UNIVERSITÉ SAAD DAHLAB BLIDA 1

Faculté des Sciences

Département de Chimie

THÈSE DE DOCTORAT - LMD

En Chimie

Spécialité : Chimie Analytique

NOUVELLES MÉTHODES DE CONTRÔLE QUALITÉ DES PRODUITS DE TABAC À CHIQUER SANS FUMÉE « CHEMMA » PAR SPECTROSCOPIE ATR-IRTF COMBINÉE À LA CHIMIOMÉTRIE

Par

Mohamed FEKHAR

Devant le jury composé de :

N. Bouchenafa-Saib	Professeur, USD-Blida 1	Présidente
A. Hadj Sadok	Professeur, USD-Blida 1	Examinateur
M. O. Boulakradeche	Maître des conférences A, USTHB	Examinateur
Y. Daghbouche	Professeur, USD-Blida 1	Directrice de thèse
N. Bouzidi	Professeur, USD-Blida 1	Co-Directrice de thèse

Blida 2024

ملخص

تتميز سوق التبغ غير المدخن (الشمة) في الجزائر بوجود عدد كبير من المنتجات المقلدة والرديئة. الطرق المعمول بها لمراقبة النيكوتين ومعلمات الجودة الأخرى بشكل موثوق في هذه المنتجات متطلبة وغير عملية للاستخدام الروتيني.

للتعامل مع هذه المسائل، قمنا بإجراء دراسة شاملة مكونة من جزئين.

ركز الجزء الأول من الدراسة على تطوير وإثبات صحة طريقة جديدة لقياس محتوى النيكوتين الإجمالي في 27 عينة من الشمة المسوقة بالإضافة إلى أنواع مختلفة من أوراق التبغ، باستخدام مطيافية الأشعة تحت الحمراء في النطاق الأوسط بتحويل فورييه مع ملحق الانعكاس الكلي المخفف (ATR-FT-MIR) وبالاقتران مع اثنتين من طرق الانحدار: أحادية المتغير ومتعددة المتغيرات. سمح هذا النهج بتحديد محتوى النيكوتين الإجمالي في المنتجات المصنعة والذي تراوح من 3.9 إلى 11.7 ملغ/غ، على أساس الوزن الجاف للمنتج.

يستكشف الجزء الثاني تطبيق التحليل الطيفي ATR-FT-MIR لدراسة مجموعتين من 105 عينات من التبغ غير المدخن والتي تم جمعها على مدار عامين متتاليين (2021 و2022م).

في التقييم النوعي، تم اعتماد نهج من خطوتين. في البداية، تم تنفيذ تحليل المكونات الرئيسية (PCA)، والتجميع الهرمي التكتلي (AHC)، والتجميع بالمتوسط (*k*-means)، كطرق غير خاضعة للإشراف متكاملة مع بعضها، لتجميع العينات التجارية في مجموعات متمايزة على أساس قياساتها الفيزيائية - الكيميائية المرجعية. ثم استخدمت هذه المجموعات بعد ذلك كفئات مستهدفة لتدريب النماذج الخاضعة للإشراف، وتحديدًا التحليل التمييزي بالمربعات الصغرى الجزئية (PLS-DA) والتصنيف بآلة المتجهات الداعمة (SVM-C)، مما يسمح بتصنيف العينات المستقبلية فقط بالاستناد إلى خصائصها الطيفية.

وفي التقييم الكمي، استُخدم تحليل الانحدار بالمربعات الصغرى الجزئية (PLSR) والانحدار بآلة المتجهات الداعمة (SVMR) لتدريب النماذج والتحقق من صحتها من أجل ا لتنبؤ المتزامن بمستوى الرطوبة والأس الهيدروجيني ونسبة الرماد ومحتويات النيكوتين الإجمالي والنيكوتين غير المتأين. باستثناء الرطوبة، كان أداء نماذج الـ SVMR مُرضيًا للغاية، حيث تفوقت على نماذج

الـ PLSR، مع معاملات تحديد تراوحت بين 0.82 و0.93، ومعاملات تحديد نسبة التنبؤ إلى الانحراف (RPD) تراوحت بين 2.4 و3.7.

الكلمات المفتاحية: مطيافية الأشعة تحت الحمراء بتحويل فوريه بانعكاس كلي مخفف، تبغ غير مدخن، معلمات الجودة، نيكوتين، تجميع وتصنيف ، المربعات الصغرى الجزئية، آلة المتجهات الداعمة.

ABSTRACT

The Algerian smokeless tobacco (ST) market is characterized by a significant presence of counterfeit and substandard products. The official methods used for reliable monitoring of nicotine and other quality parameters in STs are laborious and impractical for routine use.

To address these issues, we conducted a comprehensive study consisting of two parts.

The first part of the study focused on the development and validation of a new method for quantifying total nicotine content in 27 samples of commercial Chemma and different tobacco leaf varieties, using attenuated total reflectance Fourier transform mid-infrared (ATR-FT-MIR) spectroscopy, coupled with two regression methods: univariate and multivariate. This approach enabled us to determine the total nicotine content of the manufactured STs, which ranged from 3.9 to 11.7 mg/g, on dry weight basis of product.

The second part explores the application of ATR-FT-MIR spectroscopy for the analysis of two sets of 105 ST samples collected in consecutive years (2021 and 2022).

In the qualitative assessment, a two-step approach was adopted. Initially, Principal Component Analysis, Agglomerative Hierarchical Clustering, and *k*-means clustering were implemented as complementary unsupervised methods to group commercial samples into distinct clusters on the basis of their reference physicochemical measurements. These clusters then served as target categories for training supervised models, specifically partial least-squares discriminant analysis (PLS-DA) and support vector machine classification (SVM-C), enabling the classification of new samples solely based on their spectral features.

In the quantitative assessment, partial least-squares regression (PLSR) and support vector machine regression (SVMR) were used to train and validate models for simultaneously predicting moisture, pH, ash, total nicotine, and un-ionized nicotine contents. With the exception of moisture, SVMR performed very

satisfactorily, surpassing PLSR, with coefficients of determination ranging from 0.82 to 0.93 and prediction-to-deviation ratios (RPD) from 2.4 to 3.7.

Keywords: ATR-FT-MIR spectroscopy, smokeless tobacco, quality parameters, nicotine, clustering and classification, partial least-squares, support vector machine.

RÉSUMÉ

Le marché algérien du tabac sans fumée (ST) est caractérisé par une présence significative de produits contrefaits et de qualité inférieure. Les méthodes officielles utilisées pour un contrôle fiable de la nicotine et d'autres paramètres de qualité dans les ST sont laborieuses et peu pratiques pour une utilisation routinière.

Pour faire face à ce problème, nous avons mené une étude exhaustive comprenant deux parties.

La première partie de l'étude s'est concentrée sur le développement et la validation d'une nouvelle méthode de quantification de la teneur totale en nicotine dans 27 échantillons de Chemma commerciale et de différentes variétés de feuilles de tabac, en utilisant la spectroscopie moyen-infrarouge à transformée de Fourier à réflectance totale atténuée (ATR-FT-MIR), couplée à deux méthodes de régression: univariée et multivariée. Cette approche nous a permis de déterminer la teneur totale en nicotine des ST fabriquées qui variait de 3,9 à 11,7 mg/g, sur la base du poids sec de produit.

La deuxième partie explore l'application de la spectroscopie ATR-FT-MIR pour l'analyse de deux séries de 105 échantillons de ST collectés au cours des années consécutives (2021 et 2022).

Dans l'évaluation qualitative, une approche en deux étapes a été adoptée. Initialement, l'Analyse en Composantes Principales, la Classification Hiérarchique Ascendante et le Partitionnement en *k*-moyennes ont été mis en œuvre comme méthodes non supervisées complémentaires pour regrouper les échantillons commerciaux en clusters distincts sur la base de leurs mesures physicochimiques de référence. Ces clusters ont ensuite servi de catégories cibles pour l'entraînement de modèles supervisés, spécifiquement l'analyse discriminante par moindres carrés partiels (PLS-DA) et la classification par machine à vecteurs de support (SVM-C), permettant la classification de nouveaux échantillons uniquement en se basant sur leurs caractéristiques spectrales.

Dans l'évaluation quantitative, la régression par moindres carrés partiels (PLSR) et la régression par machine à vecteurs de support (SVMR) ont été utilisées pour entraîner et valider des modèles permettant de prédire simultanément les teneurs en humidité, pH, cendres, nicotine totale et nicotine non ionisée. À l'exception de l'humidité, la SVMR a montrée des performances très satisfaisantes, surpassant la PLSR, avec des coefficients de détermination allant de 0,82 à 0,93 et des RPD (Ratio of Prediction-to-Deviation) allant de 2,4 à 3,7.

Mots-clés : Spectroscopie ATR-FT-MIR, tabac sans fumée, paramètres de qualité, nicotine, regroupement et classification, moindres carrés partiels, machine à vecteurs de support.

REMERCIEMENTS

Je commence par exprimer ma profonde reconnaissance envers Allah le tout-puissant, qui m'a accordé la santé, la patience et la volonté nécessaires pour mener à bien ce modeste travail. Il m'a appris ce que je ne savais pas et sa grâce sur moi a été immense.

Mes remerciements les plus sincères vont à Madame ***Y. Daghbouche*** *et Madame* ***N. Bouzidi*** *pour leur direction attentive, leurs conseils, et leur intérêt marqué pour ce travail. Travailler sous leur supervision a été une opportunité exceptionnelle et enrichissante.*

Je souhaite exprimer ma haute gratitude et mes respects envers les membres du jury, Madame ***N. Bouchenafa-Saib****, Monsieur* ***A. Hadj Sadok*** *et Monsieur* ***M. O. Boulakradeche****, pour l'honneur qu'ils m'ont fait en évaluant mon travail.*

Je tiens également à remercier Monsieur ***M. El Hattab****, Directeur du laboratoire LCSN-BioM, pour m'avoir accueilli pendant la réalisation de cette recherche.*

Un merci particulier à toute ma famille pour leur soutien indéfectible, et surtout à ***A. Fekhar*** *pour son soutien et sa contribution inestimable à la conception graphique de nombreuses parties de cette étude.*

Mes remerciements vont également aux collègues ***C. Ait si said****,* ***F. Messaoudi****,* ***D. Medjdoub*** *et* ***R. Larouci*** *pour leur aide précieuse et leur disponibilité tout au long du processus de recherche.*

Je suis reconnaissant envers tous les enseignants du département de chimie et de la faculté des sciences de l'université Blida 1 qui ont contribué à ma formation.

Enfin, je souhaite remercier toutes les personnes qui ont participé de près ou de loin à la réalisation de ce travail.

TABLE DES MATIÈRE

RÉSUMÉ (arabe, anglais et français) i
REMERCIEMENTS vii
TABLE DES MATIERES viii
LISTE DES ILLUSTRATIONS GRAPHIQUES ET TABLEAUX xii

INTRODUCTION 19

CHAPITRE 1 : FONDEMENTS THÉORIQUES ET BIBLIOGRAPHIQUES 22

1.1. Tabac et produits à base de tabac 22
1.1.1. Histoire du tabac 22
1.1.2. Tabac : Taxonomie et composition chimique 23
1.1.3. Nicotine : Propriétés physicochimiques, biologiques et effets indésirables 25
1.1.4. Types de produits à base de tabac et de nicotine 26
1.1.4.1. Produits traditionnels 26
1.1.4.2. Produits récemment développés 27
1.1.5. Prévalence des produits de tabac 28
1.1.6. Tabac à chiquer : Composition, pharmacocinétique et considérations toxicologiques 29
1.1.7. Réglementation des produits de tabac sans fumée 31

1.2. Spectroscopie infrarouge : Principes et progrès 32
1.2.1. Fondements de la spectroscopie infrarouge 32
1.2.2. Spectromètre infrarouge à transformée de Fourier à réflexion totale atténuée (ATR-FTIR) 34

1.3. Méthodes chimiométriques appliquées à la spectroscopie analytique 39
1.3.1. Généralités 39
1.3.2. Loi de Beer-Lambert : Applications et limitations 40
1.3.3. Méthodes de régression et de reconnaissance de motifs 41
1.3.3.1. Régression linéaire univariée 41
1.3.3.2. Moindres carrés partiels 42
1.3.3.3. Machine à vecteurs de support 43
1.3.3.4. Analyse en composantes principales 44
1.3.3.5. Classification hiérarchique ascendante 45
1.3.3.6. Classification *k*-means 46
1.3.4. Validation du modèle 47
1.3.4.1. Validation croisée 47
1.3.4.2. Validation par un ensemble de test indépendant 48

1.3.5. Activités de pré-calibration 48
1.3.5.1. Sélection des échantillons 48
1.3.5.2. Sélection des variables spectrales 50
1.3.5.3. Prétraitement mathématique des spectres 54
1.3.6. Évaluation du modèle 63
1.3.6.1. Évaluation de la performance des modèles de régression 63
1.3.6.2. Évaluation de la performance des modèles de reconnaissance de motifs 67

1.4. Revues de la littérature et état de l'art 68
1.4.1. Travaux antérieurs sur le tabac sans fumée 68
1.4.1.1. Qualité des feuilles 69
1.4.1.2. Type du produit 69
1.4.1.3. Caractéristiques du produit 69
1.4.1.4. Composition chimique 70
1.4.2. Travaux antérieurs sur la détermination de la nicotine et d'autres paramètres de qualité dans le tabac 73

CHAPITRE 2 : SPECTROSCOPIE ATR-FTIR COMBINÉE À LA CHIMIOMÉTRIE POUR LA QUANTIFICATION DE LA NICOTINE TOTALE DANS DES PRODUITS COMMERCIAUX DU TABAC SANS FUMÉE ALGÉRIEN 80

2.1. Introduction 80

2.2. Partie expérimentale 82
2.2.1. Équipements et logiciels 82
2.2.2. Réactifs et produits chimiques 84
2.2.3. Collecte d'échantillons commerciaux 84
2.2.4. Préparation et analyse des échantillons 85
2.2.4.1. Préparation des extraits 85
2.2.4.2. Préparation des solutions étalons 86
2.2.4.3. Analyse ATR-FTIR 87
2.2.5. Analyse des données 87
2.2.5.1. Détection des échantillons aberrants 87
2.2.5.2. Méthodes de prétraitement et de calibration utilisées 88
2.2.5.3. Validation et évaluation des modèles 88

2.3. Résultats et discussion 89
2.3.1. Interprétation spectrale et analyse exploratoire 89
2.3.2. Optimisation des modèles de calibration 92
2.3.3. Évaluation de la performance des modèles optimaux 109
2.3.4. Evaluation de l'exactitude et de la précision 112
2.3.5. Analyse d'échantillons réels et validation des résultats 114

2.3.5.1. Vérification de l'ajustement spectral 118
2.3.5.2. Valeur d'écart 118
2.3.5.3. Diagramme d'influence 119
2.3.5.4. Contributions des variables 120

2.4. Conclusion 123

CHAPITRE 3 : DÉTERMINATION RAPIDE DES PARAMÈTRES DE QUALITÉ DU TABAC SANS FUMÉE PAR SPECTROSCOPIE ATR-FT-MIR : COMPARAISON DES APPROCHES MATHÉMATIQUES ET D'APPRENTISSAGE AUTOMATIQUE COVENTIONNEL 124

3.1. Introduction 124

3.2. Partie Expérimentale 126
3.2.1. Équipements et logiciels 126
3.2.2. Réactifs et produits chimiques 127
3.2.3. Collecte des échantillons 127
3.2.4. Préparation et analyse des échantillons 128
3.2.5. Mesures de référence des paramètres de qualité du ST 129
3.2.5.1. Teneur en humidité totale 129
3.2.5.2. Détermination du pH 130
3.2.5.3. Teneur en cendres totales 130
3.2.5.4. Teneur en nicotine totale 131
3.2.5.5. Teneur en nicotine non ionisée 131
3.2.6. Analyse des données 131
3.2.6.1. Prétraitement spectral 131
3.2.6.2. Analyse en composantes principales (PCA) 132
3.2.6.3. Classification hiérarchique ascendante (AHC) 134
3.2.6.4. Classification *k*-means 135
3.2.6.5. Régression par moindres carrés partiels (PLSR) 135
3.2.6.6. Régression par machine à vecteurs de support (SVMR) 136
3.2.7. Métriques d'évaluation 137
3.2.8. Limites de détection et de quantification 137

3.3. Résultats et discussion 138
3.3.1. Mesures de référence des paramètres de qualité 138
3.3.2. Interprétation spectrale 145
3.3.3. Analyses de classification 148
3.3.3.1. Méthodes non supervisées 148
3.3.3.2. Méthodes supervisées 156
3.3.4. Prédiction des paramètres de qualité 158
3.3.4.1. Humidité totale 165
3.3.4.2. Niveau de pH 166

3.3.4.3. Cendres totales 168
3.3.4.4. Nicotine totale et nicotine non ionisée 169
3.3.4.5. Analyse EJCR et LOD / LOQ 171

3.4. Conclusion 174

CONCLUSION GÉNÉRALE 175
APPENDICE
A. Liste des abréviations 177
B. Tests de vérification de performance de l'appareil 179

RÉFÉRENCES 181
PUBLICATIONS ET COMMUNICATIONS 197

LISTE DES ILLUSTRATIONS, GRAPHIQUES ET TABLEAUX

Liste des figures

Figure 1.1	Présentation de trois espèces du genre *Nicotiana*. (A) *N. tabacum*, (B) *N. rustica* et (C) *N. glauca*	24
Figure 1.2	Structures chimiques des énantiomères de la nicotine	25
Figure 1.3	Types de produits de tabac sans fumée vendus aux États-Unis	27
Figure 1.4	Produits récemment développés. (A) Composants du système de tabac chauffé électriquement (EHTP), (B) Vue en coupe transversale d'EHTP insérés dans le support, (C) Cigarettes électroniques et (D) Certains produits de thérapie de remplacement de la nicotine	28
Figure 1.5	Formation des TSNAs	30
Figure 1.6	Les modes normaux de vibration moléculaire	34
Figure 1.7	Aperçu des sources lumineuses, des matériaux de guide d'ondes et des principes de détection les plus répandus dans la région spectrale de MIR	35
Figure 1.8	Schéma d'un interféromètre de Michelson	36
Figure 1.9	Réfraction et angle critique à l'interface entre deux milieux	37
Figure 1.10	Configuration d'une ATR simple à mono-réflexion	38
Figure 1.11	Illustration de la manière dont un modèle univarié conduira à des prédictions biaisées lorsque des interférents non-suspectées contribuent de manière variable au signal	41
Figure 1.12	(A) Optimisation de la marge d'erreur pour une SVMR linéaire, (B) Représentation d'hyperplan SVM-C à travers deux classes linéairement séparables	44
Figure 1.13	Scores des PCs pour les données d'exemple. Le diagramme révèle que les composés se regroupent en deux clusters distincts	45

Figure 1.14	Regroupement de données par HCA. (A à E) Groupes formés à chaque étape, délimités par des lignes pointillées, (F) Dendrogramme représentant la hiérarchie des groupes	46
Figure 1.15	L'algorithme de Kennard-Stone en action (la croix rouge dans "A" est le centre d'étalonnage et la ligne verte continue dans "B" est le maximum des huit distances minimales)	49
Figure 1.16	(A) Système analytique idéal avec un analyte d'intérêt (spectre bleu), deux constituants supplémentaires (spectres rouge et vert), et une ligne de base saturant le détecteur dans la plage des variables de 250 à 300 (spectre bleu clair). (B) Vecteur des coefficients de régression PLS pour l'analyte d'intérêt dans ce système	52
Figure 1.17	Diagramme en barres montrant la RMSECV pour le système de la Figure 1.16 en utilisant la méthode *i*-PLS avec des intervalles de 15 variables. La ligne rouge montre le spectre de calibration moyen	53
Figure 1.18	VIP pour la sélection des bandes caractéristiques (scores VIP > 1, les régions importantes sont marquées en gris)	54
Figure 1.19	(A) Spectres de calibration de 50 échantillons contenant trois analytes et un signal de fond constant. (B) Spectres centrés sur la moyenne	55
Figure 1.20	(A) Spectre avec un niveau élevé de bruit aléatoire ; (B) Application du lissage Savitzky-Golay avec première fenêtre de 5 variables (ligne bleue et cercles), polynôme ajusté du troisième degré (ligne rouge) et valeur au point central estimée par cet ajustement (cercle noir); (C) Effet produit par le filtre en utilisant une fenêtre de 5 variables ; et (D) Spectre résultant après l'utilisation d'une fenêtre de 11 variables	58
Figure 1.21	(A) Courbes gaussiennes à décalages et intensités différents, (B) Dérivée première des courbes et (C) Dérivée seconde des courbes	58
Figure 1.22	Spectres NIR de 80 échantillons de fructo-oligosaccharides, originaux (A) et prétraités par SNV (B)	60
Figure 1.23	(A) Spectres de réflectance diffuse NIR de la cellulose, la tendance non linéaire est indiquée à peu près par la courbe rouge en pointillés ; et (B à E) Mêmes spectres prétraités par SNV et DT avec un polynôme d'ordre 1 à 4, respectivement	61

Figure 1.24	(A) Spectres avec décalage additif de la ligne de base ; (B) Spectres avec effet multiplicatif ; (C et D) Tracés des longueurs d'onde individuelles par rapport au spectre moyen respectivement aux effets de diffusion additive et multiplicative ; et (E) Spectres corrigés par MSC	62
Figure 1.25	Exemples de régions de confiance dans le plan pente-interception. L'ellipse bleue montre un modèle précis alors que l'ellipse rouge indique que le modèle correspondant est imprécis	66
Figure 2.1	Caractéristiques principales du (A) Spectromètre Nicolet iS10 et (B) Accessoire Smart iTR	83
Figure 2.2	Spectres FTIR-ATR des étalons de NCT (rouge et bleu) à des concentrations de 8 $mg.ml^{-1}$, ainsi que le spectre du blanc analytique (vert). Les zones spectrales favorables pour l'analyse de la NCT sont indiquées en gris	89
Figure 2.3	Exemple de détection des valeurs aberrantes appliquée aux répliques d'une solution étalon traitée. (A) Graphique d'influence des Q-résidus contre T^2 de Hotelling et (B) Graphique des scores PCA avec T^2 de Hotelling, tous deux ("A" et "B") à une limite critique 25 %	91 et 92
Figure 2.4	Diagrammes radiaux comparant les performances des modèles optimisés univariés (A) et PLSR (B)	107
Figure 2.5	Régions 2+7 prétraitées par LBC-BO (A) et par LBC-BO-EMSC-SGS (C) et leurs droites de régression prédites contre réelles (B) et (D) respectivement, calculées avec l'algorithme PLS-1 pour les ensembles de calibration et de validation	109
Figure 2.6	EJCR dans le plan pente-interception réalisé pour les meilleurs modèles basés sur les moindres carrés ordinaires (A) et les moindres carrés pondérés (B)	113
Figure 2.7	Spectres FTIR-ATR d'extraits de produits contenant (vert) et ne contenant pas de tabac (bleu) par rapport à l'étalon traité (rouge)	115
Figure 2.8	Spectres FTIR-ATR moyens des produits commerciaux de ST présentés en fonction de leurs teneurs en nicotine totale	116

Figure 2.9	Graphique des Q-résidus contre T^2 de Hotelling pour les échantillons réels. Les lignes rouges représentent les limites critiques associées	119
Figure 2.10	Contributions des variables spectrales au modèle (A) et aux résidus (B) en utilisant la statistique de T^2 de Hotelling et des Q-résidus, respectivement, pour un nouvel échantillon (réplique n°4 de l'échantillon 4)	121
Figure 2.11	Contribution moyenne des variables aux résidus pour les 17 échantillons commerciaux	122
Figure 3.1	Graphique d'influence des F-résidus contre T^2 de Hotelling, aux limites critiques de 5 %, appliqué sur (A) les mesures de référence et (B) les spectres FTIR-ATR, du nombre total d'échantillons pour l'élimination d'aberration inter-groupe	133
Figure 3.2	Méthode Kennard-Stone−PCA appliquée sur l'ensemble de mesures de références pour la division des échantillons	134
Figure 3.3	Distribution de fréquence et courbe normale des teneurs en humidité, pH, cendres, nicotine totale et nicotine non ionisée des échantillons de ST algérien utilisés pour (A) l'entraînement et (B) le test des modèles	144
Figure 3.4	Spectres FTIR-ATR non-traités obtenus pour les 96 produits sélectionnés de ST	145
Figure 3.5	Spectres ATR-FT-MIR représentatifs normalisés. Le moyen (A) et l'écart-type (B) des ST comparés aux ingrédients les plus susceptibles d'être utilisés dans leur préparation	146
Figure 3.6	Graphique de scores PCA obtenus sur la base des mesures de référence pour l'étude de la tendance générale des produits commerciaux sélectionnés	149
Figure 3.7	Évolution de la variance intra-classe pour (A) le regroupement hiérarchique et (B) la classification *k*-means	150
Figures 3.8	Dendrogramme AHC représentant la hiérarchie des échantillons commerciaux de Chemma obtenus sur la base des mesures de référence.	151
Figure 3.9	Graphiques des scores PCA affichant le regroupement des échantillons selon les résultats de l'AHC (A et B) et les résultats de *k*-means (C et D)	154

Figure 3.10	(A) Scores PCA pour les trois premières PCs affichant le partitionnement des échantillons selon les résultats de *k*-means. (B) Biplot des scores et des coefficients PCA	155
Figure 3.11	Matrices de confusion de l'ensemble d'entraînement des modèles (A) PLS-DA et (B) SVM-C utilisés pour la classification des produits commerciaux	158
Figure 3.12	Exemples de sélection de bandes caractéristiques par VIP et *i*-PLS pour les modèles PLSR optimaux de (A) humidité totale, (B) niveau de pH, (C) nicotine totale et (D) nicotine non ionisée	163
Figure 3.13	(A et B) Lignes de régression des valeurs réelles par rapport aux valeurs prédites calculées respectivement par PLSR et SVMR pour l'humidité. (C) Coefficients de régression pour le nombre pertinent de LVs dans le modèle PLSR correspondant	165 et 166
Figure 3.14	(A et B) Lignes de régression des valeurs réelles par rapport aux valeurs prédites calculées respectivement par PLSR et SVMR pour le pH. (C) Coefficients de régression pour le nombre pertinent de LVs dans le modèle PLSR correspondant	167
Figure 3.15	(A et B) Lignes de régression des valeurs réelles par rapport aux valeurs prédites calculées respectivement par PLSR et SVMR pour les cendres. (C) Coefficients de régression pour le nombre pertinent de LVs dans le modèle PLSR correspondant	168 et 169
Figure 3.16	(A et B) Lignes de régression des valeurs réelles par rapport aux valeurs prédites calculées respectivement par PLSR et SVMR pour la nicotine totale. (C) Coefficients de régression PLS pour le nombre pertinent de LVs	170
Figure 3.17	(A et B) Lignes de régression des valeurs réelles par rapport aux valeurs prédites calculées respectivement par PLSR et SVMR pour la nicotine libre. (C) Coefficients de régression PLS pour le nombre pertinent de LVs	171
Figure 3.18	EJCRs dans le plan pente-interception réalisés pour les modèles optimaux de (A) PLSR et (B) SVMR basés sur la méthode des moindres carrés ordinaires	172

Liste des tableaux

Tableau 1.1	Classification des espèces de tabac	24
Tableau 1.2	Niveaux maximaux admissibles de diverses toxines dans le « Snus » suédois avec une teneur en eau de 50 %	32
Tableau 1.3	Les FOMs analytiques pour les modèles de calibration univarié et multivarié	64
Tableau 1.4	La statistique RPD pour la prédiction des aliments, des sols et des facteurs de fonctionnalité	65
Tableau 1.5	Structure de la matrice de confusion	67
Tableau 1.6	Protocoles de laboratoire couramment utilisés pour l'analyse de la nicotine totale dans quelques produits de ST	75
Tableau 1.7	Divers méthodes chimiométriques développées pour l'analyse quantitative et qualitative des échantillons de tabac	77
Tableau 2.1	Caractéristiques des ensembles de calibration et de test	86
Tableau 2.2	Paramètres de régression des procédures de calibration, de validation croisée et de prédiction de la nicotine dissous directement dans le chloroforme et soumis au traitement, obtenus par la loi de Beer-Lambert (analyse univariée)	94
Tableau 2.3	Paramètres de régression des procédures de calibration, de validation croisée et de prédiction de la nicotine dissous directement dans le chloroforme et soumis au traitement, obtenus par PLS-1 (analyse multivariée)	99
Tableau 2.4	Facteurs de mérite pour les modèles univariés et multivariés sélectionnés	110
Tableau 2.5	Teneurs en nicotine totale calculée en mg/g de produit dans des échantillons réels obtenues par l'approche proposée	117
Tableau 3.1	Mesures de référence des paramètres de qualité dans les produits commerciaux de ST collectés	140

Tableau 3.2	Profil d'échantillons de ST algérien basé sur cinq paramètres de qualité avant et après la suppression des valeurs aberrantes	143
Tableau 3.3	Performances de discrimination dans les procédures de calibration et de CV en utilisant les spectres FTIR avec différentes méthodes de prétraitement	157
Tableau 3.4	Paramètres de régression des procédures de calibration, de validation croisée et de prédiction des paramètres de qualité pour les modèles calculés par PLSR en utilisant différents prétraitements spectraux et méthodes de sélection des variables spectrales	159
Tableau 3.5	Paramètres de régression des procédures de calibration, de validation croisée et de prédiction pour les modèles analytiques calculés par SVMR en utilisant différents prétraitements spectraux, fonctions de noyau et paramètres de réglage	161
Tableau 3.6	Métriques d'évaluation pour les modèles optimaux de PLSR et SVMR	164
Tableau 3.7	Limites de détection et de quantification calculées en utilisant les approches proposées dans la littérature pour les modèles PLSR et SVMR optimaux	173

INTRODUCTION

En Algérie, le tabac sans fumée, connu localement sous le nom de « Chemma », est une forme de tabac populaire consommée par voie orale. Elle se présente sous forme de poudre fine, généralement composée d'un mélange de feuilles de *Nicotiana rustica* séchées, de sels inorganiques, de cendres alcalines et/ou d'autres matières végétales, humidifiées avec de l'eau traitée [1]. La Chemma est consommée en plaçant une pincée entre les lèvres et les dents, directement ou enveloppée dans du papier à cigarette. Sa popularité parmi la jeune génération est en hausse, que ce soit pour un usage récréatif, comme alternative à la cigarette, ou simplement comme une émulation sociale entre pairs.

La production de Chemma repose principalement sur le tabac cultivé dans les régions orientales du pays, en particulier dans les régions du Grand Constantinois et des Oasis. Diverses variétés de plantes sont cultivées, donnant lieu à plusieurs sortes commerciales de qualités variées, souvent nommés selon la zone de culture [2]. Un produit typique de Chemma peut inclure un mélange de différentes variétés de tabac.

Jusqu'à il y a quelques années, le groupe MADAR Holding (anciennement SNTA) détenait le monopole de la fabrication et de la commercialisation du tabac à chiquer algérien [3]. Cependant, en raison d'un contrôle insuffisant de la contrefaçon par les autorités publiques, le secteur informel a pris une ampleur considérable sur le marché. Ce secteur va au-delà de la simple contrefaçon de produits authentiques, en produisant et en inondant le marché avec ses propres marques à bas prix, ciblant spécifiquement les communautés à faibles revenus. Plus de 50 marques différentes sont vendues par les buralistes, sans aucune réglementation sur l'emballage. Souvent, la composition, l'adresse du fabricant, la date de péremption et d'autres informations d'identification ne sont pas correctement déclarées.

Motivés par la maximisation des profits, les producteurs illégaux de Chemma n'hésitent pas à utiliser des matériaux de qualité douteuse, y compris des ingrédients périmés ou même pourris. Cette pratique soulève des inquiétudes quant aux impacts potentiels sur la santé publique, d'autant plus que la contrefaçon devient une préoccupation majeure.

Face à la prolifération croissante des produits de la contrefaçon, les associations locales de protection des consommateurs et le fabricant légal de tabac peine à suivre le rythme. En effet, leurs analyses se limitent aux caractéristiques physiques et détectables des produits, telles que le taux de nicotine, la granulométrie, la couleur, l'odeur des produits finis et à l'emballage, ainsi qu'un contrôle microbiologique [3].

L'objectif principal fixé, dans le cadre de cette thèse, concerne le développement et la validation de méthodes analytiques simples, efficaces, peu coûteuses, à haut débit et accessibles à la majorité des laboratoires d'analyse afin de contribuer à la lutte contre la contrefaçon et de garantir la qualité et la conformité des produits de tabac sans fumée. La spectrométrie moyen-infrarouge à transformée de Fourier à réflectance totale atténuée (ATR-FT-MIR) s'est imposée comme technique de choix pour cette étude. Les résultats obtenus contribueront de manière significative à l'enrichissement des connaissances sur cette matrice, en fournissant des informations précieuses aux chercheurs, aux industriels et aux professionnels travaillant dans les domaines de la sécurité alimentaire et de la santé publique.

Ce manuscrit est structuré en trois chapitres:

Le premier chapitre expose l'état de l'art, abordant: i) les bases théoriques sur le tabac et les produits à base de tabac, notamment le tabac à chiquer ; ii) les principes et les avancées de la spectroscopie ATR-FTIR et des méthodes chimiométriques connexes ; iii) une synthèse des travaux de recherche antérieurs sur le tabac sans fumée mondial, ainsi que sur la détermination de la nicotine et d'autres paramètres de qualité dans le tabac.

Le deuxième chapitre se concentre sur la quantification précise de la nicotine totale dans les extraits de Chemma algérienne en utilisant, pour la première fois, la spectroscopie ATR-FTIR à mono-réflexion couplée à deux méthodes de régression (univariée et par moindres carrés partiels). Une méthode d'extraction sélective a été utilisée dans la préparation des échantillons, une approche de micro-échantillonnage a été appliquée pour l'analyse, et la chimiométrie a été intensivement utilisée pour prétraiter les spectres, évaluer les performances des modèles et valider les résultats selon les dernières approches de la littérature.

Le troisième chapitre présente une nouvelle approche méthodologique pour une analyse de routine en utilisant la même technique analytique soutenue par des méthodes mathématiques et d'apprentissage automatique pour aborder les trois aspects suivants: i) Identification des principaux ingrédients inconnus dans les produits commerciaux du tabac à chiquer ; ii) Classification des échantillons à l'aide de différentes méthodes d'apprentissage non supervisées et supervisées ; et iii) Détermination rapide de cinq paramètres de qualité: humidité, pH, cendres, nicotine totale et nicotine non ionisée par deux méthodes de régression (moindres carrés partiels et machine à vecteurs de support).

Nous avons conclu cette thèse par une synthèse générale des résultats les plus importants, mettant en avant les avantages, les inconvénients et les perspectives liés à chaque méthode développée.

CHAPITRE 1

FONDEMENTS THÉORIQUES ET BIBLIOGRAPHIQUES

1.1. Tabac et produits à base de tabac

1.1.1. Histoire du tabac

Le tabac et l'humanité ont été associés de la même manière que la nourriture et le thé depuis avant le début de l'histoire [4]. On pense que la plante de tabac a ses origines entre l'Amérique du Nord et l'Amérique du Sud, remontant à environ 1400 − 1000 av. J.-C. Les Amérindiens auraient été les premiers à fumer et à prendre du tabac en poudre dès les années 1400 [5]. Après la découverte des Amériques par Christophe Colomb, la plante a été introduite en Europe par les premiers explorateurs, où elle était utilisée par les Portugais et les Espagnols [4]. Au milieu du XVIe siècle, Jean Nicot, ambassadeur de France au Portugal, a introduit le tabac et les graines de tabac en France, d'où le nom "Nicotiana". Plus tard, il a été adopté par la société et réexporté vers le reste du monde à mesure que la colonisation européenne progressait. Par la suite, de nombreuses civilisations ont expérimenté indépendamment les uns des autres les effets de l'auto-administration de doses de tabac [5, 6].

La pratique du tabagisme semble avoir émergé de la prise de tabac en poudre (ou « snuffing » en anglais), car les instruments pour snuffing figurent parmi les artefacts liés au tabac les plus anciens qui aient été découverts. Cependant, le tabac n'était pas seulement pris en poudre et fumé, mais aussi mâché, mangé, bu (comme du thé), étalé sur le corps (pour tuer les poux et autres parasites), et utilisé en gouttes pour les yeux et les lavements. En plus qu'il était utilisé à des fins médicinales pour ses propriétés analgésiques et antiseptiques, le tabac était aussi soufflé au visage des guerriers avant la bataille et sur les champs avant le semis (il est toujours utilisé comme insecticide en agriculture) [4].

En Algérie, l'histoire du tabac est étroitement liée à la période coloniale. Pendant l'ère coloniale française, la culture du tabac a été introduite en Algérie dans le cadre d'initiatives plus larges de développement agricole. Les colons français ont établi des plantations de tabac, notamment dans les régions fertiles du nord de l'Algérie, en particulier en Kabylie et à Constantine. Le tabac est devenu une culture rentable importante, et sa culture s'est étendue pour répondre à la demande croissante tant sur le marché intérieur qu'international [7].

Après avoir obtenu son indépendance en 1962, l'Algérie a continué la production de tabac, mais en recentrant son attention sur la satisfaction des besoins intérieurs. Au fil des ans, le gouvernement a pris des mesures pour réglementer l'industrie et contrôler la consommation de tabac. Diverses politiques ont été mises en œuvre pour décourager le tabagisme, notamment des campagnes de sensibilisation du public et la taxation des produits du tabac. Comme de nombreux autres pays, l'Algérie a été confrontée aux implications sanitaires de l'utilisation du tabac et a cherché à trouver un équilibre entre considérations économiques et préoccupations de santé publique. Aujourd'hui, le tabac reste une partie du paysage agricole algérien, mais les efforts pour réduire la prévalence du tabagisme et ses risques sanitaires associés continuent d'être une priorité pour le gouvernement [8].

1.1.2. Tabac : Taxonomie et composition chimique

Le tabac est le produit agricole des feuilles de plantes du genre *Nicotiana*, qui sont des plantes herbacées et des arbustes de la famille des Solanacées (Tableau 1.1) [9]. *Nicotiana*, avec 76 espèces naturelles, est le sixième plus grand genre de la famille, réparti naturellement dans les Amériques et en Australie, et avec une seule espèce en Afrique [10]. Cependant, la distribution actuelle est plus étendue, avec la culture de *N. tabacum* et *N. rustica* (Figures 1.1A et B) utilisées pour la fabrication des produits du tabac, et l'espèce envahissante *N. glauca* (Figure 1.1C), originaire d'Amérique du Sud, mais occupant désormais une niche mondiale [10].

Tableau 1.1 : Classification des espèces de tabac [9].

Classification scientifique	
Règne	Plantae
Clade	Trachéophytes
Clade	Angiospermes
Clade	Eudicots
Clade	Astérides
Ordre	Solanales
Famille	Solanaceae
Tribu	Nicotianeae
Genre	Nicotiana L.

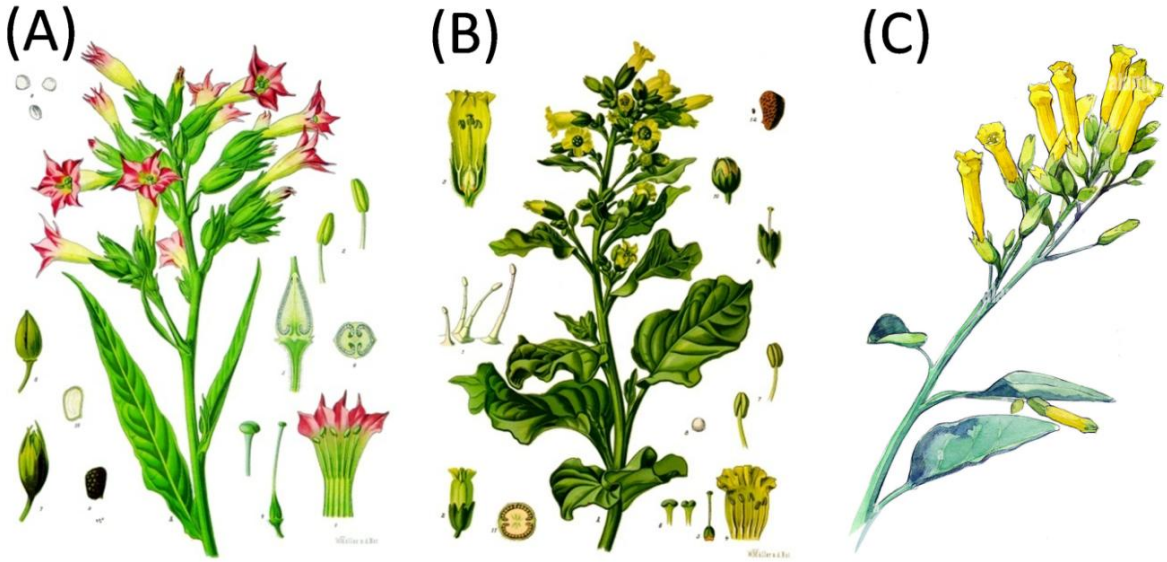


Figure 1.1 : Présentation de trois espèces du genre *Nicotiana*. (A) *N. tabacum*, (B) *N. rustica* et (C) *N. glauca* [11].

Au sein du genre *Nicotiana*, chaque espèce possède une composition unique en alcaloïdes et autres métabolites secondaires. Ces différentes combinaisons de molécules donnent à chaque espèce des propriétés et des effets spécifiques. Par conséquent, la toxicité et l'impact sur la santé de ces diverses variétés de tabac peuvent varier de manière significative [12].

En fait, le tabac cultivé (*N. tabacum*) est l'une des espèces les plus étudiées sur le plan chimique et biologique du règne végétal, avec plus de 2500 métabolites caractérisés, mis à jour par des recherches continues. Les composés étudiés comprennent des alcaloïdes, des composés aromatiques, des

flavonoïdes, des volatiles, des sesquiterpénoïdes, des alcools diterpéniques, des esters de glucides, etc [10, 13].

Les alcaloïdes sont des composés importants présents dans les tabacs, essentiels dans la défense des plantes contre les pathogènes et les herbivores [14]. Dans la plupart des souches de tabac, l'alcaloïde principal est la nicotine, suivie des alcaloïdes mineurs tels que, la nornicotine, l'anatabine et l'anabasine, respectivement, mais un seul alcaloïde prédomine généralement dans les autres espèces [14, 15]. La nicotine est formée dans un processus impliquant plusieurs enzymes dans les racines, puis translocatée vers les feuilles de la plante [14].

1.1.3. Nicotine : Propriétés physicochimiques, biologiques et effets indésirables

La nicotine est la substance la plus stimulante et pharmacologiquement active du tabac [1]. Elle se produit naturellement dans une proportion d'environ 0,5 à 14,0 % du poids sec de la plante et représente environ 95 % de la teneur totale en alcaloïdes [15]. La nicotine dans le tabac est principalement l'isomère lévogyre (S); seulement 0,1 à 1,0 % de la teneur totale en nicotine est (R)-nicotine (Figure 1.2) [15]. Ce composé chiral est également connu sous le nom de 3-(1-méthylpyrrolidin-2-yl)pyridine, dans la nomenclature IUPAC, avec la formule brute $C_{10}H_{14}N_2$ et une masse moléculaire de 162,23 g/mol [17]. C'est une base azotée dibasique, ayant pK_1 = 6,16 et pK_2 = 10,96, avec une rotation spécifique de $[\alpha]_D$ = −169°, un point d'éclair de 95 °C et un point d'ébullition de 274,5 °C à 1 atm [16-18].

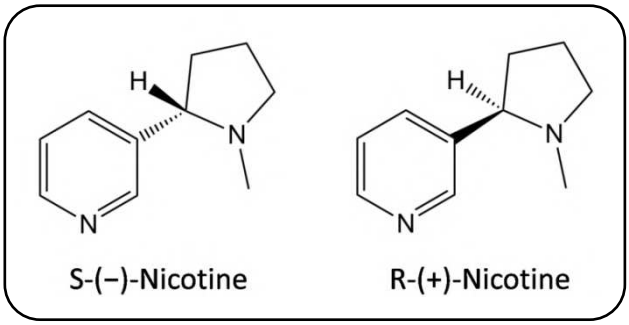


Figure 1.2 : Structures chimiques des énantiomères de la nicotine.

La nicotine pure est un liquide huileux clair qui devient jaune-brun au contact de l'air. Elle est miscible avec l'eau sous sa forme de base amine neutre entre 60 et 210 °C. Cependant, elle se partitionne préférentiellement dans les solvants organiques (le coefficient de distribution le plus bas étant avec le chloroforme). Ainsi, elle peut facilement être extraite des solutions aqueuses par extraction par solvant [18].

La dépendance à la nicotine est la principale raison pour laquelle les gens continuent d'utiliser des produits du tabac [19]. Il a été démontré que la nicotine traverse facilement les membranes biologiques et la barrière hémato-encéphalique. Son absorption peut se faire par la cavité buccale, la peau, les poumons, la vessie urinaire et le tractus gastro-intestinal [20].

La nicotine est classée comme un poison. Il a été démontré qu'elle affecte une grande variété de fonctions biologiques, allant de l'expression génique, à la régulation de la sécrétion hormonale et des activités enzymatiques [19], et qu'elle pourrait jouer un rôle dans le développement de maladies cardiovasculaires et de perturbations de la reproduction [20]. La limite inférieure estimée de dose pour des résultats fatals est de 500 à 1000 mg de nicotine ingérée pour un adulte (6,5 à 13 mg/kg) [16].

1.1.4. Types de produits à base de tabac et de nicotine

On peut classer les produits de tabac consommables en quatre catégories principales:

1.1.4.1. Produits traditionnels

- Fumés: Produits contenant du tabac destinés à être fumés. Les produits du tabac fumé courants comprennent les cigarettes, les cigares et le tabac à chicha [22].
- Sans fumée (« smokeless tobacco »): Collection très diversifiée de préparations contenant du tabac qui sont consommés en mâchant, suçant, sniffant ou sous d'autres formes de non-combustion [23]. Les produits du tabac sans fumée courants (Figure 1.3) comprennent le tabac à mâcher (en vrac « loose leaf », en

bloc « plug » ou en torsade « twist »), le tabac à priser (« snuff » humide et sec), le Snus et les produits dissolvables (pastilles « lozenges », bâtonnets « sticks », bandes « strips » et billes « orbs ») [24, 25].

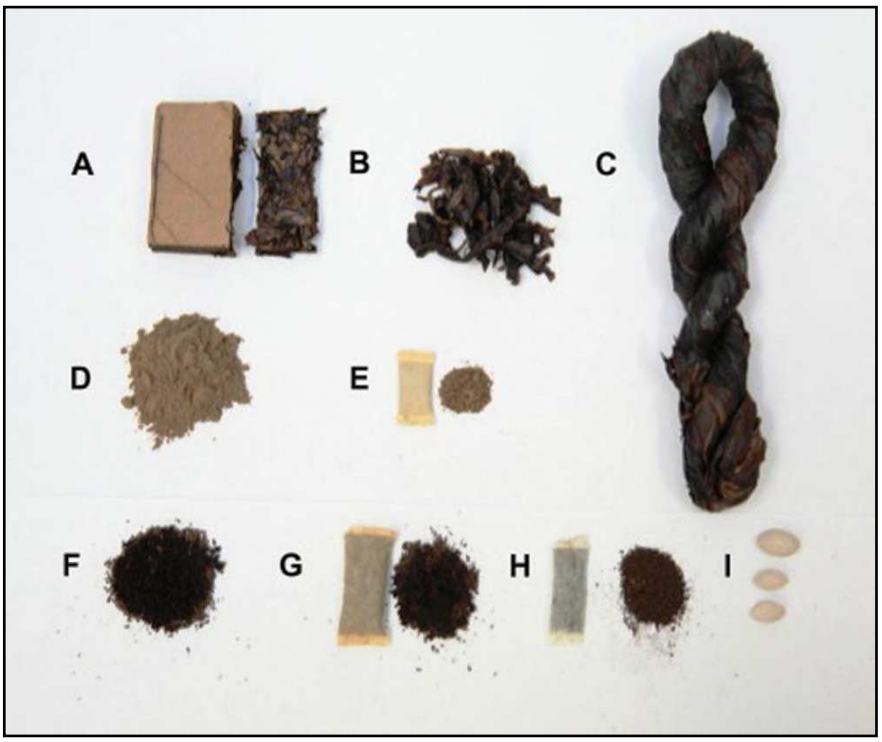


Figure 1.3 : Types de produits de tabac sans fumée vendus aux États-Unis. "A", Plug ; "B", Loose leaf ; "C", Twist ; "D", Snuff sec (loose) ; "E", Snuff sec (pouch) ; "F", Snuff humide ; "G" et "H", Snus ; et "I", Dissolvable [25].

1.1.4.2. Produits récemment développés

- Chauffés: Produits contenant du tabac qui est chauffé afin de produire un aérosol ou une suspension de particules inhalable. Aussi connus sous le nom de "produits à tabac chauffé sans combustion" (« heat-not-burn » en anglais) ou "cigarettes sans fumée" (en anglais « smokeless cigarettes ») (Figure 1.4A et B). Exemples: cylindres de tabac IQOS et préparations de tabac vaporisées dans les vaporisateurs à herbes sèches Pax [22, 26].
- Uniquement à base de nicotine (« Nicotine-only products »): produits ne contenant pas de tabac mais de la nicotine, extraite du tabac ou synthétique. Exemples: e-liquides (vaporisés à l'aide d'une cigarette électronique ou d'un vaporisateur), sachets de nicotine et certains produits de thérapie de remplacement de la nicotine [26] (Figure 1.4C et D). Une sous-catégorie de

produits à base de nicotine comprend certains produits constitués d'ingrédients à base de plantes ou de fines herbes infusés à la nicotine, comme les cigarettes à base de plantes nicotinisées et le tabac sans fumée à base d'herbes [22, 26].

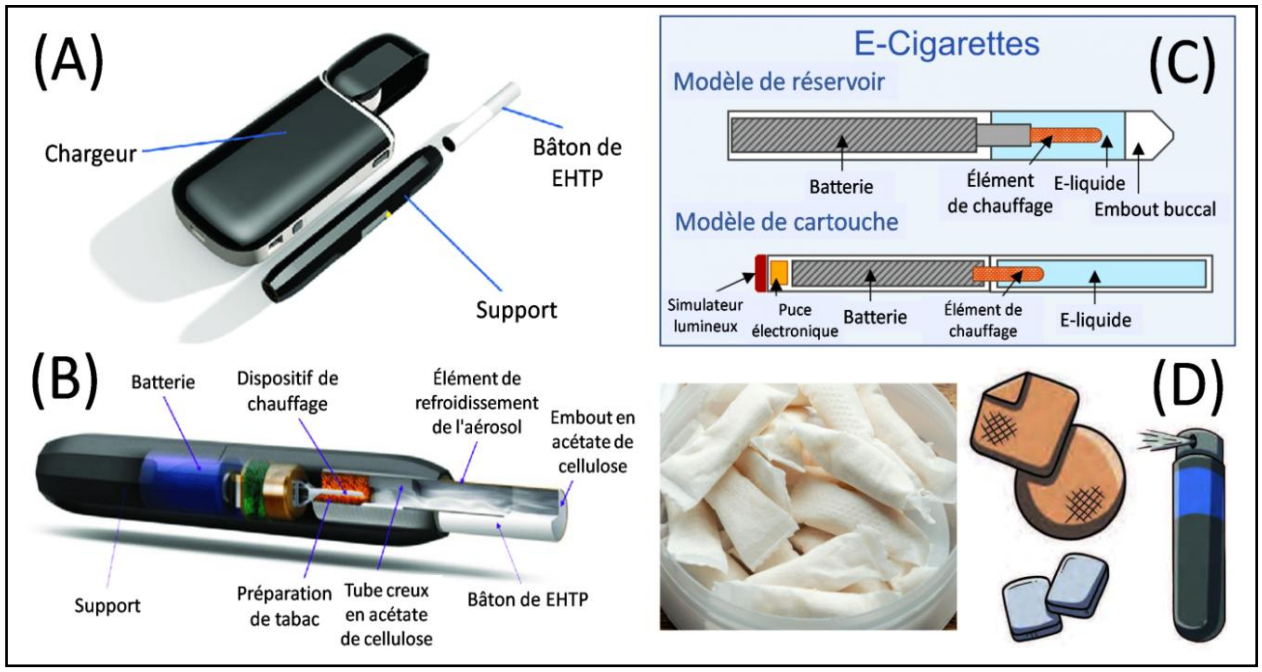


Figure 1.4 : Produits récemment développés. (A) Composants du système de tabac chauffé électriquement (EHTP) [27], (B) Vue en coupe transversale de EHTP insérés dans le support [27], (C) Cigarettes électroniques [28] et (D) Certains produits de thérapie de remplacement de la nicotine [29].

1.1.5. Prévalence des produits de tabac

Le fumage, en particulier la cigarette, est la forme la plus courante de consommation de tabac dans le monde entier. Le tabac sans fumée (ST) est la deuxième forme la plus répandue de consommation [1]. En 2015, il y avait environ un milliard de fumeurs et plus de 350 millions d'utilisateurs de ST dans le monde. La plus grande part, environ 80 %, vit dans des pays à faible et moyen revenu [23, 30].

À l'échelle mondiale, l'utilisation de ST était responsable d'environ 315 000 décès en 2016. En 2017, la charge mondiale de morbidité causée par la consommation de ST s'élevait à environ 8,7 millions d'années de vie ajustées en

fonction de l'incapacité [23]. Une prévalence moyenne de l'utilisation de ST de plus de 10 % est signalée en affichant au Myanmar (29,6 %), au Bangladesh (27,2 %), en Inde (25,9 %), au Népal (18,6 %), en Suède (17 %), au Sri Lanka (15,8 %), en Ouzbékistan (11,3 %) et au Yémen (10,7 %). Les pays présentant une prévalence d'utilisation comprise entre 5 et 10 % comprennent l'Algérie, la Mauritanie et la Tunisie [5].

1.1.6. Tabac à chiquer : Composition, pharmacocinétique et considérations toxicologiques

Le tabac à chiquer (ou en poudre humidifié, en anglais « moist snuff ») est une sous-catégorie de produits de tabac sans fumée buccaux, populaire en Amérique du Nord, en Scandinavie (où il est connu sous le nom de « Snus »), en Asie du Sud (par exemple au Bangladesh, au Bhoutan et en Inde) et dans certaines parties de l'Afrique (par exemple en Algérie, au Soudan et au Nigeria) [31]. Le tabac à chiquer est un mélange chimique complexe contenant des feuilles de tabac séchées finement broyées mélangées à de l'eau, des charges organiques / inorganiques, des conservateurs, des exhausteurs de goût (sel), des acidifiants / alkalinisants et des arômes [31-33]. Ensuite, le mélange est soumis à des procédures de fermentation afin d'obtenir un produit final aux propriétés physiques et à la composition chimique adaptées [34]. En Afrique, le tabac à chiquer est souvent mélangé à des substances alcalines telles que des cendres végétales / bois ou de la potasse, ce qui forme un produit hautement irritant [1].

Le tabac à chiquer peut être utilisé pour fournir des niveaux psychoactifs et de dépendance de nicotine. Une tolérance se développe avec une utilisation répétée, amenant l'utilisateur à augmenter la dose de nicotine par une utilisation accrue et/ou un passage à des produits à rendement nicotinique plus élevé [35]. De plus, l'utilisation de ST par les mineurs s'est également avérée associée à une probabilité accrue de fumage ultérieur, de consommation excessive d'alcool et de consommation du cannabis [36].

Généralement, lorsqu'un ST est utilisé, la nicotine est extraite du tabac, puis traverse la muqueuse buccale pour entrer dans la circulation sanguine et, par la

suite, dans le cerveau, où elle exerce des effets pharmacologiques conduisant à la dépendance [37]. La disponibilité de la nicotine à partir d'un tabac à chiquer donné est influencée par plusieurs facteurs, tels que la quantité de nicotine dans le produit, le pH du produit, le temps passé dans la bouche, la présence ou l'absence d'une poche externe et la taille de la coupe de tabac [35, 38]. Augmenter le pH d'un produit du tabac transforme la nicotine en sa forme non ionisée; cette forme est plus facilement absorbée dans la bouche, et pour la nicotine, comme pour d'autres drogues, la vitesse d'absorption est un déterminant majeur de la dépendance [32, 37].

Bien que la nicotine elle-même ne soit pas carcinogène, le ST peut délivrer à l'utilisateur un mélange complexe de carcinogènes, de promoteurs de tumeurs et de co-carcinogènes [19]. En particulier, le consommateur de tabac à chiquer est exposé à plus de 30 composés nocifs, notamment des contaminants tels que des résidus de pesticides et d'herbicides, des hydrocarbures aromatiques polycycliques, des métaux lourds ou des traces d'éléments radioactifs, des composés organiques volatils, et des nitrosamines spécifiques du tabac (TSNAs), le groupe de carcinogènes le plus actif (Figure 1.5) [39-42].

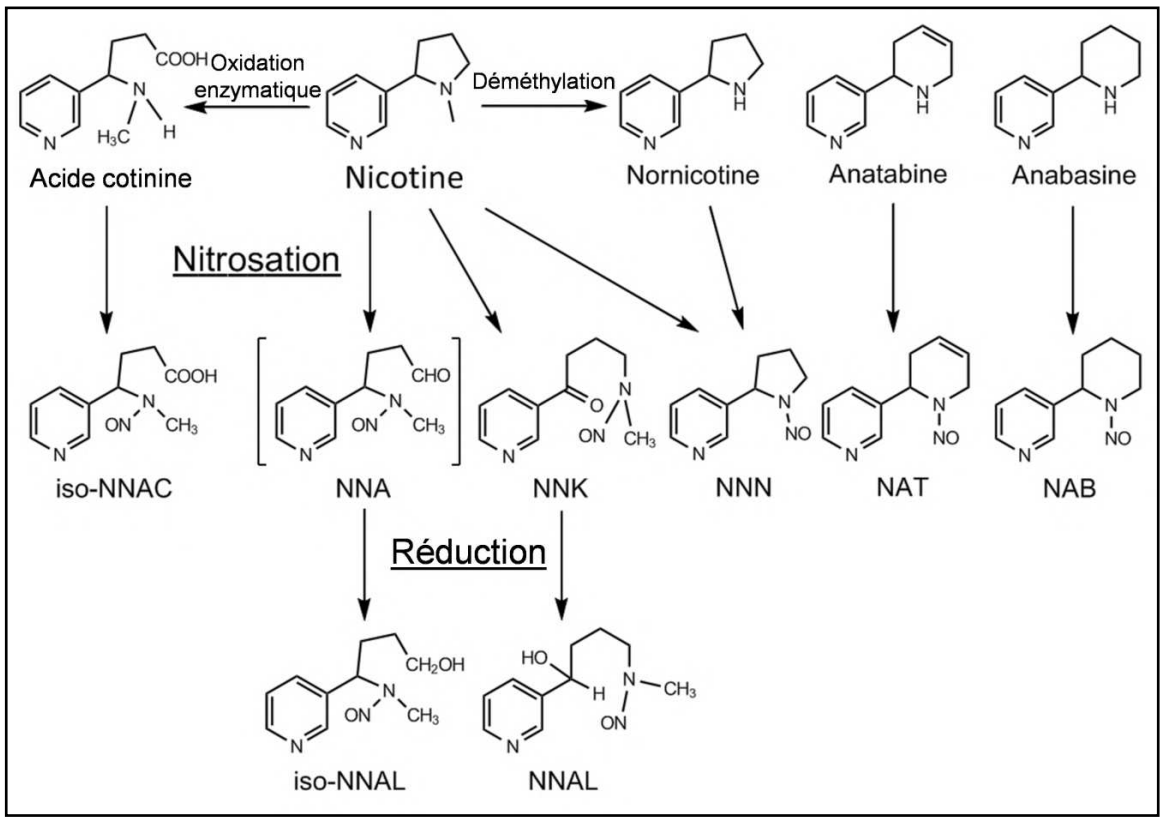


Figure 1.5 : Formation des TSNAs [39].

Les TSNAs les plus carcinogènes couramment présentes sont la 4-(méthylnitrosamino)-1-(3-pyridyl)-1-butanone (NNK) et la N'-nitrosonornicotine (NNN), tandis que la N'-nitrosoanabasine (NAB) est un carcinogène faible et la N'-nitrosoanatabine (NAT) semble être inactive [43]. Les précurseurs des TSNAs sont la nicotine et les alcaloïdes mineurs du tabac, tandis que l'agent nitrosant majeur dans le ST est le nitrite. Ce dernier est formé à partir de nitrate par activité microbienne pendant le séchage, la fermentation et le stockage du tabac à des températures élevées [39, 44]. D'autres facteurs peuvent jouer un rôle significatif dans l'induction du cancer, notamment la variété de plante, l'utilisation de tabac reconstitué et le vieillissement du produit [5].

1.1.7. Réglementation des produits de tabac sans fumée

Dans les pays occidentaux, la réglementation des ST est régie par la législation alimentaire, et les autorités mettent en œuvre diverses stratégies pour réduire son accessibilité et son utilisation. Notamment, la « Food and Drug Administration (FDA) » des États-Unis supervise rigoureusement la fabrication, le marketing et la vente des produits du tabac et impose des normes strictes pour contrôler les niveaux de nicotine et d'autres ingrédients [45]. En complément de ces efforts, les « Centers for Disease Control and prevention (CDC) » apportent un soutien essentiel à la FDA en lui fournissant une assistance technique et une expertise scientifique. En ce qui concerne le tabac à chiquer, les CDC ont établi un protocole de laboratoire standardisé pour l'analyse de la teneur en nicotine, de l'humidité totale et du pH de tous les produits fabriqués, importés ou emballés aux États-Unis [46, 47].

En Suède, environ la moitié du tabac est consommée sous forme de « Snus »; cela a réduit les niveaux de mortalité liée au tabac d'environ la moitié par rapport au reste de l'Union Européen [48]. Cela aussi souligne que la réduction des constituants nocifs est techniquement réalisable et offre une opportunité de façonner le marché du tabac et de veiller à ce que si de tels produits sont utilisés, ils soient mis sur le marché avec un niveau élevé de protection pour le consommateur [48, 49].

Les approches réglementaires pourraient inclure [48]:

- Établissement de normes maximales pour une gamme de toxines cibles impliquées dans les principales maladies liées au tabac.
- Relier la proportion de toxines à la quantité de nicotine, et réguler les additifs.
- Imposer des restrictions de durée de conservation car certains des contaminants changent avec l'âge du produit.
- Exiger que les produits soient testés selon une méthodologie convenue.

Exemple de norme:

Des normes volontaires, basées sur le marché, relatives à la toxicité des produits du tabac sans fumée existent. Par exemple, le Tableau 1.2 montre la norme GOTHIATEK® (utilisée par Swedish Match).

Tableau 1.2 : Niveaux maximaux admissibles de diverses toxines dans le « Snus » suédois avec une teneur en eau de 50 % [1].

Toxine	Limite
Nitrite	3,5 mg/kg
TSNAs totaux	5 mg/kg
N-nitroso diméthylamine	5 µg/kg
Benzo[a]pyrène	10 µg/kg
Cadmium	0,5 mg/kg
Plomb	1,0 mg/kg
Arsenic	0,25 mg/kg
Nickel	2,25 mg/kg
Chrome	1,5 mg/kg

1.2. Spectroscopie infrarouge : Principes et progrès

1.2.1. Fondements de la spectroscopie infrarouge

Initialement introduite au début du XXe siècle, la spectroscopie infrarouge (IR) est l'une des techniques spectroscopiques vibrationnelles les plus courantes et les plus largement utilisées disponibles pour les scientifiques travaillant dans une gamme complète de domaines. Le principe de cette technique réside dans

l'interaction du rayonnement IR avec la matière, en particulier les fréquences de vibration-rotation des liaisons chimiques au sein des molécules.

La région IR dans le spectre électromagnétique de la lumière est située entre la région visible et la région micro-onde. Étant donné la large gamme spectrale du rayonnement IR qui inclut des longueurs d'onde avec une grande différence d'énergie, cette région du spectre est divisée en trois autres sous-régions: le proche-infrarouge (NIR) entre 0,8 − 2,5 µm, le moyen-infrarouge (MIR) entre 2,5 − 25 µm et le lointain-infrarouge (FIR) entre 25 − 1000 µm [50, 51].

D'autre part, les molécules possèdent des modes vibrationnels avec une plage de fréquences entre 10^{13} et 10^{14} Hz environ. Lorsqu'elles sont exposées au rayonnement MIR, les molécules absorbent de l'énergie de manière dépendante de la fréquence, induisant des transitions vibrationnelles associées à des changements dans le moment dipolaire [52]. La quantité et les fréquences du rayonnement IR absorbé varient en fonction d'un certain nombre de facteurs, notamment le mode de vibration, le type de liaisons et la masse des atomes impliqués [50].

Les modes normaux de vibration impliquent deux types de mouvements vibrationnels cardinaux généralement nommés étirement (ou élongation) et déformation (ou flexion). L'étirement correspond à la variation de la distance interatomique qui peut être symétrique (ν_s) s'il se produit tout en préservant la symétrie moléculaire, ou asymétrique (ν_{as}) s'il y a perte d'un ou plusieurs éléments de symétrie avec allongement / raccourcissement différent de chaque liaison. Les vibrations de déformation correspondent à la variation de l'angle entre deux liaisons et peuvent se produire dans le plan, auquel cas les mouvements de cisaillement (scissoring, δ_s) et de basculement (rocking, ρ) peuvent être distingués, ou en dehors du plan, auquel cas les mouvements d'agitation (wagging, ω) et de torsion (twisting, τ) se produisent [53]. Les modes vibrationnels décrits sur l'exemple de la fonctionnalité $-XY_2$ sont présentés dans la Figure 1.6.

L'interprétation du spectre MIR implique la corrélation des pics avec les modes vibrationnels spécifiques des groupes fonctionnels présents dans l'échantillon. L'intensité des pics dans le spectre peut être corrélée à la concentration des groupes fonctionnels correspondants, permettant au MIR de

fournir des informations chimiques, structurales et compositionnelles quantitatives sur les molécules constitutives dans les phases gazeuse, liquide et solide en utilisant une variété de modes de mesure existants [54, 55].

