Résumés des communications

Communications orales

Pages

Mardi 13 juin 2023

Mercredi 14 juin 2023

Jeudi 15 juin

Posters

Communications orales

Optimisation de l'acquisition des spectres appliquée sur des modèles de calibration chez les arbres forestiers

[1*]Rémy GOBIN, [12]Aurélien LADET, [1]Nassim BELMOKHTAR, [2]Cécile VINCENT-BARBAROUX, [2]Régis FICHOT

[1]INRAE, UMR BIOFORA - PHENOBOIS, 45075 Orléans – France
[2]UNIVERSITE D'ORLEANS, LBLGC, 45067 Orléans – France
Email : remy.gobin@inrae.fr

***Mots-clefs** : protocole de mesure, matrice de l'échantillon, qualité du spectre*

La spectrométrie proche infrarouge (SPIR) couplée à la chimiométrie est un outil d'analyse haut-débit dont les avantages (rapidité, potentiellement non destructif et faible coût) par rapport aux approches traditionnelles permettent l'analyse d'un grand nombre d'échantillons. Dans le cadre de nos travaux de recherche sur les ressources génétiques forestières, celle-ci est mise en œuvre pour la quantification des caractères physico-chimiques du bois ou des feuilles. Toutefois, les mesures SPIR demeurent encore chronophages (récolte, préparation des échantillons et acquisition des spectres). Ce constat nous encourage à poursuivre l'optimisation de l'acquisition des spectres sans pour autant réduire la capacité prédictive des modèles de calibration. Nous avons identifié trois axes d'optimisation : i) réduire le nombre de scans par échantillon lors de l'acquisition du spectre, ii) réutiliser des échantillons anciens pour la construction de nouvelles calibrations ou prédire de nouveaux caractères, iii) acquérir des spectres sur échantillons bruts (non réduit en poudre).

Pour tester et valider ces 3 axes d'optimisation, nous avons i) comparé la qualité des modèles avec la réduction du nombre de scans par échantillon (64 par défaut à 1 scan), ii) testé le vieillissement des poudres de bois en comparant deux campagnes d'acquisition de spectres à un an d'intervalle, iii) comparé l'acquisition de spectres sur des échantillons bruts et sur les mêmes échantillons réduits en poudre.

Les résultats positifs du premier axe d'optimisation vont nous permettre d'ajuster notre protocole et d'améliorer le débit de prise de spectres. Par contre, l'optimisation de l'acquisition des spectres qui se traduit par une dégradation de la qualité des échantillons engendre une perte plus ou moins importante de la capacité prédictive des calibrations selon les caractères intérêts.

Impact de l'appareil de mesure utilisé et de la préparation des échantillons sur des modèles de diagnostic foliaire du palmier à huile par spectrométrie proche infrarouge

[1]Valentin AVIT, [1]Sylvain VRIGNON, [1]Albert FLORI, [2,3,4]Gilles CHAIX, [5]Vivien SARAZIN, [6]Marie TELLA, [1]Jean OLLIVIER

[1]CIRAD/INRAE, UMR ABSYS, 34398 Montpellier - France
[2]CIRAD, UMR AGAP Institut, F-34398 Montpellier, France
[3]UMR AGAP Institut, Univ Montpellier, CIRAD, Institut Agro, Montpellier, France
[4]ChemHouse Research Group, Montpellier, France
[5]SYMILAB, SADEF, 68700 Aspach-le-Bas - France
[6]CIRAD, US49, 34398 Montpellier – France

Email : valentin.avit@cirad.fr

***Mots-clefs** : Spectrométrie, spectromètres portables, régression PLS, palmier à huile, diagnostic foliaire, pilotage de fertilisation, agriculture villageoise*

Le pilotage de la fertilisation est un enjeu essentiel dans le fonctionnement des exploitations agricoles. Pourtant, le diagnostic foliaire, le principal outil de pilotage des exploitations de palmiers à huile, est peu accessible aux exploitations villageoises couvrant de petites superficies. La spectrométrie proche infrarouge produit des mesures liées à certaines liaisons chimiques dans la matière et permet une mesure indirecte des variables biochimiques. Ces dernières années, de nombreux modèles de spectromètres portatifs de petite taille ont été développés, offrant la possibilité d'effectuer des mesures localement et à bas coût. Nous étudions ici la possibilité d'utiliser la spectrométrie proche infrarouge, avec des appareils de mesure portatifs ou transportables, et dans différentes gammes de prix, afin d'effectuer des mesures directement en champs, en « bord de champ », ou en laboratoire mais avec des traitements des échantillons moins coûteux. Les analyses présentées ici portent sur des échantillons de folioles de palmiers à huile, la même partie de l'organe analysée avec la méthode standard de diagnostic foliaire du palmier à huile, sous forme de folioles fraîches rapidement nettoyées en champs, de folioles fraîches nettoyées et hachées fraîches en « bord de champ » et de folioles séchées et broyées en laboratoire. Des mesures ont été réalisées avec un spectromètre ASD à gamme large, un spectromètre Micronir à gamme intermédiaire et un spectromètre Nirone à gamme réduite. L'étude a porté sur la mesure sur le pourcentage en matière sèche de l'azote, du phosphore, du potassium, du calcium, du magnésium et du chlore, des éléments communément analysés lors du diagnostic foliaire.

Après avoir évalué quels prétraitements étaient les plus adaptés pour chaque spectromètre, des régressions PLS (Partial Least Squares) ont permis de comparer les erreurs obtenues sur chacun des nutriments étudiés pour chaque combinaison spectromètre-produit étudiée. Les résultats obtenus étaient relativement contrastés en fonction du nutriment considéré.

Si l'augmentation des erreurs semble peu favorable à une utilisation par les exploitations agricoles pouvant se permettre des analyses en laboratoire plus précises afin de piloter leur fertilisation, notamment compte tenu des forts coûts pouvant découler d'une augmentation des erreurs de mesure sur des parcelles de très grande taille, l'outils semble en revanche en mesure d'offrir une application intéressante dans une perspective d'aide au développement des pratiques culturales chez les petits planteurs : la précision des prédictions les moins coûteuses, avec le spectromètre Nirone et sur folioles fraîches, semble suffisante pour discriminer des états de carence, et pourrait ouvrir à la possibilité de proposer un outil d'aide à la décision portatif afin d'appuyer le travail des agents de développement.

Analyse en ligne de traces par PIR : traces d'eau dans solutions huileuses, traces de COV dans l'eau

[1*]Noémie CAILLOL, [1]Manis Gheghiani, [1]Charlotte Bocquelet, [1]Franck Baco-Antoniali, [2]Maud REY-BAYLE, [3]Sandra GRIMALDI

[1]AXEL'ONE, Analysis, 69360 Solaize – France
[2] IFPEN, Département Expérimentation, 69360 Solaize – France
[3]ARKEMA, CRRA, 69 310 Pierre-Bénite – France

Email : Noemie.caillol@axel-one.com

Mots-clefs : *analyse en ligne, analyses de traces, Phase aqueuse, phase organique, acquisition et exploitation des données*

L'analyse PIR n'est pas une analyse de traces ! C'est bien connu. Et pourtant on peut suivre l'eau à très faible teneur (autour de la dizaine de ppm) même si calibrer un modèle peut être une gageure.

On pourra présenter une étude qui compare à la fois IR et PIR. Le retour d'expérience sur les méthodologies d'acquisition de données pour la calibration pourra être présenter. Ceci afin de mettre en avant les écueils à éviter mais aussi tout l'intérêt de l'analyse en ligne pour comprendre et suivre les mécanismes de mélangeage lors de de la préparation des points de calibration par ajouts dosés.

En effet, ce travail a permis de tester différents protocoles de synthèse des standards de calibration pour les diverses matrices dont la nature a toute son importance.

L'analyse PIR est une méthode linéaire ! C'est bien connu. Et pourtant on verra très rapidement les phénomènes d'interactions entre molécules qui s'expriment d'autant plus volontiers à faible teneurs.

En effet, pour pouvoir analyser l'eau, celle-ci doit s'incorporer dans le milieu de manière assez homogène, et ce paramètre s'est avéré être déterminant sur la validation de la méthode. C'est pourquoi une étude de faisabilité pour chacune des matrices à étudier est à effectuer au préalable et qu'il n'y a pas de protocole définit mais plusieurs qui doivent être maitrisés pour que les résultats soient reproductibles et que le traitement de données soit effectif.

Enfin nous parlerons aussi de la recherche de traces d'organique dans l'eau.

Imagerie hyperspectrale proche infrarouge appliquée à l'art

Benoît JAILLAIS[1]

[1] *StatSC, Oniris INRAE, Rue de la géraudière, 44322, Nantes – France*

Email : Benoit.Jaillais@inrae.fr

***Mots-clefs** : imagerie hyperspectrale proche infrarouge, vérité terrain, peinture, chimiométrie.*

En spectroscopie proche infrarouge, la vérité terrain est toujours compliquée à obtenir, voire impossible. Ainsi lors du traitement exploratoire de données, des loadings peuvent être ininterprétables soit parce que les longueurs d'ondes identifiées sont reliées à un phénomène sous-jacent inconnu, soit parce qu'une bande peut être assignée à des fréquences de vibration de liaisons chimiques communes à de nombreuses molécules.

Il existe des matériaux de référence qui peuvent être achetés pour aider à l'interprétation. Afin d'étudier cette vérité terrain, j'ai décidé de « fabriquer » un échantillon à partir de matériau connu, à savoir une reproduction partielle de la danse de Matisse (Figure 1a).

Une image hyperspectrale proche infrarouge de ce tableau est ensuite acquise (SWIR-CL-400-N25E, SPECIM) puis analysée par ACP, MCR-ALS et ICA afin d'estimer quels sont les apports de chaque technique pour décrire au mieux l'échantillon. Le traitement le plus pertinent par rapport à la vérité terrain sera sélectionné et appliqué ultérieurement à l'étude d'autres tableaux.

Il est aisé de voir que l'image score associée à PC2 (Fig. 1b), est bien représentative du tableau étudié. Toutefois, le fond du tableau n'est pas autant homogène que dans le tableau. Ce résultat (probablement dû à la profondeur de pénétration du rayonnement proche infrarouge) se retrouve avec les autres traitements utilisés et sera détaillé dans la présentation.

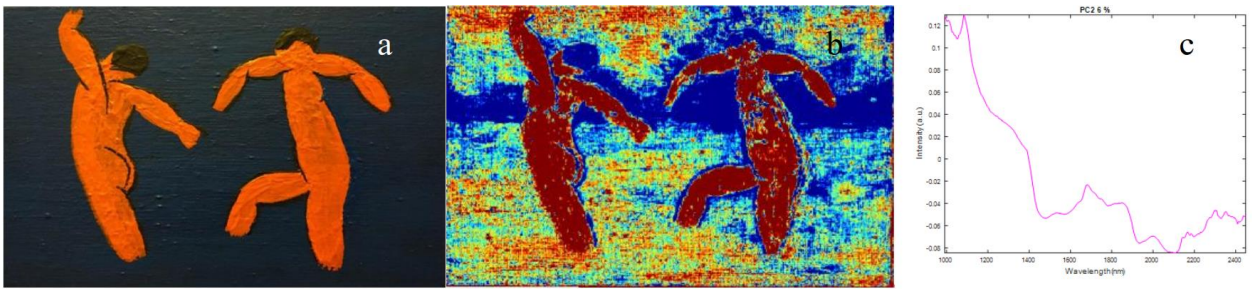


Figure 1: Image RGB (a), PC2 image-score (b) et Loadings associés à PC2 (c)

Hyperspectral imaging for the determination of relevant cooking quality traits of boiled cassava

Karima Meghar[a*], Thierry Tran[a,b], Luis Fernando Delgado[b], Maria Alejandra Ospina[b], Jhon Larry Moreno[b], Jorge Luna[b], Luis Londoño[b], Dominique Dufour[a], Fabrice Davrieux[a]

[a] *CIRAD, UMR Qualisud, F-34398 Montpellier, France.*
[b] *Alliance of Bioversity-International Center for Tropical Agriculture (CIAT), Cassava Program, Cali, Colombia.*

***Keywords**: Dry matter content, water absorption, chemometrics, high throughput phenotyping, consumer preferences*

The purpose of this study was to investigate the potential of hyperspectral imaging (HSI) for the characterization of cooking quality parameters, dry matter content (DMC) and water absorption (WAB) in cassava genotypes contrasting for their cooking quality.

Hyperspectral images were acquired on cooked and fresh intact longitudinal and transversal slices from 31 cassava genotypes harvested in March 2022 in Colombia. Different chemometric methods were tested for the quantification of DMC, WAB and texture parameters. Data analysis was conducted through Partial Least Square Regression (PLSR), K Nearest Neighbors Regression (KNNR), Support Vector Machine Regression (SVM) and CovSel Multiple Linear regression (CovSel_MLR). Efficient performances were obtained for DMC using CovSel_MLR with, R^2p = 0.94, RMSEP = 0.96 g/100g and RPD = 3.60. High heterogeneity was observed between contrasting genotypes. The predicted distribution of DMC within the root can be homogeneous or heterogeneous depending on the genotype. Weak predictions were obtained for WAB parameter.

This study showed that HSI could be used as a high throughput phenotyping tool for the visualization of DMC in contrasting cooking quality genotypes. Further improvement of protocols and larger datasets are required for WAB quality trait.

How to properly analyse spectra

Harald Martens

[1] *DEPARTMENT OF ENGINEERING CYBERNETICS, Norw. U. of Sci. & Technol. NTNU, 7034 Trondheim Norway*
[2] *IDLETECHS AS (www.idletechs.com) Trondheim Norway*

***Keywords**: Big Data, Chemometrics, Calibration*

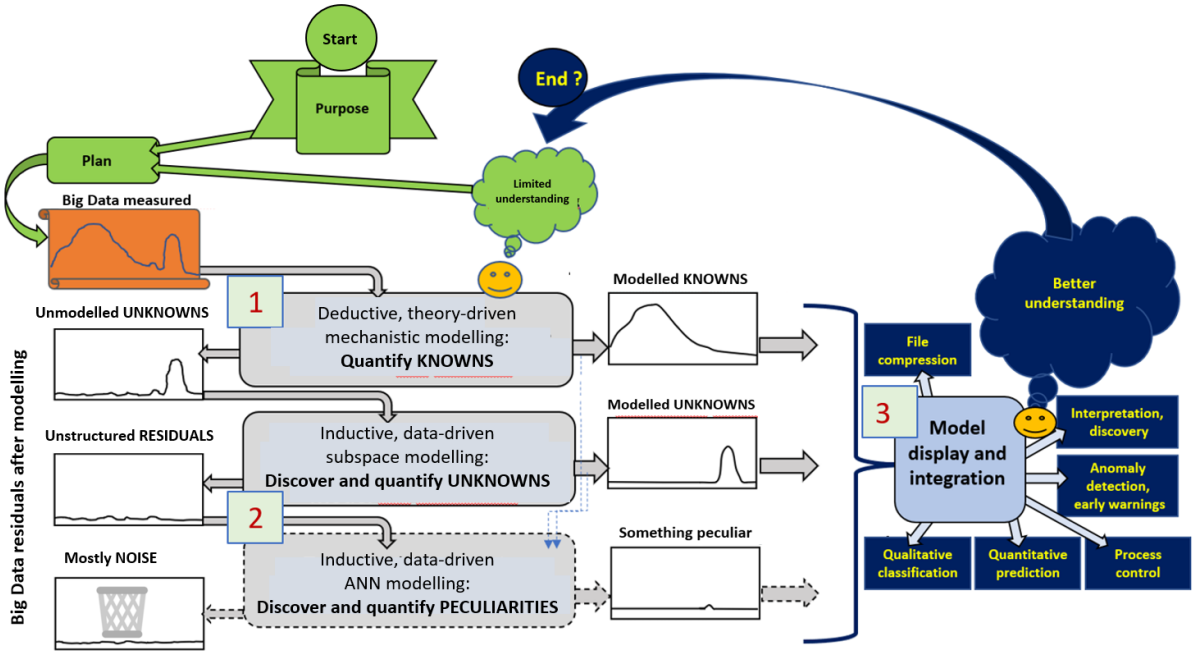
Modern high-speed measurements generate Big Data. But whether we know it or not, most technical scenes and samples change in systematic ways according to laws of nature, and so do modern multi-channel spectrometers and multi-pixel imagers. This allows us to discover and quantify various systematic change patterns by multivariate data modelling.

Thus, systematic Big Data measurements of spectral or imaging type do not call for black box AI solutions, with the alienation, risk and cost that involves. Chemometric hybrid subspace modelling is often a simpler and better alternative. But the laws of nature are not necessarily linear and additive. Response curvature, mixed multiplicative/additive effects, chemical and physical interaction effects, temperature effects, wavelength- or spatial shifts, and even totally unexpected change patterns need to be dealt with, e.g. by multivariate hybrid dual-domain IDLE modelling.

The big problem in many modern measurements is not uncertainty caused by random measurement noise, but selectivity problems caused by impurities and other uncontrolled variation types that interfere with the desired instrument signals.

Multivariate calibration based on hybrid causal/empirical subspace modelling allows us to convert torrents of per se meaningless multi-channel or multi-pixel raw "spectra", via their known and unknown but systematic patterns of covariation, into meaningful quantification and classification, in a way that combines deductive use of prior knowledge and inductive discovery of unexpected but systematic patterns.

The lecture outlines how chemometric hybrid subspace modelling provides generic tools for "human-interpretable machine learning with an eye for the physics" that can be adapted to a wide range of spectral and imaging data.



