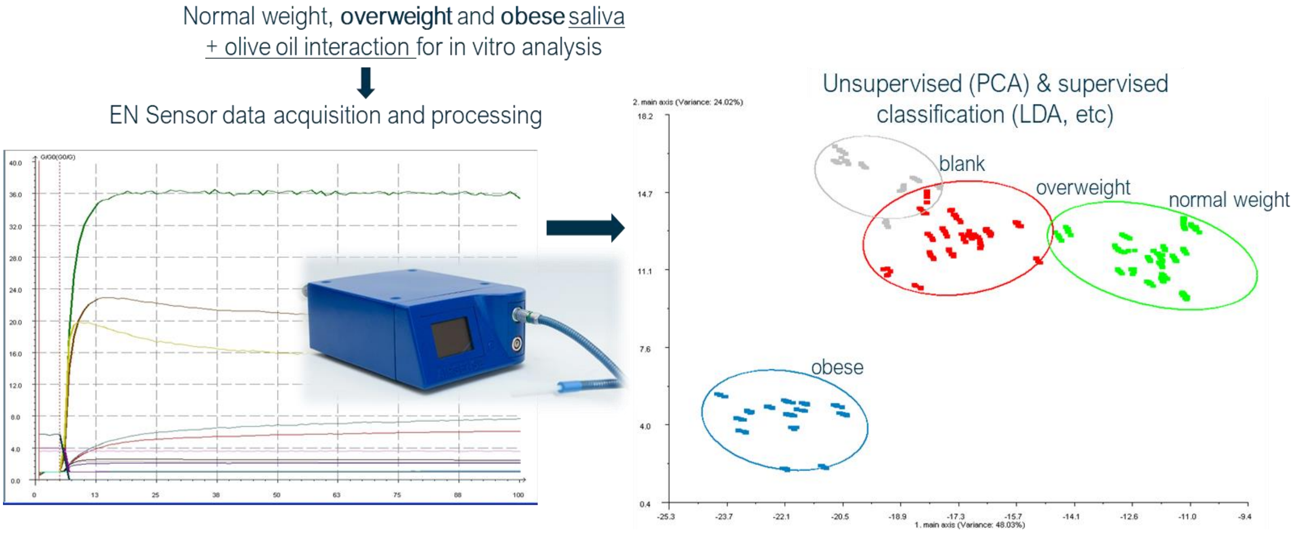
This contribution will present an experiment aimed to verify the ability of EN in detecting differences in volatiles emitted after the interaction of a food product with human saliva produced by subjects with different BMIs.

Extra virgin olive oil (EVOO) was used as our fat of preference, which was combined with saliva from 60 subjects (age range: 18-65 years), of which 23 had normal BMI, 18 was overweight and 14 was obese (water was used as the blank for the analysis). After 4 minutes of interaction between 0.3 mL of saliva and 100 mL EVOO, the released volatiles were recorded up to 100 seconds using an Electronic Nose PEN2 (Airsense Analytics, Germany) fitted with 10 sensors MOS (Metal Oxide Semiconductor).

The data analysis aimed to firstly highlight the part of the signal, then on the 'cleaned' data, several normalization methods were applied in order to remove unwanted effects and to standardize the responses of different runs, then PCA was applied as an unsupervised exploratory tool to understand the data structure, and finally some other chemometric methods were applied to generate a classification model, which was cross-validate by using a leave-one-out cross-validation by excluding one subject at a time.

Despite the inter- and intra-variability of the subjects' salivary composition and the analytical error of the EN, it was possible to find data cluster by using PCA according to the subjects BMI, and differentiating the obese individuals from the others (Figure 1). A classification model based on linear discriminant analysis (LDA) was also successfully employed, showing a correct classification rate (cross-validation) of 88.6%, 92.6 and 93.9% for overweight, obese and normal-weight subjects, respectively. A partial least square discriminant analysis (PLS-DA) was also built, showing potential to classify subjects based on 3 groups of data, with promising applications for screening of fat-saliva matrices simulating food consumption (Figure 1). This research could be useful to understand the different response of subjects to new food products, by using an objective instrumental tool which can provide fast responses, in addition to proper data treatment and chemometric analysis.

Figure 1. Schematic representation of the experiment and the results from EN analysis.



References

[1] Piombino P, Genovese A, Esposito S, Moio L, Cutolo PP, Chambery A, Ercolini D (**2014**). Saliva from obese individuals suppresses the release of aroma compounds from wine. PloS one, **9**(1), e85611.

[2] Genovese A, Rispoli T, Sacchi R (**2018**). Extra virgin olive oil aroma release after interaction with human saliva from individuals with different body mass index. *Journal of the Science of Food and Agriculture*, **98**(9), 3376-3383.

Masters of PCA: The Evolution of Metallica Through the Years with Exploratory Data Analysis. An Invitation to Teaching PCA in a Different Way.

Nicola Cavallini[1], **Jose Manuel Amigo**[2,3]

[1] Polytechnic of Turin, Department of Applied Science and Technology, Corso Duca degli Abruzzi 24, Turin, 10129, Italy
[2] Ikerbasque, Basque Foundation for Sciences, María Díaz de Haro, 3, Bilbao, 48013, Spain
[3] University of the Basque Country, Department of Analytical Chemistry, Barrio Sarriena S/N, Leioa, 48940, Spain
E-mail: nicola.cavallini@polito.it

Metallica is one of the cornerstones of metal rock music, and the go-to band for starting to explore this genre [1]. It is a band central to any discourse around the multi-faced topic of what is real rock and metal music, but also a very successful powerhouse of musical production and live performances.

Principal Component Analysis (PCA, [2]) is, arguably, the cornerstone of chemometrics, and the go-to method for exploratory data analysis. It is a technique central to any course in multivariate data analysis, but also a very efficient tool for compressing data, to the point that many people, especially those from the many branches of informatics, only use it as a compression tool, thus not fully exploiting its potential.

In our study we want to investigate the meeting point between Metallica and PCA, to answer a very common research question about the former: “Has Metallica evolved through the years in their studio albums attending to strict music indicators?”. This debate has been always a central point discussion between supporters and critics of the band. The approach considers a careful identification and selection of the data that can be used to describe the musical production (not the recording quality) of the band throughout their studio albums, coupled with the power of PCA, which is simply used to model and extract information from such dataset.

The parallel aim of the study, and the crucial message behind this research, is to convey the idea behind PCA in a clear and fresh way, making its use, potentiality and functioning as clear as possible to both the novice (who might be struggling with grasping the full power of this technique) and the expert user (who might be sick and tired of always seeing the usual chemical dataset used for explaining PCA). With the idea of eventually shaping it into a tutorial, the whole data analysis pipeline will be described together with tricks and suggestions on how to structure and conduct it in potentially any other real-world scenario, showing the strengths and weaknesses of the data collection procedure, the database construction, the application of PCA and the interpretation of the results.

Disclaimer: Why Metallica? That's an easy one. Because we love it!

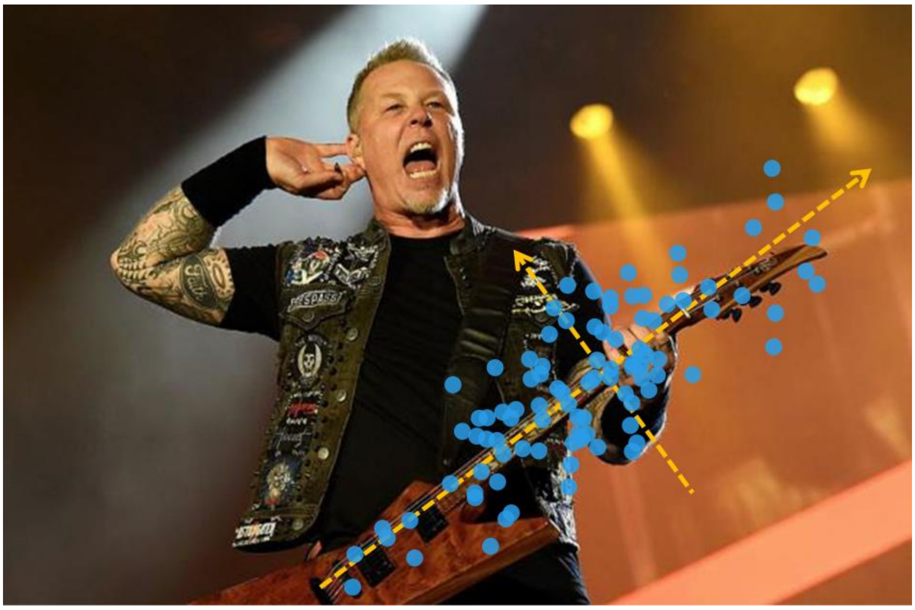


Figure 1. James Hetfield estimating the sources of maximum variance.

References

[1] Lawson D (**2016**) Metallica – the band who reinvented metal, *The Guardian Music Blog* (https://www.theguardian.com/music/musicblog/2016/apr/19/metallica-the-band-who-reinvented-metal)

[2] Hotelling H (**1933**) Analysis of a complex of statistical variables into principal components, *Journal of Educational Psychology*, **24**, 417–441.

Application of Variable-selection ANOVA Simultaneous Component Analysis in a longitudinal study on intrauterine growth restriction

Polushkina, O. A., Ferrer, P., Fernandez-Gonzalez S., Pérez-Cruz M., Gómez-Roig M.D., Camacho, J., Gómez-Llorente, C.

Addresses: Signal Theory, Networking and Communications Department, University of Granada, Granada 18014, Spain.
E-mail: josecamacho@ugr.es

Intrauterine growth restriction (IUGR) is defined as a reduction in an expected fetal growth pattern and is associated with poor perinatal and long-term postnatal outcomes (Sharma et al., 2016). Perinatal mortality rate has been decreased thanks to the prenatal identification of small fetuses for gestational age (SGA), however, a large proportion of fetuses born SGA are not diagnosed during the pregnancy despite performing third trimester ultrasounds (Figueras et al., 2008).

In this work, we use the recently developed Variable-selection ANOVA Simultaneous Component Analysis (VASCA, Camacho et al. (2023)) to infer potential biomarkers in a longitudinal study in IUGR. Variables of different classes and its possible relations, such as complete blood count, hemodynamics, immune biomarkers and fecal bacteria (metagenomics), in different stages of the pregnancy and fecal samples of the baby will be include in the model. VASCA is a multivariate alternative to the False Discovery Rate by Benjamini and Hochberg (1995), which can improve the statistical power in multivariate studies, leading to a better identification of biomarkers.

References

Sharma, D., Shastri, S., & Sharma, P. (2016). Intrauterine Growth Restriction: Antenatal and Postnatal Aspects. Clinical medicine insights. Pediatrics, 10, 67–83. https://doi.org/10.4137/CMPed.S40070

Figueras F, Eixarch E, Gratacos E, Gardosi J. Predictiveness of antenatal umbilical artery Doppler for adverse pregnancy outcome in small-for-gestational-age babies according to customised birthweight centiles: population-based study. Bjog 2008; 115(5):590-594

Camacho J, Vitale R, Morales-Jiménez D, Gómez-Llorente C. Variable-selection ANOVA Simultaneous Component Analysis (VASCA). Bioinformatics. 2023 Jan 1;39(1):btac795. doi: 10.1093/bioinformatics/btac795.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. B (Methodological), 57, 289–300.

Evaluation of Local approaches for processing Near-Infrared Spectroscopy for prediction of multi-products database

Xueping Yang [1], Paolo Berzaghi [1,2*], Matthieu Lesnoff [3], Pierre Dardenne [4], Alessandro Ferragina [5]

Addresses: [1] University of Padova, Department of Animal Medicine, Production and Health, Legnaro, Padova, 35020, Italy; [2] GraiNit s.r.l., Padova, 35129, Italy; [3] CIRAD, UMR SELMET, Montpellier, France; [4] Walloon Agricultural Research Centre CRA-W, Belgium; [5] Teagasc Food Research Centre, Ashtown, Dublin 15. D15 KN3K, Ireland
E-mail (paolo.berzaghi@unipd.it)

Near infrared (NIR) spectroscopy has been widely used in food, feed and other fields for rapid analysis nondestructive analysis. Normally, PLS calibration are developed specifically for each product (Global) and accuracy depends also on the correct choice of calibrations. A different approach (Shenk et al. 1998) is to build a multiproduct database and have LOCAL calibrations built on the fly for each sample being predicted. In practice with LOCAL calibrations the user doesn't have to select a specific calibration, but the algorithm selects the best set of samples to be used for predicting each unknown sample. The study evaluated Global, Shenk's LOCAL and a new locally weighted method based on partial least squared scores (LWPLSR; Lesnoff et al. 2020) recently published, comparing them to the traditional Global calibrations. A total of 3379 samples including forages (5 products n = 2263) and grains (4 products n = 1116) were used, with selection of a test set (n = 340) using "Kennard and Stone" algorithm. Global calibrations were developed for the 9 different products and also for the entire multiproduct database, while local methods always used the full database. Calibrations were developed after spectral math pretreatments, Standard Normal Variate and Detrend as well as Savitzky-Golay (D=1, Smooth 7, Pol. Ord 2). PLS prediction models were developed and optimized by cross validation and, prediction were evaluated by Root Mean Standard Error of Prediction (RMSEP) All computations for calibration, validation were developed using R Version 1.3.1093.

Table 1 shows the RMSEP values, for Global the single products were predicted with specific calibrations, while the entire testing set was predicted with the one multiproduct calibration. In general local methods provided more accurate predictions for the full testing set but also for the specific products than Global calibrations. The differences between global and local were smaller for DM, than Protein and ADF. Reduction of RMSEP of Local over Global for the different products was about 20-25% for Protein, 24-36% for ADF and 0-18% for DM. Between locals, LWPLSR is the method that yielded the best performances. Among the products there was a very small set (ALP n = 27) of alfalfa byproducts, which had the largest differences between Global and local for Protein and ADF. These constituents are more influenced by the matrix of the sample and having a large global dataset ensure fair accuracy. On the other hand for DM which has stronger spectral signal global calibration work as good as local ones. LOCAL algorithm was much faster (few seconds per sample) than LWPLSR run under R package ("Rchemo"), which should be improved for practical use.

Table 1. RMSEP values obtained by Global calibration and for each Local method.

Ingredient	Testing set (No.)	RMSEP of different Models		
		Global	**Local Algorithm**	**LWPLSR**
Protein (%)	All (339)	0.73	0.59	0.57
	CSL (47)	0.55	0.40	0.41
	HAY (121)	0.91	0.78	0.75
	ALP (3)	0.99	0.46	0.54
	SGS (28)	0.99	0.69	0.52
	TMR (28)	0.43	0.40	0.36
	Barley (13)	0.55	0.34	0.48
	Oat (16)	0.62	0.52	0.60
	Corn (48)	0.34	0.36	0.39
	Wheat (35)	0.62	0.37	0.40
ADF (%)	All (331)	1.44	1.31	1.03
	CSL (47)	1.35	0.95	0.76
	HAY (121)	1.49	1.30	1.25
	ALP (3)	1.70	0.92	0.65
	SGS (28)	2.05	2.54	1.40
	TMR (27)	1.90	1.88	1.13
	Barley (13)	1.35	0.85	0.86
	Oat (16)	1.54	1.05	1.10
	Corn (44)	0.80	0.46	0.44
	Wheat (32)	0.71	0.40	0.54
DM (%)	All (332)	0.72	0.63	0.50
	CSL (47)	0.92	0.73	0.46
	HAY (121)	0.72	0.68	0.60

	ALP (3)	0.38	0.78	0.42
	SGS (28)	0.79	0.72	0.40
	TMR (28)	0.90	0.85	0.69
	Barley (13)	0.25	0.32	0.41
	Oat (16)	0.38	0.35	0.26
	Corn (41)	0.54	0.40	0.30
	Wheat (35)	0.45	0.39	0.34

References

J.S.Shenk, P.Berzaghi and M. O. Westerhaus, J. Near Infrared Spectrosc. 5, 223-232 (1998).

Lesnoff, M., Metz, M., & Roger, J. M. (2020). Comparison of locally weighted PLS strategies for regression and discrimination on agronomic NIR data. *Journal of Chemometrics*, *34*(5), 10–12.

Chemometrics applied to the study of non-canonical DNA structures

Raimundo Gargallo

Addresses: University of Barcelona, Department of Chemical Engineering and Analytical Chemistry, Marti I Franques 1-11, Barcelona, E-08028, Spain
raimon_gargallo@ub.edu

Apart from the well-known double-stranded structure, DNA may form other characteristic, non-canonical, structures, such as G-quadruplex, i-motif, triplex DNA, … Recent studies on human genome have discovered DNA regions where these structures may be formed *in vivo*. Also, the i-motif structure has been shown to be a potential tool for the development of pH sensors in chemical analysis [Alba, 2016].

The *in vitro* study of the chemical and conformational equilibria, as well as the related kinetics aspects, of these structures is mainly carried out by using spectroscopic techniques, such as molecular absorption, circular dichroism, molecular fluorescence, or nuclear magnetic resonance. In this scenario, Chemometrics has been shown to be a useful in analyzing multivariate data recorded along the monitoring of these conformational changes or chemical equilibria.

In this presentation, a brief introduction to some of these structures will be given, paving the way to the explanation of many examples of application of Chemometrics (hard and soft modelling, basically) to data sets of different nature.

References

Alba J, Sadurni A, Gargallo R (**2016**) Nucleic Acid i-motif structures in Analytical Chemistry. *Critical Reviews in Analytical Chemistry* **46**, 443-454

An instrumental data fusion technique based on Self-Organizing Map algorithm for contextualizing seawater sampling

Sabina Licen[1], **Anna Bressan**[1], **Vinko Bandelj**[2], **Pierluigi Barbieri**[1]

[1]Dept. of Chemical and Pharmaceutical Sciences, University of Trieste, Trieste , Italy;
[2]National Institute of Oceanography and Experimental Geophysics-OGS, Trieste, Italy;
E-mail slicen@units.it (corresponding author)

In the last decades the enforcement of national and international regulations on water quality, such as the Water Framework Directive (WFD - 2000/60/EC), the Marine Strategy Framework Directive (MSFD - 2008/56/EC), and the Maritime Spatial Planning Directive (MSPD - 2014/89/EU), obliges EU Member States to regularly and extensively monitor marine ecosystems. Several water quality parameters can be monitored: physical properties, chlorophyll content, nutrients, toxic elements, organic pollutants, etc. The techniques can be applied in situ, ex situ (lab-based measurement), or by means of remote sensing. Huang et al. (Huang et al., 2022), in a very recent critical review, observed that currently, short-term and discrete monitoring coupled with ex situ lab-based measurement are still the mainstream in water quality measurement. In this way the information obtained is sparse along time dimension, thus there is a need of combining the high-dimensional data obtained by the different sampling methods using data fusion techniques and machine learning models. The aim of the study is to use a data fusion technique to contextualize water sampling analysis results merging them with data collected continuously, at different time spans by different instruments, at a seawater sampling site. In brief, long-term time series of data (temperature, pH, dissolved oxygen, etc...) collected at a sampling site in the Gulf of Trieste (Italy) in the past has been organized and analyzed by Self-Organizing Map (SOM) algorithm (Kohonen, 2013) for identifying seawater typical dynamic states. Among the unsupervised data analysis methods Clark et al. (Clark et al., 2020) published a review describing the use and advantages of elaborating water quality data using SOM algorithm. It proved to be a useful tool for investigating spatial and temporal patterns in the data. Moreover Li et al. (Li et al., 2020) proved that the SOM is effective in marine data clustering. The SOM outcome is a matrix containing recurrent profiles that can be considered recurrent "states" of the environmental compartment. The profiles can be easily characterized by the modeled values (weight of nodes) of the monitored parameters. The output matrix can be further analyzed by clustering algorithms such as hierarchical (HCA) or k-means (KM) clustering for obtaining few recurrent "macro-states" of the environmental compartment (Solidoro et al., 2009; Vesanto and Alhoniemi, 2000). The results obtained has been integrated by a low-level data fusion technique with results of sparse seawater sampling carried out at the same site for ex-situ lab analysis of nutrients and pollutants. The sparse samplings have been contextualized in terms of "macro-states" occurring before and during the sampling time to allow the characterization of the "macro-states" in terms of nutrient and pollutant profiles and, at the same time, to allow comparison of sampling results carried out during the same period in different years.

References

Clark, S., Sisson, S.A., Sharma, A., 2020. Tools for enhancing the application of self-organizing maps in water resources research and engineering. Adv. Water Resour. https://doi.org/10.1016/j.advwatres.2020.103676

Huang, Y., Wang, X., Xiang, W., Wang, T., Otis, C., Sarge, L., Lei, Y., Li, B., 2022. Forward-Looking Roadmaps for Long-Term Continuous Water Quality Monitoring: Bottlenecks, Innovations, and Prospects in a Critical Review. Environ. Sci. Technol. 56, 5334–5354. https://doi.org/10.1021/ACS.EST.1C07857

Kohonen, T., 2013. Essentials of the self-organizing map. Neural Networks 37, 52–65. https://doi.org/10.1016/J.NEUNET.2012.09.018

Li, Y., Yang, Z., Han, K., 2020. Research on the clustering algorithm of ocean big data based on self-organizing neural network. Comput. Intell. 36, 1609–1620. https://doi.org/10.1111/COIN.12299

Solidoro, C., Bastianini, M., Bandelj, V., Codermatz, R., Cossarini, G., Canu, D.M., Ravagnan, E., Salon, S., Trevisani, S., 2009. Current state, scales of variability, and trends of biogeochemical properties in the northern Adriatic Sea. J. Geophys. Res. Ocean. 114, 7–91. https://doi.org/10.1029/2008JC004838

Vesanto, J., Alhoniemi, E., 2000. Clustering of the self-organizing map. IEEE Trans. Neural Networks 11, 586–600. https://doi.org/10.1109/72.846731

DEVELOPMENT AND OPTIMIZATION OF AN ANALYTICAL METHOD BASED ON SPME-ARROW AND CHEMOMETRIC TECHNIQUES FOR THE CHARACTERIZATION OF THE AROMA PROFILE OF BAKERY PRODUCTS

Samuele Pellacani, Marina Cocchi, Lorenzo Strani, Caterina Durante

University of Modena and Reggio Emilia, Dept. Chemical and Geological Sciences, via Campi 103, Modena, 41125, Italy
samuele.pellacani@unimore.it

In the baking industry, a precise knowledge of aroma and its influencing parameters is crucial for the formulation of new recipes and meeting consumer needs.