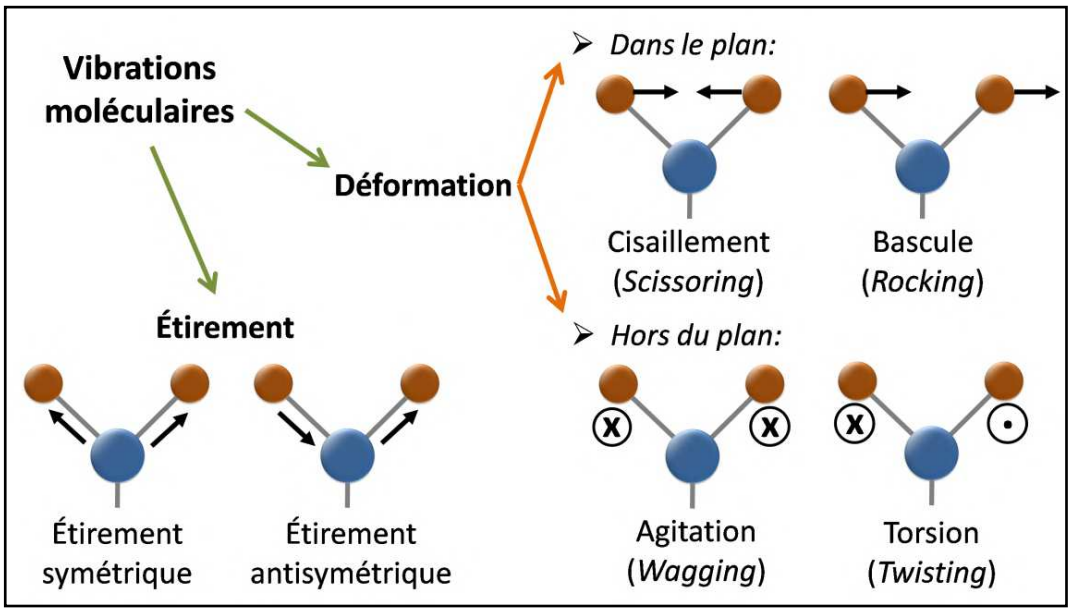


Figure 1.6 : Les modes normaux de vibration moléculaire.

La spectroscopie MIR trouve de nombreuses applications dans diverses disciplines scientifiques, notamment la chimie, la biotechnologie, la science des matériaux, la médecine, la pharmacie et la science de l'environnement. Notamment, l'IR est une technique non destructive pour la plupart des échantillons, ce qui en fait un outil polyvalent dans la recherche scientifique et les processus industriels [54].

1.2.2. Spectromètre infrarouge à transformée de Fourier à réflexion totale atténuée (ATR-FTIR)

En général, un spectromètre se compose de quatre parties principales: la source de rayonnement, les éléments de discrimination des longueurs d'onde, une interface d'échantillonnage et un détecteur [50]. Différentes sources de rayonnement et détecteurs, ainsi que différentes interfaces d'échantillonnage (ou de guides d'ondes) sont utilisés en spectroscopie MIR (Figure 1.7).

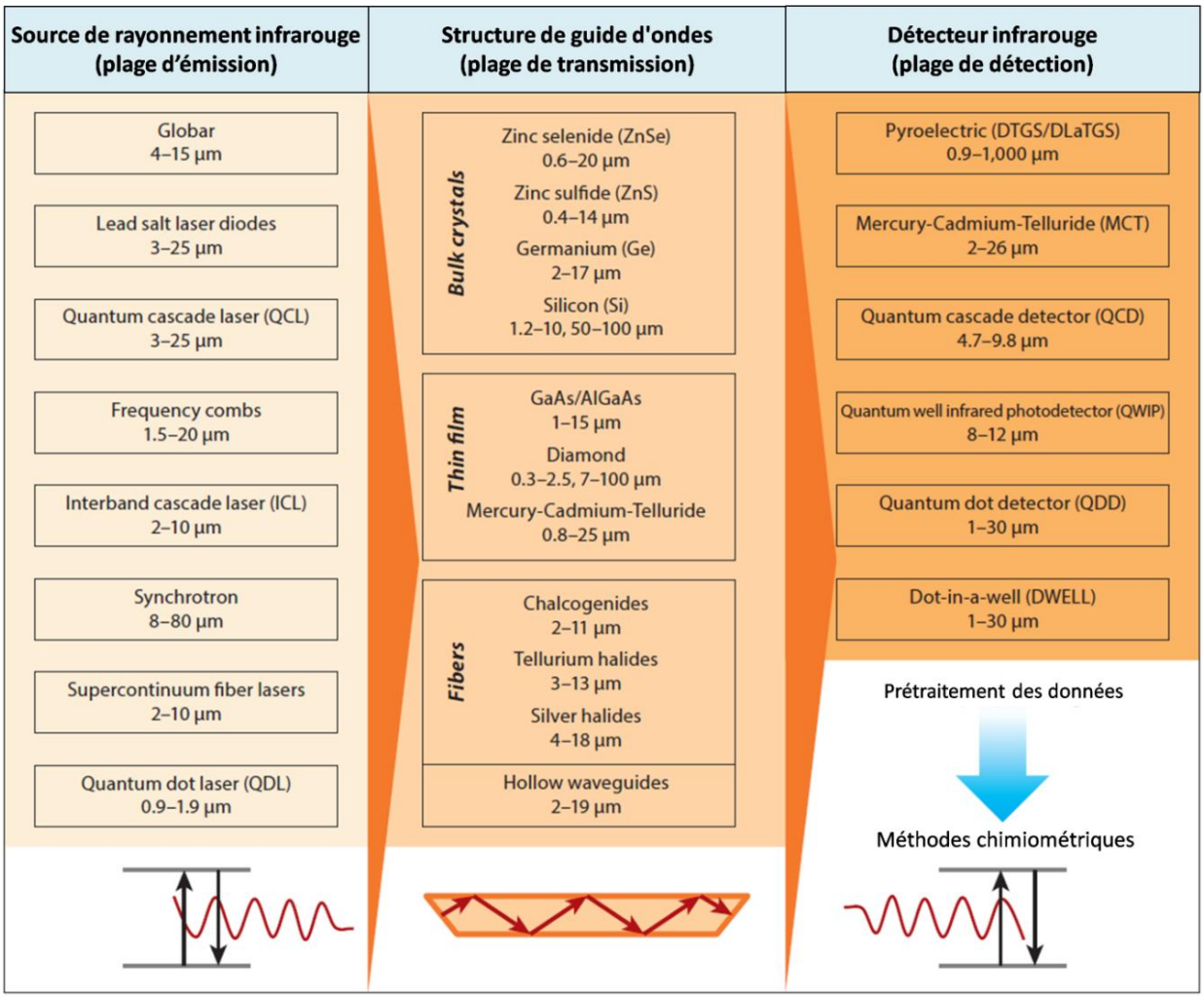


Figure 1.7 : Aperçu des sources lumineuses, des matériaux de guide d'ondes et des principes de détection les plus répandus dans la région spectrale de MIR [54].

La première version du spectrophotomètre IR était basée sur l'utilisation d'un dispositif dispersif appelé monochromateur. Ce dispositif sépare le spectre IR en bandes continues avec des longueurs d'onde définies. Par conséquent, l'échantillon testé est exposé séquentiellement au rayonnement IR avec des fréquences définies et son absorption est mesurée [52].

L'itération contemporaine de la spectroscopie infrarouge, connue sous le nom de spectroscopie infrarouge à transformée de Fourier (FTIR), a été développée dans les années 1960. Cette technique comprend une source de rayonnement infrarouge, un interféromètre, un détecteur et un système informatique. L'interféromètre, initialement conçu par Albert A. Michelson, intègre un diviseur de faisceau et une configuration constituée d'un miroir fixe et d'un

miroir mobile (Figure 1.8). Si un faisceau collimaté de rayonnement monochromatique avec une longueur d'onde "λ" est dirigé vers un diviseur de faisceau idéal, 50 % du rayonnement incident se réfléchit sur un miroir, et les 50 % restants se transmettent à l'autre. Lorsque le miroir mobile se déplace, les deux faisceaux IR identiques, ayant parcouru des distances différentes, se recombinent, générant un nouveau faisceau avec une longueur d'onde en évolution continue [52, 56]. Par conséquent, l'échantillon examiné subit une large gamme de fréquences IR au fil du temps. L'étape cruciale implique l'application d'une transformée de Fourier pour convertir l'interférogramme du domaine temporel au domaine fréquentiel, ce qui donne le spectre FTIR. Pour obtenir une haute précision du nombre d'ondes et, par conséquent, une haute résolution, un laser est utilisé comme référence interne pendant le balayage [53].

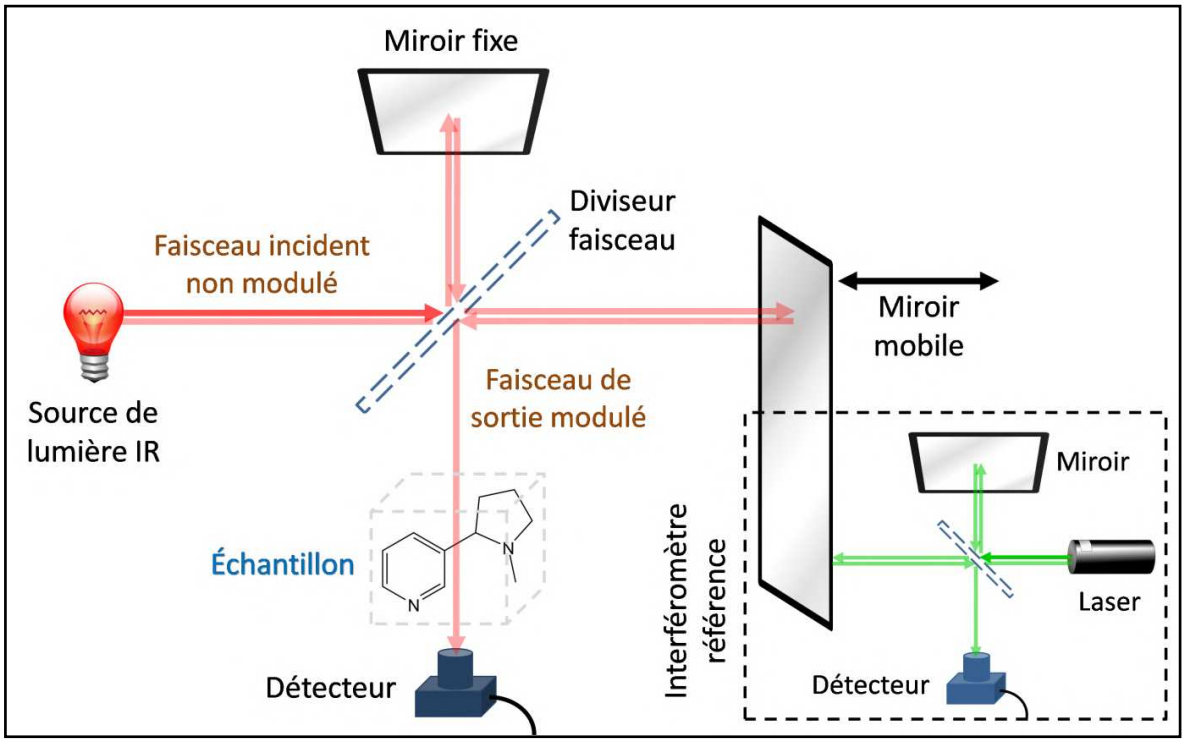


Figure 1.8 : Schéma d'un interféromètre de Michelson [56].

La spectroscopie FTIR a considérablement amélioré la qualité des spectres infrarouges en fournissant une haute résolution, un rapport signal / bruit élevé, une grande précision, ainsi qu'en minimisant le temps nécessaire pour obtenir des données [55].

La spectroscopie ATR-FTIR est une extension puissante de l'instrument FTIR traditionnel. Le principe fondamental de cette technique implique l'utilisation de la réflexion interne totale de la lumière infrarouge à l'interface entre un cristal de haut indice de réfraction et l'échantillon [56, 57]. Cela se produit uniquement lorsque l'angle du faisceau IR incident est supérieur à l'angle critique (θ_c), défini par la loi de Snell comme [57]:

$$\theta_c = \sin^{-1}\left(\frac{n_2}{n_1}\right) \qquad \text{(Éq. 1.1)}$$

où n_1 et n_2 désignent les indices de réfraction du cristal et de l'échantillon, respectivement. La Figure 1.9 montre qu'aucune réflexion ne se produit à des angles supérieurs à l'angle critique.

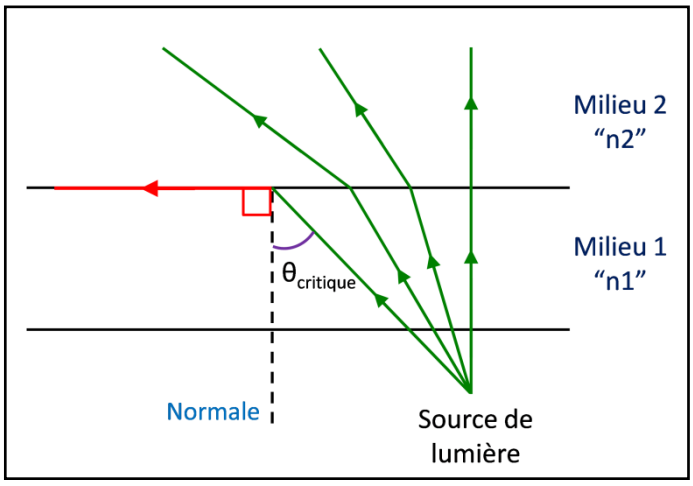


Figure 1.9 : Réfraction et angle critique à l'interface entre deux milieux.

Lorsque le faisceau IR pénètre dans le cristal, il subit, en fonction de l'instrument, une ou plusieurs réflexions internes, créant une onde évanescente qui s'étend dans l'échantillon pressé contre le cristal ATR. Cette onde évanescente interagit avec l'échantillon, perdant de l'énergie à la longueur d'onde où le matériau absorbe. Le rayonnement atténué résultant est mesuré et tracé en fonction de la longueur d'onde par le spectromètre, donnant lieu aux caractéristiques spectrales d'absorption de l'échantillon [56]. La configuration de base de l'ATR à mono-réflexion (en anglais « single bounce ATR ») est illustrée dans la Figure 1.10.

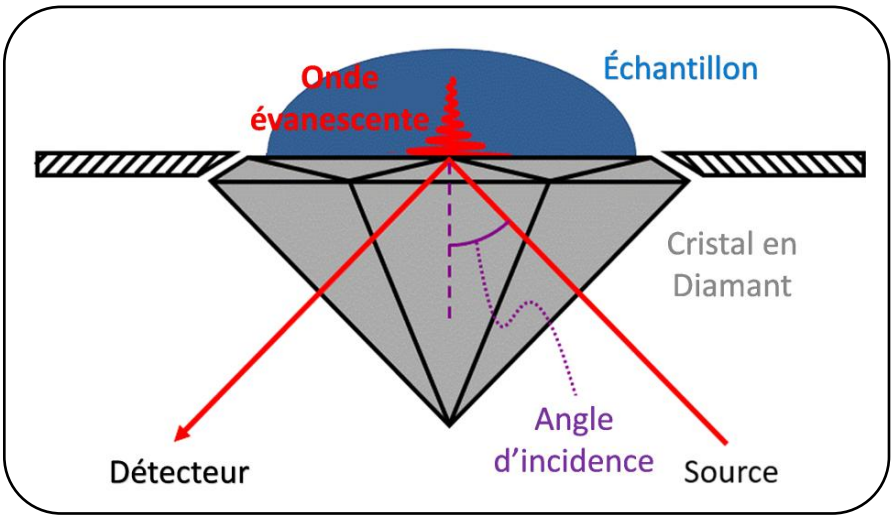


Figure 1.10 : Configuration d'une ATR simple à mono-réflexion [58].

La profondeur de pénétration (d_p) dans l'échantillon peut être ajustée en modifiant des paramètres tels que l'angle d'incidence (θ) et la longueur d'onde du rayonnement (λ), permettant l'analyse sélective des propriétés de surface ou en profondeur (cœur de l'échantillon) [53]. La formule pour "d_p" est donnée par [53]:

$$d_p = \frac{\lambda}{2\pi n_1 \sqrt{\sin^2\theta - \left(\frac{n_1}{n_2}\right)^2}} \qquad \text{(Éq. 1.2)}$$

En termes d'instrumentation, un accessoire ATR est intégré dans le spectromètre FTIR. Cet accessoire est généralement fabriqué à partir de matériaux ayant une faible solubilité dans l'eau, souvent du diamant, du germanium (Ge), du sélénium de zinc (ZnSe) ou du thallium–iodure (KRS-5) [52].

La technique ATR est particulièrement précieuse pour les échantillons difficiles à analyser dans leur forme native, pour de petites quantités d'échantillons, dans l'analyse de surface et ainsi dans la surveillance des réactions chimiques en temps réel.

1.3. Méthodes chimiométriques appliquées à la spectroscopie analytique

1.3.1. Généralités

L'amélioration de la technologie informatique associée à la spectroscopie a conduit à l'apparition des méthodes chimiométriques au début des années 1970. La chimiométrie est généralement définie comme "*la branche de la chimie qui utilise des méthodes mathématiques et statistiques avec la technologie informatique, conçoit et sélectionne les meilleures procédures de mesure et méthodes expérimentales, afin d'obtenir le maximum d'informations en interprétant les données chimiques*" [56, 59].

La principale application de la chimiométrie réside dans la possibilité de remplacer les méthodologies analytiques traditionnelles par des méthodes alternatives basées sur la combinaison de mesures optiques, électriques et autres mesures instrumentales. Cela éviterait l'utilisation de solvants toxiques, réduirait considérablement l'énergie, le coût et les déchets, diminuerait le temps d'analyse, et permettrait une détection non invasive, à distance et automatique. Ces principes sont, en termes généraux, en accord avec les nouvelles tendances vers une chimie analytique verte [60].

Une caractéristique supplémentaire de la combinaison spectroscopie / chimiométrie est la possibilité de mesurer les propriétés globales d'un échantillon, plutôt que de quantifier les analytes individuels. L'efficacité de ces méthodes repose sur la corrélation entre les concentrations des constituants de l'échantillon, générant le signal spectral, et leur impact sur la propriété globale mesurée. Les exemples incluent les propriétés organoleptiques ou les calories dans les aliments, la température de distillation dans les carburants, les propriétés rhéologiques de la farine, les paramètres de qualité des fibres textiles, etc [61].

Lors de l'utilisation de la chimiométrie, il est nécessaire de posséder une compréhension approfondie et une maîtrise du domaine concerné par le problème ou du contexte chimique pertinent. Malheureusement, il est probable que des branches entières de la chimiométrie constituent un monde inconnu pour la plupart des chimistes, en particulier pour les chimistes analystes [61].

Une variété de méthodes de calibration analytique est utilisée en conjonction avec la spectroscopie FT-MIR pour extraire des informations significatives de ses données complexes. Lorsque des mesures précises de concentration d'analytes spécifiques sont nécessaires, les méthodes des moindres carrés classiques (CLS), la régression des composantes principales (PCR) et la régression des moindres carrés partiels (PLSR) sont couramment appliquées. Cependant, si l'objectif est simplement d'identifier la présence ou l'absence d'espèces sans quantification précise, des méthodes de reconnaissance de motifs multivariées telles que l'analyse discriminante linéaire (LDA) ou la classification par machine à vecteurs de support (SVM-C) entrent en jeu. LDA et SVM-C sont connues comme des méthodes supervisées car des informations préalables sur les ensembles de données sont disponibles. Alternativement, des méthodes non supervisées telles que l'analyse en composantes principales (PCA) ou l'analyse de regroupement hiérarchique (AHC) peuvent regrouper et discerner des composants inconnus dans un ensemble de données sans aucune information préalable sur les données.

1.3.2. Loi de Beer-Lambert : Applications et limitations

La principale théorie de la quantification spectrale est basée sur la loi de Beer-Lambert. Elle décrit la relation linéaire entre la lumière absorbée (sur laquelle dépend l'intensité du signal IR) et la concentration de solution diluée, avec des absorptions négligeables des autres constituants de l'échantillon [50, 57]. En fait, comme la montre l'équation:

$$\log_{10}\left(\frac{I_0}{I}\right) = A = \varepsilon l C \qquad \text{(Éq. 1.3)}$$

où I_0 est l'intensité de la lumière incidente et I l'intensité de la lumière transmise. Ils donnent A, c'est-à-dire l'absorbance de l'échantillon; qui est égale à ε, le coefficient d'absorption molaire; l, la longueur du trajet optique; et C, la concentration de l'analyte.

Cependant, dans les applications pratiques, établir une relation fiable entre l'absorbance et la concentration de l'analyte, en particulier dans des échantillons réels caractérisés par une complexité inhérente (tels que les produits agricoles, le

pétrole, le tabac, etc.), s'avère souvent difficile en raison de l'absence d'un signal mesuré de manière sélective concernant l'analyte d'intérêt. Atteindre cette sélectivité peut impliquer des procédures complexes visant à libérer l'analyte des agents interférents potentiels qui pourraient être présents dans les échantillons testés. La Figure 1.11 met en évidence le biais introduit par les interférences dans les prédictions.

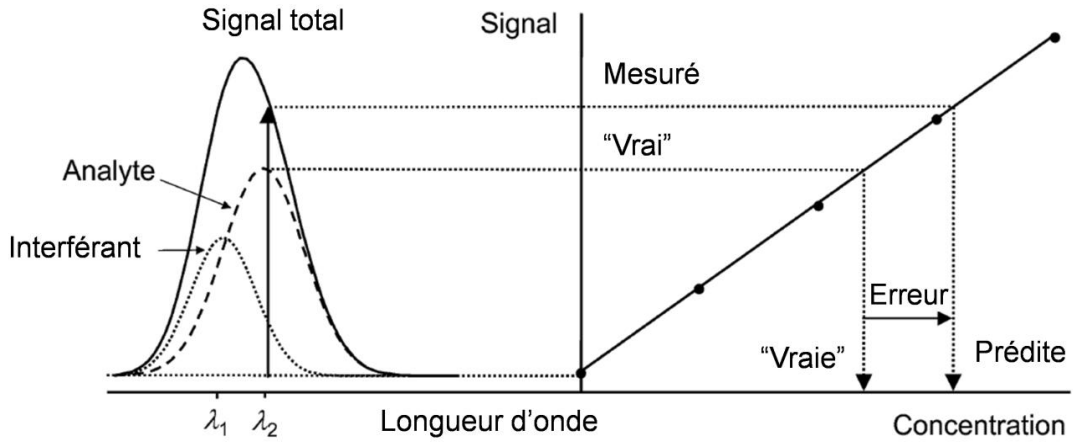


Figure 1.11 : Illustration de la manière dont un modèle univarié conduira à des prédictions biaisées lorsque des interférents non suspectées contribuent de manière variable au signal. En revanche, des mesures multivariées peuvent permettre une prédiction précise dans une telle situation [62].

En calibration multivariée, par contre, plusieurs points de données sont mesurés pour chaque échantillon. Les agents interférents, bien que possibles, ne nécessitent pas d'être éliminés ou modifiés comme dans le cas de la calibration univariée. Au lieu de cela, des modèles mathématiques sont utilisés pour compenser de manière appropriée leurs effets dans le contexte multivarié [61].

1.3.3. Méthodes de régression et de reconnaissance de motifs

1.3.3.1. Régression linéaire univariée

La régression linéaire univariée (ou monovariée) est la plus simple des régressions linéaires avec la formule:

$$y = a + bx + e \qquad \text{(Éq. 1.4)}$$

où x est une variable explicative ou le prédicteur (par exemple, l'absorbance), y est une variable à expliquer ou la réponse (par exemple, la concentration), a est l'interception à l'origine, b est la pente (sensibilité), et e est l'erreur de mesure.

Dans l'analyse de régression, l'objectif principal consiste à trouver les meilleures estimations des paramètres a et b en se basant sur un ensemble de I valeurs mesurées (x_i, y_i). Ces estimations permettent d'approcher au plus près les valeurs réelles y par les valeurs prédites $\hat{y}$. Les valeurs estimées de a et b sont souvent obtenues par la méthode des moindres carrés, qui minimise la somme des carrés des erreurs. Par la suite, ces valeurs peuvent être utilisées pour effectuer des analyses prédictives [59].

1.3.3.2. Moindres carrés partiels

La régression par moindres carrés partiels (ou « Partial Least Squares Regression, 'PLSR' » en anglais) est une technique mathématique utilisée pour modéliser les relations dans des données multidimensionnelles et/ou colinéaires. Elle décompose les prédicteurs (spectres, matrice X) et les réponses (mesures de référence, matrice Y) en « scores » et en « loadings » [59], respectivement dans les équations Éq. 1.5 et Éq. 1.6:

$$X = TP^T + E_X \quad \text{(Éq. 1.5)}$$

$$Y = UQ^T + E_Y \quad \text{(Éq. 1.6)}$$

où E_X et E_Y sont les matrices résiduelles de X et Y, respectivement.

De manière itérative, un modèle prédictif est créé en maximisant la covariance entre T et Y, ce qui donne des vecteurs de poids (W) qui sont utilisés pour calculer les coefficients de régression (B) utilisés pour prédire la variable de réponse pour de nouvelles données [59, 61]:

$$B = W(P^TW)^{-1}Q^T \quad \text{(Éq. 1.7)}$$

L'estimation de P, Q, T et W peut également être obtenue en extrayant les vecteurs propres de la plus petite taille des produits de X, X^T, Y et Y^T dans l'algorithme de base "Kernel PLS" [63].

La PLS est principalement une méthode de régression, mais elle peut également être utilisée pour l'analyse discriminante, auquel cas elle est appelée analyse discriminante par moindres carrés partiels (PLS-DA).

1.3.3.3. Machine à vecteurs de support

La régression par machine à vecteurs de support (ou « Support Vector Machine Regression, 'SVMR' », en anglais) est une méthode d'apprentissage automatique adaptée de l'algorithme SVM qui était principalement utilisé pour des tâches de classification. Cette méthode part du principe qu'au lieu de tracer une ligne exacte qui passe par tous les points de données, elle cherche à trouver une fonction optimale *f (x,w)* qui passe au plus près de la majorité des points, tout en tolérant une petite marge d'erreur (*ε*). Contrairement aux techniques de régression conventionnelles, la SVMR peut modéliser des relations plus complexes, même non linéaires, en utilisant des fonctions mathématiques appelées 'noyaux' (en anglais « kernel functions », $\phi(x_i, x_i^{'})$) [64]. En d'autres termes, elle permet à la machine d'apprendre et de comprendre des modèles de données plus subtils afin d'améliorer les prédictions. Une façon de garantir cela est de résoudre la formulation suivante [64, 65]:

$$min_{w,b,\xi,\xi^*} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{I} (\xi_i + \xi_i^*) \qquad \text{(Éq. 1.8)}$$

Sujet à:

$$\begin{cases} y_i - w^T \phi(x_i, x_i^{'}) - b \leq \varepsilon + \xi_i \\ w^T \phi(x_i, x_i^{'}) + b - y_i \leq \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 0 \; ; \forall i, i \in (1,2, \dots, I) \end{cases} \qquad \text{(Éq. 1.9)}$$

où *C* est un paramètre de régularisation qui détermine le compromis entre la complexité du modèle (aplatissement) et le degré auquel les écarts supérieurs à *ε* sont tolérés, *w* est le vecteur de poids, *b* est le terme de biais, ξ_i et ξ_i^* sont des variables d'écart qui contrastent la frontière symétrique produite par la fonction de perte "marge dure" (voir la Figure 1.12A). Le problème d'optimisation peut être résolu à l'aide de la méthode de dualisation standard utilisant des multiplicateurs de Lagrange. La solution fournit les poids et le biais optimaux pour les vecteurs de support [66].

La différence principale entre la régression par machine à vecteurs de support (SVMR) et la classification par machine à vecteurs de support (SVM-C) réside dans leurs tâches et objectifs fondamentaux. Alors que la SVMR tente de trouver un hyperplan qui minimise l'erreur de prédiction tout en maintenant une certaine marge, la SVM-C cherche à identifier un hyperplan dans l'espace des caractéristiques qui maximise la marge entre les classes distinctes, facilitant ainsi une séparation efficace des données en différentes catégories [67] (Figure 1.12B).

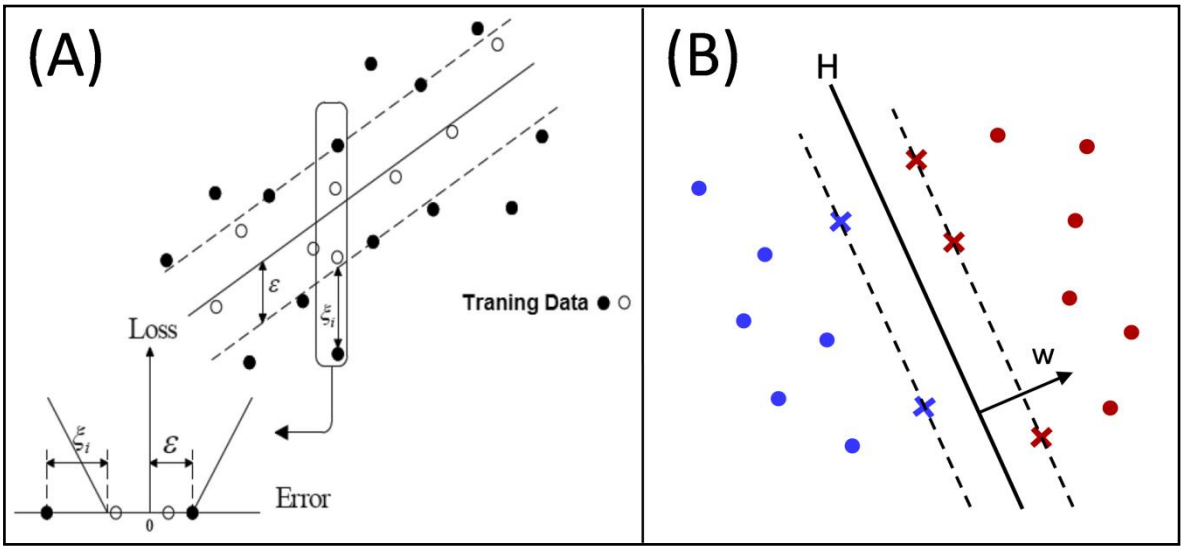


Figure 1.12 : (A) Optimisation de la marge d'erreur pour une SVMR linéaire [68], (B) Représentation d'hyperplan SVM-C à travers deux classes linéairement séparables. Les croix sur les lignes pointillées (marge maximale) sont des vecteurs de support.

1.3.3.4. Analyse en composantes principales

L'analyse en composantes principales (ou « Principal Component Analysis, 'PCA' » en anglais) est une méthode de reconnaissance de motifs non supervisée utilisée pour réduire la dimensionnalité, explorer les données, visualiser les corrélations et détecter les valeurs aberrantes [69]. Elle est particulièrement utile lorsqu'il s'agit de traiter de grands ensembles de données quantitatives, permettant une interprétation tout en minimisant la perte d'information et en traitant les problèmes de colinéarité des données [70]. La PCA génère un nouvel ensemble d'axes principaux orthogonaux, qui sont des combinaisons linéaires indépendantes des variables d'origine. La première composante principale (PC-1)

forme un seul axe dans l'espace, capturant la variance maximale possible. La deuxième composante (PC-2), perpendiculaire à la première, capture la variance suivante la plus élevée, et ainsi de suite [69, 71, 72]. La Figure 1.13 représente un exemple des scores PCA des deux premières PCs calculées pour un ensemble de composés aléatoires (A à L).

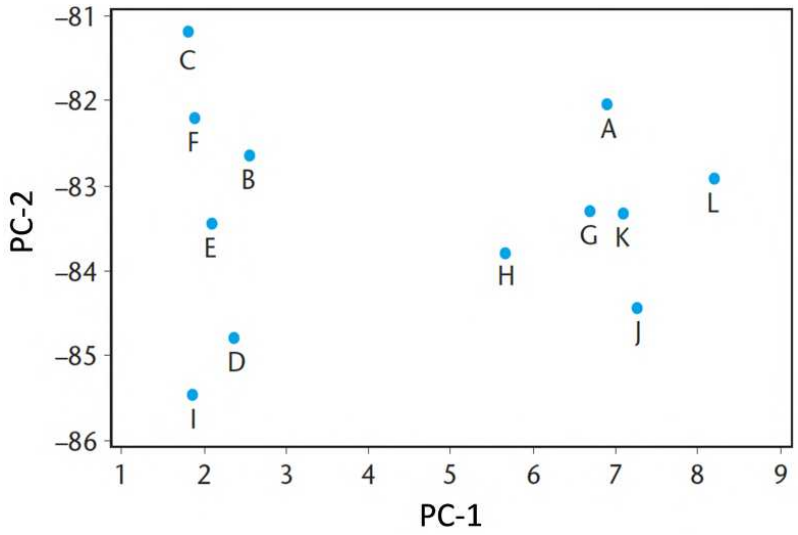


Figure 1.13 : Scores des PCs pour les données d'exemple. Le diagramme révèle que les composés se regroupent en deux clusters distincts.

1.3.3.5. Classification hiérarchique ascendante

Ou analyse de regroupement hiérarchique (ou bien « Agglomerative Hierarchical Clustering, 'AHC' » en anglais), est une technique de partitionnement consistant à construire un arbre de regroupement binaire (dendrogramme). Elle commence par calculer la dissimilarité entre les *I* échantillons stockées aux feuilles (singleton), puis procède en fusionnant deux à deux les sous-ensembles "les plus proches" (stockés aux nœuds) en fonction d'une mesure de dissimilarité entre les éléments et/ou les classes à travers une minimisation d'un critère d'agrégation donné [73-75]. Ce processus se poursuit jusqu'à ce que la racine de l'arbre, englobant toutes les échantillons, soit atteinte. La Figure 1.14 montre un exemple d'application du regroupement de données par la méthode des plus proches voisins (« single linkage method »).

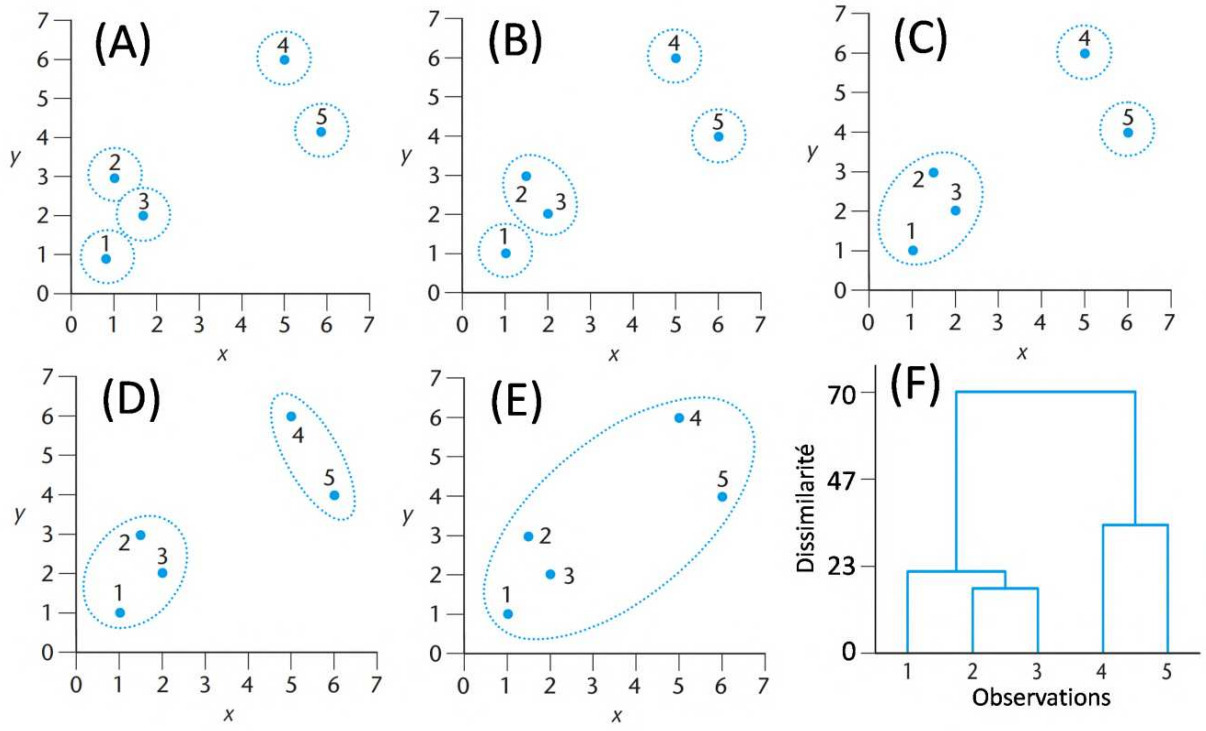


Figure 1.14 : Regroupement de données par AHC. (A à E) Groupes formés à chaque étape, délimités par des lignes pointillées, (F) Dendrogramme représentant la hiérarchie des groupes [76].

1.3.3.6. Classification k-*means*

Ou partitionnement en *k*-moyennes (« *k*-means clustering » en anglais) est une technique d'apprentissage automatique non supervisée utilisée pour regrouper des échantillons similaires en *k* clusters (ou groupes) distincts. Le processus commence par la sélection de *k* points de données comme centres de clusters initiaux (centroïdes). Ensuite, chaque point est assigné au centroïde le plus proche en fonction de la distance euclidienne. Ensuite, la position des centroïdes est mise à jour en calculant les nouvelles moyennes de clusters après chaque perte ou gain de points. Les étapes d'assignation et de mise à jour des centroïdes sont répétées jusqu'à la convergence, où aucun point ne change de cluster [76, 77]. Cette technique présente l'avantage que l'assignation des éléments est réversible d'une itération à l'autre, ce qui serait impossible avec l'AHC.

1.3.4. Validation du modèle

La robustesse d'un modèle, indiquant sa résilience aux influences externes, est un aspect critique de son évaluation de performance. Pour garantir la fiabilité du modèle, la validation est impérative, et cela peut être réalisé par des moyens internes ou externes, comprenant des ensembles de test (ou de prédiction). L'utilisation des approches de validation offre une évaluation complète de l'adaptabilité et des capacités prédictives du modèle, renforçant ainsi la confiance dans son efficacité globale.

1.3.4.1. Validation croisée

La validation croisée (« Cross-Validation, 'CV' » en anglais) est une méthode utilisée pour évaluer la fiabilité d'un modèle à travers une technique d'échantillonnage systématique. Dans ce processus, l'ensemble de données est divisé en k parties égales, appelées 'plis', chacune contenant un nombre identique d'échantillons. Initialement, un pli sert de jeu de validation, tandis que les $k - 1$ autres plis sont combinés pour former le jeu d'entraînement (calibration). Le modèle est entraîné sur cet ensemble, et sa performance est évaluée sur le jeu de validation à l'aide de métriques telles que l'erreur quadratique moyenne (ou « Root-Mean-Square Error, 'RMSECV' » en anglais) ou la somme des carrés des erreurs résiduelles prédites (ou « predicted residual error sum of squares, 'PRESS' » en anglais). Ce cycle est répété k fois, avec un pli différent agissant comme jeu de validation à chaque itération. Pour chaque configuration de variables latentes, la RMSECV ou PRESS moyennes sur tous les jeux de validation est calculée [53].

La technique de CV aide à éviter le surapprentissage (sur-ajustement ou « overfitting » en anglais), qui se produit lorsqu'un modèle mémorise trop étroitement les données d'entraînement mais ne se généralise pas bien aux nouveaux échantillons. De plus, elle évite les problèmes causés par la division des données en seulement deux ensembles, car elle garantit que tous les échantillons sont utilisés à la fois pour l'entraînement et pour le test. Le « leave-one-out CV (LOOCV) » ou « Full CV » est une instance spécifique où un seul échantillon est laissé de côté lors des itérations [53].

1.3.4.2. Validation par un ensemble de test indépendant

Habituellement, un ensemble d'échantillons connus et indépendants est utilisé pour valider l'exactitude, la stabilité, la robustesse et la transférabilité du modèle construit. Les échantillons de l'ensemble de test doivent contenir tous les composants contenus dans l'échantillon à prédire, dont la plage de concentration devrait couvrir au moins 95 % de celle de l'ensemble de calibration, avec une distribution uniforme. De plus, le nombre d'échantillons dans cet ensemble doit être suffisant pour les tests statistiques, généralement pas moins de 28 échantillons sont requis [59].

1.3.5. Activités de pré-calibration

Dans le scénario le plus simple, l'analyste commence par collecter un nombre adéquat d'échantillons représentatifs pour les ensembles de calibration (étalonnage) et de validation. Le modèle est construit en utilisant la matrice de calibration, incorporant les spectres et leurs mesures de référence correspondantes (cible). Ensuite, les performances de prédiction sont évaluées par rapport aux échantillons de validation. Une fois jugé fiable, le modèle est appliqué pour prédire l'analyte dans de futurs échantillons inconnus [59].

Dans un contexte plus large, plusieurs étapes précèdent généralement la construction du modèle. Celles-ci impliquent de déterminer (i) les échantillons particuliers désignés pour les ensembles de calibration et de validation, (ii) les variables spectrales (longueurs d'onde) à utiliser dans la modélisation, et (iii) si les spectres bruts ou prétraités subiront la construction du modèle. Bien que ces composantes − échantillons, variables spectrales et prétraitement − soient distinctes, elles sont étroitement interconnectées au sein de cette triade [61].

1.3.5.1. Sélection des échantillons

Lorsqu'il s'agit de grands ensembles de données où la composition des échantillons ne peut pas être contrôlée, le choix des échantillons pour la calibration et la validation du modèle devient crucial. Une approche courante implique une division aléatoire de l'ensemble des échantillons, allouant, par

exemple, 70% pour la calibration et 30% pour la validation. Cependant, cette sélection aléatoire présente un risque potentiel, car elle peut exclure des échantillons situés loin du centre d'étalonnage ou correspondre à des régions moins peuplées de l'espace des échantillons [61].

Pour pallier ces limitations, plusieurs algorithmes de sélection d'échantillons dédiés ont été développés. Le plus populaire est l'algorithme de Kennard-Stone, qui utilise la métrique de distance euclidienne pour répartir uniformément les échantillons sélectionnés dans l'espace des caractéristiques. Notamment, l'algorithme Kennard-Stone peut traiter soit les données spectrales, soit les scores des composantes principales issus d'une analyse PCA comme variable caractéristique. Le processus peut être résumé comme suit [61]:

i. Le premier échantillon sélectionné est celui qui est le plus proche du centre de l'espace des échantillons.
ii. Le deuxième échantillon sélectionné est le plus éloigné du premier.
iii. Pour les sélections suivantes, les distances des échantillons restants par rapport à ceux déjà sélectionnés sont d'abord calculées. Ensuite, les distances minimales sont considérées, et l'échantillon choisi est celui avec la plus grande des distances minimales.
iv. Le processus se poursuit jusqu'à un certain nombre prédéfini d'échantillons de calibration, par exemple 70% du nombre total. Les autres sont laissés pour la validation.

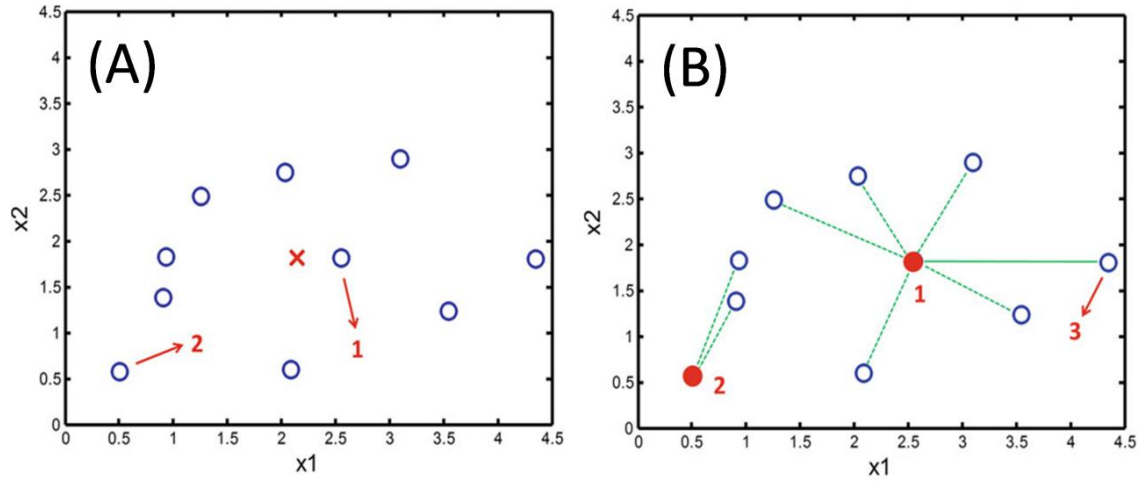


Figure 1.15 : L'algorithme de Kennard-Stone en action (la croix rouge dans "A" est le centre d'étalonnage et la ligne verte continue dans "B" est le maximum des huit distances minimales) [61].

Échantillons aberrants:

Avant de procéder au processus de sélection des échantillons de calibration, il est crucial d'identifier et d'exclure les valeurs aberrantes (ou « outliers » en anglais) − les échantillons suspects de ne pas appartenir à la population cible. Les valeurs aberrantes peuvent contenir des composants chimiques anormaux ou présenter des concentrations extrêmes, s'écartant significativement du reste des échantillons. Leur inclusion dans le développement du modèle peut compromettre la précision et la robustesse du modèle [59, 61].

L'identification des valeurs aberrantes sert à deux objectifs principaux dans l'analyse spectrale. Premièrement, cela garantit que le modèle apprend à partir des données les plus représentatives lors du développement, améliorant ainsi sa capacité de généralisation. Un modèle établi sur des données plus propres sera plus performant pour prédire des échantillons inconnus. Deuxièmement, elle assure que les échantillons à tester se situent dans le domaine d'application du modèle, conduisant à des prédictions plus fiables [59].

La littérature présente divers tests statistiques et indicateurs pour détecter les valeurs aberrantes. Dans notre étude, les Q- ou F-résidus, la statistique T^2 de Hotelling et d'autres tests sont discutés de manière approfondie avec des exemples illustratifs, ce qui constitue la première exploration de ce type dans la section expérimentale.

1.3.5.2. Sélection des variables spectrales

Les données spectrales contiennent souvent un grand nombre de longueurs d'onde (variables); cependant, toutes ne contribuent pas nécessairement de manière égale à la prédiction de l'analyte ou de la propriété d'intérêt. La sélection des longueurs d'onde pertinentes peut simplifier le modèle, réduire le bruit excessif, renforcer l'interprétabilité du modèle et potentiellement accélérer le calcul. Plus important encore, lorsque les variables non corrélées ou non linéaires sont éliminées, il devient possible d'obtenir un modèle d'étalonnage doté d'une forte capacité prédictive et d'une robustesse améliorée [59].

De nombreuses méthodes différentes existent pour la sélection des variables spectrales, chacune offrant une approche distincte, et de nouvelles techniques sont constamment introduites. Certaines méthodes privilégient des régions spectrales spécifiques en fonction de critères statistiques, tandis que d'autres utilisent des algorithmes d'apprentissage automatique pour la sélection des caractéristiques. Le choix de la méthode dépend souvent des caractéristiques de l'ensemble de données, en tenant compte des principes, de l'applicabilité, des avantages et des inconvénients de chaque méthode, ainsi que de l'alignement avec les objectifs spécifiques de l'analyse [55].

Dans la suite, nous en avons exploré trois approches différentes à savoir: la sélection basée sur le vecteur des coefficients de régression, l'interval-PLS et la méthode appelée importance des variables en projection.

1.3.5.2.a) Coefficients de régression:

La base de la sélection des longueurs d'onde actives à l'aide du vecteur des coefficients de régression se trouve dans l'expression de prédiction de la concentration de l'analyte y_n [61]:

$$y_n = b_n^T x_n = b_{1n} x_1 + b_{2n} x_2 + \cdots + b_{jn} x_j \qquad \text{(Éq. 1.10)}$$

Si les éléments du vecteur b_n sont dans une certaine région spectrale proche de zéro et/ou présentent une grande incertitude, les termes dans l'équation (Éq. 1.10) correspondant à ces éléments seront très petits et/ou présenteront une incertitude significative par rapport aux termes restants (Figure 1.16) [61].

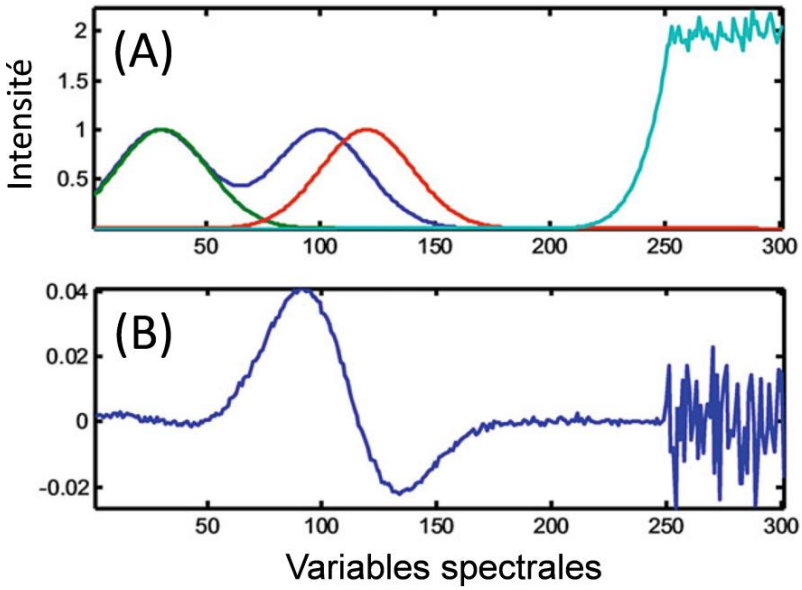


Figure 1.16 : (A) Système analytique idéal avec un analyte d'intérêt (spectre bleu), deux constituants supplémentaires (spectres rouge et vert), et une ligne de base saturant le détecteur dans la plage des variables de 250 à 300 (spectre bleu clair). (B) Vecteur des coefficients de régression PLS pour l'analyte d'intérêt dans ce système [61].

1.3.5.2.b) <u>Interval-PLS:</u>

La régression des moindres carrés partiels par intervalle (ou « interval-PLS », '*i*-PLS') consiste à diviser la gamme spectrale complète en sous-régions prédéfinies ou intervalles d'une largeur spécifiée. À l'intérieur de chacun de ces intervalles, une régression PLS distincte est réalisée en utilisant le nombre optimal de variables latentes. Ensuite, la RMSECV moyenne est calculée pour chaque intervalle, et les intervalles présentant les valeurs de RMSECV les plus faibles sont considérés comme les plus informatifs et sont recommandés pour la construction du modèle final [61].

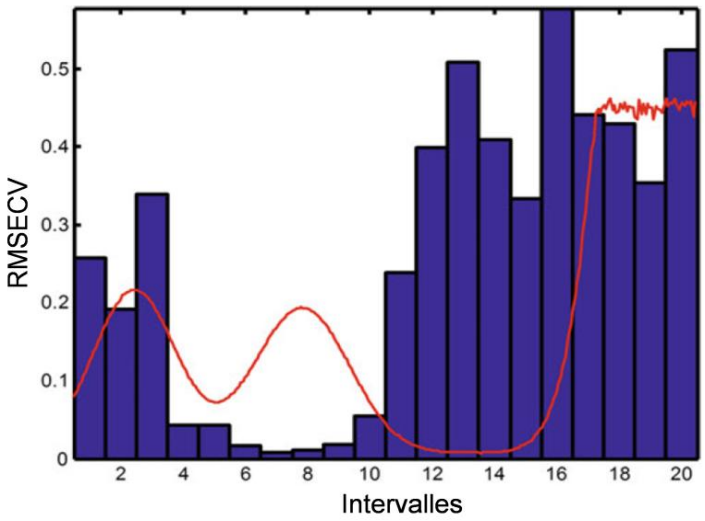


Figure 1.17 : Diagramme en barres montrant la RMSECV pour le système de la Figure 1.16 en utilisant la méthode *i*-PLS avec des intervalles de 15 variables. La ligne rouge montre le spectre de calibration moyen [61].

1.3.5.2.c) <u>Importance des variables en projection:</u>

L'importance des variables en projection (ou « Variable Importance in Projection », 'VIP') est une technique utilisée en PLSR pour évaluer l'importance relative de chaque variable spectrale dans l'explication de la cible. Les scores VIP élevés indiquent que la longueur d'onde explique fortement les changements de concentration et joue un rôle clé dans la construction du modèle. Généralement, les variables spectrales avec des scores VIP dépassant un seuil (souvent 1 ou ajusté en fonction des connaissances du domaine) sont choisies comme caractéristiques importantes, soulignant leur contribution significative à la fois à l'ajustement du modèle et à l'interprétabilité [59, 78].

Mathématiquement, le score VIP pour la *j*-ème longueur d'onde est calculé comme suit:

$$VIP_j = \sqrt{\frac{m \sum_{k=1}^{h} \left(q_k^2 t_k^T t_k \left(w_{j,k} / \|w_k\| \right)^2 \right)}{\sum_{k=1}^{h} q_k^2 t_k^T t_k}} \qquad \text{(Éq. 1.11)}$$

où m est le nombre total de longueurs d'onde, h est le nombre optimal de variables latentes, w représente le vecteur de poids, t est le vecteur de score, et

q est le vecteur de charge (« load vector » en anglais) obtenus à partir du modèle PLSR.

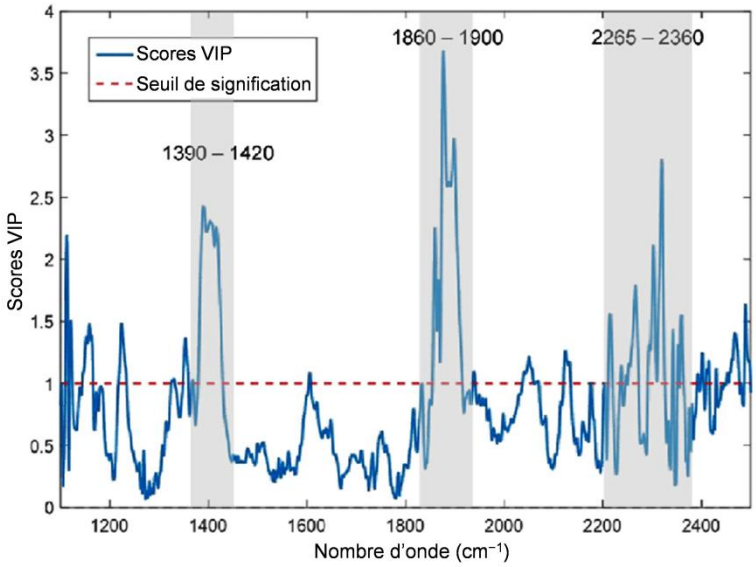


Figure 1.18 : VIP pour la sélection des bandes caractéristiques (scores VIP > 1, les régions importantes sont marquées en gris) [79].

1.3.5.3. Prétraitement mathématique des spectres

Les mesures de réflectance, obtenues par des techniques telles que l'ATR-FTIR, capturent souvent des signaux de fond indésirables qui présentent une variabilité entre les échantillons et leurs répliques, ne montrant aucune corrélation directe avec la concentration de l'analyte ou les propriétés chimiques de l'échantillon. Ces artefacts tombent généralement dans cinq catégories: données manquantes, bruit, décalages de ligne de base, effets multiplicatifs et décalages de pic. Les origines de ces artefacts peuvent être attribuées à des facteurs tels que les effets de diffusion, l'état de l'échantillon, le bruit électrique, les changements de réponse instrumentale au fil du temps et d'autres influences physiques, chimiques et environnementales externes [55].

Pour relever ce défi, un ensemble de procédures regroupées sous le terme de prétraitement mathématique a été développé. L'objectif est de réduire systématiquement, et idéalement d'éliminer, l'influence de ces artefacts, améliorant ainsi la précision et l'interprétabilité des informations chimiques extraites des spectres.

1.3.5.3.a) Centrage sur la moyenne:

Le centrage à l'aide de la valeur moyenne, connu sous le nom de centrage sur la moyenne (ou « mean centering » en anglais), permet l'interprétation des résultats en termes de variation autour de la moyenne. Cette méthode est souvent privilégiée car elle met en évidence les différences entre les observations plutôt que leurs valeurs absolues [63].

Lorsque X est une matrice spectrale avec n lignes (échantillons) et m colonnes (longueurs d'onde) et $x_{i,j}$ est la valeur de la longueur d'onde j pour l'échantillon i, le spectre centré sur la moyenne est calculé comme suit [80]:

$$x_{i,j}^{(Centr\acute{e})} = x_{i,j} - \bar{x}_j \qquad \text{(Éq. 1.12)}$$

où $\bar{x}_j$ est la valeur moyenne de chaque colonne.

La Figure 1.19 montre l'effet du centrage sur un ensemble de données comprenant un signal de fond constant superposé aux signaux de trois analytes.

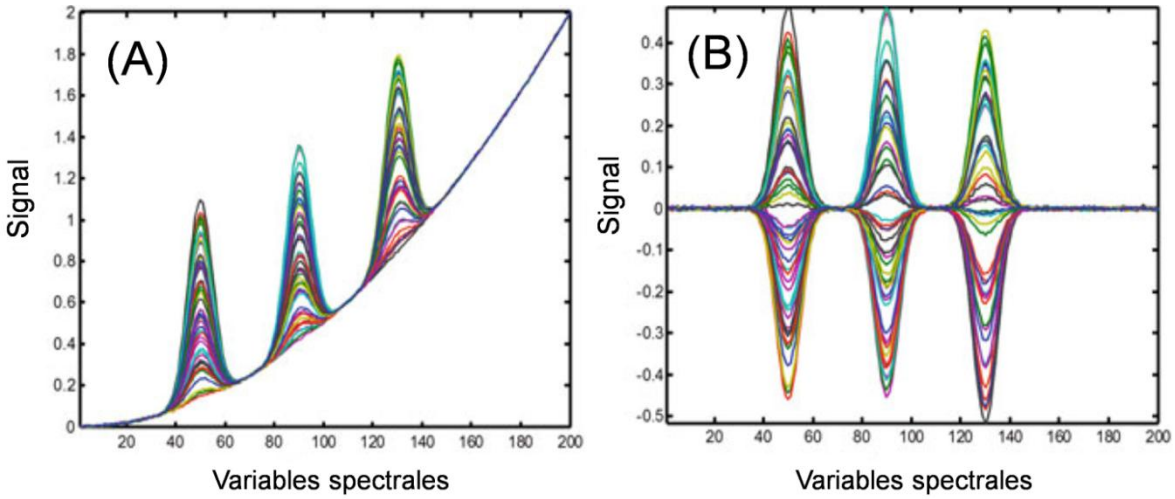


Figure 1.19 : (A) Spectres de calibration de 50 échantillons contenant trois analytes et un signal de fond constant. (B) Spectres centrés sur la moyenne [61].

1.3.5.3.b) Auto-scaling:

L' « auto-scaling », appelé mise à l'échelle de variance unitaire en français, est une technique de prétraitement des données très répandue pour la mise à l'échelle des données. Cette méthode de standardisation utilise l'écart-type comme facteur d'échelle, garantissant que les variables sont ramenées à une

échelle approximativement identique. Mathématiquement, cela peut être décrit comme suit [63, 80]:

$$x_{i,j}^{(Auto-scal\acute{e})} = \frac{x_{i,j\ (Centr\ \acute{e})}}{SD_j} = \frac{x_{i,j} - \bar{x}_j}{\sqrt{\frac{\sum_{i=1}^{n}\left(x_{i,j} - \bar{x}_j\right)^2}{n-1}}} \qquad \text{(Éq. 1.13)}$$

où SD_j représente la valeur de l'écart-type pour chaque colonne.

Les spectres auto-scalés ont une moyenne de colonne de 0 et une variance de 1. Cela est particulièrement utile pour la modélisation des composants à faible concentration car il donne le même poids à toutes les longueurs d'onde dans les spectres [59].

1.3.5.3.c) <u>Corrections de ligne de base:</u>

La correction du décalage de la ligne de base et la correction linéaire de la ligne de base sont deux transformations fondamentales utilisées pour ajuster la ligne de base des échantillons, et elles peuvent être appliquées soit séparément, soit en combinaison [81]. Dans le décalage de la ligne de base, la valeur du point le plus bas dans le spectre est soustraite à toutes les variables pour chaque échantillon, comme exprimé par l'équation suivante [63]:

$$x_i^{(Offset)} = x_i - \min_{1 \le i \le n}(x_i) \qquad \text{(Éq. 1.14)}$$

Cette opération fixe la valeur minimale à 0, ce qui entraîne des valeurs positives pour le reste.

Pour la correction linéaire de la ligne de base, la méthode implique de sélectionner deux variables qui définissent la nouvelle ligne de base, de les fixer toutes les deux à 0, et de transformer les variables restantes en conséquence par interpolation / extrapolation linéaire [63].

Il est crucial pour les deux méthodes de s'assurer que le point le plus bas correspond à la même variable pour tous les échantillons.

1.3.5.3.d) <u>Lissage et Dérivées:</u>

Le lissage spectral et les dérivées appartiennent à la catégorie du prétraitement mathématique utilisant la stratégie de fenêtre mobile (une petite région spectrale au début du spectre). L'outil le plus utilisé pour le lissage et les dérivées est le filtre de Savitzky-Golay. La philosophie de base de Savitzky-Golay est simple: dans une fenêtre mobile d'une largeur prédéfinie, les valeurs du signal sont ajustées à une fonction polynomiale, estimant dans chaque cas les coefficients des termes polynomiaux. Ensuite, le signal correspondant au point central de la fenêtre est estimé par le polynôme ajusté (lissage), ou par les dérivées polynomiales. La fenêtre est ensuite déplacée à travers la plage spectrale et le processus est répété [61, 82].

Par exemple, supposons qu'un polynôme du troisième degré soit appliqué, avec une fenêtre mobile de cinq variables, les signaux sont ajustés à la fonction suivante:

$$y = ax^3 + bx^2 + cx + d \qquad \text{(Éq. 1.15)}$$

où a, b, c et d sont des paramètres ajustables, x représente chacun des cinq variables, et y le signal à un variable spectral donné. Avec les paramètres ajustés, la valeur de y est estimée au centre de la fenêtre; ce sera la première valeur du spectre lissé (Figure 1.20).

Ayant les paramètres polynomiaux ajustés, on peut également estimer les dérivées. Dans chaque fenêtre, les valeurs suivantes pour la première et la deuxième dérivée sont estimées à partir de l'équation (Éq. 1.15):

$$\frac{dy}{dx} = 3ax^2 + 2bx + c \quad (1^{\text{ère}} \text{ dérivée}) \qquad \text{(Éq. 1.16)}$$

$$\frac{d^2y}{dx^2} = 6ax + 2b \quad (2^{\text{ème}} \text{ dérivée}) \qquad \text{(Éq. 1.17)}$$

Les spectres dérivés peuvent éliminer efficacement les interférences de la ligne de base et d'autres arrière-plans pour distinguer les pics superposés, et améliorer la résolution et la sensibilité (Figure 1.21). Cependant, cela introduit également du bruit et réduit le rapport signal / bruit. Les 1ère et 2ème dérivées sont plus courantes en pratique que les dérivées d'ordre supérieur [63, 80].

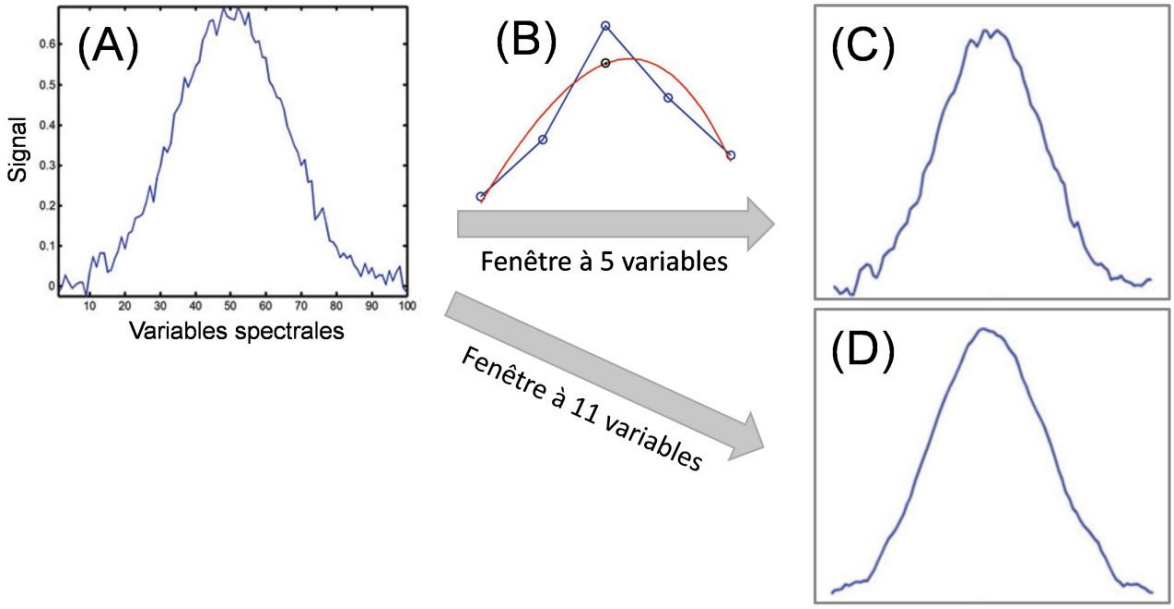


Figure 1.20 : (A) Spectre avec un niveau élevé de bruit aléatoire ; (B) Application du lissage Savitzky-Golay avec première fenêtre de 5 variables (ligne bleue et cercles), polynôme ajusté du troisième degré (ligne rouge) et valeur au point central estimée par cet ajustement (cercle noir) ; (C) Effet produit par le filtre en utilisant une fenêtre de 5 variables ; et (D) Spectre résultant après l'utilisation d'une fenêtre de 11 variables [61].

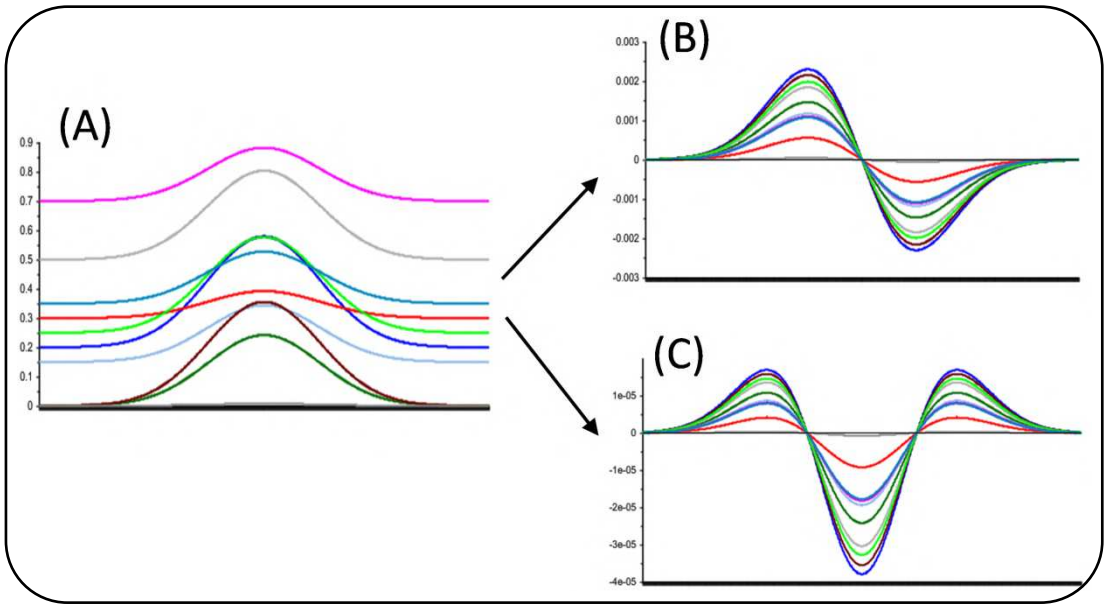


Figure 1.21 : (A) Courbes gaussiennes à décalages et intensités différents, (B) Dérivée première des courbes et (C) Dérivée seconde des courbes [63].