Utilisation de la spectroscopie proche infrarouge et la MCR-ALS pour le suivi de la cuisson des sauces béchamel

[1]Sonia Nieto-Ortega, [2,3]Silvia Mas Garcia, [1]Angela Melado-Herreros, [1]Giuseppe Foti, [1]Idoia Olabarrieta et [2,3]Jean-Michel Roger

[1]AZTI, Food Research, Basque Research and Technology Alliance (BRTA), Parque Tecnológico de Bizkaia, Astondo Bidea, Edificio 609, 48160, Derio, Spain
[2]ITAP, INRAE, Institut Agro, University Montpellier, 34196, Montpellier, France
[3]ChemHouse Research Group, 34196, Montpellier, France

Email : silvia.mas-garcia@inrae.fr

Mots-clefs : *spectroscopie proche infrarouge (SPIR), résolution de courbes multivariées par moindres carrés alternés (MCR-ALS), processus de cuisson, béchamels*

Cette étude présente une analyse exploratoire de la cinétique et des mécanismes impliqués dans le processus de cuisson de 27 sauces béchamel. Le suivi de l'évolution des sauces béchamel au cours de leur élaboration a été étudié à l'aide d'un capteur portable de spectroscopie proche infrarouge (SPIR) combiné à l'approche résolution de courbes multivariées par moindres carrés alternés (MCR-ALS). La MCR-ALS a été appliquée pour élucider le mécanisme du processus de cuisson des sauces béchamel, dans l'objectif de résoudre les signatures spectrales pures des constituants impliqués dans le processus, leurs profils cinétiques et les constantes de vitesse afférentes.

Le processus d'élaboration des sauces béchamel peut être décrit par un modèle cinétique basé sur une réaction du premier ordre (A→B). Ces deux espèces, A et B, impliquées dans le processus sont liées à des changements dans la diffusion de la lumière et dans la nature et l'état de l'eau, respectivement. Des différences dans les constantes cinétiques entre les sauces béchamel ont été associées à des différences dans la température initiale du processus de cuisson. Les résultats obtenus sont cohérents avec les images observées à l'aide d'un microscope électronique à balayage (MEB).

La méthodologie proposée dans ce travail offre une nouvelle stratégie pour étudier l'élaboration des sauces béchamel de manière non destructive, dans l'objectif de donner aux producteurs industriels une meilleure compréhension de leur processus de fabrication de manière rapide et en temps réel.

Support Vector Machines models based on PLS scores – Algorithm development and case studies

[1*]Andreas Niemoeller, [2]Jordane Poulain

[1]Bruker Optics GmbH & Co. KG, 76275 Ettlingen, Germany
[2]Ondalys, 34 830 Clapiers, France

Email : Andreas.niemoeller@bruker.com

***Mots-clefs** : Support Vector Machines, Multivariate Calibration, Regression, PLS scores*

Spectroscopic methods like NIR are established fast methods and a central element of quality control and other tasks. The usage of NIR in general is still increasing as well as the range of applications and the analyzed components. Due to the development of new handheld NIR based solutions and services and the increasing demands for process analysis the scope gets even broader. A substantial part of applying is the modelling for the required evaluation methods and here the situation influenced by contradicting trends: more NIR applications and huge amount of data collected in years of NIR activities need to be covered by decreasing manpower and less knowledge on modelling and chemometrics. Even when more instruments are networked and centralized method development is done the amount of modelling and validation work is tremendous.

One option to increase productivity of modelling and validation is to combine different sample types or products in one model instead of optimizing several models for the same data. For going beyond PLS there are a lot of possible algorithms such as Local Regression (LR), Support Vector Machines (SVM) and Artificial Neural Networks (ANN) [1]. They are more complex but offer options and work even with non-linear data.

In this paper, the application of SVR [2] based on PLS score values is proposed and tested. With this approach, users of PLS can still use their current workflow in a first step to check the data and do premodelling. PLS scores are appropriate and sparse latent variables as input for the SVR modelling. The scores represent dedicated spectral variance of the component to be calibrated and this data compression even provides a speed advantage compared to SVR modelling based on spectra. During prediction of new spectra, the scores-based SVR approach is an advantage, as the calculation of scores allows the use of outlier detection and spectral residuals as in a pure PLS evaluation, which is not possible when using SVR directly on spectra. Overall, PLS is simply extended to make best use of SVR approach while maintaining a good and well understood knowledge base of the users in industry.

In contrast to other PLS alternatives SVR has interesting advantages: Data sets can be handled over a broad concentration or component value range in which there is normally a higher variance in the spectra, especially when various products are combined. Here SVR allows the use of just one model instead of multiple PLS models and cascades. On the other hand, SVR can be used for fairly small data sets as well, which is not advised for other advance modelling methods. Moreover, SVR leads to deterministic models, meaning each calculation yields the same model or result which is not the case for ANNs with random starting values.

The approach of SVR based on PLS scores was applied to two data sets: sugar factory products [3] but with data from Bruker and a data set for feed which consists of poultry, ruminant and swine finished feeds. It could be shown that optimized PLS models could be improved with SVR. Moreover, SVR models simply based on a broad spectral range were performing almost the same than optimized PLS models without many efforts.

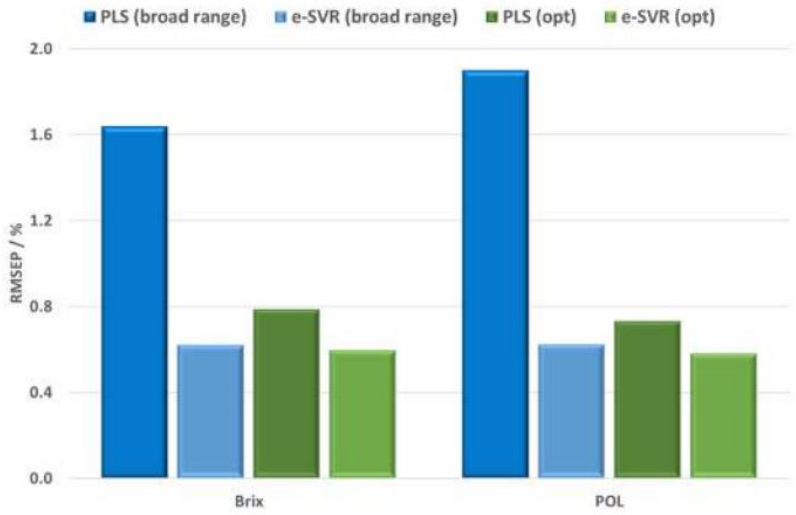


Figure 1: RMSEP values for tuning set of sugar factory calibration for optimized models and models with broad range (9080-5140 cm-1)

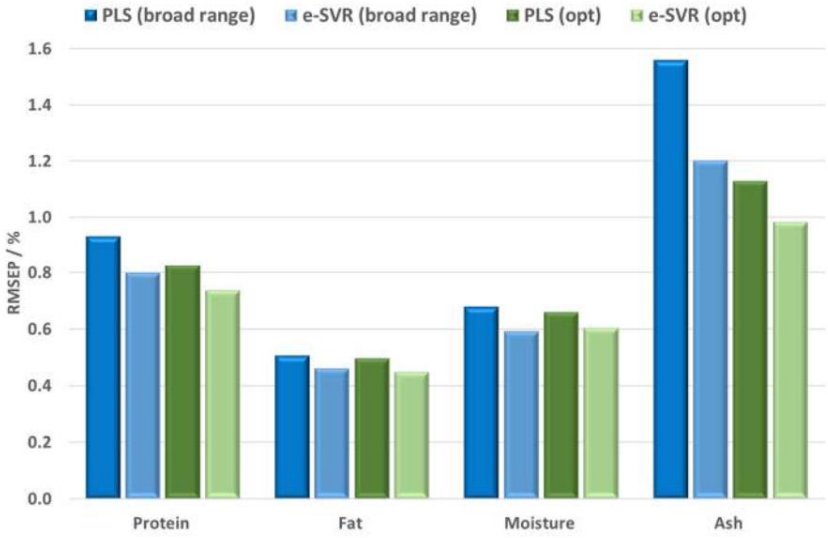


Figure 2: RMSEP values for tuning set of feed calibrations for optimized models and models with broad range (9080-5140 cm-1)

[1] J.A. Fernández Pierna, B. Lecler, J.P. Conzen, A. Niemoeller, V. Baeten and P. Dardenne, Comparison of various chemometric approaches for large near infrared spectroscopic data of feed and feed products, Analytica Chimica Acta 705, 30-34, 2011
[2] How to Apply Support Vector Machines (SVM) to predict spectroscopic data - Comparison of SVM and PLS regression models, White paper, Ondalys, http://www.ondalys.fr, 2020
[3] I. Ramírez-Morales, D. Rivero, E. Fernández-Blanco and A. Pazos, Optimization of NIR calibration models for multiple processes in the sugar industry, Chemometrics and Intelligent Laboratory Systems 159, 45-57, 2016

Combinaison de micro-spectromètres NIR pour prédire les propriétés chimiques du fourrage de canne à sucre

[1,4*] Maxime RYCKEWAERT, [2,3,4] Gilles CHAIX, [1] Daphné HERAN, [1] Ryad BENDOULA

[1] ITAP, Univ Montpellier, INRAE, Institute Agro, Montpellier, France
[2] CIRAD, UMR AGAP Institut, F-34398, Montpellier, France
[3] AGAP, Univ Montpellier, CIRAD, INRAE, Institut Agro, F-34398 Montpellier, France
[4] ChemHouse Research Group, Montpellier, France

Email : maxime.ryckewaert@inrae.fr

***Mots-clefs** : Micro-Spectromètre, Fusion de données, Multi-block*

Lors de la récolte, l'évaluation de la valeur alimentaire des cultures destinées à l'alimentation animale peut être estimée à partir de variables biochimiques. La spectroscopie proche infrarouge (NIR) est une façon de mesurer indirectement ces variables biochimiques. Au cours des dernières années, plusieurs micro-spectromètres ont été développés, offrant une opportunité de prédire ces variables biochimiques à faible coût. Dans cette étude, nous évaluons le potentiel d'une combinaison de micro-spectromètres pour prédire la teneur en protéines et en sucre total de la canne à sucre.

Tout d'abord, chaque micro-spectromètre, avec des prétraitements spectraux optimaux, a été comparé individuellement à un spectromètre de laboratoire de référence. Ensuite, nous proposons une combinaison de micro-spectromètres et avons établi des modèles de prédiction grâce à une méthode de fusion des données appelée Sequential and Orthogonalised - Partial Least Squares (SO-PLS). En ce qui concerne la teneur en protéine, la combinaison des micro-spectromètres fournit un modèle similaire (sep = 0,69 % ; biais = 0,15 % ; r2 = 0,910) à ceux obtenus avec le spectromètre de référence (sep = 0,56 % ; biais = -0,13 % ; r2 = 0,935). En ce qui concerne les prédictions de teneurs en sucre, les résultats obtenus avec cette combinaison de micro-spectromètres (sep = 2,38 % ; biais = -0,52 % ; r2 = 0,983) sont supérieurs à ceux obtenus avec le spectromètre de référence (sep = 2,59 % ; biais = 0,41 % ; r2 = 0,978).

Pour les deux variables chimiques, la combinaison des micro-spectromètres améliore les performances des modèles de prédiction par rapport aux modèles obtenus avec les micro-spectromètres utilisés individuellement. L'utilisation de plusieurs micro-spectromètres peu coûteux, combinée à une méthode multi-blocs, permet d'obtenir des résultats aussi satisfaisants qu'avec un seul spectromètre de laboratoire, mais à moindre coût.

Utilisation du SPIR portable pour la discrimination de génotypes de frêne résistant à la chalarose

[12*]Nassim BELMOKHTAR, [12]Nathalie BOIZOT, [1]Arnaud DOWKIW

[1]INRAE, UMR INRAE-ONF BioForA 0588, 45075 Orléans – France
[2]INRAE, Plateforme Phénobois, 45075 Orléans – France

Email : nassim.belmokhtar@inrae.fr

***Mots-clefs** : proche infrarouge, spir portable, frêne, génotypes, chalarose, discrimination*

La chalarose est une maladie invasive létale qui menace les frênes commun et oxyphylle en Europe occidentale[1]. Elle est causée par un champignon pathogène originaire de l'est de l'Asie nommé *Hymenoscyphus fraxineus* (forme asexuée *Chalara fraxinea*). Après avoir été signalés en Europe au début des années 1990, les dépérissements de frênes dus à ce pathogène concernent désormais toute l'Europe avec des niveaux de dépérissement et de mortalité très variables, expliqués notamment par la résistance génétique de l'individu. La préservation de cette essence forestière passe par l'identification des génotypes les plus résistants à cette maladie[2].

Dans cette étude, nous explorons le potentiel de la spectrométrie proche infrarouge portable pour discriminer les génotypes de frêne in situ en fonction de leurs niveaux de résistance à la chalarose. Pour ce faire, des spectres ont été collectés sur l'écorce et les feuilles de 103 génotypes sains âgés de 18 mois à l'aide d'un spectromètre proche infrarouge portable (908,1-1676,2 nm). Après inoculation de l'agent pathogène, les individus ont été classés en fonction de leurs symptômes : sain, moyennement atteint ou nécrosé.

Notre démarche consiste à comparer l'apport des méthodes PLS-DA, SVM et Random Forest au développement de modèles discriminant ces génotypes selon les symptômes observés tout en tenant compte de l'hétérogénéité des tissus végétaux analysés sur le terrain. Les modèles obtenus présentent des niveaux de précision allant de 70 à 90 % ouvrant ainsi la voie leur possible déploiement sur le terrain comme outil d'aide à la décision pour les acteurs de la filière.

1. Husson, C.; Dowkiw, A.; Saintonge, F.-X.; Marçais, B., La chalarose du frêne en France. Forêt Entreprise 2016, (228), 10-13.
2. Villari, C.; Dowkiw, A.; Enderle, R.; Ghasemkhani, M.; Kirisits, T.; Kjær, E. D.; Marčiulynienė, D.; McKinney, L. V.; Metzler, B.; Muñoz, F.; Nielsen, L. R.; Pliūra, A.; Stener, L.-G.; Suchockas, V.; Rodriguez-Saona, L.; Bonello, P.; Cleary, M., Advanced spectroscopy-based phenotyping offers a potential solution to the ash dieback epidemic. Scientific Reports 2018, 8 (1), 17448.

Validation de matières premières par NIR

[1] Safia MONTAGNIER, [2] Marie GEORGEL

[1,2] ARKEMA, CERDATO, 27470 Serquigny – France

Email : safia.montagnier@arkema.com

***Mots-clefs** : matières premières, analyse discriminante, test de conformité*

ARKEMA, leader dans le domaine de fabrications des polyamides longue chaine, possède différentes usines à travers le monde. Afin de protéger le savoir-faire et l'expertise développés au long de nombreuses années de recherches, les recettes des matériaux produits ne sont pas communiquées à tous les sites de fabrication. Dans certaines usines, les formules et les noms des matières premières sensibles sont codées. Ces matières premières sont ré-ensachées et ré-étiquetées dans des sacs banalisés avant d'être livrés sur le site.

Afin d'éviter des erreurs opératoires à fort impact sur la qualité des matériaux finaux, un contrôle des matières premières à leur réception sur le site de production est nécessaire pour réduire le risque. Ce contrôle est effectué par une analyse simple et discriminante, réalisable en laboratoire de contrôle de fabrication ; elle permet de vérifier que le code indiqué sur le sachet de la matière première correspond bien au produit analysé (résultat OK ou NOK).

Ces matières premières sont de natures chimiques très variées : à titre d'exemple, des antioxydants qui protègent la matière au cours de sa vie ou de sa transformation, des charges inorganiques qui ajoutent des propriétés au matériau final, ou des colorants.

Pour réaliser ce contrôle, il a été choisi de développer un test de conformité sur le logiciel OPUS de la société Bruker à partir des spectres NIR des matières premières.

Deep learning for NIR spectroscopy : overview and thoughts

[1*]Florent HAFFNER, [1]Marion LACOUE-NEGRE, [1]David GONCALVES, [2]Aurélie CHATAIGNON, [1] Julien GORNAY, [1] Maxime MOREAUD

[1] IFP Energies nouvelles, Rond-point de l'échangeur de Solaize, BP 3, 69360 Solaize, France
[2] IFP Énergies Nouvelles, 1 et 4 avenue de Bois-Préau, 92852 Rueil-Malmaison – France

Email : florent.haffner@ifpen.fr

***Keywords**: deep learning, chemometrics, near-infrared spectroscopy, machine learning*

Near-Infrared spectroscopy (NIR) combined with chemometrics has been used for years in various fields to predict properties, product concentration, classify samples, etc. At IFPEN, NIR and chemometrics are daily used since more than 30 years, both in the laboratory and on-line on pilot plant units, to predict properties of interest of petroleum cuts, and more recently, for biomass characterization and plastic recycling processes. A major task of this approach is to determine a consensus on which signal processing method applied to the spectra and which optimal settings to use for the chosen method (regression, classification, etc.). The frequent practice is a time-consuming trial-and-error experimentation that must be reiterated when changing the spectrometer, for a change in acquisition conditions, to implement in real-time for process monitoring, etc.

Deep Learning approach has a different philosophy than chemometrics since it is interested in representation learning, which builds internal abstract representation to perform its tasks. Deep learning designs a model automatically extracting knowledge from data and simplify or remove human crafted operations on datasets. Additionally, this sub-field is able not only to extract linear information from data, like traditional chemometrics learning algorithm, but also non-linearity, which helps the model extract more complex information from data. Based on previous experience, especially recently published studies that demonstrate deep networks can learn critical patterns from raw data [1], it is possible to consider Deep Learning as a challenger for Chemometrics traditional practices to simplify its processes.