In this study, an innovative method based on Arrow-SPME-GC-MS was developed with the aim of chemometrics for analyzing the flavor of four industrial bread samples and examining the influence of different yeasts (sourdough, beer's yeast, and a mixture of both) on volatile organic compounds (VOCs) produced in the final product. Since the aroma profile was sampled using a novel Arrow-SPME fiber [1], the factors affecting its adsorption and absorption mechanisms were optimized by means of Design of Experiment (DoE) technique, allowing for the rational planning of experiments to optimize specific properties of the system under investigation [2]. For this purpose, a Face-Centered Central Composite Design (FC-CCD) was exploited.

As GC-MS (hyphenate techniques in general) signals analysis requires high computational capabilities as well as efficient methods to reduce data dimensionality and to extract useful information, the novel multi-way data analysis approach, PARADISe [4] (PARAFAC2-based deconvolution and identification system [5]), was used to deconvolute the chromatographic peaks from raw chromatographic data and integrate the areas of the deconvoluted peaks for all samples simultaneously. The PARADISe approach was particularly effective in revealing co-eluted compounds that would otherwise be difficult to detect by mathematically decomposing the chromatogram. Principal Component Analysis (PCA) was used as an exploratory analysis to evaluate the effects of the investigated variables, yeasts, and leavening times, on the aroma profiles of the bread samples. The identification of the compounds that most differentiate the various samples from each other helped to explain the differences that emerged on the basis of the used recipe.

References

[1] Herrington, J.; Gómez-Ríos, G.; Myers, C.; Stidsen, G.; Bell, D. (**2020)** Hunting Molecules in Complex Matrices with SPME Arrows: A Review. *Separations*, **7**, 12,

[2] R. Leardi, (**2009**) Experimental design in chemistry: A tutorial, *Analytica Chimica Acta*. **652** 161–172.

[3] Lundstedt T., Seifert E., Abramo L., Thelin B., Nystrom A., Pettersen A. & Bergman R. (**1998**) Experimental design and optimization. *Chemometrics and Intelligent Laboratory Systems* **42**, 3-40.

[4] Amigo, J.M.; Skov, T.; Bro, R.; Coello, J.; Maspoch, S. (**2008**) Solving GC-MS Problems with PARAFAC2. *TrAC Trends in Analytical Chemistry*, **27**, 714–725.

[5] G. Baccolo, B. Quintanilla-Casas, S. Vichi, D. Augustijn, R. Bro, (**2021**) From untargeted chemical profiling to peak tables – A fully automated AI driven approach to untargeted GC-MS. *TrAC Trends in Analytical Chemistry* **145**, 116451.

Feasibility of a near-infrared spectroscopy system to estimate the microbial load of seafood products

Gonzato Elia[a], Currò Sarah[a], Facco Pierantonio[b], Novelli Enrico[a], Balzan Stefania[a], Fasolato Luca[a]

[a]Department of Comparative Biomedicine and Food Science, University of Padova Legnaro, Italy
[b]CAPE-Lab, Computer-Aided Process Engineering Laboratory, Department of Industrial Engineering, University of Padova, via Marzolo 9, Padova (Italy)
sarah.curro@unipd.it

The loss of freshness of fish and fishery products is mainly due to microbial spoilage. Particularly, Specific Spoilage Organisms (SSO), are microorganisms which, during storage, grow faster than the others, causing the formation of metabolites responsible for off odors and off-flavors, leading to sensory rejection of the product. Many of these compounds such as alcohols, aldehydes, ketones, organic acids, sulphides, esters and amines were applied to definine the spoilage status of seafood [1]. However, the conventional technique of microbial quality charaterization imply time, economic resources and skilled personnel, therfore the development of a rapid alternative method as IR spectroscopy is required. The application of IR spectroscopy to microbial load prediction should carefully consider many different confounders (i.e. physical status, seasonality, conservation temperature, etc.) that could affect bacterial and spectral features. Moreover, chemometric pre-processing approaches and features selection should be carefully evaluated to define the capability of IR spectroscopy [2].

The aim of this study is to evaluate how spectroscopic data can be used to predict microbial load of *Pseudomonas spp.* as principal indicator among other SSO in seafood products taking in consideration factors affecting prediction performances. In detail, a total of 81 cuttlefish samples (fresh, frozen-thawed, and frozen-thawed treated with peroxide hydrogen) were collected from a wholesale fish plant, and then stored at 0, 2 and 4 °C until rejection of sensory analysis. At 0, 3, 6, 9 and 12 days of storage, samples were evaluated through microbiological, sensory and spectroscopic analyis by portable NIR tools obtaining a total of 348 observations. Sensorial and microbiological analyses were combined to each respective spectrum to perform principal component analysis

(PCA) and Partial Least Square (PLS) regression as exploratory and as predictive approaches, respectively.

In particular, after proper data pretreaatment (i.e., standard normal variate, baseline removal, smoothing and second derivative) a PCA model was built to highlight the importance of factors such as physical status, geographic origin, month of fishing, treatment and days of storage. Results highlight *Pseudomonas spp.* are clustered with sensory variables, whereas batches clustered with treatment meaning that those effects affect the spectral features.

Furthermore, PLS was used to predict the load of *Pseudomonas spp.* from the sensorial and NIR spectroscopic data. With eight component (chosen according to the one-sigma heuristic method [3]), 85% and 75% of variance of the **X** and **Y** blocks, respectively, was captured. This ensures a good predictive ability with a Root Mean Squared Error Prediction RMSEP = 0.76 and R^2 = 0.75, corresponding to a relative mean error of 15%; whereas a R^2cv of 0.62 and RMSEP of 0.93 were reported in cross-validation. Figure 1a, b shows calibration and cross-validation predictions obtained from the described model. In conclusion, spectroscopic data in association with machine learning techniques appear to be a promising tool for the prediction of fish product adulteration and spoilage resulting from microbial proliferation [4].

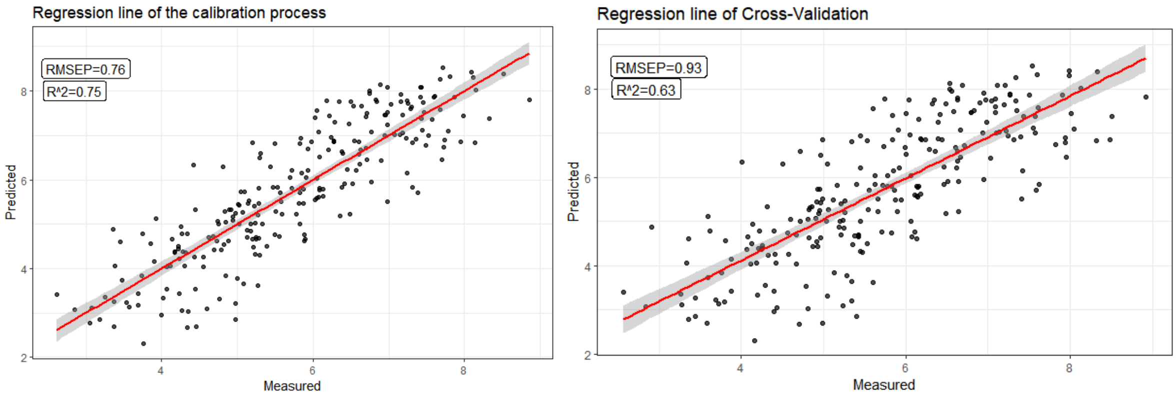


Figure 1a, b: Fitted values of the load of Pseudomonas spp. by means of Partial Least Squares Regression in calibration (left) and in cross-validation (right).

Testing set shows interestingly that predictions are, in average, more precise when microbial load is around 6 Log CFU/g, near the deterioration threshold for seafood products. At 7 Log CFU/g *Pseudomonas spp.* modify the food substrate (mainly proteolytic activity) producing different spoilage compounds that change the spectral features [5].

Furthermore, this solution represent an intertesting support to overcome spoilage level identification to reduce the number of microbiological and time of analysis (few minutes for NIR spectroscopy vs. 48/72 h of microbial analysis), reducing the cost of lab analysis on the economy of companies that operate directly in the seafood market by portable spectroscpy tools.

References

[1] Parlapani FF, Anagnostopoulos DA, Karamani E, Mallouchos A, Haroutounian SA, Boziaris IS (2023) Growth and Volatile Organic Compound Production of Pseudomonas Fish Spoiler Strains on Fish Juice Agar Model Substrate at Different Temperatures. *Microorganisms*, 11, 189.

[2] Mishra P, Ruledge DN, Roger JM, Wali K, Khan HA,(2021) Chemometric pre-processing can negatively affect the performance of near-infrared spectroscopy models for fruit quality prediction, *Talanta*, 229, 122303.

[3] Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning: data mining, inference, and prediction. 2nd Edn. *Springer Science & Business Media*.

[4] Fengou LC, Lianou A, Tsakanikas P, Gkana EN, Panagou E, Nychas GJ (2019) Evaluation of Fourier transform infrared spectroscopy and multispectral imaging as means of estimating the microbiological spoilage of farmed sea bream, *Food Microbiology*, 27-34,

[5] Govari M, Tryfinopoulou P, Parlapani FF, Boziaris IS, Panagou EZ, Nychas GJE (2021) Quest of Intelligent Research Tools for Rapid Evaluation of Fish Quality: FTIR Spectroscopy and Multispectral Imaging Versus Microbiological Analysis. *Foods*, 10, 264.

Chemometrics strategies for the implementation of surface enhanced Raman scattering of crevicular fluid in periodontics and implantology

Stefano Fornasaro[1], Antonio Rapani[2], Federico Berton[2], Claudio Stacchi[2], Valter Sergo[3], Alois Bonifacio[3]

[1]University of Trieste, Dept. of Chemical and Pharmaceutical Sciences, via L. Giorgeri 1, 34127 Trieste, Italy

[2]Maxillofacial and Dental Surgical Clinic, Dept. of Medical, Surgical and Health Sciences, University of Trieste, Piazza dell'Ospitale 1, 34125, Trieste, Italy

[3]Raman Spectroscopy Lab, Dept. of Engineering and Architecture, University of Trieste, via A. Valerio 6a, 34127 Trieste, Italy

sfornasaro@units.it (corresponding author)

Periodontal disease is one of the most frequent diseases affecting dental elements, with periods of disease activity followed by periods of remission. Traditional clinical and radiological methods only capture pre-existing destruction rather than current disease activity, and metabolomic analysis of gingival crevicular fluid (GCF) has attracted attention as a valuable tool for periodontal disease prognosis and diagnosis.

Surface enhanced Raman scattering (SERS) is a spectroscopic technique that has various advantages over standard bioanalysis for the analysis of biological samples ranging from *in vitr*o cell culture models to *ex vivo* tissues and biofluids. SERS datasets derived from biofluids are rich in metabolic fingerprint information, but access to this information is not always simple. Bioanalytical SERS is a highly complicated field that necessitates a thorough understanding of the chemical and physical interactions that exist between photons, nanomaterials, and biological systems. Discovering meaningful information in the midst of this complication is a difficult task. Chemometrics provides extremely useful tools to assist us in this effort.

Several strategies for dealing with the design, processing, and interpretation of SERS spectrum data from biofluids will be briefly discussed in this communication. In particular, a successful thorough deployment of chemometrics methodologies (from the design of experiments to multivariate analysis and data fusion) to relate the SERS metabolic fingerprint of GCF to the periodontal health of both teeth and dental implants will be presented.

References

Fornasaro S, Berton F, Stacchi C, Farina F, Esposito A, Sergo V, Di Lenarda R, Bonifacio A. (**2021**) Label-free analysis of gingival crevicular fluid (GCF) by surface enhanced Raman scattering (SERS). *Analyst*. **146**(4):1464–1471.

Fornasaro S, Beleites C, Sergo V, Bonifacio A (**2022**) Data analysis in SERS diagnostics, in *SERS for Point-Of-care and Clinical Applications*, ed.: Andrew Fales, Elsevier, 1-51.

A new chemometric method for oral and oropharyngeal cancer detection based on multivariate analysis of Surface Enhanced Raman spectra of salivary exosomes

Valentin Toma[1], Cosmin Ioan Faur[2], Gabriela Fabiola Știufiuc[3], Radu Mărginean[1], Anca Onaciu[1], Carina Culic[4], Brandusa Dragoi[5,6], Rares Ionuț Știufiuc[1,5]

1) MedFuture- Research Center for Advanced Medicine, "Iuliu Hatieganu" University of Medicine and Pharmacy, Cluj-Napoca, Romania; valentin.toma@umfcluj.ro; margi.radu@outlook.com; anca.onaciu@umfcluj.ro; rares.stiufiuc@umfcluj.ro

2) Department of Oral Radiology," Iuliu Hatieganu" University of Medicine and Pharmacy, Cluj-Napoca, Romania; cosmin.faur@umfcluj.ro ; mhedesiu@umfcluj.ro

3) Faculty of Physics, "Babes - Bolyai" University, Cluj-Napoca, Romania; gabriela.stiufiuc@ubbcluj.ro

4) Department of Odontology, Endodontics, Oral Pathology, Faculty of Dentistry, "Iuliu Hațieganu" University of Medicine and Pharmacy, Cluj-Napoca, Romania; carina.culic@umfcluj.ro

5) Regional Institute of Oncology, 2-4 General Henri Mathias Berthelot, 700483 Iasi, Romania,

6) Faculty of Chemistry, "Alexandru Ioan Cuza" University of Iasi, 11 Carol I Blvd., 700506, Iasi, Romania

---

Raman spectroscopy recently proved a tremendous capacity to identify disease specific markers in various (bio)samples being a non-invasive, rapid, and reliable method for cancer detection. In this study, we firstly aimed to record vibrational spectra of salivary exosomes isolated from oral and oropharyngeal cancer patients and healthy controls using Surface Enhancement Raman Spectroscopy (SERS). Then we assessed this method's capacity to discriminate between malignant and non-malignant samples by means of Principal Component - Linear Discrimination Analysis (PC-LDA) of the SERS spectra of salivary exosomes. The vibrational spectra were collected on a solid plasmonic substrate developed in our group, synthesized using tangential flow filtered and concentrated (TFF) silver nanoparticles, capable to generate very reproducible spectra for a whole range of bioanalytes. SERS examination identified interesting

variations of the vibrational bands assigned to thiocyanate, proteins and nucleic acids between the saliva of cancer and control groups. The chemometric analysis - performed on the whole spectra as well as on specific regions - indicated discrimination sensitivity between the two groups up to 79.3%. The sensitivity is influenced by the spectral interval used for the multivariate analysis, being lower (75.9%) when the full range spectra were used.

Acknowledgement

This work was supported by a H2020 grant – ERA-Chair, no 952390 and a grant of the Ministry of Education and Research, CNCS/CCCDI - UEFISCDI, project number PN-III-P3-3.6-H2020-2020-0105/35/2021.

A Raman-Based Chemometric study for the discrimination of tomatoes from agricultural techniques

T. Forleo[1], L.C. Giannossa[1], P. Bitsi[2], A. Mangone[1]

[1]Dipartment of Chemistry, University of Bari "Aldo Moro", Bari 70126, Italy
[2]Dipartment of Chemistry, University of Ioannina, Mpizani 45500, Greek
tiziana.forleo@uniba.it

The tomato (*Solanum lycopersicum L.*) is one of the most widely cultivated and consumed fruits globally.

Nutritionally, it is rich in substances essential for human nutrition and in particular contains carotenoids like lycopene, beta-carotene, and lutein.

Moreover, it is an easily cultivated fruit, adaptable to various climates, and used in cuisines worldwide [1].

From an economic perspective, the cost of tomato production varies depending on the cultivation method. Tomatoes grown unconventionally (in greenhouses, tents, hydroponics, organic farming, etc.) tend to have higher costs because these methods require investments in facilities, equipment, climate control systems, nutrients, and so on [2].

It is estimated that approximately 5 million hectares of land worldwide were used to grow around 189 million tonnes of tomatoes in 2021, with 24 million tonnes in Europe and 6 million tonnes in Italy [3].

In this study, a total number of 201 samples were analyzed using Raman spectroscopy, and the collected spectra were investigated using Partial-Least-Square Discriminant Analysis (PLS-DA) to evaluate the ability to distinguish tomatoes from conventional and non-conventional agriculture based on their Raman spectral profiles.

The results show that PLS-DA was capable of classifying samples according to their production technology, with a 93% rate of correctly classified samples in external prediction.

This analytical methodology has proven to be an efficient screening tool that could be used in quality control and authenticity assessment to discriminate tomatoes produced by different agricultural techniques.

References

[1] Xiaoxiao Ding, Ying Guo, Yongnian Ni, Serge Kokot (**2016**) A novel NIR spectroscopic method for rapid analyses of lycopene, total acid, sugar, phenols and antioxidant activity in

dehydrated tomato samples, Vibrational Spectroscopy, **82**, 1-9,ISSN 0924-2031, https://doi.org/10.1016/j.vibspec.2015.10.004.

[2] Mattia Rapa, Salvatore Ciano, Roberto Ruggieri, Giuliana Vinci (**2021**) Bioactive compounds in cherry tomatoes (Solanum Lycopersicum var. Cerasiforme): Cultivation techniques classification by multivariate analysis, Food Chemistry, **355**, 2021,129630,ISSN 0308-8146, https://doi.org/10.1016/j.foodchem.2021.129630.

[3] FAOSTAT 2021, https://www.fao.org/faostat/en (last visit: 14/02/2023)

Improving convex-hull estimates to better explore LIBS images of biosamples by pre-processing the spectra

Victor H.C. Ferreira[1], Vincent Gardette[2], Vincent Motto-Ros[2],
Ludovic Duponchel[1]

[1]Univ. Lille, CNRS, UMR 8516 – LASIRE, Lille F-59000, France
[2]Univ. Claude Bernard Lyon 1, CNRS, UMR5306 – iLM, Villeurbanne F-69622, France
E-mail: victor.ferreira@univ-lille.fr

The interest in using hyperspectral images generated by laser-induced breakdown spectroscopy (LIBS) on bioanalysis has been increasing due to some advantages of the technique such as its fast data acquisition, the possibility to perform multielemental analyses and to detect both major and minor elements at the same time [1], which may be well suited for clinical analyses. Some recent spectrometers can even work at kHz range acquisition rates, thus a single map has hundreds of thousands of spectra. While in one hand having that much information from a sample may allow us to dig deeper into it, on the other hand the sheer amount of spectra skews the variance and hinders minor elements found in scarce pixels. That is, even though a minor element generates a signal in a given spectrum, usual variance-based multivariate exploratory tools may not find it. One way to better explore this sort of dataset is to use the convex-hull, the minimal multidimensional geometric object that contains all the dataset. The main idea is that we can project all data as a cloud in a multidimensional space based on the variables and the points of data with the most different values will stay further away from the others. The convex hull is a geometric object that connects all outermost points and, mathematically, each point inside this object is potentially a linear combination of them. We can then consider that all different source of signals will be part of the convex-hull and use that for qualitatively finding all elements. Calculating the convex-hull, however, is not simple and for most software, it is limited to lower dimensionalities. Our group has proposed an estimation of the convex hull named Interesting features finder (IFF) [2], which attempts to find the outermost parts by using thousands of random vectors and inspecting the frequency in which the points are assumed as being the outermost ones. In the analysis of a LIBS image from a miner's lung tissue, IFF managed to identify elements such as Ni and Sn found on less than 12 pixels out of the 245100 pixels on the image. Furthermore, once we Pareto-scaled the spectra before applying IFF, we could find peaks attributed to Au and Ag. These findings are important given that some elements might prove to be harmful to a patient even if scarcely found on the tissue. As such, using pre-processing steps before IFF seems to improve the estimation of the convex-hull and its use as an exploratory tool for LIBS images on a clinical environment.

References

[1] Gardette V, Motto-Ros V, Alvarez-Lllamas C, Sancey S, Duponchel L, Busser B (**2023**) Laser-Induced Breakdown Spectroscopy Imaging for Material and Biomedical Applications: Recent Advances and Future Perspectives. *Analytical Chemistry*, **95**, 49-69

[2] Wu Q, Marina-Montes C, Cáceres JO, Anzano J, Motto-Ros V, Duponchel L (**2022**) Interesting features finder (IFF): Another way to explore spectroscopic imaging data sets giving minor compounds and traces a chance to express themselves. *Spectrochimica Acta Part B: Atomic Spectroscopy*, **195**, 106508

A PROOF-OF-CONCEPT STUDY ON THE LOGICAL ANALYSIS OF SAMPLE POOLING RESULTS FOR QUALITATIVE ANALYTICAL TESTING

L. A. Sarabia[1], O. Valencia[1], M. C. Ortiz[2]

Universidad de Burgos, Departamento de Matemáticas y Computación[1],
Universidad de Burgos, Departamento de Química[2],
Plaza. Misael Bañuelos, s/n, Burgos, 09001, Spain
E-mail (lsarabia@ubu.es)

Pooled testing is the process of testing samples together as a combined group, rather than individually, to identify those samples with a binary trait of interest, positive samples, especially when the prevalence in the whole population is low. Since the pioneer work of Dorfman [1], the pooling of samples has been widely used in infectious diseases in medical as well as veterinarian applications. The recent SARS-CoV-2 pandemic, where rapid diagnostic and screening testing through PCR were imperative, has driven numerous sample pooling strategies [2, 3], some of which have received regulatory approval.