1.3.5.3.e) Normalisation:

La normalisation est un ensemble de transformations calculées pour chaque échantillon dans le but de "mettre à l'échelle" les échantillons afin de ramener toutes les données à peu près à la même échelle. Ceci est particulièrement utile lorsque les données sont collectées à l'aide d'une méthode ou d'un système où le signal du détecteur dépend de la masse de l'échantillon (e.g., la plupart des détecteurs de CG) ou de la puissance de la source (e.g., spectroscopie Raman) plutôt que de la concentration de l'échantillon. Le logiciel « The Unscrambler » (principal logiciel utilisé dans la partie expérimentale) propose différentes méthodes de normalisation, notamment la normalisation par aire, par vecteur unitaire, par moyenne, par maximum, par gamme et la normalisation par pics [63].

En analyse spectrale, la normalisation par vecteur unitaire (ou « Unit Vector Normalization, 'UVN' » en anglais) est souvent rapportée comme fournissant des résultats optimaux dans de nombreuses études précédentes. Pour un spectre *X*, l'équation de normalisation vectorielle est la suivante [63]:

$$x_i^{(UVN)} = \frac{x_i}{\sqrt{\sum_{j=1}^{m} x_{i,j}^2}} \qquad \text{(Éq. 1.18)}$$

1.3.5.3.f) Standard Normal Variate et De-Trending:

La transformation en variable normale standardisée (ou « Standard Normal Variate, 'SNV' » en anglais) est une autre méthode de prétraitement fréquemment utilisée, connue pour sa simplicité et son efficacité dans la correction des effets de diffusion. Elle est particulièrement utile dans les situations où les décalages de ligne de base et les variations de longueur de trajet font apparaître autrement des spectres identiques comme différents. La SNV, également appelée 'z-transformation', normalise chaque spectre à une moyenne nulle et à une variance unitaire (Figure 1.22). Cela est réalisé en soustrayant la moyenne du spectre et en divisant la différence par son écart-type [63, 80, 82]:

$$x_{i,j}^{(SNV)} = \frac{x_{i,j} - \bar{x}_i}{\sqrt{\frac{\sum_{i=1}^{m}\left(x_{i,j} - \bar{x}_i\right)^2}{m-1}}} \qquad \text{(Éq. 1.19)}$$

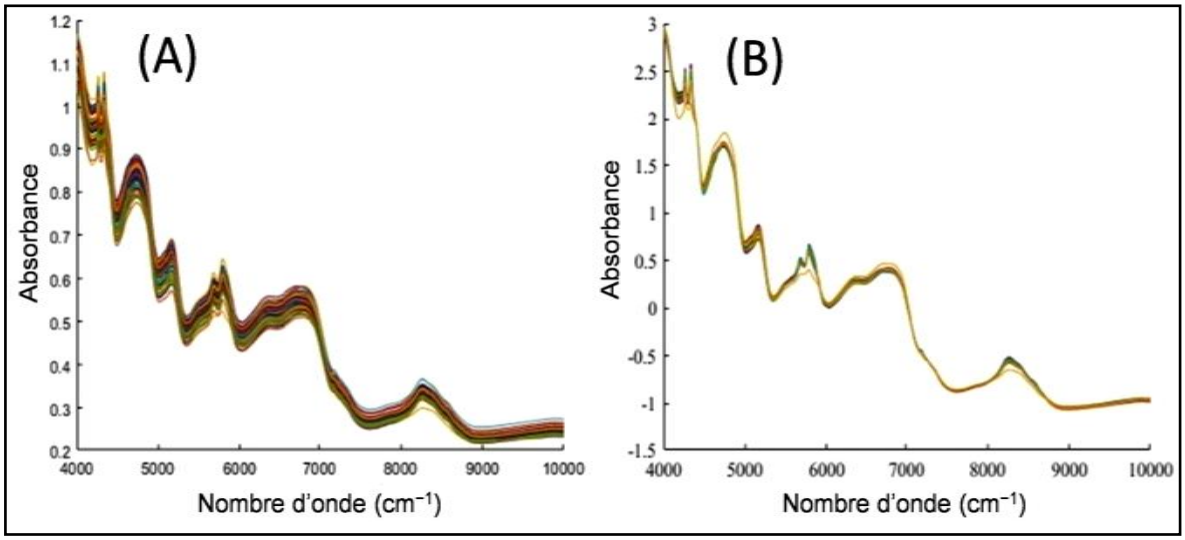


Figure 1.22 : Spectres NIR de 80 échantillons de fructo-oligosaccharides, originaux (A) et prétraités par SNV (B) [83].

Alors que la SNV corrige les interférences multiplicatives, certaines données spectrales peuvent encore présenter des tendances non linéaires, notamment une courbure de ligne de base. La dé-tendance (correction de tendance ou « De-Trending, 'DT' » en anglais), souvent utilisé après la SNV, peut réduire la multicollinéarité, les décalages de ligne de base et la courbure dans les spectres. Le processus est simple: chaque spectre est d'abord ajusté à un polynôme pour créer une ligne de tendance, qui est ensuite soustraite du spectre d'origine [59, 63]. La Figure 1.23 montre des spectres NIR après l'application du DT avec différents ordres polynomiaux.

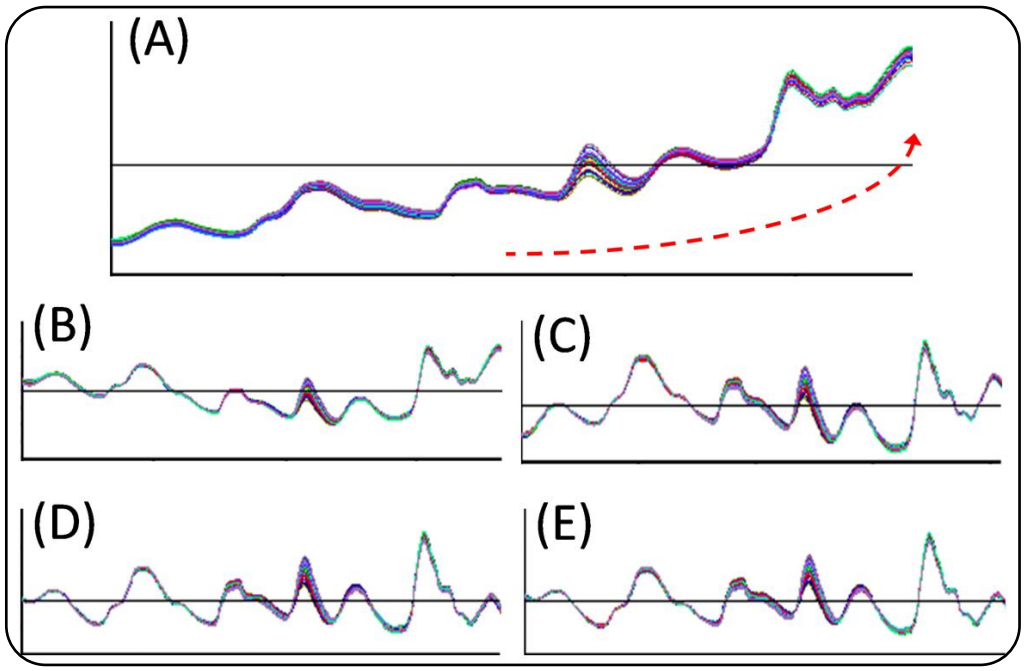


Figure 1.23 : (A) Spectres de réflectance diffuse NIR de la cellulose, la tendance non linéaire est indiquée à peu près par la courbe rouge en pointillés ; et (B à E) Mêmes spectres prétraités par SNV et DT avec un polynôme d'ordre 1 à 4, respectivement [63].

1.3.5.3.g) <u>Correction de la diffusion multiplicative et correction étendue de la diffusion multiplicative:</u>

La correction de la diffusion multiplicative (ou « Multiplicative Scatter Correction, 'MSC' » en anglais), également connue sous le nom de correction de signal multiplicatif, est une méthode de transformation utilisée pour compenser les effets purement additifs et/ou multiplicatifs dans les données spectrales résultant de la spectroscopie de réflectance. Le processus de correction implique de régresser un spectre mesuré (x_i) par rapport à un spectre de référence (souvent la moyenne, $\bar{x}_{i,j}$), puis d'ajuster le spectre mesuré en utilisant la pente et l'interception de cette équation linéaire comme suit [63, 80]:

$$x_i = a_i + b_i \bar{x}_{i,j} + e_i \qquad \text{(Éq. 1.20)}$$

$$x_{i,j}^{(MSC)} = \frac{x_i - a_i}{b_i} \qquad \text{(Éq. 1.21)}$$

où a_i et b_i sont les paramètres additifs (interception) et multiplicatifs (pente) de la régression linéaire, respectivement.

La Figure fournie 1.24 montre comment les effets additifs et multiplicatifs déforment les longueurs d'onde individuelles par rapport à un spectre de référence et visualise l'impact de la MSC pour compenser ces effets.

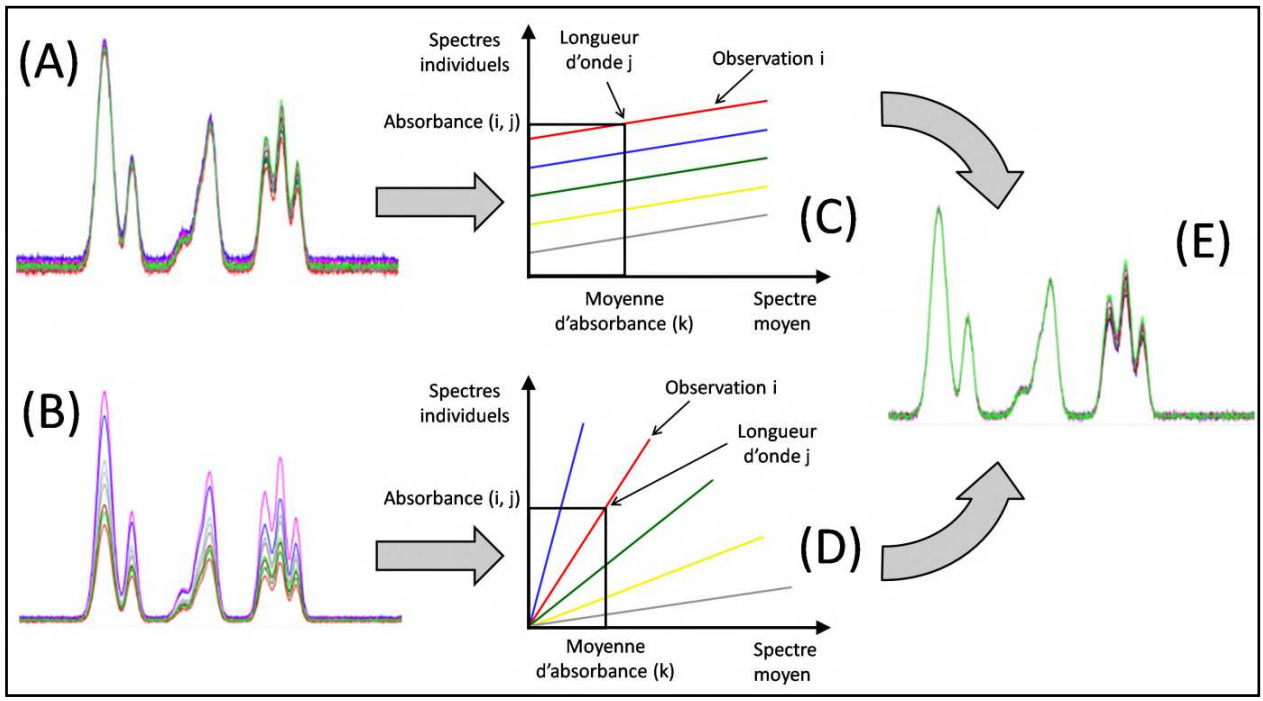


Figure 1.24 : (A) Spectres avec décalage additif de la ligne de base ; (B) Spectres avec effet multiplicatif ; (C et D) Tracés des longueurs d'onde individuelles par rapport au spectre moyen respectivement aux effets de diffusion additive (vers le haut) et multiplicative (vers le bas) ; et (E) Spectres corrigés par MSC [63].

S'appuyant sur l'algorithme MSC, plusieurs méthodes améliorées ont été introduites. Par exemple, une correction étendue de la diffusion multiplicative (« Extended MSC », 'EMSC' en anglais) étend ses capacités au-delà des effets simples additifs et multiplicatifs. En incorporant des termes dépendant de la longueur d'onde ou des informations a priori (Éq. 1.22), l'EMSC améliore la séparation de la diffusion physique de l'absorption chimique. Cela permet une correction plus nuancée et prend en compte des effets complexes [63].

$$x_i = a_i + b_i \bar{x}_{i,j} + d_i \lambda + g_i \lambda^2 + h_i BS + I_i GS + m_i m^2 + e_i \quad \text{(Éq. 1.22)}$$

où d_i, g_i, h_i, I_i et m_i sont des paramètres supplémentaires représentant des effets. Une explication plus détaillée peut être trouvée dans la référence [63].

1.3.6. Évaluation du modèle

1.3.6.1. Évaluation de la performance des modèles de régression

1.3.6.1.a) Indicateurs de performance:

Les indicateurs de performance (facteurs de mérite ou « Figures of Merit, 'FOMs' ») remplissent un rôle crucial en permettant la comparaison de différentes méthodes analytiques à travers des indicateurs numériques simples, fiables et facilement compréhensibles. Cela permet aux analystes de sélectionner efficacement la méthode la plus appropriée pour une application spécifique en pondérant les FOMs par rapport à d'autres facteurs pertinents, tels que le coût, le temps opérationnel, le potentiel d'automatisation, etc [61].

Dans la partie expérimentale de cette thèse, la performance des modèles a été vérifiée en utilisant la pente, l'interception, les erreurs quadratiques moyennes (RMSE), et les coefficients de détermination (R^2) de calibration, de validation croisée, et de prédiction. D'autres indices statistiques ont également été évalués afin de corroborer la robustesse, la justesse et la précision des modèles les plus performants. Il s'agit notamment de l'erreur absolue moyenne en pourcentage (MAPE), l'erreur relative de prédiction (REP), « Ratio of Prediction-to-Deviation (RPD) », « Range Error Ratio (RER) », et le biais (« Bias ») pour l'ensemble de test; et des FOMs tels que la sensibilité (SEN), la sensibilité analytique (γ), la sélectivité (SEL), la limite de détection (LOD) et la limite de quantification (LOQ) ont été définies et interprétées selon une approche moderne. Le calcul de ceux-ci est détaillé dans les équations 1.23 à 1.32 et le Tableau 1.3 [78, 84-88].

$$RMSE = \sqrt{\frac{1}{I}\sum_{i=1}^{I}(y_i - \hat{y}_i)^2} \quad \text{(Éq. 1.23)}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{I}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{I}(y_i - \bar{y})^2} \quad \text{(Éq. 1.24)}$$

$$PRESS = \sum_{i=1}^{I}\left(y_i - \hat{y}_i^{(CV)}\right)^2 \quad \text{(Éq. 1.25)}$$

$$Q^2 = 1 - \frac{PRESS}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad \text{(Éq. 1.26)}$$

$$Biais = \frac{1}{I}\left(\sum_{i=1}^{I}\hat{y}_i - \sum_{i=1}^{I}y_i\right) \quad \text{(Éq. 1.27)}$$

$$SEP = \sqrt{\frac{1}{I-1}\sum_{i=1}^{I}(y_i - \hat{y}_i - Biais)^2} \quad \text{(Éq. 1.28)}$$

$$MAPE = 100\sum_{i=1}^{I}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \quad \text{(Éq. 1.29)}$$

$$REP = 100\frac{RMSEP}{\bar{y}_{Cal}} \quad \text{(Éq. 1.30)}$$

$$RPD = \frac{SD_{y,Val}}{SEP} \quad \text{(Éq. 1.31)}$$

$$RER = \frac{y_{max} - y_{min}}{SEP} \quad \text{(Éq. 1.32)}$$

où I est le nombre total d'échantillons, y_i et $\hat{y}_i$ sont les valeurs réelles et prédites pour le *i*-ème échantillon, y_{max} et y_{min} sont les valeurs maximale et minimale dans les ensembles de données de référence, respectivement, $\bar{y}$ est la moyenne des valeurs réelles, $SD_{y,Val}$ est l'écart-type des valeurs de référence dans l'ensemble de données de test, SEP est la $RMSEP$ corrigé du $Biais$, $PRESS$ est la somme des carrés des erreurs résiduelles prédites, Q^2 est la proportion de la variabilité des échantillons exclus lors de l'étape de validation croisée.

Tableau 1.3 : Les FOMs analytiques pour les modèles de calibration univarié et multivarié.

Modèle	Univarié (loi de Beer)	Multivarié (PLSR)
SEN	B	$\frac{1}{\sqrt{\sum_{j=1}^{J} b_{jn}^2}}$ (Éq. 1.33)
γ	$\frac{B}{s_x}$ (Éq. 1.34)	$\frac{s_x}{\sqrt{\sum_{j=1}^{J} b_{jn}^2}}$ (Éq. 1.35)
LOD	$\frac{3.3\, s_{y/x}}{B}\sqrt{1 + h_0 + \frac{1}{I}}$ (Éq. 1.36)	$LOD_{min} = 3.3\sqrt{\frac{var(x)(1+h_{0min})}{\|b_n\|^2} + h_{0min}\, var(y_{Cal})}$ (Éq. 1.37) $LOD_{max} = 3.3\sqrt{\frac{var(x)(1+h_{0max})}{\|b_n\|^2} + h_{0max}\, var(y_{Cal})}$ (Éq. 1.38)
LOQ	$3 \times LOD_u$ (Éq. 1.39)	$3 \times [LOD_{min} - LOD_{max}]$ (Éq. 1.40)
SEL		$\frac{SEN}{SEN_0}$ (Éq. 1.41)

Là,

$$h_0 = h_{0\,min} = \frac{\bar{y}_{Cal}^2}{\sum_{i=1}^{I}(y_i - \bar{y}_{Cal})^2} \quad \text{(Éq. 1.42)}$$

$$h_{0\,max} = max\left(h_i + h_{min}\left[1 - \left(\frac{y_i - \bar{y}_{Cal}}{\bar{y}_{Cal}}\right)^2\right]\right) \quad \text{(Éq. 1.43)}$$

où B est la pente dans la calibration univariée, SEN_0 la sensibilité de l'analyte dans sa forme pure, b_n est le vecteur des coefficients de régression, s_x l'incertitude dans la réponse analytique (une mesure du bruit instrumental), $s_{y/x}$ l'écart-type résiduel dans la calibration univariée, $var(x)$ est la variance dans le signal instrumental, $var(y_{Cal})$ est la variance dans les concentrations de calibration, h_0 est le « leverage » du blanc, et $h_{0\,min}$ et $h_{0\,max}$ sont respectivement les valeurs minimale et maximale du « leverage » au niveau du blanc [86]. Les autres termes ont déjà été définis précédemment.

Dans les cas optimaux, les erreurs, l'interception et le biais devraient être proches de 0; la pente et R^2 (ou R) devraient être proches de 1, et la différence entre leurs valeurs respectives devrait être minimale pour les trois ensembles (calibration, CV et test). De plus, le modèle est optimal lorsque SEL est proche de 1; MAPE, REP, LOD et LOQ sont minimales; et SEN, γ, RPD et RER sont élevés. Ici, le RPD est une statistique non dimensionnelle couramment utilisée pour l'évaluation rapide des modèles de calibration des données spectroscopiques. Son interprétation est illustrée dans le Tableau 1.4 [88].

Tableau 1.4 : La statistique RPD pour la prédiction des aliments, des sols et des facteurs de fonctionnalité.

Valeur RPD	Classification	Application
0.0 – 1.9	Très faible	Non recommandé
2.0 – 2.4	Faible	Triage approximatif
2.5 – 2.9	Moyenne	Triage acceptable
3.0 – 3.4	Bonne	Contrôle qualité
3.5 – 4.0	Très bonne	Contrôle de processus
4.1+	Excellente	Toute application

1.3.6.1.b) Région de confiance elliptique conjointe:

Pour garantir des modèles précis et détecter tout biais constant ou proportionnel, en particulier dans un ensemble de tests à large échelle, le test recommandé est l'analyse dite de la région de confiance elliptique conjointe (ou « Elliptical Joint Confidence Region, 'EJCR' » en anglais). Il s'agit de tracer une EJCR pour la pente et l'interception de la droite de régression 'Prédite contre Réelle'. En termes pratiques, si le point idéal (pente = 1, interception = 0) se situe à l'intérieur de la région de confiance (à un niveau de confiance α choisi), cela indique l'absence de biais. Sinon, l'emplacement et la forme de l'EJCR peuvent révéler le type et l'étendue du biais présent [89-92]. L'expression spécifique décrivant l'EJCR est la suivante [86]:

$$N(Y-A)^2 + 2(X-B)(Y-A)\sum_{i=1}^{N} C_{Xn} + (X-B)^2 \sum_{i=1}^{N} C_{Xn}^2 = 2s_{Y/X}^2 F_{(0.05,2,N-2)} \quad \text{(Éq. 1.44)}$$

où B et A sont respectivement la pente et l'interception estimées pour la régression, N est le nombre d'échantillons de test, C_{Xn} est la concentration du n-ème échantillon utilisé comme référence et placé sur l'axe X de l'analyse de régression, $s_{Y/X}$ est l'écart-type résiduel de cette régression linéaire spécifique (à ne pas confondre avec celui correspondant au graphe de calibration), et $F_{(0.05,2,N-2)}$ est la valeur critique du paramètre F avec 2 et $N-2$ degrés de liberté et un niveau de confiance de 95%.

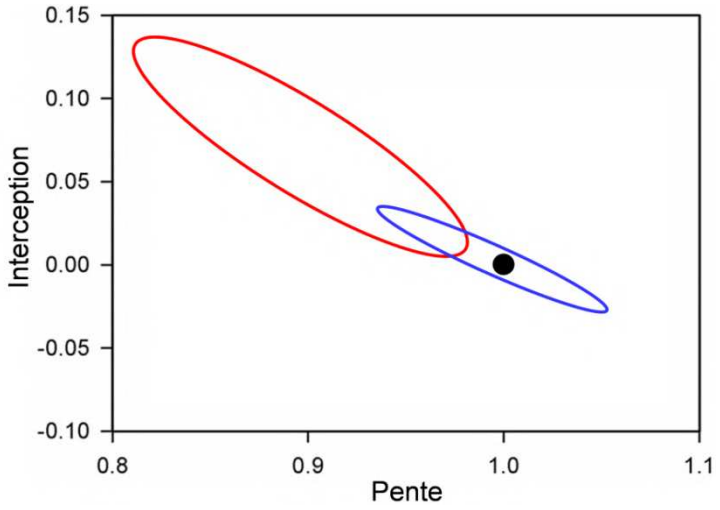


Figure 1.25 : Exemples de régions de confiance dans le plan pente-interception. L'ellipse bleue montre un modèle précis alors que l'ellipse rouge indique que le modèle correspondant est imprécis.

1.3.6.2. Évaluation de la performance des modèles de reconnaissance de motifs

La performance de l'analyse discriminante, en particulier des méthodes supervisées, peut être évaluée à la fois sur les ensembles de calibration et de validation en utilisant une matrice de confusion. Cette matrice illustre la relation entre les catégories réelles et prédites de chaque échantillon.

Pour un problème de classification impliquant G classes, la matrice de confusion est une matrice G × G (voir Tableau 1.5), où les lignes représentent les classes prédites et les colonnes représentent les classes réelles. L'élément "n_{kg}" dans la matrice indique le nombre d'échantillons de la vraie classe "g" qui ont été prédits comme classe "k". Les éléments diagonaux représentent les échantillons correctement prédits pour chaque classe [59]. Une prédiction parfaite aboutirait à une matrice diagonale, où tous les éléments hors de la diagonale sont nuls.

Tableau 1.5 : Structure de la matrice de confusion.

		Réelle			
		Classe 1	Classe 2	…	Classe G
Prédite	Classe 1	n_{11}	n_{12}	…	n_{1G}
	Classe 2	n_{21}	n_{22}	…	n_{2G}
	⋮	⋮	⋮	⋱	⋮
	Classe G	n_{G1}	n_{G2}	…	n_{GG}

La matrice de confusion peut être simplifiée en une "table de contingence", qui catégorise les prédictions en quatre résultats: Vrai Positif (TP) lorsqu'un échantillon réel est correctement identifié, Faux Positif (FP) lorsqu'un échantillon faux est incorrectement identifié, Vrai Négatif (TN) lorsqu'un échantillon faux est correctement identifié, et Faux Négatif (FN) lorsqu'un échantillon réel est incorrectement identifié. Sur la base de cette table, les paramètres suivants peuvent être calculés [93-95]:

$$Précision = \frac{TP}{TP+FP} \quad \text{(Éq. 1.45)}$$

$$Sensibilité = \frac{TP}{FN+TP} \quad \text{(Éq. 1.46)}$$

$$Spécificité = \frac{TN}{FP+TN} \quad \text{(Éq. 1.47)}$$

$$Efficacité = \sqrt{Sensiblité \times Spécificité} \quad \text{(Éq. 1.48)}$$

$$Excatitude = \frac{TP+TN}{TP+TN+FN+FP} \quad \text{(Éq. 1.49)}$$

Bien que d'autres métriques, telles que le coefficient de corrélation de Matthews (MCC) et le F1-score [96], ainsi que des courbes telles que ROC (« Receiver Operating Characteristic ») ou l'indice Kappa [97], existent, elles ne font pas partie du notre problème spécifique et à nos objectifs. Par conséquent, notre interprétation est centrée sur "l'Exactitude" en tant qu'indicateur robuste de l'efficacité du modèle pour notre tâche de classification abordée dans le Chapitre 3.

1.4. Revues de la littérature et état de l'art

1.4.1. Travaux antérieurs sur le tabac sans fumée

La recherche scientifique sur le ST est vaste et diversifiée, couvrant une large gamme de thèmes et de domaines d'investigation. À ce jour, la plupart des articles scientifiques sur le ST se sont concentrés sur la compréhension de ses effets nuisibles sur la santé et son potentiel de dépendance [5, 31, 35], l'évaluation des interventions et des programmes de cessation visant à réduire son utilisation [48], l'exploration des facteurs socio-économiques et des tendances de prévalence [12, 23, 98], le développement de biomarqueurs pour l'évaluation de l'exposition [99-101], et l'investigation des différentes caractéristiques des produits commerciaux, leurs variations et leurs ingrédients chimiques [24, 32, 102, 103].

La compréhension des caractéristiques fondamentales et de la composition chimique du ST constitue un pilier dans le domaine de la recherche. Cet aspect de l'étude est primordial car il permet d'évaluer les risques pour la santé associés à différents produits, de développer des stratégies efficaces de réduction des méfaits, d'envisager des mesures réglementaires et de fournir des informations précieuses pour les initiatives de santé publique. Les quelques sujets étudiés dans des travaux antérieurs sont résumés ci-dessous:

1.4.1.1. Qualité des feuilles

Plusieurs facteurs influencent les niveaux et types de composés chimiques et toxines présents dans les ST, à commencer par la feuille de tabac elle-même. Gupta *et al.* [104] ont passé en revue des facteurs clés tels que la génétique des plantes, les pratiques agricoles, les conditions de croissance, le séchage, le mixage des variétés, la fabrication, l'emballage et le stockage. Chaque étape contribue à la qualité du produit final et à son profil chimique.

1.4.1.2. Type du produit

Le ST se présente sous divers types et formes, chacun différant par sa texture, son contenu et ses additifs. Les tendances mondiales et régionales montrent des différences significatives dans les offres de produits. Par exemple, Stanfill *et al.* [102] a mené une étude à grande échelle dans cinq régions de l'OMS, analysant 65 produits différents du tabac oral, notamment le Chimó, le Gutkha, le Naswar, le Snus, le Toombak, et d'autres. Ainsi, une autre étude faite par Lawler *et al.* [25] est portée sur le marché américain, comparant les options établies (comme le « twist », le « loose leaf », le « plug » et le tabac à priser sec) avec les nouvelles introductions (comme le tabac à priser sec en pochettes, le Snus et les produits dissolvables).

1.4.1.3. Caractéristiques du produit

Les fabricants modifient souvent diverses caractéristiques chimiques et physiques des produits commercialisés pour les rendre plus acceptables et attrayants pour les utilisateurs. Des études menées par des équipes de recherche [105-107] ont identifié et quantifié divers additifs, notamment des agents aromatisants (menthol, eugénol, salicylate de méthyle, coumarine, camphre, et éther diphenyl, etc.), des édulcorants (sorbitol et xylitol), des liants (tels que les acides palmitique et stéarique), et des humectants (glycérine et propylène glycol). Ces constituants servent à améliorer le goût et l'arôme, masquer l'aspect désagréable et potentiellement accroître la dépendance. De plus, des caractéristiques telles que le pH, l'humidité et la taille des particules sont manipulées pour influencer des facteurs tels que la biodisponibilité de la nicotine,

la consistance, l'absorption et la distribution du produit dans la cavité buccale [38]. Des informations supplémentaires sur le pH et la teneur en humidité peuvent être trouvées dans la **sous-section 3.2.5** du Chapitre 3.

1.4.1.4. Composition chimique

Une recherche approfondie s'est penchée sur la composition chimique de divers produits de ST, en se concentrant sur l'identification d'une série de composés potentiellement nocifs. Notamment, les alcaloïdes et les agents cancérigènes sont fréquemment analysés. Ceux-ci peuvent être catégorisés en fonction de leur prévalence dans les publications scientifiques:

1.4.1.4.a) Alcaloïdes du tabac:

La nicotine, à la fois totale et non ionisée, arrive en tête de liste en raison de son impact substantiel sur les effets et le potentiel addictif de ST. Les variations dans la teneur en nicotine dépendent du type et de la marque du produit [49, 108]. De plus, d'autres alcaloïdes mineurs tels que la nornicotine, la cotinine, l'anatabine, l'anabasine, la myosmine, ainsi que leurs différents substituts et isomères, ont été documentés [109-111].

Des études ont révélé un large intervalle de concentrations de nicotine dans les produits commerciaux du monde. Par exemple, la teneur en nicotine totale dans des échantillons de ST du Pakistan et du Bangladesh variait respectivement de seulement 0,16 mg/g à un substantiel 34,1 mg/g de produit. Dans le même travail, la nicotine non ionisée variait de 0,05 à 31,0 mg/g, constituant de 0,16 à 99,1 % de la teneur en nicotine totale, selon le pH du produit [102].

Une recherche supplémentaire menée par Stepanov *et al.* [112] a exploré les concentrations d'alcaloïdes mineurs dans différents types de ST américain. Comparés aux teneurs de la nicotine dans les mêmes produits, la nornicotine, l'anatabine et l'anabasine étaient présents en moyenne de 0,95 %, 3,8 % et 0,32 %, respectivement, dans les tabacs à chiquer traditionnels. Ces valeurs étaient plus élevées dans des nouveaux produits, avoisinant en moyenne 3,3 %, 14 % et 0,77 %, respectivement. En outre, une étude plus récente confirme que

les niveaux les plus élevés d'alcaloïdes mineurs sont souvent retrouvés dans les produits de ST contenant les plus grandes quantités de nicotine [113].

1.4.1.4.b) Nitrosamines spécifiques au tabac:

Les produits de ST sont généralement soumis à une analyse de la présence de cinq TSNAs, à savoir NNK, NNN, NAT, NAB et NNAL [113, 114]. Après les alcaloïdes, cette classe de composés est la deuxième plus fréquemment rapportée dans les recherches sur les ST, mettant en évidence leurs préoccupations potentielles pour la santé.

La surveillance de ST mondial menée par Stanfill *et al.* [102] a révélé une large gamme de concentrations de TSNAs totaux, allant d'un minimum de 83,9 ng/g dans le ST pakistanais à un impressionnant 992000 ng/g dans le ST soudanais. Ces valeurs se situent sur cinq ordres de grandeur avec des concentrations de NNK. En particulier, les concentrations de NNK sont typiquement utilisées comme point de référence pour comprendre le niveau global de TSNAs dans le tabac.

1.4.1.4.c) Hydrocarbures aromatiques polycycliques:

De nombreux HAPs sont des agents cancérigènes puissants classés dans le Groupe 1 par le Centre international de recherche sur le cancer (IARC). L'analyse d'échantillons du ST a révélé la présence de divers HAPs à 2, 3, 4 et 5 cycles, notamment le naphtalène, phénanthrène, fluoranthène, pyrène, anthracène, fluorène, acénaphtène, benzo[a]pyrène et benzo[a]anthracène.

Des recherches ont montré des variations significatives dans les concentrations de HAPs en fonction du type de ST. Par exemple, le travail réalisé par McAdam *et al.* [40] a trouvé des niveaux de HAPs totaux allant de 193 à 19354 ng/g sur une base de poids sec pour les produits en pastilles dures et les tabacs à chiquer humides, respectivement. Dans un second travail [115], les niveaux de tous les HAPs détectés dans de nouveaux produits de tabac à chiquer (moyennés à 1,3 µg/g sur une base de poids sec) étaient beaucoup plus bas que ceux trouvés dans les produits de tabac à chiquer traditionnels (moyennés à

11,6 µg/g). Notamment, les HAPs carcinogènes représentaient 20 % du contenu total en HAP dans ces derniers produits.

1.4.1.4.d) Métaux toxiques:

En plus des préoccupations concernant les composés organiques, les ST disponibles commercialement contiennent également des métaux lourds et même des radionucléides. Des études menées par Kumar *et al.* et McNeill *et al.* [116, 117] ont identifié divers métaux tels que le cadmium, l'arsenic, le plomb, le chrome et le nickel. Selon Kumar *et al.* [116], les concentrations de ces métaux dans certains produits de ST international étaient 50 à 118 fois plus élevées que les limites définies par la norme GOTHIATEK (voir la **sous-section 1.1.7**). De plus, des traces d'éléments radioactifs, notamment le potassium-40, le polonium-210, le radium-226 et le thorium-228, ont été détectées dans des plages allant de $1 - 7 \times 10^{3}$, $1 - 11 \times 10^{-8}$, $1 - 24 \times 10^{-5}$ et $4 - 28 \times 10^{-8}$ ng/g sur une base de poids humide, respectivement, dans une autre publication [118].

1.4.1.4.e) Nitrate et nitrite:

Au cours du traitement des feuilles de tabac et du vieillissement de ST, les nitrates naturellement présents subissent principalement une réduction par la microflore pour former des nitrites. Le nitrite sert d'agent nitrosant majeur dans le tabac, réagissant avec les amines secondaires et tertiaires pour former les TSNAs cancérigènes [44].

Des études ont enquêté sur les niveaux de nitrate et de nitrite dans les ST, mettant en évidence leur influence potentielle sur la formation de TSNAs [43, 44, 119]. L'étude de Stepanov *et al.* [43] a trouvé une variation significative dans la concentration en nitrate, allant de 3,86 à 2950 µg/g sur une base de poids humide. De même, les niveaux de nitrite variaient considérablement, certains produits présentant des niveaux indétectables (< 0,02 µg/g) tandis que d'autres atteignaient des concentrations élevées allant jusqu'à 1410 µg/g dans les produits ST commercialisés en Inde.

1.4.1.4.f) Autres composés organiques:

Ajoutant une autre couche de préoccupation aux risques pour la santé des ST se trouve la présence d'une vaste éventail de composés méconnus, comprenant des N-nitrosamines volatiles et non volatiles [49, 120, 121], des acides nitrosamino [120, 122, 123], de l'ammoniac [24, 124], de l'acrylamide [125], des aldéhydes volatils [112], des sucres [126], etc. Alors que certains de ces composés ont des effets avérés sur la santé, beaucoup restent mal compris, soulevant des inquiétudes quant à leur impact potentiel sur les consommateurs.

1.4.2. Travaux antérieurs sur la détermination de la nicotine et d'autres paramètres de qualité dans le tabac

La quantification précise de la nicotine dans le ST est un aspect crucial de la recherche et de la réglementation sur le tabac. Les techniques chromatographiques, en particulier la chromatographie en phase gazeuse couplée à la spectrométrie de masse (GC-MS), sont largement considérées comme la référence en la matière. Le Tableau 1.6 donne un aperçu des protocoles de laboratoire couramment utilisés pour l'analyse de la teneur en nicotine dans quelques produits de ST. Ces protocoles incluent généralement la préparation des échantillons, l'extraction de la nicotine de la matrice, la séparation chromatographique et la quantification à l'aide de courbes d'étalonnage classique.

Cependant, les méthodes traditionnelles de GC-MS peuvent rencontrer des interférences d'autres constituants présents dans l'échantillon, ce qui pourrait compromettre potentiellement la précision. Pour relever ce défi, les chercheurs ont affiné les protocoles en tirant parti des progrès dans les instruments et les méthodologies analytiques. Ceux-ci incluent des techniques hybrides telles que la chromatographie en phase gazeuse couplée à la spectrométrie de masse en tandem (GC-MS/MS) et la chromatographie liquide couplée à la spectrométrie de masse en tandem (LC-MS/MS), qui offrent une sensibilité et une sélectivité accrues pour l'analyse à des niveaux de trace de la nicotine, de ses métabolites et d'autres alcaloïdes mineurs du tabac.

De plus, les recherches récentes dans ce domaine visent non seulement à développer des méthodologies standardisées pour une mesure fiable de la teneur en nicotine, mais aussi à explorer des techniques rapides et non destructives telles que la spectroscopie NIR associé à la chimiométrie. Ces méthodes permettent la détermination simultanée de plusieurs composants chimiques et propriétés de qualité à travers diverses marques et catégories de produits. Le Tableau 1.7 donne un résumé des méthodes développées dans la littérature qui ont été utilisées avec succès pour l'analyse quantitative et qualitative des échantillons de tabac et à base de tabac.

Il est important de noter que l'examen de la documentation consultée n'a identifié aucune référence se dédiant à l'analyse de paramètres de qualité des échantillons de ST en utilisant la spectroscopie ATR-FTIR. Cette lacune justifie pleinement la réalisation, dans le contexte de cette thèse, d'une étude qui explore les possibilités de cet outil, notamment lorsqu'il est combiné à des méthodes chimiométriques.

Tableau 1.6 : Protocoles de laboratoire couramment utilisés pour l'analyse de la nicotine totale dans quelques produits de ST.

Analyte	Traitement d'échantillons	Technique analytique	Échantillons de tabac étudiés	Source de l'échantillon/ année d'achat	Référence
Nicotine totale	- 1 g de tabac + 50 mL de MTBE contenant de la quinoléine (ÉI) + 5 mL de NaOH 2N. - Agiter sur un agitateur orbital pendant 2 h.	GC-MS (SIM)	56 produits Snus d'Europe du Nord et 8 produits de Snus américain.	États-Unis/ 2013-14.	[113]
			- Produits commerciaux de tabac à chiquer humide américain. - Échantillons internationaux de ST.	- États-Unis. - Divers pays d'origine.	[127]
	Méthode des CDC [45]: - 1 g de tabac broyé + 50 ml de MTBE contenant de la quinoléine (ÉI) + 5 ml de NaOH 2N. - Agiter pendant 2 h.	GC-FID	7 types de produits ST : tabac à chiquer finement coupé, longuement coupé, en pochette et à faible teneur en eau ; tabac à mâcher en vrac et en bloc ; Snus en pochette ; et un produit de type Gutkha.	Produits des États-Unis, du Royaume-Uni, de la Suède et de l'Inde importés au Canada.	[124]
	- 1 g de tabac pré-humidifié + 50 mL de méthanol (contenant de l'heptadécane 'ÉI'). - Agitation pendant 3 h. - Filtration.	GC-FID	- 5 tabacs à chiquer secs, 16 tabacs à chiquer humides, 13 tabacs à mâcher, 2 pastilles dures, 1 pastille molle et 1 produit en bloc. - 10 Snus en vrac et 22 Snus en portions.	États-Unis et Suède/ 2008-09	[128]
	- 50 mg de tabac conditionné en humidité + 20 ml de méthanol contenant 50 mg de KOH - Les échantillons ont été soniqués pendant 3 h puis centrifugés. - Une aliquote de l'extrait (200 mL) a été mélangée avec 20 µL de [CD_3]nicotine (ÉI).	GC-MS (SIM)	Un total de 216 échantillons (Snus et produits du tabac dissolvables)	États-Unis/ Avr.-Jul. 2011	[129]
			- 32 marques de ST (Zarda, Gutka, Khaini, Mishri, tabac à chiquer crémeux, dentifrice et tabac à chiquer suédois) ; - 5 marques populaires de préparations à mâcher (Supari) ; - ST de recherche humide (Standard 1S3).	Inde/ Oct.-Nov. 2003 Université de Kentucky	[43]
	- 20 mg d'échantillon broyé + 15-20 mL d'acétate d'ammonium 100 mM + 100 ppm de nicotine-d4 (ÉI). - Agitation pendant 40. - Filtration.	UPLC-ESI-MS/MS (MRM)	- 22 Zarda, 4 Gul, 3 Pan Masala, 3 produits à base de noix de bétel ; - 6 marques de tabac à chiquer ; - 8 tabacs à chiquer d'Asie du Sud-Est ; - ST de référence CORESTA (CRP-1.1 et CRP-2.1)	- Bangladesh/ Nov. 2016 - États-Unis/ Dec. 2016 - Pakistan et Inde/ 2017 - CDC, États-Unis	[108]

Alcaloïdes du tabac	Méthode de titrage AOAC [129]: - 2,5 g d'échantillon + 15 ml d'acide acétique à 5% + 100 ml de toluène-chloroforme (1:1 v/v) + 10 ml de NaOH à 36%. - Mélanger pendant 20 min puis filtrer. - 2 gouttes d'indicateur cristal violet + 25 mL d'aliquotes de la solution filtrée. - Titrage à l'aide d'une solution d'acide perchlorique toluène-chloroforme.	Méthode de titrage	6 marques de tabac à chiquer (Pan Masala)	Inde/ Août. 2015	[131]
	- La nicotine a été progressivement dissoute et libérée du Snus, déposée dans une coupelle de collecte puis analysée à λ = 260 nm. - Le signal spectral entrant dans le système a été transféré par une sonde à fibre optique.	Spectrométrie UV	16 produits commerciaux de Snus	Pays d'Europe du Nord et d'Amérique du Nord	[132]
	- 200 mg d'échantillons broyés + Isotopes de la R, S nicotine-d_4 racémique et de la R, S nornicotine-d_4 racémique comme ÉI + 200 µL de NaOH 5N. - 10 ml de méthanol à 70 % ont été ajoutés. - Après 1 h d'agitation, les extraits ont été filtrés puis dilués.	UPLC-ESI-MS/MS (MRM)	- Produits de référence à base de tabac moulu (Burley, Oriental, flue-cured, dark fire-cured et dark air-cured). - Produits de référence CORESTA.	-CTRP (Université de Kentucky). -Université d'État de la Caroline du Nord, États-Unis	[110]
	- 0,25 g de ST broyé + 25 ml d'eau déminéralisée. - Réaction à l'acide sulfanilique et au chlorure de cyanogène (la couleur développée est mesurée à 460-480 nm).	Méthode colorimétrique	Voir [128] (ci-dessus)		[133]
	- 1 g de tabac + 20 ml d'eau. -Incubation à 37 °C sous agitation pendant 24 h. - Le mélange a été centrifugé et décanté deux fois, puis le pH a été ajusté à 7,4. - Les extraits ont été stérilisés par filtration, lyophilisés et dissous dans de l'acétonitrile aqueux contenant 25 mg/ml de caféine (ÉI).	HPLC	- Feuilles de *Nicotiana gossei*. - ST de référence CORESTA (CRP2).	- Les feuilles de tabac ont été collectées à Alice Springs NT, Australie. - Université d'État de Caroline du Nord, États-Unis.	[111]

	Méthode AOAC 967.02 [133]	Spectrométrie UV	- 8 marques différentes de tabac à chiquer commercial. - 1 tabac à chiquer de référence (1S3).	- États-Unis. - Référence de l'Université de Kentucky, THRI/ 1986.	[135]

Abbreviations : AOAC, Association of Official Analytical Chemists ; CDC, Centers for Disease Control and Prevention ; CRP, CORESTA Reference Products ; CTRP, Center for Tobacco Reference Products ; ÉI, Étalon Interne ; ESI, ElectroSpray Ionization ; FID, Flame Ionization Detection ; GC, Gas Chromatography ; HPLC, High-Performance Liquid Chromatography ; MRM, Multiple Reaction Monitoring ; MS, Mass Spectroscopy ; MTBE, Methyl tert-Butyl Ether ; SIM, Selected Ion Monitoring ; ST, Tabac sans fumée ; THRI, Tobacco and Health Research Institute ; UPLC, Ultra Performance Liquid Chromatography ; UV, Ultraviolet.

Tableau 1.7 : Divers méthodes chimiométriques développées pour l'analyse quantitative et qualitative des échantillons de tabac.

Tâche	Analyte (ou propriété)	Technique analytique	Échantillons de tabac étudiés	Méthode(s) chimiométrique(s)	Métriques d'évaluation des modèles	Référence
Régression	Sucre total, sucre réducteur, nicotine et azote total.	Spectroscopie NIR	50 échantillons de tabac provenant de 11 zones de culture différentes en Chine.	Analyse comparative de: PLSR, SVMR et BP-ANN.	Meilleurs résultats obtenus en utilisant SVMR: RMSECV = 0.58, 0.65, 0.077, 0.069 et R^2 = 0.9903, 0.9817, 0.9901, 0.9724, respectivement pour chaque composantes.	[136]
	Alcaloïdes totaux, azote total et cendres totales.	Spectroscopie NIR	Un total de 322 matières premières et produits de tabac à cigare cubain.	Analyse comparative de: PCR, MLR et PLSR.	Meilleurs résultats obtenus en utilisant PLSR: SEP = 0.0011, 0.0012, 0.0049 et R^2_P = 0.987, 0.940, 0.957, respectivement pour chaque analyte.	[137]
	- 6 composés chimiques courants (sucre total, sucre réducteur, nicotine, nitrate total, K_2O et chlore) ; - 4 composés macromoléculaires (protéines, pectine,	Spectroscopie NIR	Un total d'environ 900 échantillons de tabacs reconstitués.	PLSR	- Composés chimiques courants: RMSEP = 0.65, 0.83, 0.10, 0.13, 0.10, 0.08 et R^2 = 0.90, 0.74, 0.86, 0.87, 0.89, 0.91 ; - Composés macromoléculaires: RMSEP = 0.17, 0.32, 0.15, 1.19 et R^2 = 0.96, 0.96, 0.95, 0.92 ; - Indices physiques: RMSEP = 0.89, 0.009, 0.06, 0.22, 0.85 et R^2 = 0.91,	[138]

	lignine et cellulose) ; - 5 indices physiques (grammage, épaisseur, résistance à la traction, pouvoir de remplissage et solubilité dans l'eau chaud).				0.67, 0.85, 0.88, 0.90. Toutes les valeurs sont dans l'ordre respectif de chaque analyte ou propriété.	
	Nicotine	Spectroscopie NIR	Feuilles de tabac de Chine.	Analyse comparative de: PLSR, SVMR, CNN et FCN.	Meilleurs résultats obtenus en utilisant FCN: RMSEP = 0.049 et R^2_P = 0.9969.	[139]
	Nicotine, sucre total, sucre réducteur, azote total, potassium, chlore et pH.	Hyperspectroscopie NIR	Un total de 4000 échantillons standard de feuilles de tabac chinois.	Analyse comparative de: PLSR, SVMR et LSTMNN.	Meilleurs résultats obtenus en utilisant LSTMNN: RMSEP = 0.041, 0.051, 0.041, 0.044, 0.047, 0.038, 0.028 et R^2_P = 0.9998, 0.9988, 0.9992, 0.9996, 0.9994, 0.9999, 0.9997, respectivement pour chaque composante.	[140]
	Sucre total	Spectroscopie NIR	Au total, 126 échantillons de tabac collectés dans divers provinces en Chine.	PLSR	RMSEP = 0.73 et R^2_P = 0.9566.	[141]
Classification	Discrimination entre des différentes marques de cigarettes.	Spectroscopie NIR	Un total de 259 échantillons de 3 marques de cigarettes différentes provenant de Chine.	Analyse comparative de: SVM-C, KNN et SIMCA.	Exactitude moyenne = 99.2, 96.8 et 97.7 %, respectivement pour chaque modèle d'apprentissage.	[142]
	Discrimination des mélanges de tabac à cigarettes.	Spectrophotométrie UV−Vis	Un total de 250 échantillons (extraits aqueux) de 5 marques de cigarettes différentes composées de feuilles de tabac simples et mixtes.	Analyse comparative de: DA et C&RT.	Le modèle C&RT a donné les meilleurs résultats: Exactitude globale en CV = 100 et 84 % des échantillons provenant du même lot et de lots différents, respectivement.	[143]
	Discrimination du tabac à cigarettes en fonction des marques, de l'origine géographique et de l'authenticité.	Spectroscopie NIR	Au total, 200 paquets de 3 marques de cigarettes mondialement disponibles achetés dans 55 pays différents.	PCA; PLS-DA.	Exactitude de classification = 97 % obtenue avec PLS-DA.	[144]

	Classification des tabacs en fonction de la position des tiges, de la couleur et de la qualité des feuilles.	Imagerie hyperspectrale NIR	Au total, 4052 bottes de feuilles de tabac de 2 types (flue-cured Virginia et air-cured Burley) ont été sourcées auprès de 3 régions différentes du sud du Brésil.	SVM-C	Exactitudes moyennes de prédiction = - Position de la tige: 80.4 et 88.1 % ; - Couleur de la feuille: 95.9 et 96.5 % ; - Qualité de la feuille: 78.8 et 100 %, respectivement pour les 2 types des feuilles.	[145]
	Classification des régions de culture.	Spectroscopie NIR	Au total, 13370 échantillons de tabac collectés dans trois régions différentes de la province chinoise du Guizhou (nord, nord-est et nord-ouest), ainsi qu'un emplacement externe.	Analyse comparative de: SVM-C, ANN, CNN et 1D-ResNet.	Exactitudes globales de prédiction = 90.7, 85.3, 93.2 et 97.0%, respectivement pour chaque modèle d'apprentissage.	[146]
	Identification de l'origine géographique des feuilles de tabac.	IRMS; ICP-MS	Au total, 260 échantillons collectés dans 6 zones productrices de tabac en Chine.	Analyse comparative de: OPLS-DA et RF.	RF a donné les meilleurs résultats: Exactitude globale de prédiction = 92.3, 95.4 et 98.5 % pour les modèles construits sur les ensembles de données IRMS, ICP-MS et IRMS+ICP-MS fusionnés, respectivement.	[147]

Abbreviations : 1D-ResNet, 1D Residual Neural network ; BP-ANN, Back Propagation Artificial Neural Network ; C&RT, Classification and Regression Tree ; CNN, Convolutional Neural Network ; CV, Cross-Validation ; DA, Discriminant Analysis ; FCN, Fully Convolutional Network ; ICP-MS, Inductively Coupled Plasma Mass Spectrometry ; IRMS, Stable Isotope Ratio Mass Spectrometry ; KNN, k-Nearest Neighbor ; LSTMNN, Long Short-Term Memory Neural Network ; MLR, Multiple Linear Regression ; NIR, Near Infrared ; OPLS-DA, Orthogonal Partial Least Squares-Discriminant Analysis ; PCA, Principal Component Analysis ; PCR, Principal Component Regression ; PLSR, Partial Least Squares Regression ; R^2, Determination Coefficient ; RF, Random Forest ; RMSECV, Root Mean Square Error of Cross-Validation ; RMSEP, Root Mean Square Error of Prediction ; SEP, Standard Error of Prediction ; SIMCA, Soft Independent Modeling of Class Analogies ; SVM-C, Support Vector Machine-Classification ; SVR, Support Vector Regression ; UV-Vis, Ultraviolet–Visible.

CHAPITRE 2

SPECTROSCOPIE ATR-FTIR COMBINÉE À LA CHIMIOMÉTRIE POUR LA QUANTIFICATION DE LA NICOTINE TOTALE DANS DES PRODUITS COMMERCIAUX DU TABAC SANS FUMÉE ALGÉRIEN

2.1. Introduction

La nicotine (NCT), principal alcaloïde du tabac, est une substance addictive et toxique. Bien que la nicotine elle-même ne soit pas cancérigène [148], certains de ses dérivés, tels que la 4-(méthylnitrosamino)-1-(3-pyridyl)-1-butanone (NNK) et la N′-nitrosonornicotine (NNN), ont été classés comme cancérigènes de Groupe 1 pour l'homme par le Centre international de recherche sur le cancer (IARC) [149]. Ces deux derniers et au moins 30 autres composés présents dans les produits du tabac sans fumée (ST) [102] ont été associés à un large éventail de problèmes de santé, notamment les maladies cardiovasculaires, la perte de poids des organes, les lésions buccales précancéreuses et le recul des gencives [25, 150, 151]. Ils ont également été associés à des cancers de la cavité buccale, des poumons, du pancréas, de l'œsophage, des reins et de la vessie [25, 108, 152]; alors que la Chemma algérienne a été confirmée comme cause de cancer de la cavité buccale et de l'oropharynx et ce, dans une étude qui a comparé les utilisateurs de Chemma et les non-consommateurs de tout type de tabac [153].

Pour cette raison, la détermination de la concentration en nicotine est essentielle pour une meilleure compréhension des effets généraux de la consommation de ST sur la santé. La NCT, conjointement avec les N-nitrosamines spécifiques du tabac (TSNAs), constitue les composants les plus fréquemment évoqués dans les recherches récentes [113, 154, 155]. L'analyse de la nicotine dans les produits de ST est principalement réalisée par chromatographie en phase gazeuse (GC) couplée à la spectrométrie de masse (MS) en mode de surveillance d'ions sélectionnés (SIM) [113, 127] ou couplée à la détection par ionisation de flamme (FID) comme indiqué dans le Tableau 1.5. Une

nouvelle méthode analytique a réussi à introduire la chromatographie liquide-spectrométrie de masse en tandem (LC-ESI-MS/MS) en mode de réaction multiple (MRM) dans la quantification de la NCT [108, 110]. Cependant, toutes ces techniques sont relativement laborieuses et coûteuses, prennent du temps, peuvent utiliser de grandes quantités de solvants toxiques et ne sont pas adaptées aux analyses de routine.

Au-delà des techniques chromatographiques, la spectrophotométrie n'a pas été utilisée de manière approfondie pour étudier des échantillons de ST, sauf dans quelques travaux pour la détermination des alcaloïdes totaux par colorimétrie [133], le comportement de libération de la NCT du « Snus » par UV [132], dans la méthode officielle de l'AOAC-International [134] ou pour la quantification de la nicotine totale dans les cigarettes [156]. Cette dernière était une étude typique utilisant la spectroscopie infrarouge moyenne à transformée Fourier (FT-MIR) en mode transmission avec une régression linéaire univariée.

Au cours des deux dernières décennies, la spectroscopie infrarouge à transformée de Fourier à réflexion totale atténuée (ATR-FTIR) couplée à la chimiométrie s'est imposée comme une technique robuste pour les analyses de discrimination et de quantification. Outre sa rapidité, sa simplicité, son coût abordable et sa faible exigence en quantité d'échantillon, elle offre des résultats comparables à ceux des méthodes officielles pour la quantification de divers analytes, tels que le 5-(hydroxyméthyl)furfural dans le miel [157], les adultérations à la coumarine et à l'éthylvanilline dans les extraits de vanille pure [158], et la pénicilline et l'acide phénoxyacétique pendant les fermentations de *Penicillium chrysogenum* [159]. Dans tous les cas susmentionnés, la spectroscopie FT-MIR a réussi à extraire les informations à partir de matrices très complexes, à savoir des produits naturels, directement au moyen d'analyses mathématiques multivariées simples, en particulier la régression par moindres carrés partiels (PLSR).

Cependant, ce n'est pas toujours le cas. Parfois, l'analyte est présent à l'état de traces, et aucune bande de vibration ne peut être distinguée dans les spectres. Dans de telles situations, des méthodes telles que la séparation physique, l'extraction chimique, la préconcentration, ou encore une modification de la technique d'échantillonnage (par exemple, l'utilisation de la technique du film sec

mince [160]) peuvent être nécessaires pour améliorer la sensibilité et la sélectivité de l'analyte.

En tenant compte de ce qui précède, l'objectif de cette partie était de développer une méthode simple, rapide, hautement spécifique à l'analyte, plus efficace en termes de coût, d'économie de réactifs, d'échantillonnage et d'adaptation en tant que méthode d'analyse de routine pour la détermination précise de la nicotine totale dans des échantillons commerciaux de ST en utilisant la spectroscopie ATR-FTIR associée à deux méthodes de régression: univariée (loi de Beer-Lambert) et multivariée (PLSR). Des questions concernant les différentes stratégies de prétraitement pour améliorer la qualité du modèle, la précision et la justesse des méthodes chimiométriques ainsi la validation des résultats de prédiction ont été abordées.

2.2. Partie expérimentale

2.2.1. Équipements et logiciels

Les mesures ATR-FTIR ont été réalisées à l'aide d'un spectromètre Nicolet iS10 (Figure 2.1) équipé d'un cristal de diamant ATR à réflexion unique (accessoire Smart iTR), d'un séparateur de faisceau XT-KBr et d'un détecteur au sulfate de triglycine deutéré (DTGS-KBr) contrôlé par le logiciel OMNIC^{TM} version 9.8 (Thermo Fisher Scientific, États-Unis). Chaque spectre a été enregistré sur la gamme spectrale 4000 − 525 cm^{-1} en accumulant 32 balayages co-ajoutés et une résolution spectrale de 4 cm^{-1} (espacement des données de 0,482 cm^{-1}) en mode absorbance. Le Smart iTR a un trajet optique court en raison de sa conception à réflexion unique. Cette caractéristique rend l'accessoire bien adapté aux mesures quantitatives de films secs minces [161].

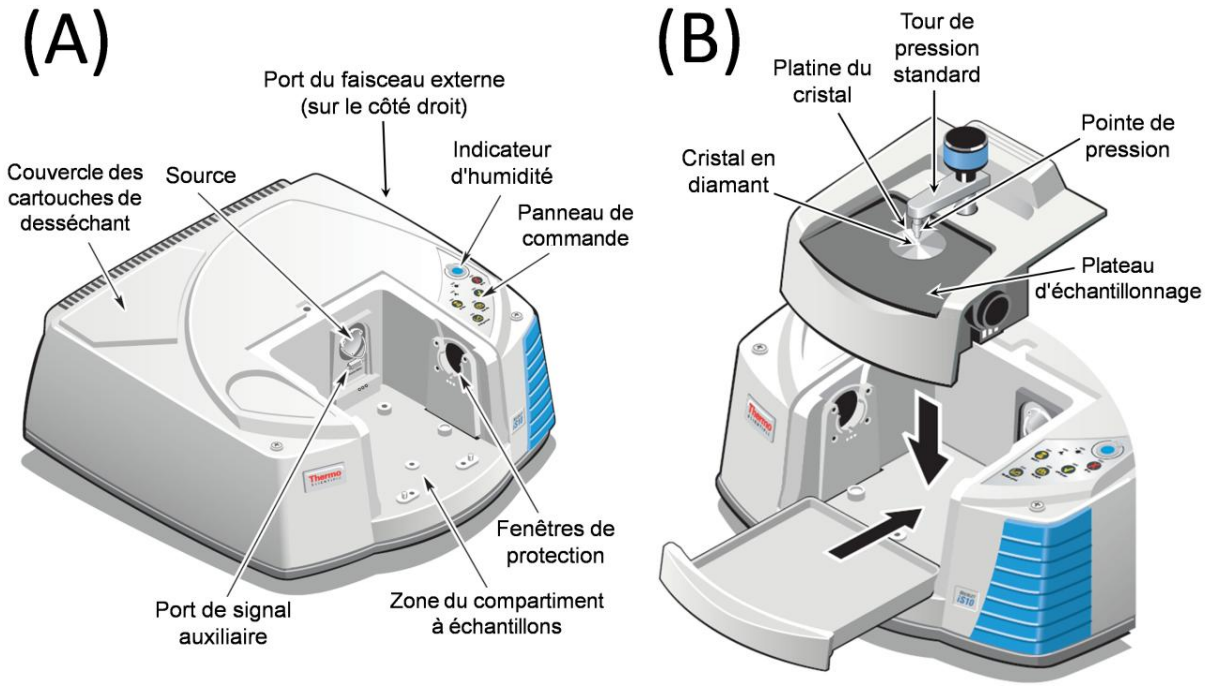


Figure 2.1 : Caractéristiques principales du (A) Spectromètre Nicolet iS10 et (B) Accessoire Smart iTR.

Pour assurer la fiabilité et la précision des résultats du spectromètre, l'utilisateur effectue régulièrement des contrôles de performance et des calibrages dans le cadre des procédures d'entretien et d'assurance qualité de l'instrument. L'analyse des matériaux de référence est un moyen courant et pratique permettant d'effectuer plusieurs tests simultanément. Au cours de notre étude, nous avons fait les vérifications présentées aux annexes 1 et 2 de l'Appendice B. Il s'agissait notamment de:

- **Énergie:** ce test évalue l'intensité globale du rayonnement infrarouge atteignant le détecteur. Il est généralement vérifié en mesurant la force du signal à un nombre d'onde spécifique ou en le comparant à une valeur de référence.
- **Bruit:** il mesure les fluctuations aléatoires du signal, qui peuvent affecter l'intensité et la résolution des pics. Il est souvent évalué en mesurant le rapport signal / bruit à des nombres d'ondes spécifiques ou en analysant l'interférogramme lui-même.
- **Précision en nombre d'onde:** il garantit l'attribution correcte des nombres d'ondes aux caractéristiques spectrales. Il est vérifié à l'aide d'un standard

connu présentant des pics d'absorption bien définis (par exemple, un film de polystyrène a été utilisé pour cela).

- **Répétabilité:** elle évalue la capacité de l'instrument à produire des résultats cohérents en intensités lors de la mesure multiple du standard (comme le NG11 glass utilisé là).

Logiciels:

Les prétraitements spectraux de base et la construction des modèles mathématiques dans la méthode univariée ont été réalisés à l'aide du logiciel TQ Analyst 9.7 (Thermo Fisher Scientific Inc., États-Unis). Le prétraitement spectral multivarié, la PCA et la PLSR ont été effectués avec The Unscrambler® X 10.4 (Camo Software AS., Norvège). Les indicateurs analytiques de performance et autres statistiques ont été calculées à l'aide des boîtes à outils UNIVAR et MVC1 v. 2018 [86] et à l'aide de fonctions-maison exécutées dans MATLAB R2012a (Mathworks Inc., États-Unis).

2.2.2. Réactifs et produits chimiques

La (−)-nicotine (≥99%) de qualité GC a été obtenue auprès de Sigma-Aldrich (Chine). Le chloroforme stabilisé à l'éthanol (≥99%) et le 2-propanol (99,8%) de qualité HPLC provenaient de Sigma-Aldrich (France). L'hydroxyde de sodium (NaOH, ≥98%), le carbonate de sodium anhydre (Na_2CO_3, 99,5%) et le sulfate de sodium anhydre (Na_2SO_4, ≥99%) étaient de qualité "Réactif analytique" et ont été obtenus auprès de Biochem Chemopharma (France) ou Panreac (Espagne).

2.2.3. Collecte d'échantillons commerciaux

Un échantillon de commodité de 17 produits algériens de ST les plus vendus a été acheté localement auprès de magasins de tabac de gros et de boutiques cosmétiques situés dans six endroits différents de deux provinces: Médéa et Blida, entre mars 2021 et mars 2022. Il s'agissait d'échantillons en double d'un produit de référence donné par le Laboratoire Central de Contrôle Qualité de l'entreprise United Tobacco Company (UTC du groupe MADAR Holding, Boumerdès), de

deux produits authentiques disponibles dans le commerce portant le même nom de la marque certifiée, de deux analogues contrefaits de la marque certifiée, de 10 produits contrefaits fabriqués illégalement, d'une Chemma traditionnelle faite à la main (échantillon 16) et d'un produit aromatisé sans tabac (mélange d'épices) destiné à faciliter l'arrêt du tabac (échantillon 17). Les échantillons ont été choisis pour refléter environ 80 % de la part de marché en Algérie à cette époque.

Les feuilles de tabac de l'espèce *Nicotiana rustica* provenant de six régions différentes dans les provinces d'Ain Mlila, de Batna, de Biskra et d'Oued Souf, ont été fournies par le Laboratoire Central de l'UTC. Une variété de tabac non identifiée (cultivée dans la région de Jijel) utilisée dans la fabrication de la Chemma traditionnelle a été obtenue sur un marché de rue populaire. Des feuilles de tabac à cigarettes, comprenant Burley, Oriental et Virginia en bandes, fournies par la Société du Tabac Algéro-Emirati (STAEM, Tipaza) ont été utilisées comme référence pour valider les résultats obtenus.

Le contenu des deux sachets d'une même marque a été mélangé et homogénéisé à l'aide d'un moulin à café, remis dans son emballage d'origine, étiqueté, scellé dans des sachets en plastique puis conservé au congélateur à −10 °C jusqu'à l'analyse.

2.2.4. Préparation et analyse des échantillons

2.2.4.1. Préparation des extraits

L'extraction acido-basique de la nicotine est basée sur la nature alcaloïde de la nicotine et sa solubilité dans différents solvants. Dans ce travail, les extraits ont été préparés selon les procédures décrites dans [156, 162] avec quelques modifications.

- Une quantité de 2,00 g de Chemma commercialisée (séchée à l'obscurité pendant 3 jours à température ambiante) a été pesées dans un flacon en verre.
- 30 ml d'eau distillée ont été ajoutés, et le flacon fermé a été soniqué pendant 20 min dans un bain à ultrasons. Le processus de sonication génère de la chaleur, et la température atteint facilement 70 °C.

- Ensuite, 0,4 g de carbonate de sodium anhydre ont été ajoutés, et le mélange a été agité puis remis dans le bain pendant encore 10 min pour reposer.
- Après une double filtration et un ajustement du pH à 12 avec de l'hydroxyde de sodium (1M), le filtrat aqueux a été vortexé à deux reprises avec 4 ml de chloroforme pendant 2 min chacune.
- Les extraits décantés ont été rassemblés dans un tube à centrifugeuse en verre et centrifugés à 3000 tr / min pendant 10 min.
- La phase chloroformique obtenue a été transférée soigneusement à l'aide d'une pipette Pasteur, en filtrant sur 0,5 g de sulfate de sodium anhydre pour retenir les traces d'eau restantes, puis concentrée sur un bain-marie sous vide à 35 °C jusqu'à siccité.
- Le volume final de l'extrait brut a été ajusté à 3 ml avec du chloroforme avant l'analyse.

En ce qui concerne les extraits des feuilles de tabac, seules 1 g de matière végétale ont été soumises à la même procédure puis diluées dans des volumes allant de 1 à 5 ml de chloroforme en fonction des rendements d'extraction estimés par gravimétrie.