The current state-of-the-art of Deep Learning applied to NIR analysis is in its infancy compared to computer vision and natural language processing. It is composed of multiple directions, from Artificial Neural Network (ANN) to Convolutional Neural Network (CNN), and other architectures such as Recurrent Neural Network (RNN) like Long Short-Term Memory (LSTM) network, or generative network like Autoencoders (AE), Transformers, and Generative Adversarial Networks (GAN).

ANNs are the simplest neural networks and have been used on NIR data mostly between 2008 and 2011 [2,3]. Their overall performance was interesting, but the absence of interpretability of these networks indicates a potential drawback. Furthermore, their difficulty of interpretability leads to a limited applicability in tasks, not only in research but also on commercial products. In the data-science community, it is well known that the strong usage of other networks such as RNNs, CNNs and more recently transformers, tends to lead the way and indicates a low trust in ANNs. Also, considering the latter do not have a clear gain in performance over Partial Least Squares, yet add a complexity in relation to tweaking, the usage of these networks has decreased over time.

Currently, the most popular networks on NIR analysis are the CNN based architectures, which is an evolution of ANN with reduced risks of overfitting. These deep networks rely on convolution operations that, when stacked in multiple layers stacks, are able to extract meaningful neighborhood features from the input data. The CNNs architectures used in NIR analysis can be associated in multiple philosophies or groups. A first group, close to traditional chemometrics, uses digital signal processing then feeds a shallow CNN. A second group, close to modern deep convolutional neural networks, uses deep architecture to delegate the manual tasks of variables selection and digital signal processing. Many studies show these

networks perform well [4–6], yet these diverse workflows with various philosophies lead to the following question: how to use deep learning for NIR analysis, which regularly involves to use small datasets. Furthermore, which architectures and learning strategy should we explore to satisfy this constraint?

The literature shown beyond CNNs, diverse architectures can manage different tasks in a more efficient and effective manner, like RNNs and AEs for qualitative/quantitative analysis, and GANs for simulating spectra. These networks and proposed usage are following:

- LSTMs are networks that excel in handling sequential and time-series data. They use memory cells and gating mechanisms to selectively retain or discard information at each step of the sequence, this approach can extract long-term dependencies present in an input spectrum. This approach has shown great performance for food industry on quantitative analysis applied to manure [7].
- AEs are networks that build new representation of data with less noise. They learn to build a compressed latent version of a spectrum, then to rebuild a decompressed spectrum without noise. This approach has shown impressive performance for chemical quantitative analysis applied to aging of product with NIR analysis [8].
- GANs are a combination of two networks to generate new representation of data, especially for improving a dataset with new samples. A first network learns to generate a representation of data by adding noise to an existing dataset, then another network learns to discriminate by deciding if the representation is good enough or to discard it and try differently. This approach has shown success for augmenting a small dataset by generating fake samples to help training PLS models and improving their robustness [9].

Considering GANs can be useful to enhance a given dataset and the networks interested in qualitative and quantitative analysis can have different types of uses and thus be combined to extract different characteristics present within the data, it is possible to consider the following hypothesis. The CNN, AE and RNN networks, each of which is interested in extracting different types of knowledge within the data, can be combined to understand neighborhood behaviors, minimizing noise within an internal representation as well as finding the long-term relationships present within the spectra, coupled with larger dataset thanks to the help of GANs. Obviously, combining these architectures is a complex and ambitious task, but it is necessary beforehand to better understand them separately. Lastly, other usage could be explored and will lead to a better understanding of how to use these technologies to better serve our domain.

Based on the overview provided, many topics needs to be both explored and discussed. Deep learning can certainly bring a lot to NIR data analysis. Further research is necessary to explore these topics. For its improvement of performance and robustness, compared to PLS, but how much and in which context is it necessary to apply both digital signal processing and variable selection? How to combine these multiple approaches to solve more complex tasks? How to use others architecture and tweak their associate hyperparameters to solve efficiently NIR issues? How does model interpretability work for such model, we know it is possible for images and natural language but what is the state of knowledge for vibrational spectral analysis?

Bibliographie

[1] Yang J., Xu J., Zhang X., Wu C., Lin T., Ying Y. Deep learning for vibrational spectral analysis: Recent progress and a practical guide, Analytica chimica acta, 2019, 1081, 6-17. DOI: 10.1016/j.aca.2019.06.012.

[2] Balabin R.M., Safieva R.Z., Lomakina E.I. Wavelet neural network (WNN) approach for calibration model building based on gasoline near infrared (NIR) spectra, Chemometrics and Intelligent Laboratory Systems, 2008, 93, 1, 58-62. DOI: 10.1016/j.chemolab.2008.04.003.

[3] Balabin R.M., Lomakina E.I., Safieva R.Z. Neural network (ANN) approach to biodiesel analysis: Analysis of biodiesel density, kinematic viscosity, methanol and water contents using near infrared (NIR) spectroscopy, Fuel, 2011, 90, 5, 2007-2015. DOI: 10.1016/j.fuel.2010.11.038.

[4] Mishra P., Passos D. A synergistic use of chemometrics and deep learning improved the predictive performance of near-infrared spectroscopy models for dry matter prediction in mango fruit, Chemometrics and Intelligent Laboratory Systems, 2021, 212, 104287. DOI: 10.1016/j.chemolab.2021.104287.

[5] Zhang X., Lin T., Xu J., Luo X., Ying Y. DeepSpectra: An end-to-end deep learning approach for quantitative spectral analysis, Analytica chimica acta, 2019, 1058, 48-57. DOI: 10.1016/j.aca.2019.01.002.

[6] Acquarelli J., van Laarhoven T., Gerretzen J., Tran T.N., Buydens L.M. C., Marchiori E. Convolutional neural networks for vibrational spectroscopic data analysis, Analytica chimica acta, 2017, 954, 22-31. DOI: 10.1016/j.aca.2016.12.010.
[7] Tan A., Wang Y., Zhao Y., Wang B., Li X., Wang A.X. Near infrared spectroscopy quantification based on Bi-LSTM and transfer learning for new scenarios, Spectrochimica acta. Part A, Molecular and biomolecular spectroscopy, 2022, 283, 121759. DOI: 10.1016/j.saa.2022.121759.
[8] Grossutti M., D'Amico J., Quintal J., MacFarlane H., Quirk A., Dutcher J.R. Deep Learning and Infrared Spectroscopy: Representation Learning with a β-Variational Autoencoder, The journal of physical chemistry letters, 2022, 13, 25, 5787-5793. DOI: 10.1021/acs.jpclett.2c01328.
[9] He K., Liu J., Li Z. Application of GAN for prediction of Gasoline Properties, 2020.

Evolution de la performance des réseaux de neurones complètement connectés en fonction de certains facteurs dans le cadre du traitement de données spectroscopiques

[1]Carl CHARLOTO, [2,4]Maxime METZ, [3,4]Matthieu LESNOFF, [1]Florent ABDELGHAFOUR, [1,4]Jean-Michel ROGER

[1]ITAP, Univ. Montpellier, INRAE, Institut Agro, Montpellier, France
[2]Pellenc Selective Technologies, Pertuis, Provence-Alpes-Côte d'Azur, France
[3]UMR SELMET, Univ Montpellier, CIRAD, INRA, Institut Agro, Montpellier, France
[4]ChemHouse Research Group, Montpellier, France

Email : carl.charloto@inrae.fr

***Mots-clés** : Données spectroscopiques, Apprentissage profond, Réseaux de neurones profond, Plan d'expériences*

Introduction :

L'apprentissage profond a permis des avancées significatives dans plusieurs domaines comme le traitement d'image [1], l'analyse de texte [2], la reconnaissance vocale [3]. De récents développements [4] ont mis en évidence la pertinence de certaines approches d'apprentissage profond dans le traitement de données spectroscopiques. Néanmoins, comparé aux outils classiques de chimiométrie, l'apprentissage profond nécessite le réglage de paramètres, ce qui rend l'étape du développement de nouvelles architectures très difficile. Cette présentation a pour but d'évaluer l'impact de deux paramètres sur la performance des architectures de neurones. Le choix a été fait de se limiter aux réseaux de neurones complètement connectés.

Matériels et méthodes :

Les données utilisées pour cette étude étaient des données fourrages (pré-traitées et totalement anonymisées) acquises au Cirad-UMR Selmet et préparées par L. Bonnal et M. Lesnoff. La base de données totale représentait 11021 individus, séparée en 2 jeux : un jeu d'entraînement (9021 individus) et un jeu de test (2000 individus). Le jeu d'entraînement a été également séparé en 2 jeux, un jeu de calibration (7217 individus) et un jeu de validation (1804 individus). Les spectres étaient constitués de 700 longueurs d'ondes. La problématique portait sur la prédiction d'une réponse continue, c'est-à-dire sur l'établissement d'une régression. Les modèles ont été créés en langage python, à l'aide des modules suivants : Tensorflow 2.8.0, Numpy 1.21.5, Pandas 1.4.4, Keras 2.8.0.

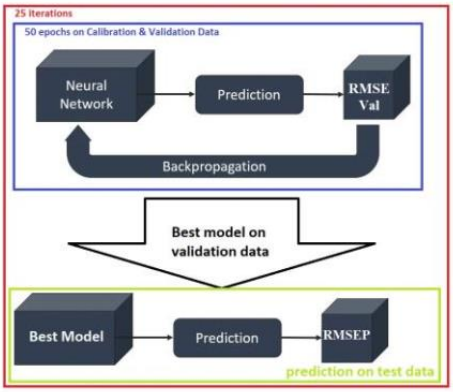


Figure 1 Schéma de la procédure suivie

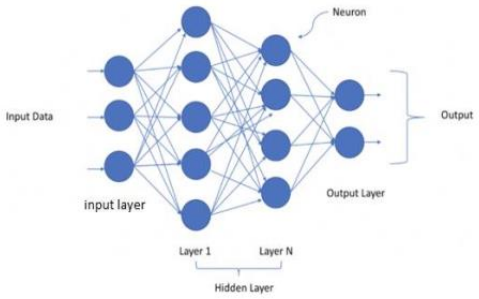


Figure 2 Schéma des architectures testées

La figure 1 représente comment chaque architecture a été entraînée et testée. Chaque architecture a été entraîné durant 50 itérations et pour chaque architecture le meilleur modèle sur le jeu de validation a été sélectionné. Toute la procédure a été répétée 25 fois pour assurer une stabilité dans les résultats. Pour chaque architecture, le meilleur modèle a été sélectionné avec comme critère la meilleure performance sur le jeu de validation.

La figure 2 représente les architectures testées, chaque architecture est composée de 3 types de couches, une couche d'entrée fixe de la taille du nombre de variables, des couches cachées et une couche de sortie d'un neurone car la valeur à prédire est une variable continue.
Deux séries d'expériences ont été mises en œuvre. La première est une série d'expériences où le nombre de couches cachées a varié de 1 à 10 par pas de 1, le nombre de neurones par couches était fixé à 128.
La deuxième est une série d'expériences où le nombre de neurones a varié de 128 à 256 par pas de 128, le nombre de couches cachés était fixé à 3

Résultats :

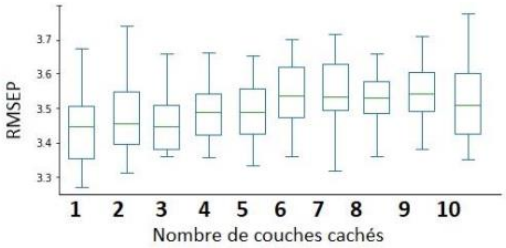


Figure 3 Distribution des RMSEP obtenues sur le jeu en fonction du nombre de couches cachées

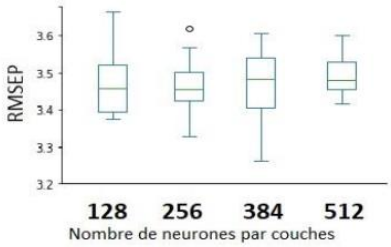


Figure 4 Distribution des RMSEP obtenues sur le jeu en fonction du nombre de neurones par couches cachées

La figure 3 montre la RMSEP en fonction du nombre de couches cachées, on observe que la RMSEP varient de façon très faible, on conclut que le nombre de couches cachées seule ne permet pas de dégager de tendance générale.
La figure 4 montre la RMSEP en fonction du nombre de neurones par couches cachées, on observe que la RMSEP varient de façon très faible, on conclut que le nombre de neurones par couches cachées seule ne permet pas de dégager de tendance générale.

Au vu des résultats, la méthode d'expérimentation de faire varier les paramètres de création d'architecture ne semble pas être pertinent.

Conclusion et perspectives :

Au vu des résultats, nous avons pu constater que faire varier les paramètres de création d'architectures de réseaux un à un n'est pas pertinent. En effet, dans la création d'architectures de réseaux, les paramètres ne sont pas indépendants. Ce type de plan d'expérience n'en tient pas compte. Pour aller plus loin, il faudrait créer des plans d'expériences qui en tiennent compte. Néanmoins, cette étude montre la nécessité d'une compétence métier pour la création d'architecture de réseaux de par la complexité de l'influence des paramètres sur les performances d'un réseau.

Références :

[1] Lyu, Yuting, Junghui Chen, et Zhihuan Song. 2019. « Image-Based Process Monitoring Using Deep Learning Framework ». Chemometrics and Intelligent Laboratory Systems 189 (juin): 8-17. https://doi.org/10.1016/j.chemolab.2019.03.008.

[2] Kim, Yoon. 2014. « Convolutional Neural Networks for Sentence Classification ». arXiv. https://doi.org/10.48550/arXiv.1408.5882.

[3] Sak, Haşim, Andrew Senior, Kanishka Rao, et Françoise Beaufays. 2015. « Fast and Accurate Recurrent Neural Network Acoustic Models for Speech Recognition ». arXiv. https://doi.org/10.48550/arXiv.1507.06947.

[4] Cui, Chenhao, et Tom Fearn. 2018. « Modern Practical Convolutional Neural Networks for Multivariate Regression: Applications to NIR Calibration ». Chemometrics and Intelligent Laboratory Systems 182 (novembre): 9-20. https://doi.org/10.1016/j.chemolab.2018.07.008.

Vers la conception d'un spectromètre infrarouge peu onéreux afin de valoriser les fourrages produits à la ferme

[1*] Nicolas BIRON, [2*]Charlène BAROTIN, [3*]Philippe BARRE, [4*]Jean-Michel ROGER

[1]INRAE, URP3F et ITAP, 86600 Lusignan – France et 34000 Montpellier – France
[2]INRAE, URP3F, 86600 Lusignan – France
[3]INRAE, URP3F, 86600 Lusignan – France
[4]INRAE, ITAP, 34000 Montpellier – France

Email :
nicolas.biron@inrae.fr
charlene.barotin@inrae.fr
philippe.barre@inrae.fr
Jean-Michel.Roger@inrae.fr

***Mots-clés** : spectrométrie proche infrarouge, transfert d'étalonnage, régression linéaire, projection orthogonale*

L'apport nutritionnel des cultures à apporter aux différents animaux dédiés à l'élevage est un point crucial pour l'éleveur. Cette valeur alimentaire peut être estimée à partir de variables biochimiques en utilisant la spectroscopie proche infrarouge (NIR) comme méthode indirecte. Dernièrement, plusieurs spectromètres ont été acquis au sein de l'unité de recherche, mais donnant des résultats différents. L'objectif de cette présentation est de traiter ce problème en comparant le potentiel de 2 outils de transfert d'étalonnage sur la calibration des différents spectromètres NIR en prédisant la teneur en matière grasse de farines de maïs sur 2 spectromètres différents. Un outil de transfert d'étalonnage utilise la régression linéaire pour corriger le spectre mesuré alors que l'autre utilise une projection orthogonale.

La base de données utilisée provient de chez CARGILL et comporte 80 spectres NIR acquis sur 3 différents spectromètres. Dans un premier temps, un instrument d'analyse sur les 3 disponibles a été choisi comme étant la référence sur lequel calibrer un autre instrument. Ensuite, le prétraitement optimal des spectres a été déterminé pour chaque variable chimique. Enfin, la matière grasse a été prédite après l'application de la Piecewise Direct Standardisation (PDS) et de Transfer by Orthogonal Projection (TOP). L'erreur de prédiction (RMSEP), la linéarité (R2) ainsi que le nombre de variables latentes (LV) utilisées ont été comparées.

En conclusion, PDS fournit un modèle de prédiction moins bon (RMSEP = 0.12, R2 = 0.93 , LV = 4) que TOP (RMSEP = 0.08, R2 = 0.96, LV = 5) pour l'élaboration du modèle de la matière grasse.

Development of prediction models for Fe, Al, C and N in costarician soils using NIR spectroscopy

[1,2,3*]Ulysse CHABROUX, [3] Juan-Carlos MENDEZ FERNANDEZ, [4]Aurélie CAMBOU, [5,6,7] Gilles CHAIX, [1,8,9] Julien DEMENOIS,

[1] AIDA, Univ Montpellier, CIRAD, Montpellier, France
[2] ENS-PSL, département de Géosciences, 75005 Paris 5e, France
[3] CENTRO DE INVESTIGACIÓNES AGRONÓMICAS, Universidad de Costa Rica, Laboratorio de Suelos y Foliares, San Pedro, Montes de Oca, Costa Rica
[4] Eco&Sols, Univ Montpellier, CIRAD, INRAE, IRD, Institut Agro, 34060 Montpellier, France
[5] CIRAD, UMR AGAP Institut, F-34398, Montpellier, France
[6] AGAP Institut, Univ Montpellier, CIRAD, INRAE, Institut Agro, F-34398 Montpellier, France
[7] ChemHouse Research Group, Montpellier, France
[8] CIRAD, UPR AIDA, Turrialba 30501, Costa Rica.
[9] CATIE, Centro Agronómico Tropical de Investigación y Enseñanza, Turrialba 30501, Costa Rica.