Within the field of chemometrics, Cela [4] has developed a different approach to obtain the pooled samples in the analytical laboratory, based on supersaturated (SS) designs of experiments.

In this work, we suggest firstly, a procedure to make the pooled samples and secondly, a logical, not numerical, algorithm to identify the positive samples from the outcomes of the analysis of the pooled samples.

For the former, the 'half' of a Plackett-Burman design is used that includes the mixture of all individual samples. For the latter, an algorithm built from the logical structure of the matrix of the design of the mixtures and the experimental response (positive/negative) for each of them determines which individual samples cannot be positive. To unequivocally identify the positive samples, a hierarchical three-stage method is applied: the first stage consists of pooling all samples, the second one applies the proposed algorithm and, if necessary, a third stage is carried out for individual identification.

As a proof of concept, 10 individual samples have been considered to build the pooled samples. In this case, if the prevalence of positive equals 0.05, the expected average number of trials is 2.552, lower than 5.987, the best expected value found in the literature for a hierarchical method.

The procedure is applied to the detection of food samples contaminated with a pathogen (*Salmonella*) and others contaminated with an allergen (*nut*) by means of polymerase chain reaction, PCR, test. The *Salmonella* is regulated by Commission Regulation (EC) No 2073/2005 of 15 November 2005 on microbiological criteria for foodstuffs, whereas the presence of *nut* by Regulation (EU) No 1169/2011 of the European Parliament and of the Council of 25 October 2011 on food information provided to consumers. All contaminated samples have been correctly identified.

Acknowledgment

This work is part of the project with reference BU052P20 financed by Junta de Castilla y León with the aid of European Regional Development Funds.

References

[1] Dorfman R (**1943**) The Detection of Defective Members of Large Populations. *The Annals of Mathematical Statistics*, **14(4)**, 436–440.

[2] Crone M, Randell P, Herm Z, *et al.* (**2021)** Rapid design and implementation of an adaptive pooling workflow for SARS-CoV-2 testing in an NHS diagnostic laboratory: a proof-of-concept study. *Wellcome Open Research,* **6**:268.

[3] Hao-Ru L, Fang-Yu W, Jia-Ling L, Teh-Ying C, Hsiang-Ling H (**2022**) Real-world Evaluation of a Sample Pooling Strategy for Large-Scale Rapid COVID-19 Testing. *Journal of Clinical Virology,* **149**, 105133.

[4] Cela R, Claeys-Bruno M, Phan-Tan Luu R (**2009**) Screening Strategies In: Brown S, Tauler R, Walczak R (eds.) *Comprehensive Chemometrics* **1,** 251–300, Oxford, Elsevier.

DEEP LEARNING COMBINED WITH EXPERIMENTAL DESIGN TO LIMIT DEPENDENCE ON ANNOTATED DATA

Magalie Claeys-Bruno[2], Maxime Metz[1], Nicolas Grotus[1], Michelle Sergent[2]

[1] Pellenc Selective Technologie, 125 Rue Francois Gernelle, BP124, 84120 Pertuis, France
E-mail : m.metz@pellencst.com
[2] Institut Méditerranéen de Biodiversité et d'Ecologie marine et continentale, Aix-Marseille Université, UMR CNRS IRD Avignon Université, Site de l'Etoile, Marseille, France
E-mail : m.claeys-bruno@univ-amu.fr

Today, deep learning approaches [1] are increasingly used to address industrial and scientific issues. Indeed, for several years, deep learning methods have been increasingly integrated for various applications of spectroscopic data processing [2] due to the development of open-source tools such as Pytorch, Tensorflow or Keras. These tools allow the use of deep learning models that are efficient both in terms of inference time and predictive performance. Nevertheless, these supervised approaches require a large volume of annotated data. In chemometrics, deep learning approaches have already been used to process spectroscopic data. However, approaches such as partial least squares (PLS) regression [3] or local PLS [4] are still widely used because they are very data-efficient. In the case of chemical data we also observe a high cost of annotating the data. Indeed, it is sometimes necessary to carry out an in-depth spectral study to characterise a sample or even to associate a reference measurement (protein content, sugar, etc.) with a spectral measurement. This annotation cost often limits chemometricians to using deep learning approaches.

To enable chemometricians to use deep learning approaches on chemical data one solution is to reduce the number of data to be annotated. In order to limit this decrease in performance, sampling methods have been developed for spectroscopic data processing [5]. Optimal experimental design (or active learning) makes it possible to select a set of samples to be annotated that is as relevant as possible for the model calibration. However, as the field of deep learning combined with optimal experimental design is still under-researched, very few studies have been developed.

In this study, we propose to test different optimal experimental design methods developed which allow the selection of representative samples from a database. More precisely, we propose to integrate Space Filling Design algorithms [6] to make them more appropriate for chemical data processing.

References

1. LeCun Y, Bengio Y, Hinton G, (**2015**) Deep learning, *Nature*, **521**, 7553, 436-444,
2. M. Said, A. Wahba, et D. Khalil, (**2022**) Semi-supervised deep learning framework for milk analysis using NIR spectrometers , *Chemometrics and Intelligent Laboratory Systems*, **228**, 104619
3. Wold S *et al*, (**2001**) PLS-regression: a basic tool of chemometrics, *Chemometrics and Intelligent Laboratory Systems*, **58**, 2, 109–130.
4. Centner V. et Massart D. L., (**1998**) Optimization in Locally Weighted Regression, *Anal. Chem.*,vol. **70**, 19, 4206–4211.
5. Douak F, Melgani F, Alajlan N, Pasolli E, Bazi Y, Benoudjit N, (**2012**) Active learning for spectroscopic data regression: Active learning for spectroscopic data regression, *Journal of Chemometrics,* **26**, 7, 374-383.
6. Santiago J, Claeys-Bruno M., Sergent M, (**2012**) Construction of space filling designs using WSP algorithm for high dimensional spaces. *Chemometrics and Intelligent Laboratory System*, **113**, 26-31.

PREDICTION OF SUPERCRITICAL CO2 EXTRACTION KINETICS OF COMPOUNDS OF INTEREST USING EXPERIMENTAL DESIGN

Magalie Claeys-Bruno[1], **Adil Mouahid**[2], **Sébastien Clercq**[2], **Michelle Sergent** [1]

[1] Institut Méditerranéen de Biodiversité et d'Ecologie marine et continentale, Aix-Marseille Université, UMR CNRS IRD Avignon Université, Site de l'Etoile, Marseille, France

E-mail : m.claeys-bruno@univ-amu.fr

[2] Aix Marseille Université, CNRS, Centrale Marseille, M2P2, Marseille, France

---

Molecules such as lipids, antioxidants, anti-inflammatory, contained in natural matrices such as seeds, plants, fruits or microalgae represent compounds of interest for the food, cosmetic and pharmaceutical industries. The recovery of these molecules is achieved by conventional extraction methods involving toxic organic solvents or by hydro distillation. Sustainable development and environmental protection are pushing companies to move towards green and efficient technologies as extraction using supercritical CO2 (CO2-SC). This clean process allows a selective extraction of weakly polar compounds with a very high efficiency, with low temperatures and extraction times around 3 hours compared to 8 to 12 hours for classical techniques. At the end of the process, the CO2 becomes gaseous again and is totally separated from the extract, thus avoiding separation steps. Within the framework of an industrial scale development, the determination of the autoclave volumes, and the choice of the optimal operating conditions are made using the knowledge of the solubilities of the extracts in the CO2-SC, of the transfer and diffusion coefficients, the optimal duration of extraction and the maximum quantity that can be extracted [1,2]. These parameters are accessible via the modeling of experimental extraction kinetics curves obtained, most often, from laboratory scale studies. Many experimental and modeling studies are presented in the literature, however few simulation studies have been performed. The results are specific to the studied matrix and the parameters useful for industrial development are not considered. That's the reason why, we propose the development of a versatile predictive tool of the extraction kinetics curves by CO2-SC for a wide range of solid matrices. This simulation tool would make it possible to get rid of the long and costly experiments or at least to limit the number of experiments to be carried out. For this purpose, the methodology consists in establishing a database as exhaustive as possible, differentiating each solid matrix and integrating the operating conditions as well as the parameters of the experimental points from the literature to develop predictive tools. The work will be based on the Sovova model (Figure 1) which takes into account the biological nature of solid matrices.

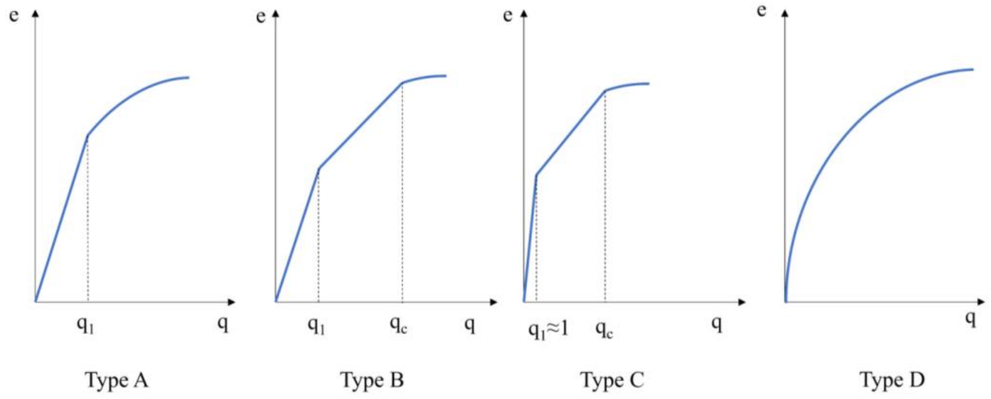


Figure 1 : Sovova models and parameters to be modeled

The information in the database will be exploited using design of experiments methodology in order to establish predictive models giving access to the values of the Sovova model parameters q1, qc, K... as a function of the type of natural matrix and the operating conditions. These tools will then be put together in the form of a software coded in Matlab. This software will simulate the extraction kinetics by CO2-SC from the choice of a natural matrix and the extraction operating conditions chosen by the user. It will also indicate the co-solvent used to extract the polar compounds of interest from the selected matrix. The tool will thus be totally predictive and versatile. The first results obtained on argan [3-5] and primerose are promising.

References

[1] López-Padilla A, Ruiz-Rodriguez A, Reglero G, Fornari T, (**2017**) Supercritical carbon dioxide extraction of Calendula officinalis : Kinetic modeling and scaling up study, *The Journal of Supercritical Fluids.* **130,** 292–300.

[2] Taher H, Al-Zuhair S, Al-Marzouqi A.H, Haik Y, Farid M, Tariq S, (**2014**) Supercritical carbon dioxide extraction of microalgae lipid: Process optimization and laboratory scale-up, *The Journal of Supercritical Fluids.* **86** 57–66.

[3] Mouahid A, Claeys-Bruno M, Bombarda I, Amat S, Ciavarella A, Myotte E, Nisteron J-P, Crampon C, Badens E, (**2022**) Supercritical CO2 extraction of oil from Moroccan unroasted Argan Kernels: Effects of process parameters to produce cosmetic oil, *Journal of CO2 Utilization,* **59,** 101952.

[4] Mouahid A, Bombarda I, Claeys-Bruno M, Amat S, Myotte E, Nisteron J-P, Crampon C, Badens E, (**2021**) Supercritical CO_2 extraction of Moroccan argan (Argania spinosa L.) oil: Extraction kinetics and solubility determination, *Journal of CO2 Utilization*, **46**.

[5] Mouahid A, Claeys-Bruno M, Bombarda I, Amat S, Ciavarella A, Myotte E, Crampon C, Nisteron J-P, Badens E, (**2023**) Valorisation of handmade Argan press cake by supercritical CO_2 extraction, *Food and Bioproducts Processing*, **137,** 168-176.

AQbD USING A D-OPTIMAL DESIGN AND PARAFAC IN AN AUTOMATIC SPE SYSTEM COUPLED TO HPLC-FLD

M.C. Ortiz[1], L. Valverde-Som[1], M.M. Arce[1], L.A. Sarabia[2]

Universidad de Burgos, [1]Dep. Chemistry, Pza. Misael Bañuelos s/n, Burgos, 09001, Spain
[2]Dep. Mathematics and Computation, Pza. Misael Bañuelos s/n, Burgos, 09001, Spain
E-mail (mcortiz@ubu.es)

Optimizing a multi-residue analysis when using an automatic SPE (Solid Phase Extraction) system and complex matrices becomes a difficult problem to address due to the large number of experimental factors that can influence the recovery of an analyte. Furthermore, in most cases the conditions that enhance the response for one analyte are in conflict with another.

In this work, AQbD (Analytical Quality by Design) is applied to the development of an analytical procedure based on automatic solid phase extraction system coupled to HPLC-FLD in a multi-residue analysis. To obtain the CMP (Control Method Parameter) that maximizing the desired CQA (Control Quality Attributes) a D-optimal design and PARAFAC have been used. Specifically, this methodology is applied in the simultaneous quantification of nine PAHs (Polycyclic Aromatic Hydrocarbons) in coffee samples. The first problem is to handle a large number of factors (with different levels) to optimize the responses. This task has been carried out using a D-optimal design [1] which, starting from a full factorial design of 4 factors with 72 experiments, allows to reduce to 19, maintaining the reliability of the estimates, saving time and costs in the laboratory. Table 1 shows each factor and their levels.

The second problem addressed is related to the choice of the type of response (Y_i). A complex matrix such as coffee contains impurities that interfere with the target analytes and in the chromatographic determination, they can coelute with them. A PARAFAC decomposition allows to avoid this problem and use the '*second order advantage*' to unequivocally identify each analyte. Then, the obtained sample loadings were used as response (specifically, each Y_i response is the difference between the obtained sample loading for a spiked coffee sample minus the sample loading of the pure coffee sample).

The multi-residue analysis has been performed with nine PAHs: PHE (phenanthrene), ANT (anthracene), FLN (fluoranthene), PYR (pyrene), CHR (chrysene), BaA (benzo[a]anthracene), PER (perylene), BbF (benzo[b]fluoranthene) and BaP (benzo[a]pyrene). PAHs are a family of organic compounds with potential toxicity (carcinogenicity) to humans and, they are mainly generated during the incomplete combustion of organic matter (such as coffee roasting). European regulations amending foodstuff set maximum levels for BaP and the sum of the content of four compounds (PAH4): BaP, BaA, BbF and CHR.

Table 1. The four factors (CMP) and their levels in the SPE optimization procedure.

Code	**Factor**	**Num. levels**	**Code Level**
X_1	Elution volume	2	A: 2 mL B: 5 mL
X_2	Dry time with nitrogen in the wash stage	3	A: 5 min B: 7.5 min C: 10 min
X_3	Wash volume	3	A: 1 mL B: 3 mL C: 5 mL
X_4	Organic solvent	4	A: 0% B: 20 % methanol C: 20 % acetonitrile D: 10 % methanol +10% acetonitrile

With the D-optimal model built, 13 coefficients are fitted that estimate the effect of going from one level to another in each factor, as well as the interaction between the levels of two of them (X_2, X_3). Once the 9 models [1] have been validated, it is possible to choose the levels of each factor that maximize the majority of nine responses (Y_i), which are shown in Table 2. The last column shows the experimental condition chosen.

Table 2. Levels of the factors that maximize each response Y_i. *significant at 0.05

	Y_{PHE}	Y_{ANT}	Y_{FLN}	Y_{PYR}	Y_{CHR}	Y_{BaA}	Y_{PER}	Y_{BbF}	Y_{BaP}	Level chosen
X_1	B	B	B	A	B	-	B	B	B	5 mL
X_2	C	A	B	B	B	B	B	B*	B	7.5 min
X_3	B	A	A	A	A	A	A	A	A	1 mL
X_4	A*	D	D	D*	D*	D*	D	D	D*	10%MeOH +10% ACN

Acknowledgements

Authors acknowledge the funding of Junta de Castilla y León under project BU052P20, co-financed with European Regional Development Funds.

References

[1] Mathieu D, Nony J, Phan-Than-Luu R, Beal A., NemrodW®, ver 2017, Marseille, France.

Optimization lithium recovery from $LiFePO_4$ batteries based on agri-food wastes through experimental design

Raffaele Emanuele Russo, Martina Fattobene, Paolo Conti, Silvia Zamponi, Mario Berrettoni and Gabriele Giuli

Addresses: University of Camerino, School of Science and Technology, ChIP, via
Madonna delle Carceri 62032 Camerino (MC) Italy
E-mail (raffaele.russo@unicam.it)

Lithium recovery from batteries is becoming increasingly important due to the growing demand for lithium as a critical material in the production of rechargeable batteries used in electric vehicles and consumer electronics.[1] The global lithium iron phosphate battery market is projected to grow from $10.12 billion in 2021 to $49.96 billion by 2028.[2]
Recycling lithium from spent batteries not only conserves the limited resources of lithium but also helps reduce the environmental impact associated with mining and refining the metal. In addition, lithium recovery from batteries can also help reduce the cost of production of new batteries and improve their sustainability.[3]

The best process to recover lithium from lithium iron phosphate batteries can be done through the hydrometallurgical process that involves the dissolution of the battery material in an aqueous solution, followed by separation and purification of the lithium. This process can be more environmentally friendly than the pyrometallurgical process and is often used for recycling batteries on a larger scale.[4]

The aim of this work is developing a cost-effective and sustainable hydrometallurgical process. For this reason, we propose the use of tartaric acid in order to recover lithium from $LiFePO_4$ pouch cells. Tartaric acid is an eco-friendly weak organic acid that occurs naturally in many fruits (mainly in grapes but also in bananas, tamarinds and citrus) and can be recovered from various natural by-products, mostly from winery ones.[5] It is an effective leaching agent for metal extraction because of their ability to complex with metal ions and it is also a perfect substitute for common strong inorganic acids (H_2SO_4, HCl, HNO_3) and strong basis (NaOH, NH_4OH).

For this study, tartaric acid was supplied by Distillerie Mazzari S.p.A. (Italy), which recovers it from wine waste (vinasses).

In other words, using agri-food wastes it is possible to process another waste, the industrial one, to recover and reintegrate precious metals for an efficient circular economy.

A representative dataset of pouch cells was considered to ensure that the project outcome was representative, including the possibility of exporting it to a pilot and/or industrial process. So, $LiFePO_4$ pouch cells were characterized by X-ray Diffraction (XRD) to determine the mineralogical composition and the degree of crystallinity; by Scanning Electron Microscopy (SEM) with Energy Dispersive X-ray Spectroscopy (EDS) for a morphological study followed by an elementary analysis; by inductively coupled plasma (ICP-OES) to investigate the concentration of Li and Fe in the batteries. For the recovery process were considered a series of independent variables: solid to liquid ratio (g/L), reaction times (min), temperature (°C), acid concentrations (mol/L) and H_2O_2 concentration (vol %). To determine the optimal conditions for an efficient metal recovery and to maximize profits, an experimental design (DOE) was carried out. In particular way, a two-step experimental design was conducted to determine the optimal conditions for metal recovery. First, a screening design was carried out to identify the most significant variables. Then, a response surface design, through a central composite design was performed to find the optimal conditions for the selected independent variables.

References

1. Mahandra, H. & Ghahreman, A. A sustainable process for selective recovery of lithium as lithium phosphate from spent LiFePO 4 batteries. *Resour. Conserv. Recycl.* **175**, 105883 (2021).
2. Lithium Iron Phosphate Batteries Market. (2022). Available at: https://www.fortunebusinessinsights.com/lithium-ion-li-ion-phosphate-batteries-market-102152.
3. Vehicles, E. & Environmental, P. I. I. Literature Review, Recycling of Lithium-Ion Batteries from Electric Vehicles, Part II: Environmental and Economic Perspective. (2022).
4. Forte, F., Pietrantonio, M., Pucciarmati, S. & Fontana, D. Lithium iron phosphate batteries recycling : An assessment of current status. *Crit. Rev. Environ. Sci. Technol.* 1–28 (2020).
5. Kontogiannopoulos, K. N., Patsios, S. I. & Karabelas, A. J. Tartaric acid recovery from winery lees using cation exchange resin : Optimization by Response Surface Methodology. *Sep. Purif. Technol.* **165**, 32–41 (2016).

When DoE goes Untargeted: optimization and validation of an extraction procedure for emerging contaminants monitoring in a complex matrix

Paravano Tommaso, **Boscolo Samuele**, **Roverso Marco**, **Centomo Paolo**

University of Padua, Department of chemical sciences, Via Marzolo 1, Padua, 35121, Italy
tommaso.paravano@phd.unipd.it

Hundreds of new chemicals are produced and released in the environment every year. Many of these, defined as emerging contaminants, are potentially harmful even at low concentrations, for humans, animals, and the environment itself. The monitoring of emerging contaminants in different matrices is crucial to unearth contaminated areas or compartments, to assess bioaccumulation and biomagnification in the biota and to evaluate the associate risk. **[1]**

To this regard, bivalves are considered good bioindicators to evaluate environmental contaminations, as these animals exclusively feed by filtering high volumes of sea water. Together with nutrients, they potentially accumulate contaminants present in the water and in the soil **[2]** and, consequently, the determination of contaminants in bivalve tissues is representative of the environmental contamination itself.