2.2.4.2. Préparation des solutions étalons

Pour déterminer la quantité de nicotine dans les extraits, une série de solutions étalons de concentrations connues a été préparée en diluant directement la nicotine dans du chloroforme. Idéalement, cette série devrait couvrir une large gamme, avec suffisamment de points pour définir avec précision la relation entre le signal et la concentration dans les échantillons analysés (voir Tableau 2.1).

Tableau 2.1 : Caractéristiques des ensembles de calibration et de test.

	Niveaux de concentration ($mg.ml^{-1}$)	SD
Domaine dynamique	1,0 – 20,0	–
Domaine linéaire	1,0 – 15,0	–
Ensemble de calibration (N = 30)	1,0; 2,0; 4,0; 8,0; 10,0 et 15,0	5,0
Ensemble de test (n = 15)	3,0; 6,0 et 12,0	3,9

Abréviation : SD, Écart-type.

Ensuite, les solutions étalons préparées précédemment ont été distillées à sec, puis traitées de la même manière que les extraits d'échantillons commerciaux. Cette démarche vise à confirmer la sélectivité de la procédure vis-à-vis de l'analyte d'intérêt.

2.2.4.3. Analyse ATR-FTIR

- Une micro-seringue Hamilton avec butée réglable a été utilisée pour déposer une goutte de 0,4 µl d'échantillon sur le cristal ATR. Cette goutte a été laissée à sécher à l'air ambiant (à une température de 25 ± 3 °C et une humidité de 45 à 60 %).
- Après évaporation du solvant, en 1 min exactement, le film sec obtenu a été recouvert d'une pointe concave de la tour de pression standard permettant la stabilisation de la nicotine pendant le temps d'acquisition (tout contrôlé par une surveillance en temps réel des étalons).
- Avant chaque mesure, le cristal a été nettoyé deux fois à l'aide d'un papier doux imbibé d'éthanol à 96° puis d'isopropanol.
- L'interférogramme de fond (ou le « background » en anglais) a été enregistré une fois pour chaque triplicat contre du chloroforme évaporé dans des conditions expérimentales identiques, puis automatiquement soustrait par le logiciel. 15 répliques ont été acquises pour chaque échantillon, ce qui augmente le nombre total à environ 700 spectres.

2.2.5. Analyse des données

2.2.5.1. Détection des échantillons aberrants

En premier lieu, une analyse en composantes principales (PCA) préliminaire a été effectuée sur les données spectrales de la région de l'empreinte digitale, corrigées par « Standard Normal Variate (SNV) », de la ligne de base et centrées sur la moyenne, en utilisant l'algorithme de décomposition en valeurs singulières (SVD) avec 7 composantes principales (PCs). L'objectif était de détecter et d'éliminer les répliques aberrantes à l'aide du graphique des scores et des valeurs Q-résidus / T^2 de Hotelling dépassant le niveau de signification variait entre

5 et 25 % [163]. Ensuite, les modèles de régression ont été construits à partir des spectres restants, divisés de manière ordonnée dans un rapport de 2:1 comme indiqué précédemment dans le Tableau 2.1.

2.2.5.2. Méthodes de prétraitement et de calibration utilisées

Des méthodes appropriées de prétraitement spectral, y compris la correction du décalage et la correction linéaire de la ligne de base (BO et LBC, respectivement), la correction de la diffusion multiplicative (MSC) avec son option étendue (EMSC), le lissage Savitzky-Golay (SGS), les dérivées Savitzky-Golay du premier et du second ordre (SG FD et SG SD, respectivement) à 11 points de chaque côté avec un second ordre polynomial, et la dé-tendance (DT) avec un second ordre polynomial, ainsi que leurs diverses combinaisons, ont été utilisées pour réduire les dérives de la ligne de base, les effets de diffusion de la lumière, la non-linéarité, les erreurs aléatoires et le bruit, les pics chevauchés et d'autres facteurs externes incontrôlés.

Au stade de la régression, sept bandes spécifiques à 716, 807, 903, 1025, 1189, 1315 et 1428 cm^{-1} ont été choisies sur la base de leurs valeurs d'intensité élevées. Leurs maximales hauteurs ou aires associées, corrigées ou non pour la ligne de base, ont été utilisées dans la calibration univariée (monovariée), tandis que les régions réduites correspondantes aux bandes précitées ont été utilisées dans la PLSR.

2.2.5.3. Validation et évaluation des modèles

Les modèles ont été validés par la méthode de validation croisée « leave-one-out », testés pour prédire les concentrations dans l'ensemble d'échantillons indépendants, et la performance de chaque modèle a été vérifiée à l'aide des indices suivants: pente, interception, RMSE, R^2 (ou R), biais, Q^2, REP, RPD, RER, SEN, γ, SEL, LOD et LOQ. Les définitions théoriques de ces facteurs de mérite se trouvent au niveau du Chapitre 1 (**sous-section 1.3.6.1**).

2.3. Résultats et discussion

2.3.1. Interprétation spectrale et analyse exploratoire

Le spectre entier de la nicotine présente différentes bandes caractéristiques, dont les attributions proposées, telles que décrites dans [164-168], sont les suivantes. Les bandes faibles à 3403 (large) et 1640 cm^{-1} sont respectivement dues aux vibrations d'étirement et de déformation O−H produites par l'eau résiduelle. Les pics dans la région 3000 - 2725 cm^{-1} correspondent aux étirements asymétriques et symétriques des groupes CH_2 pyrrolidiniques et CH_3 méthylamino pyrrolidiniques, ainsi qu'à d'autres bandes combinées diverses. La région de l'empreinte digitale se situe entre 1750 et 600 cm^{-1} et contient une trentaine de signaux significatifs. Le premier, à 1689 cm^{-1}, peut être associé au groupe amine pyrrolidinique protoné (NH^+) qui disparaît partiellement après l'ajout de NaOH, chevauchée probablement avec la déformation angulaire O−H de l'eau.

La Figure 2.2 montre les spectres ATR-FTIR moyens des films secs de la nicotine dissoute dans du chloroforme et de la nicotine traitée selon la procédure d'extraction à la même concentration, sur la gamme de nombres d'onde comprise entre 1600 et 600 cm^{-1}.

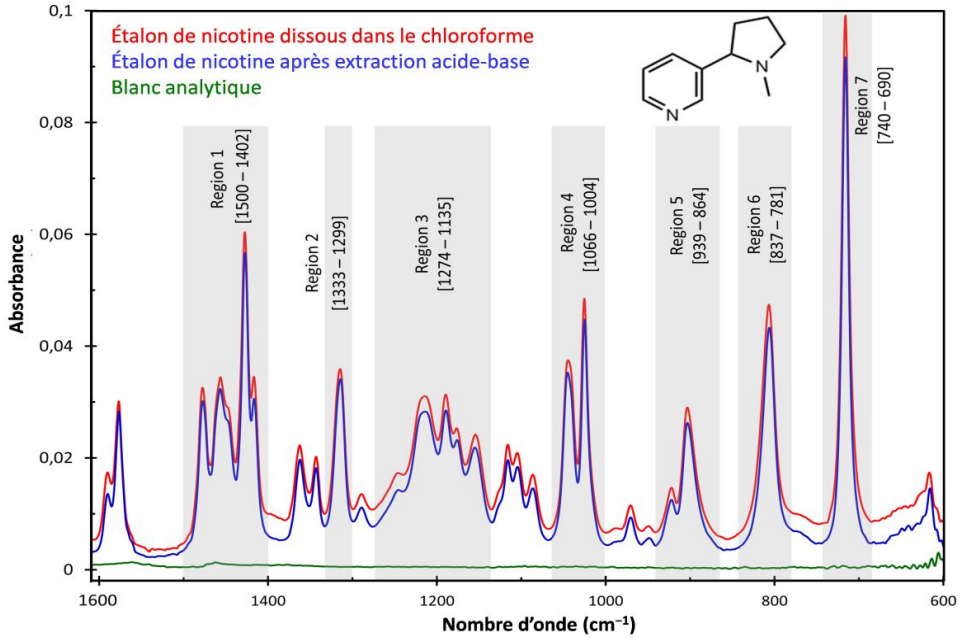


Figure 2.2 : Spectres FTIR-ATR des étalons de NCT (rouge et bleu) à des concentrations de 8 $mg.ml^{-1}$, ainsi que le spectre du blanc analytique (vert). Les zones spectrales favorables pour l'analyse de la NCT sont indiquées en gris.

Dans cette figure, plusieurs bandes d'absorption peuvent être distinguées:

- 1590 et 1577 cm^{-1}: Ces pics sont probablement corrélés aux vibrations d'étirement C−N, C−C, C=N et/ou C=C du cycle pyridinique.
- 1478, 1456, 1447, 1428 et 1417 cm^{-1}: Ces pics correspondent aux déformations asymétriques / symétriques des groupes CH_2 pyrrolidiniques et CH_3 terminal (1447 cm^{-1}), et pourraient se chevaucher avec la déformation H−C−H de la pyrrolidine à 1478 cm^{-1} et avec la déformation C−H de la pyridine à 1428 cm^{-1}.
- 1362 et 1343 cm^{-1}: Ces deux bandes sont liées aux mouvements d'agitation asymétrique du groupe CH_2 du cycle pyrrolidinique ou à l'étirement C−N des amines aromatiques tertiaires.
- 1315 cm^{-1}: Ce pic distinct correspond aux vibrations d'agitation C−N−C du méthyle pyrrolidinique.
- 1289 cm^{-1}: Il correspond à l'agitation symétrique du groupe CH_2 dans la même chaîne.
- 1275 - 1167 cm^{-1}: Ces bandes d'absorption sont principalement attribuées aux vibrations de torsion asymétrique et symétrique du CH_2 du cycle pyrrolidinique.
- 1116 cm^{-1}: Cette bande résulte de l'étirement asymétrique C−N−C.
- 1154 et 1086 cm^{-1}: Ces bandes sont associées à la déformation H−C−C−H et à la déformation symétrique C−H du cycle pyridinique, respectivement.
- 1045 cm^{-1}: Ce pic est dû à la déformation asymétrique C−H pyridinique et aux déformations C=C−C et C=N−C.
- 1025 cm^{-1}: Ce pic dominant peut être attribué à la vibration des chaînes de pyridine et de pyrrolidine ou à l'étirement C−N de pyridine.
- 971 cm^{-1}: Cette bande est due aux vibrations d'agitation asymétrique C−H de la pyridine.
- 922, 903 et 807 cm^{-1}: Ces bandes sont liées aux mouvements d'agitation symétrique et asymétrique des groupes CH_3 et CH_2.
- 716 cm^{-1}: C'est le plus intense de la région d'empreinte digitale est attribué à la déformation hors du plan C−H du cycle pyridinique monosubstitué.
- 616 cm^{-1}: Cette bande est attribuée à l'étirement dans le plan C−N−C.

Comme le montre la Figure 2.2, le blanc analytique ne contient aucune interférence ni contamination provenant du protocole d'extraction. En conséquence, le spectre de la nicotine traitée présente exactement les mêmes bandes caractéristiques que la nicotine pure, à l'exception d'un faible bruit irrégulier (bruit RMS = 1,6 10^{-4} unité d'absorbance « AU » dans toute la région d'empreinte) et d'une dérive de la ligne de base, ce qui confirme la nécessité d'une correction de la ligne de base pour des analyses quantitatives appropriées. Par conséquent, les spectres des étalons traités sont utilisés ultérieurement pour le développement de modèles et la quantification, car ils représentent mieux la nicotine extraite des échantillons réels.

Comme indiqué dans la **sous-section 2.2.5.1**, la PCA a été utilisée en tant qu'outil exploratoire pour exclure les répliques spectrales soupçonnés de ne pas appartenir à la population d'intérêt. Les première et deuxième composantes principales ont expliqué plus de 80 % de la variance totale dans tous les cas. Les Figures 2.3A et B montrent respectivement un exemple de résultat du graphique Q-résidus / T^2 de Hotelling et du graphique des scores avec la limite T^2 de Hotelling.

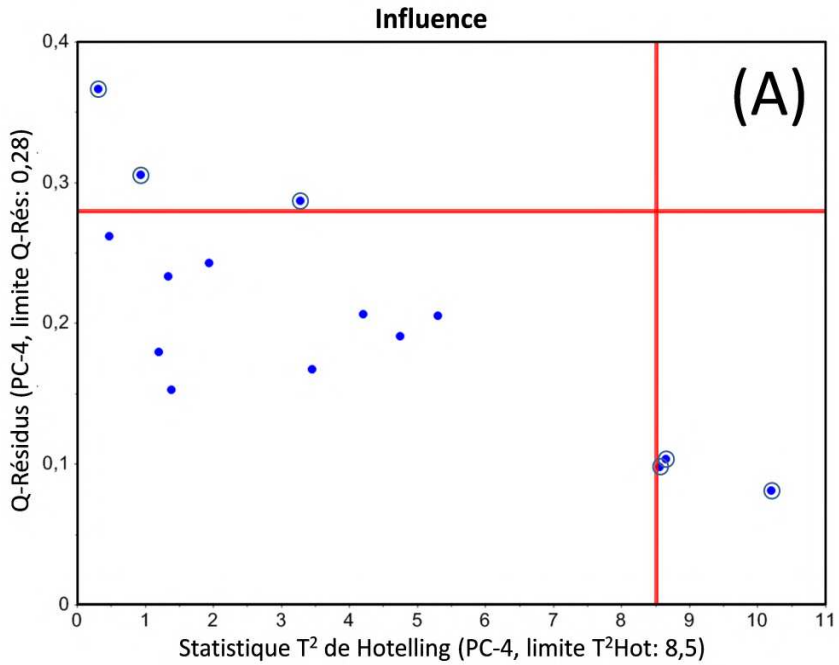


Figure 2.3 : Exemple de détection des valeurs aberrantes appliquée aux répliques d'une solution étalon traitée. (A) Graphique d'influence des Q-résidus contre T^2 de Hotelling.

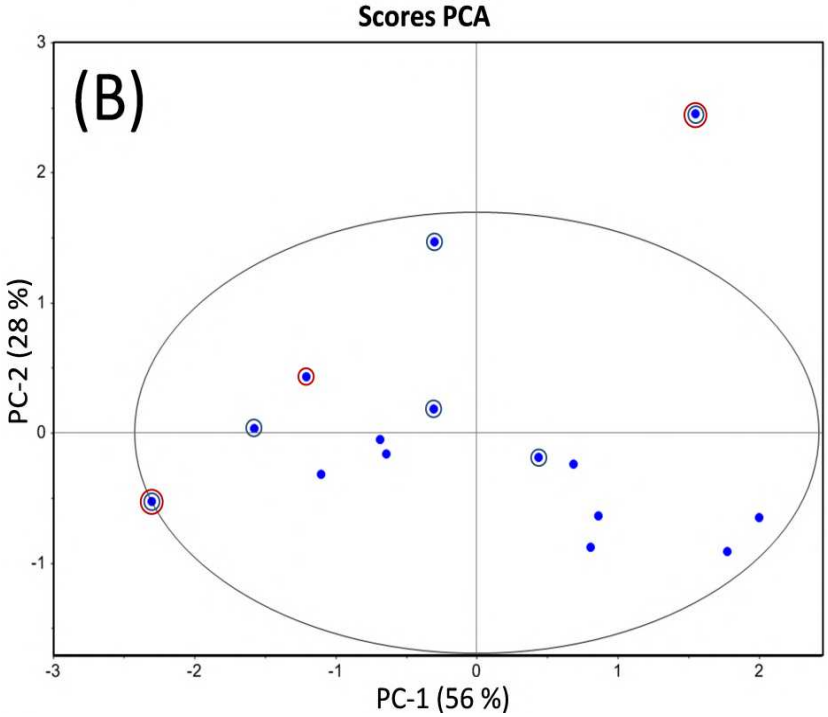


Figure 2.3 (suite) : (B) Graphique des scores PCA avec T^2 de Hotelling, tous deux ("A" et "B") à une limite critique 25 %.

Dans cet exemple, six points ont été identifiés à partir du graphique des Q-résidus contre la statistique T^2 de Hotelling et un point à partir du graphique des scores PCA. L'utilisation d'un niveau de signification aussi élevé s'explique par le fait qu'il n'y a pratiquement aucune différence entre les répliques. D'autres échantillons ont pu être éliminés du graphique des scores PLS, après la régression préliminaire, en conservant que cinq points à chaque niveau de concentration.

2.3.2. Optimisation des modèles de calibration

En utilisant une approche par essais et erreurs, les meilleurs modèles analytiques pour les méthodes univariées et multivariées ont été évalués en sélectionnant les mesures spectrales et les prétraitements mathématiques appropriés pour les étalons de nicotine: directement dissous dans le chloroforme et ceux soumis à la procédure d'extraction proposée.

Essentiellement, la MSC a été appliquée "séparément" aux répliques de chaque niveau de concentration, seule ou combinée avec d'autres techniques de prétraitement pour éliminer les effets de diffusion et les variations aléatoires dues aux volumes des gouttelettes déposées. Comme cette transformation normalise en fonction du spectre moyen dans l'ensemble de données, elle est préférée à la SNV qui recentre grossièrement chaque spectre entre −2 et +2 (− selon le logiciel utilisé −) [63], obtenant ainsi les mêmes intensités pour les spectres corrélés. La MSC ou l'EMSC peuvent améliorer la qualité du modèle tout comme la technique de la moyenne spectrale [157], conduisant à des modèles de calibration de meilleure qualité.

Les paramètres de régression complets et les stratégies suivies dans les analyses univariée et multivariée sont présentés respectivement aux Tableaux 2.2 et 2.3. Les meilleurs modèles ont été sélectionnés principalement en comparant les valeurs de pentes, d'interceptions, de RMSE et de R^2 (ou R) dans les ensembles de calibration et de validation.

En ce qui concerne la méthode univariée, le logiciel « TQ Analyst » présente l'avantage d'estimer la RMSE de manière similaire à l'approche utilisée dans les algorithmes multivariés. Il dispose également d'un paramètre distinct unique appelé indice de performance (PI) qui mesure, en pourcentage, la précision d'un modèle développé pour quantifier les échantillons de l'ensemble de test [169]. Tous les modèles sélectionnés avaient un PI supérieur à 90 %.

Tableau 2.2 : Paramètres de régression des procédures de calibration, de validation croisée et de prédiction de la nicotine dissous directement dans le chloroforme et soumis au traitement, obtenus par la loi de Beer-Lambert (analyse univariée).

Région spectrale	Format des données	Mesure spectrale (cm^{-1})	Ligne de base (cm^{-1})	Étalons	Calibration			Validation croisée			Test		Indice de performance
					Équation	RMSEC	R_C	Équation	RMSECV	R_{CV}	RMSEP	R_P	
Région 7 739,5-689,0	Spectre d'ordre zéro	Hauteur 716,0	Aucun	NCT dissous dans $CHCl_3$	y=0,0123x+ 0,00104	0,243	0,9988	y=0,0123x+ 0,00124	0,250	0,9988	0,201	0,9986	94,3
				NCT après traitement	y=0,0120x+ (-0,00393)	0,345	0,9975	y=0,0119x+ (-0,00354)	0,420	0,9974	0,347	0,9964	90,1
		Hauteur 716,0	739,5-689,0	NCT dissous dans $CHCl_3$	y=0,0112x+ (0,923e-3)	0,253	0,9987	y=0,0112x+ 0,00112	0,259	0,9987	0,182	0,9988	94,8
				NCT après traitement	**y=0,0114x+ (-0,00458)**	**0,343**	**0,9975**	**y=0,0113x+ (-0,00421)**	**0,449**	**0,9975**	**0,321**	**0,9968**	**90,8**
		Aire 739,5-689,0	Aucun	NCT dissous dans $CHCl_3$	y=0,195x+ 0,0186	0,285	0,9983	y=0,194x+ 0,0230	0,312	0,9981	0,303	0,9974	91,3
				NCT après traitement	y=0,161x+ (-0,0495)	0,367	0,9972	y=0,160x+ 0,00111	0,407	0,9965	0,474	0,9936	86,5
		Aire 739,5-689,0	739,5-689,0	NCT dissous dans $CHCl_3$	y=0,138x+ 0,0142	0,292	0,9982	y=0,137x+ 0,0175	0,317	0,9980	0,205	0,9986	94,1
				NCT après traitement	y=0,129x+ (-0,0369)	0,335	0,9976	y=0,129x+ (-0,0328)	0,384	0,9976	0,346	0,9962	90,1
	Dérivée première	Hauteur 712,8	Aucun	NCT dissous dans $CHCl_3$	y=0,00151x+ (0,847e-4)	0,292	0,9982	y=0,00150x+ (0,120e-3)	0,317	0,9979	0,185	0,9988	94,7
				NCT après traitement	y=0,00170x+ (-0,908e-3)	0,410	0,9965	y=0,00169x+ (-0,828e-3)	0,592	0,9962	0,312	0,9969	91,1
		Aire 715,5-695,2	Aucun	NCT dissous dans $CHCl_3$	y=0,0109x+ (0,547e-3)	0,249	0,9987	y=0,0108x+ (0,736e-3)	0,252	0,9987	0,191	0,9987	94,6
				NCT après traitement	y=0,0110x+ (-0,00458)	0,346	0,9975	y=0,0109x+ (-0,00421)	0,455	0,9975	0,336	0,9966	90,4
	Dérivée seconde	Hauteur 715,7	Aucun	NCT dissous dans $CHCl_3$	y=(-0,780e-3)x +(0,134e-3)	0,333	0,9977	y=(-0,777e-3)x +(0,110e-3)	0,476	0,9963	0,274	0,9982	92,2
				NCT après traitement	y=(-0,912e-3)x +(0,663e-3)	0,459	0,9956	y=(-0,904e-3)x +(0,609e-3)	0,803	0,9944	0,327	0,9967	90,6
		Aire 719,8-712,6	Aucun	NCT dissous dans $CHCl_3$	y=(-0,00297)x +(-0,312e-3)	0,258	0,9986	y=(-0,00297)x +(-0,367e-3)	0,278	0,9985	0,205	0,9987	94,2
				NCT après traitement	y=(-0,00325)x +0,00156	0,371	0,9971	y=(-0,00323)x +0,00143	0,524	0,9969	0,318	0,9968	90,9
Région 4 1033,2-1007,6	Spectre d'ordre zéro	Hauteur 1025,4	Aucun	NCT dissous dans $CHCl_3$	y=0,00579x + 0,00153	0,269	0,9985	y=0,00578x + 0,00164	0,370	0,9980	0,206	0,9990	94,1
				NCT après traitement	y=0,00558x + (-0,543e-3)	0,329	0,9977	y=0,00555x + (-0,375e-3)	0,338	0,9977	0,333	0,9962	90,5
		Hauteur 1025,4	1033,2-1007,6	NCT dissous dans $CHCl_3$	y=0,00397x + (0,738e-3)	0,274	0,9984	y=0,00395x + (0,821e-3)	0,336	0,9981	0,185	0,9990	94,7
				NCT après traitement	y=0,00407x + (-0,00119)	0,326	0,9978	y=0,00405x + (-0,00107)	0,378	0,9977	0,307	0,9970	91,2

		Aire 1033,2-1007,6	Aucun	NCT dissous dans $CHCl_3$	y=0,0754x + 0,0214	0,287	0,9983	y=0,0751x + 0,0231	0,399	0,9976	0,245	0,9985	93,0
				NCT après traitement	y=0,0678x + 0,00580	0,348	0,9975	y=0,0675x + 0,00809	0,356	0,9974	0,410	0,9941	88,3
		Aire 1033,2-1007,6	1033,2-1007,6	NCT dissous dans $CHCl_3$	y=0,0347x + 0,00445	0,300	0,9981	y=0,0346x + 0,00532	0,349	0,9977	0,204	0,9988	94,2
				NCT après traitement	y=0,0346x + (-0,00922)	0,323	0,9978	y=0,0344x + (-0,00821)	0,366	0,9978	0,320	0,9970	90,9
	Dérivée première	Hauteur 1022,6	Aucun	NCT dissous dans $CHCl_3$	y=(0,695e-3)x+ (0,169e-3)	0,253	0,9987	y=(0,693e-3)x+ (0,182e-3)	0,331	0,9984	0,185	0,9992	94,7
				NCT après traitement	y=(0,723e-3)x+ (-0,195e-3)	0,346	0,9975	y=(0,719e-3)x+ (-0,171e-3)	0,389	0,9974	0,283	0,9973	91,9
		Aire 1025,5-1004,2	Aucun	NCT dissous dans $CHCl_3$	y=0,00492x + 0,00112	0,268	0,9985	y=0,00491x + 0,00122	0,363	0,9979	0,192	0,9991	94,5
				NCT après traitement	y=0,00493x + (-0,00106)	0,340	0,9976	y=0,00491x + (-0,900e-3)	0,371	0,9974	0,295	0,9973	91,6
	Dérivée seconde	Hauteur 1025,4	Aucun	NCT dissous dans $CHCl_3$	y=(-0,427e-3)x +(-0,807e-4)	0,303	0,9981	y=(-0,425e-3)x +(-0,917e-4)	0,349	0,9978	0,239	0,9982	93,2
				NCT après traitement	y=(-0,490e-3)x +(0,244e-3)	0,373	0,9971	y=(-0,487e-3)x +(0,225e-3)	0,521	0,9971	0,320	0,9967	89,8
		Aire 1028,4-1022,1	Aucun	NCT dissous dans $CHCl_3$	y=(-0,00142)x +(-0,302e-3)	0,264	0,9985	y=(-0,00142)x +(-0,390e-3)	0,327	0,9983	0,172	0,9991	95,1
				NCT après traitement	y=(-0,00149)x +(0,477e-3)	0,337	0,9976	y=(-0,00147)x +(0,341e-3)	0,437	0,9964	0,294	0,9970	91,6
Région 2 1332,8-1299,3	Spectre d'ordre zéro	Hauteur 1314,6	Aucun	NCT dissous dans $CHCl_3$	y=0,00404x+ 0,00214	0,320	0,9979	y=0,00402x+ 0,00226	0,555	0,9965	0,308	0,9991	91,2
				NCT après traitement	y=0,00415x+ (0,272e-3)	0,310	0,9980	y=0,00413x+ (0,383e-3)	0,353	0,9975	0,365	0,9956	89,6
		Hauteur 1314,6	1332,8-1299,3	NCT dissous dans $CHCl_3$	y=0,00288x+ 0,00133	0,322	0,9978	y=0,00287x+ 0,00141	0,524	0,9966	0,250	0,9995	92,9
				NCT après traitement	**y=0,00314x+ (-0,240e-3)**	**0,336**	**0,9976**	**y=0,00312x+ (-0,141e-4)**	**0,390**	**0,9968**	**0,305**	**0,9970**	**91,3**
		Aire 1332,8-1299,3	Aucun	NCT dissous dans $CHCl_3$	y=0,0777x+ 0,0462	0,320	0,9978	y=0,0774x+ 0,0485	0,592	0,9962	0,366	0,9985	89,6
				NCT après traitement	y=0,0751x+ 0,0158	0,309	0,9980	y=0,0748x+ 0,0178	0,369	0,9977	0,560	0,9907	84,0
		Aire 1332,8-1299,3	1332,8-1299,3	NCT dissous dans $CHCl_3$	y=0,0394x+ 0,0191	0,316	0,9979	y=0,0393x+ 0,0202	0,542	0,9964	0,271	0,9993	92,3
				NCT après traitement	y=0,0415x+ (-0,624e-3)	0,341	0,9976	y=0,0413x+ (0,721e-3)	0,412	0,9965	0,294	0,9972	87,7
	Dérivée première	Hauteur 1309,5	Aucun	NCT dissous dans $CHCl_3$	y=(0,378e-3)x+ (0,177e-3)	0,322	0,9978	y=(0,376e-3)x+ (0,188e-3)	0,520	0,9967	0,198	0,9995	94,3
				NCT après traitement	y=(0,432e-3)x+ (-0,622e-4)	0,327	0,9978	y=(0,430e-3)x+ (-0,493e-4)	0,350	0,9975	0,313	0,9967	91,1
		Aire 1314,7-1298,8	Aucun	NCT dissous dans $CHCl_3$	y=0,00283x+ 0,00134	0,329	0,9977	y=0,00281x+ 0,00143	0,537	0,9964	0,259	0,9995	92,6
				NCT après traitement	y=0,00312x+ (-0,318e-3)	0,319	0,9979	y=0,00310x+ (-0,230e-3)	0,360	0,9973	0,277	0,9973	92,1

	Dérivée seconde	Hauteur 1312,9	Aucun	NCT dissous dans $CHCl_3$	y=(-0,111e-3)x +(-0,510e-4)	0,417	0,9964	y=(-0,110e-3)x +(-0,564e-4)	0,574	0,9953	0,306	0,9976	91,2
				NCT après traitement	y=(-0,137e-3)x +(0,342e-4)	0,368	0,9972	y=(-0,136e-3)x +(0,290e-4)	0,412	0,9969	0,591	0,9918	83,1
		Aire 1320,8-1309,4	Aucun	NCT dissous dans $CHCl_3$	y=(-0,679e-3)x +(-0,316e-3)	0,345	0,9975	y=(-0,676e-3)x +(-0,339e-3)	0,538	0,9962	0,210	0,9996	94,0
				NCT après traitement	y=(-0,761e-3)x +(0,868e-4)	0,326	0,9978	y=(-0,758e-3)x +(0,642e-4)	0,356	0,9974	0,304	0,9969	91,3
	Spectre d'ordre zéro	Hauteur 1189,5	Aucun	NCT dissous dans $CHCl_3$	y=0,00361x + 0,00158	0,292	0,9982	y=0,00360x + 0,00167	0,476	0,9973	0,297	0,9989	91,5
				NCT après traitement	y=0,00344x + (0,493e-3)	0,323	0,9978	y=0,00342x + (0,592e-3)	0,405	0,9969	0,379	0,9949	89,2
		Hauteur 1189,5	1273,8-1135,4	NCT dissous dans $CHCl_3$	y=0,00238x + (0,864e-3)	0,289	0,9982	y=0,00238x + (0,920e-3)	0,439	0,9975	0,253	0,9993	92,8
				NCT après traitement	y=0,00248x + (-0,174e-3)	0,319	0,9979	y=0,00247x + (-0,103e-3)	0,372	0,9971	0,300	0,9968	91,4
		Aire 1273,8-1135,4	Aucun	NCT dissous dans $CHCl_3$	y=0,334x + 0,167	0,298	0,9981	y=0,333x + 0,176	0,518	0,9970	0,308	0,9987	91,2
				NCT après traitement	y=0,305x + 0,0830	0,334	0,9977	y=0,303x + 0,0924	0,449	0,9967	0,424	0,9938	87,9
		Aire 1273,8-1135,4	1273,8-1135,4	NCT dissous dans $CHCl_3$	y=0,166x + 0,0623	0,291	0,9982	y=0,165x + 0,0662	0,459	0,9972	0,224	0,9993	93,6
Région 3 1273,8-1135,4				NCT après traitement	y=0,175x + (-0,0114)	0,331	0,9977	y=0,174x + (-0,00603)	0,384	0,9969	0,306	0,9968	91,2
	Dérivée première	Hauteur 1186,0	Aucun	NCT dissous dans $CHCl_3$	y=(0,146e-3)x+ (0,607e-4)	0,439	0,9959	y=(0,144e-3)x+ (0,686e-4)	0,552	0,9952	0,454	0,9966	87,0
				NCT après traitement	y=(0,153e-3)x+ (-0,139e-4)	0,347	0,9975	y=(0,152e-3)x+ (-0,877e-5)	0,405	0,9966	0,462	0,9938	86,8
		Aire 1189,4-1180,7	Aucun	NCT dissous dans $CHCl_3$	y=(0,788e-3)x+ (0,356e-3)	0,380	0,9970	y=(0,784e-3)x+ (0,387e-3)	0,545	0,9959	0,289	0,9993	91,8
				NCT après traitement	y=(0,828e-3)x+ (-0,991e-4)	0,320	0,9979	y=(0,825e-3)x+ (-0,754e-4)	0,348	0,9975	0,421	0,9941	88,0
	Dérivée seconde	Hauteur 1189,5	Aucun	NCT dissous dans $CHCl_3$	y=(-0,685e-4)x +(-0,363e-4)	0,452	0,9957	y=(-0,679e-4)x +(-0,402e-4)	0,588	0,9953	0,486	0,9978	86,1
				NCT après traitement	y=(-0,712e-4)x +(0,266e-5)	0,555	0,9935	y=(-0,703e-4)x +(-0,348e-5)	0,628	0,9917	0,554	0,9901	84,2
		Aire 1193,3-1185,5	Aucun	NCT dissous dans $CHCl_3$	y=(-0,316e-3)x +(-0,119e-3)	0,378	0,9970	y=(-0,314e-3)x +(-0,131e-3)	0,488	0,9964	0,356	0,9970	89,8
				NCT après traitement	y=(-0,324e-3)x +(0,270e-4)	0,346	0,9975	y=(-0,322e-3)x +(0,162e-4)	0,430	0,9962	0,387	0,9949	88,9
	Spectre d'ordre zéro	Hauteur 903,0	Aucun	NCT dissous dans $CHCl_3$	y=0,00353x + (0,658e-3)	0,307	0,9980	y=0,00352x + (0,751e-3)	0,393	0,9973	0,252	0,9981	92,8
				NCT après traitement	y=0,00323x + (-0,209e-3)	0,338	0,9976	y=0,00322x + (-0,106e-3)	0,370	0,9972	0,378	0,9950	89,2
Région 5 939,2-863,5		Hauteur 903,0	939,2-863,5	NCT dissous dans $CHCl_3$	y=0,00278x + (0,398e-3)	0,314	0,9979	y=0,00276x + (0,475e-3)	0,384	0,9973	0,229	0,9984	93,5
				NCT après traitement	y=0,00274x + (-0,633e-3)	0,354	0,9974	y=0,00272x + (-0,537e-3)	0,401	0,9970	0,351	0,9962	90,0

		Aire 939,2-863,5	Aucun	NCT dissous dans $CHCl_3$	y=0,125x + 0,0259	0,324	0,9978	y=0,125x + 0,0296	0,409	0,9971	0,290	0,9975	91,7
				NCT après traitement	y=0,107x + 0,00878	0,338	0,9976	y=0,106x + 0,0122	0,361	0,9974	0,516	0,9905	85,2
		Aire 939,2-863,5	939,2-863,5	NCT dissous dans $CHCl_3$	y=0,0682x + 0,00666	0,364	0,9972	y=0,0678x + 0,00918	0,408	0,9967	0,249	0,9979	92,9
				NCT après traitement	y=0,0693x + (-0,0229)	0,346	0,9975	y=0,0689x + (-0,0206)	0,405	0,9975	0,355	0,9958	89,9
	Dérivée première	Hauteur 898,0	Aucun	NCT dissous dans $CHCl_3$	y=(0,165e-3)x+ (0,267e-4)	0,459	0,9956	y=(0,163e-3)x+ (0,364e-4)	0,499	0,9951	0,447	0,9941	87,2
				NCT après traitement	y=(0,165e-3)x+ (-0,653e-4)	0,525	0,9942	y=(0,163e-3)x+ (-0,526e-4)	0,606	0,9934	0,603	0,9933	82,8
		Aire 903,0-876,0	Aucun	NCT dissous dans $CHCl_3$	y=0,00258x+ (0,545e-3)	0,314	0,9979	y=0,00257x+ (0,617e-3)	0,420	0,9970	0,231	0,9984	93,4
				NCT après traitement	y=0,00248x+ (-0,276e-3)	0,389	0,9968	y=0,00246x+ (-0,171e-3)	0,479	0,9953	0,345	0,9970	90,2
	Dérivée seconde	Hauteur 904,0	Aucun	NCT dissous dans $CHCl_3$	y=(-0,827e-4)x +(-0,147e-4)	0,724	0,9890	y=(-0,809e-4)x +(-0,268e-4)	0,871	0,9847	0,560	0,9909	84,0
				NCT après traitement	y=(-0,843e-4)x +(0,281e-4)	0,654	0,9910	y=(-0,828e-4)x +(0,181e-4)	0,715	0,9900	0,694	0,9836	80,2
		Aire 908,3-897,7	Aucun	NCT dissous dans $CHCl_3$	y=(-0,451e-3)x +(-0,536e-4)	0,343	0,9975	y=(-0,449e-3)x +(-0,685e-4)	0,366	0,9974	0,249	0,9981	92,9
				NCT après traitement	y=(-0,441e-3)x +(0,852e-4)	0,350	0,9974	y=(-0,439e-3)x +(0,702e-4)	0,410	0,9967	0,294	0,9971	91,6
Région 6 837,4-781,0	Spectre d'ordre zéro	Hauteur 806,6	Aucun	NCT dissous dans $CHCl_3$	y=0,00586x+ (0,754e-3)	0,295	0,9982	y=0,00584x+ (0,896e-3)	0,337	0,9979	0,229	0,9983	93,5
				NCT après traitement	y=0,00548x+ (-0,00106)	0,336	0,9976	y=0,00546x+ (-0,889e-3)	0,360	0,9977	0,370	0,9855	89,4
		Hauteur 806,6	837,4-781,0	NCT dissous dans $CHCl_3$	y=0,00475x+ (0,590e-3)	0,291	0,9982	y=0,00473x+ (0,702e-3)	0,329	0,9980	0,204	0,9985	94,2
				NCT après traitement	y=0,00472x+ (-0,00142)	0,345	0,9975	y=0,00470x+ (-0,00126)	0,403	0,9975	0,381	0,9955	89,1
		Aire 837,4-781,0	Aucun	NCT dissous dans $CHCl_3$	y=0,151x+ 0,0224	0,309	0,9980	y=0,150x+ 0,0264	0,368	0,9975	0,277	0,9976	92,1
				NCT après traitement	y=0,128x+ (-0,00345)	0,352	0,9974	y=0,127x+ (0,972e-3)	0,357	0,9974	0,468	0,9924	86,6
		Aire 837,4-781,0	837,4-781,0	NCT dissous dans $CHCl_3$	y=0,0896x+ 0,0125	0,294	0,9982	y=0,0893x+ 0,0147	0,350	0,9977	0,216	0,9986	93,8
				NCT après traitement	y=0,0854x+ (-0,0227)	0,371	0,9971	y=0,0849x+ (-0,0194)	0,421	0,9971	0,385	0,9949	89,0
	Dérivée première	Hauteur 801,5	Aucun	NCT dissous dans $CHCl_3$	y=(0,448e-3)x+ (0,470e-4)	0,368	0,9972	y=(0,446e-3)x+ (0,640e-4)	0,381	0,9971	0,178	0,9989	94,9
				NCT après traitement	y=(0,471e-3)x+ (-0,207e-3)	0,391	0,9968	y=(0,468e-3)x+ (-0,186e-3)	0,512	0,9967	0,428	0,9938	87,8
		Aire 806,6-781,0	Aucun	NCT dissous dans $CHCl_3$	y=0,00455x+ (0,652e-3)	0,297	0,9982	y=0,00453x+ (0,764e-3)	0,338	0,9979	0,193	0,9987	94,5
				NCT après traitement	y=0,00457x+ (-0,00149)	0,345	0,9975	y=0,00455x+ (-0,00134)	0,419	0,9975	0,379	0,9953	89,2

	Dérivée seconde	Hauteur 806,6	Aucun	NCT dissous dans $CHCl_3$	y=(-0,123e-3)x +(-0,417e-4)	0,562	0,9934	y=(-0,121e-3)x +(-0,525e-4)	0,611	0,9930	0,598	0,9890	82,9
				NCT après traitement	y=(-0,137e-3)x +(0,816e-4)	0,384	0,9969	y=(-0,136e-3)x +(0,760e-4)	0,575	0,9970	0,592	0,9926	83,1
		Aire 811,9-801,1	Aucun	NCT dissous dans $CHCl_3$	y=(-0,773e-3)x +(-0,926e-4)	0,325	0,9978	y=(-0,770e-3)x +(-0,115e-3)	0,351	0,9976	0,225	0,9983	93,6
				NCT après traitement	y=(-0,824e-3)x +(0,322e-3)	0,386	0,9969	y=(-0,819e-3)x +(0,288e-3)	0,482	0,9968	0,417	0,9940	88,1
	Spectre d'ordre zéro	Hauteur 1427,7	Aucun	NCT dissous dans $CHCl_3$	y=0,00659x + 0,00433	0,357	0,9973	y=0,00656x+ 0,00456	0,648	0,9953	0,374	0,9992	89,3
				NCT après traitement	y=0,00745x + (-0,00125)	0,321	0,9978	y=0,00742x + (-0,00103)	0,365	0,9976	0,445	0,9935	87,3
		Hauteur 1427,7	1500,0-1402,5	NCT dissous dans $CHCl_3$	y=0,00564x + 0,00351	0,365	0,9972	y=0,00561x+ 0,00372	0,637	0,9953	0,337	0,9995	90,4
				NCT après traitement	y=0,00656x + (-0,00133)	0,309	0,9980	y=0,00653x + (-0,00116)	0,342	0,9979	0,317	0,9966	90,9
		Aire 1438,0-1402,5	Aucun	NCT dissous dans $CHCl_3$	y=0,120x + 0,0815	0,343	0,9975	y=0,119x + 0,0854	0,657	0,9955	0,415	0,9988	88,1
				NCT après traitement	y=0,130x + (-0,0192)	0,409	0,9965	y=0,129x + (-0,0132)	0,533	0,9951	0,725	0,9836	79,3
		Aire 1438,0-1402,5	1500,0-1402,5	NCT dissous dans $CHCl_3$	y=0,0842x + 0,0510	0,345	0,9975	y=0,0838x + 0,0538	0,621	0,9956	0,331	0,9995	90,6
Région 1 1500,0-1402,5				NCT après traitement	y=0,0969x + (-0,0214)	0,303	0,9981	y=0,0966x + (-0,0189)	0,371	0,9978	0,405	0,9950	88,4
	Dérivée première	Hauteur 1424,3	Aucun	NCT dissous dans $CHCl_3$	y=(0,777e-3)x+ (0,479e-3)	0,412	0,9964	y=(0,772e-3)x+ (0,516e-3)	0,650	0,9946	0,356	0,9992	89,8
				NCT après traitement	y=(0,942e-3)x+ (-0,160e-3)	0,369	0,9972	y=(0,937e-3)x+ (0,124e-3)	0,435	0,9962	0,303	0,9972	91,4
		Aire 1427,6-1420,0	Aucun	NCT dissous dans $CHCl_3$	y=0,00313x + 0,00219	0,413	0,9964	y=0,00311x + 0,00234	0,719	0,9937	0,363	0,9991	89,6
				NCT après traitement	y=0,00376x + (-0,633e-3)	0,367	0,9972	y=0,00374x + (0,492e-3)	0,418	0,9965	0,283	0,9973	91,9
	Dérivée seconde	Hauteur 1428,2	Aucun	NCT dissous dans $CHCl_3$	y=(-0,423e-3)x +(-0,260e-3)	0,371	0,9971	y=(-0,420e-3)x +(-0,276e-3)	0,628	0,9953	0,383	0,9993	89,1
				NCT après traitement	y=(-0,507e-3)x +(0,133e-3)	0,399	0,9967	y=(-0,504e-3)x +(0,110e-3)	0,464	0,9960	0,282	0,9976	91,9
		Aire 1430,3-1424,2	Aucun	NCT dissous dans $CHCl_3$	y=(-0,00157)x +(-0,874e-3)	0,389	0,9968	y=(-0,00156)x +(-0,941e-3)	0,602	0,9953	0,316	0,9993	91,0
				NCT après traitement	y=(-0,00187)x +(0,354e-3)	0,364	0,9972	y=(-0,00186)x +(0,285e-3)	0,427	0,9964	0,292	0,9974	91,7

Les meilleurs modèles analytiques sont indiqués en gras et en couleur.

Tableau 2.3 : Paramètres de régression des procédures de calibration, de validation croisée et de prédiction de la nicotine dissous directement dans le chloroforme et soumis au traitement, obtenus par PLS-1 (analyse multivariée).

Gamme spectrale	Prétraitement	Étalons	Variables spectrales	Calibration				Validation croisée				Test			
				Pente	Intercept.	RMSEC	R^2_C	Pente	Intercept.	RMSECV	R^2_{CV}	Pente	Intercept.	RMSEP	R^2_P
Empreinte digitale 1605,0-680,3	Aucun	NCT dissous dans $CHCl_3$	1919	0,997	0,0203	0,270	0,9970	0,997	0,0198	0,288	0,9968	0,960	0,389	0,289	0,9941
		NCT après traitement	1919	0,996	0,0284	0,319	0,9957	0,995	0,0324	0,339	0,9955	0,959	0,388	0,433	0,9867
	BO	NCT dissous dans $CHCl_3$	1919	0,998	0,0153	0,234	0,9977	0,997	0,0169	0,250	0,9976	1,00	0,0589	0,240	0,9959
		NCT après traitement	1919	0,995	0,0380	0,369	0,9947	0,994	0,0380	0,369	0,9947	0,926	0,711	0,499	0,9823
	BO-MSC	NCT dissous dans $CHCl_3$	1919	1,00	1,05e-3	0,0612	0,9998	0,999	1,77e-3	0,0663	0,9998	1,00	0,0445	0,0783	0,9996
		NCT après traitement	1919	0,999	9,63e-3	0,186	0,9986	0,998	0,0135	0,198	0,9985	0,929	0,690	0,378	0,9899
	LBC-BO-EMSC-SGS	NCT dissous dans $CHCl_3$	1897	1,00	2,06e-3	0,0858	0,9997	0,999	2,30e-3	0,0931	0,9997	0,982	0,259	0,153	0,9983
		NCT après traitement	1897	0,998	0,0121	0,208	0,9982	0,998	0,0157	0,221	0,9981	0,914	0,826	0,493	0,9828
	BO-MSC-SG FD	NCT dissous dans $CHCl_3$	1897	1,00	3,11e-3	0,105	0,9995	1,00	1,45e-3	0,113	0,9995	0,986	0,107	0,0769	0,9996
		NCT après traitement	1897	0,999	6,81e-3	0,156	0,9990	0,998	0,0112	0,168	0,9989	1,02	-0,0809	0,147	0,9985
	BO-MSC-SG SD	NCT dissous dans $CHCl_3$	1897	1,00	5,74e-3	0,143	0,9991	1,00	3,77e-3	0,152	0,9991	0,994	0,0668	0,0753	0,9996
		NCT après traitement	1897	0,999	7,88e-3	0,168	0,9988	0,998	0,0130	0,180	0,9987	1,01	-0,0435	0,142	0,9986
	DT-LBC-BO-MSC	NCT dissous dans $CHCl_3$	1919	1,00	2,22e-3	0,0891	0,9997	1,00	2,37e-3	0,0962	0,9996	0,966	0,346	0,177	0,9978
		NCT après traitement	1919	0,998	0,0105	0,194	0,9984	0,998	0,0142	0,205	0,9984	0,931	0,707	0,468	0,9845
	DT-BO-SG FD-MSC	NCT dissous dans $CHCl_3$	1897	1,00	2,88e-3	0,101	0,9996	1,00	1,21e-3	0,108	0,9995	0,986	0,108	0,0699	0,9997
		NCT après traitement	1897	0,999	6,54e-3	0,153	0,9990	0,998	0,0109	0,164	0,9989	1,02	-0,0819	0,141	0,9986
	DT-BO-SG SD-MSC	NCT dissous dans $CHCl_3$	1897	0,999	4,54e-3	0,128	0,9993	0,999	2,48e-3	0,135	0,9993	0,994	0,0669	0,0393	0,9999
		NCT après traitement	1897	0,999	7,49e-3	0,164	0,9989	0,998	0,0126	0,175	0,9988	1,01	-0,0445	0,131	0,9988
Région 7 739,5-689,5	Aucun	NCT dissous dans $CHCl_3$	105	0,997	0,0213	0,276	0,9968	0,997	0,0218	0,295	0,9966	0,967	0,211	0,283	0,9943
		NCT après traitement	105	0,995	0,0330	0,344	0,9950	0,994	0,0390	0,367	0,9947	1,03	-0,141	0,358	0,9909

	BO	NCT dissous dans $CHCl_3$	105	0,997	0,0195	0,264	0,9971	0,997	0,0200	0,282	0,9969	0,975	0,123	0,255	0,9954
		NCT après traitement	105	0,996	0,0265	0,308	0,9960	0,995	0,0322	0,329	0,9958	1,02	-0,118	0,337	0,9920
	BO-MSC	NCT dissous dans $CHCl_3$	105	1,00	1,39e-3	0,0706	0,9998	1,00	7,91e-4	0,0744	0,9998	0,978	0,105	0,108	0,9992
		NCT après traitement	105	**0,999**	**5,85e-3**	**0,145**	**0,9991**	**0,998**	**9,96e-3**	**0,155**	**0,9991**	**1,02**	**-0,139**	**0,126**	**0,9989**
	LBC-BO-EMSC-SGS	NCT dissous dans $CHCl_3$	83	1,00	1,51e-3	0,0736	0,9998	1,00	6,79e-4	0,0778	0,9998	0,982	0,0739	0,0992	0,9993
		NCT après traitement	83	0,999	6,33e-3	0,151	0,9991	0,998	0,0106	0,161	0,9990	1,02	-0,140	0,138	0,9987
	BO-MSC-SG FD	NCT dissous dans $CHCl_3$	83	1,00	2,37e-3	0,0922	0,9996	1,00	4,04e-4	0,0978	0,9996	0,993	-6,18e-4	0,0705	0,9996
		NCT après traitement	83	0,999	8,32e-3	0,173	0,9988	0,998	0,0126	0,184	0,9987	1,02	-0,147	0,134	0,9987
	BO-MSC-SG SD	NCT dissous dans $CHCl_3$	83	1,00	6,09e-3	0,148	0,9991	1,00	3,32e-3	0,158	0,9990	1,01	-0,0878	0,0770	0,9996
		NCT après traitement	83	0,998	0,0124	0,211	0,9981	0,998	0,0173	0,224	0,9980	1,02	-0,0804	0,0884	0,9994
	DT-LBC-BO-MSC	NCT dissous dans $CHCl_3$	105	1,00	2,61e-3	0,0966	0,9996	1,00	4,11e-4	0,102	0,9996	0,986	0,0551	0,0760	0,9996
		NCT après traitement	105	0,999	7,93e-3	0,169	0,9988	0,998	0,0123	0,180	0,9987	1,03	-0,170	0,146	0,9985
	DT-BO-SG FD-MSC	NCT dissous dans $CHCl_3$	83	1,00	2,92e-3	0,102	0,9996	1,00	4,44e-4	0,109	0,9995	0,996	-0,0199	0,0579	0,9998
		NCT après traitement	83	0,999	9,24e-3	0,182	0,9986	0,998	0,0136	0,194	0,9985	1,02	-0,139	0,123	0,9989
	DT-BO-SG SD-MSC	NCT dissous dans $CHCl_3$	83	1,00	4,70e-3	0,130	0,9993	1,00	1,83e-3	0,139	0,9992	1,01	-0,0873	0,0481	0,9998
		NCT après traitement	83	0,998	0,0115	0,203	0,9983	0,998	0,0164	0,216	0,9982	1,02	-0,0873	0,0865	0,9995
Région 4 1065,5-1004,3	Aucun	NCT dissous dans $CHCl_3$	128	0,997	0,0218	0,279	0,9967	0,996	0,0222	0,299	0,9965	0,996	0,282	0,259	0,9953
		NCT après traitement	128	0,995	0,0309	0,333	0,9954	0,995	0,0348	0,353	0,9951	1,00	0,0651	0,351	0,9913
	BO	NCT dissous dans $CHCl_3$	128	0,997	0,0200	0,267	0,9970	0,996	0,0206	0,287	0,9968	0,979	0,194	0,228	0,9963
		NCT après traitement	128	0,995	0,0339	0,349	0,9949	0,994	0,0395	0,372	0,9946	1,02	-0,0838	0,321	0,9927
	BO-MSC	NCT dissous dans $CHCl_3$	128	0,999	5,37e-3	0,139	0,9992	0,999	5,14e-3	0,150	0,9991	0,981	0,180	0,106	0,9992
		NCT après traitement	128	0,999	7,66e-3	0,166	0,9989	0,998	0,0111	0,179	0,9988	1,02	-0,111	0,154	0,9983
	LBC-BO-EMSC-SGS	NCT dissous dans $CHCl_3$	106	0,999	4,87e-3	0,132	0,9993	0,999	4,65e-3	0,143	0,9992	0,977	0,195	0,116	0,9991
		NCT après traitement	106	0,999	6,70e-3	0,155	0,9990	0,998	9,79e-3	0,167	0,9989	1,03	-0,147	0,174	0,9979

	BO-MSC-SG FD	NCT dissous dans $CHCl_3$	106	0,999	4,14e-3	0,122	0,9994	0,999	3,24e-3	0,131	0,9993	0,981	0,135	0,0826	0,9995
		NCT après traitement	106	0,999	7,51e-3	0,164	0,9989	0,998	0,0114	0,176	0,9988	1,01	-0,0915	0,193	0,9974
	BO-MSC-SG SD	NCT dissous dans $CHCl_3$	106	0,999	4,03e-3	0,120	0,9994	0,999	2,86e-3	0,129	0,9994	0,982	0,125	0,0890	0,9994
		NCT après traitement	106	0,999	8,45e-3	0,174	0,9987	0,998	0,0129	0,187	0,9986	1,01	-0,0426	0,192	0,9974
	DT-LBC-BO-MSC	NCT dissous dans $CHCl_3$	128	0,999	4,69e-3	0,130	0,9993	0,999	3,68e-3	0,140	0,9992	0,971	0,198	0,110	0,9991
		NCT après traitement	128	0,999	6,90e-3	0,157	0,9990	0,998	0,0109	0,169	0,9989	1,02	-0,108	0,224	0,9964
	DT-BO-SG FD-MSC	NCT dissous dans $CHCl_3$	106	0,999	3,49e-3	0,112	0,9995	0,999	2,34e-3	0,120	0,9994	0,982	0,114	0,0667	0,9997
		NCT après traitement	106	0,999	7,45e-3	0,163	0,9989	0,998	0,0115	0,175	0,9988	1,01	-0,0509	0,198	0,9972
	DT-BO-SG SD-MSC	NCT dissous dans $CHCl_3$	106	1,00	3,33e-3	0,109	0,9995	0,999	2,15e-3	0,117	0,9995	0,982	0,121	0,0680	0,9997
		NCT après traitement	106	0,999	7,94e-3	0,169	0,9988	0,998	0,0123	0,181	0,9987	1,01	-0,0450	0,190	0,9974
Région 5 939,2-863,5	Aucun	NCT dissous dans $CHCl_3$	158	0,995	0,0307	0,332	0,9954	0,995	0,0323	0,354	0,9951	0,957	0,262	0,319	0,9928
		NCT après traitement	158	0,996	0,0292	0,323	0,9956	0,995	0,0327	0,344	0,9954	0,983	0,196	0,411	0,9880
	BO	NCT dissous dans $CHCl_3$	158	0,995	0,0318	0,337	0,9952	0,995	0,0339	0,362	0,9949	0,981	0,105	0,262	0,9951
		NCT après traitement	158	0,995	0,0339	0,349	0,9949	0,994	0,0391	0,371	0,9946	1,01	-0,0380	0,370	0,9903
	BO-MSC	NCT dissous dans $CHCl_3$	158	0,999	6,22e-3	0,149	0,9991	0,999	6,77e-3	0,159	0,9990	0,985	0,0793	0,121	0,9990
		NCT après traitement	158	0,999	6,89e-3	0,157	0,9990	0,998	9,87e-3	0,169	0,9989	1,01	-0,0653	0,184	0,9976
	LBC-BO-EMSC-SGS	NCT dissous dans $CHCl_3$	136	0,999	7,78e-3	0,167	0,9988	0,998	8,71e-3	0,178	0,9988	0,964	0,209	0,174	0,9979
		NCT après traitement	136	0,999	7,81e-3	0,167	0,9988	0,998	0,0111	0,180	0,9987	1,02	-0,138	0,226	0,9964
	BO-MSC-SG FD	NCT dissous dans $CHCl_3$	136	0,999	6,65e-3	0,154	0,9990	0,998	7,89e-3	0,166	0,9989	0,971	0,141	0,174	0,9978
		NCT après traitement	136	0,998	0,0112	0,201	0,9983	0,997	0,0160	0,217	0,9982	1,02	-0,132	0,227	0,9963
	BO-MSC-SG SD	NCT dissous dans $CHCl_3$	136	0,999	8,28e-3	0,172	0,9988	0,998	0,0112	0,186	0,9986	0,973	0,115	0,200	0,9972
		NCT après traitement	136	0,998	0,0106	0,195	0,9984	0,997	0,0180	0,210	0,9983	0,993	0,0993	0,152	0,9984
	DT-LBC-BO-MSC	NCT dissous dans $CHCl_3$	158	0,999	5,04e-3	0,134	0,9992	0,999	6,11e-3	0,144	0,9992	0,966	0,180	0,149	0,9984
		NCT après traitement	158	0,999	5,61e-3	0,142	0,9992	0,999	7,59e-3	0,150	0,9991	0,965	0,284	0,218	0,9966

	DT-BO-SG FD-MSC	NCT dissous dans $CHCl_3$	136	0,999	5,67e-3	0,143	0,9991	0,999	6,70e-3	0,154	0,9991	0,969	0,137	0,185	0,9976
		NCT après traitement	136	0,998	0,0119	0,207	0,9982	0,997	0,0173	0,223	0,9980	1,02	-0,0635	0,202	0,9971
	DT-BO-SG SD-MSC	NCT dissous dans $CHCl_3$	136	0,999	5,22e-3	0,137	0,9992	0,998	8,23e-3	0,148	0,9991	0,974	0,112	0,179	0,9977
		NCT après traitement	136	0,999	8,65e-3	0,176	0,9987	0,997	0,0160	0,190	0,9986	0,995	0,0887	0,129	0,9988
	Aucun	NCT dissous dans $CHCl_3$	288	0,997	0,0227	0,285	0,9966	0,996	0,0217	0,305	0,9964	0,965	0,404	0,294	0,9939
		NCT après traitement	288	0,996	0,0295	0,325	0,9956	0,994	0,0336	0,348	0,9953	0,967	0,320	0,406	0,9883
	BO	NCT dissous dans $CHCl_3$	288	0,997	0,0194	0,264	0,9971	0,997	0,0188	0,283	0,9969	0,965	0,337	0,227	0,9963
		NCT après traitement	288	0,995	0,0322	0,340	0,9952	0,994	0,0386	0,364	0,9948	1,02	-0,0678	0,334	0,9921
	BO-MSC	NCT dissous dans $CHCl_3$	288	0,999	5,38e-3	0,139	0,9992	0,999	4,12e-3	0,149	0,9991	0,968	0,324	0,158	0,9982
		NCT après traitement	288	0,998	0,0110	0,199	0,9983	0,997	0,0156	0,215	0,9982	1,02	-0,0893	0,172	0,9979
	LBC-BO-EMSC-SGS	NCT dissous dans $CHCl_3$	266	0,999	6,51e-3	0,153	0,9990	0,999	5,31e-3	0,164	0,9989	0,978	0,268	0,156	0,9983
		NCT après traitement	266	0,998	0,0107	0,196	0,9984	0,997	0,0143	0,212	0,9982	0,995	0,0726	0,170	0,9979
Région 3 1273,8-1135,4	BO-MSC-SG FD	NCT dissous dans $CHCl_3$	266	0,999	6,97e-3	0,158	0,9990	0,999	5,73e-3	0,169	0,9989	0,975	0,269	0,177	0,9978
		NCT après traitement	266	0,998	0,0112	0,200	0,9983	0,997	0,0159	0,216	0,9982	1,01	-0,0142	0,186	0,9975
	BO-MSC-SG SD	NCT dissous dans $CHCl_3$	266	0,999	7,76e-3	0,167	0,9988	0,998	7,31e-3	0,178	0,9988	0,972	0,260	0,213	0,9968
		NCT après traitement	266	0,998	0,0137	0,222	0,9979	0,996	0,0201	0,239	0,9978	0,997	0,109	0,208	0,9969
	DT-LBC-BO-MSC	NCT dissous dans $CHCl_3$	288	0,999	3,71e-3	0,115	0,9994	0,999	2,86e-3	0,124	0,9994	0,963	0,291	0,194	0,9973
		NCT après traitement	288	0,999	7,50e-3	0,164	0,9989	0,998	0,0105	0,177	0,9988	0,994	0,0733	0,166	0,9980
	DT-BO-SG FD-MSC	NCT dissous dans $CHCl_3$	266	0,999	6,67e-3	0,155	0,9990	0,999	5,49e-3	0,166	0,9989	0,972	0,279	0,174	0,9979
		NCT après traitement	266	0,998	0,0115	0,203	0,9983	0,997	0,0165	0,219	0,9981	1,01	-0,0353	0,190	0,9974
	DT-BO-SG SD-MSC	NCT dissous dans $CHCl_3$	266	0,999	6,37e-3	0,151	0,9990	0,999	5,90e-3	0,161	0,9990	0,972	0,263	0,170	0,9980
		NCT après traitement	266	0,998	0,0130	0,216	0,9980	0,996	0,0194	0,233	0,9979	0,996	0,117	0,182	0,9977
Région 2 1332,8-1299,3	Aucun	NCT dissous dans $CHCl_3$	70	0,996	0,0265	0,308	0,9960	0,995	0,0251	0,329	0,9958	0,958	0,492	0,324	0,9925
		NCT après traitement	70	0,996	0,0253	0,301	0,9962	0,995	0,0288	0,323	0,9959	0,936	0,593	0,470	0,9843

	BO	NCT dissous dans $CHCl_3$	70	0,996	0,0238	0,292	0,9964	0,996	0,0225	0,311	0,9962	0,968	0,330	0,237	0,9960
		NCT après traitement	70	0,994	0,0367	0,363	0,9945	0,993	0,0427	0,388	0,9941	0,995	0,0884	0,370	0,9903
	BO-MSC	NCT dissous dans $CHCl_3$	70	0,998	0,0115	0,203	0,9983	0,998	9,64e-3	0,217	0,9982	0,970	0,318	0,167	0,9980
		NCT après traitement	70	0,998	0,0161	0,240	0,9976	0,996	0,0204	0,258	0,9974	0,998	0,0679	0,219	0,9966
	LBC-BO-EMSC-SGS	NCT dissous dans $CHCl_3$	48	0,998	0,0129	0,215	0,9981	0,998	0,0111	0,229	0,9979	0,970	0,336	0,180	0,9977
		NCT après traitement	48	**0,999**	**9,86e-3**	**0,188**	**0,9985**	**0,997**	**0,0137**	**0,203**	**0,9984**	**1,01**	**-0,0572**	**0,204**	**0,9971**
	BO-MSC-SG FD	NCT dissous dans $CHCl_3$	48	0,998	0,0137	0,221	0,9980	0,997	0,0115	0,236	0,9978	0,974	0,295	0,163	0,9981
		NCT après traitement	48	0,999	9,58e-3	0,185	0,9986	0,997	0,0135	0,199	0,9984	1,01	-0,0234	0,204	0,9970
	BO-MSC-SG SD	NCT dissous dans $CHCl_3$	48	0,998	0,0143	0,226	0,9979	0,998	0,0116	0,240	0,9977	0,981	0,225	0,133	0,9987
		NCT après traitement	48	0,999	9,07e-3	0,180	0,9986	0,997	0,0131	0,193	0,9985	0,999	7,49e-3	0,219	0,9966
	DT-LBC-BO-MSC	NCT dissous dans $CHCl_3$	70	0,998	0,0103	0,193	0,9984	0,999	6,98e-3	0,204	0,9984	0,984	0,146	0,0877	0,9995
		NCT après traitement	70	0,998	0,0101	0,190	0,9985	0,997	0,0137	0,203	0,9984	0,986	0,0681	0,280	0,9944
	DT-BO-SG FD-MSC	NCT dissous dans $CHCl_3$	48	0,998	0,0132	0,217	0,9980	0,998	0,0101	0,231	0,9979	0,986	0,187	0,105	0,9992
		NCT après traitement	48	0,999	8,70e-3	0,177	0,9987	0,998	0,0128	0,189	0,9986	0,997	0,0115	0,225	0,9964
	DT-BO-SG SD-MSC	NCT dissous dans $CHCl_3$	48	0,998	0,0143	0,226	0,9979	0,998	0,0114	0,241	0,9977	0,981	0,218	0,121	0,9990
		NCT après traitement	48	0,999	8,92e-3	0,179	0,9987	0,997	0,0131	0,192	0,9986	0,999	0,0123	0,215	0,9967
Région 6 837,4-781,0	Aucun	NCT dissous dans $CHCl_3$	118	0,996	0,0285	0,319	0,9957	0,995	0,0303	0,341	0,9955	0,963	0,217	0,295	0,9938
		NCT après traitement	118	0,995	0,0319	0,338	0,9952	0,995	0,0358	0,359	0,9950	1,01	0,0330	0,367	0,9904
	BO	NCT dissous dans $CHCl_3$	118	0,996	0,0247	0,298	0,9963	0,996	0,0271	0,317	0,9961	0,968	0,108	0,304	0,9935
		NCT après traitement	118	0,995	0,0367	0,362	0,9945	0,994	0,0416	0,385	0,9942	1,03	-0,225	0,392	0,9891
	BO-MSC	NCT dissous dans $CHCl_3$	118	0,999	5,56e-3	0,141	0,9992	0,999	6,74e-3	0,149	0,9991	0,971	0,0894	0,178	0,9978
		NCT après traitement	118	0,999	7,14e-3	0,160	0,9989	0,998	9,68e-3	0,172	0,9988	1,04	-0,256	0,244	0,9958
	LBC-BO-EMSC-SGS	NCT dissous dans $CHCl_3$	96	0,999	4,78e-3	0,131	0,9993	0,999	5,53e-3	0,139	0,9992	0,978	0,0452	0,146	0,9985
		NCT après traitement	96	0,999	6,90e-3	0,157	0,9990	0,999	8,89e-3	0,168	0,9989	1,02	-0,0948	0,187	0,9975