Email : mail ulysse.chabroux@ens.psl.eu

***Mots-clefs** : Sols tropicaux, Modèle PLSR, Spectroscopie Proche-Infrarouge, Carbone du sol, Fer du sol, Aluminium du sol.*

La spectroscopie infrarouge a déjà fait les preuves de son efficacité pour l'étude des sols. Elle permet d'obtenir à moindre coût des informations clés sur la chimie et la physique de tous les types de sols, et beaucoup plus rapidement que les analyses chimiques en laboratoire utilisées plus classiquement. La littérature fournit des exemples de modèles de prédiction dans le cas de sols tropicaux, et particulièrement pour l'étude de sols volcaniques tropicaux (Andosols).

Cependant, afin de pouvoir utiliser la spectroscopie infrarouge pour prédire les caractéristiques physico-chimiques des sols, il est nécessaire de développer un modèle de prédiction basé à la fois sur les mesures de référence de laboratoire et sur les spectres infrarouges d'échantillons d'étalonnage.

Dans le cas d'études locales ou régionales, la constitution d'une base de données (incluant à la fois les spectres et les mesures de références) suffisamment grande pour permettre la construction d'un modèle sans biais et avec une erreur de prédiction faible peut être un véritable défi. Ajouter dans le modèle, en plus des spectres infrarouges, des variables indépendantes facilement mesurables sur le terrain qui covarient avec les variables d'intérêt, comme l'altitude ou la profondeur de prélèvement, constitue une voie potentielle, et généralement peu explorée, pour améliorer les performances d'un modèle de prédiction.

Dans cette étude, nous construisons un modèle de prédiction régional pour le C, le N, le Fe et l'Al grâce à des échantillons d'Andosols (collectés à deux profondeurs ou plus sur chaque site) issus des agrosystèmes de la région de Cartago, au Costa Rica. Les mesures de Fe et d'Al en laboratoire ont consisté en une extraction à l'oxalate d'ammonium, et les mesures de C et N ont été effectuées par combustion sèche. La méthode utilisée pour l'étalonnage est la régression des moindres carrés partiels (PLSR) entre des mesures de laboratoire et les spectres infrarouges de 108 échantillons de sols prélevées sur le flanc sud du volcan Irazú, une zone intensivement utilisée pour le maraîchage et l'élevage bovin (prairies), et où les enjeux de gestion des sols sont importants. En effet, cette région présente une utilisation de phytosanitaires et d'engrais record au Costa Rica, ce qui entraîne une pollution des sols et des nappes phréatiques environnants.

La zone d'échantillonnage, constituée par 39 sites de collecte, est large de 16 km d'Est en Ouest et de 14 km du Sud au Nord et couvre un important gradient d'altitude (allant de 500m a.s.l. à 3500m a.s.l.), associé à un important gradient de températures moyennes et de précipitations, ce qui permet d'entraîner le modèle de prédiction sur une gamme de valeurs de paramètres physico-chimiques représentatifs des variations de la région.

Pour chaque élément (C, N, Fe, Al), un modèle de prédiction indépendant a été effectué. Pour chacun des quatre éléments, sept prétraitements spectraux ont été testés, ainsi que l'ajout un à un de quatre paramètres environnementaux (i.e. altitude, profondeur de l'échantillon, altitude et profondeur, sans ajout de paramètre environnemental). Ainsi, pour chaque élément, 28 modèles PLSR ont été construits. Les échantillons ont été séparés en deux sous-groupes, en utilisant l'algorithme Duplex sur les spectres prétraités et en conservant les différents horizons d'un même point d'échantillonnage dans un même sous-groupe : l'un pour l'étalonnage du modèle (83 échantillons) et l'autre pour la validation du modèle (25 échantillons). L'étalonnage a été réalisé en validation croisée avec 3 groupes répétés 10 fois en respectant les règles précédentes.

La performance d'un modèle de prédiction a été évaluée en regardant la Déviation Résiduelle de Prédiction (RPD). Dans le domaine du sol, un modèle est considéré satisfaisant si le RPD est supérieur à 1,6. Dans la majorité des cas, ajouter une ou deux variables environnementales aux spectres IR dans le modèle de prédiction améliore sa qualité en validation. L'Al est l'élément le mieux prédit, avec un RPD allant jusqu'à 2,8 en utilisant les spectres seuls et jusqu'à 3,0 en y ajoutant l'altitude et la profondeur comme co-variables. Le RPD de validation est de 2,1 pour la prédiction de C et de 2,0 pour la prédiction de N en ajoutant l'altitude aux spectres, contre respectivement 1,6 et 1,5 en utilisant les spectres seuls. Le Fe est l'élément le moins bien prédit par le modèle, avec un RPD ne dépassant pas 1,6 avec seulement les spectres IR, et de 1,8 avec l'ajout de l'altitude et de la profondeur.

Le C et le N sont très liés à la présence de matières organiques et de racines dans les sols. Ils sont donc concentrés près de la surface, et positivement corrélés à la température (donc inversement corrélés à l'altitude). Le Fe est présent sous de nombreuses formes à l'état naturel dans les Andosols, et la méthode d'extraction à l'oxalate permet d'obtenir la totalité du Fe du sol, indépendamment des proportions de ces différentes formes, ce qui pourrait expliquer les difficultés observées à prédire cet élément seulement à partir des spectres infrarouges.

Au-delà des résultats encourageants de notre étude, celle-ci souligne aussi l'intérêt d'ajouter des variables environnementales facilement accessibles sur le terrain, comme l'altitude ou la profondeur d'échantillonnage du sol, pour améliorer la qualité des modèles de prédiction. Toutefois, la prédiction avec ces modèles nécessitera d'avoir enregistré les données des variables environnementales des nouveaux échantillons à prédire.

Ces travaux s'inscrivent dans une volonté pour les différents laboratoires du Costa Rica d'être en mesure d'effectuer des prédictions de propriétés du sol grâce à la spectrométrie infrarouge : la présente étude, issue d'une collaboration entre l'Université du Costa Rica et le CIRAD, mais également dans le cadre d'une campagne de caractérisation des sols et de cartographie du territoire (utilisant les outils spectroscopiques) menée par l'Instituto Nacional de Innovación y Transferencia en Tecnología Agropecuaria (INTA).

Traitement conjoint des informations spectrales et spatiales des images hyperspectrales pour la détection du feu bactérien sur des plants de pommiers

Belal Gaci[1,2,3], Florent Abdelghafour[2,3], Silvia Mas-Garcia[2,3], Marine Louargant[1], Yohana Laloum, Ryad Bendoula[2,3], Jean-Michel Roger[2,3].

[1] *CTIFL, France*
[2] *ITAP-INRAE, Institut Agro, University Montpellier, Montpellier, France*
[3] *ChemHouse Research Group, Montpellier, France*

Email : Belal.gaci@ctifl.fr

Mots-clés : *Imagerie hyperspectrale, classification, méthode multi-bloc, détection du feu bactérien sur les plants de pommiers.*

Ce travail présente une méthodologie pour le traitement des images hyperspectrales (Figure 1). L'approche consiste à extraire les informations spatiales et spectrales des zones d'intérêt ou "imagettes" dans l'image hyperspectrale. L'objectif est de caractériser chaque imagette à l'aide d'un ensemble de scores. Les imagettes sont des carrés de largeur impaire, mettant ainsi l'accent sur le pixel central.

Pour extraire les informations spatiales et spectrales, deux approches sont utilisées. La première approche se concentre sur la dimension spatiale. Elle réduit la dimension spectrale en utilisant l'analyse en composantes principales (ACP) sur un ensemble d'apprentissage, transformant ainsi les imagettes en imagettes monochromatiques. Ensuite, les caractéristiques spatiales sont extraites à l'aide de l'analyse de texture basée sur les tenseurs de structures, ce qui permet d'obtenir trois indices spatiaux caractérisant la texture de chaque imagette.

La deuxième approche se concentre sur la dimension spectrale. Les imagettes sont dépliées pour obtenir des matrices, où le nombre de lignes correspond au nombre de pixels dans chaque imagette, et le nombre de colonnes correspond au nombre de longueurs d'onde. Ensuite, le spectre moyen de chaque imagette est choisi comme signature spectrale.

Les informations spatiales et spectrales obtenues par les deux approches sont fusionnées en utilisant une fusion supervisée avec la méthode ROSA-DA. Cette méthode a été appliquée à la discrimination du feu bactérien du pommier. L'étude a utilisé 64 images hyperspectrales dans la plage VNIR [400-1000nm] de feuilles de pommiers, comprenant des échantillons sains, infectés et sous stress hydrique.

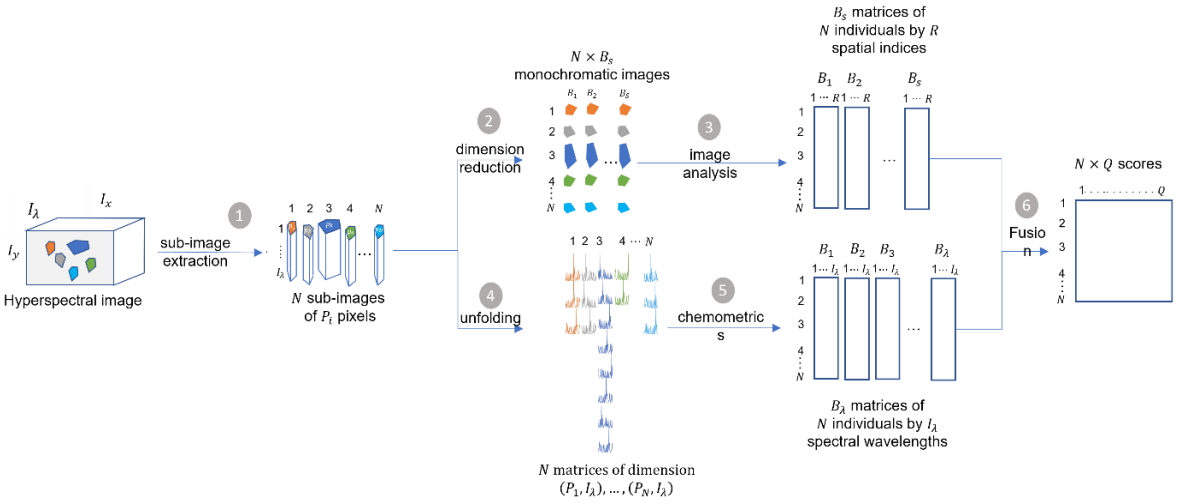


Figure 1: General scheme of the spatial-spectral method of processing hyperspectral images.

Segmentation non supervisée d'image hyperspectrale dans l'espace des matrices de covariance

[1*] Zexing YAO, Florent Abdelghafour

[1]ITAP, Univ Montpellier, INRAE, Institute Agro, Montpellier, France

Email : zexing.yao@inrae.fr

***Mots-clefs** : Imagerie hyperspectrale, Segmentation non supervisée*

L'imagerie hyperspectrale permet de mesurer un grand nombre de longueurs, notamment dans le visible et l'infrarouge. Nous souhaitons explorer une nouvelle méthode pour évaluer et segmenter des images hyperspectrale de végétation avec l'exemple de feuilles de tomate à l'aide de l'espace des matrices de covariance.

Tout d'abord, il faut réaliser une décomposition l'imagerie hyperspectrale, nous choisissons deux méthodes ACP et ondelettes. Ensuite, Il faut modéliser la distribution de la covariance des scores résultants de l'ACP et des ondelettes. Pour ce faire on utilise l'algorithme d'Expectation-Maximisation pour classifier les caractères dans chaque famille de covariance qui correspondent aux classes dans l'image. Finalement, nous réalisons une analyse des paramètres de la méthode pour la qualité de segmentation des images.
Pour la méthode à l'aide de la matrice covariance, on peut bien segmenter les images hyperspectrale et visualiser les détails en direct.

Figure 1 : Image hyperspectral de tomate

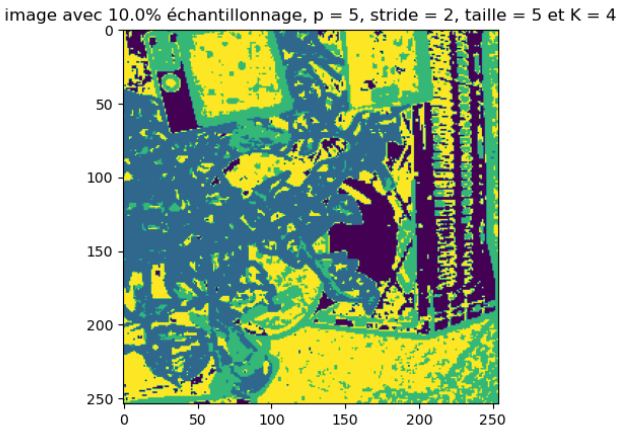


Figure 2 : La segmentation d'image

Transfert d'étalonnage entre différents spectromètres en réflexion diffuse dans le proche infrarouge appliqué aux sols

[1*] Vova Martirosyan, [1] Aurélie CAMBOU, [1] Bernard G. BARTHÈS, [2*] Jean-Michel ROGER

[1]Eco&Sols, Université de Montpellier, CIRAD, INRAE, IRD, Institut Agro, 34060 Montpellier – France
[2] ITAP, INRAE, Institut Agro, University Montpellier, 34196, Montpellier – France

Orateur : Vova Martirosyan (stage M2) – vovmart78@gmail.com ; encadrante : Aurélie Cambou – aurelie.cambou@ird.fr

Mots-clefs : *carbone organique, sols tropicaux, orthogonalisation*

Le carbone organique des sols (COS) est l'un des principaux piliers de la fonctionnalité des sols. En effet, le COS joue un rôle prépondérant et positif sur leur fertilité physique (e.g., résistance à l'érosion, développement racinaire), chimique (fourniture de nutriments, pouvoir tampon) et biologique (activités et diversités microbienne, animale et végétale). De plus, la quantification du COS suscite beaucoup d'intérêt car l'augmentation ou la perte de COS a un effet direct sur la concentration en CO2 dans l'atmosphère, et donc sur le climat. Depuis une trentaine d'années, l'intérêt de la spectroscopie en réflexion diffuse dans le proche infrarouge (SPIR, 800-2500 nm) utilisée en condition de laboratoire a été rapporté dans la littérature pour quantifier le COS. En effet, il s'agit d'un outil d'acquisition rapide, non destructif et peu coûteux, qui n'implique que peu de préparation d'échantillon. De nombreux modèles de prédiction des teneurs en COS (gC.kg-1 sol) d'échantillons de sol ont alors été construits à partir des spectres dans le proche infrarouge (PIR) acquis par différents appareils de mesure et pour différents jeux d'échantillons de sol. Aujourd'hui, dans un objectif d'ouverture des données, un fort enjeu concerne la mise à disposition et l'interopérabilité de bases de données spectrales obtenues avec différents appareils de mesure, ce qui permettrait notamment d'élargir nos capacités de prédiction du COS dans une diversité de contextes. Les perturbations liées à la diversité des instruments limitent l'interopérabilité des bases de données et différentes stratégies de correction permettent de réduire leur impact. Une méthode de correction simple, dite de spiking, consiste à enrichir le jeu d'étalonnage scanné avec un appareil A (spectres A) avec quelques échantillons du jeu test scanné avec un appareil B (spectres B) avant la construction d'un modèle de prédiction d'une variable d'intérêt. La standardisation directe par segment (piecewise direct standardisation ; PDS) est une autre méthode de correction a priori utilisant un jeu d'échantillons de transfert scanné par les deux appareils A et B. Une fonction de correction mathématique est alors construite pour ajuster les spectres B aux spectres A au sein de ce jeu de transfert. Puis un modèle de prédiction de la variable d'intérêt construit à partir d'un jeu d'étalonnage scanné avec le spectromètre A peut être appliqué sur les spectres B d'un jeu test corrigés par la fonction PDS. La correction de biais et pente (CBP) est quant à elle une méthode a posteriori qui modifie les prédictions : les paramètres de correction sont obtenus par régression linéaire entre variable prédite à partir des spectres A et variable prédite à partir des spectres B sur le jeu d'échantillons de transfert. Cette fonction peut ensuite être utilisée pour corriger la variable d'intérêt prédite par un modèle construit à partir de spectres A et appliqué sur des spectres B. Une autre approche consiste à définir dans l'espace des spectres A et B du jeu de transfert, un sous-espace impacté par les perturbations qui est alors retiré de l'espace initial : on définit alors le sous-espace résiduel, non impacté par les perturbations. Le modèle de prédiction de la variable d'intérêt est ensuite développé dans ce sous-espace résiduel (méthodes d'orthogonalisation). L'objectif de la présente étude est de tester et d'optimiser ces différentes approches de transfert d'étalonnage pour assurer l'interopérabilité de bases de données spectrales obtenues avec deux spectromètres PIR afin de prédire la teneur en COS.

Cette étude repose sur un total de 178 échantillons de sols issus de 24 sites (au Brésil, Bénin, Burkina Faso, Cameroun, Congo, Côte d'Ivoire, Madagascar, Mali et Sénégal) caractérisés par des textures variées (de sableuse à argileuse) et différentes profondeurs. Les spectres PIR ont été acquis avec deux spectromètres proche infrarouge (marques Foss et ASD) sur échantillons séchés à l'air et tamisés à 2 mm. La teneur en COS a été analysée sur échantillons tamisés à 2 mm puis broyés à 0,2 mm à l'aide d'un analyseur élémentaire

CHN (combustion sèche). Ce jeu de données a ensuite été divisé entre (i) le jeu d'étalonnage permettant de construire le modèle de prédiction de la teneur en COS (67 échantillons), (ii) le jeu de transfert permettant de construire le modèle de transfert d'étalonnage (32 échantillons), et (iii) deux jeux d'échantillons tests indépendants (ne provenant pas des sites utilisés pour l'étalonnage et le transfert) : le premier provenant du Burkina Faso (texture sableuse ; n = 38) et le second du Brésil (texture argileuse ; n = 41). L'efficacité des quatre approches de transfert présentées ci-dessus (i.e. spiking, PDS, CBP et une méthode d'orthogonalisation) sera donc évaluée et comparée à travers cette étude.