From an analytical point of view, the characterization and quantification of emerging contaminants in animal tissues is doubly challenging, as analytes are unknown, and the matrix is complex. The non-targeted workflow is generally based on high performance liquid chromatography coupled with high resolution mass spectrometry (HPLC-HRMS) **[3]**. In this light, along with the optimization of the instrumental parameters, it is also crucial to optimize and validate a general and simple extraction protocol to cover the widest possible range of analytes in terms of chemical properties (i.e. molecular weight and polarity). Dispersive solid phase extraction, based on QuEChERS (Quick, Easy, Cheap, Effective, Rugged and Safe) approach, is a good compromise, but several parameters should be evaluated.

In the present work, the optimization and validation, on 33 selected analytes of a general extraction procedure based on QuEChERS, is proposed. A d-Optimal Design approach, based on four variables (matrix to solvent ratio and percentage of acid in the solvent being the continuous ones and salting-out buffer composition and adsorbent for clean up being the discreet ones), three levels each, was chosen to optimize the number of experiments. The responses were analyzed with CAT [4] software to obtain coefficients, confidence intervals and isoresponse curves for every variable and every analyte (Figure 1).

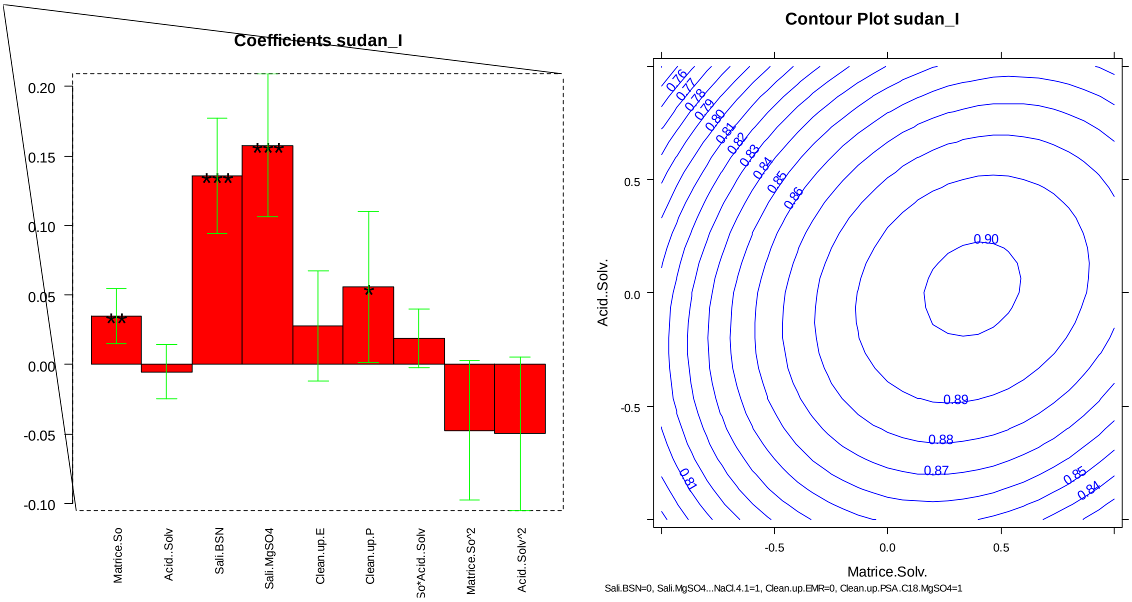


Figure 1: Example of coefficients values (on the left) and isoresponse curve (on the right).

In the isoresponse curves plot, the two continuous variables are on the abscissa and ordinate and the levels of the two discreet variables were selected according to the maximum response. This allowed, for each analyte, the determination of the best set of values for the four variables under study. The values of the parameters leading to the extraction, at least to some extent, of all the analytes was considered the optimal one.

References

[1] Ouda, M. et al. (**2021**) Emerging contaminants in the water bodies of the Middle East and North Africa. *Science of the Total Environment*, **754**.
[2] Chiesa, L. M., Nobile, M., Malandra, R., Panseri, S. & Arioli, F. (**2018**) Occurrence of antibiotics in mussels and clams from various FAO areas. *Food Chem,* **240**, 16–23.
[3] Pang, X. et al. (**2021**) Wide-scope multi-residue analysis of pesticides in beef by UHPLC coupled with quadrupole time-of-flight mass spectrometry. *Food Chem,* **351**.
[4] Leardi, R.; Melzi, C.; Polotti, G. CAT (Chemometric Agile Tool), freely downloadable from http://gruppochemiometria.it/index.php/software.

Quantitative Structure-Property Relationships for the prediction and understanding of liquid phase oxidation in hydrocarbons

Adrián Venegas-Reynoso[a,b], Lucia Giarracca-Mehl[a], Benoît Creton[a],
Marion Lacoue-Negre[a], Cyril Ruckebusch[b], Ludovic Duponchel[b]

Addresses: [a] IFP Energies nouvelles, 1-4 avenue de Bois-Préau, Rueil-Malmaison, 92852, France
[b]Univ. Lille, CNRS, UMR 8516 – LASIRE – Laboratoire de Spectroscopie pour Les Interactions, La Réactivité et L'Environnement, Lille, F-59000, France
E-mail: adrian.venegas-reynoso@ifpen.fr

The aviation sector is one of the most important industries in the world, by 2019 it supported around 4.1% of the world's Gross Domestic Product, transported around 4.5 billion passengers and 0.5% of international shipment, representing 35% of worldwide trade by value.

However, 80% of CO_2 emissions come from long-haul flights, representing a limitation for the low energy density of current batteries. With the present projections, it is not expected that cryogenic hydrogen or electric-powered planes will be commercially available before 2030 [1-3]. On the other hand, intense research has been performed for the development and certification of synthetic fuels coming from renewable feedstock, also known as Sustainable Aviation Fuels (SAF). Since SAF can reduce CO_2 emissions up to 80% over their life cycle, they are considered key elements to reach the sector's climate objectives.

In general, it is expected that both conventional jet fuels and SAF preserve their properties over time, nevertheless the exposure pollutants and thermal stress during the length of their lifetime can result in oxidation. Oxidation leads to fuel degradation, which deteriorate the quality of the product, which can thus limit the efficiency of the system or even lead to failures [4].

Different approaches have been used to model autoxidation; for example, several detailed mechanisms for different hydrocarbons have been developed [5–7]. However, these approaches cannot consider solvent effects in liquid phase oxidation. Even though some corrections have been [8, 9] used to incorporate the solvent effects, the number of parameters to be calculated in real fuels with thousands of mixtures is computationally unfeasible [10].

An alternative to the traditional kinetic modelling is the use of Quantitative Structure-Property Relationships (QSPR), which has already been used for the prediction of several hydrocarbons and fuels, such as the freezing point, flash point, octane and cetane number [11-13].

In this work, we propose to use QSPR to predict and understand the factors responsible of oxidation stability in liquid phase.

We identified 105 relevant and commercially available pure hydrocarbons present in conventional and alternative jet fuels. These hydrocarbons were analyzed by the Rapid Small Scale Oxidation Test (PetroOxy/RapidOxy) with oxygen at 700 kPa and temperatures between 100 and 160 ^{O}C. This method measures the time required for the consumption of 10% of the oxygen present in the system, also known as induction period (IP). Relevant 2D molecular descriptors were calculated from the hydrocarbons' structures, such as the number of primary, secondary, tertiary and quaternary carbons, among other molecular fragments

A first model was obtained by using Partial Least Squares (PLS) [14] and the calculated molecular descriptors. The resulting model was able to accurately predict the stability of linear, branched and cyclic alkanes. However, the model provided inaccurate predictions for di-aromatics like 1-methylnaphthalene, and some mixtures of di-aromatics and naphtheno-aromatics with alkanes, suggesting the presence of non-linear blending effects. To address these limitations, non-linear modeling was performed with Support Vector Machine (SVM) regression [15].

The interpretation of the PLS model's coefficient allowed to determine the parameters related to a compound's reactivity, like the experimental temperature, the presence of tertiary carbons, secondary carbons in alpha position in naphthene-aromatics and alkyl-benzenes, and a high number of hydrogen atoms, among others. On the other hand, it was found that the presence of quaternary carbons, aromatic and aliphatic rings was related to high stability. While these trends have been already found in other works [16,17], to our knowledge, this work is the first to establish to generate a database of the IP of jet fuel-related hydrocarbons with a standardized method, as well to quantitively determine the relationship between IP and several molecular features.

Future work will focus on the use Near Infrared Spectroscopy and chemometrics to study relations between the formation of oxidized species and the induction period, as well as to try to predict the IP from spectral information.

References

[1] A. N. Sarkar. Evolving Green Aviation Transport System: A Holistic Approach to Sustainable Green Market Development. American Journal of Climate Change, 01(03): 164–180, 2012. ISSN 2167-9495. doi: 10.4236/ajcc.2012.13014

[2] Air Transport Action Group. Facts & figures, 31/07/2022. URL https://www.atag.org/facts-figures.html

[3] Johnathan Holladay, Zia Abdullah, and Joshua Heyne. Sustainable Aviation Fuel: Review of Technical Pathways.

[4] James Pullen and Khizer Saeed. An overview of biodiesel oxidation stability. Renewable and Sustainable Energy Reviews, 16(8):5924–5950, 2012. ISSN 13640321. doi: 10.1016/j.Rser.2012.06.024.

[5] Yachao Chang, Ming Jia, Yaopeng Li, Yaodong Liu, Maozhao Xie, Hu Wang, and Rolf D. Reitz. Development of a skeletal mechanism for diesel surrogate fuel by using a decoupling methodology. Combustion and Flame, 162(10):3785–3802, 2015. ISSN 00102180. doi: 10.1016/j.combustflame.2015.07.016.

[6] Wenbin Yu, Wenming Yang, Kunlin Tay, and Feiyang Zhao. Development of a new skeletal mechanism for decalin oxidation under engine relevant conditions. Fuel, 212: 41–48, 2018. ISSN 00162361. doi: 10.1016/j.fuel.2017.10.009.

[7] Wenbin Yu, Feiyang Zhao, Wenming Yang, Kunlin Tay, and Hongpeng Xu. Development of an optimization methodology for formulating both jet fuel and diesel fuel surrogates and their associated skeletal oxidation mechanisms. Fuel, 231:361–372, 2018. ISSN 00162361. doi: 10.1016/j.fuel.2018.05.121.

[8] M. H. Abraham. Application of solvation equations to chemical and biochemical processes. Pure and Applied Chemistry, 65(12):2503–2512, 1993. ISSN 0033-4545. doi: 10.1351/pac199365122503.

[9] Christina Mintz, Tara Ladlie, Katherine Burton, Michael Clark, William E. Acree, and Michael H. Abraham. Enthalpy of Solvation Correlations for Gaseous Solutes Dissolved in Alcohol Solvents based on the Abraham Model. QSAR & Combinatorial Science, 27(5):627–635, 2008. ISSN 1611020X. doi: 10.1002/qsar.200730128.

[10] M. D. Le, V. Warth, L. Giarracca, E. Moine, R. Bounaceur, R. Privat, J-N Jaubert, R. Fournet, P-A Glaude, and B. Sirjean. Development of a Detailed Kinetic Model for the Oxidation of n-Butane in the Liquid Phase. The journal of physical chemistry. B, 125 (25):6955–6967, 2021. doi: 10.1021/acs.jpcb.1c02988.

[11] Li, Runzhao; Herreros, Jose Martin; Tsolakis, Athanasios; Yang, Wenzhao (2020) Machine learning regression based group contribution method for cetane and octane numbers prediction of pure fuel compounds and mixtures. In : Fuel, vol. 280, p. 118589. DOI: 10.1016/j.fuel.2020.118589.

[12] Le, M. D.; Warth, V.; Giarracca, L.; Moine, E.; Bounaceur, R.; Privat, R. et al. (2021) Development of a Detailed Kinetic Model for the Oxidation of n-Butane in the Liquid Phase. In : The journal of physical chemistry. B, vol. 125, n° 25, p. 6955–6967. DOI: 10.1021/acs.jpcb.1c02988.

[13] Hall, Clemens; Creton, Benoit; Rauch, Bastian; Bauder, Uwe; Aigner, Manfred (2022) Probabilistic Mean Quantitative Structure–Property Relationship Modeling of Jet Fuel Properties. In : Energy & Fuels, vol. 36, n° 1, p. 463–479. DOI: 10.1021/acs.energyfuels.1c03334.

[14] Geladi, Paul; Kowalski, Bruce R. (1986) Partial least-squares regression: a tutorial. In : Analytica chimica acta, vol. 185, p. 1–17. DOI: 10.1016/0003-2670(86)80028-9.

[15] Cortes, Corinna; Vapnik, Vladimir (1995) Support-vector networks. In : Machine Learning, vol. 20, n° 3, p. 273–297. DOI: 10.1007/BF00994018.

[16] Ben Amara, Arij; Kaoubi, Skander; Starck, Laurie (2016) Toward an optimal formulation of alternative jet fuels: Enhanced oxidation and thermal stability by the addition of cyclic molecules. In : Fuel, vol. 173, p. 98–105. DOI: 10.1016/j.fuel.2016.01.040.

[17] Larsen, R. G.; Thorpe, R. E.; Armfield, F. A. (1942) Oxidation Characteristics of Pure Hydrocarbons. In : Industrial & Engineering Chemistry, vol. 34, n° 2, p. 183–193. DOI: 10.1021/ie50386a012.

TIRESIA, a friendly XAI website for the assessment of developmental toxicity

F. Mastrolorito,[a] M.V. Togo,[a] N. Gambacorta,[a] D. Trisciuzzi,[a] A.R. Tondo,[a] C.D. Altomare,[a] F. Ciriaco,[b] N. Amoroso,[c] and O. Nicolotti.[a]

[a]University of Aldo Moro, Department of Pharmacy, Via Orabona, 4, Bari, 70125, Italy
[b]University of Aldo Moro, Department of Chemistry, Via Orabona, 4, Bari, 70125, Italy
[c]National Institute for Nuclear Physics, Bari Division, Via Orabona, 4, Bari, 70125, Italy
fabrizio.mastrolorito@uniba.it

In the last years, Artificial Intelligence has played an increasingly relevant role for the prediction of human health toxicological endpoints.[1] Among others, the developmental toxicity[2] is of utmost importance for the protection of children from the risks posed by chemicals. This work presents TIRESIA (an acronym standing for Toxicology Intelligence and Regulatory Evaluations for Scientific and Industry Application), a free website available at http://tiresia.uniba.it/.[3] TIRESIA employs the established CAESAR[4] training set made of 234 chemicals for model learning. As many as 212 free available molecular descriptors by RDKit[5] and 606 free available autocorrelator descriptors by Mordred[6] were calculated. Various classifiers such as Random Forest (RF),[7] Support Vector Machine (SVM)[8] and eXtreme Gradient Boostin (XGB)[9] were challenged and based on cross-validation analyses, the latter was selected along with 27 molecular descriptors.

The CAESAR test set and the Procter&Gamble (P&G)[10] industrial dataset, including as a whole 585 chemicals, were finally used for model validation and generalization.

TIRESIA provides, in agreement with OECD principles,[11] a density-based applicability domain to improve the prediction reliability. Notably, TIRESIA generates results in a standard portable document format, detailing a wealth of information about SHAP explainability analyses to increase the prediction transparency.[12,13]

References

(1) Lavecchia, A. Machine-Learning Approaches in Drug Discovery: Methods and Applications. *Drug Discov. Today* **2015**, *20* (3), 318–331. https://doi.org/10.1016/j.drudis.2014.10.012.

(2) Zhang, H.; Mao, J.; Qi, H.-Z.; Ding, L. In Silico Prediction of Drug-Induced Developmental Toxicity by Using Machine Learning Approaches. *Mol. Divers.* **2020**, *24* (4), 1281–1290. https://doi.org/10.1007/s11030-019-09991-y.

(3) Togo, M. V.; Mastrolorito, F.; Ciriaco, F.; Amoroso, N.; Nicolotti, O. TIRESIA: An EXplainable Artificial Intelligence Platform for Predicting Developmental Toxicity. *J. Chem. Inf. Model.* **2022**. https://doi.org/10.1021/acs.jcim.2c01126.

(4) Cassano, A.; Manganaro, A.; Martin, T.; Young, D.; Piclin, N.; Pintore, M.; Bigoni, D.; Benfenati, E. CAESAR Models for Developmental Toxicity. *Chem. Cent. J.* **2010**, *4* (S1), S4. https://doi.org/10.1186/1752-153X-4-S1-S4.

(5) Landrum, G. *RDKit: Open-Source Cheminformatics Software*. https://www.rdkit.org/.

(6) Moriwaki, H.; Tian, Y.-S.; Kawashita, N.; Takagi, T. Mordred: A Molecular Descriptor Calculator. *J. Cheminformatics* **2018**, *10* (1), 4. https://doi.org/10.1186/s13321-018-0258-y.

(7) Biau, G.; Scornet, E. A Random Forest Guided Tour. *TEST* **2016**, *25* (2), 197–227. https://doi.org/10.1007/s11749-016-0481-7.

(8) Noble, W. S. What Is a Support Vector Machine? *Nat. Biotechnol.* **2006**, *24* (12), 1565–1567. https://doi.org/10.1038/nbt1206-1565.

(9) Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; KDD '16; Association for Computing Machinery: New York, NY, USA, 2016; pp 785–794. https://doi.org/10.1145/2939672.2939785.

(10) Wu, S.; Fisher, J.; Naciff, J.; Laufersweiler, M.; Lester, C.; Daston, G.; Blackburn, K. Framework for Identifying Chemicals with Structural Features Associated with the Potential to Act as Developmental or Reproductive Toxicants. *Chem. Res. Toxicol.* **2013**, *26* (12), 1840–1861. https://doi.org/10.1021/tx400226u.

(11) Nicolotti, O.; Gadaleta, D.; Mangiatordi, G.; Catto, M.; Carotti, A. Applicability Domain for QSAR Models: *Int. J. Quant. Struct.-Prop. Relatsh.* **2016**, *1*, 45–63. https://doi.org/10.4018/IJQSPR.2016010102.

(12) Dosilovic, F. K.; Brčič, M.; Hlupic, N. Explainable Artificial Intelligence: A Survey. *2018 41st Int. Conv. Inf. Commun. Technol. Electron. Microelectron. MIPRO* **2018**. https://doi.org/10.23919/MIPRO.2018.8400040.

(13) Shapley, L. S. 17. A Value for n-Person Games. In *17. A Value for n-Person Games*; Princeton University Press, 2016; pp 307–318. https://doi.org/10.1515/9781400881970-018.

TEXTURE ANALYSIS AND CHEMOMETRICS OF HYPERSPECTRAL IMAGES OF BREAD

Prost Carole, Fillonneau Catherine, Bertrand Dominique, Jaillais Benoît

Consultant, 25 rue Stendhal, 44300 Nantes, France.
Oniris, UMR Cnrs 6144 GEPEA, équipe MAPS2, 44300 Nantes, France
Oniris, INRAE, StatSC, 44300 Nantes, France
E-mail: Benoit.Jaillais@inrae.fr

Introduction. Bread-making quality is an important issue in the food industry and much research has been conducted to characterize it (MacRitchie, 2016). The use of non-contact and fast spectroscopic sensors is a competitive advantage. Hyperspectral imaging combines spatial and spectral measurement, and increases the discriminating power of image analysis. The aim of this study was to evaluate breadmaking quality by using texture analysis of hyperspectral imaging and chemometrics.

Material and methods. In this study, the slices of bread are issued from breads made from an experimental design with the recipe (P), the baking(C), the position of the slice and the staling time as factors. To characterize these products, we used a Visible-Near InfraRed (VNIR) imaging system, which is inexpensive, easy to use, easy to transport, and does not require a conveyor. The chemometrical part of this work focuses on the analysis of the image texture using co-occurrence matrix (Haralick et al., 1973) by extracting parameters (contrast, correlation, energy and homogeneity), which will be used as new predictive features. These data are analyzed by principal component analysis (PCA) and analysis of variance (ANOVA) for the selection of relevant variables.

Results. The PCA results (Figure1a) show that the breads are clearly different by recipe on the PC1-PC2 factorial map. Bread with recipe #1 was correlated with the feature “Contrast” regardless the angle tested. Contrast is a measure of the difference in intensity between a pixel and its neighbor across the image. The more the texture is contrasted, the higher the feature. Breads with recipes #2 and #3 are associated with the correlation and homogeneity/energy variables, respectively.

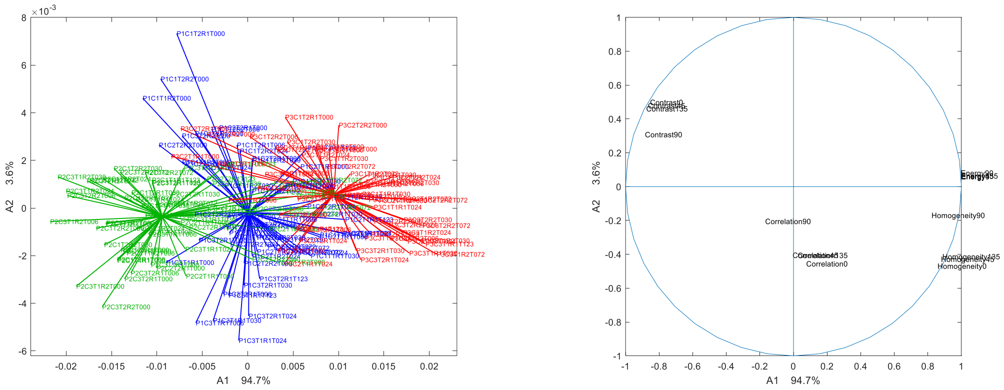


Figure 5: PC1-PC2 factorial map with barycenters of group "recipes" (a) and correlation circle (b).