	BO-MSC-SG FD	NCT dissous dans $CHCl_3$	96	0,999	3,61e-3	0,114	0,9995	0,999	4,01e-3	0,120	0,9994	0,983	0,0584	0,0922	0,9994
		NCT après traitement	96	0,999	7,06e-3	0,159	0,9989	0,998	9,92e-3	0,170	0,9989	1,02	-0,187	0,244	0,9958
	BO-MSC-SG SD	NCT dissous dans $CHCl_3$	96	0,999	6,04e-3	0,147	0,9991	0,999	6,26e-3	0,154	0,9991	0,999	-0,0768	0,126	0,9989
		NCT après traitement	96	0,999	8,55e-3	0,175	0,9987	0,998	0,0121	0,188	0,9986	0,987	0,0560	0,224	0,9965
	DT-LBC-BO-MSC	NCT dissous dans $CHCl_3$	118	0,999	5,17e-3	0,136	0,9992	0,999	6,51e-3	0,142	0,9992	0,985	0,0307	0,0950	0,9994
		NCT après traitement	118	0,999	7,45e-3	0,163	0,9989	0,999	9,30e-3	0,174	0,9988	1,05	-0,427	0,365	0,9906
	DT-BO-SG FD-MSC	NCT dissous dans $CHCl_3$	96	1,00	2,78e-3	0,0997	0,9996	0,999	2,62e-3	0,105	0,9996	0,985	0,0595	0,0724	0,9996
		NCT après traitement	96	0,999	7,25e-3	0,161	0,9989	0,998	0,0106	0,172	0,9988	1,03	-0,260	0,296	0,9938
	DT-BO-SG SD-MSC	NCT dissous dans $CHCl_3$	96	0,999	3,61e-3	0,114	0,9995	0,999	3,75e-3	0,119	0,9994	0,998	-0,0576	0,0774	0,9996
		NCT après traitement	96	0,999	7,16e-3	0,160	0,9989	0,998	0,0107	0,171	0,9989	0,990	0,0389	0,215	0,9967
Région 1 1500,0-1402,5	Aucun	NCT dissous dans $CHCl_3$	204	0,995	0,0340	0,349	0,9949	0,994	0,0324	0,372	0,9946	0,954	0,632	0,402	0,9885
		NCT après traitement	204	0,992	0,0559	0,448	0,9916	0,993	0,0566	0,482	0,9909	0,876	0,886	0,697	0,9656
	BO	NCT dissous dans $CHCl_3$	204	0,996	0,0293	0,324	0,9956	0,995	0,0285	0,345	0,9953	0,976	0,448	0,343	0,9917
		NCT après traitement	204	0,994	0,0379	0,369	0,9943	0,995	0,0401	0,396	0,9939	0,898	0,826	0,579	0,9762
	BO-MSC	NCT dissous dans $CHCl_3$	204	0,997	0,0175	0,250	0,9974	0,997	0,0163	0,266	0,9972	0,977	0,437	0,295	0,9938
		NCT après traitement	204	0,998	0,0131	0,217	0,9980	0,999	0,0140	0,232	0,9979	0,902	0,804	0,480	0,9837
	LBC-BO-EMSC-SGS	NCT dissous dans $CHCl_3$	182	0,997	0,0213	0,276	0,9968	0,996	0,0198	0,293	0,9966	0,973	0,452	0,292	0,9940
		NCT après traitement	182	0,998	0,0122	0,209	0,9982	0,999	0,0130	0,223	0,9980	0,923	0,550	0,336	0,9920
	BO-MSC-SG FD	NCT dissous dans $CHCl_3$	182	0,996	0,0251	0,300	0,9962	0,996	0,0231	0,317	0,9961	0,974	0,406	0,254	0,9954
		NCT après traitement	182	0,999	9,06e-3	0,180	0,9986	0,997	0,0144	0,195	0,9985	1,00	0,0453	0,143	0,9985
	BO-MSC-SG SD	NCT dissous dans $CHCl_3$	182	0,996	0,0265	0,308	0,9960	0,995	0,0247	0,326	0,9959	0,977	0,359	0,225	0,9964
		NCT après traitement	182	0,998	0,0139	0,223	0,9979	0,996	0,0199	0,240	0,9977	1,02	-0,0649	0,233	0,9961
	DT-LBC-BO-MSC	NCT dissous dans $CHCl_3$	204	0,997	0,0208	0,273	0,9969	0,996	0,0188	0,290	0,9967	0,966	0,460	0,266	0,9950
		NCT après traitement	204	0,998	0,0121	0,208	0,9982	0,997	0,0178	0,225	0,9980	0,983	0,341	0,230	0,9962

	DT-BO-SG FD-MSC	NCT dissous dans $CHCl_3$	182	0,996	0,0248	0,298	0,9963	0,996	0,0227	0,315	0,9961	0,974	0,392	0,239	0,9959
		NCT après traitement	182	0,998	0,0151	0,232	0,9977	0,996	0,0214	0,251	0,9975	1,02	-0,0840	0,210	0,9969
	DT-BO-SG SD-MSC	NCT dissous dans $CHCl_3$	182	0,996	0,0260	0,305	0,9961	0,996	0,0241	0,323	0,9959	0,978	0,359	0,220	0,9966
		NCT après traitement	182	0,998	0,0133	0,218	0,9980	0,996	0,0193	0,235	0,9978	1,02	-0,0655	0,210	0,9969
	Aucun	NCT dissous dans $CHCl_3$	175	0,997	0,0194	0,264	0,9971	0,997	0,0195	0,282	0,9969	0,966	0,248	0,272	0,9948
		NCT après traitement	175	0,995	0,0307	0,332	0,9954	0,994	0,0364	0,354	0,9951	1,01	-0,0224	0,348	0,9914
	BO	NCT dissous dans $CHCl_3$	175	0,997	0,0187	0,259	0,9972	0,997	0,0190	0,276	0,9970	0,975	0,139	0,247	0,9957
		NCT après traitement	175	0,996	0,0260	0,306	0,9961	0,995	0,0318	0,326	0,9958	1,01	-0,100	0,336	0,9920
	LBC-BO	NCT dissous dans $CHCl_3$	175	0,997	0,0188	0,259	0,9972	0,997	0,0190	0,277	0,9970	0,979	0,109	0,244	0,9958
		NCT après traitement	175	**0,996**	**0,0275**	**0,314**	**0,9959**	**0,995**	**0,0334**	**0,335**	**0,9956**	**1,02**	**-0,111**	**0,336**	**0,9920**
	BO-MSC	NCT dissous dans $CHCl_3$	175	1,00	1,34e-3	0,0692	0,9998	1,00	6,09e-4	0,0734	0,9998	0,997	0,122	0,105	0,9992
		NCT après traitement	175	0,999	5,56e-3	0,141	0,9992	0,998	9,68e-3	0,151	0,9991	1,02	-0,121	0,131	0,9988
	LBC-BO-EMSC-SGS	NCT dissous dans $CHCl_3$	131	1,00	1,51e-3	0,0735	0,9998	1,00	5,67e-4	0,0782	0,9998	0,981	0,0939	0,0949	0,9994
Régions 2+7 1332,8-1299,3 739,5-689,5		NCT après traitement	131	**0,999**	**6,03e-3**	**0,147**	**0,9991**	**0,998**	**0,0102**	**0,158**	**0,9990**	**1,02**	**-0,133**	**0,142**	**0,9986**
	BO-MSC-SG FD	NCT dissous dans $CHCl_3$	131	1,00	2,34e-3	0,0915	0,9996	1,00	3,38e-4	0,0973	0,9996	0,992	0,0195	0,0653	0,9997
		NCT après traitement	131	0,999	7,68e-3	0,166	0,9988	0,998	0,0119	0,177	0,9988	1,02	-0,138	0,137	0,9987
	BO-MSC-SG SD	NCT dissous dans $CHCl_3$	131	0,999	5,84e-3	0,145	0,9991	1,00	3,08e-3	0,154	0,9991	1,01	-0,0749	0,0718	0,9996
		NCT après traitement	131	0,998	0,0116	0,204	0,9983	0,998	0,0164	0,217	0,9982	1,01	-0,077	0,0909	0,9994
	DT-LBC-BO-MSC	NCT dissous dans $CHCl_3$	175	1,00	2,80e-3	0,100	0,9996	1,00	5,54e-4	0,106	0,9996	0,986	0,0570	0,0728	0,9996
		NCT après traitement	175	0,999	7,66e-3	0,166	0,9986	0,998	0,0120	0,177	0,9988	1,02	-0,160	0,147	0,9985
	DT-BO-SG FD-MSC	NCT dissous dans $CHCl_3$	131	1,00	2,93e-3	0,102	0,9996	1,00	4,42e-4	0,109	0,9995	0,996	-0,0153	0,0555	0,9998
		NCT après traitement	131	0,999	8,97e-3	0,179	0,9987	0,998	0,0133	0,191	0,9986	1,02	-0,135	0,124	0,9989
	DT-BO-SG SD-MSC	NCT dissous dans $CHCl_3$	131	0,999	4,57e-3	0,128	0,9993	1,00	1,71e-3	0,137	0,9993	1,01	-0,0752	0,0423	0,9999
		NCT après traitement	131	0,998	0,0108	0,197	0,9984	0,998	0,0156	0,209	0,9983	1,02	-0,0834	0,0893	0,9994

	Aucun	NCT dissous dans $CHCl_3$	391	0,997	0,0216	0,278	0,9968	0,997	0,0222	0,297	0,9966	0,966	0,234	0,277	0,9946
		NCT après traitement	391	0,995	0,0309	0,309	0,9954	0,995	0,0361	0,354	0,9951	1,02	-0,0489	0,355	0,9911
	BO	NCT dissous dans $CHCl_3$	391	0,997	0,0194	0,264	0,9971	0,997	0,0201	0,282	0,9969	0,977	0,137	0,243	0,9958
		NCT après traitement	391	0,996	0,0276	0,315	0,9959	0,995	0,0332	0,336	0,9956	1,02	-0,105	0,333	0,9921
	BO-MSC	NCT dissous dans $CHCl_3$	391	1,00	1,94e-3	0,0833	0,9997	1,00	1,54e-3	0,0886	0,9997	0,979	0,119	0,0990	0,9993
		NCT après traitement	391	0,999	5,61e-3	0,142	0,9992	0,998	9,48e-3	0,152	0,9991	1,02	-0,127	0,134	0,9987
	LBC-BO-EMSC-SGS	NCT dissous dans $CHCl_3$	325	1,00	2,11e-3	0,0870	0,9997	1,00	1,59e-3	0,0927	0,9997	0,979	0,113	0,102	0,9993
		NCT après traitement	325	0,999	5,78e-3	0,144	0,9991	0,998	9,67e-3	0,155	0,9991	1,02	-0,142	0,152	0,9984
Régions 4+5+7 1065,5-1004,3 939,2-863,5 739,5-689,5	BO-MSC-SG FD	NCT dissous dans $CHCl_3$	325	1,00	2,25e-3	0,0897	0,9997	1,00	5,86e-4	0,0952	0,9996	0,990	0,0277	0,0727	0,9996
		NCT après traitement	325	0,999	7,65e-3	0,166	0,9989	0,998	0,0119	0,177	0,9988	1,02	-0,137	0,145	0,9985
	BO-MSC-SG SD	NCT dissous dans $CHCl_3$	325	0,999	4,32e-3	0,124	0,9994	1,00	2,04e-3	0,132	0,9993	1,00	-0,0355	0,0706	0,9996
		NCT après traitement	325	0,998	0,0103	0,192	0,9985	0,998	0,0151	0,205	0,9984	1,01	-0,0700	0,106	0,9992
	DT-LBC-BO-MSC	NCT dissous dans $CHCl_3$	391	1,00	2,45e-3	0,0938	0,9996	1,00	7,15e-4	0,0998	0,9996	0,981	0,0955	0,0810	0,9995
		NCT après traitement	391	0,999	6,62e-3	0,154	0,9990	0,998	0,0108	0,165	0,9989	1,02	-0,135	0,158	0,9982
	DT-BO-SG FD-MSC	NCT dissous dans $CHCl_3$	325	1,00	2,42e-3	0,0932	0,9996	1,00	3,05e-4	0,0990	0,9996	0,993	7,40e-3	0,0573	0,9998
		NCT après traitement	325	0,999	8,12e-3	0,171	0,9988	0,998	0,0125	0,182	0,9987	1,02	-0,123	0,134	0,9987
	DT-BO-SG SD-MSC	NCT dissous dans $CHCl_3$	325	1,00	3,16e-3	0,106	0,9995	1,00	8,24e-4	0,113	0,9995	1,00	-0,0367	0,0338	0,9999
		NCT après traitement	325	0,999	9,44e-3	0,184	0,9986	0,998	0,0142	0,196	0,9985	1,01	-0,0763	0,103	0,9992

Les meilleurs modèles analytiques sont indiqués en gras et colorés.

Abréviations : BO, Correction du décalage de la ligne de base ; DT, Dé-tendance avec ordre polynomial (PO) = 2 ; LBC, Correction linéaire de la ligne de base ; MSC, Correction de la diffusion multiplicative ; EMSC, Correction étendue de la diffusion multiplicative ; NCT, Nicotine ; SGS, Lissage Savitzky-Golay à 11 points / côté et PO=2 ; SG FD, 1ère dérivée de Savitzky-Golay à 11 points / côté et PO=2 ; SG SD, 2ème dérivée de Savitzky-Golay à 11 points / côté et PO=2.

Les modèles correspondant aux hauteurs des pics à 716, 903, 1025, 1190 et 1315 cm^{-1} présentaient des RMSE inférieures à 0,500 et des coefficients de corrélation supérieurs à 0,9960. Malgré les valeurs d'erreur élevées, ces modèles pouvaient être acceptés pour une telle technique d'échantillonnage (film sec mince) à une gamme de concentrations aussi large. La Figure 2.4A fournit une comparaison visuelle à l'aide d'un diagramme radar, inspiré d'une étude antérieure [170], des résultats les plus pertinents (hauteurs à 716 et 1315 cm^{-1}). Ce résultat est cohérent avec celui rapporté dans l'étude précédente [156] sur la même bande (1315 cm^{-1}).

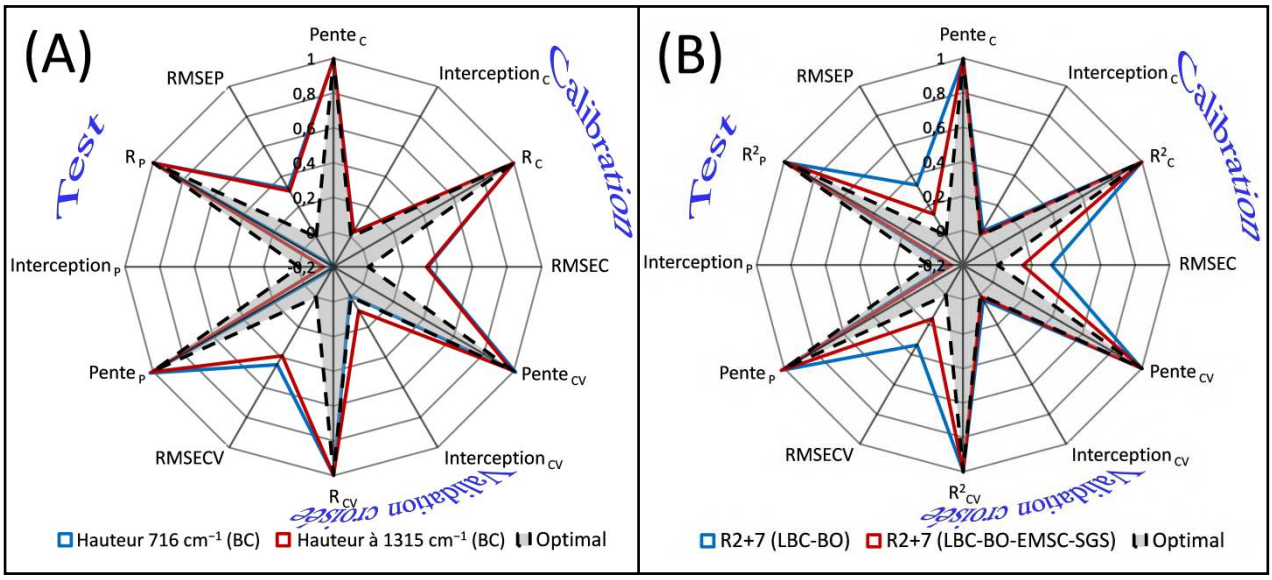


Figure 2.4 : Diagrammes radiaux comparant les performances des modèles optimisés univariés (A) et PLSR (B). Les valeurs optimales préétablies pour l'interception et le RMSE sont 0, et pour la pente et le R^2 (ou R) sont 1.

D'autre part, étant donné qu'un seul composant a été considéré dans la calibration multivariée, un seul facteur a suffi pour décrire les données dans tous les modèles PLSR. Ces modèles sont donc mieux caractérisés par le nombre de variables utilisées que par les variables latentes. Dans l'ensemble, par rapport aux modèles univariés, les modèles PLSR ont fourni des paramètres de qualité nettement supérieure (Figure 2.4B), ce qui peut s'expliquer par les prétraitements spectraux avancés appliqués à chaque région réduite, minimisant les valeurs d'erreur de 2 à 3 fois, tandis que certains R^2 ont atteint 0,9999.

Parmi tous les modèles PLSR développés, ceux de la région 1 (1500 - 1402 cm^{-1}) ont présenté les erreurs les plus élevées car ils contiennent un bruit irrégulier provenant du protocole d'extraction, et ont donc été les premiers à être exclus. La région 3 (1274 - 1135 cm^{-1}) comprend plusieurs bandes chevauchantes qui pourraient ne pas être différenciées ultérieurement des autres agents interférents lors de la prédiction, c'est pourquoi elle a également été éliminée. Bien que les régions 4 (1066 - 1004 cm^{-1}) et 6 (837 - 781 cm^{-1}) aient présenté de bons paramètres de régression, elles sont associées respectivement à des modes de vibration très courants des cycles aromatiques et des mouvements d'agitation symétriques / asymétriques des groupes CH_3 et CH_2, ce qui semble ne pas en faire d'excellents candidats pour une prédiction spécifique de la nicotine dans des matrices complexes. La région 5 (939 - 864 cm^{-1}) est spécifique et a donné de très bons paramètres, mais son intensité est relativement faible par rapport aux autres bandes et pourrait réduire la sensibilité de la méthode. Par la suite, seuls les modèles des régions 2 (1333 - 1299 cm^{-1}) et 7 (740 - 690 cm^{-1}), ainsi que leurs combinaisons, peuvent être étudiés en profondeur car; en outre, les deux régions semblent particulièrement être spécifiques à la nicotine.

L'arrangement de DT-BO-SG SD-MSC dans l'ordre, par exemple, a fourni la meilleure valeur de RMSEP (0,0865) dans la région 7. Cependant, le fait que cette correction peut réduire significativement la sensibilité, le nombre de variables à considérer et entraîner des valeurs d'erreurs de calibration et de prédiction non corrélées, impose la satisfaction avec des techniques de prétraitement moins complexes.

En conséquence, les régions 2 et 7 ont été combinées et prétraitées en utilisant des techniques simples. La combinaison de LBC-BO-EMSC-SGS a donné une pente de 0,999, une interception de 0,006, une RMSEC de 0,147 et un R^2 de 0,9991, garantissant un bon ajustement de la ligne de calibration aux données spectrales. Ce modèle a donc été évalué pour d'autres tests. Les régions spectrales 2+7 prétraitées et leurs droites de régression correspondantes sont visualisées dans les Figures 2.5A à D. À partir de la droite de calibration des données corrigées pour la ligne de base (Figure 2.5B), on remarque que la variance diminue lors de la dilution, ce qui confirme la précision de

l'échantillonnage et de l'analyse, et à partir des données prétraitées par EMSC (Figure 2.5D), la proximité des droites de régression des ensembles de calibration et de test entre elles illustre la grande capacité prédictive du modèle.

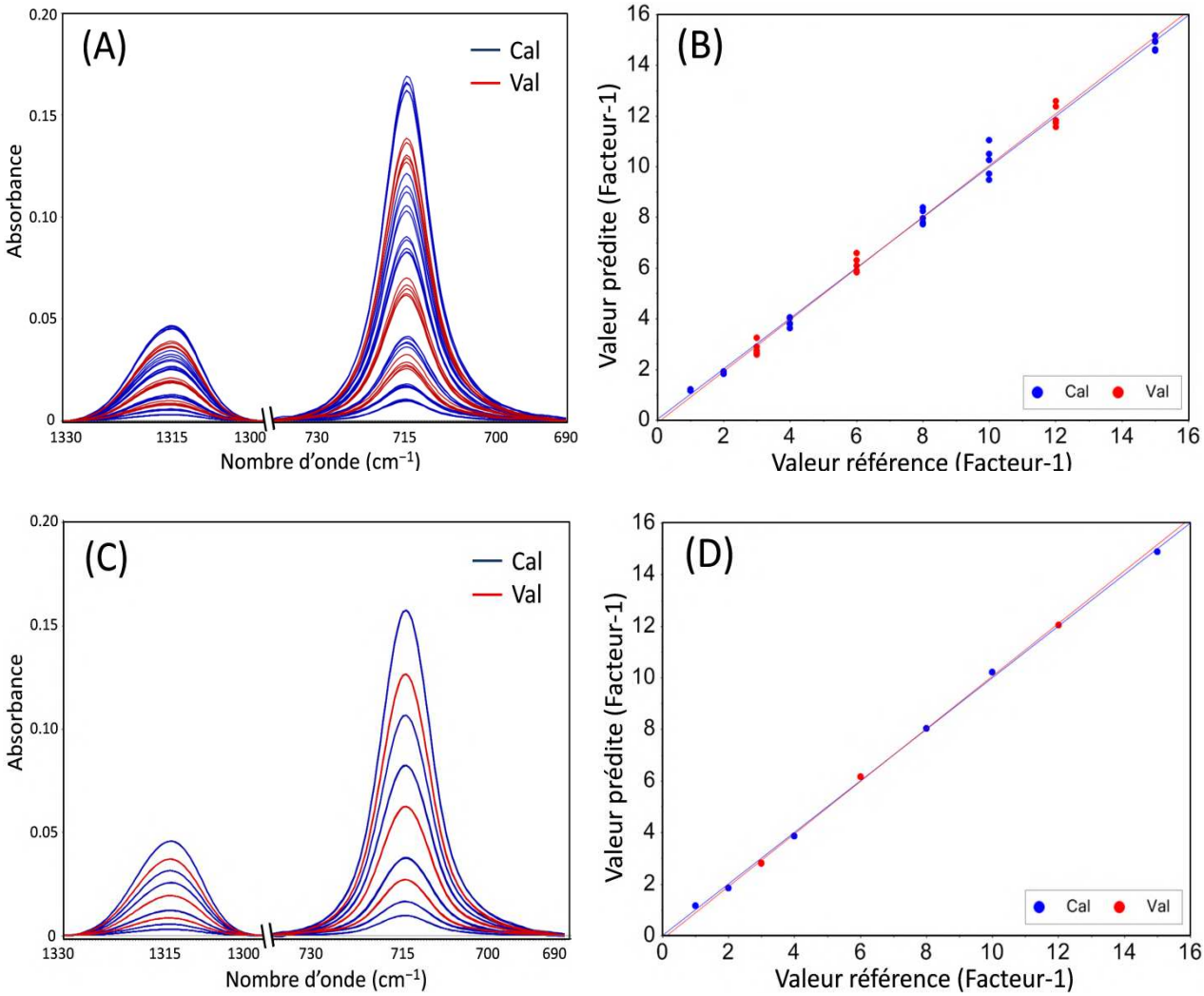


Figure 2.5 : Régions 2+7 prétraitées par LBC-BO (A) et par LBC-BO-EMSC-SGS (C) et leurs droites de régression prédites contre réelles (B) et (D) respectivement, calculées avec l'algorithme PLS-1 pour les ensembles de calibration et de validation.

2.3.3. Évaluation de la performance des modèles optimaux

Avant d'utiliser l'un des modèles présentés, ceux-ci ont été évalués à l'aide de paramètres statistiques supplémentaires estimés selon les dernières approches et directives proposées dans la littérature [61, 85, 86]. Le Tableau 2.4

présente les facteurs de mérite analytiques (FOMs) évaluées pendant la validation pour les modèles optimisés des méthodes univariée et multivariée.

Tableau 2.4 : Facteurs de mérite pour les modèles univariés et multivariés sélectionnés.

Méthode	**Loi de Beer-Lambert**		**PLS-1**	
Mesure spectrale	Hauteur à 716 cm^{-1}	Hauteur à 1315 cm^{-1}	Régions 2+7	Régions 2+7
Prétraitement	BC: 740-690 cm^{-1}	BC: 1333-1299 cm^{-1}	LBC-BO	LBC-BO-EMSC-SGS
Ensemble de calibration				
PRESS	4,0	3,9	3,4	0,75
Q^2	0,9944	0,9946	0,9953	0,9990
Test F (F_{exp}) [$F_{crit\ (0.05,\ 28,\ 24)}$ = 1,95]	1,10	1,23	1,00	1,24
Valeur p	0,41	0,30	0,50	0,72
SEN ($AU.ml.mg^{-1}$)	0,0114	0,00314	0,0456	0,0452
Anal. SEN ($ml.mg^{-1}$)	2,81	2,87	103	131
SEL	1,02	1,09	0,986	0.981
LOD [¶] ($mg.ml^{-1}$)	1,3	1,2	0,33 – 0,35	0,15 – 0,16
LOQ [§] ($mg.ml^{-1}$)	3,7	3,6	0,98 – 1,0	0,46 – 0,47
Ensemble de test				
RMSEP ($mg.ml^{-1}$)	0,32	0,31	0,34	0,14
REP (%)	4,8	4,6	5,0	2,1
RPD	12,1	12,7	11,5	27,1
RER	28,0	29,5	26,8	63,1
Biais ($mg.ml^{-1}$)	−0,063	−0,073	$-1{,}5\ 10^{-3}$	$-3{,}0\ 10^{-4}$
Récupération $_{moy}$ ± SD (%)	98,1 ± 5,6	97,8 ± 6,6	98,9 ± 6,8	98,8 ± 3,9
Test t (t_{exp}) [$t_{crit\ (0.025,\ 14)}$ = 2,15]	1,34	1,31	0,609	1,15

[¶] Limite de détection univariée (LOD_u) et limites de détection multivariées [LOD_{min} − LOD_{max}].

[§] Limite de quantification univariée (LOQ_u) et limites de quantification multivariées [LOQ_{min} − LOQ_{max}].

Abréviations : BC, correction simple de la ligne de base ; Régions 2+7, zones spectrales combinées: 1333 -1299 cm^{-1} + 740 - 690 cm^{-1}.

Commençant par corroborer la linéarité sur le domaine linéaire, nous avons calculé la valeur expérimentale de F, définie comme le rapport au carré de l'écart-type résiduel sur l'erreur pure, puis l'avons comparée à la valeur tabulée au seuil de signification α [86]. Tous les modèles sélectionnés ont donné $F_{exp} < F_{critique}$,

reconfirmant la linéarité des droites de calibration. Un paramètre connexe, le critère de Haaland [171], recherche un niveau de probabilité $p < 0,75$ pour la valeur de *F*. Ce test a été effectué pour valider que les modèles ne nécessitaient qu'une (1) seule variable latente pour prédire le composant unique dans la solution étalon même après sa soumission au protocole d'extraction.

Le calcul des valeurs de PRESS et Q^2 des échantillons qui ne sont pas utilisés pour l'estimation des paramètres de calibration a été effectué. En général, lorsque la valeur de PRESS est minime, cela signifie une capacité de prédiction plus élevée. Pour le deuxième indice, une grande différence entre le Q^2 et le R^2 montre que le modèle est sensible à la présence ou à l'absence de certains échantillons [78]. Les meilleures valeurs de PRESS (0,75) et de Q^2 (0,9990) ont été obtenues avec le modèle PLSR sur les régions 2+7 prétraitées par LBC-BO-EMSC-SGS, montrant une capacité prédictive nettement supérieure à celle de la simple régression par la loi de Beer-Lambert.

En calibration univariée classique, la sensibilité est la pente de la droite d'étalonnage [62], alors qu'en PLSR, elle correspond à l'inverse de la longueur du vecteur des coefficients de régression [61]. De plus, le rapport entre les sensibilités de l'analyte dans la matrice et dans sa forme pure est appelé sélectivité [86, 172]. Cette dernière n'est généralement pas rapportée dans les méthodes multivariées car aucune approximation n'est disponible pour le spectre de l'analyte pur dans le mélange [86]. Toutefois, ce n'est pas le cas ici, c'est pourquoi elle a été calculée pour prouver qu'aucune contamination ou perte ne sont survenues, à une mesure spectrale spécifique, après avoir soumis les étalons à la procédure d'extraction recommandée. Les sensibilités dans les modèles PLSR se sont avérées quatre fois supérieures, et plus, à celles des modèles univariés. Des valeurs de sensibilité analytique encore plus élevées ont été calculées pour la méthode PLSR, atteignant 131 $ml.mg^{-1}$, en raison du bruit minimal dans ces régions caractérisées par un grand nombre de variables spectrales. Les valeurs de sélectivité n'étaient pas très différentes de 1 dans les deux méthodes de régression, ce qui confirme l'adéquation qualitative de l'ensemble des étalons traités pour la quantification de la nicotine.

Il est important de signaler un FOM essentiel "la limite de détection (LOD)", reconnue comme la concentration minimale détectable d'un analyte, liée à un certain risque de fausses détections et de fausses non-détections (erreurs α et β) [61]. Selon une définition moderne proposée par Allegrini et Olivieri [173], la LOD en PLSR comporte deux limites: une limite minimale et une limite maximale, associées respectivement aux échantillons situés près du centre et aux extrémités de l'espace de calibration multivariée. La limite de quantification (LOQ), à son tour, peut être estimée sur la base de la même approche, correspondant à trois fois la LOD associée [86]. Les modèles univariés ont atteint des LOD de l'ordre de 1,2 et 1,3 $mg.ml^{-1}$ respectivement aux hauteurs 1315 et 716 cm^{-1}, alors que les modèles PLSR ont permis d'obtenir des LOD de 0,33 − 0,35 $mg.ml^{-1}$ pour le prétraitement LBC-BO et de 0,15 − 0,16 $mg.ml^{-1}$ pour LBC-BO-EMSC-SGS.

Enfin, à partir des statistiques obtenues à l'étape de prédiction, les valeurs suivantes ont été trouvées pour le modèle monovarié à la hauteur de 1315 cm^{-1}: RMSEP = 0,31 $mg.ml^{-1}$, REP = 4,6 %, RPD = 12,7 et RER = 29,5. Dans le cas des modèles PLS-1, les meilleures valeurs de RMSEP, REP, RPD et RER ont été obtenues avec le prétraitement spectral LBC-BO-EMSC-SGS des régions 2+7, soient respectivement 0,14 $mg.ml^{-1}$, 2,1 %, 27,1 et 63,1. Cela indique une excellente capacité (RPD > 8,1 [88] et RER > 20 [174]) du modèle correspondant à prédire la concentration de NCT sous les conditions proposées dans l'ensemble de test, à condition qu'il ne soit pas biaisé.

2.3.4. Evaluation de l'exactitude et de la précision

Cette étape consiste à évaluer l'exactitude et la précision des modèles les plus performants à l'aide de trois paramètres analytiques: le biais, le taux de récupération et l'analyse de la région de confiance elliptique conjointe (EJCR) pour l'ensemble de test.

La valeur moyenne de la différence entre les concentrations prédites et réelles pour les échantillons de test, appelée biais, a révélé les valeurs suivantes: −0,073, −0,063, −0,0015 et −0,0003 $mg.ml^{-1}$, encore une fois en faveur des modèles multivariés. Ces faibles valeurs suggèrent une distribution aléatoire des points autour de la ligne de régression pour les quatre modèles.

La récupération de l'analyte, généralement évaluée dans la méthode des ajouts dosés [86, 170], a néanmoins été estimée ici pour l'ensemble de test traité comme le rapport des concentrations prédites aux concentrations réelles. Pour tous les modèles, la récupération moyenne ne diffère pas statistiquement de 100 % puisque la valeur t_{exp} calculée est inférieur à la valeur tabulée; en accord avec les valeurs de sélectivité obtenues dans le Tableau 2.2. Ces résultats justifient, une fois de plus, le choix approprié de l'ensemble d'étalons traités pour établir les droites d'étalonnage en vue de quantifier la nicotine dans les échantillons commerciaux.

Maintenant, afin de vérifier davantage l'exactitude des modèles proposés et d'inspecter la présence de biais ou non, nous avons envisagé le test EJCR pour l'ensemble de données de test. La Figure 2.6A illustre le test EJCR réalisé sur les modèles les plus performants basés sur les moindres carrés ordinaires (OLS).

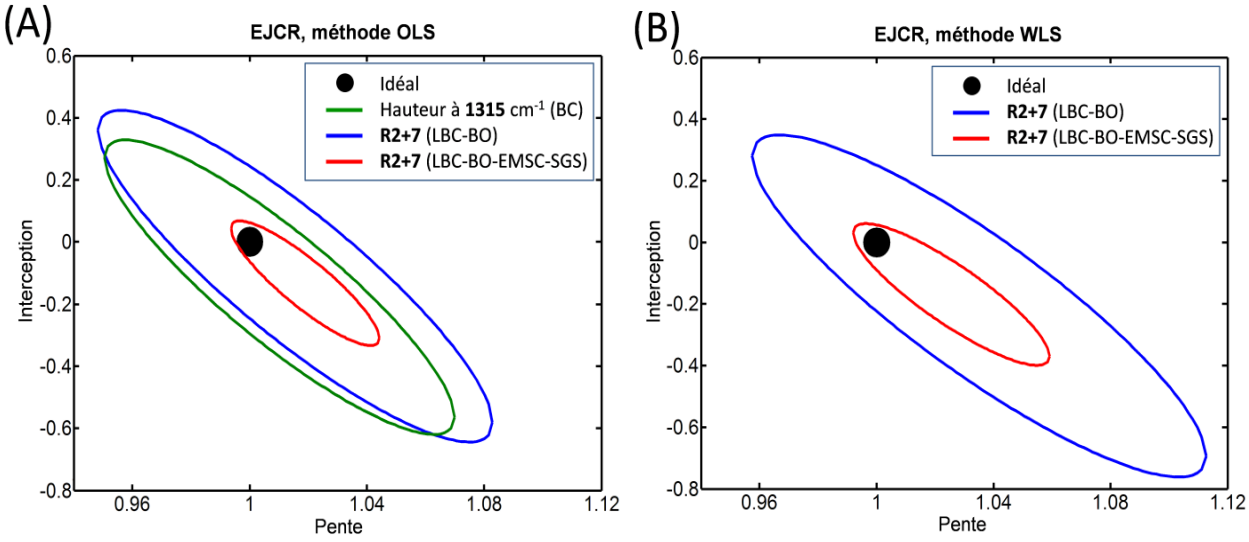


Figure 2.6 : EJCR dans le plan pente-interception réalisé pour les meilleurs modèles basés sur les moindres carrés ordinaires (A) et les moindres carrés pondérés (B).

Remarquons que le point théorique est situé à l'intérieur des régions de confiance pour tous les modèles, ce qui confirme l'absence de biais ou d'autres erreurs. Néanmoins, les modèles corrigés uniquement pour la ligne de base montrent des ellipses plus larges, mettant en jeu la précision de la technique d'échantillonnage. De plus, le modèle correspondant à la hauteur de pic

1315 cm^{-1} corrigé pour la ligne de base a montré un léger décalage vers la gauche, loin du point idéal, révélant l'efficacité des techniques de prétraitement multivariées même dans la correction de la ligne de base. En revanche, le modèle PLS prétraité avec EMSC montre une région de confiance plus étroite confirmant la précision obtenue mais pas l'exactitude des résultats.

Pour obtenir une meilleure compréhension de l'exactitude des modèles PLS développés, le test EJCR utilisant les moindres carrés pondérés (WLS) a été effectué. Cette approche est généralement utilisée lorsqu'une variance non constante peut être supposée [90, 91]. Pour ce faire, une nouvelle méthodologie a été étudiée en tenant compte des valeurs d'écart (ou « deviation » en anglais) estimées par le logiciel « The Unscrambler » pour pondérer individuellement chaque concentration prédite des échantillons de test. Cela permet une plus grande robustesse étant donné que les techniques de prétraitement mathématique peuvent conduire à une sous-estimation des valeurs d'écart-type (voir la **sous-section 2.3.5** ci-dessous).

La Figure 2.6B présente le nouvel EJCR basé sur les régressions WLS. Encore une fois, le point idéal était à l'intérieur mais pas au centre des régions de confiance, qui étaient légèrement plus larges mais ne différaient pas significativement de celles calculées par OLS; cette observation pourrait s'expliquer par l'incertitude élevée provenant du faible volume de la gouttelette déposée d'échantillon, un résultat qui peut être accepté pour une telle technique d'échantillonnage et d'analyse.

L'utilisation de l'ECJR pour la pente et l'interception, à la fois pour OLS et WLS, conduit à la conclusion qu'il n'y a pas de biais et que le taux de récupération peut donc être considéré comme 100 % pour les modèles choisis.

2.3.5. Analyse d'échantillons réels et validation des résultats

Dans cette section, les modèles validés utilisant la hauteur à 1315 cm^{-1} corrigée de la ligne de base (pour le modèle univarié) et celui utilisant les régions 2+7 prétraitées avec LBC-BO-EMSC-SGS (pour le modèle multivarié) ont été appliqués pour déterminer la teneur en nicotine dans 17 échantillons commerciaux

de tabac à chiquer et de 10 variétés de feuilles de tabac, traités selon le protocole décrit dans la **sous-section 2.2.4.1**. La Figure 2.7 illustre les spectres FTIR-ATR de deux extraits d'échantillons contenant et ne contenant pas de tabac (considérés comme des échantillons représentatifs) comparés à l'étalon de NCT traité, tandis qu'une comparaison complète de tous les spectres des produits commerciaux (de l'échantillon 1 à l'échantillon 17) est présentée à la Figure 2.8.

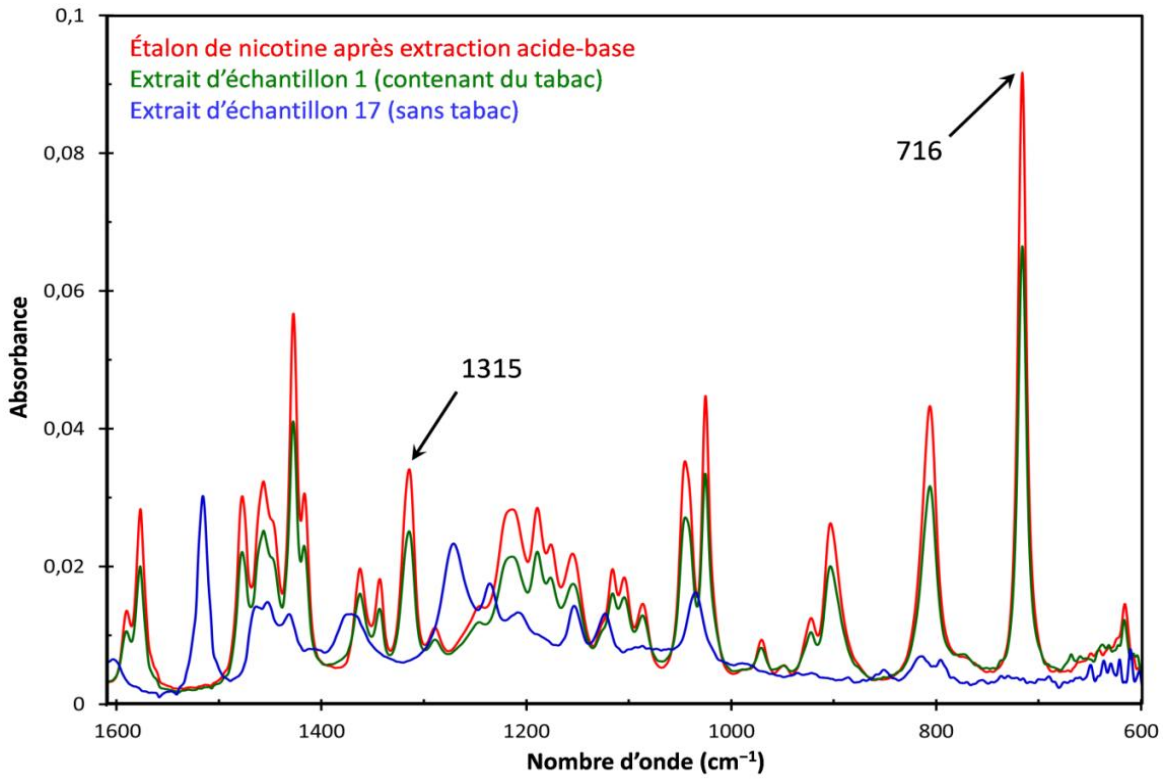


Figure 2.7 : Spectres FTIR-ATR d'extraits de produits contenant (vert) et ne contenant pas de tabac (bleu) par rapport à l'étalon traité (rouge).

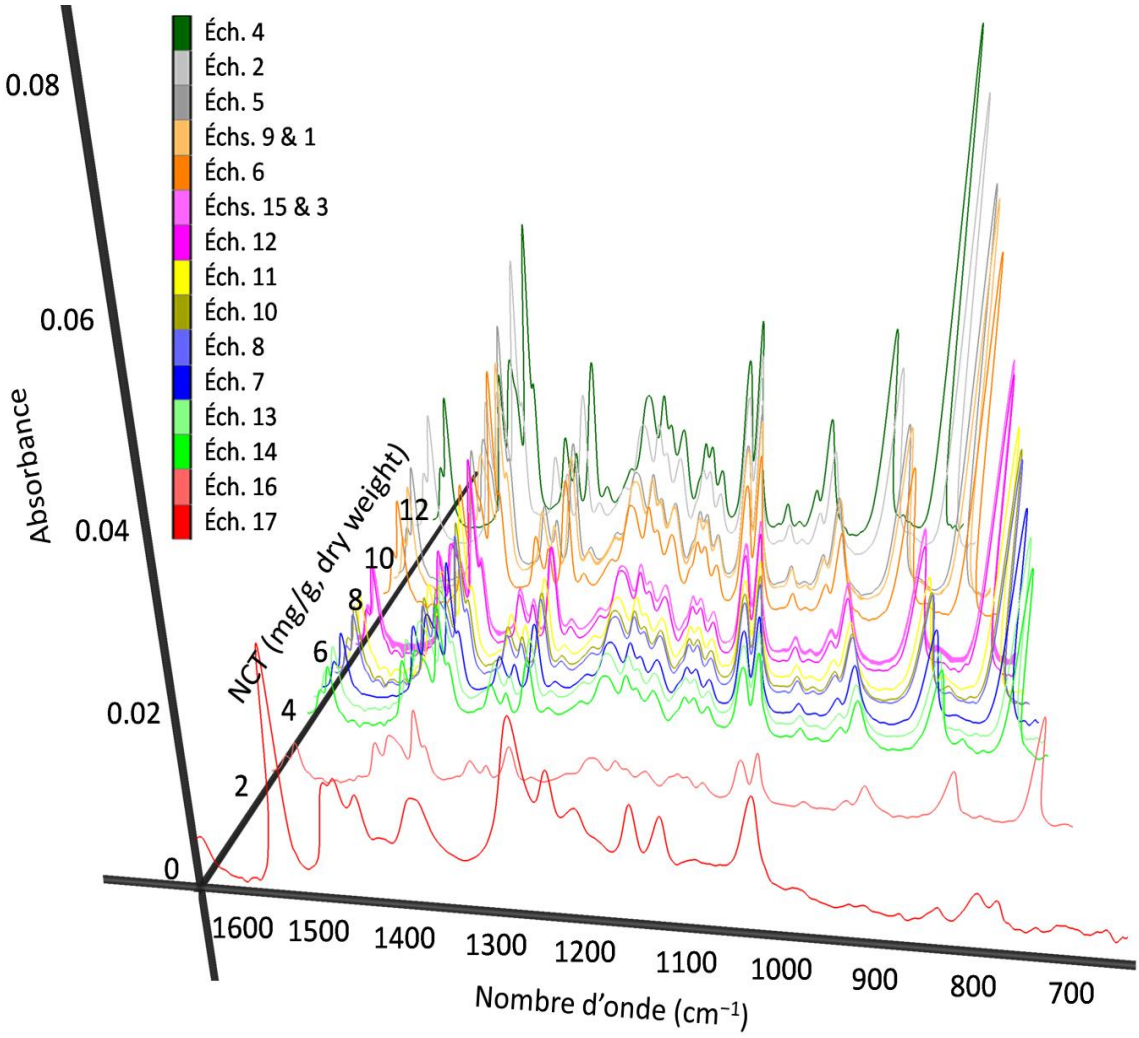


Figure 2.8 : Spectres FTIR-ATR moyens des produits commerciaux de ST présentés en fonction de leurs teneurs en nicotine totale.

Comme le montrent les spectres ci-dessus, la procédure recommandée fait preuve d'une très grande spécificité à l'égard de la nicotine en l'isolant de tous les produits commercialisés contenant du tabac, ce qui démontre la haute qualité de la détermination de l'analyte pouvant être obtenue lors de l'étape de prédiction. Une description détaillée des teneurs en nicotine totale en milligramme par gramme de poids sec d'échantillon est présentée dans le Tableau 2.5, conformément aux lignes directrices des Centres Américains de Contrôle et de Prévention des Maladies (CDC) relatives à la déclaration de la quantité de NCT dans les produits du ST [46, 47].

Tableau 2.5 : Teneurs en nicotine totale calculée en mg/g de produit dans des échantillons réels obtenues par l'approche proposée.

Méthode	**Loi de Beer**		**PLS-1**				
Mesure spectrale	Hauteur à 1315 cm^{-1}		Régions 2+7				
Prétraitement spectral	BC: 1333-1299 cm^{-1}		LBC-BO-EMSC-SGS		Ajustement spectral (%)		
Teneur en NCT (mg/g de poids sec)	Moy. ¶	SD	Moy. ¶	Écart moy.	Moy. ¶	SD	
Produits commerciaux							
Échantillon 1	8,5	0,4	8,7	0,2	99,46	0,27	
Échantillon 2	10,0	0,2	10,4	0,5	98,88	0,28	
Échantillon 3	6,0	0,3	6,4	0,3	99,06	0,24	
Échantillon 4	12,0	0,4	11,7	0,6	98,82	0,22	
Échantillon 5	9,3	0,1	9,0	0,5	98,36	0,21	
Échantillon 6	8,1	0,6	8,0	0,5	98,14	0,48	
Échantillon 7	4,5 *	0,2	4,6	0,2	98,36	0,58	
Échantillon 8	5,0 *	0,1	5,2	0,2	99,26	0,31	
Échantillon 9	8,3	0,3	8,7	0,4	98,96	0,27	
Échantillon 10	5,1 *	0,2	5,3	0,2	99,60	0,12	
Échantillon 11	5,4 *	0,5	5,6	0,2	99,50	0,25	
Échantillon 12	6,6	0,2	6,3	0,4	97,42	0,68	
Échantillon 13	4,0 *	0,4	4,2	0,2	98,58	0,52	
Échantillon 14	3,7 *	0,2	3,9	0,2	98,3	1,1	
Échantillon 15	6,1	0,4	6,4	0,3	98,64	0,50	
Éch. 16 (traditionnelle)	2,2 *	0,2	2,4	0,2	97,46	0,89	
Éch. 17 (sans tabac)	< LOD	-	< LOD	-	13,4	4,0	
Tabac à ST							
Berzili-Ain Mlila	5,2	0,3	5,9	0,4	95,5	2,3	
Berzili-Batna	< LOD	-	0,8	0,3	42,9	3,1	
Berzili-Biskra	24,9	0,7	25,3	1,4	98,52	0,37	
Chergui-Ain Mlila	9,6	0,7	9,7	0,6	96,7	1,2	
Jijel	45,2	2,1	45,0	1,6	98,88	0,35	
Soufi-Oued Souf	18,1	0,8	18,0	1,1	97,82	0,42	
Zeribet El Oued-Biskra	16,1	0,2	16,4	1,0	98,02	0,33	
Tabac à cigarettes							**Références** §
Burley strips	10,1	0,5	10,2	0,4	99,23	0,25	6,50-47,7 [175]
Oriental strips	9,2	0,3	9,3	0,6	97,3	1,2	1,80-12,6 [176]
Virginia strips	8,2	0,4	8,9	0,6	97,76	0,78	6,52-60,4 [176]

¶ Valeurs moyennes de cinq répétitions issues de deux mesures indépendantes de l'échantillon.

* La valeur de concentration calculée était inférieure à la $LOQ_{univariée}$ pour le modèle considéré.

§ Valeurs rapportées dans la littérature.

Abréviations : BC, Correction simple de la ligne de base ; Régions 2+7, Zones spectrales combinées: 1333 -1299 cm^{-1} + 740 - 690 cm^{-1}.

Contrairement à l'approche univariée qui se limite à la prédiction directe de la concentration de l'analyte, l'approche multivariée exploite de multiples variables et des tests statistiques pour évaluer les spectres de manière quantitative et qualitative. Cela permet une compréhension plus complète et approfondie des données analysées, améliore la capacité à discriminer les informations pertinentes et conduit à des prédictions plus fiables et robustes. Parmi ces tests statistiques, nous avons exploité la vérification de l'ajustement spectral, l'écart, le diagramme d'influence (Q-résidus par rapport au T^2 de Hotelling), ainsi que les contributions des variables au modèle et aux résidus.

2.3.5.1. Vérification de l'ajustement spectral

L'ajustement du spectre ou de la région de mesure, calculée par le logiciel « TQ Analyst », est une fonction attribuée pour comparer qualitativement les régions spectrales spécifiées de chaque échantillon quantifié avec les spectres de la matrice de calibration afin de déterminer leur degré de similitude, exprimé sous forme de valeur d'ajustement allant de 0 (pas de similitude) à 100 (concordance parfaite) [169].

Tous les produits commerciaux contenant du tabac présentent une excellente corrélation (> 97,4 %) entre leurs propres spectres et ceux de la matrice de calibration, permettant une discrimination précise entre les échantillons contenant et ne contenant pas de tabac. Pour les feuilles de tabac brut, celles utilisées dans la fabrication des cigarettes affichent de meilleurs coefficients d'ajustement (> 97,3 %) que celles utilisées pour la fabrication des ST (de 42,9 à 98,9 %). Cela pourrait s'expliquer par la différence entre les variétés ou le traitement auquel certains lots ont été soumis.

2.3.5.2. Valeur d'écart

Le deuxième paramètre inspecté, l'écart, est une évaluation de l'incertitude de prédiction pour chaque réplique d'échantillon inconnu. Il est estimé en fonction de l'erreur globale du modèle, de l'effet de « leverage » de l'échantillon et de la variance résiduelle de l'échantillon en X, sans tenir compte de la concentration référence [63].

Bien que les techniques de prétraitement mathématique aient amélioré la qualité des données de régression, elles ont eu un impact négatif sur les écarts-types calculés (de l'ordre de 10^{-4} à 10^{-3} mg/g, résultats non présentés). Ces valeurs ne reflétant pas la réalité, rendant l'écart mentionné ci-dessus la meilleure alternative. La plupart des valeurs d'écart calculées se situaient entre 0,2 et 0,6 mg/g, et des valeurs plus élevées ont été observées pour des concentrations plus élevées, signe d'une bonne qualité de prédiction. Cependant, il convient de noter que certaines de ces valeurs n'étaient pas corrélées avec l'ajustement spectral associé.

2.3.5.3. Diagramme d'influence

Ce test, issu des résultats de PLSR, repose sur les Q-résidus et la statistique de Hotelling T^2 appliqués aux échantillons réels. Il permet de décrire respectivement la distance du nouvel échantillon par rapport au modèle de calibration et le degré auquel cet échantillon est expliqué par le modèle [63]. La Figure 2.9 montre le tracé des Q-résidus en fonction de T^2 de Hotelling avec une valeur *p* de 0,05 comme limite critique.

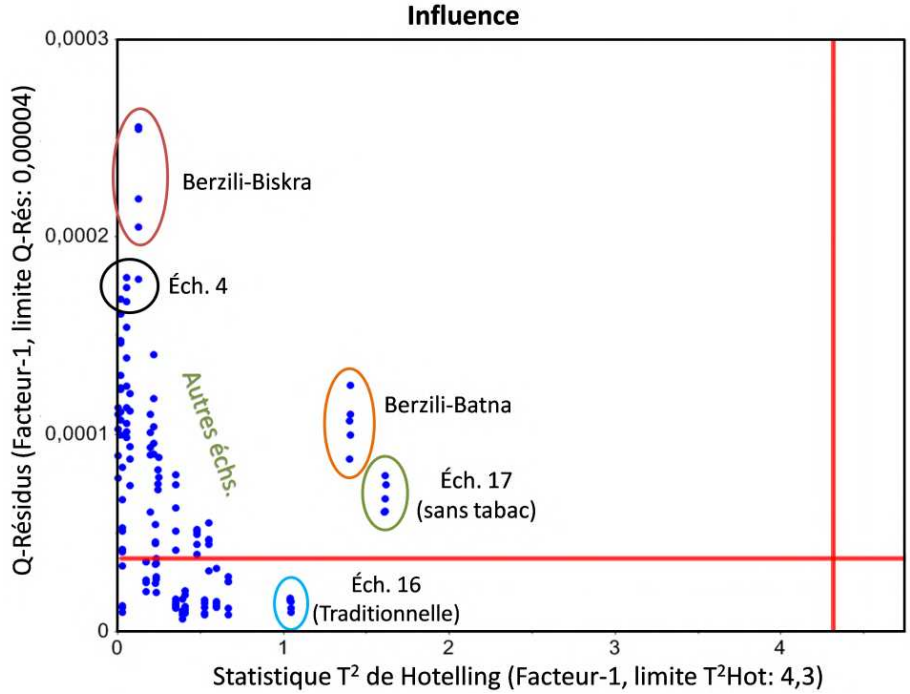


Figure 2.9 : Graphique des Q-résidus contre T^2 de Hotelling pour les échantillons réels. Les lignes rouges représentent les limites critiques associées.

Sur l'axe des abscisses (« leverages »), tous les échantillons se situent en dessous de la limite critique, indiquant une bonne modélisation. De gauche à droite, les points à l'extrême droite correspondent à l'extrait de Chemma traditionnelle suivi des feuilles de Berzili-Batna puis du produit sans tabac. Ces deux derniers ont montré de faibles valeurs d'ajustement spectral, ce qui est cohérent avec leur position sur le graphique et leur différence par rapport aux autres échantillons.

Sur l'axe des ordonnées (Q-résidus), la plupart des points dépassent la limite critique, ce qui pourrait suggérer un problème à première vue. Après avoir revérifié les données d'entrée pour des erreurs, il s'est avéré que ces échantillons n'étaient pas différents mais avaient simplement des valeurs extrêmes de concentration. Et tant que ces échantillons ne sont pas influents, la forte variance résiduelle est probablement due à des contributions mineures d'autres composants dans les extraits ou simplement une modélisation du bruit, ce qui sera discuté plus en détail avec le prochain paramètre.

2.3.5.4. Contributions des variables

Les contributions de T^2 de Hotelling et des Q-résidus, pour un échantillon donné, détaillent la contribution de chaque variable spectrale au modèle de calibration et aux résidus, respectivement [63]. La Figure 2.10 illustre comment les régions 2+7 contribuent au modèle, bien que de manière inégale. La région 2 démontre une performance supérieure dans la prédiction de cet échantillon exemple. En général, plus la valeur de contribution est petite, meilleure est la description de la variable, et par conséquent, de l'échantillon par le modèle.

Le graphique des contributions aux résidus (Figure 2.11) fait apparaître de multiples bandes autour de la zone d'intérêt. Ces bandes peuvent être liées à des décalages de pics, à des prétraitements spectraux, au bruit instrumental ou à des interférents inconnus dans la matrice (peut-être d'autres alcaloïdes mineurs). Heureusement, toutes les contributions aux résidus des échantillons étaient 100 à 2000 fois plus faibles que les contributions correspondantes au modèle, et n'ont donc pratiquement aucune influence sur les résultats de prédiction. De plus, aucune contribution résiduelle fréquente ou cohérente n'a été observée entre les

répliques, ce qui confirme que ces variances sont dues à des bandes non importantes du spectre, c'est-à-dire au bruit de signal.

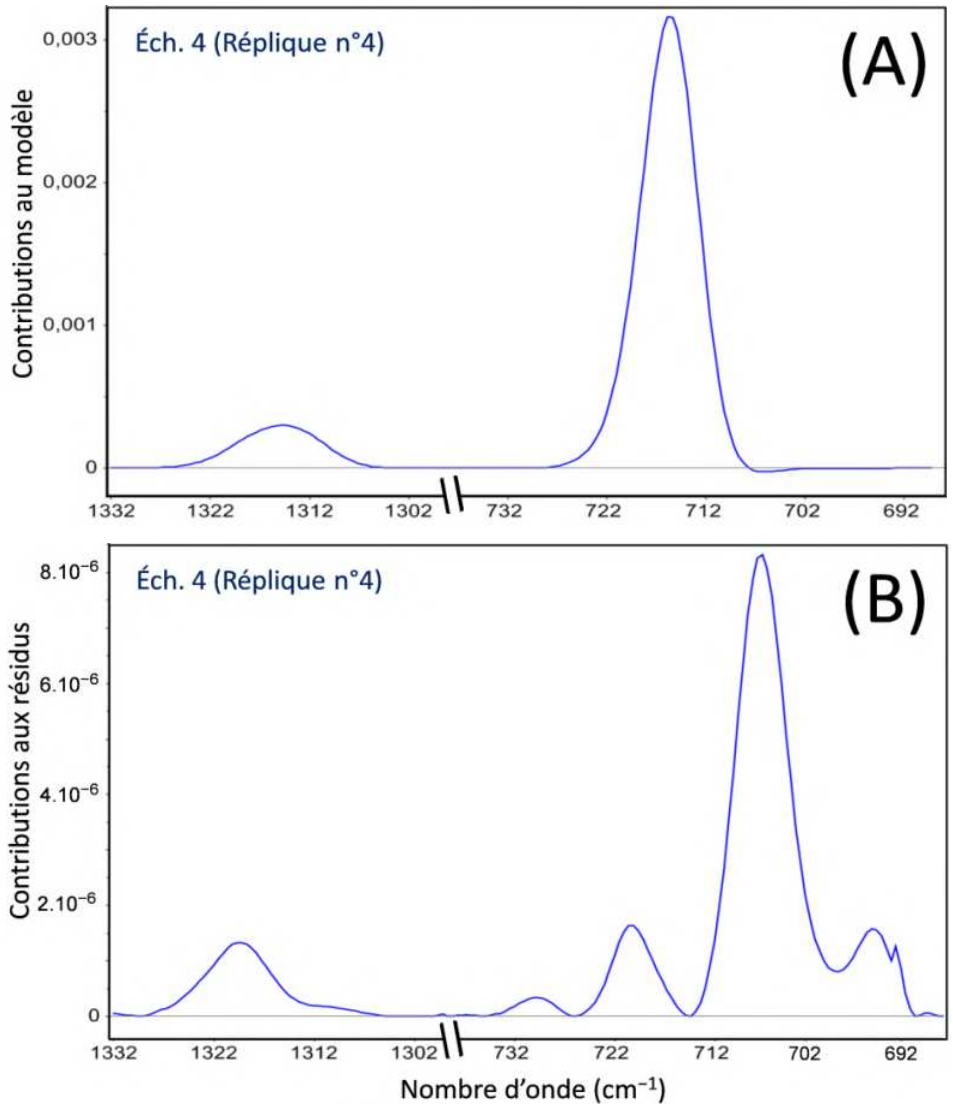


Figure 2.10 : Contributions des variables spectrales au modèle (A) et aux résidus (B) en utilisant la statistique de T^2 de Hotelling et des Q-résidus, respectivement, pour un nouvel échantillon (réplique n°4 de l'échantillon 4).

Il convient de noter que de telles observations seraient impossibles à vérifier avec la régression univariée et sans chimiométrie. Par conséquent, il est fortement recommandé d'effectuer toutes les calibrations à composante unique classiques en utilisant une approche multivariée afin de valider la méthode développée et les résultats obtenus.

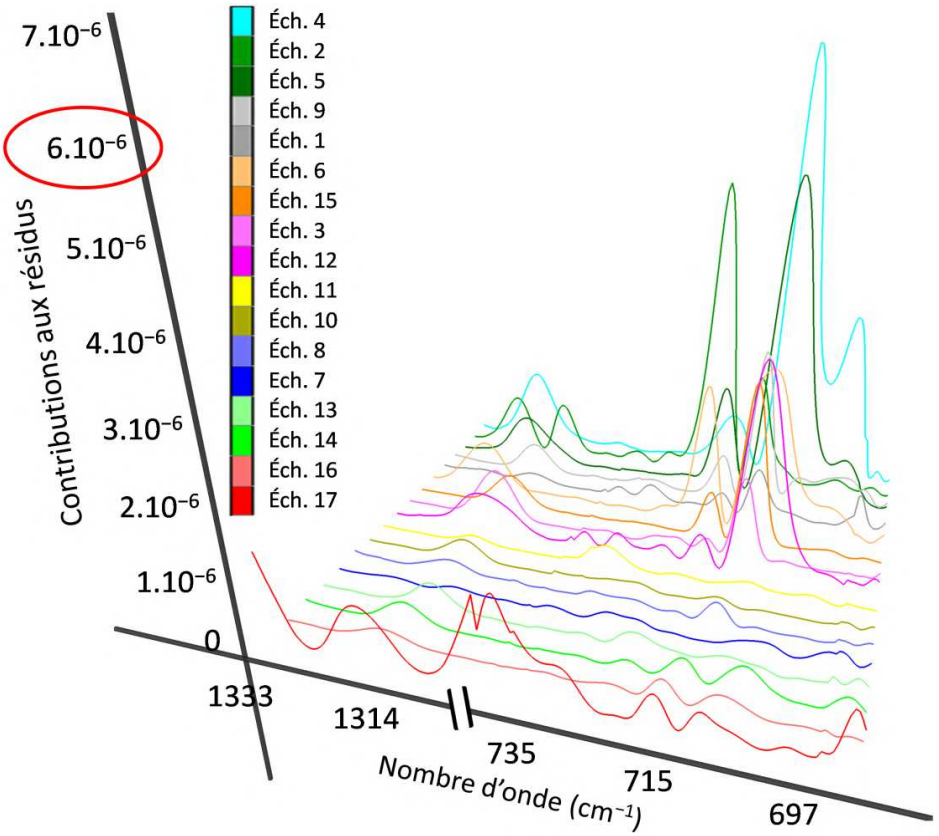


Figure 2.11 : Contribution moyenne des variables aux résidus pour les 17 échantillons commerciaux.

En utilisant le modèle PLS-ATR-FTIR optimal, la teneur en nicotine totale dans les produits de ST fabriqués s'est avérée varier de 3,9 à 11,7 mg/g de poids sec. Dans le produit traditionnel fait main, la teneur en NCT n'était que de 2,4 mg/g de produit sec, probablement en raison de la quantité importante d'additifs parmi ses ingrédients. En revanche, il a été confirmé que le produit aromatisé sans tabac ne contenait aucune trace détectable de NCT. Pour les feuilles de tabac, celles utilisées dans la fabrication du ST avaient des concentrations de NCT allant de 5,9 à 45,0 mg/g, à l'exception d'un échantillon atypique (0,8 mg/g). Les feuilles de tabac à cigarettes ont donné des concentrations variant entre 8,9 et 10,2 mg/g de feuille sèche, conformément aux valeurs de référence rapportées dans la littérature, ce qui valide partiellement la capacité de prédiction de la méthodologie développée.

2.4. Conclusion

Dans cette étude, nous avons démontré la faisabilité d'utiliser la technique ATR-FTIR à une réflexion unique combinée à la chimiométrie en tant qu'approche simple, peu coûteuse et spécifique pour l'analyse quantitative de la nicotine dans le tabac sans fumée commercial algérien "Chemma", marquant ainsi une première dans ce domaine.

La procédure d'extraction exposée a montré une haute sélectivité chimique vis-à-vis de la nicotine. L'approche multivariée utilisant l'algorithme PLS-1 a réussi à fournir les meilleurs résultats analytiques et une meilleure robustesse, et a donc été considérée comme la plus appropriée pour détecter et quantifier la nicotine dans une matrice complexe telle que le tabac à chiquer. Néanmoins, l'approche univariée utilisant la hauteur à 1315 cm^{-1} corrigée peut être considérée comme une alternative plus simple pour la détermination de la concentration de nicotine en utilisant la procédure susmentionnée.

Ainsi, avec un nombre croissant de nouveaux produits de ST contenant des agents de saveur complexes, des matières végétales non issues du tabac et/ou des adultérants, la spectroscopie ATR-FTIR est très prometteuse. Elle ne se veut pas une alternative aux solutions existantes, mais plutôt une méthode d'analyse de routine complémentaire et durable pour déterminer rapidement la nicotine dans divers produits du tabac. Cela peut contribuer à une meilleure compréhension des risques pour la santé associés à l'utilisation de ces produits et aider à l'élaboration de stratégies de réduction des méfaits.

CHAPITRE 3

DÉTERMINATION RAPIDE DES PARAMÈTRES DE QUALITÉ DU TABAC SANS FUMÉE PAR SPECTROSCOPIE ATR-FT-MIR : COMPARAISON DES APPROCHES MATHÉMATIQUES ET D'APPRENTISSAGE AUTOMATIQUE COVENTIONNEL

3.1. Introduction

La détermination efficace des paramètres de qualité du tabac sans fumée (ST) est indispensable pour garantir la qualité et conformité du produit. Ces paramètres reflètent directement la composition chimique du produit, qui est à son tour influencée par la qualité des ingrédients et la présence de substances nocives. Cependant, bien qu'offrant une grande fiabilité dans la surveillance des ST, les procédures existantes [46, 47] sont entravées par leur nature destructive des échantillons, des protocoles chronophages, des exigences énergétiques élevées et la nécessité d'une expertise technique spécialisée et de divers équipements. Ces limitations rendent ces procédures inadaptées à une utilisation routinière.

En tant qu'alternative rapide et non destructive, la spectroscopie proche infrarouge (NIR) gagne en popularité dans l'analyse du tabac. Une étude précédente [138] a rapporté une utilisation réussie de la technique NIR pour prédire la nicotine, les indices physiques, ainsi qu'un certain nombre de macromolécules, y compris la cellulose, la lignine, la pectine et les protéines, dans le tabac reconstitué. Plus récemment, Geng *et al.* [141] et Shu *et al.* [177] ont introduit de nouvelles méthodes basées sur le transfert de calibration pour déterminer, respectivement, les sucres totaux et les alcaloïdes totaux dans les feuilles de tabac. En utilisant la même technique spectroscopique, des résultats optimaux ont été obtenus grâce à des approches d'apprentissage profond (ou « deep learning » en anglais) pour quantifier la nicotine [139] et prédire simultanément la nicotine, les sucres, l'azote total et le pH [140].

Néanmoins, ces dernières solutions ont démontré leur efficacité principalement dans la prédiction de la composition du tabac brut, et leur performance n'est pas encore étudiée de manière approfondie pour des matrices commerciales complexes telles que le ST. De plus, les limitations inhérentes aux régressions basées sur l'apprentissage profond, telles que les problèmes de sur-ajustement [136] et l'absence de cadres mathématiques pour la validation des résultats, peuvent compromettre la robustesse des modèles développés, limitant ainsi leur application plus large.

La spectroscopie infrarouge moyenne à transformée de Fourier (FT-MIR), renommée pour sa grande capacité à cartographier les composés chimiques, offre une multitude d'informations précieuses par spectre en comparaison avec la spectroscopie NIR [178]. Malgré cette capacité, son application dans l'étude du ST a été principalement limitée à l'identification des espèces de tabac, des composants végétaux autres que le tabac, des produits chimiques individuels et des additifs spécifiques [102, 103, 154].