Ce travail ouvre le champ d'application des méthodes de transfert d'étalonnage sur les sols, à travers un jeu d'échantillons aux propriétés hétérogènes.

Prédiction de l'indice d'iode pour des charges lipidiques par SPIR : apport de la sélection de variables

[1,2] Chloé GICQUEL, [1] Joana FERNANDES, [1] David GONCALVES, [1] Marion LACOUE-NEGRE

[1] *IFP Energie nouvelles, 69360 Solaize – France*
[2] *ECPM, 67000 Strasbourg – France*

Email : chloe.gicquel@ifpen.fr

Mots-clefs : *PIR, PLS, indice d'iode, charges lipidiques, biocarburants*

1. Introduction

IFP Energies nouvelles travaille depuis plusieurs années sur le développement et l'optimisation de procédés capables de produire des biocarburants pour tous types de transports par hydrotraitement de charges lipidiques. Les charges considérées dans ces procédés peuvent être de première génération telles que les huiles végétales alimentaires (huile de colza, l'huile de soja, l'huile de tournesol, etc.), de seconde génération comme les huiles de cuisson usagées, les graisses animales, les huiles de poisson, etc., ou encore des déchets et/ou sous-produits de l'industrie (par exemples l'huile de pin de l'industrie papetière). Afin de pouvoir convertir efficacement ces charges, il est indispensable de les caractériser. L'indice d'iode, l'indice d'acide, la composition en glycérides, la composition en acides gras et la viscosité sont des paramètres importants caractérisant les charges lipidiques et qui peuvent influencer la qualité des biocarburants produits. Des méthodes standardisées existent pour mesurer ces propriétés, néanmoins ces méthodes sont souvent, destructives, coûteuses et chronophages. La spectroscopie couplée à la chimiométrie apparait alors comme une solution plus rapide et moins coûteuse[1]. Dans le cadre de cette étude, la spectroscopie proche infra-rouge (PIR) couplée à de la partial least squares (PLS) a été utilisée pour prédire l'indice d'iode des charges lipidiques. Afin d'optimiser au maximum notre modèle, différentes méthodes de pré-traitement des spectres ont été testées. De plus, l'apport d'une sélection de variables ainsi que l'impact de la température d'acquisition des spectres ont été étudiés.

2. Matériels et méthodes

Au total, 34 échantillons représentant une large gamme de charges lipidiques (huiles végétales, graisses animales, huile de poisson, huile de friture, …) ont été sélectionnés pour cette étude. Les spectres ont été acquis sur un spectromètre ABB FT-NIR MB3600 couvrant une gamme spectrale de 4000 à 12000cm^{-1}. Des cellules avec un trajet optique de 8 mm ont été utilisées. Les échantillons ont été chauffés à 70°C pour chaque mesure. Les spectres ont été enregistrés avec une résolution spectrale de 4 cm^{-1}, et 80 scans pour chaque spectre.

L'indice d'iode a été mesuré sur chaque charge suivant la méthode standard EN ISO 3961[2]. Les incertitudes sur cette méthode sont les suivantes : ±1.1 pour Y < 20 g/100g, ± 4.5 pour 20 g/100g < Y < 100 g/100g et ±7.5 pour Y > 100 g/100g. La gamme d'indice d'iode sur ces échantillons s'étendait de 1.8 à 185 g/100g.

Le logiciel MATLAB et la PLS Toolbox (Eigenvector) ont été utilisés afin de développer les modèles PLS. Un choix a été fait de scinder la base de données en une base de calibration et une base de validation. Ainsi, 26 spectres ont ainsi été utilisés pour développer les modèles PLS, qui ont ensuite été appliqués sur 8 spectres de la base de validation. Afin d'évaluer les performances de chaque modèle développé, les critères statistiques suivants ont été calculés : RMSEC, RMSECV et RMSEP, ainsi que les coefficients de détermination (R^2).

3. Résultats

Différents pré-traitements ont été appliqués sur les spectres, en considérant la gamme spectrale 4464-12000 cm^{-1} : Automatic Weighted Least Squares Baseline (AWLS), Savitsky-Golay Derivative (1er ordre de dérivé), Detrend, Multiplicative Signal Correction (MSC) et Standard Normal Variate (SNV). Les meilleures performances ont été trouvées en utilisant la correction de ligne de base AWLS, suivie d'un centrage par la moyenne. Les résultats obtenus pour les RMSEC, RMSECV et RMSEP sont respectivement 2.6, 5.5 et 2.7 et pour les coefficients de détermination : 0.997, 0.986 et 0.992. Ces résultats sont conformes avec l'incertitude de la méthode de référence (Figure 1)

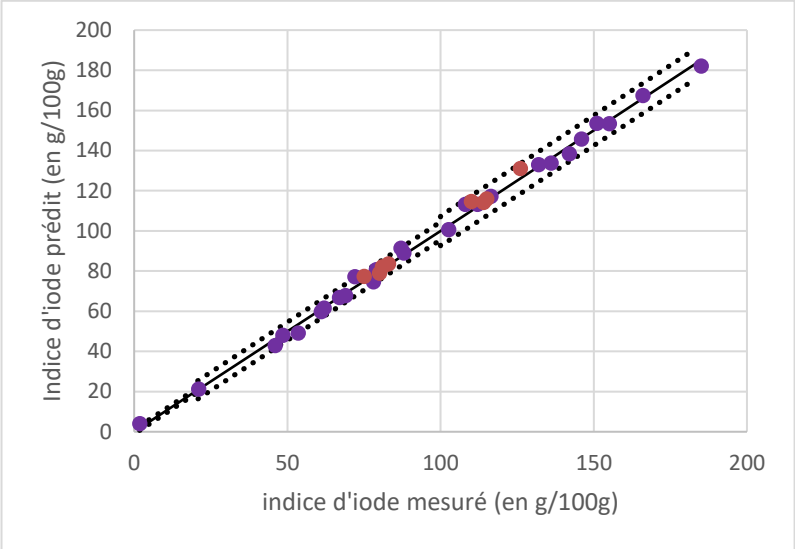


Fig 1. Droite de parité de l'indice d'iode avec des points en validation. En orange, les points des spectres en validation et en violet, les points des spectres en calibration. La ligne en pointillés représente l'incertitude de la méthode de référence.

Néanmoins, les résultats sur la cross-validation restent insatisfaisants. En effet quelques spectres se retrouvent en dehors de l'intervalle de confiance établi par la méthode de référence lors de la cross-validation, comme visible sur la Figure 2.

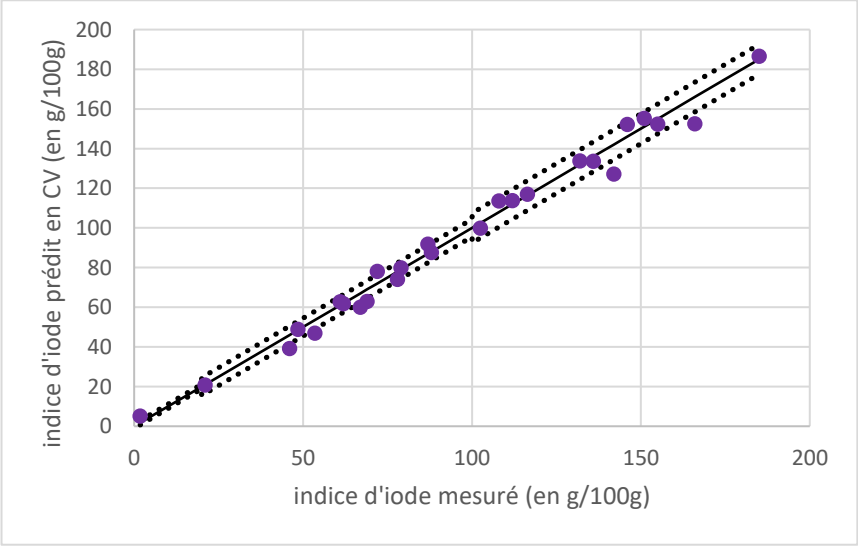


Fig 2. Droite de parité de l'indice d'iode prédit en cross-validation. En pointillés l'incertitude de la méthode de référence

Afin d'améliorer la cross-validation, deux méthodes de sélection de variables, interval PLS et Genetic Algorithm ont été testées et comparées au modèle précédent. Il en résulte de meilleurs résultats en cross-validation. En effet, la moyenne des ratios RMSECV/RMSEC de chaque modèle construit avec sélection de variables est de 1.6 contre 2.2 pour le modèle sans sélection de variables. Néanmoins, les résultats montrent que les modèles ont tendance à sur-paramétrer avec des différences entre RMSEC et RMSEP beaucoup plus importantes par rapport au modèle sans sélection de variables. De ce fait, les modèles créés par sélection de variables vont manquer de robustesse dans le temps.
Enfin, pour vérifier la robustesse de notre modèle face à des variabilités d'acquisition de spectres, les indices d'iode d'une base de données composée uniquement de spectres acquis à 40°C ont été prédits. Les résultats montrent que la température a un impact minime sur la prédiction de l'indice d'iode.

4. Conclusion

La faisabilité de prédiction de l'indice d'iode par PIR et chimiométrie a été évaluée au cours de cette étude. La méthode utilisée permet de prédire de manière satisfaisante l'indice d'iode avec une erreur proche voire inférieure à celle de la méthode de référence. Ainsi, pour prédire l'indice d'iode la spectroscopie PIR, méthode rapide et non destructive, peut remplacer la méthode de référence. Pour les autres propriétés d'intérêt, des modèles sont en cours de développement ou à développer, il reste donc à prouver que la spectroscopie pourra alors remplacer ces autres méthodes d'analyses.

Références :

1 Moreira, S. A.; Sarraguça, J.; Saraiva, D. F.; Carvalho, R.; Lopes, J. A. Optimization of NIR spectroscopy based PLSR models for critical properties of vegetable oils used in biodiesel production. Fuel [Online] 2015, 150, 697–704.
2 EN ISO 3961 « Corps gras d'origines animale et végétale – Détermination de l'indice d'iode »

Suivi in situ de cristallisations par Spectroscopie Résolue Spatialement (SRS)

[1*]Romain KERSAUDY, [1]Emilie GAGNIERE, [2] Noémie CAILLOL, [1] Didier COLSON, [1]Stéphane LABOURET, [1] Denis MANGIN

[1] Université Claude Bernard Lyon 1, CNRS, LAGEPP UMR 5007, 69100 Villeurbanne – France
[2] Axel'One, 69360 Solaize – France

Email : romain.kersaudy@univ-lyon1.fr

Mots-clefs : *Cristallisation, Suivi in situ, Spectroscopie Résolue Spatialement,*

Introduction

La cristallisation est un processus de séparation et de purification utilisé dans de nombreux domaines industriels, conduisant à la production de cristaux aux spécifications bien définies en fonction de l'application souhaitée. Ce processus comprend plusieurs mécanismes tels que la nucléation, la croissance, l'agglomération. Par leur complexité, ces mécanismes nécessitent une supervision des paramètres tant physiques que chimiques pour permettre un bon contrôle de la cristallisation. De nombreuses études ont été menées pour contrôler la concentration de la phase liquide pendant la cristallisation, montrant l'efficacité de la spectroscopie ATR-FTIR (Attenuated Total Reflectance - Fourier Transformed Infrared). En ce qui concerne le contrôle des paramètres physiques, il est beaucoup plus difficile de trouver une méthode permettant des analyses précises et robustes. Nous proposons d'étudier les capacités de la Spectroscopie Résolue Spatialement, dans la zone du proche infrarouge pour prédire le taux de solide et les descripteurs de la distribution de taille des cristaux pendant des cristallisations par refroidissement de l'acide adipique dans un milieu aqueux.

Matériels et méthodes

Deux approches ont été envisagées pour l'élaboration de modèles de prédiction. Une première consiste en l'utilisation de données mesurées sur des échantillons synthétiques. La base de données a alors été construite à partir de suspensions de cristaux non sphériques impliquant différentes concentrations en solide (1 – 16 %m/m) et différentes distributions de taille de cristaux (0-90 µm, 90-125 µm, 125-200 µm, 200-315 µm). La seconde approche consiste en l'utilisation de données mesurées au cours de plusieurs cristallisations par refroidissement en variant la concentration initiale et les propriétés de la semence introduite (nature, masse).
Les mesures pour les deux approches ont été effectuées dans un réacteur agité à double-enveloppe de 2,5 L équipé d'une sonde ATR-FTIR pour mesurer la concentration de la phase liquide, d'une sonde SamFlex SRS (spectres PIR sur 4 angles : 3 en transmission et 1 en réflexion), d'une sonde PT100 et d'une sonde vidéo. La distribution de la taille des cristaux des échantillons a été déterminée par une mesure ex situ à l'aide du granulomètre laser Mastersizer 3000.

Résultats et discussions

La première approche a mis en évidence la difficulté d'élaborer des modèles de prédiction des caractéristiques du solide. Les méthodes linéaires (PLS) et non-linéaires (SVM-R) n'ont pas permis d'établir une corrélation entre les données de référence (taux de solide et descripteurs de taille) et les spectres SRS. Les échantillons synthétiques sont sujets à des mécanismes difficiles à maitriser, tel que l'agglomération des petits cristaux, ce qui engendre une difficulté supplémentaire à la modélisation ainsi qu'un biais par rapport aux cristaux obtenus lors des cristallisations (différences de faciès, d'état de surface etc …). Toutefois, cette approche a permis de mettre en avant la nécessité d'utiliser une base de données large et variée (avec des distributions de taille de cristaux monomodales et multimodales) pour établir un modèle convenable de prédiction du taux de solide par SVM-R.

La seconde approche s'est révélée beaucoup plus adéquate pour la prédiction du taux de solide. La base de calibration a été établie à partir de données acquises au cours de 5 cristallisations. Les valeurs de référence du taux de solide ont été calculées par bilan de matière à partir des concentrations initiales mesurées par extrait sec et des valeurs de solubilité de l'acide adipique dans l'eau. Un modèle PLS à 4 facteurs a été établi permettant une RMSECV de 1,0 %m/m. Ce modèle a par la suite été utilisé pour prédire le taux de solide au cours de 3 autres expériences de cristallisation révélant de bons résultats de prédiction pour deux d'entre eux. En ce qui concerne la troisième expérience, une sous-estimation du taux de solide a été notifiée par rapport au taux théorique calculé, en raison d'un manque de représentation de ce type de cristallisation (ajout de semence humide) dans le modèle de calibration. En effet, il a été observé lors des essais de cristallisations que la nature de la semence introduite a un impact non négligeable sur le déroulé de la cristallisation.

Concernant la prédiction de la taille des cristaux, des modèles linéaires (PLS) et non linéaires (SVM,) ont été testés sans permettre l'élaboration d'un modèle satisfaisant. Le choix d'un descripteur unique des distributions de taille de cristaux est très complexe, notamment lorsqu'il s'agit de distributions multimodales. A cela, s'ajoute la forte non-linéarité des données recueillies avec la sonde SRS causée par la diffusion complexe de la lumière sur des objets de grande taille (plusieurs centaines de micromètres) et accentuée par les caractéristiques physiques des cristaux (faciès, état de surface, opacité ou transparence). Pour conclure, la SRS basée sur le proche infrarouge permet l'acquisition de données très riches mais leur complexité rend difficile l'élaboration de modèle de prédiction. Le taux de solide peut être prédit au cours de cristallisations par refroidissement, et l'ajout de données supplémentaires pourraient permettre d'obtenir un modèle PLS plus précis et robuste. Néanmoins, la prédiction de taille des cristaux est beaucoup plus complexe et nécessiterait l'utilisation d'algorithmes plus puissants, mais nécessitant un grand nombre de données, pour corriger les fortes non-linéarités (Réseaux de neurones, Deep Learning).

Bénéfices et limites de la méthode de visualisation t-SNE pour la spectroscopie proche infrarouge

* François STEVENS, Vincent BAETEN, Juan Antonio FERNÁNDEZ PIERNA

CENTRE WALLON DE RECHERCHES AGRONOMIQUES, UNITÉ QUALITÉ ET AUTHENTIFICATION DES PRODUITS, 5030 Gembloux – Belgique

Email : f.stevens@cra.wallonie.be

Mots-clefs : *t-distributed stochastic neighbour embedding, analyse en composantes principales, effet batch, prétraitement, analyse exploratoire, interface visuelle*

L'algorithme de t-distributed stochastic neighbour embedding ou t-SNE [1] est une méthode non-linéaire de réduction de dimensions utilisée pour visualiser des données multivariées. Il permet de représenter un jeu de données de grande dimensionalité, tel qu'un ensemble de spectres infrarouge, sur un unique graphique typiquement à deux dimensions, et d'en révéler la structure locale et globale. Les tendances et groupement de points observés sur le graphique t-SNE peuvent être interprétés à la lumière de différentes variables quantitatives et/ou catégoriques associées à ces spectres, typiquement en les utilisant comme couleur, symbole ou taille de points, afin d'évaluer sommairement l'influence de ces variables sur le spectre.

t-SNE est une méthode très populaire dans la communauté du machine learning où elle a été appliquée dans de nombreux domaines, généralement dans le but de visualiser des jeux de données de grande taille [2], [3]. De nombreuses applications ont aussi été mises au point afin de fournir des services, en particulier dans le domaine de la médecine (identification de tumeurs) ou des biotechnologies, par exemple en utilisant des données de GC-MS [4]. En particulier, la méthode t-SNE s'est imposée dans le domaine de la transcriptomique comme la pierre angulaire de l'exploration et de l'analyse des données issues de séquençages ARN unicellulaires [5].