The ANOVA shows a significant effect for the factor "recipe" (figure 2), which is associated with the "correlation" parameter. The factor "cooking" is significant to a lesser extent and an interaction between these two factors was also noted.

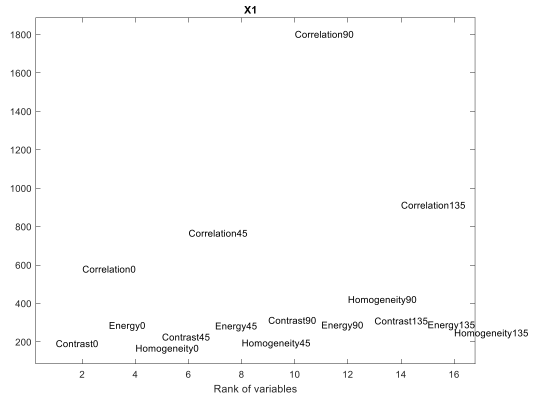


Figure 6: Independent analyses of variance: Fisher test values

Discussion and conclusion. This first study of the potential of VNIR imaging for spatial analysis of bread characteristics with different manufacturing parameters could become a complementary tool to those of sourdough rheology on the one hand, and near infrared spectroscopy on the other hand.

Future developments concern a generalization of the calculation of the co-occurrence matrix, by integrating different distances between the pixels considered. Taking into account all spectral frames will allow to complete this study by combining the spatial and spectral aspects, and to select relevant variables, i.e. wavelengths.The presented methodology can be appropriated for characterizing breadmaking quality.

References

Haralick, R.M., Shanmugan, K., Dinstein, I., 1973. Textural Features for Image Classification. IEEE Trans. Syst. Man Cybern. 610–621.

MacRitchie, F., 2016. Seventy years of research into breadmaking quality. J. CEREAL Sci. 70, 123–131. https://doi.org/10.1016/j.jcs.2016.05.020.

WHEN DO HETEROGENEOUS SAMPLES REALLY NEED HYPERSPECTRAL IMAGING TECHNIQUES?

Daniele Tanzilli[1,2], Alessandro D'Alessandro[1,3], José Manuel Amigo[4,5], Marina Cocchi[1]

1 University of Modena e Reggio Emilia, DSCG, Via Campi 103, Modena, 41121, Italy

2 Université de Lille, LASIRE, Cité Scientifique, Villeneuve-d'Ascq, 59650, France

3 Barilla G. e R. Fratelli, Via Mantova 166, Parma, 43122, Italy

4 University of the Basque Country, Department of Analytical Chemistry, Barrio Sarriena S/N, Leioa, 48940, Spain.

5 IKERBASQUE, Basque Society for the Promotion of Science, Plaza Euskadi 5, Bilbao, 48009, Spain.

daniele.tanzilli@unimore.it

The continuous technological evolution and the interest of industries in monitoring their own industrial processes have allowed the installation of in-line or on-line of Near Infrared (NIR) sensors that, together with the potential of chemometrics, provide real-time information of the quality of the final product.

Regarding quality, one of the main parameters that the industry is concerned with when quantifying and classifying is the heterogeneity of the material and the type of the NIR sensor needed to asses that quality. When the material is suspected to be heterogeneous, NIR hyperspectral cameras might be desired instead of NIR classical single-spot spectrophotometers. Nevertheless, it is not clear yet the extent to which a NIR hyperspectral camera is needed or the spatial information that a single spot classical NIR device can provide.

The case study here concerns the industrial production of *pesto alla genovese* sauce in Barilla G. e R. Fratelli S.p.A food company (Italy). *Pesto alla genovese* sauce is one of the most popular Italian foods, and it is composed of an apparent heterogeneous mixture of basil, garlic, extra-virgin olive oil, parmesan cheese and other minor ingredients.

Considering heterogeneous products such as *pesto*, Hyperspectral imaging techniques (HSI) have the advantages of optical spectroscopy combined with spatial information, thanks to the acquisition of spectral information for each pixel of the image, taking into consideration the distribution of components which enables objects to be found, materials to be identified or processes to be detected [1].

This work aims at defining the application range of single-spot NIR spectrophotometers and NIR hyperspectral devices for the classification of different *pesto* sauces in-line. Points like the need for the HSI technique against NIR single spot will be comprehensively evaluated considering the component distribution in order not to overestimate the area or interest.

References

1 Amigo, J. M. (**2019**) Hyperspectral and multispectral imaging: Setting the scene. In Data Handling in Science and Technology, Elsevier, **32**.

IDENTIFICATION AND LOCALIZATION OF DEHYDRATED BIOLOGICAL FLUIDS IN FORENSIC EVIDENCES: THE ROLE OF CHEMOMETRIC STRATEGIES IN THE PROCESSING OF MULTIBLOCK HYPERSPECTRAL DATA

G. Sciutto[1], P. Oliveri[2], E. Alladio[3], A. Mazzoleni[3], E. Catelli[1], C. Malegori[2], S. Prati[1], R. Mazzeo[1], M. Occhipinti[4], P. Garofano[5], R. Alberti[4], T. Frizzi[4]

[1] Department of Chemistry, University of Bologna, Via Guaccimanni 42, 48121 – Ravenna
[2] DIFAR Department of Pharmacy, University of Genoa, Viale Cembrano 4, 16148 – Genoa
[3] Department of Chemistry, University of Torino, Via Giuria 5, 10125 – Torino
[4] XGLab SRL – Bruker Nano Analytics, Via Conte Rosso 23, 20134 – Milano
[5] Centro Regionale Antidoping "A. Bertinaria", Regione Gonzole 10/1, 10043 – Orbassano (Torino)
giorgia.sciutto@unibo.it (corresponding author)

From a criminalistic point of view, traditional body fluid identification and analysis present severe limitations, essentially due to the critical conditions (e.g., complex heterogeneous sample matrices and/or degradation effects) in which these evidences are often found. Nowadays, the development of new techniques allowing to detect minute biological traces in a selective and non-destructive way is an urgent research issue in the forensic field.

The present study delineates the possibility of localizing and identifying dehydrated biological fluids by using a novel multiblock platform (IRIS, XGLab SRL – Bruker Nano Analytics) that simultaneously acquires X-ray fluorescence (XRF, 1-40 keV) data together with visible & near-infrared (VNIR, 380–1100 nm) and short-wave infrared reflectance spectra (SWIR, 1100–2500 nm), coupled with chemometric methods. To exploit the complementary information embedded in the different hyperspectral data blocks, an efficient multivariate strategy based on principal component analysis (PCA) and multivariate curve resolution alternating least squares (MCR-ALS) was applied. Moreover, preliminary classification models involving partial least squares – discriminant analysis (PLS-DA) were evaluated, too. The chemometric strategy proved to be highly efficient in the analysis of 3D data arrays by resolving overlapped vibrational bands and correlating elemental information.

The performances of the chemometric strategy were evaluated for the differentiation of different biological fluids, such as blood, urine, semen and saliva, applied on wood and cotton tissue, producing multivariate chemical distribution maps, with and without the exploitation of reference spectra of pure components.

EXPLOITING THE CORRELATION BETWEEN DIFFERENT SPECTRAL BLOCKS IN A MULTIPLATFORM HYPERSPECTRAL IMAGING SYSTEM

P. Oliveri[1], G. Sciutto[2], E. Catelli[2], Z. Li[2], C. Malegori[1], S. Prati[2], M. Occhipinti[3], R. Alberti[3], T. Frizzi[3], R. Mazzeo[2]

[1]DIFAR Department of Pharmacy, University of Genova, Viale Cembrano 4, 16148 – Genova
[2]Department of Chemistry, University of Bologna, Via Guaccimanni 42, 48121 – Ravenna
[3]XGLab SRL – Bruker Nano Analytics, Via Conte Rosso 23, 20134 – Milano
paolo.oliveri@unige.it (corresponding author)

Hyperspectral imaging (HSI) systems are today well established in a wide range of analytical chemistry applications, driven by technological advancements. In the present study a multiblock platform that simultaneously acquire X-ray fluorescence (XRF, 1-40 keV) data together with visible & near-infrared (VNIR, 380–1100 nm) and short-wave infrared reflectance spectroscopy (SWIR, 1100–2500 nm) was applied (IRIS, XGLab SRL – Bruker Nano Analytics). A critical point that limits implementation and application of such a platform is the complexity of data processing, trying to find a good compromise between advanced computational operations and straightforward graphical interfaces for data management and visualisation. Ideally, outputs should be easily readable and interpretable both by the users, who aim at digging inside the hyperspectral data, to reveal and discover the composition of the investigated surface, and by the research community, with whom the results are shared. In the present study, to exploit the complementary information embedded in the different hyperspectral data blocks, an efficient multivariate strategy based on the combination of principal component analysis (PCA), brushing, correlation diagrams and maps (within and between spectral blocks) was purposedly developed and presented. The chemometric strategy proved to be highly efficient in data reduction and for the extraction and integration of the most useful information coming from the three different spectroscopies, through an effective data visualization and inter-correlation assessment. As an illustrative example, data obtained from the analysis of a multi-layered painting were used.

MONITORING THE COLOUR OF STRAWBERRY YOGHURT PURÉE: A COMBINED APPROACH OF MIXTURE DESIGN AND MULTIVARIATE IMAGE ANALYSIS

Pier Lorenzo Rolando[1], Rosalba Calvini[1,2], Giorgia Foca[1,2], Alessandro Ulrici [1,2]

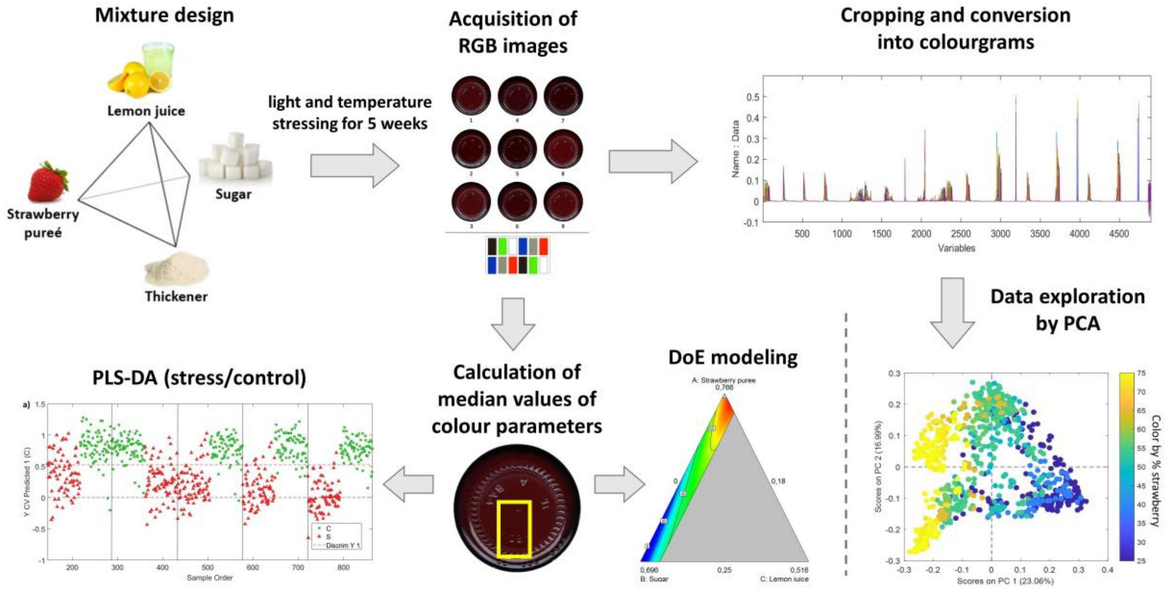
[1] Dipartimento di Scienze della Vita, Università di Modena e Reggio Emilia, Padiglione Besta, Via Amendola, 2 – 42122 Reggio Emilia, Italy

[2] Centro Interdipartimentale BIOGEST-SITEIA, Università degli Studi di Modena e Reggio Emilia, Piazzale Europa, 1 – 42122 Reggio Emilia, Italy

pierlorenzo.rolando@unimore.it

Food colour is the first appealing factor for consumers, thus representing a commercial added value. In this work we investigated how the formulation of strawberry yogurt purée (SYP) – a strawberry-based preparation widely used in industry – affects its colour and how the colour changes over time when SYP is subjected to stress conditions due to the combined effect of light and temperature. In fact, the colour of anthocyanins can vary due to several factors, such as pH, temperature, light exposure, oxidase enzyme activity, and so on (Holzwarth, 2013), which cause degradation and browning phenomena.

A combined approach involving mixture design and multivariate analysis of RGB images was used to achieve these objectives. Different proportions of strawberry purée, sugar, lemon juice and two types of thickener were mixed according to an I-optimal mixture design to obtain 44 SYP mixtures. The samples were stressed for five weeks under light and 35 °C of temperature, while a second aliquot of each sample was stored in the dark at 0-4 °C as control; every week RGB images of stressed and control samples were acquired by a flatbed scanner. The collected images were converted to reduce their dimensionality using two different approaches: i) by conversion into colourgrams, i.e., signals which can be seen as the colour fingerprint of the samples (Calvini, 2020), and ii) by calculation of the median values of various colour-related parameters. The colourgrams dataset was examined by Principal Component Analysis, while the median values of colour-related parameters were analysed by Response Surface Methodology and Partial Least Squares-Discriminant Analysis. Data analysis was aimed both at identifying the colour parameters most suitable to describe colour variability over time, and at investigating the cause-effect relation between mixture composition and colour parameters.



The results highlighted that the median red value is the most informative parameter to monitor SYP colour variation, which strongly depends on the amount of strawberry purée. The comparison between stressed and control samples showed a progressive browning over time, due to the thermal degradation of the anthocyanins, which becomes increasingly marked as the amount of strawberry purée increases. Results also showed that the thickener, although present in small quantities, influenced the colour degradation kinetics of the product. Finally, the median value of relative green was observed to increase over time for stressed samples. These results suggest that red and relative green parameters could be used as indices of the SYP colour degradation process.

References

Calvini R, Orlandi G, Foca G, Ulrici A (**2020**) Colourgrams GUI: A graphical user-friendly interface for the analysis of large datasets of RGB images. *Chemometrics and Intelligent Laboratory Systems*, **196**, 1–11.

Holzwarth M, Korhummel S, Siekmann T, Carle R Kammerer DR (**2013**) Influence of different pectins process and storage conditions on anthocyanin and colour retention in strawberry jams and spreads. *LWT - Food Science and Technology*, **52**, 131–138.

WHIM Descriptors for Medical Images Analysis

Bertani Pietro, Maiocchi Alessandro

Bracco SpA, Via Egidio Folli 50, Milan, 20134, Italy
pietro.bertani@bracco.com, alessandro.maiocchi@bracco.com

Radiomics is a field of medical imaging that aims the extraction and analysis of quantitative features from medical images, such as CT, MRI, and PET scans, with the goal of improving diagnostic and therapeutic decision-making, in direction to enable personalized medicine [1]. Quantitative features are used to characterize the texture, shape, and intensity of a defined region of interest, ROI, of the image. Those includes but are not limited to: first order statistical features, as mean and standard deviation; shape based features as volume and surface area; gray-level co-occurrence matrix (GLCM) features that describes lengths and orientations of runs of pixels with similar gray levels [2]. In overall, in literature hundreds of radiomic features have been described. Weighted Holistic Invariant Molecular, WHIM, descriptors are a set of molecular descriptors based on statistical indices calculated on the projections of the atoms along principal axes [3]. WHIMs are built in such a way as to capture relevant molecular 2D/3D information being invariant to translation, rotation, and scaling. As far as we know the application of such descriptors among traditional radiomic ones has not been explored yet. In proposed work, we transpose WHIM application from chemical to 2D or 3D image analysis. To do that, atomic coordinates of original WHIMs are substituted by pixel's coordinates and weights of the calculated covariance matrix are represented by image's pixels values, allowing to introduce a pixel intensity distribution dependence to shape descriptors, not possible with standard radiomic analysis. We started by computing the directional and global WHIMs based on the eigenvalues of a weighted covariance matrix. The directional ones determine the size of the image content along each principal direction, also considering the scores distribution, while the global descriptors are obtained by combining the directionals to account for changes in image properties across the main directions. To demonstrate the roto-translational invariance of this method, we firstly tested it on toy images. Application on Medical Images was performed on BrainMetShare dataset [4]. Starting from

11 descriptors, an analysis on a subset of ≈ 1200 2D pre-contrast slices with unsupervised clustering, hierarchal and k-means techniques, was performed to identify similar brain anatomies. To do that each slice was scaled between 0 and 255. Enlarged set of WHIM molecular descriptors can be defined after any binning screen procedure using grayscale intervals. Finally, application of WHIM was tested to characterize shapes of segmented tumour volumes, at different grayscale levels, of the BRATS2020 dataset [5]. All these analyses were developed adapting WHIM formulas to images exploiting Python 3.8.0 libraries.

Proposed WHIM descriptors applied to images, still not available in radiomics libraries, have been demonstrated to be invariant to roto translation. Moreover, this approach illustrates the potential of using this analysis to distinguish between images that represent different contexts of same body organ by using pixel-weighted shapes features. Preliminary studies for clinical outcome prediction are ongoing to be tested. Further potential successful application may involve the tracking of changes in the shape and size of tumours during and after treatment.

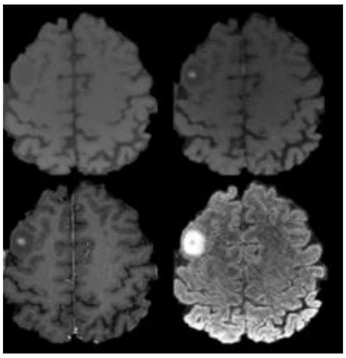


Figure 1. Examples of BrainMetShare MRI images

References

[1] Van Timmeren, J., Cester, D., Tanadini-Lang, S. et al. Radiomics in medical imaging-“how-to” guide and critical reflection. Insights Imaging 11, 91 (2020).

[2] Rizzo et al. “Radiomics: the facts and the challenges of image analysis”, European Radiology Experimental (2018) 2:36

[3] Todeschini R., Consonni V., - “Handbook of Molecular Descriptors” – Wiley VCH

[4] Grøvik et al., Deep Learning Enables Automatic Detection and Segmentation of Brain Metastases on Multisequence MRI, JMRI 2019; 51(1):175-182

[5] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, et al. "The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS)", IEEE Transactions on Medical Imaging 34(10), 1993-2024 (2015)

APPLICATION OF NIR-HYPERSPECTRAL IMAGING FOR THE MANAGEMENT OF THE BROWN MARMORATED STINK BUG (BMSB) PEST

V. Ferrari[1], R. Calvini[1], C. Menozzi[1], P. Offermans[2], L. Maistrello[1], G. Foca[1], A. Ulrici[1]

[1] University of Modena and Reggio Emilia, Department of Life Sciences, Pad. Besta, Via Amendola, 2, Reggio Emilia, 42122, Italy
[2] IMEC OnePlanet, Bronland 10, Wageningen, the Netherlands
veronica.ferrari@unimore.it

Halyomorpha halys, commonly known as Brown Marmorated Stink Bug (BMSB), is an insect pest of global importance which causes serious damages to the agri-food supply chain. In Europe, the Emilia-Romagna region was one of the areas of BMSB first occurrence, where it has caused major economic losses to pear orchards due to punctures on harvested fruits (Maistrello et al., 2017). In the frame of the EU Project HALY.ID, several strategies were proposed both upstream to prevent BMSB activity on crops and downstream to manage harvested product quality (HALY.ID, 2023).

With the aim of implementing innovative pest management strategies, Near Infrared Hyperspectral Imaging (NIR-HSI) was considered as a potential field monitoring method for BMSB detection. Starting from hyperspectral images of BMSB specimens acquired on different vegetal backgrounds in the 980-1660 nm spectral range, a library of representative spectra of both BMSB and vegetal background types was used to calculate pixel-level classification models. The considered approach focused on modelling spectral information and selecting relevant spectral regions for discrimination by means of sparse-based variable selection coupled with Soft Partial Least Squares Discriminant Analysis (s-Soft PLS-DA) algorithm. Besides leading to promising results (EFF>90%), the spectral regions selected by s-Soft PLS-DA can be viewed as a first step toward the development of multispectral imaging systems for on-field BMSB detection (Ferrari et al., 2023).

Moreover, NIR-HSI was also evaluated as a postharvest sorting system for the identification of damages caused by BMSB feeding activity on pears. To this aim, pear fruits (*cv.* «Williams» and «Abate Fétel») were collected in an organic orchard in Carpi (Modena, Italy). About half of the collected fruits were kept in contact with BMSB specimens for one week before harvesting, while the remainder fruits were used as control samples. The hyperspectral images of the collected fruits were acquired in the 1156-1674 nm range the same day of harvesting and at seven subsequent times, over a period of six weeks. Following the same procedure, RGB images of the pear samples were acquired as reference both before and after peeling the fruits. Indeed, the RGB images of unpeeled and peeled samples were used to visually classify the fruits based on damage type and intensity.

A preliminary exploratory analysis performed on the hyperspectral images using Principal Component Analysis enabled the identification of damages. However, a clear identification of the Regions Of Interest (ROIs) ascribable to punctures was challenging due to their highly irregular shapes and blurred edges between sound and damaged areas. To address this issue, a supervised annotation of punctured regions based on data dimensionality reduction using the hyperspectrograms approach and spatial feature selection was successfully adopted (Ferrari et al., 2013, 2015)

References

Ferrari C, Foca G, Ulrici A (**2013**) Handling large datasets of hyperspectral images: Reducing data size without loss of useful information. *Analytica Chimica Acta*, **802**, 29–39.