Dans d'autres contextes, la technique MIR, lorsqu'elle est associée à la réflexion totale atténuée (ATR) pour l'échantillonnage et utilisant de moindres carrés partiels (PLS) ou de machines à vecteurs de support (SVM) pour la modélisation, s'est révélée être un outil robuste pour les analyses qualitatives et quantitatives. Parmi les applications notables, citons la prédiction des réserves d'azote et d'amidon dans les vignes [163], des sucres et des acides organiques dans les pêches [179], et le profilage de l'adultération dans les saisies de cocaïne [180]. Dans des études comparatives, le MIR a démontré une performance supérieure à celle du NIR et d'autres techniques spectroscopiques dans la détermination des propriétés physicochimiques et rhéologiques des purées de pommes [181] et dans la quantification des adultérants dans la poudre de cumin [178].

À notre connaissance, aucun travail antérieur n'a fait état de la surveillance des paramètres de routine dans les ST à l'aide de la spectroscopie MIR. Par conséquent, la présente étude vise à explorer le potentiel de la technique ATR-FT-MIR, en combinaison avec la chimiométrie, en tant que méthode rapide, rentable, à haut débit et conviviale afin d'atteindre trois objectifs principaux: i) identifier les

principaux ingrédients du ST algérien ; ii) classifier les différentes marques commerciales largement disponibles sur les marchés ; iii) prédire cinq paramètres de qualité cruciaux, à savoir l'humidité, le pH, les cendres, la nicotine totale et la nicotine libre dans ces produits.

3.2. Partie Expérimentale

3.2.1. Équipements et logiciels

- **Spectroscopie FTIR:** Toutes les mesures ATR-FTIR ont été réalisées à l'aide du spectromètre Nicolet iS10 équipé d'un accessoire Smart iTR avec cristal de diamant (Thermo Fisher Scientific, États-Unis), comme décrit dans la **sous-section 2.2.1**. (Les détails sur les paramètres de mesure spécifiques peuvent être trouvés dans la section 'Analyse FTIR').
- **Séchage:** Une étuve de laboratoire MEMMERT UN30 (MEMMERT GmbH+Co.KG, Allemagne) d'un volume de 32 litres a été utilisée pour déterminer la teneur en humidité. Cette étuve offre une plage de température de +20 à +300 °C avec une précision de ±1 °C, un clapet de sortie d'air à commande électronique et une minuterie digitale avec programmation de l'heure d'arrêt, assurant un contrôle précis de la durée de chauffage et donc un processus de séchage efficace.
- **Détermination du pH:** Un pH-mètre HANNA HI 2209 (HANNA Instruments, Roumanie) équipé d'une électrode pH SenTix 62 a été utilisé pour effectuer toutes les mesures de pH. Le pH-mètre a été calibré à l'aide de solutions tampons standard avant chaque utilisation.
- **Calcination:** Un four à moufle de laboratoire de la marque Nabertherm LV 11/9/B180 (Nabertherm S.A.S, Allemagne) a été utilisé pour un processus d'incinération complet des échantillons. D'un volume de 9 litres, ce four est doté d'éléments chauffants intégrés dans le moufle en céramique, capable d'atteindre des températures allant jusqu'à 1100 °C.

Logiciels:

Le logiciel intégré OMNIC™ version 9.8 (Thermo Fisher Scientific, États-Unis) a été utilisé pour piloter le spectromètre FTIR et acquérir les spectres. Le logiciel XLSTAT v.2016 (Addinsoft SARL, France) exécuté dans Microsoft Office (MS) Excel 2010 (Microsoft, États-Unis) a permis de réaliser la PCA (analyse en composantes principales), l'AHC (classification hiérarchique ascendante) et le partitionnement en *k*-moyennes.

Le logiciel The Unscrambler® X 10.4 (Camo Software AS., Norvège) a été employé pour prétraiter les spectres et mettre en œuvre les méthodes PLS-DA (analyse discriminante par moindres carrés partiels), PLSR (régression par moindres carrés partiels), SVM-C (classification par machine à vecteurs de support) et SVMR (régression par machine à vecteurs de support).

Enfin, les analyses *i*-PLS (interval-PLS) et EJCR (région de confiance elliptique conjointe) ont été mis en œuvre à l'aide des codes MATLAB référencés dans [86], alors que d'autres paramètres statistiques ont été calculés avec MS Excel ou par des fonctions personnalisées exécutées dans MATLAB R2012a (Mathworks Inc., États-Unis).

3.2.2. Réactifs et produits chimiques

Les détails concernant l'étalon de nicotine (NCT), le chloroforme ($CHCl_3$), le 2-propanol, NaOH, Na_2CO_3 anhydre et Na_2SO_4 anhydre ont été décrits de manière plus détaillée au niveau de la **sous-section 2.2.2**.

Une eau tri-distillée a été préparée dans notre laboratoire et utilisée à la place de l'eau distillée déionisée pour les mesures de pH.

3.2.3. Collecte des échantillons

L'échantillonnage s'est déroulé en deux parties. La première partie consistait en l'achat de 46 échantillons de tabac à chiquer algériens natifs entre janvier et juin 2021. Ces échantillons ont été obtenus auprès de diverses sources, y compris

des magasins de tabac en gros et au détail (boutiques cosmétiques), ainsi que des vendeurs de rue, dans 10 endroits différents dans deux provinces: Médéa et Blida. Cette série d'échantillons comprenait un produit de contrôle fourni par le Laboratoire Central de Contrôle Qualité de l'entreprise United Tobacco Company (UTC du Groupe MADAR, Boumerdès), deux produits authentiques disponibles commercialement portant le même nom de la marque certifiée, deux analogues contrefaits des produits authentiques, 39 produits fabriqués illégalement, une variante traditionnelle de Chemma fabriquée maison, et un produit aromatisé à base de plantes (sans tabac) destiné à aider à l'arrêt du tabagisme.

Dans la deuxième partie, menée entre janvier et juin 2022, 59 échantillons supplémentaires ont été collectés auprès des mêmes sources que la première série, à l'exception de deux nouveaux produits étrangers. Cette série d'échantillons comprenait les deux produits commerciaux certifiés de l'UTC, 54 produits contrefaits, deux produits de ST fabriqués en Belgique et importés de la France, ainsi qu'un autre produit de Chemma traditionnelle. Il convient de noter que seuls 33 produits de l'ensemble initial ont été ré-échantillonnés, car certains échantillons n'étaient plus disponibles à la vente ou avaient fait l'objet d'un changement de marque.

Les noms de marque ne sont pas indiqués pour des raisons de confidentialité. Tous les échantillons étaient estimés représenter collectivement plus de 90% de la part de marché du ST en Algérie.

3.2.4. Préparation et analyse des échantillons

La Chemma se présente généralement sous une forme broyée et ne nécessite pas de traitement avant l'analyse. Toutefois, certains produits de contrefaçon peuvent contenir de grosses particules agglomérées en raison de méthodes de préparation inappropriées ou de l'incorporation de déchets de tabac (tiges, par exemple). Afin d'assurer une analyse précise, le contenu du produit a été mélangé soigneusement à l'aide d'un moulin à café. Ce processus a été répété trois fois ou plus, pendant 20 secondes à chaque fois, afin d'éviter tout échauffement et d'obtenir un échantillon homogène avec une granulométrie minimale. Ensuite, le contenu de chaque produit a été remis dans son emballage

ou sa boîte d'origine, étiqueté, scellé dans des sachets en plastique et stocké à −10 °C jusqu'à l'analyse.

Pour l'analyse, une petite quantité d'échantillons, généralement quelques milligrammes, a été pressée directement contre le cristal de diamant à l'aide de la pointe de dispositif de pression standard. Les spectres ont été enregistrés dans la région MIR (4000 - 525 cm^{-1}) avec une résolution de 4 cm^{-1} et une moyenne de 32 balayages par spectre en mode absorbance.

Avant chaque mesure, le cristal a été nettoyé deux fois avec de l'éthanol à 96° suivi d'isopropanol, et l'interférogramme de fond a été enregistré une fois après chaque triplicat dans des conditions identiques d'analyse des échantillons. Chaque échantillon a été analysé 10 fois, ce qui a permis d'obtenir un total de plus de 1000 spectres.

3.2.5. Mesures de référence des paramètres de qualité du ST

3.2.5.1. Teneur en humidité totale

Les teneurs en humidité des ST ont été déterminés selon la méthode de Centres Américains de Contrôle et de Prévention des Maladies (CDC) [46, 47] en utilisant un étuvage gravimétrique d'une portion pesée (précisément 5,00 g) à 99,0 ± 1,0 °C pendant 3 heures.

Cette méthode est appelée « humidité totale » car elle mesure l'eau et tous les autres constituants volatils du produit à une température de 99 °C. Il existe d'autres approches pour déterminer uniquement la teneur en eau, utilisant la méthode de Karl Fischer [133], la spectroscopie NIR [118], la co-distillation dans un appareil Dean-Stark [109], lyophilisation, ou simplement le séchage à température ambiante pendant 3 à 5 jours [105].

Ce paramètre est lié à l'activité microbienne globale, et plus particulièrement à certains micro-organismes transformateurs d'azote qui peuvent favoriser indirectement la formation de nitrosamines spécifiques du tabac (TSNAs) au cours des différents processus de production du tabac [44, 154]. La teneur en eau peut

également affecter la durée de conservation du produit et d'autres paramètres de qualité, tels que le pH et les concentrations de NCT.

3.2.5.2. Détermination du pH

Exactement 2,00 g de produit homogénéisé ont été placés dans un bécher de 50 ml avec 20 ml d'eau tri-distillée et la suspension a été agitée magnétiquement pendant 30 minutes comme décrit dans le protocole révisé du CDC [47]. Lors de la filtration du surnageant dans l'obscurité, celui-ci a reposé pendant encore 20 minutes avant l'analyse.

Les mesures de pH ont été effectuées en double avec un pH-mètre étalonné quotidiennement à deux points avec des solutions tampons à 7,00 et 10,01, ainsi qu'une compensation de température. Aucune variation significative des valeurs de pH n'a été observée entre les intervalles de temps d'agitation de 30 et 60 minutes.

En plus d'autres facteurs, le pH joue un rôle crucial dans la détermination de la proportion de NCT sous sa forme non ionisée [37]; à des niveaux de pH plus élevés, une plus grande proportion de NCT est non ionisée. La manipulation du pH du produit peut amplifier de manière significative les effets pharmacologiques et le potentiel de dépendance du NCT (Voir la **sous-section 3.2.5.5** ci-dessous).

3.2.5.3. Teneur en cendres totales

La teneur en cendres a été déterminée suivant une procédure précédemment rapportée [128] avec quelques modifications. Une quantité de 5,00 g du produit entier a été brûlée à l'air sur une plaque chauffante à 300 °C dans un creuset en silice pendant 30 minutes. Elle a ensuite été transférée dans un four à moufle et incinérée à 600 °C pendant 2 heures pour assurer l'élimination complète de toute particule de carbone.

Cette analyse est appelée « cendres totales » car elle mesure les cendres résiduelles restant après l'incinération du tabac, ainsi que tous les composants minéraux. Cela peut fournir une estimation des additifs ou des adultérants inorganiques qui ont été ajoutés au produit final.

3.2.5.4. Teneur en nicotine totale

La quantification du NCT totale a été réalisée en utilisant la méthode développée par notre équipe de recherche exposée dans le Chapitre 2.

La NCT peut subir une déméthylation pour former d'autres alcaloïdes mineurs qui peuvent se transformer via la nitrosation en TSNAs pendant le traitement ou le stockage du tabac à haute température [44, 116]. Par conséquent, l'analyse du NCT est essentielle pour évaluer à la fois les effets addictifs et identifier les produits à haute teneur en TSNAs cancérigènes.

3.2.5.5. Teneur en nicotine non ionisée

Considérant la nicotine est une base faible: $B + H^+ \rightleftharpoons BH^+$.

La concentration de la forme non ionisée (*B*) de NCT peut être calculée en utilisant le pH de l'échantillon et la constante d'ionisation (pK_a = 8,02) substituée dans l'équation de Henderson-Hasselbalch [46]:

$$pH = pK_a + \log\frac{[B]}{[BH^+]} \qquad \text{(Éq. 3.1)}$$

Après réarrangement:

$$NCT\ libre\ (mg/g) = NCT\ totale\ (mg/g) \times \frac{\frac{[B]}{[BH^+]}}{\frac{[B]}{[BH^+]}+1} \qquad \text{(Éq. 3.2)}$$

En augmentant l'alcalinité du produit, les sels de NCT se transforment en forme libre, qui traverse plus facilement la muqueuse buccale. Cela peut entraîner une absorption plus rapide de la NCT vers le système nerveux central, pouvant potentiellement en affecter ses effets [37, 182].

3.2.6. Analyse des données

3.2.6.1. Prétraitement spectral

Plusieurs stratégies de prétraitement, incluant la correction du décalage de la ligne de base (BO), la correction étendue de la diffusion multiplicative (EMSC), la

« Standard Normal Variate » (SNV), deux types de normalisation (normalisation par vecteur unitaire et normalisation par gamme spectrale), la dérivée de premier ordre de Savitzky-Golay (SG FD) à 9 points / côté et un ordre polynomial de 2 (PO = 2), la dérivée de deuxième ordre de Savitzky-Golay (SG SD) à 7 points / côté et un PO = 3, la correction de tendance (DT) avec PO = 2, et certaines de leurs combinaisons ont été étudiées afin de réduire les effets externes indésirables.

3.2.6.2. Analyse en composantes principales (PCA)

Une PCA centrée sur la moyenne a été réalisée en utilisant l'algorithme de décomposition en valeurs singulières avec 10 composantes principales (PCs) sur:

i. Les répliques spectrales à gamme complète, corrigées par SNV, pour détecter les valeurs aberrantes résultant de divers problèmes de collecte de données (aberration intra-groupe). L'identification des échantillons aberrants s'est faite par examen du graphique des scores et du graphique d'influence (F-résidus contre statistique T^2 de Hotelling) à un niveau de signification de 0,05 (Plus de détails peuvent être trouvés dans la **sous-section 2.2.5.1**).
ii. Les mesures de référence normalisées des échantillons (sur matrice de corrélation), ainsi que leurs spectres correspondants corrigés par SNV (à l'exclusion des régions CO_2, cristal ATR, blanc et bruit) comme mentionné ci-dessus, ont été utilisés pour éliminer les échantillons influents présentant un écart significatif de la tendance générale des données (aberration inter-groupe, Figure 3.1).
iii. Le jeu de données normalisé de mesures références des échantillons restants de l'étape (ii) pour examiner les tendances de distribution, révélant tout groupe visible pouvant être discriminé en fonction des similitudes et des différences entre les échantillons.
iv. Le jeu de données normalisé des données restants après l'élimination des valeurs aberrantes (étape ii) pour diviser les échantillons en ensembles de calibration (67 % des données) et de test (33 % des données) en utilisant la technique de sélection d'échantillon de Kennard-Stone appliquée sur le graphique des scores.

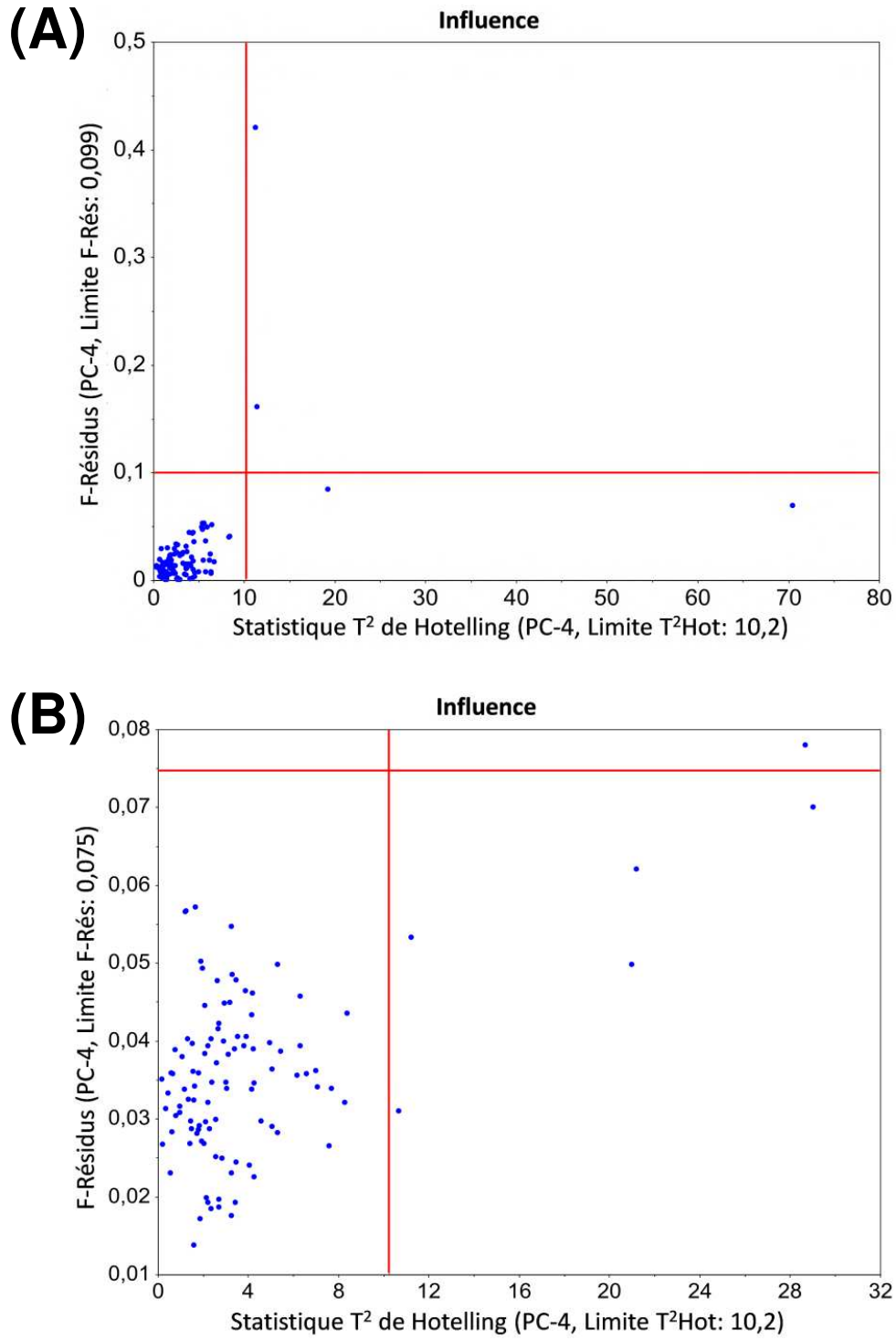


Figure 3.1 : Graphique d'influence des F-résidus contre T^2 de Hotelling, aux limites critiques de 5 %, appliqué sur (A) les mesures de référence et (B) les spectres FTIR-ATR du nombre total d'échantillons commerciaux pour l'élimination d'aberration inter-groupe. Les points situés au-delà des lignes rouges sont considérés comme des valeurs aberrantes.

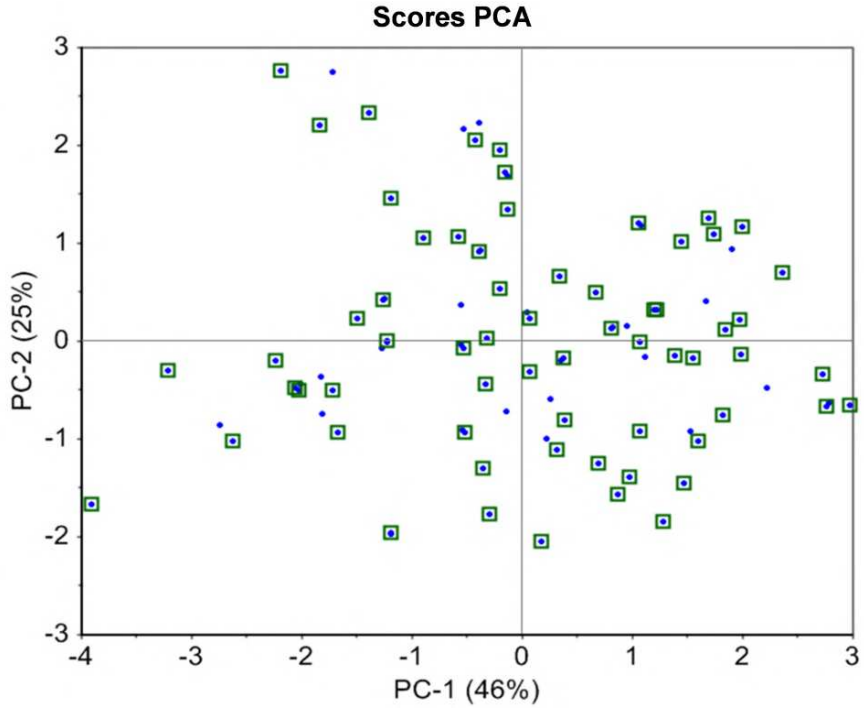


Figure 3.2 : Méthode Kennard-Stone−PCA appliquée sur l'ensemble de mesures de référence pour la division des échantillons. Les points marqués avec des carrés ont été utilisés pour l'entraînement et les restants ont été utilisés pour tester les modèles.

3.2.6.3. Classification hiérarchique ascendante (AHC)

Outre la nature des données, le résultat d'une classification AHC dépend du type de proximité (similarité ou dissimilarité), de l'indice choisi pour le calcul (distance euclidienne, distance de Mahalanobis, …) et de la stratégie d'agrégation.

Dans ce travail, l'AHC a été testée en utilisant les stratégies d'agrégation par les liens: simple, complet, moyen, proportionnel, flexible et la méthode de Ward. De plus, la troncature du dendrogramme à un niveau donné a été définie automatiquement en fonction de critères tels que les valeurs d'entropie ou d'inertie, ou manuellement en évaluant les relations entre les individus à tous les niveaux ainsi que l'évolution de la variance intra-classe.

3.2.6.4. Classification k-*means*

Concernant cette méthode, quatre critères de classification ont été utilisés pour parvenir à la solution la plus acceptable, à savoir la Trace(W), le Déterminant(W), le lambda de Wilks et la Trace(W) / Médiane.

Le nombre pertinent de classes a été déterminé via la méthode de recherche du "coude" à partir du tracé de la variance intra-classe en fonction du nombre de classe *k* variant de 1 à 10. Sur ce dernier, l'évolution de la variance diminue mathématiquement lorsque le nombre de classes augmente. Si les données sont distribuées de manière homogène, la décroissance est linéaire. S'il y a réellement une structure de groupes, un coude sera observé pour le nombre de classes pertinent.

Étant donné que la classification *k*-means est une méthode itérative, les calculs doivent être arrêtés si une valeur maximale (500 itérations) a été atteinte, ou si le modèle a convergé pour une valeur minimale (choisie égale à 0,00001).

3.2.6.5. Régression par moindres carrés partiels (PLSR)

Pour développer des modèles PLSR robustes, la(les) région(s) spectrale(s) appropriée(s), le prétraitement spectral et le nombre de variables latentes (LVs) ont été soigneusement définis.

En plus de considérer la gamme spectrale complète, deux méthodes ont été comparées pour sélectionner les nombres d'ondes les plus pertinents: la régression des moindres carrés partiels par intervalle (*i*-PLS) et l'importance des variables en projection (VIP).

Le nombre optimal de LVs a été déterminé à l'aide de deux approches: i) Examen du graphique de variance totale expliquée de la validation croisée (CV), recherchant la valeur la plus élevée avec le moins de facteurs possible; et ii) Détermination du critère de Haaland, qui implique d'utiliser le premier modèle calculant pour une valeur $p < 0,75$ de la valeur F expérimentale (rapport entre la somme des carrés des erreurs résiduelles prédites (PRESS) pour chaque nombre de LV et le PRESS minimum possible [171]). Une valeur $p < 0,75$ indique que le

modèle avec ce nombre de LV n'est pas statistiquement différent du modèle avec le PRESS minimum.

Tous les modèles analytiques ont été construits en utilisant l'algorithme « Kernel PLS » avec un maximum de 15 facteurs et un centrage sur la moyenne des données.

3.2.6.6. Régression par machine à vecteurs de support (SVMR)

Au cours de l'étape de régression, tous les paramètres de qualité ont été optimisés pour différentes stratégies de prétraitement, fonctions de noyau (linéaire, polynomial et fonction de base radiale) et trois hyperparamètres (C, ε et γ) en utilisant l'algorithme ε-SVR sur des données mises à l'échelle de [−1, 1].

L'hyperparamètre γ, qui détermine la largeur du noyau gaussien, a été ajusté que dans la SVMR à des fonctions de noyau non linéaire (ou « custom kernelized SVMR » en anglais). La valeur initiale de ε a été fixée à environ 5 à 10 % de la valeur correspondante d'écart-type calculée pour chaque paramètre de qualité.

Une recherche exhaustive sur grille, à échelle logarithmique dans l'intervalle de $[10^{-4}, 10^{4}]$, a ensuite été effectuée pour déterminer les valeurs optimales de C et γ. Chaque combinaison d'hyperparamètres a été rigoureusement évaluée en utilisant un cadre de CV à 10 plis, et la combinaison produisant la plus faible RMSECV a été choisie. En cas de performances comparables, le modèle avec le moins de vecteurs de support (SVs) était préféré en raison de sa sensibilité réduite aux fluctuations mineures dans les données d'entraînement et d'un risque moindre de sur-ajustement.

En général, la méthodologie adoptée a suivi une approche systématique d'essais et d'erreurs, visant à identifier la meilleure configuration qui permet au modèle de se généraliser efficacement à de nouveaux échantillons tout en s'ajustant toujours avec précision les données d'entraînement. En cas de problèmes de convergence, les résultats actuels étaient renvoyés avec un avertissement d'échec associé.

3.2.7. Métriques d'évaluation

Les modèles PLSR et SVMR ont été validés en utilisant la méthode de CV complète (leave-one-out), testés pour prédire les valeurs cibles dans un ensemble de tests indépendant, et la performance de tous les modèles a été vérifiée à l'aide des erreurs quadratiques moyennes (RMSE) et des coefficients de détermination (R^2) de calibration, de CV et de prédiction.

Des statistiques critiques, y compris l'erreur absolue moyenne en pourcentage (MAPE), l'erreur relative de prédiction (REP), « Ratio of Prediction-to-Deviation (RPD) », « Range Error Ratio (RER) » et le biais pour l'ensemble de test, ont été évaluées afin de corroborer la robustesse des modèles les plus performants. Tous ces éléments ont été calculés comme indiqué dans la **sous-section 1.3.6.1** du Chapitre 1.

De plus, la justesse de la prédiction a été déduite en effectuant l'analyse de la région de confiance elliptique conjointe (EJCR) à un niveau de 95 % pour la pente et l'interception de la régression "Prédite contre Réelle" pour l'ensemble de tests, en se basant sur la méthode des moindres carrés ordinaires.

3.2.8. Limites de détection et de quantification

Parmi les indicateurs de performance analytiques, la limite de détection (LOD) revêt une importance primordiale car elle révèle la plus faible concentration détectable d'un analyte à l'aide du modèle correspondant, permettant la comparaison avec d'autres méthodes.

Selon Allegrini et Olivieri [173], la LOD en PLSR comprend deux limites: inférieure et supérieure, en considérant à la fois les risques de fausses détections (erreur α) et de fausses non-détections (erreur β). Comme indiqué dans le Tableau 1.3 (du Chapitre 1), elle peut être calculée facilement à partir de la sensibilité et des incertitudes dans les signaux et les concentrations. Cependant, ce n'est pas le cas avec la SVMR qui fournit des modèles plus complexes avec des relations potentiellement non linéaires entre les caractéristiques d'entrée, implicitement transformées dans un espace de dimension supérieure, et la cible.

Par conséquent, des méthodes plus simples pour le calcul de la LOD sont souhaitables. Plusieurs approches existent dans la littérature, mais les méthodes présentées dans Ortiz *et al.* [183] (Voir Éq. 3.3 et l'explication ci-après) et les directives de l'ICH [§] [184] (Éq. 3.4) semblent les plus adaptées à ce travail.

$$\mathrm{LOD_{pu}} = \frac{2t_{(0.05,I-2)}}{b'}\sqrt{\left(\frac{1}{K}+\frac{1}{I}+\frac{\bar{y}^2}{\sum_{i=1}^{I}(y_i-\bar{y})^2}\right)\frac{\sum_{i=1}^{I}\left(\hat{y}_i-\hat{y}_i'\right)^2}{I-2}} \qquad \text{(Éq. 3.3)}$$

$$\mathrm{LOD_{ICH}} = \frac{3.3\times SD_a'}{b'} \qquad \text{(Éq. 3.4)}$$

où *I* est le nombre total d'échantillons, y_i et $\hat{y}_i$ sont respectivement les valeurs réelles et prédites pour le i-ème échantillon, $t_{(0.05,\ I-2)}$ représente la distribution *t* de Student pour le niveau de signification donné (*α* et *β*) et les degrés de liberté (*I* − *2*), *b* est la pente, SD_a est l'écart-type de l'interception, *K* est le nombre de répliques, et l'apostrophe (′) indique les valeurs estimées à partir de la droite de régression pseudo-univariée (pu). Le LOQ a été estimé à 3 fois la LOD correspondante.

L'idée d'utiliser une approche simple pour le calcul de la LOD était basée à l'origine sur le travail antérieur d'Ortiz *et al.* [183] où le principe était exprimé comme suit: « *Si les données (x_i, y_i) sont transformées en (x_i, y_i') au moyen de $y_i' = m_{yi} + n$, la régression de Y' sur la concentration X donne la même capacité de détection, x_d, que la régression de Y sur la concentration X* ».

3.3. Résultats et discussion

3.3.1. Mesures de référence des paramètres de qualité

L'analyse des graphiques d'influence PCA (Figure 3.1) a révélé neuf échantillons suspects aberrants, y compris le produit sans tabac (à base de plantes), la Chemma traditionnelle et les deux produits importés de l'étranger. En effet, ces ST présentaient des valeurs extrêmes probablement dues à leurs

[§] Conseil international pour l'harmonisation des exigences techniques pour l'enregistrement des médicaments à usage humain.

composants chimiques anormaux, les rendant influents et s'écartant significativement de la tendance générale des données. Bien qu'ils ne représentent que 8,6% des données, l'inclusion de ces échantillons dans le développement du modèle pourrait compromettre sa précision et sa robustesse. Leur suppression a permis d'améliorer l'homogénéité des données au sein de chaque ensemble.

L'application de la méthode Kennard-Stone−PCA sur les deux séries d'échantillons combinés (comprenant les échantillons collectés en 2021 et 2022) a été réalisée pour garantir la formation d'un ensemble de données équilibré. Cette approche facilite l'inclusion d'échantillons représentatifs dans le processus de calibration, ce qui permet de contourner efficacement les problèmes liés à la variabilité de la composition chimique entre les plantes de tabac. Cette variabilité découle généralement de divers facteurs, tels que la génétique des plantes, les conditions de croissance (climat, sol) et les pratiques agricoles. En combinant des échantillons de différentes années, une plus grande partie de cette variabilité naturelle est prise en compte par le modèle. Cela, à son tour, améliore la capacité de généralisation du modèle aux données inconnues. La méthode de Kennard-Stone renforce encore cet avantage en garantissant une distribution uniforme des échantillons sélectionnés dans l'espace des caractéristiques.

Les résultats des mesures de référence des paramètres physicochimiques sont présentés dans le Tableau 3.1. Le contrôle statistique de processus avant et après la suppression des valeurs aberrantes est résumé dans le Tableau 3.2, tandis que les distributions de fréquence des données pour les ensembles de calibration et de prédiction sont présentées dans la Figure 3.3. Comme le montre le Tableau 3.2, les sous-ensembles de calibration présentent des plages de valeurs équivalentes ou légèrement plus larges que les ensembles de test pour chaque paramètre. Des moyennes, médianes et écarts-types comparables ont été obtenues pour les deux ensembles, confirmant à nouveau l'efficacité de l'approche de partitionnement dans le maintien de l'uniformité des données et garantissant que les échantillons sélectionnés représentent les variations anticipées dans les produits commerciaux.

Tableau 3.1 : Mesures de référence des paramètres de qualité dans les produits commerciaux de ST collectés.

Produits commerciaux		Humidité (%)	pH	Cendres (%)	NCT totale (mg/g, ps)	NCT libre (mg/g, ps)
Échantillons sélectionés par la méthode de Kennard-Stone	S01	45,4	10,29	20,8	8,0	8,0
	S02	46,7	9,48	25,5	3,9	3,7
	S03	50,4	10,15	25,7	10,4	10,3
	S04	47,9	10,90	24,8	8,0	8,0
	S05	47,5	11,05	23,1	10,0	10,0
	S06	49,8	9,89	24,3	6,4	6,3
	S07	48,5	10,68	23,1	8,1	8,1
	S08	46,9	10,64	23,1	8,1	8,1
	S09	51,0	9,89	26,2	9,9	9,7
	S10	48,8	9,86	26,2	9,9	9,7
	S11	50,3	9,76	26,6	4,0	3,9
	S12	51,3	9,85	26,6	4,0	3,9
	S13	46,9	9,35	24,1	11,7	11,2
	S14	46,1	9,60	24,1	11,7	11,4
	S15	47,4	10,24	23,8	11,9	11,8
	S16	45,5	10,30	23,9	13,5	13,4
	S17	50,9	9,42	26,6	7,1	6,8
	S18	49,3	9,44	26,6	7,1	6,9
	S19	50,0	10,92	25,8	9,0	9,0
	S20	49,6	10,95	24,3	8,0	8,0
	S21	46,6	10,23	26,8	9,0	8,9
	S22	45,7	10,26	22,6	8,5	8,5
	S23	46,9	10,26	22,6	8,5	8,5
	S24	50,0	9,00	27,4	4,1	3,7
	S25	49,9	10,83	27,3	8,2	8,1
	S26	51,2	10,75	27,3	8,2	8,1
	S27	46,3	10,12	25,3	13,2	13,1
	S28	44,4	10,06	19,8	8,8	8,7
	S29	45,4	10,02	20,4	8,8	8,7
	S30	48,7	11,48	24,3	5,3	5,3
	S31	50,7	11,43	24,6	5,3	5,3
	S32	49,1	11,51	28,6	8,7	8,7
	S33	47,7	10,70	26,2	6,0	6,0
	S34	48,8	11,14	23,4	7,3	7,3
	S35	46,8	11,16	23,4	7,3	7,3
	S36	49,7	10,80	24,9	3,9	3,9
	S37	47,4	10,75	25,8	3,9	3,9
	S38	47,8	10,16	25,0	4,2	4,2
	S39	48,6	10,20	26,4	5,0	5,0
	S40	49,0	11,31	28,1	5,4	5,4

S41	48,3	10,04	25,4	5,8	5,7
S42 (Produit certifié N°1)	49,7	9,99	25,4	5,8	5,7
S43 (Produit certifié N°1)	47,4	10,23	24,3	5,2	5,2
S44	50,1	9,98	23,9	9,9	9,8
S45	48,6	9,67	24,8	8,2	8,0
S46 (Produit de contrôle)	48,5	9,58	26,7	11,8	11,5
S47	49,7	10,16	22,6	8,7	8,6
S48	48,2	10,38	24,5	11,8	11,7
S49 (Produit certifié N°2)	47,3	10,43	25,1	9,1	9,0
S50 (Produit certifié N°2)	46,3	9,91	22,1	6,4	6,3
S51	48,6	11,32	23,4	11,8	11,8
S52	51,9	10,76	25,0	6,3	6,3
S53	50,1	10,74	25,0	6,3	6,3
S54	47,4	11,64	22,1	7,2	7,2
S55	51,1	11,40	26,3	7,1	7,1
S56	47,9	10,46	25,8	6,7	6,7
S57	50,4	9,86	25,8	8,2	8,1
S58	49,1	9,86	25,8	8,2	8,1
S59	47,4	9,68	23,7	15,9	15,6
S60	46,7	10,35	23,8	10,3	10,2
S61	46,1	10,45	22,0	6,8	6,8
S62	47,2	9,85	23,1	7,7	7,6
S63	48,1	9,71	24,4	6,4	6,3
S64	49,1	10,48	22,2	10,3	10,2
S65	44,7	10,28	20,2	8,1	8,1
S66	47,6	9,48	25,5	3,9	3,7
S67	50,4	10,13	25,7	10,4	10,3
S68	49,1	10,80	24,8	8,0	8,0
S69	47,5	11,11	23,1	10,0	10,0
S70	49,3	9,87	24,3	6,4	6,3
S71	47,4	10,28	23,8	11,8	11,7
S72	49,6	10,85	25,8	9,0	9,0
S73	48,2	10,90	24,3	8,0	8,0
S74	47,4	10,23	26,8	9,0	8,9
S75	50,0	9,10	27,4	4,0	3,7
S76	45,8	10,11	25,3	13,2	13,1
S77	50,6	11,40	23,9	5,6	5,6
S78	48,0	11,34	28,1	5,4	5,4
S79	47,5	10,27	24,3	5,2	5,2
S80	50,0	9,96	23,9	9,9	9,8
S81	49,4	9,65	24,8	8,2	8,0

	S82	48,5	9,55	26,7	11,8	11,5
	S83	48,7	10,15	22,6	8,7	8,6
	S84	47,6	10,36	24,5	11,8	11,7
	S85	47,2	10,45	25,1	9,1	9,0
	S86	45,5	9,95	22,1	6,4	6,3
	S87	48,2	11,29	23,4	11,8	11,8
	S88	47,5	11,62	22,1	7,2	7,2
	S89	50,8	11,42	26,3	7,1	7,1
	S90	47,9	10,41	25,8	6,7	6,7
	S91	47,4	9,70	23,7	15,9	15,6
	S92	47,6	10,38	23,8	10,3	10,2
	S93	45,7	10,36	22,0	6,8	6,8
	S94	47,2	9,89	23,1	7,7	7,6
	S95	50,0	9,69	24,4	6,4	6,3
	S96	49,1	10,47	22,2	10,3	10,2
Échantillons aberrants	S97	44,4	8,35	24,0	7,1	4,8
	S98 (Produit étranger N°1)	47,2	9,85	19,8	8,9	8,8
	S99	46,5	10,30	23,9	12,0	11,9
	S100	52,6	8,30	30,4	4,8	3,1
	S101	46,0	8,40	23,1	16,5	11,6
	S102 (Produit étranger N°2)	47,2	9,64	19,7	8,3	8,1
	S103 (Chemma traditionnelle 1)	47,5	10,28	26,3	2,4	2,4
	S104 (Chemma traditionnelle 1)	47,8	10,22	26,3	2,4	2,4
	S105 (Produit sans tabac)	39,9	4,75	4,8	0,0	0,0

Abréviations : NCT, Nicotine ; ps, Base de poids sec.

Pour le test de normalité, un graphique de probabilité normale et un test bilatéral de Kolmogorov-Smirnov à un niveau de signification de 5 % avec correction de Lilliefors ont été exploités. Ce dernier test compare la distribution empirique des données à la distribution théorique normale. Si la valeur de probabilité (Valeur p) obtenue est supérieure à 0,1, cela signifie que la distribution des données ne peut pas être rejetée et est considérée comme normale au niveau de signification de 0,05. Les résultats indiquent que tous les paramètres de qualité présentaient une distribution normale après l'élimination des valeurs aberrantes et l'application de la méthode de Kennard-Stone.

Tableau 3.2 : Profil d'échantillons de ST algérien basé sur cinq paramètres de qualité avant et après la suppression des valeurs aberrantes.

Allocation des échantillons	Paramètre (unité)	Ensemble	Nombre d'échantillons	Min.	Max.	Moy.	Med.	SD	Valeur p *
Avant élimination des valeurs aberrantes	Humidité (%)	Tous les échantillons	105	39,9	52,6	48,2	48,0	1,9	0,52
	pH		105	4,75	11,64	10,22	10,24	0,87	0,071
	Cendres (%)		105	4,8	30,4	24,3	24,4	2,7	0,024
	NCT totale (mg/g, ps)		105	0,0	16,5	8,1	8,1	2,9	0,37
	NCT libre (mg/g, ps)		105	0,0	15,6	7,9	8,0	2,8	0,58
Après élimination des valeurs aberrantes	Humidité (%)	Calibration	64	44,4	51,9	48,4	48,5	1,7	0,85
		Test	32	44,7	50,8	48,2	48,0	1,5	0,55
	pH	Calibration	64	9,00	11,64	10,34	10,25	0,61	0,68
		Test	32	9,10	11,62	10,36	10,28	0,63	0,45
	Cendres (%)	Calibration	64	19,8	28,6	24,7	24,8	1,8	0,97
		Test	32	20,2	28,1	24,4	24,3	1,7	0,99
	NCT totale (mg/g, ps)	Calibration	64	3,9	15,9	8,0	8,1	2,6	0,70
		Test	32	3,9	15,9	8,6	8,2	2,7	0,97
	NCT libre (mg/g, ps)	Calibration	64	3,7	15,6	7,9	8,0	2,6	0,70
		Test	32	3,7	15,6	8,5	8,1	2,7	0,96

* Valeurs de probabilité issues du test de Kolmogorov-Smirnov à deux queues pour la normalité. Une valeur de $p > 0,1$ indique que la distribution des données ne peut être rejetée de l'hypothèse de normalité au seuil de signification de 0,05.

Abréviations : Med, Médiane ; Moy, Moyenne ; NCT, Nicotine ; ps, Base de poids sec ; SD, Écart-type.

Sur un autre point, l'augmentation observée du pH et de la teneur en cendres parmi les produits commerciaux est normalement attribuée à l'ajout d'agents alcalins et de charges inorganiques, respectivement. Alors que l'augmentation de la teneur en nicotine totale est directement liée à la qualité du mélange des feuilles de tabac et inversement liée à la proportion d'autres ingrédients. Bien que la teneur en humidité soit moins pertinent car elle peut être manipulée pour altérer le poids final, elle peut parfois servir d'indicateur de la durée de conservation du produit et de ses conditions de stockage.

Une comparaison préliminaire de nos résultats sur la Chemma algérienne avec des produits du tabac à usage oral provenant de diverses régions du monde

étudiés dans [102], suggère une similarité potentielle avec le « Toombak » soudanais. Toutefois, des analyses supplémentaires, telles que la mesure des concentrations de TSNAs, sont nécessaires pour étayer cette conclusion.

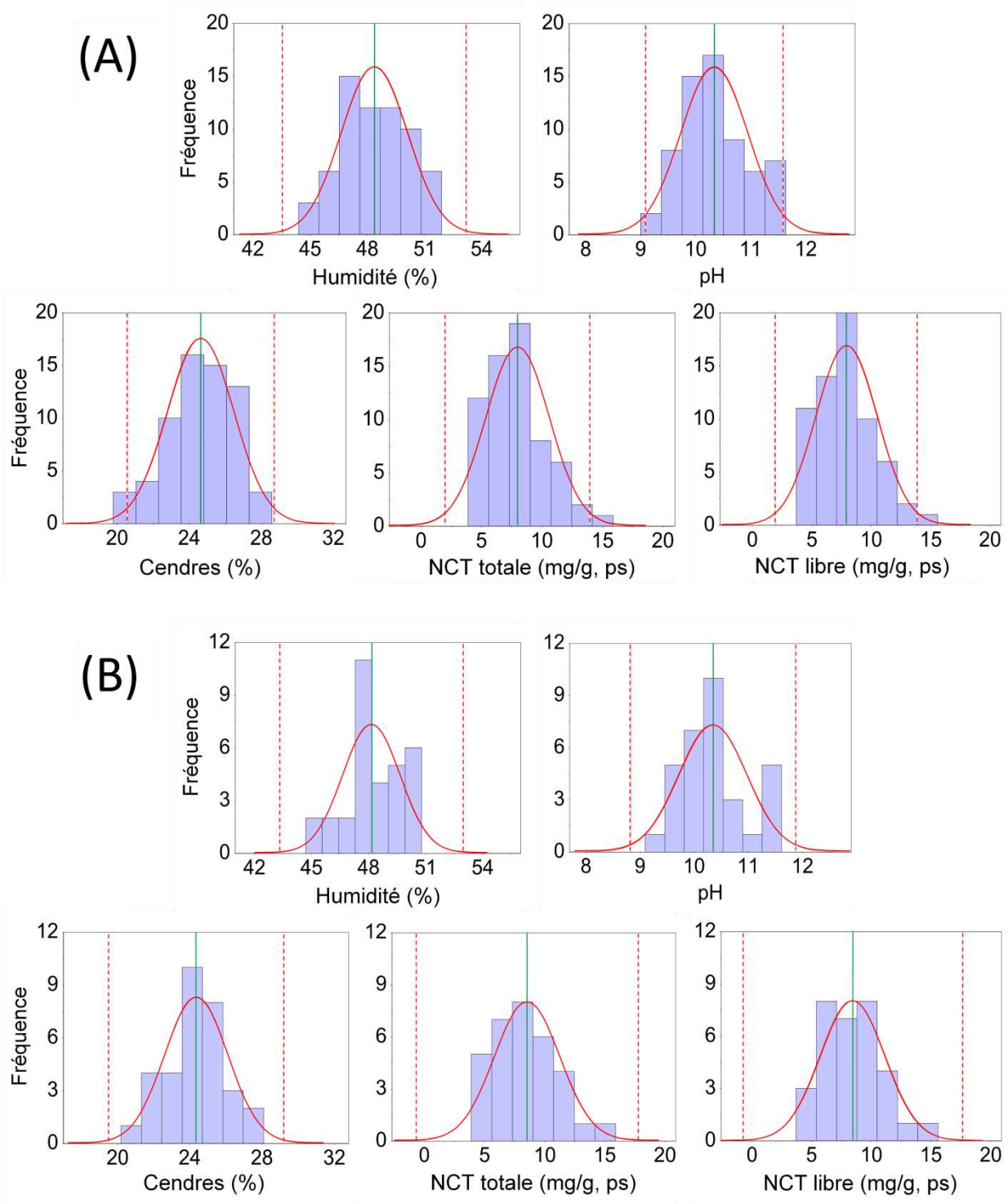


Figure 3.3 : Distribution de fréquence et courbe normale des teneurs en humidité, pH, cendres, nicotine totale et nicotine non ionisée des échantillons de ST algérien utilisés pour (A) l'entraînement et (B) le test des modèles.

3.3.2. Interprétation spectrale

Comme de nombreux produits de tabac à chiquer mondial, la Chemma peut contenir des matières végétales autres que le tabac et d'autres additifs. Au sein des plantes, des réseaux complexes de microfibrilles de cellulose, d'hémicelluloses, de lignine et de protéines structurelles sont présents [185]. Tous ces composants peuvent être détectés par spectroscopie MIR et contribuent collectivement à la complexité du spectre, rendant son interprétation qualitative difficile.

Grâce à l'application de traitements mathématiques simples tels que la soustraction ou le spectre de variance sur des échantillons appropriés, l'interprétation des résultats peut être considérablement améliorée. La Figure 3.4 présente les spectres non traités des 96 échantillons sélectionnés de ST. Comme on peut le voir, les régions présentant les variations les plus importantes entre les échantillons se situent entre 3700 - 2430 et 1860 - 615 cm^{-1}. La Figure 3.5 illustre les spectres moyen (A) et écart-type (B), à une échelle normalisée, comparés aux ingrédients les plus susceptibles d'être utilisés dans la préparation des ST algériens.

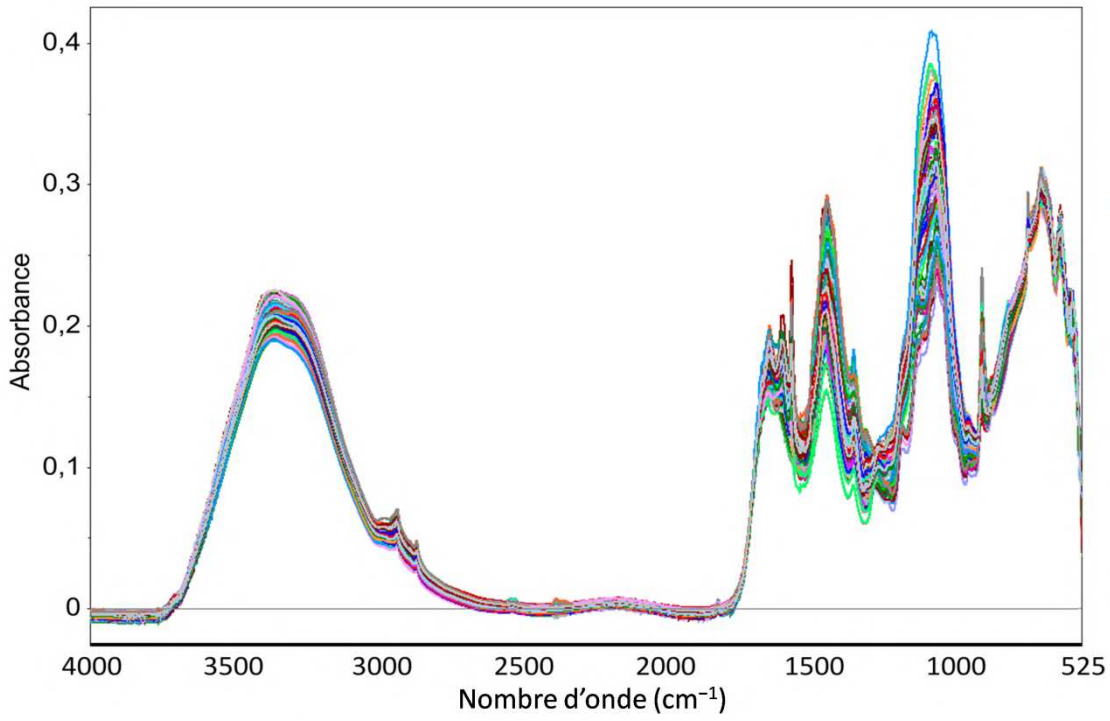


Figure 3.4 : Spectres FTIR-ATR non-traités obtenus pour les 96 produits sélectionnés du tabac à chiquer.

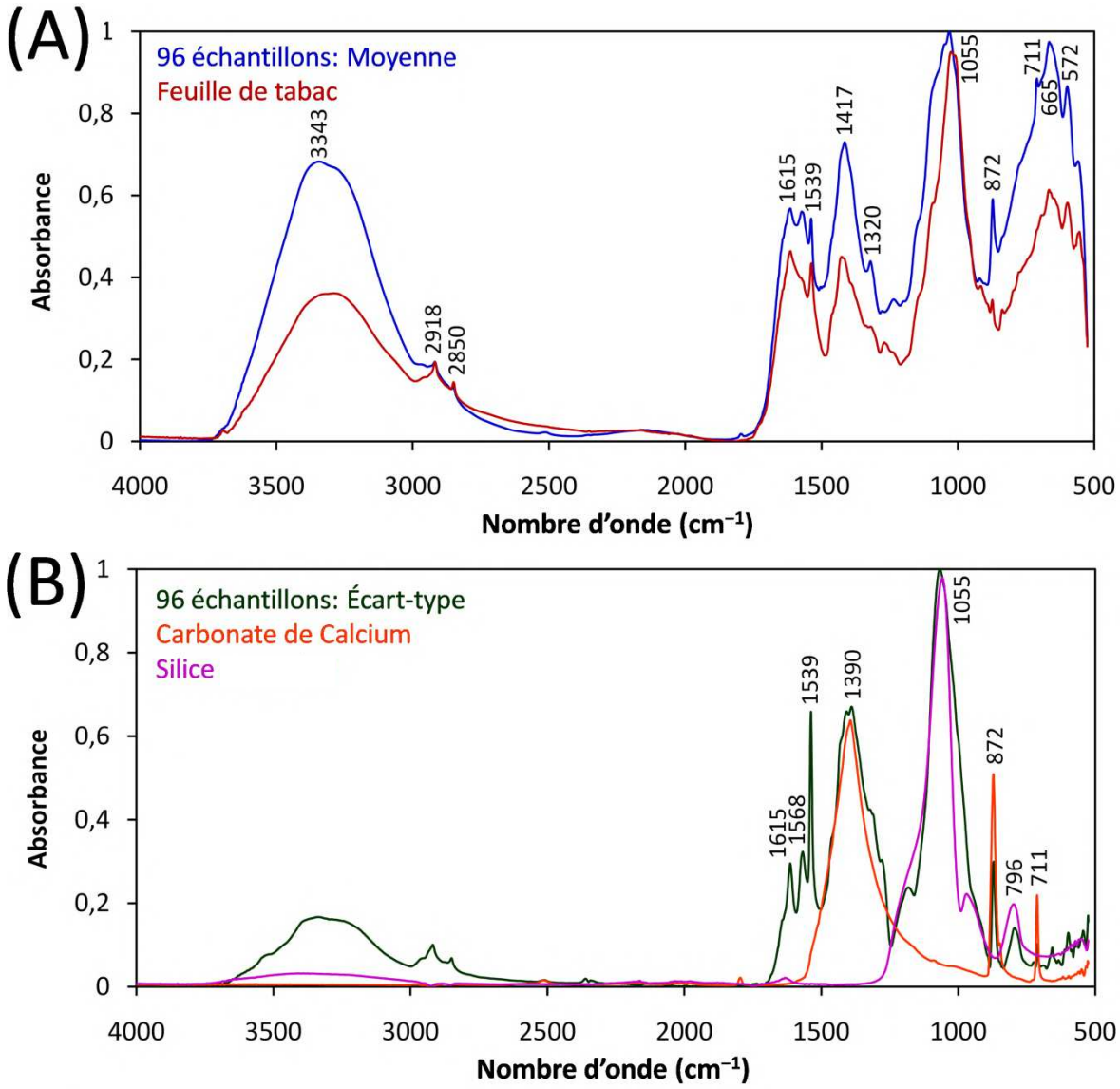


Figure 3.5 : Spectres ATR-FT-MIR représentatifs normalisés. Le moyen (A) et l'écart-type (B) des ST comparés aux ingrédients les plus susceptibles d'être utilisés dans leur préparation.

Plusieurs bandes d'absorption importantes peuvent être identifiées à partir du spectre moyen. Leurs attributions respectives [56, 185, 186] sont détaillées ci-dessous:

- Une bande d'absorption large et intense s'étendant de 3700 à 3000 cm^{-1}, culminant à 3343 cm^{-1}, peut être attribuée aux étirements O−H de l'eau, aux étirements O−H de la cellulose et des glucides, et aux étirements N−H des protéines dans les plantes.

- Les deux pics bien connus à 2918 et 2850 cm^{-1} proviennent des étirements asymétrique et symétrique du C−H des groupes méthyle et méthylène présents dans divers composés organiques.
- De petits pics à 2513 et 1795 cm^{-1} sont des indicateurs spécifiques de la présence de carbonate de calcium.
- Une bande faible et masquée observée à 1732 cm^{-1}, discernée à partir du spectre de la feuille de tabac pur, est probablement associée aux hémicelluloses.
- La bande légèrement large s'étendant de 1700 à 1600 cm^{-1}, centrée à 1615 cm^{-1}, est principalement attribuée à l'étirement C=C dans la lignine, et éventuellement dans la pectine, certainement chevauchée par la déformation O−H de l'eau (1640 cm^{-1}).
- En examinant la variance des pics dans le spectre d'écart-type (SD), des bandes caractéristiques à 1568 et 1539 cm^{-1}, pouvant être superposées aux vibrations d'absorption de la lignine, sont liées à la déformation N−H et à l'étirement C−N dans l'amide II. Ce lien est soutenu par l'observation que les plantes (ou les produits commerciaux) plus âgées présentaient généralement une teneur en protéines plus faible.
- Dans la gamme spectrale de 1500 à 1340 cm^{-1}, plusieurs bandes chevauchées sont discernables. Ces bandes sont attribuables aux vibrations de déformation de C−O−H et de C−H, ainsi qu'à l'étirement asymétrique de C−N−C dans la cellulose et la lignine. De plus, elles coïncident avec l'étirement asymétrique C−O de l'ion carbonate (CO_3^{2-}) à environ 1390 cm^{-1}.
- Le pic d'intensité moyenne à 1320 cm^{-1} est probablement corrélé à la déformation C−H de la cellulose, à l'étirement C−N dans l'amide I des protéines ou à divers composants végétaux, y compris la lignine, la xyloglucane ou les composés aliphatiques de type cire.
- Une large bande intense (1270 - 900 cm^{-1}) avec un pic central à 1055 cm^{-1}, confirmée par comparaison avec le spectre SD comme étant attribuée à l'étirement asymétrique Si−O de la silice. Cette gamme englobe également des chevauchements importants de pics interférents à 1147, 1093 et 1031 cm^{-1} correspondant aux étirements C−C, C−O, C−O−C et aux vibrations de flexion C−O−H dans la cellulose, la lignine et d'autres glucides. Malgré la variabilité

observée dans cette région, l'emplacement du pic et les caractéristiques spectrales globales, y compris la présence d'une bande supplémentaire à 796 cm^{-1}, ressemblent étroitement au spectre de référence de la silice pure et apportent un soutien supplémentaire à notre assignation.

- La région entre 1000 et 525 cm^{-1} est remarquable par son décalage de ligne de base dans tous les échantillons contenant de l'eau, ainsi que par les effets de surdimensionnement des nombres d'onde plus faibles dus à la réflexion du cristal ATR. La présence d'un nombre substantiel de bandes faibles provenant des constituants mineurs communs des plantes, y compris les glucides (915 - 840 cm^{-1}), les protéines (780 cm^{-1}) et les composés minéraux (615 - 550 cm^{-1}), aux côtés des constituants majeurs, complique l'interprétation visuelle de cette région. Exceptionnellement, deux pics étroits distincts à 872 et 711 cm^{-1} sont respectivement identifiés comme étant attribués aux vibrations hors du plan et dans le plan de l'ion CO_3^{2-}.

3.3.3. Analyses de classification

3.3.3.1. Méthodes non supervisées

Lorsqu'il s'agit de produits illicites, les mesures quantifiables traditionnelles telles que la popularité ou la part de marché sont souvent difficiles à obtenir en raison de la nature clandestine de ces marchés. À la place, des facteurs tels que la qualité perçue ou la réputation au sein des marchés souterrains prennent de l'importance. La classification des échantillons sur la base de ces facteurs peut s'avérer cruciale pour le contrôle de la conformité. Cependant, l'intégration directe d'évaluations qualitatives ou de perceptions parfois subjectives dans un modèle d'apprentissage supervisé peut aboutir à des classifications moins fiables que les modèles fondés sur des mesures physicochimiques objectives.

En revanche, il est évident que deux échantillons indépendants n'ont jamais des spectres FTIR similaires à moins qu'ils ne soient identiques, ce qui peut être extrêmement avantageux dans les tâches de discrimination. Toutefois, l'application de méthodes non supervisées, telles que la PCA, aux différences

spectrales de produits de type anonyme risque de regrouper les échantillons sur la base d'informations non pertinentes contenues dans les spectres.

Pour relever ces défis, nous avons proposé une approche en deux étapes qui combine l'apprentissage non supervisé et l'apprentissage supervisé. Cela implique d'appliquer des méthodes non supervisées aux mesures physicochimiques de référence afin d'extraire des caractéristiques informatives, qui sont ensuite utilisées comme données d'entrée pour le modèle d'apprentissage supervisé. Cette méthodologie permet de créer un modèle à la fois ancré dans des données empiriques et moins sensible aux biais et aux incertitudes des catégories étiquetées par l'homme. De plus, elle ouvre la possibilité de réaliser une classification future directement à partir de mesures spectrales, même lorsque les mesures de qualité traditionnelles ne sont pas disponibles.

Dans un premier temps, une PCA a été appliquée à l'ensemble des 96 échantillons afin d'étudier la tendance générale des produits du ST local par rapport à l'échantillon de contrôle. Un graphique des scores à deux dimensions (2-d) des deux premières PCs, qui expliquaient 71 % de la variance totale, a été principalement généré (Figure 3.6 ci-dessous).

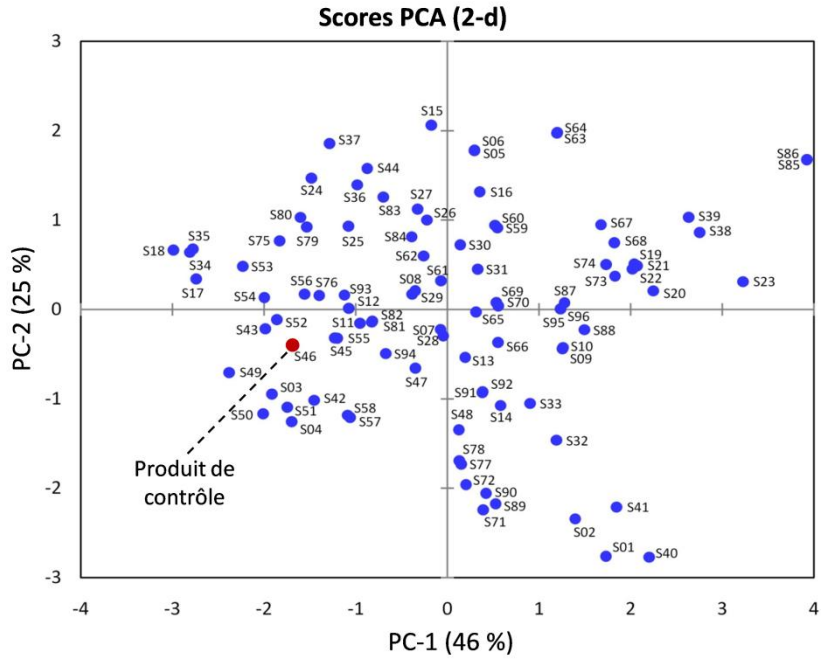
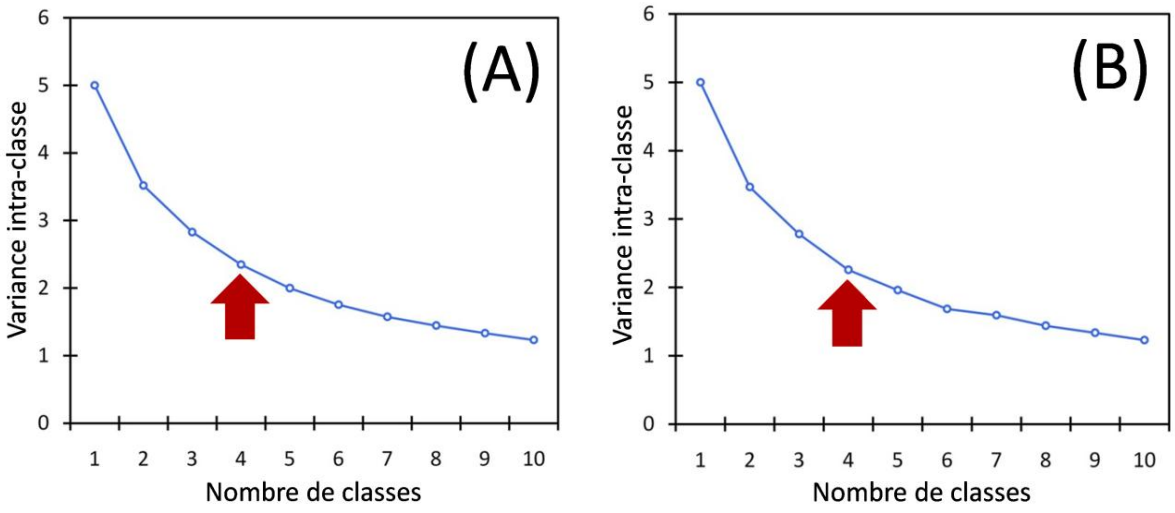


Figure 3.6 : Graphique des scores PCA obtenus sur la base des mesures de référence pour l'étude de la tendance générale des échantillons sélectionnés.

Au moins deux groupes peuvent être différenciés sur les parties positives et négatives de la PC-1, et au plus, trois groupes peuvent être à peine distingués en divisant la partie positive de la PC-1 en deux sous-groupes. Globalement, aucun regroupement clair d'échantillons ou séparation nette des données n'a été perçu entre ces groupes, en particulier sur les axes.

Par conséquent, afin d'avoir une idée du nombre approprié de classes, il convient d'effectuer une analyse AHC. Le dendrogramme AHC a été créé en utilisant le même ensemble de données (mesures de qualité de référence) précédemment utilisé pour la PCA. Après plusieurs essais, une partition satisfaisante des observations a été obtenue en utilisant la méthode de Ward. Cette dernière méthode a été signalée comme offrant les meilleurs résultats dans des études antérieures [187]. La troncature automatique du dendrogramme en fonction de la valeur d'entropie calculée montre une valeur de dissimilarité élevée (égale à 60), divisant tous les échantillons en trois classes principales. Ce résultat est en accord avec les résultats précédents de la PCA à bien des égards. Néanmoins, l'examen des variances intra-classes (Figure 3.7A), à la recherche du "coude", suggère de mieux tronquer les données en quatre classes (Figure 3.8).



Figures 3.7 : Évolution de la variance intra-classe pour (A) le regroupement hiérarchique et (B) la classification *k*-means. Un "coude" est observé pour le nombre approprié de classes.

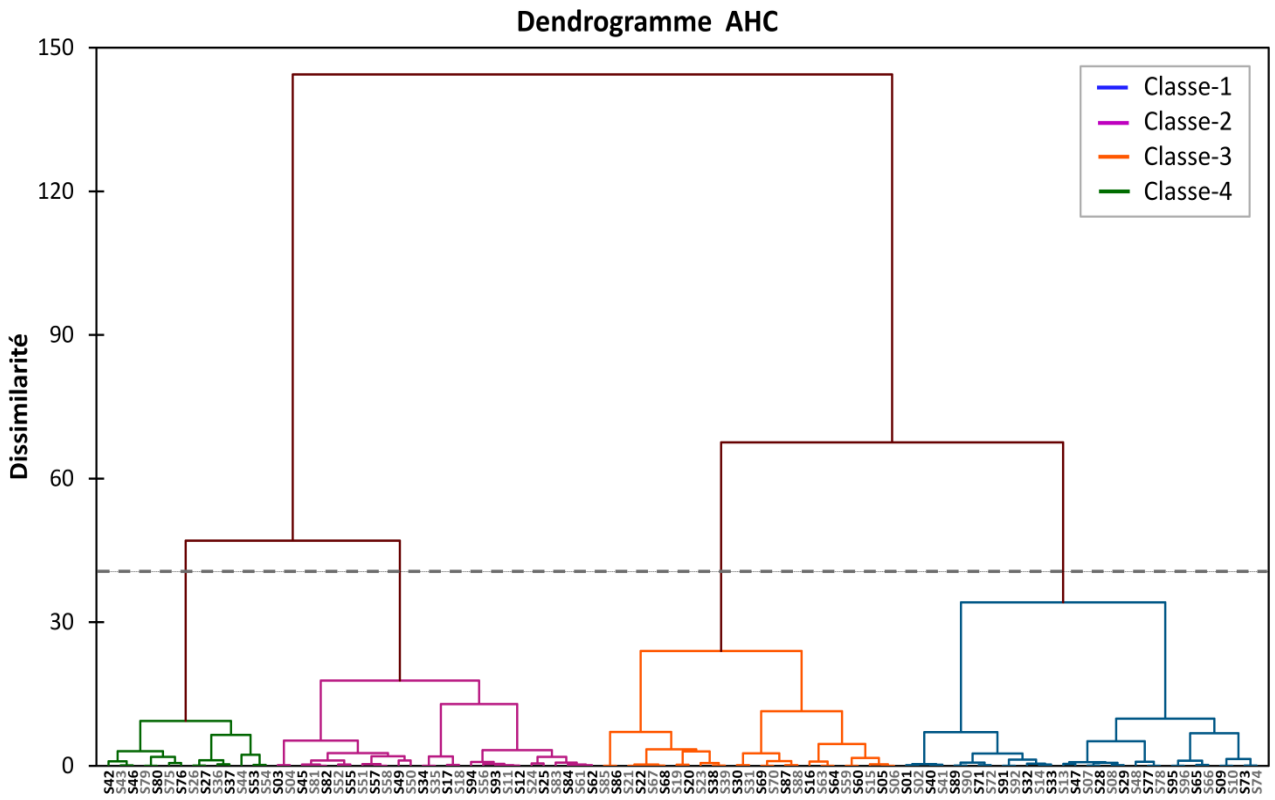


Figure 3.8 : Dendrogramme AHC représentant la hiérarchie des échantillons commerciaux de Chemma obtenus sur la base des mesures de référence. La ligne en pointillé indique le dendrogramme tronqué au niveau de quatre classes.