En spectroscopie vibrationnelle, t-SNE gagne en notoriété [6–8], mais l'analyse en composantes principales (PCA) reste de loin la méthode de référence pour l'analyse exploratoire et la réduction de dimension. Pourtant, les deux méthodes sont très différentes et chacune peut faire valoir quelques avantages comparatifs. Tandis que PCA est une méthode linéaire basée sur la variance globale, t-SNE est aussi capable de modéliser les relations non linéaires et les variations locales, tout en étant moins sensible aux valeurs extrêmes ou aberrantes. De son côté, contrairement à t-SNE, PCA fournit également des informations dans le domaine spectral sous forme de loadings et permet par ailleurs de projeter directement de nouvelles observations.

Pour l'utilisateur, un avantage notable de t-SNE est de permettre de visualiser le jeu de données au moyen d'une seule figure, là où une analyse PCA impliquerait d'inspecter les scores de plusieurs composantes principales. Néanmoins, une réduction de dimension aussi drastique a ses limites, et parfois des facteurs de variation dominants mais peu pertinents, tels qu'un « effet batch », peuvent masquer des effets d'ampleur très limitée, mais d'intérêt majeur. Différentes approches ont été mises en œuvre pour pallier à ce problème, soit en adaptant l'algorithme pour le rendre insensible à ces effets dominants [9], soit en fournissant des interfaces de visualisation incluant de nombreux indicateurs de performance et permettant une interprétation plus profonde [10, 11]. Néanmoins, ces approches viennent ajouter une couche de complexité et risquent dès lors de rebuter l'utilisateur novice de t-SNE pour qui la simplicité reste le principal attrait de la méthode.

Nous proposons ici une solution alternative simple basée sur une synergie entre les méthodes t-SNE et PCA et permettant d'exploiter les avantages respectifs de chacune. Concrètement, une PCA est d'abord appliquée avec un nombre de composantes principales suffisamment grand pour garantir que toute la variabilité pertinente soit bien captée par le modèle. Ensuite, le choix est donné à l'utilisateur d'exclure une ou un petit nombre de composantes principales soupçonnée(s) de masquer la variabilité pertinente. t-SNE est alors appliqué sur les spectres reconstruits à partir des autres composantes principales. Une interface permettant de visualiser les scores et loadings des composantes principales exclues et le résultat de

l'algorithme t-SNE sur les composantes conservées (le tout enrichi par de l'information sur les variables associées) permet d'appréhender simultanément les deux facettes du problème : un contrôle de la part de variabilité exclue et une synthèse de celle qui a été conservée, permettant de mettre en évidence un effet éventuel des variables d'intérêt sur le spectre. Actuellement, une interface interactive basée sur le langage R et le package Shiny [12] est en développement.

[1] L. van der Maaten and G. Hinton, "Visualizing Data using t-SNE," J. Mach. Learn. Res. 9, vol. 9, pp. 2579–2605, 2008.

[2] P. Hajibabaee, F. Pourkamali-Anaraki, and M. A. Hariri-Ardebili, "An Empirical Evaluation of the t-SNE Algorithm for Data Visualization in Structural Engineering," Proc. - 20th IEEE Int. Conf. Mach. Learn. Appl. ICMLA 2021, pp. 1674–1680, 2021, doi: 10.1109/ICMLA52953.2021.00267.

[3] Y. Wang, H. Huang, C. Rudin, and Y. Shaposhnik, "Understanding how dimension reduction tools work: An empirical approach to deciphering T-SNE, UMAP, TriMap, and PaCMAP for data visualization," J. Mach. Learn. Res., vol. 22, pp. 1–73, 2021.

[4] N. Kessler et al., "Learning to classify organic and conventional wheat - A machine learning driven approach using the MeltDB 2.0 metabolomics analysis platform," Front. Bioeng. Biotechnol., vol. 3, no. MAR, pp. 0–10, 2015, doi: 10.3389/fbioe.2015.00035.

[5] G. C. Linderman, M. Rachh, J. G. Hoskins, S. Steinerberger, and Y. Kluger, "Fast interpolation-based t-SNE for improved visualization of single-cell RNA-seq data," Nat. Methods, vol. 16, no. 3, pp. 243–245, 2019, doi: 10.1038/s41592-018-0308-4.

[6] N. Luo, X. Yang, C. Sun, B. Xing, J. Han, and C. Zhao, "Visualization of vibrational spectroscopy for agro-food samples using t-Distributed Stochastic Neighbor Embedding," Food Control, vol. 126, p. 107812, 2021, doi: 10.1016/j.foodcont.2020.107812.

[7] E. P. Mwanga et al., "Using transfer learning and dimensionality reduction techniques to improve generalisability of machine-learning predictions of mosquito ages from mid-infrared spectra," BMC Bioinformatics, vol. 24, no. 1, pp. 1–15, 2023, doi: 10.1186/s12859-022-05128-5.

[8] B. Xie, W. Njoroge, L. M. Dowling, J. Sulé-Suso, G. Cinque, and Y. Yang, "Detection of lipid efflux from foam cell models using a label-free infrared method," Analyst, vol. 407, pp. 5372–5385, 2022, doi: 10.1039/d2an01041k.

[9] P. G. Poličar, M. Stražar, and B. Zupan, "Embedding to reference t-SNE space addresses batch effects in single-cell classification," Mach. Learn., vol. 112, no. 2, pp. 721–740, 2023, doi: 10.1007/s10994-021-06043-1.

[10] A. Chatzimparmpas, R. M. Martins, and A. Kerren, "T-viSNE: Interactive Assessment and Interpretation of t-SNE Projections," IEEE Trans. Vis. Comput. Graph., vol. 26, no. 8, pp. 2696–2714, 2020, doi: 10.1109/TVCG.2020.2986996.

[11] R. J. Rainer, M. Mayr, J. Himmelbauer, and R. Nikzad-Langerodi, "Opening the black-box of Neighbor Embeddings with Hotelling's T2 statistic and Q-residuals," Chemom. Intell. Lab. Syst., vol. 238, 2023, doi: 10.1016/j.chemolab.2023.104840.

[12] W. Chang et al., "shiny: Web Application Framework for R." 2023, [Online]. Available: https://shiny.rstudio.com/.

Modélisation quantitative à partir d'images hyperspectrales proche infrarouge : cas pratique de l'igname

[1] Julien BOYER, [2] Karima MEGHAR, [1] Jordane POULAIN,

[1] ONDALYS, 34830 Clapiers – France
[2] CIRAD, PERSYST/UMR QUALISUD, 34090 Montpellier – France

Email : jboyer@ondalys.fr

Mots-clefs : *Imagerie hyperspectrale, Modélisation quantitative, SPIR, Igname.*

Les données traitées dans cette étude ont été acquises au cours d'un stage de fin d'études de Master 2, au sein de l'UMR Qualisud du CIRAD. Ce stage a été réalisé dans le cadre du projet RTBfoods qui a pour objectif d'assurer et de généraliser l'adoption de variétés améliorées de Racines, Tubercules et Bananes à cuire (RTB), et de ce fait, de renforcer la sécurité alimentaire. Le traitement de données a ensuite été poursuivi au sein de la société Ondalys.

L'igname est un tubercule qui constitue la base de l'alimentation de plus de 500 millions de personnes dans certains pays tropicaux d'Afrique, des Caraïbes, d'Océanie et d'Amérique latine. Il est une source majeure de glucides et de fibres, et est considéré comme ayant un grand potentiel pour améliorer la sécurité alimentaire.

Le stage, intitulé « Prédiction du comportement à la cuisson (bonne/mauvaise) de l'igname bouillie par imagerie hyperspectrale », avait pour objectif de réaliser des modèles de classification et de quantification pour la prédiction de la qualité de cuisson de l'igname à partir d'images hyperspectrales, afin de fournir aux sélectionneurs de variétés, des outils rapides et non destructifs pour le phénotypage à haut débit de l'igname. L'imagerie hyperspectrale proche infrarouge a été choisie car c'est une méthode rapide, simple et qui nécessite peu de préparation d'échantillon. De plus, par rapport à la spectroscopie classique, elle permet d'obtenir une mesure plus représentative et une visualisation spatiale des composés d'intérêt et donc de détecter des hétérogénéités.

Les images hyperspectrales (HSI) ont été acquises avec la caméra FX17 (SPECIM), qui a une gamme spectrale de 935 à 1720 nm. Le traitement des données a été réalisé sur MATLAB avec la PLS Toolbox et la MIA Toolbox (Eigenvector Research Inc).

Cette présentation concerne la prédiction de la teneur en matière sèche de l'igname frais, exprimée en pourcentage (%MS). Le but de cette étude est la comparaison de différentes approches de traitements d'images hyperspectrales ainsi que l'application de ces modèles à l'ensemble des pixels des images.

Une approche de segmentation non-supervisée des images hyperspectrales basée sur les métriques de l'espace des matrices de covariance

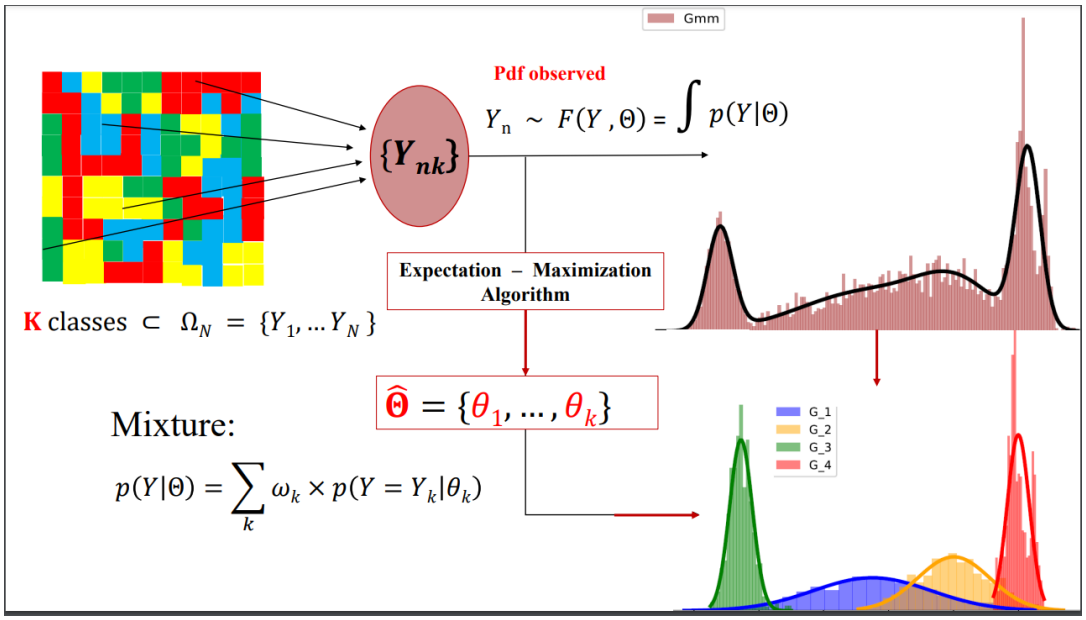
Florent Abdelghafour[1]

[1] *ITAP, Univ. Montpellier, INRAE, Institut Agro,34196 Montpellier, France*

***Keywords**: Hyperspectral imaging, unsupervised segmentation, Covariance matrices*

On propose de cartographier une image hyperspectrale comme un champs de matrices de covariances issus de scores locaux.

Il s'agit de considérer une image comme un ensemble de sous-régions centrées sur un pixel d'intérêt. Dans chaque ROI, on calcul des scores locaux, issus d'une transformation locale qui permet une réduction de dimension (e.g. PCA, MCR-ALS, ICA, Ondelettes, Fourier). A chaque ROI, correspond une matrice de covariance des scores associés. Puis, on redéfini les algorithmes EM, Kmeans et KNN grâce aux métriques associés à l'espace des matrices de covariances (variété riemannienne). Ainsi, on propose à la fois des descripteurs spectraux-spatios et on propose des outils pour les modéliser.



An overview on advanced chemometric approaches for NIR spectroscopy1*

Federico MARINI[1*]

[1]*UNIVERSITY OF ROME LA SAPIENZA, DEPT. CHEMISTRY, 00185 Rome – Italy*

Email : federico.marini@uniroma1.it

***Keywords** : Designed experiments, ANOVA-simultaneous component analysis (ASCA), Sequential preprocessing through Orthogonalization (SPORT), Class-modeling*

In the present communication, the potential of coupling some advanced chemometric tools with vibrational spectroscopy, in particular in the Near infrared range, to solve problems in different fields of application will be discussed and illustrated by means of selective examples. In particular, attention will be focused on methods which span all the phases of the analytical pipeline, from experimental design/sampling to variable selection/putative marker identification and validation of the whole process.

Methods like ANOVA-simultaneous component analysis (ASCA) [1] or ANOVA-Target projection (ANOVA-TP [2]), which allow to partition the variability of multivariate data matrices collected according to an underlying experimental design, to evaluate the significance of the different model terms and to provide a straightforward interpretation of the observed effects, will be illustrated.

On the other hand, some recently proposed preprocessing approaches aimed at reducing the impact of unwanted variability on the data, at the same time preserving as much as possible of the relevant information will also be discussed [3], together with the possibility of exploiting the multi-block concept to avoid the need for selection of the optimal pre-processing strategy [4].

Lastly, a few words will be spent on some recent development in the field of class modeling.

[1] Smilde, A.K., Jansen, J.J., Hoefsloot, H.C.J., Lamers, R.-J.A.N., van der Greef, J., Timmerman, M.E. (2005) Bioinformatics, 21, 3043-3048.
[2] Marini, F., de Beer, D., Joubert, E., Walczak, B. (2015) J. Chromatogr. A, 1405, 94-102.
[3] Rabatel, G, Marini, F, Walczak, B, Roger, J-M. (2020) J. Chemometr., 34, e3164.
[4] Roger, J-M, Biancolillo, A., Marini, F. (2020) Chemometr. Intell. Lab. Syst., 199, 103975.

Le point sur les plans d'expériences ; applications et review de leur utilisation pour le NIR

[1] Magalie Claeys-Bruno, [2]Yohann Clément, [2]Pierre Lantéri, [1]Michelle Sergent

[1] *INSTITUT MEDITERRANEEN DE BIODIVERSITE ET D'ECOLOGIE MARINE ET CONTINENTALE, AIX-MARSEILLE UNIVERSITE, UMR CNRS IRD AVIGNON UNIVERSITE, SITE DE L'ETOILE, 13013 Marseille, France*
[2] *UNIV LYON, UNIVERSITE CLAUDE BERNARD LYON 1, CNRS, ISA (INSTITUT DES SCIENCES ANALYTIQUES), UMR CNRS N°5280, VILLEURBANNE, France*

Email : mailmailmailmail

Mots-clefs : Plan d'expériences, NIR

1. Méthodologie des plans d'expériences

Dans la plupart des situations, l'expérimentateur se trouve devant une démarche classique que l'on peut schématiser ainsi : l'expérimentateur émet des hypothèses, il en déduit des conséquences et si les informations nécessaires pour vérifier ses hypothèses ne sont pas disponibles, il doit conduire une expérimentation afin d'obtenir ces informations. L'expérimentation ne peut pas être quelconque : elle doit fournir l'information désirée. Cette relation entre l'hypothèse testée et l'expérimentation qui doit fournir l'information nécessaire est essentielle.

Les problèmes scientifiques se répartissent en trois classes selon la question posée :

- QUI ? Quels sont les paramètres dont les variations entraînent une variation du phénomène ?
- COMMENT ? Quelle est la "forme" de la variation du phénomène lorsque les paramètres influents varient ?
- POURQUOI ? Quelles sont les explications « mécanistiques » qui peuvent permettre de relier les phénomènes et les paramètres étudiés ?

Tout expérimentateur a donc besoin d'outils qui l'aideront à exprimer au mieux ses objectifs et à établir des stratégies expérimentales optimales en fonction des moyens dont il dispose. Pour cela, il peut faire appel à un ensemble de méthodes et de modes de raisonnement qui ont pour but d'optimiser l'efficacité de sa recherche. Il est donc nécessaire d'utiliser une approche méthodologique qui permet non seulement de réduire le coût de l'expérimentation mais aussi d'établir une organisation optimale des expériences. Cette approche est basée sur le fait que toute la qualité de l'information acquise par l'expérimentation n'est pas contenue dans le résultat des expériences mais dans les conditions expérimentales. En outre, la qualité de l'information ne dépend pas du nombre d'expériences. Cette approche méthodologique de l'expérimentation utilise des outils mathématiques et statistiques, les plans d'expériences, qui permettent d'organiser les essais de telle façon que l'information obtenue soit celle désirée, avec la meilleure qualité possible.

Le choix du plan d'expériences adéquate est directement lié à l'objectif de l'étude et la plupart des études expérimentales appartiennent généralement à l'une des classes suivantes :

- le criblage des facteurs :

Cette phase consiste à rechercher très rapidement quels sont, parmi un ensemble de facteurs potentiellement influents, ceux qui le sont effectivement dans un domaine expérimental fixé.

- les études quantitatives des facteurs :

L'hypothèse d'additivité est abandonnée et les facteurs sont étudiés plus précisément en prenant en compte les effets d'interaction possibles entre les différents facteurs.

- les études d'optimisation :

La suite logique consiste à rechercher l'optimum d'une ou plusieurs réponses expérimentales. Pour cela, l'expérimentateur a besoin de connaître en n'importe quel point du domaine expérimental d'intérêt la valeur d'une ou plusieurs réponses expérimentales, ce qui lui permettra de déterminer la zone de compromis acceptable. Ceci est souvent obtenu en modélisant le phénomène, c'est-à-dire en le simplifiant sous la forme d'un modèle mathématique et l'expérimentation a alors pour but de déterminer la forme de ces relations et les valeurs des coefficients.