Ferrari C, Foca G, Calvini R, Ulrici A (**2015**) Fast exploration and classification of large hyperspectral image datasets for early bruise detection on apples. *Chemometrics and Intelligent Laboratory Systems*, **146**, 108–119.

Ferrari V, Calvini R, Boom B, Menozzi C, Rangarajan AK, Maistrello L, Offermans P, Ulrici A (**2023**). Evaluation of the potential of near infrared hyperspectral imaging for monitoring the invasive brown marmorated stink bug. *Chemometrics and Intelligent Laboratory Systems*, **104751**.

HALY.ID, project part of ERA-NET ICT-AGRI-FOOD, funded by the European Union's Horizon 2020 and by MIPAAF, https://www.haly-id.eu/, **2023**.

Maistrello L, Vaccari G, Caruso S, Costi E, Bortolini S, Macavei L, Foca G, Ulrici A, Bortolotti PP, Nannini R, Casoli L, Fornaciari M, Mazzoli GL, Dioli P (**2017**) Monitoring of the invasive *Halyomorpha halys*, a new key pest of fruit orchards in northern Italy. *Journal of Pest Science*, **90**, 1231–1244.

AN UPGRADE OF MVC2, A MATLAB GRAPHICAL USER INTERFACE FOR SECOND-ORDER MULTIVARIATE CALIBRATION

A. Muñoz de la Peña[1], F. A. Chiappini[2,3], H. C. Goicoechea[2,3], A. C. Olivieri[3,4]

[1]Universidad de Extremadura, Departamento de Química Analítica, Avda. de Elvas, s/n, Badajoz, 06006, Spain, [2]Universidad Nacional de Litoral, Laboratorio de Desarrollo Analítico y Quimiometría (LADAQ), Ciudad Universitaria, Santa Fe, S3000ZAA, Argentina,[3]Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Godoy Cruz 2290 CABA, C1425FQB, Argentina, [4]Universidad Nacional de Rosario, Departamento de Química Analítica, Suipacha 531, S2002LRK, Argentina

E-mail: arsenio@unex.es

In the present Communication, an upgrade of a MATLAB graphical user interface (GUI) toolbox for implementing second-order multivariate calibration models is presented, named Multivariate Calibration 2 (MVC2). Due to the advances in the chemometric field, this software has undergone numerous changes since its last release in 2009 [1]. The new freely available MVC2 is presented and described, including the most recent findings regarding second-order calibration algorithms and their associated analytical figures of merit (AFOMs), incorporating novel features and models that make it a very practical tool for data processing. Overall, the current version presents different types of algorithms, designed to provide specific degrees of model flexibility for different multi-way data scenarios: in addition to trilinear decomposition (TLD) and residual bilinearization (RBL) models, which were previously included in the 2009 version, parallel factor analysis 2 (PARAFAC2), a variant of the classic PARAFAC method [2], and multivariate curve resolution (MCR) [3] methods are now available. All these chemometric methodologies are integrated in the software for quantitative purposes, and if needed, they permit to exploit the so-called analytical second-order advantage. The program offers practical tools for data loading and visualization, model development and validation, and prediction in unknown samples. The software output also includes a comprehensive report on the AFOMs applicable to second-order calibration models i.e., sensitivity, analytical sensitivity, selectivity, limit of detection, limit of quantitation and prediction uncertainty, following IUPAC recommendations [4].

In particular, the program also offers the possibility of estimating the extent of rotational ambiguity (RA) and its impact on the prediction uncertainty for a given test sample in MCR-ALS calibration models [5]. In the present report, the new features of MVC2 are illustrated through simulated examples and real calibration datasets. For the readers interested in testing and using the application it is possible to do so on a fully functional application running on a public available server.The GUI is freely available, along with examples and manual, and it can be downloaded from www.iquir-conicet/descargas/mvc2.zip

---

Acknowledgments

The authors are grateful to Universidad Nacional del Litoral (Project CAI+D 2020- 50620190100020LI), Ministerio de Ciencia e Innovación of Spain Project PID2020-661 112996GB-I00, and ANPCyT (Agencia Nacional de Promoción Científica y Tecnológica, Projects PICT 2020-0105 and 2020-00179) for financial support.

References

[1] Olivieri AC, Wu HL, Yu, RQ **(2009)** MVC2: A MATLAB graphical interface toolbox for second-order multivariate calibration, *Chemom. Intell. Lab. Syst.*, **96**, 246-251

[2] Kiers HAL, ten Berge JMF, Bro R **(1999)** PARAFAC2-Part I. A direct fitting algorithm for the PARAFAC2 model, *J. Chemom.*, **13**, 275-294

[3] Tauler R, de Juan A **(2015)** Multivariate Curve Resolution for Quantitative Analysis, Chapter 5, in: Muñoz de la Peña A, Goicoechea HC, Escandar GM, Olivieri AC (Eds.) Fundamentals and Analytical Applications of Multiway Calibration, Elsevier, Boston, pp. 247-292

[4] A.C. Olivieri, **(2014)** Analytical Figures of Merit: From Univariate to Multiway Calibration, *Chem. Rev.*, **114**, 5358-5378

[5] Olivieri AC **(2022)** Evaluation of the ambiguity in second-order analytical calibration based on multivariate curve resolution. A tutorial, *Microchem. J.*, **179**, 107455

Calcium, magnesium, and potassium determination in soy and corn leaves combining laser-induced breakdown spectroscopy (LIBS) direct analysis and partial least squares (PLS)

João Victor Borges Assis [1,2], Dennis da Silva Ferreira[1], Daniela de Assis Bócoli[2],
Carlos Henrique Hoff Brait[2], Edenir R. Pereira-Filho[1]

Addresses: [1] Group of Analytical Instrumental Analysis (GAIA), Chemistry Department, Federal University of São Carlos (UFSCar), São Carlos, São Paulo 13565-905, Brazil
[2] Exata Brasil, Jataí, GO, 75802-005, +55 64 3631-7774, ped@exatabrasil.com.br, Brazil.
E-mail erpf@ufscar.br

Plant nutrition is a critical aspect of evaluating a specific cultures' productivity. The determination of some macronutrients such as calcium, magnesium, and potassium in leaves is performed after acid digestion to convert the solid sample into a homogeneous aqueous solution. This procedure is time-consuming, and several experimental errors can be introduced. This study proposes a direct solid sample method combining laser-induced breakdown spectroscopy (LIBS) and partial least squares (PLS). Soy (with and without petiole) and corn leaves were collected, dried, milled, and pressed to obtain pellets around 15 mm in diameter. After acid digestion, reference concentrations for Ca, Mg, and K were obtained using inductively coupled plasma optical emission spectrometry (ICP OES), and a total of 896 samples were analyzed. An Applied Spectra (model J200) LIBS system was employed for emission spectra acquisition, and around 300 spectra (from 186 up to 1042 nm) were obtained for each sample. The LIBS system configuration was 70 mJ pulse energy, 1.0 µs delay time, and 100 µm spot size [1]. Twelve normalization modes [2] were tested, and calibration models for each type of sample and analyte were proposed using the signal area for some selected emission lines of the analytes. The noteworthy normalization mode for most models was the Euclidean norm. The standard error of calibration (SEC) varied from 0.5 g/kg (Mg in corn leaves with 2 latent variables) up to 4.9 g/kg (K in soy with petiole leaves with 2 latent variables). The proposed calibration models have been tested in validation and prediction data sets with promising results for direct plant routine analysis. The authors are grateful to Fapesp, CNPq (process number 302719/2020-2), and Capes (finance code 001).

References

[1] D. F. Andrade, E. R. Pereira-Filho, P. Konieczynski (**2017**) Comparison of ICP OES and LIBS analysis of medicinal herbs rich in flavonoids from Eastern Europe. *J. Braz. Chem. Soc.*, **28**, 838-847.

[2] J. P. Castro, E. R. Pereira-Filho (**2016**) Twelve different types of data normalization for the proposition of classification, univariate and multivariate regression models for the direct analyses of allows by laser-induced breakdown spectroscopy (LIBS). *J. Anal. At. Spectrom.*, **31**, 2005-2014.

GC-IMS chromatogram alignment using the retention indexes of Analytical Standards

E. Caballero[1], L. Fernández[1,2], S. Oller-Moreno[1,2], C Mallafré-Muro[1,2], A. Pardo[2], S. Marco[1,2].

[1]Institute for Bioengineering of Catalonia -The Barcelona Institute of Science and Technology (IBEC-BIST), Signal and Information Processing for Sensing Systems, Baldri Reixac 10-12, Barcelona, 08028, Spain
[2]Department of Electronics and Biomedical Engineering, Universitat de Barcelona, Marti i Franqués, Barcelona, 08028, Spain
ecaballero@ibecbarcelona.eu

Gas-Chromatography coupled with Ion Mobility Spectrometry (GC-IMS) is a fast and sensitive analytical tool to extract chemical information from complex volatile samples. GC-IMS has been widely used in the food industry [3] and recently has been explored in biomedical applications [6] because of its ability to separate volatile organic compounds, fast response, easy operation, and portability [5]. Despite the advantages it presents it still presents several drawbacks, such as non-linearities, baseline drift, misalignments among samples, and peak overlaps, among others [4]. Correct chromatogram alignment is crucial to track data features among samples. Otherwise, subsequent data processing results would not be as reliable. There are several ways to align samples such as COW, PTW [1] and icoshift [2]. These techniques depend on parameter optimization that must be performed by an expert and can be computationally costly. To avoid this optimization, we propose a piecewise alignment with a set of reference peaks, we use a mixture of 6 ketones (2-butanone, 2-pentanone, 2-hexanone, 2-heptanone, 2-octanone and 2-nonanone) with known retention indexes. To perform this alignment the retention time of each ketone is automatically recorded and matched to its retention index, and then with an interpolation we can obtain the new spectrum. To test this technique a set of 11 Samples were used, the samples were collected with a GC-IMS a FlavourSpec® from G.A.S. Dortmund (Dortmund, Germany), the samples were incubated 15 minutes at 60^{O}C, and the GC and the IMS temperatures were also set at 60^{O}C finally the flow rate for the drift gas and the carrier gas was set at 200ml/min and 11 ml/min respectively. We can see the results of this alignment in Figure 1, where we can see that the dispersion in each peak is reduced.

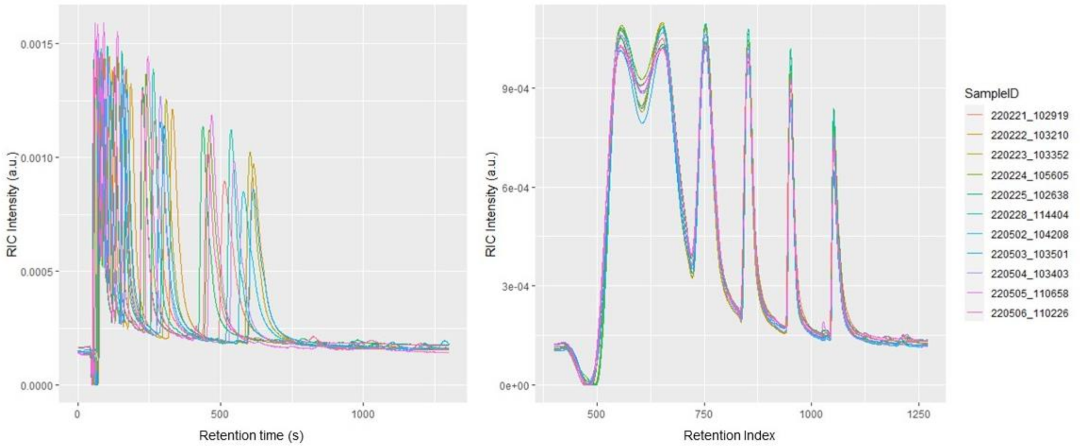


Figure 1. Reverse Ion Chromatogram of 11 ketone mixture samples before and after alignment.

References

[1] Drees,C. *et al.* (2019) GC-IMS headspace analyses allow early recognition of bacterial growth and rapid pathogen differentiation in standard blood cultures. *Appl Microbiol Biotechnol*, **103**, 9091–9101.

[2] Giorgio Tomasi; Francesco Savorani; Søren B. Engelsen (2011). icoshift: An effective tool for the alignment of chromatographic data. , 1218(43), 7832–7840. doi:10.1016/j.chroma.2011.08.086

[3] L. Fernandez, A. Martín-Gómez, M. Mar Contreras, M. Padilla, S. Marco and L. Arce (2017), Ham quality evaluation assisted by gas chromatography ion mobility spectrometry. *ISOCS/IEEE International Symposium on Olfaction and Electronic Nose (ISOEN)*, Montreal, QC, Canada, 2017, pp. 1-3, doi: 10.1109/ISOEN.2017.7968852.

[4] L. Fernandez, A. Blanco, C. Mallafré-Muro and S. Marco (2022), Towards batch correction for GC-IMS data. *IEEE International Symposium on Olfaction and Electronic Nose (ISOEN)*, Aveiro, Portugal, 2022, pp. 1-3, doi: 10.1109/ISOEN54820.2022.9789646.

[5] Wang, Shuqi; Chen, Haitao; Sun, Baoguo (2020). Recent progress in food flavor analysis using gas chromatography–ion mobility spectrometry (GC–IMS). Food Chemistry, (), 126158–. doi:10.1016/j.foodchem.2019.126158

[6] Wei Jiang, Zhi-Min Zhang, YongHuan Yun, De-Jian Zhan, Yi-Bao Zheng, Yi-Zeng Liang, Zhen Yu Yang, Ling Yu (2013). Comparisons of Five Algorithms for Chromatogram Alignment. *, 76(17-18), 1067–1078.* doi:10.1007/s10337-013-2513-8

NIR SPECTROSCOPY AND CHEMOMETRICS FOR THE IDENTIFICATION OF ADULTERATED OREGANO SAMPLES

Giorgia Foca[1,2], Camilla Menozzi[1], Veronica Ferrari[1], Rosalba Calvini[1,2], Marco Bragolusi[3], Alessandra Tata[3], Alessandro Ulrici[1,2]

[1]Università di Modena e Reggio Emilia, Dipartimento di Scienze della Vita, Via Amendola 2, Reggio Emilia, 42122, Italy
[2]Università degli Studi di Modena e Reggio Emilia, Centro Interdipartimentale BIOGEST-SITEIA, Piazzale Europa 1, Reggio Emilia, 42122, Italy
[3]Laboratorio di Chimica Sperimentale, Istituto Zooprofilattico Sperimentale delle Venezie, Viale Fiume 78, Vicenza, 36100, Italy
giorgia.foca@unimore.it

Oregano is the herb with the greatest risk of adulteration (Maquet, 2021), which usually consists of the addition of undeclared plant material of lesser commercial value, such as olive, hazel and myrtle leaves. To tackle this issue, the European Spice Association suggests carrying out regular checks to verify the purity of the product through non-targeted analytical techniques combined with chemometric methods for data processing (European Spice Association, 2018). Among these, NIR spectroscopy has found wide application in the identification of food fraud also in herbal and spice-based products thanks to its numerous advantages: reduced costs, rapidity in carrying out the analysis and limited or no sample preparation (Kucharska-Ambrożej, 2020; Massaro, 2023).

There are various types of NIR sensors on the market, which differ in size, acquisition methods and cost. The choice of the optimal configuration also depends on the nature of the sample: for powders with fine particle size and homogeneous distribution it may be sufficient to carry out punctual acquisitions with bench or portable NIR spectrometers, while for samples consisting of coarse and heterogeneous particles, such as dried oregano leaves, it is more appropriate to use imaging systems (NIR-HSI) which provide both spatial and spectral information. However, NIR-HSI systems are expensive, while portable NIR spectrometers are much cheaper and easier to handle even if they generally have lower performance as compared to benchtop instruments.

The main goal of this work is to develop multivariate models for the identification of adulterants in oregano leaves starting from the spectral data acquired with three NIR systems: benchtop NIR spectrometer, portable NIR spectrometer and NIR-HSI. The comparison of the performances obtained from the various instruments is aimed at i.) evaluating the minimum quantity of adulterant detectable with the various acquisition systems, ii.) identifying the most significant spectral variables for adulteration detection and, more generally, iii.) obtaining useful information on the cost-benefit ratio of the three types of NIR sensors when the analysis involves coarse-grained samples.

References

Adulteration Awareness Document, European Spice Association (**2018**) https://www.esa-spices.org/download/esa-adulteration-awareness-document2

Kucharska-Ambrożej K, Karpinska J (**2020**) The application of spectroscopic techniques in combination with chemometrics for detection adulteration of some herbs and spices. *Microchemical Journal*, **153**, 104278.

Maquet A, Lievens A, Paracchini V, Kaklamanos G, De La Calle Guntinas MB, Garlant L, Papoci S, Pietretti D, Ždiniaková T, Breidbach A, Omar Onaindia J, Boix Sanfeliu A, Dimitrova T, Ulberth F (**2021**) Results of an EU wide coordinated control plan to establish the prevalence of fraudulent practices in the marketing of herbs and spices, EUR 30877 EN, *Publications Office of the European Union*, Luxembourg, JRC126785.

Massaro A, Bragolusi M, Tata A, Zacometti C, Lefevre S, Frégière-Salomon A, Lafeuille JL, Sammarco G, Fiordaliso Candalino I, Suman M, Piro R (**2023**) Non-targeted authentication of black pepper using a local web platform: Development, validation and post-analytical challenges of a combined NIR spectroscopy and LASSO method, *Food Control*, **145**, 109477.

J-Score: a new joint parameter for PLSR model performance evaluation

<u>Jokin Ezenarro</u>, Daniel Schorn-García, Laura Aceña, Montserrat Mestres, Olga Busto, Ricard Boqué

Universitat Rovira i Virgili, Department of Analytical Chemistry and Organic Chemistry, Campus Sescelades, Edifici N4, C/Marcel·lí Domingo 1, Tarragona, 43007, Spain
jokin.ezenarro@urv.cat

Partial Least Squares Regression (PLSR) is a widely used multivariate data analysis technique that has gained popularity due to its simplicity, versatility, and applicability in various fields, especially in spectroscopy. This method is used to build a regression equation that fits multiple response variables in a single model by finding new dimensions or Latent Variables (LVs) from the original data. PLSR maximizes the covariance between the response variables (**X**) and the value to be predicted (**Y**), reducing the dimensionality and making the interpretation of the results more intuitive (Geladi & Kowalski, 1986).

Nevertheless, data analysis needs to be supervised by an expert analyst with enough experience to build regression models and ensure good performance, by choosing the right spectral preprocessing and the rank of the model, that is, the optimal number of LVs. To decide if the regression model is adequate, the analyst must choose the most suitable quality assessment parameters among the existing ones. Some of these parameters account for the same properties of the model, while some are complementary (Clark, 2003). For example, typical properties of the model are the shape of the curves of RMSE of Calibration ($RMSE_{Cal}$), Cross Validation ($RMSE_{CV}$), or Prediction ($RMSE_{P}$) vs. the number of LVs, the values of Explained Variance (EV) for each LV, the shape of the loadings and the Regression Vector (RV), or the plot of Q residuals vs Hotelling T^2 values. However, the adequacy of these values and plots is usually decided based on the opinion of the analyst, which comes from experience and training. Sometimes, this may not be objective enough, and more quantitative tools are needed to assess the performance of the model (Liemohn et al., 2021).

In this work, a novel tool is proposed to quantitatively evaluate the overall performance of PLSR models for spectroscopic data and potentially other sources of continuous data.

This tool considers several model properties and describes their suitability in a global way. The proposed tool enables non-experienced analysts to choose the optimal number of LVs and the adequate data preprocessing. It is calculated by converting some of the aforementioned parameters into numerical indexes and then combining them into a general score, the J-score, which can be used for the performance assessment of a model or for the comparison between models in a fast and objective way.

In summary, the proposed tool is especially useful for non-expert analysts, helping them to evaluate the performance of PLSR models and can be easily implemented in existing software. The tool has been shown to be efficient and effective on real-world spectroscopic data, even if further improvements and extensions can be made in future work.

References

Clark, R. D. (2003). Boosted leave-many-out cross-validation: the effect of training and test set diversity on PLS statistics. *Journal of Computer-Aided Molecular Design*, *17*, 265–275.

Geladi, P., & Kowalski, B. R. (1986). PARTIAL LEAST-SQUARES REGRESSION: A TUTORIAL. *Analytica Chimica Acta*, *186*.

Liemohn, M. W., Shane, A. D., Azari, A. R., Petersen, A. K., Swiger, B. M., & Mukhopadhyay, A. (2021). RMSE is not enough: Guidelines to robust data-model comparisons for magnetospheric physics. *Journal of Atmospheric and Solar-Terrestrial Physics*, *218*, 105624. https://doi.org/10.1016/J.JASTP.2021.105624

expertPLS: a MATLAB toolbox for spectral preprocessing selection

Jokin Ezenarro, Daniel Schorn-García, Laura Aceña, Montserrat Mestres, Olga Busto, Ricard Boqué

Universitat Rovira i Virgili, Department of Analytical Chemistry and Organic Chemistry,
Campus Sescelades, Edifici N4, C/Marcel·lí Domingo 1, Tarragona, 43007, Spain
jokin.ezenarro@urv.cat

Multivariate data analysis is an indispensable tool when working with spectroscopic techniques as it helps extract the information of interest from the data obtained. Often data preprocessing is needed for this purpose, as its goal is to remove any unwanted systematic variation from the spectra, such as baseline shifts and scatter effects, while retaining the information of interest. Systematic variation in the descriptor matrix **X** (spectra) that is independent of the chemical variation in the response matrix **Y** (value to predict) introduces modeling problems for subsequent multivariate projection and regression methods such as Partial Least Squares (PLS) and negatively affects prediction and interpretation (Jonsson & Gabrielsson, 2009).