La Classe-1, composée de 30 observations, a été la première discriminée sur le bras droit lointain et comprend les produits de contrefaçon les moins chers et impopulaires. Fusionnée avec la Classe-1 au premier niveau, la Classe-3 contenait les marques illégales les plus populaires à bas prix, tout en comprenant 25 observations. Le deuxième niveau du dendrogramme, incorporant des échantillons montrant une grande similitude avec les deux marques certifiées de l'UTC, a été divisé en deux sous-groupes. Le premier sur le bras droit, désignés sous le nom Classe-2, se compose de 27 observations et représente des produits de contrefaçon plus chers et abondants, analogues au produit commercial certifié N°2. Les derniers échantillons séparés, désignés comme Classe-4, comprenaient l'échantillon de contrôle, le produit certifié N°1, les produits analogues contrefaits et les échantillons de contrefaçon partageant des caractéristiques physicochimiques similaires. Ces résultats étaient inattendus car on s'attendait à ce que les deux marques authentiques se regroupent en raison de leurs constituants fondamentalement identiques prédéterminés par le fabriquant (UTC).

Dans une troisième étape, un partitionnement en *k*-moyennes a été réalisé pour évaluer davantage les résultats obtenus avec l'AHC. Le critère couramment utilisé dans les classifications *k*-means, la Trace(W) (matrice des sommes de carrés et produits croisés regroupés), a fourni le résultat le plus favorable dans ce contexte. La minimisation de la Trace(W) minimise la variance intra-classe totale pour un nombre donné de classes, et donc l'hétérogénéité des groupes. En effet, cette technique a permis d'obtenir une inertie plus faible (2,26) par rapport à l'AHC (2,35), recommandant ainsi quatre classes pour l'ensemble de données étudié (Figure 3.7B). Le taux de concordance des échantillons, qui correspond au pourcentage d'échantillons affectés à la même catégorie par les deux méthodes, était de 83,3 %. Il est intéressant de noter que ces regroupements *k*-means reflètent étroitement les structures sous-jacentes associées aux distinctions de qualité entre les différentes marques. Cela est confirmé par un taux de concordance de 90,6 % entre la classification *k*-means et une catégorisation de référence basée sur trois attributs qualitatifs clés des produits: le prix, la réputation et la qualité perçue. Nous pensons que l'écart restant de 9,4 % pourrait être dû à de la confusion ou à des tentatives trompeuses de la part des vendeurs. Le modèle *k*-means, qui s'appuie sur des mesures physico-chimiques objectives, a probablement été capable de détecter ces divergences.

Bien que la classification *k*-means ait démontré des performances supérieures, elle ne permet pas de visualiser la proximité entre les classes ou les observations, une capacité que possèdent l'AHC et la PCA. En conséquence, les résultats de *k*-means ont été exploités qualitativement pour afficher les observations en différentes couleurs, localiser les centroïdes et présenter des ellipses de confiance autour de chaque classe. De plus, le fait de s'appuyer exclusivement sur les deux premières PCs s'est avéré insuffisant pour capturer entièrement la complexité des données. Dans ce cas là, l'incorporation du PC-3 a permis d'améliorer sa représentation, expliquant plus de 91 % de la variance totale.

Un examen approfondi de la catégorisation de référence et les sorties des méthodes d'apprentissage (Figures 3.9 A-D) a révélé des résultats qui, bien que pas radicalement différents, offraient des informations plus significatives par rapport à l'AHC. Notamment, les produits commerciaux authentiques sont

logiquement regroupés dans la même classe (Classe-4), ainsi que leurs analogues contrefaits, et seul un petit nombre d'échantillons ont présenté un changement de classe, ce qui a conduit à une homogénéité améliorée au sein des groupes. Essentiellement, les produits regroupés avec des marques authentiques au sein du même groupe sont ceux qui ont réussi le mieux à imiter les caractéristiques du produit authentique. Cela suggère que ces produits ont reproduit efficacement la formule originale, ce qui pourrait poser un problème pour les distinguer des marques authentiques en se basant uniquement sur les paramètres de qualité évalués. En excluant quelques échantillons qui peuvent être considérés comme atypiques, ces résultats renforcent les conclusions tirées de l'analyse AHC. En outre, les Figures 3.9C et D représentant respectivement les scores PC-2 contre PC-1 et PC-3 contre PC-1 ont permis de distinguer la Classe-1 de la Classe-3 et la Classe-4 de la Classe-2. D'après le tracé en 3-d des scores PCA des trois premières PCs (Figure 3.10A), la distribution des données présente un motif uniforme ressemblant à un tétraèdre régulier.

Maintenant, pour une compréhension approfondie des variations physicochimiques qui ont conduit à la séparation observée, les coefficients PCA (« loadings » en anglais) assument un rôle crucial en quantifiant l'influence individuelle de chaque variable (paramètre) sur les PCs désignées. Les coefficients PCA correspondants ont été représentés avec les centroïdes des classes dans un graphique en 3-d, comme le montre la Figure 3.10B.

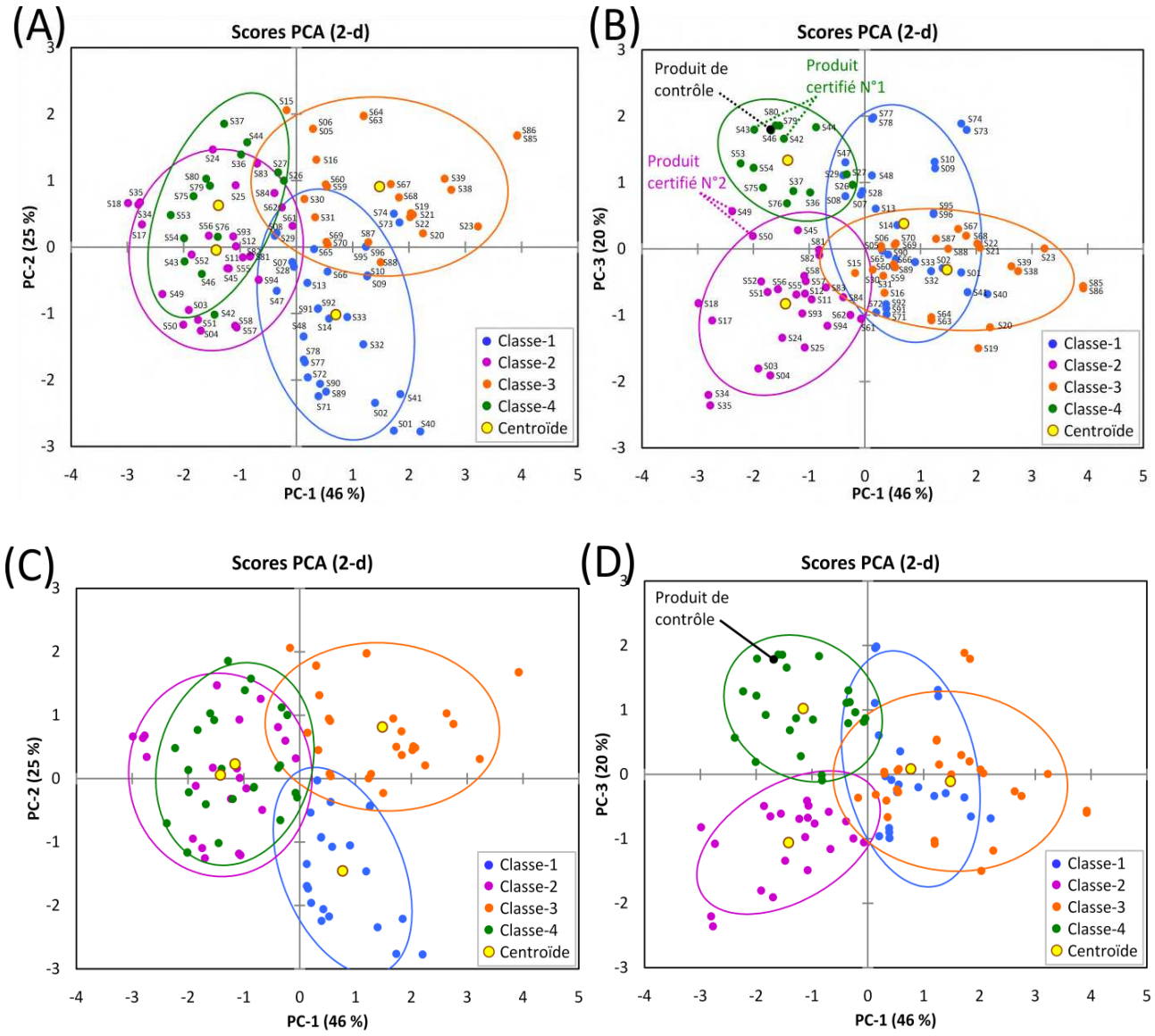


Figure 3.9 : Graphiques des scores PCA affichant le regroupement des échantillons selon les résultats de l'AHC (A et B) et les résultats de *k*-means (C et D). (Les ellipses sont à un intervalle de confiance de 80%.)

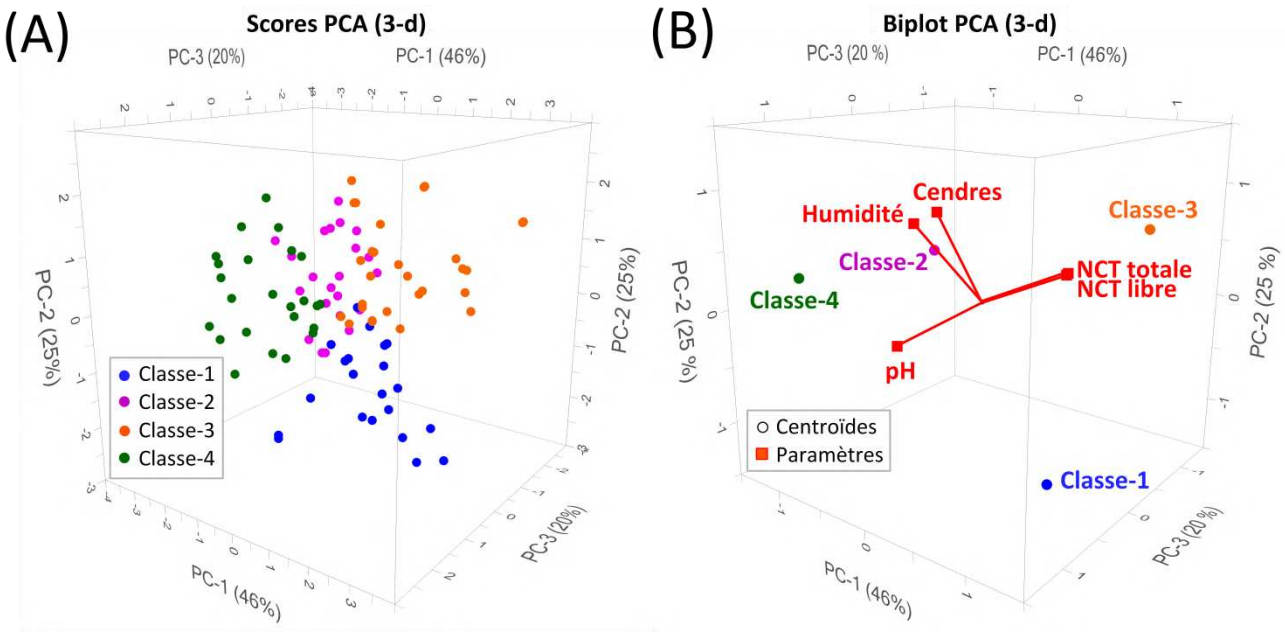


Figure 3.10 : (A) Scores PCA pour les trois premières PCs affichant le partitionnement des échantillons selon les résultats de *k*-means. (B) Biplot des scores et des coefficients PCA.

Sur ce graphique, le type de corrélation peut être déduit à partir des angles, soit entre les variables, soit entre les variables et les axes PCA. Comme on peut le voir, la nicotine était principalement responsable du regroupement des échantillons à travers la PC-1. On note également que la NCT totale et la NCT libre coïncidaient, ce qui s'explique par le fait que la majorité des produits avaient des valeurs de pH > 9.4, à ce niveau, 96 % de la nicotine est sous sa forme non ionisée. L'humidité et les cendres étaient positivement corrélées et contribuaient toutes deux au PC-2, avec des variations indépendantes envers la nicotine. Cela suggère que les teneurs en cendres totales sont principalement liées au "kieselguhr" inorganique utilisé comme agent hygroscopique et comme charge dans la fabrication de la Chemma. Le kieselguhr, composé jusqu'à 90 % de silice, est cohérent avec la bande d'étirement Si−O distincte observée dans les spectres FTIR. Le pH était fortement chargé sur la PC-3, et il était perpendiculaire à la fois aux vecteurs de nicotine et d'humidité / cendres, ce qui signifie qu'ils ne sont pas liés entre eux.

Dans le but d'interpréter les caractéristiques de chaque classe, les échantillons situés dans la même direction qu'un paramètre donné ont des valeurs élevées pour ce paramètre. Les échantillons de la Classe-3 présentent des concentrations extrêmes en NCT et des niveaux de pH relativement plus faibles. La Classe-1 présente des concentrations moyennes en NCT et des teneurs en humidité et en cendres les plus faibles. Étant donné leur alignement avec la PC-3 (fortement chargée en pH), les échantillons de la Classe-4 présentent les valeurs de pH les plus élevées, indiquant leur nature alcaline. Dans d'autres aspects, la Classe-4 partage des similitudes avec la Classe-2, comme une faible teneur en NCT et des teneurs élevées en humidité et en cendres. Ces résultats confirment une fois de plus que non seulement la proportion d'autres ingrédients affecte la concentration de nicotine dans un produit final, mais aussi la variété de tabac utilisée.

3.3.3.2. Méthodes supervisées

En dernière étape, les sorties de *k*-means ont été utilisées pour entraîner deux méthodes supervisées, à savoir l'analyse discriminante par moindres carrés partiels (PLS-DA) et la classification par machine à vecteurs de support (SVM-C). Cette approche permet de tirer parti du regroupement non supervisé basé sur les mesures physicochimiques et de le traduire en une tâche de classification supervisée. Malgré la limitation potentielle selon laquelle les groupes de *k*-means pourraient ne pas refléter parfaitement les catégories de référence, cette méthode fournit un point de départ fiable pour une caractérisation plus objective des échantillons.

La PLS-DA repose essentiellement sur les mêmes principes que la PLSR, bien que la tâche soit différente (catégorielle dans le cas de la PLS-DA). La différence clé entre la SVMR et la SVM-C réside dans leurs tâches et objectifs fondamentaux (Revoir le Chapitre 1, **sous-sections 1.3.3.2 et 1.3.3.3**).

De manière similaire à une régression, divers prétraitements spectraux ont été testés (Tableau 3.3). Le pouvoir discriminant des modèles a été évalué en calculant l'exactitude (rapport des échantillons correctement classés sur le nombre total d'échantillons) de la calibration et de la CV.

Tableau 3.3 : Performances de discrimination dans les procédures de calibration et de CV en utilisant les spectres FTIR avec différentes méthodes de prétraitement.

Méthode de classification	Prétraitement spectral	Valeur de C	No. de LVs / SVs	Exactitude (%)	
				Calibration	CV
PLS-DA	None	–	16 §	86,5	65,6
	SNV	–	16 §	85,4	65,6
	EMSC	**–**	**14 §**	**85,4**	**67,8**
	SG FD	–	10 §	83,3	60,4
SVM-C *	None	100	53	95,8	81,3
	BO	100	50	94,8	80,2
	SNV	**1**	**53**	**95,8**	**84,4**
	EMSC	100	54	93,8	84,4

§ Valeur sélectionnée en utilisant le critère de Haaland (premier modèle avec $p < 0,75$ pour le rapport $PRESS_k$ / min(PRESS)).

* Seule la fonction de noyau linéaire a été utilisée dans la SVM-C.

Les modèles optimaux sont en gras.

Abréviations : BO, Correction du décalage de la ligne de base ; C, Fonction de coût dans les modèles SVM ; EMSC, Correction étendue de la diffusion multiplicative ; No. de LVs / SVs, Nombre de variables latentes dans les modèles PLS ou de vecteurs de support dans les modèles SVM ; SG FD, 1ère dérivée de Savitzky-Golay à 9 points / côté avec un ordre polynomial de deuxième degré ; SNV, Standard Normal Variate.

Les matrices de confusion (Figure 3.11) révèlent que les deux modèles prédictifs ont rencontré des difficultés à regrouper les échantillons de la Classe-1 ; cela est peut-être dû à la grande similarité entre certains échantillons ou à la complexité de la composition des produits. Notamment, malgré l'utilisation d'un nombre élevé de LVs (n = 14) dans la PLS-DA pour capturer la complexité des données, la SVM-C a surpassé, atteignant des pourcentages d'exactitude de calibration / validation égaux à 95,8 / 84,4 contre 85,4 / 67,7 pour la PLS-DA. Cela signifie que la SVM-C a le potentiel d'être un outil précieux pour la classification future d'échantillons inconnus basée sur leurs spectres MIR, sans recours à des jugements subjectifs.

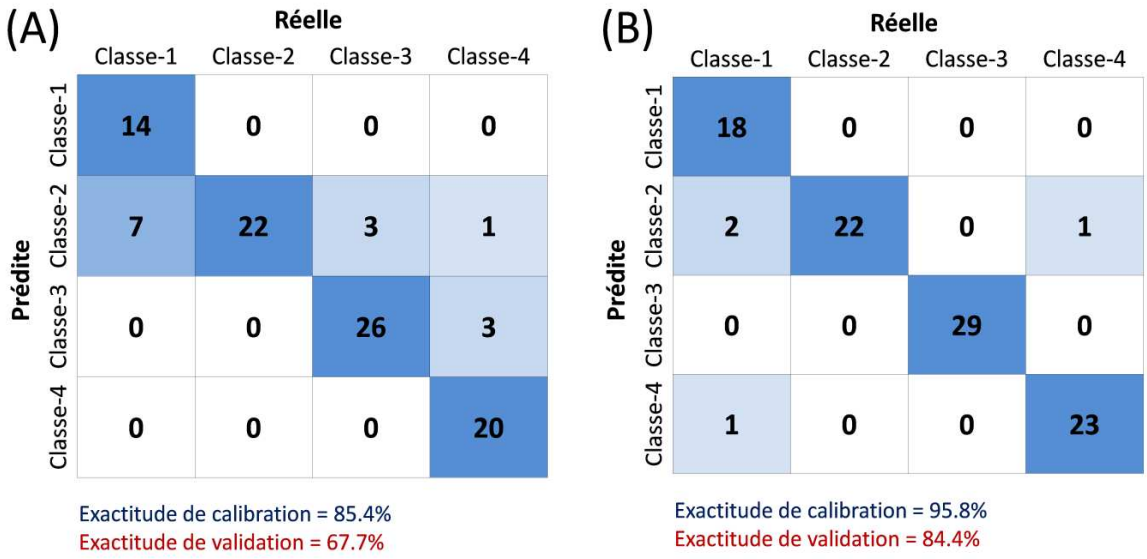


Figure 3.11 : Matrices de confusion de l'ensemble d'entraînement des modèles (A) PLS-DA et (B) SVM-C utilisés pour la classification des produits commerciaux.

3.3.4. Prédiction des paramètres de qualité

Dans la partie principale de cette étude, des modèles PLSR et SVMR ont été entraînés et validés pour déterminer cinq paramètres de qualité, à savoir la teneur en humidité, le pH, les cendres, la nicotine totale et la nicotine libre dans les produits commercialisés de Chemma. La détermination des trois premiers paramètres physicochimiques peut être établie de manière plus directe à partir des spectres FT-MIR, étant donné que plusieurs bandes correspondantes ont déjà été distinguées (**sous-section 3.3.2**). En revanche, en raison de sa faible abondance dans une matrice complexe, la détermination de la nicotine à partir des spectres des échantillons entiers est principalement indirecte. Cette détermination indirecte repose sur la recherche d'une corrélation entre les bandes spectrales indiquant le matériel végétal du tabac et la teneur globale en nicotine. Étant donné que le tabac contient naturellement de la nicotine, la présence et la quantité de la plante de tabac dans le spectre peuvent servir de marqueur de substitution pour estimer le niveau de nicotine dans l'échantillon. Les résultats détaillés du processus d'optimisation sont présentés dans les Tableaux 3.4 et 3.5 pour la PLSR et la SVMR, respectivement.

Tableau 3.4 : Paramètres de régression des procédures de calibration, de validation croisée et de prédiction des paramètres de qualité pour les modèles calculés par PLSR en utilisant différents prétraitements spectraux et méthodes de sélection des variables spectrales.

Paramètre de qualité (unité)	Pré-traitement spectral	Gamme spectrale (cm^{-1})	LVs	Calibration		Validation croisée		Prédiction	
				RMSEC	R^2_C	RMSECV	R^2_{CV}	RMSEP	R^2_P
Humidité (%)	Aucun	3700-2430	9 *	0,918	0,7132	1,34	0,4063	1,06	0,5039
		1860-615	7 §	0,991	0,6651	1,40	0,3532	1,07	0,4992
	BO	3700-2430	8 *	0,983	0,6710	1,38	0,3722	1,07	0,5002
		1860-615	5 §	1,17	0,5350	1,49	0,2678	1,37	0,1730
	UVN	3700-2430	8 *	0,922	0,7102	1,33	0,4135	1,07	0,4967
		1860-615	7 §	0,965	0,6830	1,39	0,3620	1,10	0,4735
	RN	3700-2430 1860-615	7 *§	0,929	0,7061	1,28	0,4556	1,11	0,4585
	EMSC	3700-2430	8 *	0,951	0,6917	1,34	0,4076	0,994	0,5675
		1860-615	6 §	1,09	0,5923	1,45	0,3082	1,14	0,4284
	SNV	3700-2430	8 *	0,951	0,6920	1,37	0,3829	1,03	0,5399
		1860-615	5 §	1,16	0,5407	1,45	0,3013	1,33	0,2244
	SG FD	3700-2430	7 *	0,805	0,7790	1,35	0,3958	1,04	0,5264
		1860-615	5 §	1,02	0,6470	1,40	0,3517	1,13	0,4446
		i-PLS 3700-3440 1830-1080	7 * 5 §	1,03 1,15	0,6416 0,5506	1,47 1,48	0,2869 0,2782	1,12 1,25	0,4479 0,3176
		VIP 1703-841 721-615	7 * **5 §**	0,902 **1,04**	0,7226 **0,6300**	1,37 **1,39**	0,3847 **0,3608**	1,06 **1,14**	0,5131 **0,4358**
	SG SD	3700-2430 1860-615	1 *§	1,53	0,2049	1,59	0,1624	1,43	0,1034
	EMSC-SG FD	3700-2430	6 *	0,928	0,7063	1,36	0,3883	0,988	0,5731
		1860-615	4 §	1,13	0,5664	1,44	0,3122	1,29	0,2739
pH	Aucun	3700-2430	8 *	0,210	0,8818	0,290	0,7771	0,198	0,8981
		1860-615	5 §	0,262	0,8114	0,334	0,7035	0,269	0,8118
	BO	3700-2430	8 *	0,207	0,8824	0,285	0,7839	0,195	0,9011
		1860-615	5 §	0,257	0,8194	0,324	0,7206	0,287	0,7855
	UVN	3700-2430	7 *	0,211	0,8785	0,285	0,7836	0,193	0,9031
		1860-615	5 §	0,232	0,8523	0,300	0,7607	0,257	0,8289
	RN	3700-2430	8 *	0,199	0,8918	0,276	0,7980	0,181	0,9154
		1860-615	6 §	0,229	0,8557	0,299	0,7625	0,252	0,8349
	EMSC	3700-2430	7 *	0,212	0,8765	0,282	0,7884	0,203	0,8926
		1860-615	5 §	0,246	0,8336	0,305	0,7527	0,270	0,8111
	SNV	3700-2430	6 *	0,222	0,8653	0,286	0,7821	0,233	0,8587
		1860-615	5 §	0,239	0,8433	0,300	0,7613	0,268	0,8133
	SG FD	3700-2430	4 *	0,227	0,8585	0,285	0,7840	0,233	0,8588
		1860-615	3 §	0,250	0,8283	0,301	0,7599	0,255	0,8307
		***i*-PLS 1700-680**	4 * **3 §**	0,231 **0,248**	0,8543 **0,8311**	0,283 **0,297**	0,7880 **0,7663**	0,236 **0,251**	0,8557 **0,8362**
		VIP 1680-637	4 * 3 §	0,229 0,250	0,8561 0,8291	0,283 0,299	0,7868 0,7619	0,229 0,254	0,8561 0,8329
	SG SD	3700-2430	5 *	0,195	0,8957	0,346	0,6828	0,238	0,8530
		1860-615	3 §	0,295	0,7620	0,384	0,6088	0,319	0,7352
	SG FD-SNV-DT	3700-2430	5 *	0,192	0,8992	0,265	0,8136	0,195	0,9015
		1860-615	3 §	0,244	0,8365	0,269	0,8116	0,295	0,7690
Cendres (%)	Aucun	3700-2430	3 *	1,32	0,4749	1,45	0,3774	1,06	0,6266
		1860-615	2 §	1,35	0,4472	1,49	0,3492	1,04	0,6418
	BO	3700-2430	3 *	1,29	0,4933	1,41	0,4127	1,05	0,6334
		1860-615	2 §	1,32	0,4705	1,45	0,3780	1,02	0,6584
	UVN	3700-2430	3 *	1,31	0,4756	1,43	0,3949	1,04	0,6455
		1860-615	2 §	1,38	0,4187	1,49	0,3436	1,05	0,6379
	EMSC	**3700-2430**	3 *	1,30	0,4876	1,42	0,4037	1,05	0,6352
		1860-615	**2 §**	**1,33**	**0,4611**	**1,44**	**0,3921**	**1,02**	**0,6582**

		i-PLS	3 *	1,26	0,5167	1,44	0,3901	1,22	0,5083
		2600-1760 1500-615	2 §	1,42	0,3908	1,50	0,3412	1,23	0,4989
		VIP 1625-1523 1442-1284 1188-982	2 *§	1,43	0,3765	1,50	0,3344	1,17	0,5547
	SNV	3700-2430 1860-615	3 *	1,32	0,4745	1,43	0,3969	1,04	0,6418
			2 §	1,39	0,4154	1,49	0,3461	1,05	0,6327
	SG FD	3700-2430 1860-615	3 *	1,24	0,5269	1,43	0,3945	1,08	0,6150
			1 §	1,43	0,3777	1,51	0,3286	1,24	0,4877
	SG SD	3700-2430 1860-615	8 *	0,373	0,9577	1,26	0,5311	1,08	0,6139
			5 §	0,670	0,8636	1,35	0,4653	0,935	0,7110
	SNV-DT	3700-2430 1860-615	2 *§	1,34	0,4551	1,44	0,3866	1,02	0,6541
	EMSC-DT	3700-2430 1860-615	3 *	1,30	0,4876	1,42	0,4037	1,05	0,6352
			2 §	1,33	0,4611	1,44	0,3921	1,02	0,6582
Nicotine totale (mg/g, ps)	Aucun	3700-2430 1860-615	6 *	1,27	0,7586	1,57	0,6408	1,43	0,7271
			10 §	0,932	0,8701	1,50	0,6744	1,32	0,7676
	BO	3700-2430 1860-615	5 *	1,30	0,7473	1,57	0,6442	1,45	0,7216
			10 §	0,920	0,8733	1,50	0,6733	1,35	0,7563
	UVN	3700-2430 1860-615	5 *§	1,26	0,7622	1,56	0,6467	1,42	0,7330
	EMSC	3700-2430 1860-615	5 *§	1,28	0,7552	1,52	0,6630	1,41	0,7364
	SNV	3700-2430 1860-615	5 *	1,25	0,7677	1,53	0,6609	1,42	0,7325
			4 §	1,32	0,7386	1,58	0,6385	1,48	0,7095
		i-PLS 2997-2789 1754-1502 1339-701	**4 *§**	**1,27**	**0,7586**	**1,46**	**0,6891**	**1,37**	**0,7499**
		VIP 1669-965 714-689	4 *§	1,28	0,7564	1,50	0,6725	1,42	0,7303
	SG FD	3700-2430 1860-615	11 *	0,417	0,9739	1,27	0,7643	1,10	0,8397
			9 §	0,572	0,9510	1,38	0,7218	1,06	0,8508
	SG SD	3700-2430 1860-615	4 *§	1,14	0,8066	1,73	0,5649	1,49	0,7027
	SNV-SG FD	3700-2430 1860-615	3 *	1,21	0,7826	1,52	0,6636	1,32	0,7691
			9 §	0,569	0,9516	1,41	0,7130	1,02	0,8620
	SNV-DT	3700-2430 1860-615	5 *	1,23	0,7726	1,51	0,6691	1,41	0,7358
			4 §	1,29	0,7503	1,54	0,6564	1,43	0,7267
Nicotine non ionisée (mg/g, ps)	Aucun	3700-2430 1860-615	6 *	1,26	0,7557	1,57	0,6353	1,41	0,7295
			10 §	0,917	0,8709	1,50	0,6639	1,30	0,7718
	BO	3700-2430 1860-615	5 *	1,29	0,7434	1,56	0,6384	1,43	0,7227
			10 §	0,912	0,8723	1,51	0,6624	1,35	0,7549
	UVN	3700-2430 1860-615	5 *§	1,25	0,7597	1,55	0,6423	1,40	0,7348
	EMSC	3700-2430 1860-615	5 *§	1,27	0,7526	1,52	0,6581	1,40	0,7370
	SNV	3700-2430 1860-615	5 *	1,24	0,7645	1,52	0,6551	1,40	0,7337
			4 §	1,31	0,7358	1,57	0,6352	1,46	0,7109
		i-PLS 2997-2789 1754-1502 1339-701	**4 *§**	**1,26**	**0,7568**	**1,46**	**0,6842**	**1,35**	**0,7521**
		VIP 1669-965 714-689	4 *§	1,27	0,7535	1,49	0,6690	1,41	0,7331
	SG FD	3700-2430 1860-615	11 *	0,411	0,9740	1,25	0,7666	1,10	0,8367
			10 §	0,508	0,9604	1,34	0,7321	1,13	0,8265
	SG SD	3700-2430 1860-615	4 *§	1,14	0,8020	1,73	0,5561	1,46	0,7101
	SNV-DT	3700-2430 1860-615	4 *§	1,28	0,7479	1,53	0,6537	1,42	0,7286

* Valeur estimée à partir de la méthode des variances expliquées de la validation croisée.

(Suite des notes du Tableau 3.4)

§ Valeur estimée selon le critère de Haaland (premier modèle avec p < 0,75 pour le rapport $PRESS_k$ / min(PRESS)).

Les modèles optimaux sont en gras et colorés.

Abréviations : BO, Correction du décalage de la ligne de base ; DT, Correction de tendance avec un ordre polynomial (PO) = 2 ; EMSC, Correction étendue de la diffusion multiplicative ; *i*-PLS, Moindres carrés partiels par intervalle ; ps, Base de poids sec ; RN, Normalisation par gamme ; SG FD, 1ère dérivée de Savitzky-Golay à 9 points / côté et un PO = 2 ; SG SD, 2ère dérivée de Savitzky-Golay à 7 points / côté et un PO = 3 ; SNV, Standard Normal Variate ; UVN, Normalisation par vecteur unitaire ; VIP, Importance des variables en projection.

Tableau 3.5 : Paramètres de régression des procédures de calibration, de validation croisée et de prédiction pour les modèles analytiques calculés par SVMR en utilisant différents prétraitements spectraux, fonctions de noyau et paramètres de réglage (d'apprentissage).

Paramètre de qualité (unité)	Pré-traitement spectral	Fonction de noyau	Hyperparamétrage			SVs	Calibration		Validation croisée		Prédiction	
			C	ε	γ		RMSEC	R^2_C	RMSECV	R^2_{CV}	RMSEP	R^2_P
Humidité (%)	Aucun	Linéaire	0,03	0,15	–	45	0,787	0,7946	1,31	0,4431	0,964	0,5858
		PO = 2	0,01	0,15	0,01	53	1,15	0,5925	1,69	0,1163	1,39	0,1431
		PO = 3	0,01	0,15	0,01	50	0,553	0,9034	1,92	0,1920	1,18	0,3763
		RBF	1	0,15	0,01	52	0,350	0,9830	1,52	0,2435	1,07	0,4855
	EMSC	Linéaire	0,03	0,15	–	45	0,823	0,7747	1,28	0,4729	0,990	0,5631
		PO = 2	1	0,15	1	57	0,541	0,9061	1,97	0,1453	1,28	0,2682
		PO = 3	0,1	0,15	1	53	0,534	0,9049	2,63	0,0366	1,60	0,1408
		RBF	1	0,1	0,001	51	0,873	0,7563	1,45	0,2961	0,993	0,5607
	SNV	Linéaire	0,03	0,15	–	42	0,837	0,7662	1,37	0,4102	0,948	0,5991
		PO = 2	1	0,15	1	50	0,517	0,9150	1,73	0,2558	1,35	0,1836
		PO = 3	0,01	0,15	0,01	48	0,634	0,8707	2,09	0,1041	1,31	0,2302
		RBF	0,1	0,1	0,001	48	0,898	0,7414	1,46	0,2821	1,02	0,5391
	SG FD	Linéaire	0,6	0,15	–	43	0,485	0,9295	1,51	0,2753	0,981	0,5711
		PO = 2	1	0,15	10	46	0,502	0,9423	1,44	0,2974	1,07	0,4882
		PO = 3	0,1	0,15	10	43	0,502	0,9530	1,45	0,2883	1,01	0,5418
		RBF	10	0,1	0,001	48	0,344	0,9754	1,37	0,3672	0,951	0,5964
	SNV-DT	**Linéaire**	**0,03**	**0,15**	**–**	**43**	**0,819**	**0,7816**	**1,34**	**0,4367**	**0,916**	**0,6258**
		PO = 2	1	0,15	1	54	0,537	0,9068	2,10	0,1131	1,21	0,3453
		PO = 3	0,01	0,15	0,01	51	0,703	0,8412	2,26	0,0646	1,09	0,4738
		RBF	0,1	0,1	0,001	49	0,899	0,7434	1,43	0,3068	1,03	0,5289
pH	Aucun	Linéaire	0,03	0,08	–	48	0,176	0,9198	0,303	0,7510	0,188	0,9085
		PO = 2	0,1	0,08	0,01	46	0,129	0,9573	0,4927	0,4416	0,191	0,9054
		PO = 3	0,1	0,08	0,01	49	0,0982	0,9737	0,6307	0,4166	0,417	0,5479
		RBF	10	0,05	0,0001	45	0,200	0,9002	0,303	0,7511	0,224	0,8700
	EMSC	Linéaire	0,03	0,08	–	41	0,167	0,9246	0,308	0,7433	0,176	0,9199
		PO = 2	1	0,08	0,1	46	0,0955	0,9754	0,461	0,5174	0,236	0,8557
		PO = 3	0,1	0,08	0,1	46	0,0993	0,9732	0,6126	0,3625	0,269	0,8119
		RBF	10	0,04	0,0001	51	0,207	0,8905	0,320	0,7212	0,238	0,8533
	SNV	**Linéaire**	**0,03**	**0,08**	**–**	**41**	**0,171**	**0,9220**	**0,297**	**0,7606**	**0,168**	**0,9266**
		PO = 2	1	0,08	1	49	0,0962	0,9755	0,410	0,5887	0,177	0,9186
		PO = 3	1	0,08	1	51	0,977	0,9740	0,489	0,4981	0,274	0,8049
		RBF	10	0,04	0,0001	50	0,221	0,8804	0,309	0,7428	0,244	0,8455
	SG FD	Linéaire	0,6	0,08	–	43	0,0920	0,9784	0,327	0,7085	0,221	0,8730
		PO = 2	0,1	0,08	1	42	0,0906	0,9821	0,359	0,6627	0,235	0,8560
		PO = 3	1	0,08	0,1	48	0,0981	0,9831	0,392	0,5928	0,247	0,8410
		RBF	10	0,04	0,0001	51	0,0491	0,9939	0,316	0,7262	0,197	0,8996

Cendres (%)	Aucun	Linéaire	0,03	0,15	–	40	0,859	0,7797	1,41	0,4497	1,09	0,5970
		PO = 2	0,1	0,15	0,1	42	0,576	0,9046	1,44	0,4407	1,01	0,6541
		PO = 3	0,1	0,15	1	43	0,5811	0,9005	1,77	0,3461	0,794	0,7852
		RBF	0,1	0,1	0,001	42	0,617	0,9147	1,27	0,5140	0,841	0,7588
	EMSC	Linéaire	0,03	0,15	–	42	0,860	0,7827	1,32	0,5001	1,02	0,6441
		PO = 2	0,1	0,15	0,1	41	0,577	0,9008	1,50	0,4184	1,11	0,5769
		PO = 3	0,1	0,15	0,1	44	0,572	0,9019	1,56	0,4540	1,30	0,4275
		RBF	0,6	0,1	0,001	44	0,790	0,8345	1,31	0,4766	0,758	0,8041
	SNV	Linéaire	0,03	0,15	–	44	0,903	0,7564	1,46	0,4138	1,07	0,6120
		PO = 2	0,1	0,15	0,1	44	0,589	0,9022	1,56	0,3850	0,990	0,6660
		PO = 3	0,1	0,15	1	43	0,596	0,8921	1,61	0,4424	1,32	0,4044
		RBF	**0,1**	**0,1**	**0,001**	**43**	**0,710**	**0,8587**	**1,23**	**0,5451**	**0,728**	**0,8197**
	SG FD	Linéaire	0,2	0,15	–	41	0,585	0,9011	1,22	0,5528	0,999	0,6599
		PO = 2	0,1	0,15	0,1	41	0,569	0,9339	1,15	0,6341	0,833	0,7637
		PO = 3	0,1	0,15	0,1	41	0,565	0,9332	1,17	0,6275	0,811	0,7758
		RBF	1	0,1	0,001	44	0,397	0,9652	1,11	0,6505	0,765	0,8008
Nicotine totale (mg/g, ps)	Aucun	Linéaire	0,2	0,2	–	26	0,915	0,8804	1,50	0,6637	1,16	0,8142
		PO = 2	1	0,2	0,1	37	1,03	0,8581	2,36	0,2386	1,30	0,7656
		PO = 3	1	0,2	0,01	35	0,982	0,8749	2,51	0,3743	1,31	0,7607
		RBF	10	0,1	0,0001	38	1,00	0,8576	1,79	0,5233	1,25	0,7818
	EMSC	Linéaire	0,6	0,2	–	26	0,941	0,8734	1,47	0,6807	1,18	0,8068
		PO = 2	0,1	0,2	0,01	34	1,02	0,8623	1,94	0,4429	1,16	0,8145
		PO = 3	1	0,2	0,01	30	0,929	0,8761	1,56	0,6430	1,20	0,8002
		RBF	**10**	**0,1**	**0,001**	**39**	**0,526**	**0,9685**	**1,67**	**0,5898**	**0,835**	**0,9031**
	SNV	Linéaire	0,6	0,2	–	26	0,955	0,8677	1,47	0,6779	1,17	0,8111
		PO = 2	0,1	0,2	0,1	36	1,03	0,8511	2,00	0,4122	1,23	0,7903
		PO = 3	0,1	0,2	10	28	0,917	0,9058	1,48	0,6878	1,21	0,7957
		RBF	10	0,1	0,001	45	0,536	0,9663	1,60	0,6239	0,864	0,8963
	SG FD	Linéaire	0,08	0,2	–	28	0,969	0,8685	1,51	0,6610	1,23	0,7894
		PO = 2	1	0,2	1	36	0,991	0,8969	1,89	0,5034	1,28	0,7715
		PO = 3	1	0,2	1	37	0,972	0,9097	1,81	0,5599	1,26	0,7794
		RBF	10	0,1	0,0001	44	0,531	0,9622	1,46	0,6833	0,964	0,8709
Nicotine non ionisée (mg/g, ps)	Aucun	Linéaire	0,2	0,2	–	29	0,904	0,8791	1,49	0,6586	1,15	0,8146
		PO = 2	1	0,2	0,1	37	1,03	0,8554	2,33	0,2323	1,28	0,7678
		PO = 3	0,01	0,2	0,01	36	0,994	0,8629	2,23	0,4029	1,23	0,7880
		RBF	10	0,1	0,0001	40	0,988	0,8576	1,74	0,5359	1,23	0,7863
	EMSC	Linéaire	0,2	0,2	–	29	0,927	0,8744	1,47	0,6701	1,17	0,8075
		PO = 2	0,1	0,2	0,01	34	1,01	0,8620	1,92	0,4386	1,15	0,8131
		PO = 3	1	0,2	0,01	30	0,921	0,8747	1,56	0,6331	1,19	0,8015
		RBF	**10**	**0,1**	**0,001**	**42**	**0,521**	**0,9682**	**1,67**	**0,5785**	**0,831**	**0,9024**
	SNV	Linéaire	0,2	0,2	–	26	0,940	0,8675	1,49	0,6636	1,14	0,8156
		PO = 2	0,1	0,2	0,01	37	1,03	0,8490	1,99	0,4050	1,20	0,7957
		PO = 3	1	0,2	0,1	28	0,908	0,9009	1,49	0,6698	1,20	0,7957
		RBF	1	0,1	0,001	45	0,794	0,9225	1,69	0,5715	1,11	0,8250
	SG FD	Linéaire	0,08	0,2	–	27	0,963	0,8680	1,51	0,6536	1,21	0,7916
		PO = 2	1	0,2	1	37	0,983	0,8937	1,88	0,4870	1,28	0,7682
		PO = 3	1	0,2	0,01	35	0,962	0,9072	1,80	0,5480	1,26	0,7762
		RBF	10	0,1	0,0001	43	0,533	0,9620	1,45	0,6774	0,949	0,8729

Les modèles optimaux sont en gras et colorés.

Abréviations : C, Fonction de coût ; ε, Marge d'erreur ; γ, Diamètre radial ; DT, Correction de tendance avec un ordre polynomial = 2 ; EMSC, Correction étendue de la diffusion multiplicative ; PO, Ordre polynomial ; ps, Base de poids sec ; RBF, Fonction de base radiale ; SG FD, 1ère dérivée de Savitzky-Golay à 9 points / côté et un PO = 2 ; SNV, Standard Normal Variate ; SVs, Nombre de vecteurs de support.

Quel que soit la méthode de régression employée, la sélection des modèles optimaux reposait sur des critères tels que le RMSEP, le R^2_P et le rapport de RMSEP / RMSEC, associés aux modèles avec le plus petit nombre de LVs ou SVs. Notamment, l'application individuelle de SNV, EMSC et SG FD a démontré une amélioration des performances des deux méthodes de régression. En revanche, si les méthodes *i*-PLS et VIP (voir des exemples dans la Figure 3.12) ont apporté des améliorations marginales aux résultats de la PLSR, ils ont eu un effet négatif sur la capacité prédictive de la SVMR (données non présentées). Cette constatation était attendue, étant donné que les deux méthodes de sélection de variables spectrales sont basées sur les principes de PLS et ne sont donc compatibles qu'avec la PLS.

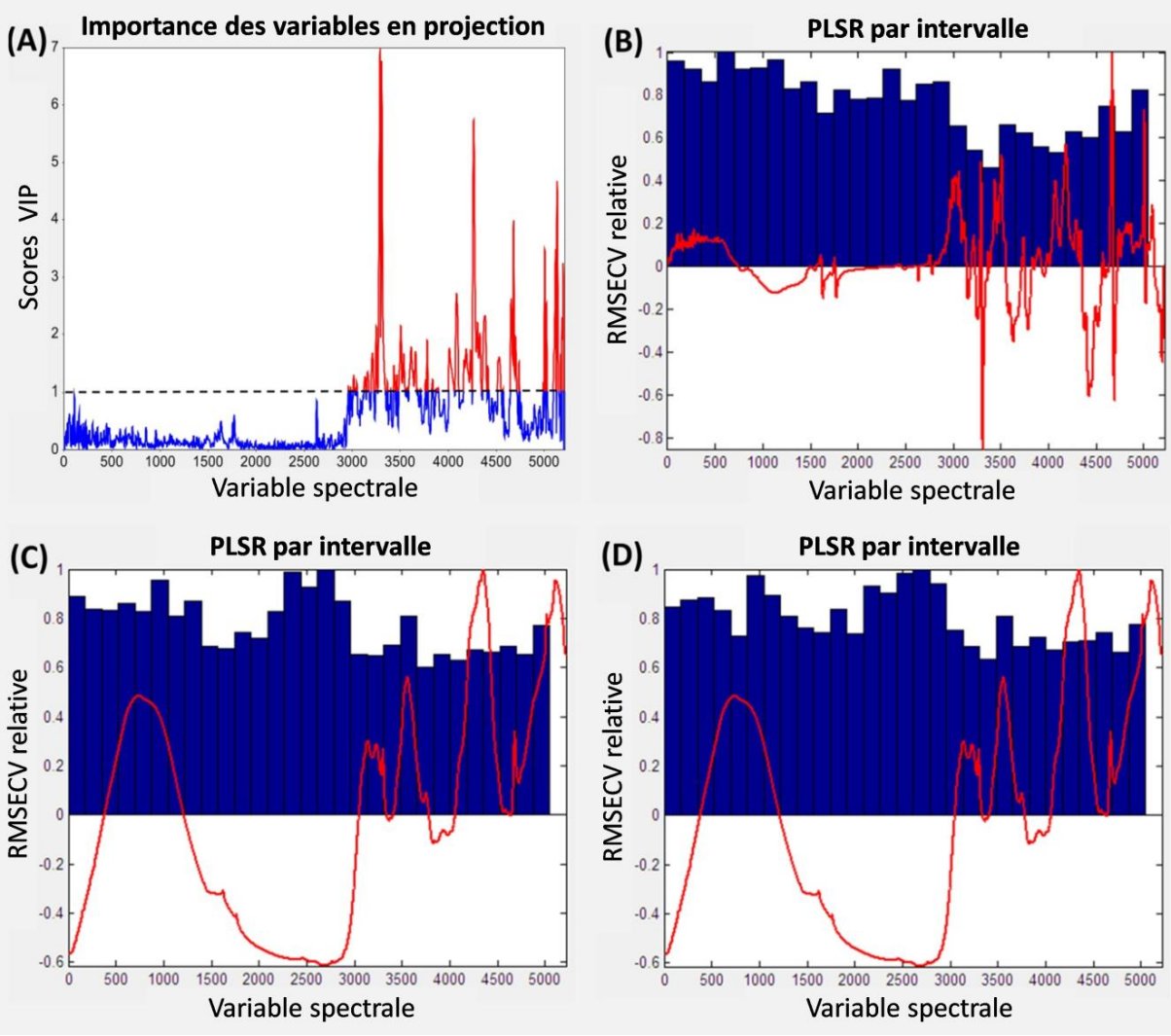


Figure 3.12 : Exemples de sélection de bandes caractéristiques par VIP et *i*-PLS pour les modèles PLSR optimaux de (A) humidité totale, (B) niveau de pH, (C) nicotine totale et (D) nicotine non ionisée.

En plus des méthodes *i*-PLS et VIP, les coefficients de régression ont été utilisés pour évaluer la contribution de nombres d'onde spécifiques dans le modèle de régression pour le nombre désigné de LVs. En ce qui concerne les LVs, le critère de Haaland a donné des résultats très satisfaisants, avec un moins de facteurs par rapport à la méthode des variances de CV couramment utilisée. Toutes ces considérations mathématiques soulignent la robustesse de la PLSR en tant qu'outil puissant et fiable en régression linéaire, préservant ses modèles des problèmes de sur-ajustement ou de sous-ajustement.

Contrairement à la PLSR, il n'existe pas de critères valables pour évaluer l'ajustement des données dans la SVMR. Cependant, maintenir le rapport RMSECV / RMSEC, le R^2_{CV}, et les valeurs de C et de γ aussi proches que possible de un (1) peut garantir une bonne capacité de généralisation. Un autre paramètre important qui ne doit pas être ignoré est le nombre de SVs. Dans cette étude, les modèles les plus performants ont montré que 61 à 67 % des échantillons d'entraînement étaient utilisés comme SVs. Les valeurs semblent relativement élevées en comparaison avec les résultats de Schmidtke *et al.* [163] qui ont signalé un pourcentage maximal de 54 %; néanmoins, cela peut être accepté pour un ensemble de données aussi petit et complexe. Des statistiques critiques supplémentaires (Tableau 3.6) et des graphiques de relation entre les valeurs prédites et références (Figures de 3.13 à 3.17) ont été mis en œuvre dans le but d'évaluer l'exactitude de prédiction dans l'ensemble de test indépendant.

Tableau 3.6 : Métriques d'évaluation des modèles optimaux de PLSR et SVMR.

Méthode	Paramètre (unité)	MAPE (%)	REP (%)	RPD *	RER *	Biais #
PLSR	Humidité (%)	1,8	2,3	1,3	5,4	0,22
	pH	1,86	2,43	2,5	9,9	$1{,}4\ 10^{-3}$
	Cendres (%)	3,4	4,1	1,8	8,2	0,36
	NCT totale (mg/g, ps)	13,4	17,1	2,0	8,9	−0,36
	NCT libre (mg/g, ps)	13,6	12,1	2,0	9,0	−0,36
SVMR	Humidité (%)	1,4	1,9	1,7	6,7	0,16
	pH	1,29	1,63	3,7	14,8	−0,018
	Cendres (%)	2,5	2,9	2,4	11,0	0,17
	NCT totale (mg/g, ps)	9,5	10,4	3,2	14,2	−0,067
	NCT libre (mg/g, ps)	9,5	10,5	3,2	14,2	−0,084

* Métrique sans unité ; # Même unité du paramètre.

Abréviations : NCT, Nicotine ; ps, Base de poids sec.

3.3.4.1. *Humidité totale*

Concernant la teneur en humidité totale, tant la PLSR que la SVMR n'ont pas permis d'obtenir des estimations satisfaisantes de l'humidité du ST, se traduisant par des RMSEP > 0,91, R^2_P < 0,63 et RPD < 1,7. Selon l'échelle RPD référencée dans [88], les modèles développés ne conviennent qu'à des fins de triage préliminaire.

À partir des graphiques "Prédite contre Réelle" (Figures 3.13A et B), la plupart des points étaient dispersés loin des lignes de régression, et la forte divergence entre les ajustements de calibration et de test indique une faible capacité prédictive. L'examen du graphique des coefficients de régression (Figure 3.13C), à la recherche des variations spectrales responsables de ce manque d'ajustement dans la PLSR, a révélé des caractéristiques de bruit irrégulier et l'utilisation de nombres d'onde non pertinents dans la construction du modèle au lieu des bandes d'absorption de l'eau.

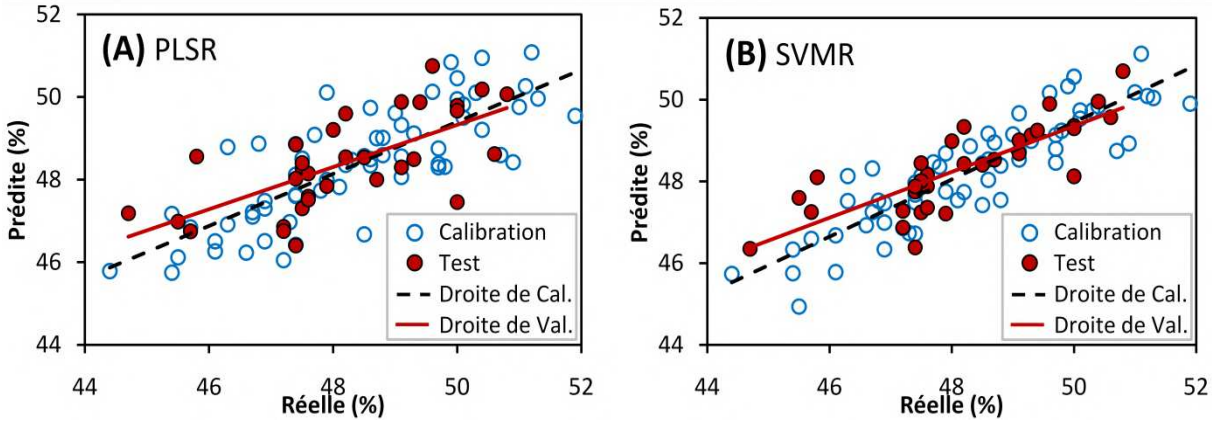


Figure 3.13 : (A et B) Lignes de régression des valeurs réelles par rapport aux valeurs prédites calculées respectivement par PLSR et SVMR pour l'humidité.

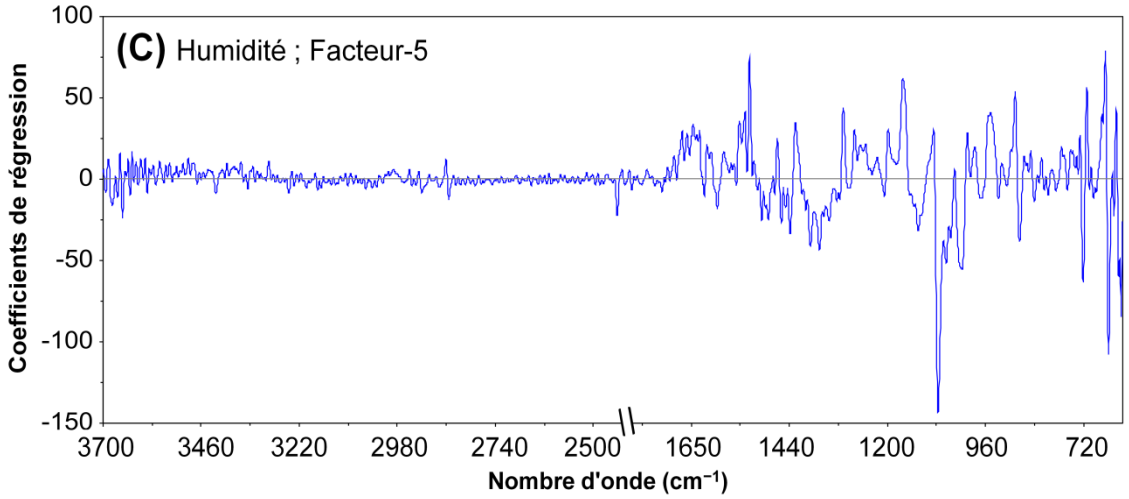


Figure 3.13 (suite) : (C) Coefficients de régression pour le nombre pertinent de LVs dans le modèle PLSR correspondant.

Une explication de ce résultat peut être attribuée au fait que les bandes d'eau dans les spectres ont été arbitrairement superposées avec différents agents interférents, ou probablement au fait que la procédure de référence mesure non seulement l'eau mais une somme de composés volatils qui peuvent être relativement influents. Si ce dernier cas est avéré, l'étude de méthodes alternatives pour mesurer l'eau / l'humidité dans les ST, comme celles décrites par McAdam *et al.* [128], pourrait fournir des valeurs mieux corrélées avec les spectres de nos échantillons. Étonnamment, de faibles valeurs de MAPE (1,4 %) et de REP (1,9 %) ont été obtenues.

3.3.4.2. Niveau de pH

Concernant le pH, le meilleur résultat a été obtenu pour la SVMR linéaire avec RMSEP = 0,17, $R^2_P \approx 0,93$ et RPD = 3,7. La PLSR, quant à elle, a également présenté de très bons résultats dans la région *i*-PLS (1700 - 680 cm^{-1}) prétraitée par SG FD, donnant RMSEP = 0,25, $R^2_P \approx 0,84$ et RPD = 2,5 en utilisant trois facteurs. La superposition parfaite des deux lignes d'ajustement (Figures 3.14A et B) avec des valeurs de MAPE et REP du même ordre de grandeur faible implique une capacité de prédiction élevée.

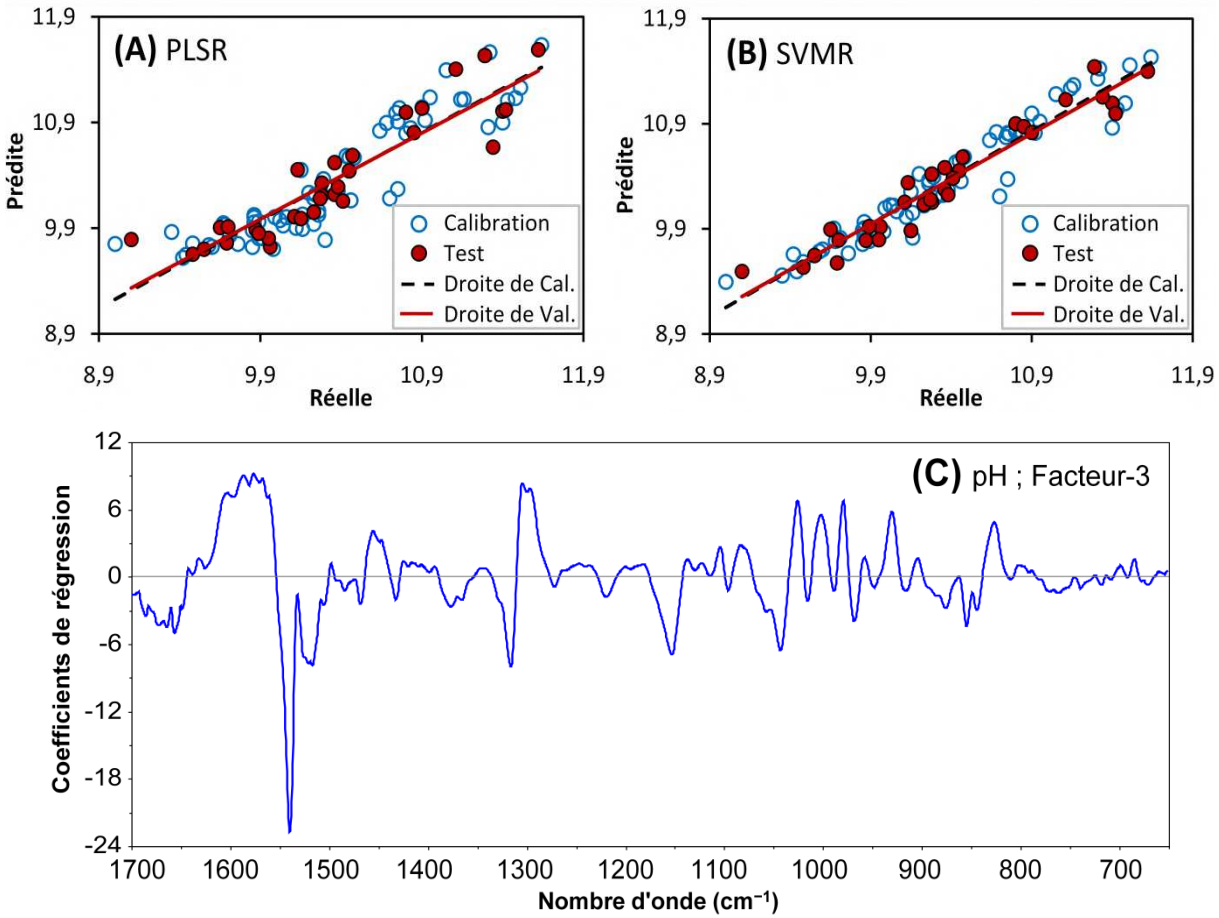


Figure 3.14 : (A et B) Lignes de régression des valeurs réelles par rapport aux valeurs prédites calculées respectivement par PLSR et SVMR pour le pH. (C) Coefficients de régression pour le nombre pertinent de LVs dans le modèle PLSR correspondant.

Selon l'échelle RPD mentionnée précédemment, le modèle SVMR peut être appliqué avec succès au contrôle de processus pour prédire le niveau de pH. Les coefficients de régression PLS (Figure 3.14C) font apparaître des corrélations positives avec les bandes de l'ion CO_3^{2-} (à 1315, 854 et 717 cm^{-1}) et des corrélations négatives avec l'amide II dans la plante (1540 cm^{-1}) ainsi qu'avec les vibrations SiO_2 du kieselguhr (1168 et 1063 cm^{-1}) dans les spectres dérivés, ce qui établit un lien logique entre le pH mesuré et la proportion de $CaCO_3$ (modificateur alcalin) dans les ingrédients du produit final.

3.3.4.3. *Cendres totales*

Pour la teneur en cendres totales, la PLSR a fourni une valeur peu élevée de RMSEP (1,0), $R^2_P \approx 0,66$ et RPD = 1,8 en utilisant seulement deux LVs. Par contre, la SVMR basée sur la fonction de base radiale (RBF) a clairement améliorée la performance avec une RMSEP (0,73) diminuée de 28,6 % et un R^2_P (0,82) augmenté de 24,5 %, tandis que le RPD est devenu 2,4, ce qui signifie une précision acceptable et une adéquation du modèle à des fins de triage.

La qualité des lignes de régression "Prédite contre Réelle" a démontré une qualité d'ajustement satisfaisante pour la SVMR par rapport à la PLSR, qui présentait une divergence résiduelle de la ligne de l'ensemble de test loin de la ligne de l'ensemble de calibration, comme le montrent les Figures 3.15A et B. Notez également que des valeurs plus faibles de MAPE et de REP ont été obtenues pour le modèle SVMR.

La Figure 3.15C montre que le large pic centré à 1120 cm^{-1}, qui chevauche et masque l'étirement Si−O de la silice à environ 1055 cm^{-1}, contribue de manière significative à la prédiction de la teneur en cendres; ceci reconfirme la relation étroite entre le paramètre actuel et la quantité de "kieselguhr" dans les échantillons. Il est à noter qu'on peut observer des légères variations dans les positions des pics attribuées aux effets de matrice dans ce graphique et dans d'autres.

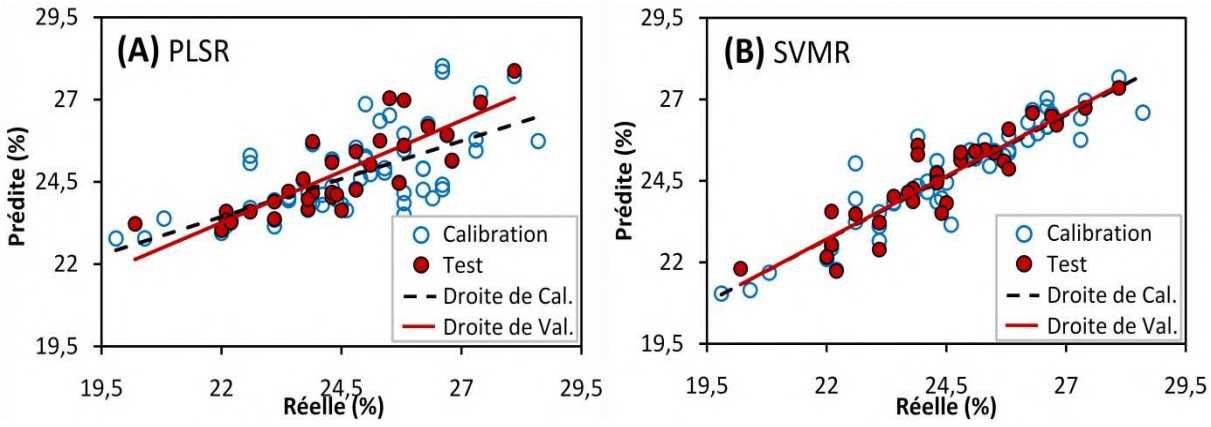


Figure 3.15 : (A et B) Lignes de régression des valeurs réelles par rapport aux valeurs prédites calculées respectivement par PLSR et SVMR pour les cendres.

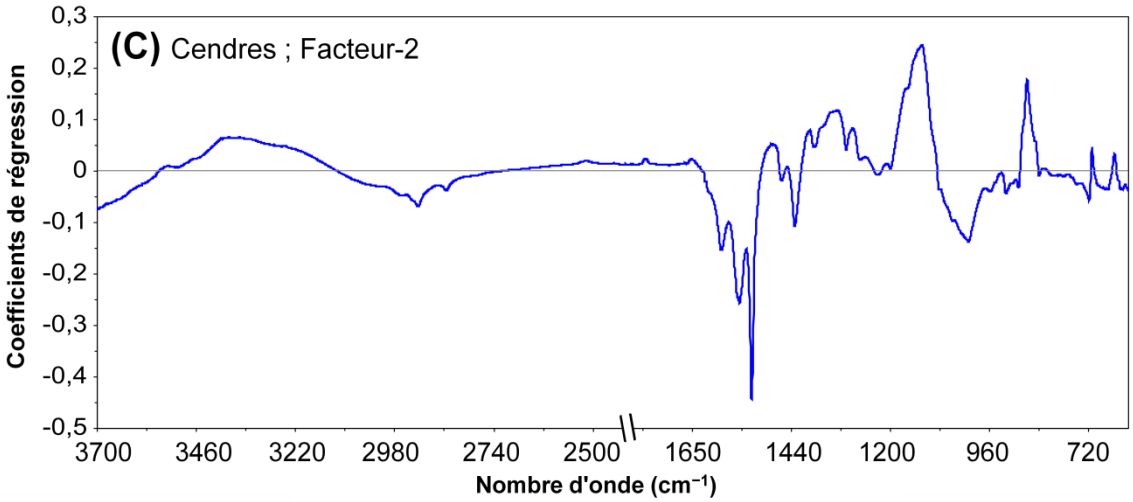


Figure 3.15 (suite) : (C) Coefficients de régression pour le nombre pertinent de LVs dans le modèle PLSR correspondant.

3.3.4.4. Nicotine totale et nicotine non ionisée

Puisque la NCT totale et la NCT non ionisée ont présenté des résultats comparables pour nos échantillons, elles ont été examinées collectivement comme un seul paramètre. Bien que présentant un RMSEP relativement plus élevé (0,84), les modèles SVMR basés sur RBF ont démontré des capacités prédictives acceptables avec un R^2_P d'environ 0,90 et un RPD de 3,2. En comparaison, PLSR a donné une RMSEP maximale de 1,4, un R^2_P de 0,75 et un RPD de 2,0.

Contrairement aux cendres, les concentrations de NCT dans les échantillons examinés ont révélé une plus grande variabilité (SD = 2,6). Cette variabilité soutient l'adéquation des modèles SVMR pour des applications de contrôle qualité, compte tenu de l'échelle RPD [88]. Il est intéressant de noter que des valeurs plus faibles de RMSECV ont été obtenues pour la PLSR. Cela suggère un léger problème de sur-ajustement dans la SVMR, néanmoins, d'excellentes corrélations entre les prédictions de SVMR et les valeurs de la méthode référence peuvent être observées dans les Figures 3.16B et 3.17B.