- les études de "formulation" :

Lorsque la nature des produits est fixée, le formulateur désire optimiser sa "formule", c'est à dire déterminer les "bonnes" proportions des constituants dans la formule. Pour cela, il a besoin d'avoir une connaissance de l'évolution des propriétés étudiées dans tout le domaine de variation des proportions des constituants.

2. Etude Bibliographique et applications au proche infrarouge

a. Etude Bibliographique

De nombreuses bases de données publiques ont été interrogées : Google Scholar, Web of Sciences, Scopus, Scinapse …

Les interrogations ont été faites sur la période 2012-2022 en croisant les mots clés Near –Infrared (NIR) et différents « Designs » :

Type de plans	Nombre de référence identifiées
Supersaturated Design	9
Full Factorial design	1230
Fractional design	502
Composite Design	1960
Box-Behnken Design	1000
Doehlert Design	65
D-optimal design	299
Mixture design	709
Taguchi	232
Space Filling Design	32

Tableau n°1 : Statistiques de Design of Experiment (DOE) et Near infrared

L'utilisation des plans d'expériences au sens large dans le domaine du NIR n'est pas négligeable ; à peu près tous les types de plan ont été utilisés, souvent associés à d'autre techniques chimiométriques (PLS, MCR-ALS, ICA,Space Filling Design …).

Type de plans	Nombre de référence identifiées
Supersaturated Design	9
Full Factorial design	1230
Fractional design	502
Composite Design	1960
Box-Behnken Design	1000
Doehlert Design	65
D-optimal design	299
Mixture design	709
Taguchi	232
Space Filling Design	32

Tableau n°2 : Statistiques de type de plans et objectifs

Les objectifs recherchés sont variés. On retrouve l'importance de la calibration de de l'optimisation mais aussi de la modélisation dans des buts bien précis comme le Qbd (Quality by Design) et le PAT (Process Analytical Technology).

b. Applications :

Nous avons retenu quelques exemples d'applications qui illustrent cette méthodologie. Une bibliographie des articles les plus cités dans les différentes catégories de plans d'expériences est disponible.

i. Plans de mélange, QbD et PAT :

Optimization of a pharmaceutical tablet formulation based on a design 2 space approach and using vibrational spectroscopy as PAT tool: https://doi.org/10.1016/j.ijpharm.2015.03.025

ii. Box-Behnken Design et PAT:

A novel application of pulsed electric field (PEF) processing for improving glutathione (GSH) antioxidant activity: https://doi.org/10.1016/j.foodchem.2014.04.027

iii. D-optimal Design, modélisation et PLS

Partial Least Squares, Experimental Design, and Near-Infrared Spectrophotometry for the Remote Quantification of Nitric Acid Concentration and Temperature: https://doi.org/10.3390/molecules28073224

Bibliography NIR and DOE: Highly cited papers in different items

"factorial design"

Jakkula, P., Reinikainen, M., Hästbacka, J., Loisa, P., Tiainen, M., Pettilä, V., Toppila, J., Lähde, M., Bäcklund, M., Okkonen, M., Bendel, S., Birkelund, T., Pulkkinen, A., Heinonen, J., Tikka, T., & Skrifvars, M. B. (2018). Targeting two different levels of both arterial carbon dioxide and arterial oxygen after cardiac arrest and resuscitation: a randomised pilot trial. Intensive Care Medicine, 44(12), 2112–2121. https://doi.org/10.1007/s00134-018-5453-9

Bian, X., Wang, K., Tan, E., Diwu, P., Zhang, F., & Guo, Y. (2020). A selective ensemble preprocessing strategy for near-infrared spectral quantitative analysis of complex samples. Chemometrics and Intelligent Laboratory Systems, 197, 103916. https://doi.org/10.1016/j.chemolab.2019.103916

Kolluru, L. P., Rizvi, S. A. A., D'Souza, M., & D'Souza, M. J. (2012). Formulation development of albumin based theragnostic nanoparticles as a potential delivery system for tumor targeting. Journal of Drug Targeting, 21(1), 77–86. https://doi.org/10.3109/1061186x.2012.729214

Arango, O., Trujillo, A. J., & Castillo, M. (2013). Influence of fat replacement by inulin on rheological properties, kinetics of rennet milk coagulation, and syneresis of milk gels. Journal of Dairy Science, 96(4), 1984–1996. https://doi.org/10.3168/jds.2012-5763

Grassi, S., Amigo, J. M., Lyndgaard, C. B., Foschino, R., & Casiraghi, E. (2014). Assessment of the sugars and ethanol development in beer fermentation with FT-IR and multivariate curve resolution models. Food Research International, 62, 602–608. https://doi.org/10.1016/j.foodres.2014.03.058

"plackett-burman" Design

Ahmed, O. A. A., Kurakula, M., Banjar, Z. M., Afouna, M. I., & Zidan, A. S. (2015). Quality by Design Coupled with Near Infrared in Formulation of Transdermal Glimepiride Liposomal Films. Journal of Pharmaceutical Sciences, 104(6), 2062–2075. https://doi.org/10.1002/jps.24448

Boateng, I. D., Yang, X.-M., & Li, Y.-Y. (2021). Optimization of infrared-drying parameters for Ginkgo biloba L. seed and evaluation of product quality and bioactivity. Industrial Crops and Products, 160, 113108. https://doi.org/10.1016/j.indcrop.2020.113108

Luo, Y., Li, W., Huang, W., Liu, X., Song, Y., & Qu, H. (2017). Rapid quantification of multi-components in alcohol precipitation liquid of Codonopsis Radix using near infrared spectroscopy (NIRS). Journal of Zhejiang University-SCIENCE B, 18(5), 383–392. https://doi.org/10.1631/jzus.b1600141

Box Behnken Design

Wang, J., Wang, K., Wang, Y., Lin, S., Zhao, P., & Jones, G. (2014). A novel application of pulsed electric field (PEF) processing for improving glutathione (GSH) antioxidant activity. Food Chemistry, 161, 361–366. https://doi.org/10.1016/j.foodchem.2014.04.027

Srivastava, P., Ajayakumar, P. V., & Shanker, K. (2014). Box-Behnken Design for Optimum Extraction of Biogenetic Chemicals fromP. lanceolatawith an Energy Audit (Thermal × Microwave × Acoustic): A Case Study of HPTLC Determination with Additional Specificity Using On-line/Off-line Coupling with DAD/NIR/ESI-MS. Phytochemical Analysis, 25(6), 551–560. https://doi.org/10.1002/pca.2529

Luo, Y., Li, W., Huang, W., Liu, X., Song, Y., & Qu, H. (2017). Rapid quantification of multi-components in alcohol precipitation liquid of Codonopsis Radix using near infrared spectroscopy (NIRS). Journal of Zhejiang University-SCIENCE B, 18(5), 383–392. https://doi.org/10.1631/jzus.b1600141

Composite Design
Lima, A. B. S. de, Batista, A. S., Jesus, J. C. de, Silva, J. de J., Araújo, A. C. M. de, & Santos, L. S. (2020). Fast quantitative detection of black pepper and cumin adulterations by near-infrared spectroscopy and multivariate modeling. Food Control, 107, 106802. https://doi.org/10.1016/j.foodcont.2019.106802
Guzmán-Ortiz, F. A., Hernández-Sánchez, H., Yee-Madeira, H., San Martín-Martínez, E., Robles-Ramírez, M. del C., Rojas-López, M., Berríos, J. D. J., & Mora-Escobedo, R. (2014). Physico-chemical, nutritional and infrared spectroscopy evaluation of an optimized soybean/corn flour extrudate. Journal of Food Science and Technology, 52(7), 4066–4077. https://doi.org/10.1007/s13197-014-1485-5
Boateng, I. D., Yang, X.-M., & Li, Y.-Y. (2021). Optimization of infrared-drying parameters for Ginkgo biloba L. seed and evaluation of product quality and bioactivity. Industrial Crops and Products, 160, 113108. https://doi.org/10.1016/j.indcrop.2020.113108

D-Optimal Design
El-Hagrasy, A. S., D'Amico, F., & Drennen, J. K. (2006). A Process Analytical Technology approach to near-infrared process control of pharmaceutical powder blending. Part I: D-optimal design for characterization of powder mixing and preliminary spectral data evaluation. Journal of Pharmaceutical Sciences, 95(2), 392–406. https://doi.org/10.1002/jps.20467
Ebrahimi-Najafabadi, H., Leardi, R., Oliveri, P., Chiara Casolino, M., Jalali-Heravi, M., & Lanteri, S. (2012). Detection of addition of barley to coffee using near infrared spectroscopy and chemometric techniques. Talanta, 99, 175–179. https://doi.org/10.1016/j.talanta.2012.05.036
Sulub, Y., & DeRudder, J. (2013). Determination of polymer blends composed of polycarbonate and rubber entities using near-infrared (NIR) spectroscopy and multivariate calibration. Polymer Testing, 32(4), 802–809. https://doi.org/10.1016/j.polymertesting.2013.03.008

Mixture Design
Brereton, R. G. (2007). Applied Chemometrics for Scientists. https://doi.org/10.1002/9780470057780
Wu, H., Tawakkul, M., White, M., & Khan, M. A. (2009). Quality-by-Design (QbD): An integrated multivariate approach for the component quantification in powder blends☆. International Journal of Pharmaceutics, 372(1–2), 39–48. https://doi.org/10.1016/j.ijpharm.2009.01.002
Burger, J., & Geladi, P. (2006). Hyperspectral NIR image regression part II: dataset preprocessing diagnostics. Journal of Chemometrics, 20(3–4), 106–119. https://doi.org/10.1002/cem.986

Space filling design
Puchert, T., Holzhauer, C.-V., Menezes, J. C., Lochmann, D., & Reich, G. (2011). A new PAT/QbD approach for the determination of blend homogeneity: Combination of on-line NIRS analysis with PC Scores Distance Analysis (PC-SDA). European Journal of Pharmaceutics and Biopharmaceutics, 78(1), 173–182. https://doi.org/10.1016/j.ejpb.2010.12.015

Khorasani, M., Amigo, J. M., Sun, C. C., Bertelsen, P., & Rantanen, J. (2015). Near-infrared chemical imaging (NIR-CI) as a process monitoring solution for a production line of roll compaction and tableting. European Journal of Pharmaceutics and Biopharmaceutics, 93, 293–302. https://doi.org/10.1016/j.ejpb.2015.04.008
Ma, W., Xu, Y., Xiong, B., Deng, L., Peng, R., Wang, M., & Liu, Y. (2022). Pushing the Limits of Functionality-Multiplexing Capability in Metasurface Design Based on Statistical Machine Learning. Advanced Materials, 34(16), 2110022. Portico. https://doi.org/10.1002/adma.202110022

Doelhert Design
Foca, G., Ferrari, C., Sinelli, N., Mariotti, M., Lucisano, M., Caramanico, R., & Ulrici, A. (2010). Minimisation of instrumental noise in the acquisition of FT-NIR spectra of bread wheat using experimental design and signal processing techniques. Analytical and Bioanalytical Chemistry, 399(6), 1965–1973. https://doi.org/10.1007/s00216-010-4431-z

Approche multiblocs pour l'analyse de données de spectroscopie vibrationnelle

Benoît JAILLAIS, Mohamed HANAFI*

ONIRIS, INRAE, StatSC, 44300 Nantes, France.

Email : Benoit.jaillais@inrae.fr; Mohamed.hanafi@oniris-nantes.fr

***Mots-clefs** : Données multiblocs, Spectroscopie Vibrationnelle, Analyse en Composantes Principales.*

Dans le domaine des sciences biologiques, comme dans bien d'autres domaines scientifiques, l'intégration de données multi-sources est plus que jamais d'actualité. Parallèlement, les données collectées sont de plus en plus complexes et leur volume ne cesse de croître du fait du développement des plateformes analytiques, des techniques d'imagerie, de l'essor des données omiques, etc. Naturellement, ce contexte a stimulé la recherche autour des méthodes pour l'analyse conjointe de plusieurs tableaux de données (données structurées, multi-blocs, multi-voies). Aujourd'hui, les mesures multimodales par spectroscopie vibrationnelle des propriétés des processus et des produits sont devenues très populaires. Les méthodes chimiométriques classiques telles que l'analyse en composantes principales (ACP) et la régression des moindres carrés partiels (PLS) sont étendues alors pour être plus efficace pour analyser des données multiblocs. Différentes approches avec une visée exploratoire sont proposées : généralisations de l'analyse canonique, algorithme NIPALS, et décompositions tensorielles. La prise en compte de structures connues sur les observations (groupes, multi-niveaux, hiérarchiques) par ces approches est également considérée dans de nombreux travaux. Par ailleurs, des stratégies avec une visée prédictive sont également très populaires et permettent une grande flexibilité dans la prise en compte du statut (expliqué / explicatif) des blocs par le biais d'un graphe orienté entre ces blocs.

Devant la multitude des approches proposées et des méthodes qui en résultent pour répondre à tel ou tel objectif, l'utilisateur non spécialiste pourrait rester perplexe. La présente communication présente une introduction didactique à l'analyse de données multi-blocs lorsque ces techniques sont appliquées sur données issues de la spectroscopie vibrationnelle. Le processus d'analyse sera présenté sur la base de plusieurs études de cas. Les avantages et inconvénients des différentes méthodes seront également discutées. De plus, les tâches de base allant de la visualisation de données multi-blocs aux applications innovantes seront brièvement mises en évidence. Enfin, un résumé des ressources logicielles disponibles pour l'analyse de données multi-blocs est fourni.

Transformers : une alternative au CNN pour le traitement de données spectrales

[1,4] Maxime Metz, [2,4]Matthieu lesnoff, [3,4]Jean-Michel Roger, [1] Nicolas Grotus

[1] *Pellenc Selective Technologies, Pertuis, Provence-Alpes-Côte d'Azur, France*
[2] *UMR SELMET, Univ Montpellier, CIRAD, INRA, Institut Agro, Montpellier, France*
[3] *ITAP, Univ. Montpellier, INRAE, Institut Agro, Montpellier, France*
[4] *ChemHouse Research Group, Montpellier, France*

Email : m.metz@pellencst.com

***Mots-clefs** : Apprentissage profond, transformers, attention, spectroscopie*

Le deep learning a révolutionné de nombreux domaines scientifiques grâce à sa capacité à traiter des données complexes. Ses avancées constantes ont permis des percées significatives dans de nombreux domaines tels que la reconnaissance d'images, la traduction automatique et l'analyse de texte. La communauté du deep learning est vaste et dynamique, et elle propose désormais de nouveaux outils et frameworks (Keras, TensorFlow, PyTorch) pour faciliter la mise en œuvre et l'expérimentation des modèles de deep learning. Ces développements ont permis d'identifier le deep learning comme une famille de méthodes potentiellement pertinente pour le traitement des données spectrales [1-2]. L'une des méthodes de deep learning les plus utilisées dans le traitement des données spectrales est le réseau de neurones convolutifs (Convolutional Neural Network ou CNN) [2].

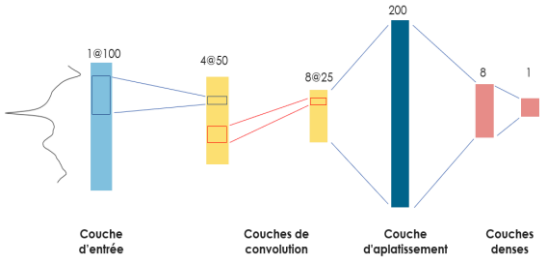


Figure 2 : Exemple d'un CNN avec deux couches de convolution, une couche d'aplatissement et deux couches denses

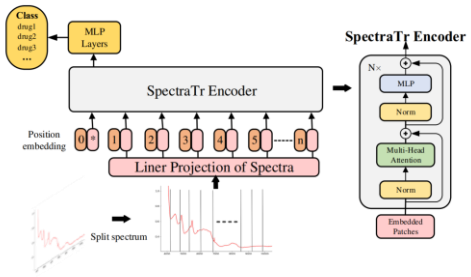


Figure 3 : SpectraTr

Bien que le CNN ait démontré de bonnes performances pour le traitement de données spectrales, il souffre néanmoins de plusieurs limitations, dont une majeure : le manque de contextualisation (spectrale, spatial, temporel, etc…). En effet, appliqué à la spectroscopie, le CNN a pour objectif de définir des descripteurs pertinents à partir de spectres bruts ou prétraités. Ces descripteurs sont cependant concaténés de manière naïve. Cela signifie qu'après la couche d'aplatissement dans le réseau (voir figure 1), la structure spectrale n'est plus mise en évidence. Les transformations du spectre par chacun des filtres sont mis bout à bout sans considérer les corrélations lointaines entre certaines zones spectrales.

Récemment, les « mécanismes d'attention » ont été mis en œuvre pour lever cette limitation méthodologique. Ces mécanismes d'attention ont été intégrés dans une méthode appelée « Transformer » [3]. Cette dernière a tout d'abord été utilisée pour le traitement naturel du langage, puis dans la vision par ordinateur. Aujourd'hui, ces modèles font partie des outils récemment développés tels que ChatGPT, DALL-E, LLaMA, etc. Récemment, une architecture de type Transformer a été développée pour le traitement des données spectrales. Cette architecture s'inspire fortement d'un modèle issu de la vision par ordinateur appelé ViT (Vision Transformers) [4]. SpectraTr [5] remplace tout simplement la stratégie d'encodage de l'image par une stratégie d'encodage du spectre (voir figure 2). Dans cette présentation, l'objectif est de comparer le modèle SpectraTr avec un modèle moins complexe de type CNN et un modèle couramment utilisé en chimiométrie : la PLS locale [6]. Les modèles ont été testés avec une base de données prétraitée et anonymisée.