Traditionally, choosing the appropriate preprocessing steps involves trial-and-error approaches, where the prediction uncertainty (Root Mean Square Error) or other metrics can be used to determine the best preprocessing method. However, as datasets become larger and preprocessing and analysis algorithms more complex, the computation time required to iterate all available options becomes non-practical. To overcome this challenge, various approaches have been proposed, such as Design of Experiment (Gerretzen et al., 2015), orthogonalization (SPORT) (Roger et al., 2020), or spectral signal-to-error ratio (Skibsted et al., 2004).

Instead, the present research introduces a new approach to filtering preprocessing algorithms based on the properties of the raw data. The proposed method creates rules for selection criteria that an expert analyst would use, but in an automated and objective way, eliminating the need for trial-and-error methods. The MATLAB toolbox presented quantifies the noise in the data and the multiplicative and additive scatter-effects, to decide which preprocessing algorithms should be iterated to correct the spectra. The toolbox helps the user choose the best preprocessing methodology for their data, resulting in optimal PLS models that are better than what a non-expert analyst would obtain by trial-and-error and in considerably less time.

References

Gerretzen, J., Szymańska, E., Jansen, J. J., Bart, J., van Manen, H. J., van den Heuvel, E. R., & Buydens, L. M. C. (2015). Simple and Effective Way for Data Preprocessing Selection Based on Design of Experiments. *Analytical Chemistry*, *87*(24), 12096–12103. https://doi.org/10.1021/acs.analchem.5b02832

Jonsson, H., & Gabrielsson, J. (2009). Evaluation of Preprocessing Methods. In S. D. Brown, B. Walczak, & R. Tauler (Eds.), *Comprehensive Chemometrics* (Vol. 2, pp. 199–206). Elsevier.

Roger, J. M., Biancolillo, A., & Marini, F. (2020). Sequential preprocessing through ORThogonalization (SPORT) and its application to near infrared spectroscopy. *Chemometrics and Intelligent Laboratory Systems*, *199*. https://doi.org/10.1016/j.chemolab.2020.103975

Skibsted, E. T. S., Boelens, H. F. M., Westerhuis, J. A., Witte, D. T., & Smilde, A. K. (2004). New Indicator for Optimal Preprocessing and Wavelength Selection of Near-Infrared Spectra. *Applied Spectroscopy*, *58*(3), 264–271. https://doi.org/10.1366/000370204322886591

NIRS MEASUREMENT OF SOIL ORGANIC CARBON BY USING A COMPOSITIONAL DATA METHOD. AN APPROACH TO AVOID MOISTURE INTERFERENCE

José A. Cayuela-Sánchez[1], Rafael López-Núñez[2]

[1]Instituto de la Grasa, Dpt. Biochemistry and Molecular Biology of Plant Products, Spanish Council for Scientific Research (CSIC), Campus Universidad Pablo de Olavide, Ed. N^{O} 46, Ctra. Utrera, Km 1, 41013, Sevilla, Spain

[2]Instituto de Recursos Naturales y Agrobiología, Dpt. Protection of the Soil, Plant, Water System, Spanish Council for Scientific Research, (CSIC), Avda. Reina Mercedes 10, 41012, Sevilla, Spain

jacayuela@ig.csic.es

Soil organic carbon (SOC) is an indicator of soil organic matter content, which has an important role in modulating atmospheric CO_2. Soil components are interdependent, that means no single soil component is free to vary independently from the rest and carry only relative information. It is, soil properties have a compositional nature. Thus, VisNIRS correlations with compositional reference data needs specific compositional data (CoDa) methods (Aitchison 1986; Pawlowsky-Glahn et al. 2015). The soil moisture variation supposes a handicap to VisNIRS measurement of SOC. Considering moisture as a soil component, a compositional model can integrate the moisture prediction along with the other components considered, as soil organic carbon content (SOC). The interference of moisture content in determining SOC could therefore be avoided by using a compositional model. The preliminary calibrations from four ILR (isometric logarithmic coordinates as defined by Egozcue et al., 2003), implying as soil components moisture (SM), SOC, inorganic carbon (NOC), the textural fractions 'clay' and 'silt', and the rest of soil mass 'sand+rest', were quite fine, with R^2 0.94, 0.87, 0.81 and 0.79 and R^2_{cv} 0.86, 0.77, 0.71, 0.67. Therefore, this method to estimate SOC could be satisfactory. More research is ongoing to verify this approach within the frame of the ProbeField project (https://ejpsoil.eu/soil-research/probefield), granted by the EJP-Soil program, focused on proximal sensing techniques for SOC measurement.

References

Aitchison, J. 1986. The statistical analysis of compositional data, Chapman and Hall: London, 1986.

Egozcue, J. J., Pawlowsky-Glahn, V., Mateu-Figueras, G., Barceló-Vidal, G. 2003. Isometric Logratio Transformations for Compositional Data Analysis, Math Geol. 35, 279-300. https://doi.org/10.1023/A:1023818214614

Pawlowsky-Glahn, V., Egozcue, J. J., Tolosana-Delgado, R. 2015. Modeling and analysis of compositional data. Wiley&Sons, Chichester.

DRUG EXPIRATION STUDY USING RAMAN SPECTROSCOPY AND SUPER-PARAMAGNETIC CLUSTERING

González-Solís JL[a], Anaya-Martin E[a], Martínez-Zérega BE[a],
Oseguera-Galindo DO[b], Torres-González LA[c], Guizar-Ruiz JI[a]

[a]*Biophysics and Biomedical Sciences Laboratory,*
Centro Universitario de Lagos, Universidad de Guadalajara
Enrique Díaz de León S/N, Paseo de la Montaña, 47460, Lagos de Moreno, Jal., México.
[b]*Departamento. Cs. Naturales y Exactas,*
Centro Universitario de los Valles, Universidad de Guadalajara, Carretera Guadalajara - Ameca Km. 45.5, Ameca 46600, Jalisco, México
[c]*Departamento de Ingeniería, Universidad Iberoamericana León,*
Blvd. Jorge Vértiz Campero, Fracciones Canadá de Alfaro, 37238, León, Guanajuato, México
jluis0968@gmail.com (corresponding author)

Currently, there is a belief that drugs used after the expiration date no longer work or even these can cause some type of damage. In the present work, the expiration date of several types of drugs on the market is analyzed, to find out if these, with the expired expiration date, present spectroscopic differences with those drugs whose expiration date is still valid. For this study, the Raman spectroscopy technique and Super-Paramagnetic Clustering (SPC) method were used to analyze spectral differences (or chemical differences) between drugs. The Raman spectroscopy technique is the best technique to study the chemical composition of a substance or material. To measure the drug Raman spectra, Horiba equipment, LabRam HR800, was used with an Olympus confocal microscope focusing a laser of 830 nm and 17 mW on the drugs through a 100X Leica long-range objective [1]. In the SPC method, used to classify Raman spectra, the clustering process is based on a phenomenon of clustering observed in nature at the atomic level and perfectly described by a statistical physics model known as the Potts model, which describes the interacting spins on a crystalline lattice [2]. In this novel method, a Potts spin is assigned to each data point (Raman spectrum) and introduces an interaction between neighboring points, whose strength or coupling is a decreasing function of the distance between the nearest neighboring sites [3-5]. This clustering method allows identifying hierarchical structures in the spectra data banks. Thirteen drugs were analyzed, including capsules, tablets, syrups, ointments, and injections with 1 to 3 expired expiration dates. Comparing drugs with expiration dates still valid and expired, the SPC method results showed that, although some drugs clearly indicated chemical differences (formation of different clusters), others indicated no existence of chemical differences (formation of a single cluster), even among those with up to two expired expiration dates. The results show that SPC is an excellent mathematical tool for pattern recognition in databases (spectra), allowing to discrimination of expired drugs from non-expired ones. To the best of our knowledge, this is the first report of preliminary results evaluating the usefulness of Raman spectroscopy and SPC in the study of the expiration of drugs distributed on the market.

References

1. Anaya-Martin E (**2022**) Estudio de la caducidad de medicamentos usando espectroscopia Raman y Análisis de Componentes Principales [Bachelor's thesis, Centro Universitario de los Lagos, Universidad de Guadalajara].
2. Blatt M, Wiseman S, and Domany E (**1996**) Superparamagnetic clustering of data. Physical Review Letters, **76 (18)**, 3251–3254.
3. González-Solís JL, Torres-González LA and Villafán-Bernal JR (**2019**) Superparamagnetic Clustering of Diabetes Patients Raman Spectra. Journal of Spectroscopy, **2019**, 4296153.
4. González-Solís JL (**2019**) Discrimination of different cancer types clustering Raman spectra by a super paramagnetic stochastic network approach. PLoS ONE, **14(3)**, e0213621.
5. González-Solís JL, Guizar-Ruiz JI, Martínez-Espinosa JC, Martínez-Zerega BE, Juárez-López HA, Vargas-Rodríguez H, Gallegos-Infante LA, González-Silva RA, Espinoza-Padilla PB, Palomares-Anda P (**2016**) Cancer detection based on Raman spectra super-paramagnetic clustering. Physica A, **455**, 52-64.

Emerging FTIR-chemometric approach for ALS patients' discrimination based on selected spectra biomarkers

J.M González-Sáiz[1], K. Tkachenko[1], I. Esteban-Díez[1], Ana C. Calvo[2], R. Osta[2], C. Pizarro[1]

[1]University of La Rioja, Department of Chemistry, Madre de Dios, 53, (E26006) Logroño, La Rioja (Spain)

[2]Instituto de Investigación Sanitaria Aragón (IISA), C. de San Juan Bosco, 13, 50009 Zaragoza, Aragón, (Spain)

josemaria.gonzalez@unirioja.es

Amyotrophic lateral sclerosis (ALS) continues to be a challenging neurodegenerative disease lacking reliable biomarkers and early diagnosis (1). Moreover, despite the progress in discovering genetic mutations, misdiagnosis with other-neurodegenerative diseases still occurs. In this study, we used human supernatant blood to obtain infrared spectroscopic signatures of ALS and other motor-related neuro disorders, comparing them to healthy controls. Since blood reflects many metabolic changes and disarrangements in the organism, spectral differences associated with health and disease patients or ALS-specific pathological events could be readily detectable. The recorded spectra classification analysis was performed to establish apparent differences between the groups (2). Our results showed that FTIR spectroscopy is a powerful tool in discriminating between ALS and controls, providing high specificity for ALS (100%) and other neuropathies (98%). In addition, the results after six months of ALS progression are discussed. Overall, the study suggests that infrared spectroscopy could be a promising diagnostic tool for ALS and other neurodegenerative diseases.

References

1. Xu, R. S.; Yuan, M. **(2021)** Considerations on the Concept, Definition, and Diagnosis of Amyotrophic Lateral Sclerosis. *Neural Regen Res*, **16** (9), 1723–1729.
2. Tkachenko, K.; Espinosa, M.; Esteban-Díez, I.; González-Sáiz, J. M.; Pizarro, C. **(2022)** Extraction of Reduced Infrared Biomarker Signatures for the Stratification of Patients Affected by Parkinson's Disease: An Untargeted Metabolomic Approach. *Chemosensors*, ***10*** (6) 229.

Identification of lipidomic traits in plasma samples for the discrimination of patients affected by Parkinson's disease: UPLC-MS untargeted approach

K. Tkachenko, C. Pizarro, I. Esteban-Díez, J.M González-Sáiz

University of La Rioja, Department of Chemistry, Madre de Dios, 53, (E26006) Logroño, La Rioja (Spain)
consuelo.pizarro@unirioja.es

CRYPTED FOR CONFIDENTIALITY REASON

ADJUSTED PARETO SCALING

Kurt Varmuza, Peter Filzmoser

Vienna University of Technology, Institute of Statistics and Mathematical Methods in Economics, Wiedner Hauptstrasse 8-10/E105/CSTAT, Vienna, 1040, Austria
kurt.varmuza@tuwien.ac.at

The performance of multivariate models in chemometrics often depends on the applied method of scaling the x-variables. Autoscaling and Pareto scaling are widely used; here an *adjusted Pareto scaling* is suggested – covering the range from no scaling via classical Pareto scaling to autoscaling.

(1) *Autoscaling* of a variable is performed by x_c / s with x_c for the centred original variable, and s the standard deviation of the variable. (2) *Pareto scaling* is performed by $x_c / s^{0.5}$. The scaling effect is weaker than with autoscaling, noise is less amplified, and variables with a high original variance retain part of their importance for the model. (3) *Adjusted Pareto scaling* is performed by x_c / s^{P} with P varying between 0 (no scaling) and 1 (autoscaling), typically in steps of 0.1, thus including classical Pareto scaling with $P = 0.5$. An optimum P can be selected by evaluating the model performances by repeated double cross validation [1].

Adjusted Pareto scaling has been applied together with PLS regression for various data sets: (A) Heating value of biomass modelled by elemental content data. (B) GC retention indices modelled by molecular descriptors. (C) Glucose content of fermentation samples modelled by NIR absorbances. Results indicate that classical Pareto scaling often can be improved by searching for an optimum exponent P.

References

[1] Filzmoser P, Liebmann B, Varmuza K (**2009**) Repeated double cross validation. *Journal of Chemometrics,* **23**, 160-171.

Establishing applicability limits of regression neural networks using decoders. Application to predict diesel properties using infrared spectra

M. Suliany Rodríguez-Barrios[1], Joan Ferré[1], M. Soledad Larrechi[1],
Enric Ruiz[2]

Addresses: [1]Universitat Rovira i Virgili, Department of Analytical and Organic Chemistry, Faculty of Chemistry, Campus Sescelades, Carrer Marcel·lí Domingo, 1, Tarragona, 43007, Spain.
[2]Repsol-Petróleo, Tarragona, Spain
E-mail: mariasuliany.rodriguez@urv.cat

Knowledge of the applicability domain (AD) of regression models based on artificial neural networks (ANNs) is a requisite for applying ANNs in routine analyses using spectroscopic data. The AD limits can be defined from different metrics that condition the confidence of the predictions of the established regression model.

In multivariate regression methods such as the Partial Least Squares (PLS) calibration, the applicability limits are commonly based on Hotelling T^2 and Q statistics [1,2]. These limits allow the detection of samples that are rare compared to those used to establish the model. Other limits based on criteria such as ASTM's RMSSR (Root Mean Square Spectral Residuals) and NND (Nearest Neighbor Distance) [3] have also been used to flag spectra outside the established limits as discordant spectra. Although measures based on the similarity among spectra apply to all types of models, those that consider the specific form of the model, such as Hotelling's T^2 and Q statistics, are preferred, since they are related to how the spectrum is being used by the model. A similar system for defining the limits of applicability of multivariate regression based on ANNs has not been reported yet. This work proposes a methodology to establish the limits of applicability of regression neural networks and shows its implementation for the prediction of a diesel property using infrared spectra.

A test set was created by randomly selecting samples from a data set of 2172 diesel samples. The rest of the samples were split into training and validation sets using the Kennard-Stone algorithm. A feed-forward neural network (FFNN) was trained to predict the density of the diesel samples from their infrared spectra. The activations of the hidden layer of the FFNN were used to train a decoder network to reconstruct the training spectra. The squared Mahalanobis distance (MD^2) of the hidden layer activations and the spectral residuals (Q residuals) from the decoder network were used to define the applicability limits of the calibration model. The FFNN model provided a high determination coefficient between real and predicted density values (R^2 = 0.99), with low prediction errors (RMSEP = 0.72 kg/m^3) comparable to other reported results [4-6]. The warning limits for the MD^2 and Q residuals were decided from the empirical cumulative distribution function of both metrics. These warning limits define the applicability domain of the FFNN regression model. Beyond these limits, the spectrum of a new sample is flagged as atypical.

References

[1] J. F. MacGregor and T. Kourti, "Statistical process control of multivariate processes," Control Eng. Pract., 3 (1995) 403–414.
[2] R. P. Cogdill, C. A. Anderson, and J. K. Drennen, "Process analytical technology case study, part III: Calibration monitoring and transfer," AAPS PharmSciTech, 6 (2005) 284-297.
[3] ASTM E1655-17, 2017. ASTM E1655-17 standard practices for infrared multivariate quantitative analysis. ASTM Int. 05, 30
[4] R. M. Balabin, E. I. Lomakina, and R. Z. Safieva, "Neural network (ANN) approach to biodiesel analysis : Analysis of biodiesel density, kinematic viscosity, methanol, and water contents using near-infrared (NIR) spectroscopy," *Fuel*, 90 (2010) 2007–2015, doi: 10.1016/j.fuel.2010.11.038.
[5] R. M. Balabin and S. V. Smirnov, "Variable selection in near-infrared spectroscopy: Benchmarking of feature selection methods on biodiesel data," *Anal. Chim. Acta*, 692 (2011) 63–72, doi: 10.1016/j.aca.2011.03.006.
[6] L. de Fátima Bezerra de Lira, F. V. C. de Vasconcelos, C. F. Pereira, A. P. S. Paim, L. Stragevitch, and M. F. Pimentel, "Prediction of properties of diesel/biodiesel blends by infrared spectroscopy and multivariate calibration," *Fuel*, 89 (2010) 405–409, doi: 10.1016/j.fuel.2009.05.028.

Developing a Local RoBoost PLS framework

Daniele Tanzilli[1,2], Lorenzo Strani[1], Maxime Metz[3], Maxime Ryckewaert[4,5], Jean Michel Roger[4,5], Matthieu Lesnoff[5,6], Marina Cocchi[1]

Addresses: [1]University Modena and Reggio Emilia, Department of Chemical and Geological Sciences, Via Campi 103, Modena, 41125, Italy
[2]Université de Lille, LASIRE, Cité Scientifique, Villeneuve-d'Ascq, 59650, France
[3]Pellenc ST, Pertuis, Provence-Alpes-Côte d'Azur, France
[4]ITAP, Univ. Montpellier, INRAE, Institut Agro, Montpellier, France
[5]ChemHouse Research Group, Montpellier, France
[6]UMR SELMET, Univ Montpellier, CIRAD, INRA, Institut Agro, Montpellier, France
E-mail marina.cocchi@unimore.it

Recently, some of us proposed a novel RoBoost PLS algorithm to assess the relevance of different samples in PLS calibration based on a weighting scheme defined according to three criteria calculated for each latent variable: X-residuals, Y-residuals and leverage. The method (extended also to multiple responses) [1] has proven to be efficient to detect both X and Y outliers and obtaining more robust models. Here we aim extending the methodology also to local regression scenario [2], and multiblock-PLS. In particular, by using simulated spectroscopic data [3] we investigated different situations ranging from clear presence of grouping to smooth transitions, such as in process monitoring. In both cases, local regression is best suited with respect to standard regression, however the influent/outlier samples may be of different nature (strong outlier for the overall data set, outlier for specific group, a combination of both situation) and have a different impact on the model robustness.

Since two weighting schemes are combined, i.e. the Roboost and the local, several parameters need to be tuned, thus a double cross-validation is implemented. The code has been developed in MATLAB environment.

References

[1] M. Metz, F. Abdelghafour, J.-M. Roger, M. Lesnoff, Analytica Chimica Acta (2021) 338823. doi:434 10.1016/j.aca.2021.338823.
[2] M. Metz, M. Lesnoff, J.-M. Roger, J. Chemometrics (2020) e3209. https://doi.org/10.1002/cem.3209
[3] M. Metz, A. Biancolillo, M. Lesnoff, J.-M. Roger, A note on spectral data simulation, Chemom. Intell. Lab. Syst. (2020) 103979. doi:10.1016/j.chemolab.2020.103979.

EXTENDED MULTIVARIATE COMPARISON OF VALIDITY INDICES FOR CLUSTER ANALYSIS

R. Todeschini, V. Consonni, F. Gosetti, V. Termopoli, D. Ballabio

University of Milano-Bicocca, Department of Earth and Environmental Sciences P.zza della Scienza 1, Milano, 20126, Italy

E-mail (roberto.todeschini@unimib.it)

Clustering is an unsupervised machine learning methodology widely used in several research fields to find groups of similar patterns in complex data. The results generated by clustering algorithms generally depend on user-defined input parameters such as the number of expected groups, which can have a great impact on the homogeneity of the identified clusters.

Clustering validity indices (CVIs) are an effective method for determining the optimal number of clusters that best fits the natural partition of a dataset. They do not require any underlying assumption nor a priori knowledge about the real dataset structure. Since 1965, many cluster validity indices have been proposed in the literature and used in several different applications.

We evaluated the performance of 68 cluster validity indices on several real-life and simulated datasets. CVIs were compared on the same optimal partition for each dataset, which was previously searched for by the k-means clustering algorithm. Principal Component Analysis (PCA) and Minimum Spanning Tree (MST) methods were applied to disclose mutual relationships among the indices and to select those that are more promising in terms of accuracy and reliability.

Raman technology implementation PAT for Cell culture development: Chemometrics Raman models for accelerated and easier new process development

Gaelle Bordes[1], M. David Paul[2], M. Laëtitia Macon[1]

[1] *Sanofi / Cell Culture Development laboratory, 13 Quai Jules Guesde, 94400 Vitry sur Seine,*
Gaelle.Bordes@sanofi.com, Laetitia.Macon@sanofi.com
[2] *Ividata Life Sciences for Sanofi / CMC Data Sciences, 79 Rue Baudin, 92300 Levallois Perret,*
DavidOlivier.Paul-ext@sanofi.com

CRYPTED FOR CONFIDENTIALITY REASON

Calibration Transfer to Facilitate Knowledge Sharing and Speed Deployment

Art Hamfeldt[1], Lorenza Broccardo[2]

1 Sartorius North America Inc., 565 Johnson Avenue, Bohemia, NY 11716, USA
2 S-IN Soluzioni Informatiche srl, via Ferrari 14, 36100 Vicenza, Italy
lorenza.broccardo@s-in.it *(corresponding author)*

Transferring models between instruments or processes speeds deployment and facilitates knowledge sharing. Challenges often arise when transferring models between instruments and process equipment.