Les premiers résultats mettent en évidence la polyvalence de SpectraTr. En l'état, pour ce jeu de données, cette approche ne permet pas d'obtenir des performances supérieures aux approches traditionnelles de chimiométrie comme la PLS locale. Pour améliorer les performances de l'approche SpectraTr, plusieurs perspectives sont envisageables. La première consiste à augmenter la taille de la base de données afin de permettre au modèle de capturer plus d'informations. La deuxième consiste à introduire des connaissances et des méthodes issues de la chimiométrie dans les stratégies d'entrainement du modèle (augmentations de données par exemple) ou dans l'architecture du réseau de neurones entre autres.

[1] Bjerrum, E.J., Glahder, M., Skov, T., 2017. Data Augmentation of Spectral Data for Convolutional Neural Network (CNN) Based Deep Chemometrics.
[2] Cui, C., Fearn, T., 2018. Modern practical convolutional neural networks for multivariate regression: Applications to NIR calibration. Chemometrics and Intelligent Laboratory Systems 182, 9–20.
[3] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I., 2017. Attention Is All You Need.
[4] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale.
[5] Fu, P., Wen, Y., Zhang, Y., Li, L., Feng, Y., Yin, L., Yang, H., 2022. SpectraTr: A novel deep learning model for qualitative analysis of drug spectroscopy based on transformer structure. J. Innov. Opt. Health Sci. 15, 2250021.
[6] Lesnoff, M., Metz, M., Roger, J., 2020. Comparison of locally weighted PLS strategies for regression and discrimination on agronomic NIR data. Journal of Chemometrics 34.

A little journey through Causality

Philippe Bastien

L'Oréal

Email : Philippe.bastien@loreal.com

Surprisingly today causality can be seen as an emerging field, particularly due to the work of Judea Pearl at the turn of the last century. Even if the notion of causality was preexisting to the development of probability theory (De Moivre, The Doctrine of Chances, 1718), asking the questions in terms of causality until recently could even be considered unscientific. Causality almost disappeared as a specific concept at the end of the 19ème century with the appearance of the notion of correlation (Galton, 1888).

Correlation is not causation, but we will see that some correlation implied causation.

Causality is not a statistical notion, but an enrichment of the Statistic to uncover part of the world that traditional methods cannot approach. Traditional statistical methods are oriented towards inference (Fisher, 1922), associated more with a parsimonious description of the data than a description of the process responsible for the data. As a result, when using conventional statistical methodology without causal lenses some paradox appears like the most known Simpson's paradox. We will explain this paradox and show that it is no more a paradox when using a causal approach.

We will show, by taking up the work of Judea Pearl, how one can access to causality from simple observed data, without necessarily resorting to a randomized trial, by using simple graphic rules associated with a representation of the world.

Let's note that the notion of causality extends beyond the observable domain, mixing observed and unobserved worlds, through the notion of counterfactual to answer questions that statistics cannot answer, as in the case of mediation.

The purpose of this presentation will be to introduce in a succinct way some essential notions around causality.

[1] Dana Mackenzie, Judea Pearl, The book of Why, Penguin publisher, 2018.
[2] Judy Pearl, Causality, Cambridge university Press, 2000

Posters

Suivi non destructif de l'accumulation des sucres et de l'acide malique dans le raisin par spectroscopie proche infra-rouge

[1*]Flora Tavernier, [1,2]Charles Romieu, [1,2]Elias Motelica-Heino, [1,2]Miguel Thomas, [1,2]Theresa Herbold, [1,2]Loic Le Cunff, [1,2]Patrice This, [1,2]Vincent Segura

[1] *INRAE-Montpellier CIRAD, UMR AGAP DAAV, 34398 Montpellier - France*
[2] *IFV-INRAE, UMT Geno-Vigne, 34398 Montpellier – France*

Email : flora.tavenier@inrae.fr

***Mots-clefs** : Vigne, Spectroscopie proche infra-rouge, Baie unique, Sucre, Acides organiques*

Il devient particulièrement urgent de décrypter les mécanismes physiologiques sous-jacents de l'impact du changement climatique sur la maturation des baies et de sélectionner de nouveaux génotypes conservant un équilibre adéquat entre les sucres et les acides malgré l'augmentation de la température estivale. Les études de génétique classique consistent à étudier la diversité des traits d'intérêt, éventuellement en réponse à des contraintes, afin d'identifier des allèles d'intérêt. Ce type d'étude est particulièrement complexe à mettre en œuvre en ce qui concerne la composition du raisin qui évolue fortement au cours de la maturation des baies, cette variation phénologique étant elle-même contrôlée génétiquement. Ceci est encore plus compliqué par le développement asynchrone des baies au sein des grappes, il est donc essentiel de phénotyper les baies de différentes variétés au même stade de développement, afin que les différences observées ne reflètent pas uniquement des écarts incontrôlés de niveau de maturité. Dans ce contexte, il est crucial de développer des outils non destructifs de suivi de la maturation et du développement des baies à haut débit. Nous rapportons ici l'utilisation de la spectroscopie proche infrarouge (NIRS), à l'aide d'un appareil portable (MicroNIR OnSite-W, VIAVI), pour étudier l'accumulation de sucres et d'acides dans les baies de dix variétés de vigne pendant deux ans. Nous avons acquis séquentiellement des spectres sur des baies individuelles de 50 grappes tout au long de leur développement, du stade vert à la surmaturité, en collectant un sous-ensemble de ces baies chaque semaine, pour quantifier les concentrations de sucres et d'acides par HPLC. Ces données ont été utilisées pour construire des modèles d'étalonnage entre les spectres et les concentrations de sucres ou d'acides par régression PLS. Nous avons divisé les données en un sous-groupe d'apprentissage (⅘ des données) et un sous-groupe de validation (⅕ restant), en utilisant l'algorithme de Kennard-Stone. Ces modèles se sont avérés assez précis pour les teneurs en sucres et en acide malique, et ont donc ensuite été appliqués pour prédire l'évolution de ces composés sur les baies qui ont été suivies par NIRS mais non collectées pour les mesures HPLC. Cela a permis de reconstruire les trajectoires de développement des baies individuelles pendant toute la période de maturation. Ces résultats ouvrent des pistes pour des études génétiques et physiologiques de la maturation des baies qui sont essentielles pour la sélection et le développement de nouvelles variétés dans le contexte du changement climatique.

Utilisation de trois spectromètres pour prédire le fonctionnement foliaire des vignes en réponse à un déficit hydrique

[1,2*]Eva Coindre, [2]Laurine Chir, [2]Romain Boulord, [2]Thomas Laisné, [2]Gaëlle Rolland, [2]Maëlle Lis, [2]Mélyne Falcon, [2]Llorenç Cabrera-Bosquet, [2]Thierry Simonneau, [1]Agnès Doligez, [2]Benoît Pallas, [2]Aude Coupel-Ledru, [1]Vincent Segura

[1] *AGAP Institut, Univ Montpellier, CIRAD, INRAE, Institut Agro, 34398 Montpellier, France*
[2] *LEPSE, Univ Montpellier, INRAE, Institut Agro, 34060 Montpellier, France*

Email : eva.coindre@umontpellier.fr

***Mots-clefs** : Spectromètres micro-portables, Prédiction, Traits écophysiologiques, Déficit hydrique, Vigne*

Le changement climatique se matérialise par l'augmentation de la température et la perturbation des précipitations, favorisant les risques et la durée des sécheresses. La demande évaporative atmosphérique et le déficit hydrique du sol vont s'accentuer dans les années à venir devenant un facteur limitant majeur à la production et la qualité des raisins dans le Sud de la France. Le compromis entre les pertes en eau et l'assimilation de carbone est un enjeu primordial pour la plante afin de limiter sa déshydratation tout en maintenant son rendement. Dans ce contexte, étudier la diversité génétique de la vigne en réponse à différents scénarios hydriques à l'échelle du fonctionnement foliaire devrait permettre de mettre en évidence des régions génétiques intéressantes pour la sélection de nouveaux cultivars mieux adaptés à la sécheresse.

Cependant, les études de diversité génétique nécessitent l'évaluation de caractères d'intérêt chez un grand nombre d'individus. Or les méthodes habituellement utilisées pour caractériser le fonctionnement écophysiologique des plantes sont typiquement lourdes à mettre à œuvre et ne sont donc pas vraiment compatibles avec des études génétiques. Des études récentes soulignent l'intérêt de la NIRS pour l'évaluation à haut-débit de caractères de fonctionnement foliaire. Nos travaux visent à mettre au point et déployer de telles méthodes pour le phénotypage de la diversité génétique de la vigne.

Lors d'une première expérience, des spectres ont été collectés sur des feuilles de vignes n'ayant pas subi de déficit hydrique. Ils ont ensuite été utilisés pour construire des modèles de prédiction de traits de fonctionnement foliaire liés au fonctionnement hydrique et carbonée, tels que le LMA (Leaf Mass Area), le WC (Water Content) et le SPAD (un proxy de la quantité de chlorophylle).

Une seconde expérience a ensuite été menée dans le cadre du projet ANR G2WAS dans la plateforme de phénotypage PhénoArch avec un panel de 250 génotypes de vignes, représentatif de la diversité de l'espèce. Chaque génotype a été mesuré dans trois scénarios hydriques différents : i/ bien irrigué, ii/ déficit hydrique modéré et iii/ déficit hydrique sévère. Des spectres NIRS ont été collectés sur des feuilles intactes à l'aide de deux spectromètres micro-portables (MicroNIR™ Onsite-W, VIAVI, et NeoSpectra™ Scanner, Si-Ware) présentant des gammes de longueur d'ondes restreintes mais complémentaires (950-1650nm et 1350-2500nm respectivement). Des spectres ont également été pris sur des disques de feuilles séchés avec un spectromètre de laboratoire (ASD LabSpec® 4) présentant une gamme de longueur d'ondes plus large (350-2500nm) et une meilleure résolution que les appareils micro-portables.

Ces spectres ont d'abord été analysés pour évaluer les effets associés notamment au génotype, au scénario hydrique et à l'interaction Génotype x Scénario. Des analyses en composantes principales (ACP) sur les spectres permettent déjà de mettre en évidence un effet du scénario hydrique. Ces mêmes spectres ont ensuite été utilisés avec les modèles de prédiction précédemment établis, afin d'évaluer la robustesse de ces modèles vis à vis de la contrainte hydrique notamment. Les valeurs prédites ont alors été comparées aux valeurs mesurées de LMA, WC et SPAD sur les mêmes individus afin d'étudier la précision de prédiction

des modèles en fonction des trois scénarios hydriques. Les résultats obtenus sont prometteurs pour l'utilisation de spectromètres au vignoble comme proxy des caractères d'intérêt. Les analyses en cours permettront d'affiner les modèles de prédiction des caractères considérés ainsi que d'en étudier la variabilité et le contrôle génétique, dans l'optique d'identifier des cultivars d'intérêt.

Développement d'un modèle SPIR multi-données et multi-espèces appliqué aux arbres forestiers

[1,2]Mélissa SANCHARME, [1*]Rémy GOBIN, [1]Nassim BELMOKHTAR, [2]Cécile VINCENT-BARBAROUX, [2]Régis FICHOT

[1]INRAE, UMR BIOFORA - PHENOBOIS, 45075 Orléans – France
[2]UNIVERSITE D'ORLEANS, LBLGC, 45067 Orléans – France

Email : remy.gobin@inrae.fr

***Mots-clefs** : modèle global, teneur en amidon, chêne, peuplier*

La spectrométrie proche infra-rouge associée à des méthodes de chimiométrie est un outil de caractérisation de la composition chimique des échantillons de bois. Jusqu'à présent, une nouvelle calibration était établie à chaque campagne de mesures. Ainsi, au fil des projets de recherche sur les arbres, nous avons acquis des jeux de données et établi des courbes de calibration pour des caractères similaires mais de manière indépendante. Ces jeux de données sont constitués d'échantillons prélevés dans différents types d'environnement, de tissus (tronc, branche, racine) et d'espèces. A travers cette étude, où nous avons regroupé ces différents jeux de données (plus de 1 500 échantillons), nous essayons de tendre vers un modèle de calibration global et potentiellement utilisable à chaque nouvelle campagne de mesures. Pour cela, nous avons testé la capacité prédictive des différents modèles obtenus sur chaque jeu de données et leurs combinaisons avec des analyses simples de chimiométrie (prétraitement, détection des outliers, PLS). Le caractère étudié correspond à la concentration en amidon dans les tissus ligneux, qui est potentiellement un indicateur de la sensibilité à la sécheresse des arbres. Pris séparément, nous observons des capacités prédictives variables selon les jeux de données. Lorsqu'ils sont combinés, la capacité prédictive n'est pas supérieure au meilleur résultat obtenu par un jeu de données analysé séparément. De plus, l'application du modèle globale sur chaque jeu de données se traduit par une erreur de prédiction variable selon les jeux de données. Ainsi, le développement d'une calibration plus globale semble possible pour ce caractère intérêt. Toutefois, des approches de chimiométrie plus complexes seraient nécessaire pour contribuer à l'amélioration de la capacité prédictive d'un modèle global.

Influence de l'étendue de la gamme spectrale sur la performance des modèles de discrimination SPIR : cas de six espèces de *Diospyros* de Madagascar

[1,4]Andry Clarel RAOBELINA, [2,3,4]Gilles CHAIX, [1] Tahiana RAMANANANTOANDRO

[1] *Mention Foresterie et Environnement – Université d'Antananarivo, BP 175 Antananarivo 101, Madagascar*
[2] *CIRAD, UMR AGAP Institut, F-34398 Montpellier, France*
[3] *UMR AGAP Institut, Univ Montpellier, CIRAD, Institut Agro, Montpellier, France*
[4] *ChemHouse Research Group, Montpellier, France*

Email : andryclarel@gmail.com

***Mots clefs** : Spectrométrie Proche Infrarouge, Diospyros, Madagascar, gamme spectrale*

Le genre *Diospyros* de Madagascar regroupe actuellement 255 espèces dont 82 sont considérées comme des grands arbres. Plusieurs de ces espèces sont exploitées illégalement, ce qui a entrainé l'insertion de toutes les espèces de *Diospyros* malgaches dans l'annexe II de la Convention sur le Commerce international des espèces de faune et de flore sauvage menacées d'extinction (CITES) en 2013. Le développement d'outil d'identification comme la Spectrométrie Proche InfraRouge (SPIR) qui permet d'identifier les ébènes malgaches à partir de leurs bois est en cours à Madagascar pour mettre en œuvre lu plan d'action de la CITES dans l'objectif d'une exploitation et d'un commerce durable de ces espèces.

La SPIR est parmi les techniques utilisés dans l'identification des espèces anatomiquement proches et dont la commerce est réglementée par la CITES. L'utilisation des spectromètres portatifs en particulier s'est beaucoup développée ces dernières années. Pour les spectromètres portables, l'étendue de la gamme spectrale peut être différente d'un modèle à un autre. La question principale adressée dans le cadre de cette étude est alors la suivante : quelle plage de longueurs d'onde dans le proche infrarouge est idéale pour discriminer les essences d'ébènes malagasy sur la base des spectres PIR de leurs bois ?

Les spectres des 78 carottes ont été mesurés à l'aide d'un spectromètre Bruker MPA II entre 12500 cm-1 à 4000 cm-1 de nombre d'ondes, correspondant à 800 – 2500 nm de longueurs d'ondes, et à une résolution de 8 cm-1. Six spectres ont été acquis sur la partie duramen du bois de chaque carotte depuis la moelle vers l'écorce afin de considérer la variabilité radiale du bois, ce qui fait un total de 468 spectres pour l'ensemble des échantillons.

Six bases de données ont été extraites à partir du jeu de données initial (468 spectres et identification botanique). La première est constituée par les données d'origine (spectres et données de références Y_((468 ×6))), c'est-à-dire des spectres comprenant les absorbances dans toute la gamme spectrale (800-2500 nm). Les cinq autres bases de données sont extraites des données originales, sur des gammes de longueur d'onde différente selon le type de spectromètres portables existants et courants : 950-1650 correspondants aux spectromètres avec détecteur InGaAs, et les autres gammes correspondantes aux spectromètres de la marque NIROne.

Une validation croisée à 4 blocs répétée 20 fois avec 30 Variables Discriminates (VDs) au maximum a été effectuée afin d'identifier les meilleurs prétraitements et le nombre optimal de VDs. Les meilleurs modèles ont ensuite été testés pour classifier les spectres du jeu de validation. La performance des modèles a été évaluée à partir de trois métriques de classification dont « Accuracy » qui est le pourcentage global de spectres bien classés par le modèle pour les six espèces, « Recall » qui exprime le pourcentage de spectres bien classés pour une classe déterminée et « Precision » qui exprime la probabilité qu'un spectre inconnu appartient à la classe prédite par le modèle.

Le choix de l'étendue de la gamme spectrale sur laquelle les modèles de discrimination sont étalonnés influence considérable la performance de ces derniers. La région au-delà de 2000 nm est la plus importante

pour étalonner des modèles pour le cas des six espèces de *Diospyros*. Le meilleur modèle dans la plage de longueurs d'onde 2000-2450 nm a une performance globale de 99,9%, alors qu'elle diminue jusqu'à 40-45% pour les régions avant 2000 nm. Tester un spectromètre PIR portable qui couvre la région 2000 – 2450 nm tel le NIRone 2.5 pour discriminer les six espèces de *Diospyros*, puis comparer les résultats avec ceux de la présente étude serait intéressant dans le futur. Cela permettrait de renforcer le choix d'un spectromètre couvrant cette gamme pour avoir des modèles performants, et prendre en compte les caractéristiques techniques du spectromètre comme la résolution spectrale, l'optique, l'éclairage.