A methodology will be presented which will reduce the transfer burden while identifying key parameters helping mitigate common issues when transferring calibration or reactor models.

Quality prediction with sequential-orthogonalized PLS: an explorative method to capture process topology

Qiang Zhu[1,2], Pierantonio Facco[2], Zhonggai Zhao[1], Massimiliano Barolo[2]

[1] Key Laboratory of Advanced Process Control for Light Industry (Ministry of Education),
Institute of Automation, Jiangnan University, Wuxi 214122, P.R. China
[2] CAPE-Lab – Computer-Aided Process Engineering Laboratory, Department of Industrial
Engineering, University of Padova, 35131 Padova PD, Italy
E-mail (max.barolo@unipd.it)

In decades, numerous data-driven models have been extensively employed for quality prediction in modern process engineering, and sufficient process measurements are the basis for applying these models. Nevertheless, some available process information, such as the process topology (i.e., the process flow diagram), has been neglected in previous research, even if it would be highly desirable to include it in a regression model that is used for product quality prediction. Unfortunately, including process information in a mathematical model is quite challenging since fusing two types of data (measurements and graphical information) together is very complex[1]. In this research, a state-of-art structural model, sequential-orthogonalized PLS (SO-PLS)[2], is applied to address this problem and predict product quality. SO-PLS is an extension of a popular chemometric method, multi-block PLS, which can block the complex process system into several simple subdivisions for modeling[3]. Differing from multi-block PLS, SO-PLS has the potential to capture the process topology, track the information flow, and translate it into the block permutation[4]. The orthogonalization procedure, performed in a predefined block order, eliminates the overlapping information between the connecting blocks and retains the independent information of each block for quality prediction[5]. This extra "feature extraction" from orthogonalization empowers SO-PLS to provide a better prediction performance than conventional PLS for complex manufacturing processes including several unit operations and connections between them. The potential of the proposed SO-PLS model is demonstrated by a simple but representative numerical case.

References

1. Yang, F., Duan, P., Shah, S. L., & Chen, T. (**2014**). Capturing connectivity and causality in complex industrial processes. In *SpringerBriefs in Applied Sciences and Technology* Springer.
2. Menichelli, E., Almøy, T., Tomic, O., Olsen, N. V., & Næs, T. (**2014**). SO-PLS as an exploratory tool for path modelling. *Food Quality and Preference*, **36**, 122–134.
3. Smilde, A. K., Næs, T., & Liland, K. H. (**2022**). Multiblock data fusion in statistics and machine learning. In *Multiblock Data Fusion in Statistics and Machine Learning*. Wiley.
4. Næs, T., Romano, R., Tomic, O., Måge, I., Smilde, A., & Liland, K. H. (**2021**). Sequential and orthogonalized PLS (SO-PLS) regression for path analysis: Order of blocks and relations between effects. *Journal of Chemometrics*, **35**(10), 1–24.
5. Jia, R., Mao, Z., Wang, F., & He, D. (**2016**). Sequential and orthogonalized partial least-squares model based real-time final quality control strategy for batch processes. *Industrial & Engineering Chemistry Research*, **55**(19), 5654–5669.

Automated Exploratory Analysis of Spectroscopic Data for Raw Material Identification (RMID) in the Tyre Industry

Riccardo Voccio [a], Cristina Malegori [a], Paolo Oliveri [a], Marco Arimondi [b], Andrea Bernardi [b], Giorgio Luciano [c], Mattia Cettolin [b].

[a] *University of Genova, Department of Pharmacy (DIFAR), Viale Cembrano, 4. Genova, 16148, Italy.*
riccardo.voccio@edu.unige.it (corresponding author)
[b] *Pirelli Tyre SpA, Viale Piero e Alberto Pirelli, 25. Milano, 20126, Italy.*
[c] *Istituto di Scienze e Tecnologie Chimiche "Giulio Natta", Consiglio Nazionale delle Ricerche, SCITEC-CNR, Via De Marini 6. 16149, Genova, Italy.*

The correct identification of raw materials plays a key role in the quality control of the tyre industry production chain. Especially when dealing with a large number of raw materials having different chemical compositions, and with analysis performed by several operators in different plants, it is important to perform raw material identification (RMID) being as more accurate and automated as possible. In this context, near-infrared (NIR) spectroscopy is one of the most suitable techniques, as it is non-destructive, user-friendly and timesaving.

In the present study, performed in collaboration with Pirelli Tyre S.p.A., FT-NIR spectra (Perkin Elmer, Frontier NIR, 10000 – 4000 cm^{-1}) were acquired for most of the raw materials used for tyre production in two of the main Pirelli Tyre's factories.
Exploratory analysis by means of principal component analysis (PCA) was performed on the data with the aim to investigate data structures and groupings within samples. Due to the huge amount of data collected, it was necessary to automate the exploratory analysis through the creation of a dedicated routine in the Matlab ® (Version 2022b, The MathWorks, Inc.) environment. This automation strategy includes incoming controls for each spectrum acquired, with the aim of detecting anomalous data and for the transformation of the measurement scale (from reflectance to pseudo-absorbance). If there are measurement replicates, the average spectrum is then calculated. Supervised classification strategies were then applied on this data, in order to define the class

membership of raw material analyzed, in terms of both material specification and supplier.
With the aim to apply this strategy for the routine analysis performed in the context of quality assurance, a report including all steps of data import and processing is then generated.
A further improvement of this RMID protocol will be the development of predictive models in order to estimate chemical and physical characteristics of the raw material directly during the acceptance phase.

Once developed and validated, the proposed strategy will be implemented in a graphical user interface (GUI) created in the Python environment for an immediate application in the industrial production chain; its usage will not require the operator to have previous knowledge about spectroscopy and chemometrics.

PARTICIPANTS

Riccardo	Aigotti	riccardo.aigotti@unito.it	Università di Torino
Gorka	Albizu Martinez	gorka.albizu@ehu.eus	Universidad del País Vasco
Faisal	Algethami	falgethami@imamu.edu.sa	University of Riyadh
Eugenio	Alladio	eugenio.alladio@unito.it	Università di Torino
Jose Manuel	Amigo Rubio	josemanuel.amigo@ehu.eus	Universidad del País Vasco
Elia	Arnese Feffin	elia.arnesefeffin@phd.unipd.it	Università di Padova
Alejandra	Arroyo Cerezo	arroyoc@ugr.es	University of Granada
Eloisa	Bagnulo	e.bagnulo@unito.it	Università di Torino
Andrea	Balivo	andrea.balivo@unina.it	Università di Napoli Federico II
Davide	Ballabio	davide.ballabio@unimib.it	Università di Milano-Bicocca
Marcella	Barbera	marcella.barbera@unipa.it	Università di Palermo
Gianmarco	Barberi	gianmarco.barberi@unipd.it	Università di Padova
Issam	Barra	issam.barra@um6p.ma	Universitè Mohammed VI Polytechnique
Martina	Beese	martina.beese@uni-rostock.de	Universität Rostock
Barbara	Benedetti	barbara.benedetti@unige.it	Università di Genova
Nicolas	Bernard	nicolas.bernard@totalenergies.com	TotalEnergies OneTech
Pietro	Bertani	pietro.bertani@bracco.com	Bracco
Paolo	Berzaghi	paolo.berzaghi@unipd.it	Università di Padova
Alessandra	Biancolillo	alessandra.biancolillo@univaq.it	Università dell'Aquila
Rebecca	Birolo	rebecca.birolo@unito.it	Università di Torino
Flavia	Bisozzi	flavia.bisozzi@student.unisi.it	Università di Siena
Davide	Boldini	davide.boldini@tum.de	Technische Universität München
Ricard	Boqué	ricard.boque@urv.cat	Universitat Rovira i Virgili
Ane	Bordagaray	ane.bordagaray@ehu.eus	Universidad del País Vasco
Gaelle	Bordes	Gaelle.Bordes@sanofi.com	Sanofi
Joao Victor	Borges Assis	joao.victor@exatabrasil.com.br	Exata Brasil
Joan	Borras-Ferris	joaborfe@eio.upv.es	Universidad Politecnica de Valencia
Andrea	Botton	andrea.botton.1@phd.unipd.it	Università di Padova
Marion	Brandolini-Bunlon	marion.brandolini-bunlon@inrae.fr	Inrae
Lorenza	Broccardo	lorenza.broccardo@s-in.it	S-in
Michela	Burico	mburico@aboca.it	Aboca S.p.A.
Eduardo	Caballero Saldivar	ecaballero@ibecbarcelona.eu	IBEC
Maria	Cairoli	maria.cairoli@ru.nl	Radboud University
Marco	Calderisi	info@kode-solutions.net	Kode
M. Pilar	Callao	mariapilar.callao@urv.cat	Universitat Rovira i Virgili
Rosalba	Calvini	rosalba.calvini@unimore.it	Università Modena/Reggio Emilia
Josè	Camacho	josecamacho@ugr.es	Universidad de Granada
Vicky	Caponigro	vcaponigro@unisa.it	Università di Salerno
Nicola	Caporaso	nicola.caporaso@buhlergroup.com	Bühler Group
Monica	Casale	casale@difar.unige.it	Università di Genova
Lorenzo	Castellino	lorenzo.castellino@unito.it	University of Turin
Marco	Cattaldo	marco.cattaldo@nofima.no	Nofima
Nicola	Cavallini	nicola.cavallini@polito.it	Politecnico di Torino
José Antonio	Cayuela-Sànchez	jacayuela@ig.csic.es	Idaea-csic
Elena	Cazzaniga	elena.cazzaniga@polito.it	Politecnico di Torino
Francesca	Cenci	francesca.cenci@phd.unipd.it	Università di Padova
Magalie	Claeys-Bruno	magalieclaeys@gmail.com	Universitè de Aix-Marseille
Marina	Cocchi	marina.cocchi@unimore.it	Università Modena/Reggio Emilia
Guido	Cozzolino	guido.cozzolino@mi.camcom.it	Camera di Commercio MMB
Jordi	Cruz	jcruz@euss.cat	Euss - School of Engineering
Andres	Cruz Conesa	andres.cruz@urv.cat	Universitat Rovira i Virgili
Enmanuel	Cruz Muñoz	e.cruzmunoz@campus.unimib.it	Università di Milano-Bicocca
Alessandro	D'Alessandro	alessandro.dalessandro@unimore.it	Università Modena/Reggio Emilia

Pierre	Dardenne	dardennepaj@gmail.com	Centre Wallon de Recherches Agronomiques
Daniela	de Assis Bocoli	daniela@exatabrasil.com.br	Exata Brasil
Anna	de Juan	anna.dejuan@ub.edu	Universitat de Barcelona
Marco S.	do Reis	marco@eq.uc.pt	University of Coimbra
Laurence	Dujourdy	laurence.dujourdy@agrosupdijon.fr	AgroSup Dijon
Isabel	Duran Meras	iduran@unex.es	Universidad de Extremadura
Caterina	Durante	caterina.durante@unimore.it	Università Modena/Reggio Emilia
Youssra	Ech Chadhi	youssrachahdi2015@gmail.com	
Jokin	Ezenarro	jokin.ezenarro@urv.cat	Universitat Rovira i Virgili
Pierantonio	Facco	pierantonio.facco@unipd.it	Università di Padova
Emanuele	Farinini	emanuele.farinini@edu.unige.it	Università di Genova
Martina	Fattobene	martina.fattobene@unicam.it	Università di Camerino
Veronica	Ferrari	veronica.ferrari@unimore.it	Università Modena/Reggio Emilia
Joan	Ferré	joan.ferre@urv.cat	Universitat Rovira i Virgili
Victor Hugo	Ferreira	victor.ferreira@univ-lille.fr	Universitè de Lille
Alberto José	Ferrer-Riquelme	aferrer@eio.upv.es	Universidad Politecnica de Valencia
Giorgia	Foca	giorgia.foca@unimore.it	Università Modena/Reggio Emilia
Michele	Forina		Università di Genova
Tiziana	Forleo	tiziana.forleo@uniba.it	Università di Bari
Stefano	Fornasaro	sfornasaro@units.it	Università di Trieste
Martina	Foschi	martina.foschi@univaq.it	Università dell'Aquila
Nicola	Gambacorta	nicola.gambacorta1@uniba.it	Università di Bari
Sergio	Garcia Carrion	sergarc6@doctor.upv.es	Universidad Politecnica de Valencia
Raimundo	Gargallo	raimon_gargallo@ub.edu	Universitat de Barcelona
Carolina	Gastone	carolina@eigengran.it	Eigengran S.r.l.
Emiliano	Genorini	emiliano.genorini@viavisolutions.com	VIAVI Solutions
Beatrice	Gironi	bgironi@aboca.it	Aboca S.p.A.
Barbara	Giussani	barbara.giussani@uninsubria.it	Università dell'Insubria
Adrian	Gomez Sanchez	aderegomez@gmail.com	Universitat de Barcelona
Josè Luis	Gonzalez Solis	jluis0968@gmail.com	Universidad de Guadalajara
Jose Maria	Gonzalez-Saiz	josemaria.gonzalez@unirioja.es	Universidad de la Rioja
Elia	Gonzato	e.gonzato@campus.unimib.it	Università di Milano-Bicocca
Giulia	Gorla	ggorla@uninsubria.it	Università dell'Insubria
Francesca	Grisoni	f.grisoni@tue.nl	University of Eindhoven
Ruggero	Guerrini	ruggero.guerrini.etu@univ-lille.fr	Universitè de Lille
Amanda	Guilardelli	awg@usp.org	US Pharmacopeia
Lenka	Halamkova	lenka.halamkova@ttu.edu	Texas Tech University
Harun	Hano	h.hano@nanogune.eu	CIC nanoGUNE
Nunzia	Iaccarino	nunzia.iaccarino@unina.it	Università di Napoli Federico II
Benoit	Jaillais	benoit.jaillais@inrae.fr	Inrae
Joaquim	Jaumot	joaquim.jaumot@idaea.csic.es	Idaea-csic
Ana maria	Jimenez Carvelo	amariajc@ugr.es	Universidad de Granada
Nicholas	Kassouf	nicholas.kassouf2@unibo.it	Università di Bologna
Sergey	Kucheryavskiy	svk@bio.aau.dk	University of Aalborg
Silvia	Lanteri	silvialanteri@gmail.com	Università di Genova
Riccardo	Leardi	leardi@difar.unige.it	Università di Genova
Pierre	Lebrun	pierre.lebrun@pharmalex.com	Universitè de Liege
Matthieu	Lesnoff	matthieu.lesnoff@cirad.fr	Cirad
Sabina	Licen	slicen@units.it	Università di Trieste
Eneko	Lopez	e.lopez@nanogune.eu	CIC nanoGUNE
Mohammed	Loudiyi	mohammed.loudiyi@geves.fr	GEVES
Henry	Mac Keown	henry.mackeown@edu.unige.it	Università di Genova
Eleonora	Macchia	eleonora.macchia@uniba.it	Università di Bari
Lisa Rita	Magnaghi	lisarita.magnaghi@unipv.it	Università di Pavia
Cristina	Malegori	malegori@difar.unige.it	Università di Genova

Manuela	Mancini	manuela.mancini@staff.univpm.it	Università Politecnica delle Marche
Federico	Marini	federico.marini@uniroma1.it	Università La Sapienza Roma
Silvia	Mas Garcia	silvia.mas-garcia@inrae.fr	Inrae
Fabrizio	Mastrolorito	fabrizio.mastrolorito@uniba.it	Università di Bari
Miriam	Medina Garcia	miriammedina@ugr.es	Universidad de Granada
Camilla	Menozzi	camilla.menozzi@unimore.it	Università Modena/Reggio Emilia
Ayten Ekin	Mese Sezen	aytenekin.mese@tupras.com.tr	Tupras
Alessandro	Monari	alessandro.monari@unimore.it	Università Modena/Reggio Emilia
Flavio	Montenegro de Andrade	flavio.montenegro@ufpe.br	Federal University of Pernambuco
Arsenio	Muñoz de la Pena	arsenio@unex.es	Universidad de Extremadura
Eleonora	Mustorgi	eleonora.mustorgi@viavisolutions.com	VIAVI Solutions
Alessandra	Olarini	alessandra.olarini@unimore.it	Università Modena/Reggio Emilia
Paolo	Oliveri	oliveri@difar.unige.it	Università di Genova
Cruz	Ortiz	mcortiz@ubu.es	Universidad de Burgos
Tommaso	Paravano	tommaso.paravano@phd.unipd.it	Università di Padova
David-Olivier	Paul	davidolivier.paul@ividata.com	Ividata
Samuele	Pellacani	samuele.pellacani@unimore.it	Università Modena/Reggio Emilia
Edenir Rodrigues	Pereira Filho	erpf@ufscar.br	Universidade Federal de Sao Carlos
Fabiola	Pereira Verbi	fabiola.verbi@unesp.br	Universidade Estadual Paulista
Consuelo	Pizarro	consuelo.pizarro@unirioja.es	Universidad de la Rioja
Oliver Aleksandrei	Polushkina	oliverp@ugr.es	Universidad de Granada
Josè Manuel	Prats Montalbàn	jopramon@eio.upv.es	Universidad di Valencia
El Mostafa	Qannari	elmostafa.qannari@oniris-nantes.fr	ONIRIS INRAE StatSC
Aina	Queral-Beltran	aina.queral@idaea.csic.es	Idaea-csic
Beatriz	Quintanilla Casas	beatriz@food.ku.dk	University of Copenhagen
Claudio Marzio	Quintiero	cmquintiero@aboca.it	Aboca S.p.A.
Stiufiuc	Rares Ionut	rares.stiufiuc@umfcluj.ro	Regional Insitute of Oncology
Jordi	Riu	jordi.riu@urv.cat	Universitat Rovira i Virgili
Elisa	Robotti	elisa.robotti@uniupo.it	Università del Piemonte Orientale
Rodrigo	Rocha de Oliveira	rodrigo.rocha@ub.edu	Universitat de Barcelona
Maria Suliany	Rodriguez Barrios	mariasuliany.rodriguez@urv.cat	Universitat Rovira i Virgili
Jean-Michel	Roger	jean-michel.roger@inrae.fr	Inrae
Pier Lorenzo	Rolando	pierlorenzo.rolando@unimore.it	Università Modena/Reggio Emilia
Gloria	Rovira Garrido	gloria.rovira@urv.cat	Universitat Rovira i Virgili
Cyril	Ruckebusch	Cyril.ruckebusch@univ-lille.fr	Universitè de Lille
Itziar	Rui Sanchez	itziar.ruisanchez@urv.cat	Universitat Rovira i Virgili
Raffaele Emanuele	Russo	raffaele.russo@unicam.it	Università di Camerino
Sagrario	Sanchez	ssanchez@ubu.es	Universidad de Burgos
Eugenio	Sandrucci	eugenio.sandrucci@uniroma1.it	Università La Sapienza Roma
Luis	Sarabia	lsarabia@ubu.es	Universidad de Burgos
Francesco	Sartori	francesco.sartori.5@phd.unipd.it	Università di Padova
Francesco	Savorani	francesco.savorani@polito.it	Politecnico di Torino
Claudia	Scappaticci	claudia.scappaticci@graduate.univaq.it	Università dell'Aquila
Giorgia	Sciutto	giorgia.sciutto@unibo.it	Università di Bologna
Andreas	Seifert	a.seifert@nanogune.eu	CIC nanoGUNE
Michelle	Sergent	michelle.sergent@univ-amu.fr	Universitè de Aix-Marseille
Marino	Sergi	marino@eigengran.it	Eigengran S.r.l.
Leonardo	Sibono	leonardo.sibono@unica.it	Università di Cagliari
Mattia	Sozzi	mattia.sozzi@polito.it	Politecnico di Torino
Giacomo	Squeo	giacomo.squeo@uniba.it	Università di Bari
Lorenzo	Strani	lostrani@unimore.it	Università Modena/Reggio

			Emilia
Daniele	Tanzilli	daniele.tanzilli@unimore.it	Università Modena/Reggio Emilia
Romà	Tauler	roma.tauler@idaea.csic.es	Idaea-csic
Kateryna	Tkachenko	kateryna.tkachenko@unirioja.es	Universidad de la Rioja
Roberto	Todeschini	roberto.todeschini@unimib.it	Università di Milano-Bicocca
Berta	Torres Cobos	bertatorres@ub.edu	Universitat de Barcelona
Alessandro	Ulrici	alessandro.ulrici@unimore.it	Università Modena/Reggio Emilia
Kurt	Varmuza	kurt.varmuza@tuwien.ac.at	Università di Vienna
Adrian	Venegas-Reynoso	adrian.venegas-reynoso@ifpen.fr	IFPEN
Guido	Ventura	ventura@optoprim.it	Optoprim
Raffaele	Vitale	raffaele.vitale@univ-lille.fr	Universitè de Lille
Riccardo	Voccio	riccardo.voccio@edu.unige.it	Università di Genova
Xueping	Yang	xueping.yang@studenti.unipd.it	Università di Padova
Alessandro	Zappi	alessandro.zappi4@unibo.it	Università di Bologna
Qiang	Zhu	q.zhu@stu.jiangnan.edu.cn	Università di Padova