L'examen des graphiques des coefficients de régression (Figures 3.16C et 3.17C) montre que la bande autour de 1540 cm^{-1}, attribuable à l'amide II, est apparue comme une variable importante pour identifier la variété spécifique de tabac utilisée dans les échantillons. Des contributions plus faibles ont été observées à partir des bandes situées à 1150 et 1080 cm^{-1}, qui correspondent à la cellulose dans les plantes. Les pics à 1640 et 1320 cm^{-1}, déjà liés à la présence d'eau et aux vibrations d'amide I / lignine, respectivement, étaient chargés négativement à travers les facteurs utilisés.

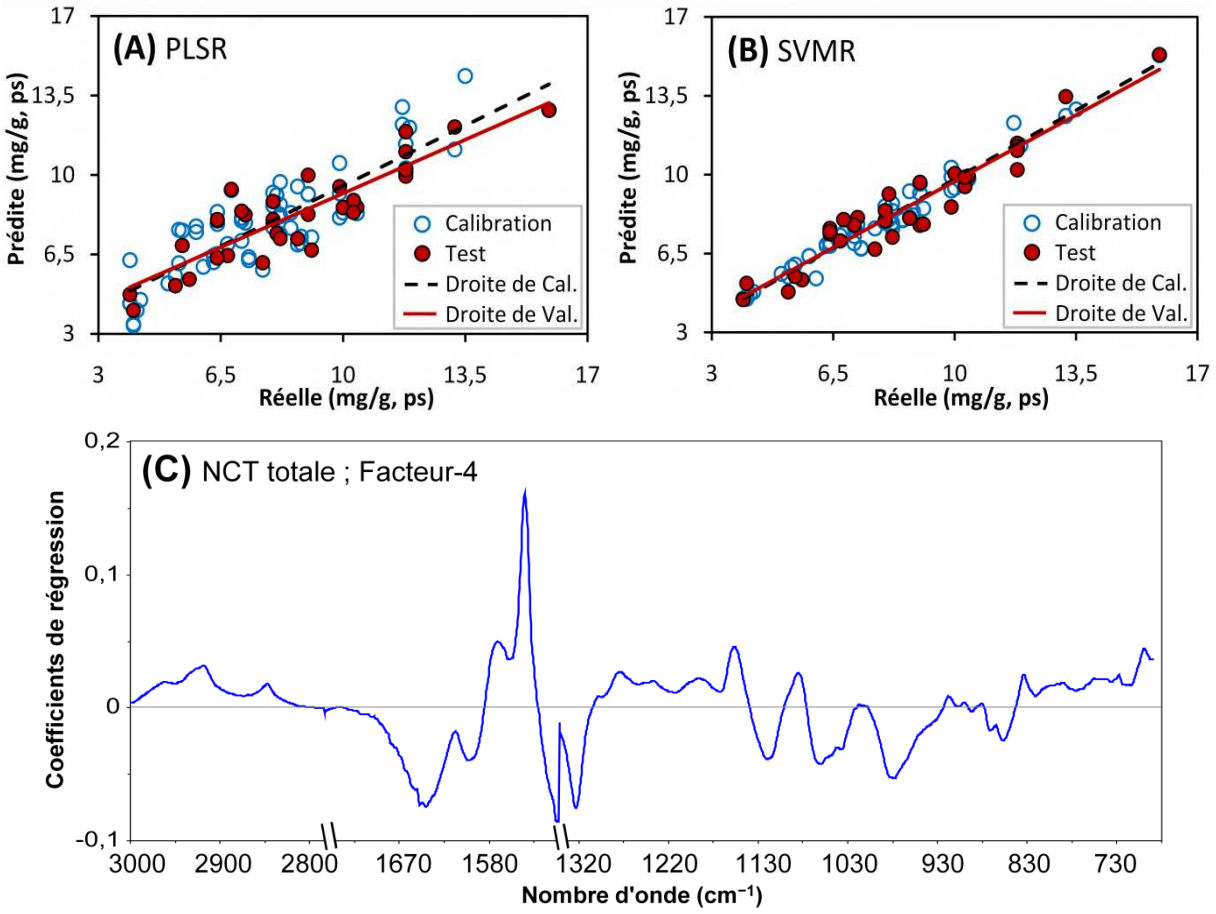


Figure 3.16 : (A et B) Lignes de régression des valeurs réelles par rapport aux valeurs prédites calculées respectivement par PLSR et SVMR pour la nicotine totale. (C) Coefficients de régression PLS pour le nombre pertinent de LVs.

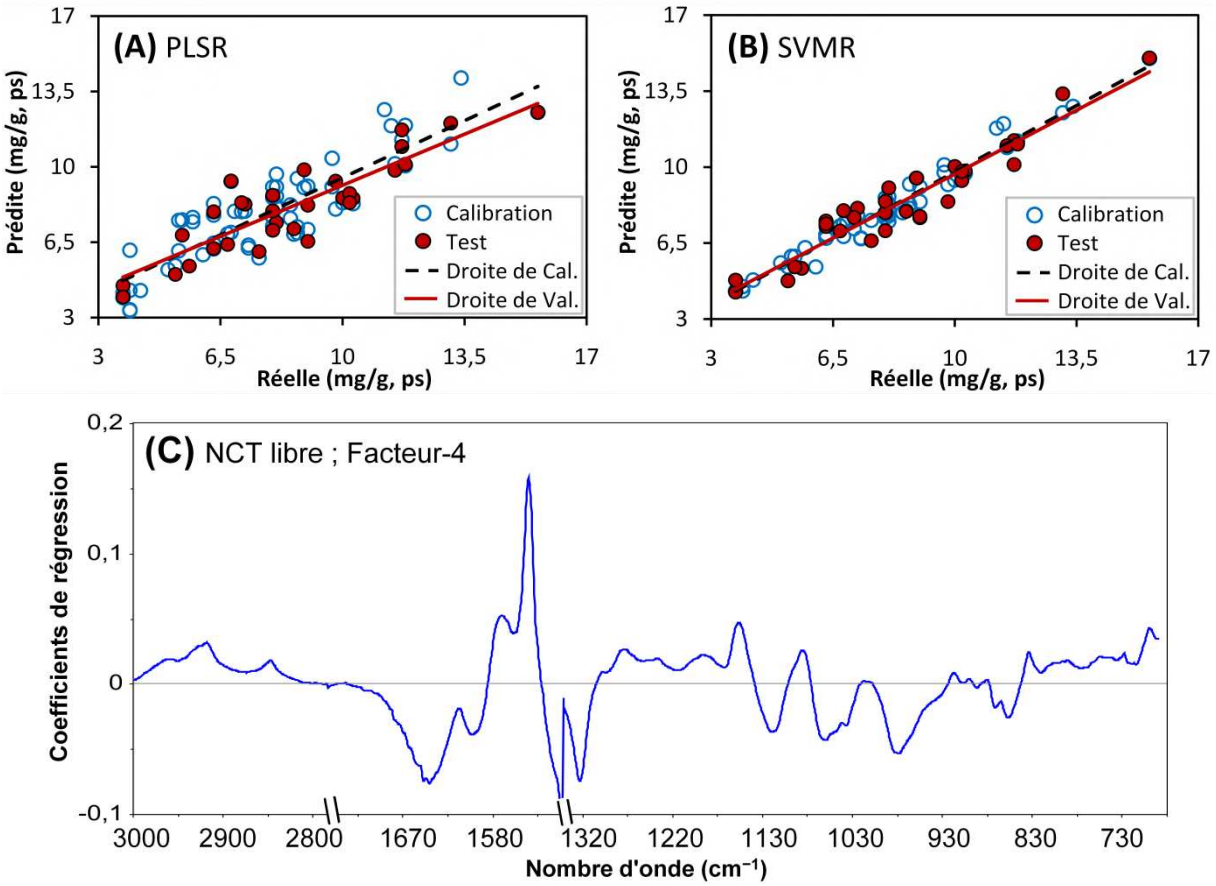


Figure 3.17 : (A et B) Lignes de régression des valeurs réelles par rapport aux valeurs prédites calculées respectivement par PLSR et SVMR pour la nicotine libre. (C) Coefficients de régression PLS pour le nombre pertinent de LVs.

Les MAPE ont diminué de 13,6 % à 9,5 %, et les REP ont diminué de 17,1 % à 10,4 % lors de l'utilisation de la SVMR. Toutefois, il est important de noter que ces valeurs, bien qu'améliorées, dépassent toujours les valeurs souhaitées.

3.3.4.5. Analyse EJCR et LOD / LOQ

Pour une évaluation plus approfondie de la justesse et de la précision des modèles optimaux, le test de la région de confiance elliptique conjointe (EJCR) a été recommandé pour vérifier la présence de biais plutôt que les autres tests individuels tels que le taux de récupération de l'analyte. Comme le montrent les Figures 3.18A et B, l'humidité présente les régions de confiance les plus larges et les plus éloignées du point idéal, ce qui implique leur inaptitude pour l'estimation de la teneur en humidité dans les produits.

Concernant les quatre paramètres restants, la SVMR a montré des régions de confiance à la fois plus étroites et plus proches du point (1, 0). Néanmoins, aucune des ellipses n'a englobé le point idéal. Dans une telle situation et pour considérer que les valeurs prédites de ces modèles sont significativement biaisées, des EJCRs ultérieurs basés sur les moindres carrés pondérés (WLS) ou les moindres carrés bilinéaires (BLS) devraient être effectués. Malheureusement, en raison de la quantité limitée de certains échantillons, certains paramètres n'ont été analysés qu'une seule fois, ce qui a entraîné des valeurs d'écart-type manquantes empêchant d'appliquer ces dernières méthodes.

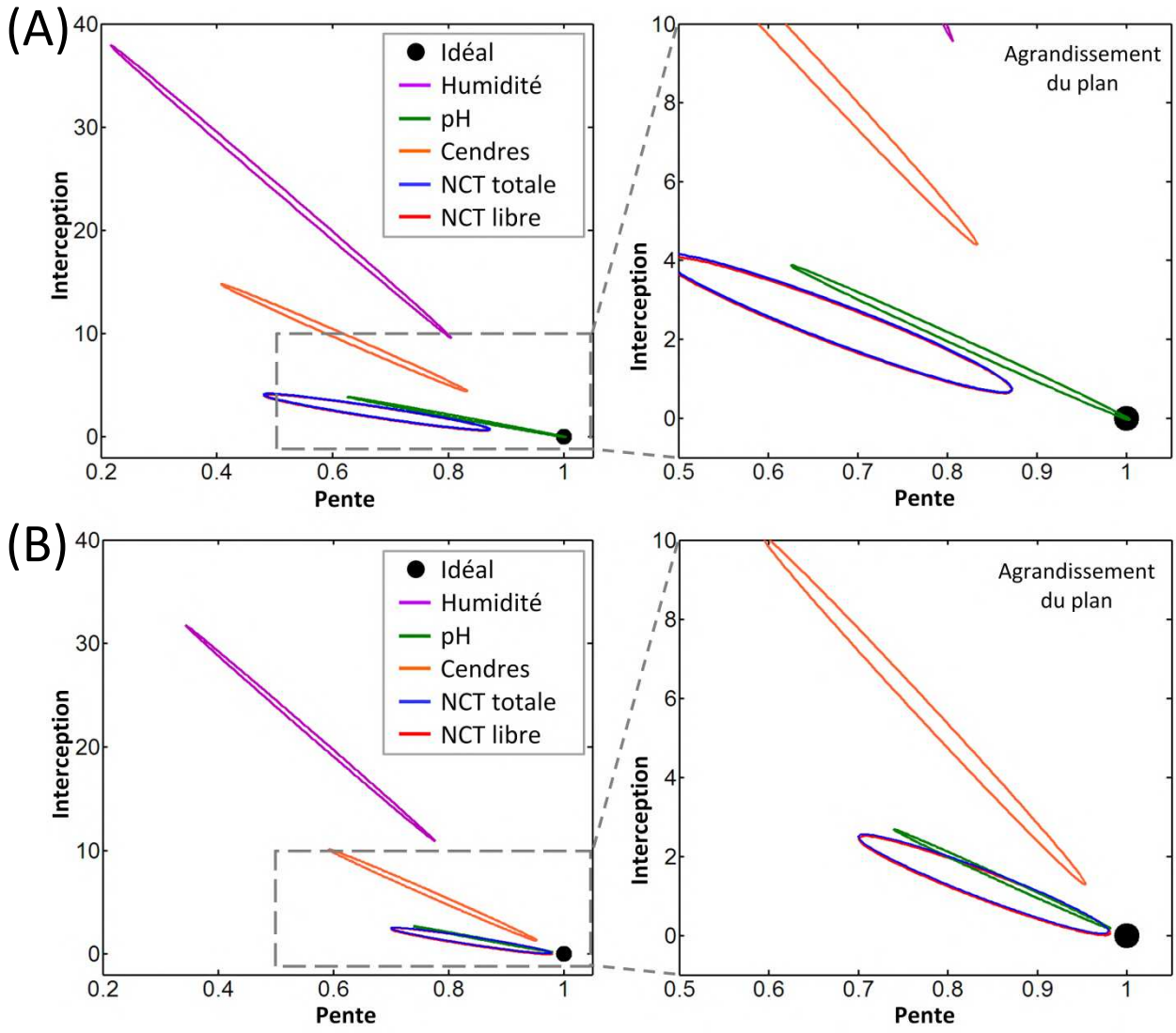


Figure 3.18 : EJCRs dans le plan pente-interception réalisés pour les modèles optimaux de (A) PLSR et (B) SVMR basés sur la méthode des moindres carrés ordinaires.

Le Tableau 3.7 montre les LOD et LOQ estimées en utilisant les deux approches analytiques introduites dans la **sous-section 3.2.8**. Lorsqu'elles sont calculées en utilisant l'approche de référence proposée par [173], les valeurs de LOD démontrent une comparabilité acceptable avec les valeurs de LOD_{ICH}. Par contre, les valeurs de LOD_{pu} ont été notablement surestimées, ce qui était attendu en raison des incertitudes finies dans les valeurs prédites. Toutefois, cela ne signifie pas l'invalidité de cette dernière approche pour d'autres cas. Il est intéressant aussi d'avertir des faibles écarts observés entre certaines valeurs obtenues entre les approches d'ICH et de référence, soulignant la nécessité de poursuivre les recherches afin d'envisager des facteurs correcteur pour améliorer la précision des formules existantes.

Tableau 3.7 : Limites de détection et de quantification calculées en utilisant les approches proposées dans la littérature pour les modèles PLSR et SVMR optimaux.

Méthode	Paramètre de qualité (unité)	LOD_{ICH} §	LOQ_{ICH} §	LOD_{pu} ¶	LOQ_{pu} ¶	LOD_{min} – LOD_{max}	LOQ_{min} – LOQ_{max}
PLSR	Humidité (%)	15,5	47,1	16,3	49,0	12,3 – 12,4	36,9 – 37,2
	pH	1,96	5,93	2,19	6,5	1,77 – 1,80	5,3 – 5,4
	Cendres (%)	11,2	34,0	13,1	39,4	7,6 – 7,7	22,7 – 23,1
	NCT totale (mg/g, ps)	2,0	6,0	5,3	16,0	1,7 – 2,5	5,2 – 7,4
	NCT libre (mg/g, ps)	2,0	6,0	5,3	15,9	1,7 – 2,4	5,2 – 7,3
SVMR	Humidité (%)	10,7	32,5	11,3	33,8	–	–
	pH	1,26	3,83	1,41	4,22	–	–
	Cendres (%)	4,2	12,7	4,9	14,8	–	–
	NCT totale (mg/g, ps)	0,6	1,9	1,7	5,1	–	–
	NCT libre (mg/g, ps)	0,6	1,9	1,7	5,1	–	–

§ Valeur déterminée selon les lignes directrices de l'ICH.

¶ Valeur déterminée à partir de la ligne de régression pseudo-univariée.

Abréviations : NCT, Nicotine ; ps, Base de poids sec.

Globalement, les résultats obtenus à partir de l'analyse EJCR et des valeurs LOD / LOQ étaient en accord avec les autres métriques d'évaluation, ce qui démontre la stabilité et la robustesse de la méthodologie proposée.

3.4. Conclusion

Cette étude présente pour la première fois une nouvelle méthodologie d'analyse rapide des produits du tabac oral à l'aide de la spectroscopie ATR-FT-MIR, soutenue par des méthodes mathématiques-théoriques et d'apprentissage automatique conventionnel.

En se basant sur des paramètres de qualité de référence, il a été possible de regrouper les produits de Chemma algérienne en quatre classes en utilisant de manière complémentaire trois techniques non supervisées (PCA, AHC et *k*-means clustering). Ensuite, deux techniques supervisées (PLS-DA et SVM-C) ont été menées, en adoptant les résultats de *k*-means, pour classer les échantillons directement sur la base de leurs spectres MIR. Des exactitudes de calibration de 85,4 % et 95,8 % ont été obtenues respectivement pour PLS-DA et SVM-C. Ces informations qualitatives peuvent être d'une grande importance aux autorités réglementaires afin d'identifier et de suivre les produits suspects pouvant présenter des risques pour la santé des consommateurs.

La spectroscopie ATR-FTIR combinée à la SVMR permet, par une mesure simple, de déterminer simultanément le pH, les cendres, la nicotine totale et la nicotine non ionisée. De bons résultats ont été obtenus avec $R^2_P \geq 0,82$ et RPD ≥ 2,4, permettant l'incorporation réussie des modèles optimaux au contrôle de la qualité dans l'industrie du tabac. De plus, la cohérence entre tous les indicateurs de performance souligne la grande stabilité, fiabilité et robustesse des modèles.

La méthodologie proposée a permis de réduire considérablement la quantité et le temps nécessaires à l'analyse des échantillons par rapport aux procédures de référence. Toutefois, l'un des inconvénients de la méthode est qu'elle génère des modèles prédictifs d'une précision partielle, ce qui souligne la nécessité de poursuivre le développement pour obtenir des performances optimales.

CONCLUSION GÉNÉRALE

Cette thèse présente, pour la première fois, de nouvelles méthodes pour l'analyse rapide de la nicotine et d'autres paramètres de qualité dans les produits commerciaux de tabac sans fumée en utilisant la spectroscopie ATR-FT-MIR couplée à des approches mathématiques et d'apprentissage automatique.

Dans la première partie de l'étude expérimentale, une méthode calibrée et validée sur une large gamme de concentrations a été développée pour la quantification fiable de la nicotine. Les paramètres analytiques et statistiques mettent en évidence des performances optimales avec l'application des méthodes de prétraitement LBC et EMSC, suivies de la PLSR. Cette approche exploite la puissance de la chimiométrie multivariée pour une analyse des données et une prédiction précise de l'analyte cible dans les spectres d'échantillons réels, offrant une quantification améliorée par rapport à la loi de Beer-Lambert traditionnelle. La capacité prédictive de la méthode a été examinée par le biais d'un test de la région de confiance elliptique conjointe et d'une comparaison avec la littérature existante. Les principaux avantages de cette méthode sont sa simplicité, sa rentabilité et sa spécificité par rapport à l'analyte. Un inconvénient potentiel est lié aux incertitudes sur le volume des gouttelettes déposées dues à l'utilisation de microseringue, qui peuvent être atténuées en effectuant des répétitions multiples.

La deuxième partie a consisté en une évaluation rapide et durable des paramètres de qualité de ST dans le cadre d'une étude chimiométrique comparative. En exploitant les caractéristiques spectrales, nous avons réussi à caractériser la présence de différentes qualités de feuilles de tabac dans les échantillons et à identifier la présence de carbonate de calcium et de kieselguhr en tant qu'ingrédients majeurs de la Chemma.

Dans le cadre des analyses de classification, l'utilisation du *k*-means clustering et de la SVM-C a facilité la classification des produits commerciaux en quatre catégories reflétant les attributs qualitatifs réels des échantillons (comprenant le prix, la réputation et la perception de la qualité), en se basant sur des paramètres physicochimiques de référence. L'approche proposée, qui repose

sur des mesures spectrales simples, offre une solution prometteuse pour contrôler et identifier rapidement les produits non conformes aux normes qui pourraient présenter des risques pour la santé. Un inconvénient distinct, cependant, est l'incapacité de cette méthode à différencier les produits authentiques de leurs contrefaits en s'appuyant uniquement sur des paramètres de qualité conventionnels. Cela est probablement dû aux efforts méticuleux d'imitation. Pour résoudre ce problème, des recherches futures se concentreront sur l'exploration de méthodes telles que la DD-SIMCA (« Data-Driven Soft Independent Modeling of Class Analogy ») pour développer des modèles robustes construits sur des ensembles de données provenant de la spectrométrie FTIR ou d'une combinaison de FTIR et d'autres techniques d'analyse.

Dans la prédiction à l'aide de méthodes de régression, la SVMR s'est révélée plus performante que la PLSR pour prédire avec précision quatre paramètres de qualité, à savoir le pH, les cendres, la nicotine totale et la nicotine libre, dans les tabacs à chiquer provenant des marchés algériens. Néanmoins, il convient de noter les défis posés par la complexité des ingrédients dans les formulations d'échantillons, l'interférence significative de l'humidité et la taille relativement petite de l'ensemble de données, qui peuvent avoir affecté les performances des modèles. Par conséquent, des recherches futures seront menées avec des ensembles de données plus larges et diversifiés, en employant une technique de transfert de calibration efficace pour pallier l'interférence de l'humidité, et en explorant des méthodes d'apprentissage profond pour améliorer les performances des modèles et intégrer des paramètres de qualité supplémentaires. De plus, le développement de logiciels plus avancés, accompagnés de guides d'utilisation, sera envisagé afin de faciliter leur adoption et leur mise en œuvre pratique.

Enfin, la qualité des produits du tabac est une question complexe. Outre la variabilité inhérente aux plantes et aux processus de traitement, les paramètres de qualité d'un produit dépendent principalement de sa composition, qui est souvent non révélée et établie par des non-professionnels au sein d'un marché non réglementé où sévit la contrefaçon. Partant de ce fait, les autorités publiques doivent prendre des mesures rigoureuses pour lutter contre les produits illicites et obliger les fabricants à divulguer les taux de nicotine et tout autre composant ou ingrédient susceptible d'avoir une incidence sur la qualité du produit et la santé.

APPENDICE A

LISTE DES ABRÉVIATIONS

AHC : Classification hiérarchique ascendante
ATR : réflexion totale atténuée
BO : correction du décalage de la ligne de base
CDC : centres américains de contrôle et de prévention des maladies
$CHCl_3$: chloroforme
CV : validation croisée
DT : correction de tendance
EJCR : région de confiance elliptique conjointe
EMSC : correction étendue de la diffusion multiplicative
FOMs : facteurs de mérite
FT-MIR : spectroscopie infrarouge moyenne à transformée Fourier
g : gramme
GC-MS : chromatographie en phase gazeuse couplée à la spectrométrie de masse
HAPs : hydrocarbures aromatiques polycycliques
IARC : centre international de recherche sur le cancer
i-PLS : moindres carrés partiels par intervalle
k-means : partitionnement en *k*-moyennes
L : litre
LBC : correction linéaire de la ligne de base
LC-MS/MS : chromatographie liquide couplée à la spectrométrie de masse en tandem
LOD : limite de détection
LOQ : limite de quantification
LVs : variables latentes
MAPE : erreur absolue moyenne en pourcentage
MSC : correction de la diffusion multiplicative
NAB : N'-nitrosoanabasine
NAT : N'-nitrosoanatabine
NCT : nicotine
NIR : proche-infrarouge
NNK : 4-(méthylnitrosamino)-1-(3-pyridyl)-1-butanone
NNN : N'-nitrosonornicotine

OLS : moindres carrés ordinaires
PC : composante principale
PCA : analyse en composantes principales
PLS-DA : analyse discriminante par moindres carrés partiels
PLSR : régression par moindres carrés partiels
PO : ordre polynomial
PRESS : somme des carrés des erreurs résiduelles prédites
ps : base de poids sec
R : coefficient de corrélation
R^2 : coefficients de détermination
RBF : fonction de base radiale
REP : erreur relative de prédiction
RER : Range Error Ratio
RMSE : erreur quadratique moyenne
RN : normalisation par gamme
RPD : Ratio of Prediction-to-Deviation
SD : écart-type
SEL : sélectivité
SEN : sensibilité
SG FD : dérivée de premier ordre de Savitzky-Golay
SG SD : dérivée de deuxième ordre de Savitzky-Golay
SGS : lissage de Savitzky-Golay
SNV : Standard Normal Variate
ST : tabac sans fumée
SVM-C : classification par machine à vecteurs de support
SVMR : régression par machine à vecteurs de support
SVs : vecteurs de support
TSNAs : nitrosamines spécifiques du tabac
UTC : United Tobacco Company
UV-Vis : Ultraviolet-Visible
UVN : normalisation par vecteur unitaire
VIP : Importance des variables en projection
WLS : moindres carrés pondérés
γ : sensibilité analytique

APPENDICE B

Appendice B / Annexe 1 : Exemple de rapport de vérification de performance exécuté à la fin de notre étude.

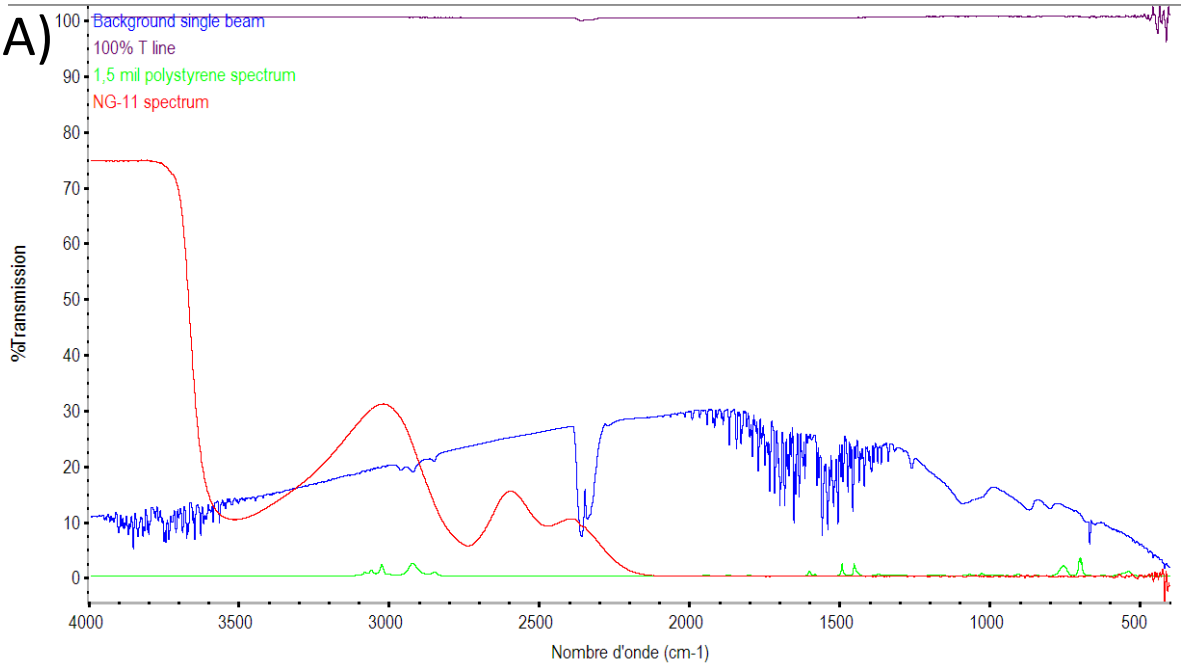
Vérification de performance

Opérateur: pc
Date: Lun; 27 Fév 2023; 12:11 (GMT+01:00)
Instrument: Nicolet iS10 Nr Série: AKY1910221

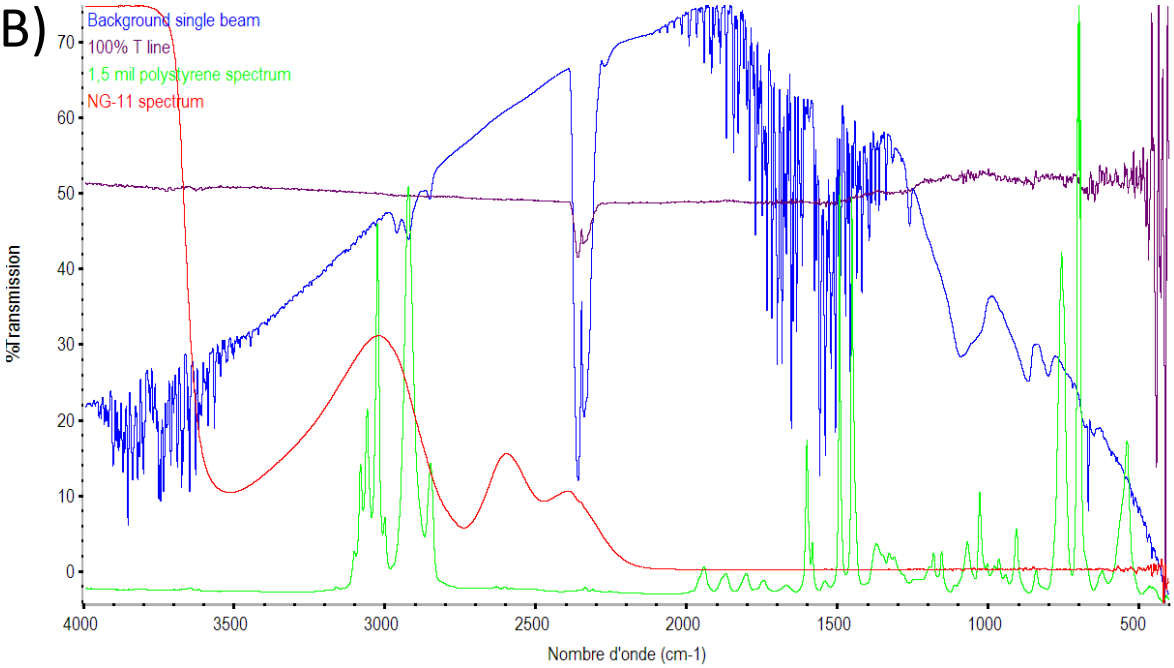
Description du Test	Limite haute	Limite basse	Mesuré	Résultat
Rapport d'énergie (Simple faisceau)				
Rapport d'énergie 4000 / 2000	1,0	0,2	0,365	Passe
Rapport d'énergie 2000 / 1000	4,0	0,9	1,872	Passe
Niveau de bruit (100 %T)				
Bruit pic à pic 4050 - 3950 (%T)	0,3	0,0	0,022	Passe
Bruit pic à pic 2050 - 1950 (%T)	0,3	0,0	0,017	Passe
Bruit pic à pic 1050 - 950 (%T)	0,3	0,0	0,149	Passe
Bruit pic à pic 550 - 450 (%T)	3,0	0,0	2,657	Passe
Bruit RMS 4050 - 3950	0,05	0,0	0,005	Passe
Bruit RMS 2050 - 1950	0,05	0,0	0,004	Passe
Bruit RMS 1050 - 950	0,05	0,0	0,029	Passe
Bruit RMS 550 - 450	0,6	0,0	0,388	Passe
Précision en nombre d'onde (1.5 mil Polystyrene)				
Pic à 3060.0 (cm-1)	3061,0	3059,0	3059,795	Passe
Pic à 1601.2 (cm-1)	1602,2	1600,2	1601,164	Passe
Pic à 1028.3 (cm-1)	1029,3	1027,3	1028,405	Passe
Répétabilité en Y(NG11 glass)				
Intensity (%T) à 3990.0	85,0	65,0	74,666	Passe
Intensity (%T) à 3031.0	42,0	22,0	30,909	Passe
Intensity (%T) à 2598.0	25,0	5,0	15,312	Passe
Intensity (%T) à 2010.0	10,01	-9,99	0,007	Passe

Appendice B / Annexe 2 : Spectres résultants de la vérification du bruit, de la ligne de base et des matériaux de référence. (A) Échelle commune et (B) Pleine échelle.

(A)



(B)



RÉFÉRENCES

1. Klus H., Kunze M., König S., Pöschl E., "Smokeless tobacco-An overview", Contributions to Tobacco & Nicotine Research, V. 23, n° 5, (2009), 248-76.

2. "Le tabac en Algérie", (2005), http://alger-roi.fr/Alger/documents_algeriens/economique/pages/63_tabac.htm. (Accès le 5 Mars 2023).

3. Laboratoire Central de Contrôle Qualité de l'entreprise United Tobacco Company (UTC du groupe MADAR Holding, Boumerdès).

4. Musk A. W., De Klerk N. H., "History of tobacco and health", Respirology, V. 8, n° 3, (2003), 286-90.

5. Bakdash A., "Shammah (smokeless tobacco) and public health", Asian Pacific Journal of Cancer Prevention: APJCP, V. 18, n° 5, (2017), 1183.

6. Severson H. H., Hatsukami D., "Smokeless tobacco cessation", Primary Care: Clinics in Office Practice, V. 26, n° 3, (1999), 529-51.

7. McDougall J, "Chapter 2: French colonization", de: "A History of Algeria", Cambridge University Press, (2017).

8. Ministère du Commerce et de la Promotion des Exportations, "Arrêté n° 04-331 : Réglementation de la fabrication, de l'importation et de la distribution des produits du tabac", (18 Octobre 2004), https://www.commerce.gov.dz/fr (Accès le 18 Avril 2022).

9. Plants of the World Online, "Nicotiana L.", Royal Botanic Gardens, Kew., https://powosciencekeworg/taxon/325974-2 (Accès le 29 Décembre 2023).

10. Reza J. A., Somayeh Z., Mojtaba A., "Ecological Roles and Biological Activities of Specialized Metabolites from the Genus Nicotiana", Chemical Reviews, V. 117, (2017), 12227−12280.

11. Köhler F. E, "Köhler's Medizinal-Pflanzen", http://pharm1.pharmazie.uni-greifswald.de/allgemei/koehler/koeh-eng.htm (Accès le 5 Janvier 2024).

12. Alsanosy R. M., "Smokeless tobacco (shammah) in Saudi Arabia: a review of its pattern of use, prevalence, and potential role in oral cancer", Asian Pacific Journal of Cancer Prevention, V. 15, n° 16, (2014), 6477-83.

13. Molina-Hidalgo F. J., Vazquez-Vilar M., D'Andrea L., Demurtas O. C., *et al.*, "Engineering metabolism in Nicotiana species: a promising future", Trends in biotechnology, V. 39, n° 9, (2021), 901-13.

14. Zenkner F. F., Margis-Pinheiro M., Cagliari A., "Nicotine biosynthesis in Nicotiana: a metabolic overview", Tobacco Science, V. 56, n° 1, (2019), 1-9.

15. Benowitz N. L., Hukkanen J., Jacob III P., "Nicotine chemistry, metabolism, kinetics and biomarkers"; Dans: Henningfield, J. E., London, E. D., Pogun, S. (editors), "Nicotine Psychopharmacology", Handbook of Experimental Pharmacology, Springer, Berlin, Heidelberg, V. 192, (2009), 29-60.

16. PubChem Compound Database, "Nicotine", United States National Library of Medicine – National Center for Biotechnology Information, (2019), https://pubchemncbinlmnihgov/compound/nicotine (Accès le 7 Janvier 2024).

17. Armstrong D. W., Wang X., Ercal N., "Enantiomeric composition of nicotine in smokeless tobacco, medicinal products, and commercial reagents", Chirality, V. 10, n° 7, (1998), 587-91.

18. Banyasz J. L., "The physical chemistry of nicotine"; Dans: Gorrod J. W., Jacob 3rd P. (editors), "Analytical determination of nicotine, and related compounds and their metabolites", Elsevier, Amsterdam, (1999), 149-90.

19. Wogan G. N., Hecht S. S., Felton J. S., Conney A. H., *et al.*, "Environmental and chemical carcinogenesis", Seminars in cancer biology, Elsevier, (2004), 473-486.

20. Yildiz D., "Nicotine, its metabolism and an overview of its biological effects", Toxicon, V. 43, n° 6, (2004), 619-32.

21. Jacob 3rd P., Yu L., Shulgin A. T., Benowitz N. L., "Minor tobacco alkaloids as biomarkers for tobacco use: comparison of users of cigarettes, smokeless tobacco, cigars, and pipes", American journal of public health, V. 89, n° 5, (1999), 731-6.

22. Michigan Department of Health & Human Services, "Types of tobacco products", https://www.michigan.gov/mdhhs/keep-mi-healthy/chronicdiseases/tobacco/types-of-tobacco-products (Accès le 19 Janvier 2024).

23. Yang H., Ma C., Zhao M., Magnussen C. G., Xi B., "Prevalence and trend of smokeless tobacco use and its associated factors among adolescents aged 12–16 years in 138 countries/territories, 1999–2019", BMC medicine, V. 20, n° 1, (2022), 460.

24. Song M.-A., Marian C., Brasky T. M., Reisinger S., *et al.*, "Chemical and toxicological characteristics of conventional and low-TSNA moist snuff tobacco products", Toxicology letters, V. 245, (2016), 68-77.

25. Lawler T. S., Stanfill S. B., Zhang L., Ashley D. L., Watson C. H., "Chemical characterization of domestic oral tobacco products: total nicotine, pH, unprotonated nicotine and tobacco-specific *N*-nitrosamines", Food and chemical toxicology, V. 57, (2013), 380-6.

26. STOP organization, "5 tobacco products that are fueling a global epidemic", https://exposetobaccoorg/news/tobacco-products/ (Accès le 7 Janvier 2024).

27. Scientific Figure on ResearchGate, "An experimental investigation into the operation of an electrically heated tobacco system", ResearchGate, https://www.researchgate.net/figure/a-The-three-components-of-the-Electrically-Heated-Tobacco-System-EHTS-and-b_fig1_337750942 (Accès le 23 Janvier 2024).

28. Electronic cigarette description, "E-cigarettes: The Basics", https://tobaccotactics.org/article/e-cigarettes-the-basics/ (Accès le 21 Janvier 2024).

29. "Nicotine Replacement Therapy Market to Reach USD 52.1 Bn by 2031 | TMR", https://www.openpr.com/news/3422227/nicotine-replacement-therapy-market-to-reach-usd-52-1-bn (Accès le 31 Janvier 2024).

30. WHO, "Tobacco: Key facts", (2020), https://www.who.int/news-room/fact-sheets/detail/tobacco#:~:text=Around%2080%25%20of%20the%20world's,(WHO%20FCTC)%20in%202003 (Accès le 2 Février 2023).

31. Sand L., Wallström M., Hirsch J.-M., "Smokeless tobacco, viruses and oral cancer", Oral Health Dent. Manag., V. 13, n° 2, (2014), 372-8.

32. Richter P., Spierto F. W., "Surveillance of smokeless tobacco nicotine, pH, moisture, and unprotonated nicotine content", Nicotine & tobacco research, V. 5, n° 6, (2003), 885-9.

33. Singh A., Thomas S., Dagli R., Bhateja G. A., *et al.*, "Prevalence oral mucosal lesions among moist snuff users in Jodhpur, India", Journal of Health Research and Reviews (In Developing Countries), V. 1, n° 2, (2014), 54-8.

34. Ali H., Pätzold R., Brückner H., "Determination of L-and D-amino acids in smokeless tobacco products and tobacco", Food chemistry, V. 99, n° 4, (2006), 803-12.

35. Henningfield J., Fant R., Tomar S., "Smokeless tobacco: an addicting drug", Advances in dental research, V. 11, n° 3, (1997), 330-5.

36. Fant R. V., Henningfield J. E., Nelson R. A., Pickworth W. B., "Pharmacokinetics and pharmacodynamics of moist snuff in humans", Tobacco Control, V. 8, n° 4, (1999), 387-92.

37. Tomar S. L., Henningfield J. E., "Review of the evidence that pH is a determinant of nicotine dosage from oral use of smokeless tobacco", Tobacco Control, V. 6, n° 3, (1997), 219-25.

38. Li P., Zhang J., Sun S.-H., Xie J.-P., Zong Y.-L., "A novel model mouth system for evaluation of In Vitro release of nicotine from moist snuff", Chemistry Central Journal, V. 7, n° 1, (2013), 1-9.

39. Hoffmann D., Djordjevic M. V., Fan J., Zang E.*, et al.*, "Five leading US commercial brands of moist snuff in 1994: assessment of carcinogenic N-nitrosamines", JNCI: Journal of the National Cancer Institute, V. 87, n° 24, (1995), 1862-9.

40. McAdam K. G., Faizi A., Kimpton H., Porter A., Rodu B., "Polycyclic aromatic hydrocarbons in US and Swedish smokeless tobacco products", Chemistry Central Journal, V. 7, (2013), 1-18.

41. Pappas R., Stanfill S., Watson C., Ashley D., "Analysis of toxic metals in commercial moist snuff and Alaskan iqmik", Journal of analytical toxicology, V. 32, n° 4, (2008), 281-91.

42. Walsh P. M., Epstein J. B., "The oral effects of smokeless tobacco", J. Can. Dent. Assoc., V. 66, n° 1, (2000).

43. Stepanov I., Hecht S. S., Ramakrishnan S., Gupta P. C., "Tobacco-specific nitrosamines in smokeless tobacco products marketed in India", International Journal of Cancer, V. 116, n° 1, (2005), 16-9.

44. Wang J., Yang H., Shi H., Zhou J.*, et al.*, "Nitrate and nitrite promote formation of tobacco-specific nitrosamines via nitrogen oxides intermediates during postcured storage under warm temperature", Journal of Chemistry, V. 2017, (2017).

45. FDA, "Family Smoking Prevention and Tobacco Control Act - An Overview", (2020), https://www.fda.gov/tobacco-products/rules-regulations-and-guidance/family-smoking-prevention-and-tobacco-control-act-overview/ (Accès le 1 Octobre 2023).

46. CDC, "Protocol to measure the quantity of nicotine contained in smokeless tobacco products manufactured, imported, or packaged in the United States", Federal Register, V. 62, n° 85, (1997), 24116-19.

47. CDC, "Notice regarding revisions to the laboratory protocol to measure the quantity of nicotine contained in smokeless tobacco products manufactured, imported, or packaged in the United States", Federal Register, V. 74, n° 4, (2009), 712–9.

48. Bates C., Fagerström K., Jarvis M., Kunze M.*, et al.*, "European Union policy on smokeless tobacco: a statement in favour of evidence based regulation for public health", Tobacco Control, V. 12, n° 4, (2003), 360-7.

49. Brunnemann K. D., Genoble L., Hoffmann D., "*N*-Nitrosamines in chewing tobacco: An international comparison", Journal of agricultural and food chemistry, V. 33, n° 6, (1985), 1178-81.

50. Freitag S., Sulyok M., Logan N., Elliott C. T., Krska R., "The potential and applicability of infrared spectroscopic methods for the rapid screening and routine analysis of mycotoxins in food crops", Comprehensive Reviews in Food Science and Food Safety, V. 21, n° 6, (2022), 5199-224.

51. Sen R. K., Karthikeyan K., Prabhakar P., Vishwakarma J.*, et al.*, "Fast tracking of adulterants and bacterial contamination in food via Raman and infrared spectroscopies: paving the way for a healthy and safe world", Sensors & Diagnostics, V. 1, n° 4, (2022), 673-85.

52. Alkhuder K., "Attenuated total reflection-Fourier transform infrared spectroscopy: A universal analytical technique with promising applications in forensic analyses", International Journal of Legal Medicine, V. 136, n° 6, (2022), 1717-36.

53. Tomasi E., "Rapid, non-destructive, and on-site capable detection and quantification of protein adulteration in insect-based food products with benchtop and miniaturized vibrational spectroscopic methods", Master thesis, Leopold-Franzens-Universität Innsbruck, (2009), 86 pages.

54. Haas J., Mizaikoff B., "Advances in mid-infrared spectroscopy for chemical analysis", Annual Review of Analytical Chemistry, V. 9, (2016), 45-68.

55. Xu Y., Zhang J., Wang Y., "Recent trends of multi-source and non-destructive information for quality authentication of herbs and spices", Food chemistry, V. 398, (2023), 133939.

56. Stuart B. H., "Infrared spectroscopy: fundamentals and applications", John Wiley & Sons, (2004).

57. Pagnin L., "Influence of atmospheric aging on the stability of binding media in paintings: Study of their degradation behaviour when exposed to UV-light, corrosive gases, and humidity", Doctoral thesis, eingereicht an der Technischen Universität Wien, (2021), 166 pages.

58. Specac, "Attenuated Reflectance Spectroscopy"; Dans: "FTIR – Transmission vs ATR spectroscopy", https://specac.com/theory-articles/transmission-vs-atr-spectroscopy-animated-guides/ (Accès le 11 Février 2024).

59. Chu X., Huang Y., Yun Y.-H., Bian X., "Chemometric methods in analytical spectroscopy technology", Springer, (2022), 596 pages.

60. de la Guardia M., Garrigues S., "Handbook of green analytical chemistry", Wiley Online Library, (2012), 546 pages.

61. Olivieri A. C., "Introduction to multivariate calibration: A practical approach", Springer, (2018), 250 pages.

62. Olivieri A. C., Faber N. M., Ferré J., Boqué R.*, et al.*, "Uncertainty estimation and figures of merit for multivariate calibration (IUPAC Technical Report)", Pure and Applied Chemistry, V. 78, n° 3, (2006), 633-61.

63. Camo Software, "Help documentation of The Unscrambler® X 10.4", Camo Software AS., Norway, (2016).

64. Cheng K., Lu Z., Zhou Y., Shi Y., Wei Y., "Global sensitivity analysis using support vector regression", Applied Mathematical Modelling, V. 49, (2017), 587-98.

65. Ortaç-kabaoğlu R., "A support vector regression method for reducing the high-order systems to first-order plus time-delay forms", IU-Journal of Electrical & Electronics Engineering, V. 11, n° 1, (2012), 1305-9.

66. Smola A. J., Schölkopf B., "A tutorial on support vector regression", Statistics and computing, V. 14, (2004), 199-222.

67. Awad M., Khanna R. "Efficient learning machines: theories, concepts, and applications for engineers and system designers", Springer nature, (2015), 263 pages.

68. Soltanikazemi M., Abdanan Mehdizadeh S., Heydari M., Faregh S. M., "Development of a smart spectral analysis method for the determination of mulberry (Morus alba var. nigra L.) juice quality parameters using FT-IR spectroscopy", Food Science & Nutrition, V. 11, n° 4, (2023), 1808-17.

69. Tharwat A., "Principal component analysis-a tutorial", International Journal of Applied Pattern Recognition, V. 3, n° 3, (2016), 197-240.

70. Smith L. I., "A tutorial on principal components analysis", (2002), 27 pages. https://ourarchive.otago.ac.nz/bitstream/handle/10523/7534/OUCS-2002-12.pdf

71. Abdi H., Williams L. J., "Principal Component Analysis", Wiley interdisciplinary reviews: computational statistics, V. 2, n° 1, (2010), 97-106.

72. Shlens J., "A tutorial on principal component analysis", arXiv preprint arXiv:14041100, (2014), 12 pages. https://doi.org/10.48550/arXiv.1404.1100

73. Khan L., Luo F., "Hierarchical clustering for complex data", International Journal on Artificial Intelligence Tools, V. 14, n° 05, (2005), 791-809.

74. Murtagh F., Contreras P., "Algorithms for hierarchical clustering: an overview", Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, V. 2, n° 1, (2012), 86-97.

75. Nielsen F., "Chapter 8: Hierarchical clustering"; Dans: "Introduction to HPC with MPI for Data Science", Springer, (2016), 195-211.

76. Miller J., Miller J. C., "Statistics and chemometrics for analytical chemistry", 6 ed, Pearson education, (2010), 297 pages.

77. Sisodia D., Singh L., Sisodia S., Saxena K., "Clustering techniques: a brief survey of different clustering algorithms", International Journal of Latest Trends in Engineering and Technology (IJLTET), V. 1, n° 3, (2012), 82-7.

78. Addinsoft, "XLSTAT version 2016.02 Help topics", Addinsoft SARL, France, (2016).

79. da Mata M. M., Rocha P. D., de Farias I. K. T., da Silva J. L. B.*, et al.*, "Distinguishing cotton seed genotypes by means of vibrational spectroscopic methods (NIR and Raman) and chemometrics", Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, V. 266, (2022), 120399.

80. Esteki M., Memarbashi N., Simal-Gandara J., "Classification and authentication of tea according to their harvest season based on FT-IR fingerprinting using pattern recognition methods", Journal of Food Composition and Analysis, V. 115, (2023), 104995.

81. Sharma C. P., Sharma S., Rawat G. S., Singh R., "Rapid and non-destructive differentiation of Shahtoosh from Pashmina/Cashmere wool using ATR FT-IR spectroscopy", Science & Justice, V. 62, n° 3, (2022), 349-57.

82. Agustika D. K., Mercuriani I., Purnomo C. W., Hartono S.*, et al.*, "Fourier transform infrared spectrum pre-processing technique selection for detecting PYLCV-infected chilli plants", Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, V. 278, (2022), 121339.

83. Liu S., Lei T., Li G., Liu S.*, et al.*, "Rapid detection of micronutrient components in infant formula milk powder using near-infrared spectroscopy", Frontiers in Nutrition, V. 10, (2023), 1-12.

84. Fearn T., "Assessing calibrations: SEP, RPD, RER and R^2", NIR news, V. 13, n° 6, (2002), 12-3.

85. Olivieri A. C., "Analytical figures of merit: from univariate to multiway calibration", Chemical reviews, V. 114, n° 10, (2014), 5358-78.

86. Olivieri A. C., "Practical guidelines for reporting results in single-and multi-component analytical calibration: A tutorial", Analytica Chimica Acta, V. 868, (2015), 10-22.

87. Tamiji Z., Habibi Z., Pourjabbar Z., Khoshayand M. R.*, et al.*, "Detection and quantification of adulteration in turmeric by spectroscopy coupled with chemometrics", Journal of Consumer Protection and Food Safety, V. 17, n° 3, (2022), 221-30.

88. Williams P., "The RPD statistic: A tutorial note", NIR news, V. 25, n° 1, (2014), 22-6.

89. Amorim M. V., Costa F. S., Aragão C. F., Lima K. M., "The use of near infrared spectroscopy and multivariate calibration for determining the active principle of olanzapine in a pharmaceutical formulation", Journal of the Brazilian Chemical Society, V. 28, n°, (2017), 920-6.

90. Franco V. G., Mantovani V. E., Goicoechea H. C., Olivieri A. C., "Teaching chemometrics with a bioprocess: analytical methods comparison using bivariate linear regression", The Chemical Educator, V. 7, (2002), 265-9.

91. Galea-Rojas M., de Castilho M. V., Bolfarine H., de Castro M., "Detection of analytical bias", Analyst, V. 128, n° 8, (2003), 1073-81.

92. González A. G., Herrador M. A., Asuero A. n. G., "Intra-laboratory testing of method accuracy from recovery assays", Talanta, V. 48, n° 3, (1999), 729-36.

93. Elfiky A. M., Shawky E., Khattab A. R., Ibrahim R. S., "Integration of NIR spectroscopy and chemometrics for authentication and quantitation of adulteration in sweet marjoram (Origanum majorana L.)", Microchemical Journal, V. 183, (2022), 108125.

94. Menevseoglu A., Gumus-Bonacina C. E., Gunes N., Ayvaz H., Dogan M. A., "Infrared spectroscopy-based rapid determination of adulteration in commercial sheep's milk cheese via n-hexane and ethanolic extraction", International Dairy Journal, V. 138, (2023), 105543.

95. Xu W., Xia J., Min S., Xiong Y., "Fourier transform infrared spectroscopy and chemometrics for the discrimination of animal fur types", Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, V. 274, (2022), 121034.

96. De Andrade J. C., Galvan D., Effting L., Lelis C.*, et al.*, "An Easy-to-Use and Cheap Analytical Approach Based on NIR and Chemometrics for Tomato and Sweet Pepper Authentication by Non-volatile Profile", Food Analytical Methods, V. 16, n° 3, (2023), 567-80.

97. Rozali N. L., Azizan K. A., Singh R., Jaafar S. N. S.*, et al.*, "Fourier transform infrared (FTIR) spectroscopy approach combined with discriminant analysis and prediction model for crude palm oil authentication of different geographical and temporal origins", Food Control, V. 146, (2023), 109509.

98. Cullen D., Keithly L., Kane K., Land T.*, et al.*, "Smokeless tobacco products sold in Massachusetts from 2003 to 2012: trends and variations in brand availability, nicotine contents and design features", Tobacco control, V. 24, n° 3, (2015), 256-62.

99. Stepanov I., Hecht S. S., "Tobacco-specific nitrosamines and their pyridine-N-glucuronides in the urine of smokers and smokeless tobacco users", Cancer Epidemiology Biomarkers & Prevention, V. 14, n° 4, (2005), 885-91.

100. Prasad G. L., Jones B. A., Schmidt E., Chen P., Kennedy A. D., "Global metabolomic profiles reveal differences in oxidative stress and inflammation pathways in smokers and moist snuff consumers", Journal of metabolomics, V. 1, (2015), 2.

101. Rostron B. L., Chang C. M., van Bemmel D. M., Xia Y., Blount B. C., "Nicotine and toxicant exposure among US smokeless tobacco users: results from 1999 to 2012 National Health and Nutrition Examination Survey Data", Cancer Epidemiology, Biomarkers & Prevention, V. 24, n° 12, (2015), 1829-37.

102. Stanfill S. B., Connolly G. N., Zhang L., Jia L. T.*, et al.*, "Global surveillance of oral tobacco products: total nicotine, unionised nicotine and tobacco-specific *N*-nitrosamines", Tobacco control, V. 20, n° 3, (2010), e2-e2.

103. Stanfill S. B., da Silva A. L. O., Lisko J. G., Lawler T. S.*, et al.*, "Comprehensive chemical characterization of Rapé tobacco products: Nicotine, un-ionized nicotine, tobacco-specific *N'*-nitrosamines, polycyclic aromatic hydrocarbons, and flavor constituents", Food and Chemical Toxicology, V. 82, (2015), 50-8.

104. Gupta A. K., Tulsyan S., Bharadwaj M., Mehrotra R., "Grass roots approach to control levels of carcinogenic nitrosamines, NNN and NNK in smokeless tobacco products", Food and Chemical Toxicology, V. 124, (2019), 359-66.

105. Chen C., Isabelle L., Pickworth W., Pankow J., "Levels of mint and wintergreen flavorants: smokeless tobacco products vs. confectionery products", Food and chemical toxicology, V. 48, n° 2, (2010), 755-63.

106. Lisko J. G., Stanfill S. B., Duncan B. W., Watson C. H., "Application of GC-MS/MS for the analysis of tobacco alkaloids in cigarette filler and various tobacco species", Analytical chemistry, V. 85, n° 6, (2013), 3380-4.

107. Rainey C. L., Conder P. A., Goodpaster J. V., "Chemical characterization of dissolvable tobacco products promoted to reduce harm", Journal of agricultural and food chemistry, V. 59, n° 6, (2011), 2745-51.

108. Nasrin S., Chen G., Watson C. J., Lazarus P., "Comparison of tobacco-specific nitrosamine levels in smokeless tobacco products: High levels in products from Bangladesh", PloS one, V. 15, n° 5, (2020), e0233111.

109. Brunnemann K., Qi J., Hoffmann D., "Chemical profile of two types of oral snuff tobacco", Food and chemical toxicology, V. 40, n° 11, (2002), 1699-703.

110. Ji H., Wu Y., Fannin F., Bush L., "Determination of tobacco alkaloid enantiomers using reversed phase UPLC/MS/MS", Heliyon, V. 5, n° 5, (2019), e01719.

111. Moghbel N., Ryu B., Cabot P. J., Steadman K. J., "In vitro cytotoxicity of Nicotiana gossei leaves, used in the Australian Aboriginal smokeless tobacco known as pituri or mingkulpa", Toxicology letters, V. 254, (2016), 45-51.

112. Stepanov I., Jensen J., Hatsukami D., Hecht S. S., "New and traditional smokeless tobacco: comparison of toxicant and carcinogen levels", Nicotine & Tobacco Research, V. 10, n° 12, (2008), 1773-82.

113. Lawler T. S., Stanfill S. B., Tran H. T., Lee G. E.*, et al.*, "Chemical analysis of snus products from the United States and northern Europe", PloS one, V. 15, n° 1, (2020), e0227837.

114. Prokopczyk B., Wu M., Cox J. E., Amin S.*, et al.*, "Improved methodology for the quantitative assessment of tobacco-specific N-nitrosamines in tobacco by supercritical fluid extraction", Journal of Agricultural and Food Chemistry, V. 43, n° 4, (1995), 916-22.

115. Stepanov I., Villalta P. W., Knezevich A., Jensen J.*, et al.*, "Analysis of 23 polycyclic aromatic hydrocarbons in smokeless tobacco by gas chromatography− mass spectrometry", Chemical research in toxicology, V. 23, n° 1, (2010), 66-73.

116. Kumar A., Bhartiya D., Kaur J., Kumari S.*, et al.*, "Regulation of toxic contents of smokeless tobacco products", The Indian Journal of Medical Research, V. 148, n° 1, (2018), 14.

117. McNeill A., Bedi R., Islam S., Alkhatib M., West R., "Levels of toxins in oral tobacco products in the UK", Tobacco control, V. 15, n° 1, (2006), 64-7.

118. McAdam K., Kimpton H., Porter A., Liu C.*, et al.*, "Comprehensive survey of radionuclides in contemporary smokeless tobacco products", Chemistry Central Journal, V. 11, (2017), 1-20.

119. Fisher M. T., Bennett C. B., Hayes A., Kargalioglu Y.*, et al.*, "Sources of and technical approaches for the abatement of tobacco specific nitrosamine formation in moist smokeless tobacco products", Food and Chemical Toxicol., V. 50, n° 3-4, (2012), 942-8.

120. Djordjevic M. V., Fan J., Bush L. P., Brunnemann K. D., Hoffmann D., "Effects of storage conditions on levels of tobacco-specific N-nitrosamines and N-nitrosamino acids in US moist snuff", J. of Agricultural and Food Chemistry, V. 41, n° 10, (1993), 1790-4.

121. Lv F., Guo J., Yu F., Zhang T.*, et al.*, "Determination of nine volatile N-nitrosamines in tobacco and smokeless tobacco products by dispersive solid-phase extraction with gas chromatography and tandem mass spectrometry", Journal of Separation Science, V. 39, n° 11, (2016), 2123-8.

122. Djordjevic M. V., Brunnemann K. D., Hoffmann D., "Identification and analysis of a nicotine-derived N-nitrosamino acid and other nitrosamino acids in tobacco", Carcinogenesis, V. 10, n° 9, (1989), 1725-31.

123. Hoffmann D., Diordjevic M., Brunnemann K., "New brands of oral snuff", Food and Chemical Toxicology, V. 29, n° 1, (1991), 65-8.

124. Rickert W., Joza P., Trivedi A., Momin R.*, et al.*, "Chemical and toxicological characterization of commercial smokeless tobacco products available on the Canadian market", Regulatory Toxicology and Pharmacology, V. 53, n° 2, (2009), 121-33.

125. McAdam K., Kimpton H., Vas C., Rushforth D.*, et al.*, "The acrylamide content of smokeless tobacco products", Chemistry Central Journal, V. 9, (2015), 1-14.

126. Chamberlain W. J., Schlotzhauer W. S., Chortyk O. T., "Chemical composition of nonsmoking tobacco products", Journal of Agricultural and Food Chemistry, V. 36, n° 1, (1988), 48-50.

127. Stanfill S. B., Jia L. T., Ashley D. L., Watson C. H., "Rapid and chemically selective nicotine quantification in smokeless tobacco products using GC-MS", Journal of Chromatographic Science, V. 47, (2009), 902-9.

128. McAdam K. G., Kimpton H., Faizi A., Porter A., Rodu B., "The composition of contemporary American and Swedish smokeless tobacco products", BMC chemistry, V. 13, n° 1, (2019), 1-15.

129. Stepanov I., Biener L., Yershova K., Nyman A. L.*, et al.*, "Monitoring tobacco-specific N-nitrosamines and nicotine in novel smokeless tobacco products: findings from round II of the new product watch", Nicotine & tobacco research, V. 16, n° 8, (2014), 1070-8.

130. Helrich K., "Official methods of analysis of the Association of Official Analytical Chemists", 15th ed., Association of official analytical chemists, (1990), 66 pages.

131. Amith H., Agrawal D., Gupta A., Shrivastava T. P.*, et al.*, "Assessing the nicotine content of smoked and smokeless forms of Tobacco Available in Bhopal", Indian Journal of Dental Research, V. 29, n° 3, (2018), 341.

132. Li P., Zeng S., Zhang J., Shen Y.*, et al.*, "Real-Time Monitoring of Nicotine Release Behavior from Smokeless Tobacco (Snus) Based on Fiber Optic Sensing Technology", Dissolution Technologies, V. 26, n° 4, (2019), 24-30.

133. McAdam K., Vas C., Kimpton H., Faizi A.*, et al.*, "Ethyl carbamate in Swedish and American smokeless tobacco products and some factors affecting its concentration", Chemistry Central Journal, V. 12, n° 1, (2018), 1-17.

134. AOAC International, "AOAC Official Method 960.07 Alkaloids (Total as Nicotine) in Tobacco Distillation Method", Gaithersburg, MD, (1995), 30-1.

135. Ciolino L. A., McCauley H. A., Fraser D. B., Barnett D. Y.*, et al.*, "Reversed phase ion-pair liquid chromatographic determination of nicotine in commercial tobacco products. 1. Moist snuff", Journal of agricultural and food chemistry, V. 47, n° 9, (1999), 3706-12.

136. Zhang Y., Cong Q., Xie Y., Zhao B., "Quantitative analysis of routine chemical constituents in tobacco by near-infrared spectroscopy and support vector machine", Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, V. 71, n° 4, (2008), 1408-13.

137. Borges Miranda A., Pérez Martínez C., Jiménez Chacón J., Álvarez Prieto M., "Near infrared spectroscopic analysis of total alkaloids as nicotine, total nitrogen and total ash in Cuban cigar tobacco", Journal of Near Infrared Spectroscopy, V. 27, n° 2, (2019), 123-33.

138. Wu L., Wang B., Zhang L., Duan R.*, et al.*, "Determination of routine chemicals, physical indices and macromolecular substances in reconstituted tobacco using near infrared spectroscopy combined with sample set partitioning", Journal of Near Infrared Spectroscopy, V. 28, n° 3, (2020), 153-62.

139. Jiang D., Hu G., Qi G., Mazur N., "A fully convolutional neural network-based regression approach for effective chemical composition analysis using near-infrared spectroscopy in cloud", J. of Artif. Intell. and Tech., V. 1, n° 1, (2021), 74-82.

140. Zhu Z., Qi G., Lei Y., Jiang D.*, et al.*, "A long short-term memory neural network based simultaneous quantitative analysis of multiple tobacco chemical components by near-infrared hyperspectroscopy images", Chemosensors, V. 10, n° 5, (2022), 164.

141. Geng Y., Shen H., Ni H., Tian Y.*, et al.*, "Non-destructive determination of total sugar content in tobacco filament based on calibration transfer with parameter free adjustment", Microchemical Journal, V. 181, (2022), 107797.

142. Tan C., Qin X., Li M., "Comparison of chemometric methods for brand classification of cigarettes by near-infrared spectroscopy", Vibrational Spectroscopy, V. 51, n° 2, (2009), 276-82.

143. Giokas D. L., Thanasoulias N. C., Vlessidis A. G., "Multivariate chemometric discrimination of cigarette tobacco blends based on the UV–Vis spectrum of their hydrophilic extracts", Journal of hazardous materials, V. 185, n° 1, (2011), 86-92.

144. Omar J., Slowikowski B., Boix A., "Chemometric approach for discriminating tobacco trademarks by near infrared spectroscopy", Forensic science international, V. 294, (2019), 15-20.

145. Marcelo M. C., Soares F. L., Ardila J. A., Dias J. C.*, et al.*, "Fast inline tobacco classification by near-infrared hyperspectral imaging and support vector machine-discriminant analysis", Analytical Methods, V. 11, n° 14, (2019), 1966-75.

146. Jiang D., Qi G., Hu G., Mazur N.*, et al.*, "A residual neural network based method for the classification of tobacco cultivation regions using near-infrared spectroscopy sensors", Infrared Physics & Technology, V. 111, (2020), 103494.

147. Cui L.-L., Chen H., Chen Z.-P., Yuan Y.-W.*, et al.*, "Geographical origin classification of tobacco by stable isotope and multi-elemental analysis in combination with chemometric methods", Microchemical Journal, V. 193, (2023), 109163.

148. Zhang H., Pang Y., Luo Y., Li X.*, et al.*, "Enantiomeric composition of nicotine in tobacco leaf, cigarette, smokeless tobacco, and e-liquid by normal phase high-performance liquid chromatography", Chirality, V. 30, n° 7, (2018), 923-31.

149. IARC, "Smokeless Tobacco and Some Tobacco-Specific N-Nitrosamines", World Health Organization International Agency for Research on Cancer, (2007), https://monographs.iarc.fr/ENG/recentpub/mono89.pdf (Accès le 6 Janvier 2018).

150. Stepanov I., Biener L., Knezevich A., Nyman A. L.*, et al.*, "Monitoring tobacco-specific *N*-nitrosamines and nicotine in novel Marlboro and Camel smokeless tobacco products: findings from Round 1 of the New Product Watch", Nicotine & Tobacco Research, V. 14, n° 3, (2012), 274-81.

151. Alhazmi H. A., Khalid A., Sultana S., Abdelwahab S. I.*, et al.*, "Determination of phytocomponents of twenty-one varieties of smokeless tobacco using gas chromatography-mass spectroscopy (GC-MS)", South African Journal of Chemistry, V. 72, (2019), 47-54.

152. Stepanov I., Jensen J., Biener L., Bliss R. L.*, et al.*, "Increased pouch sizes and resulting changes in the amounts of nicotine and tobacco-specific *N*-nitrosamines in single pouches of Camel Snus and Marlboro Snus", Nicotine & Tobacco Research, V. 14, n° 10, (2012), 1241-5.

153. Oudjehih M., Deltour I., Bouhidel M. L., Bouhidel A.*, et al.*, "Smokeless Tobacco Use, Cigarette Smoking, and Upper Aerodigestive Tract Cancers: A Case-Control Study in the Batna Region, Algeria, 2008-2011", Tobacco use insights, V. 13, (2020), 1-11.

154. Stanfill S. B., Croucher R. E., Gupta P. C., Lisko J. G.*, et al.*, "Chemical characterization of smokeless tobacco products from South Asia: Nicotine, unprotonated nicotine, tobacco-specific *N'*-Nitrosamines, and flavor compounds", Food and chemical toxicology, V. 118, (2018), 626-34.

155. Stepanov I., Gupta P. C., Parascandola M., Yershova K.*, et al.*, "Constituent variations in smokeless tobacco purchased in Mumbai, India", Tobacco Regulatory Science, V. 3, n° 3, (2017), 305-14.

156. Garrigues J. M., Pérez-Ponce A., Garrigues S., de la Guardia M., "Fourier-transform infrared determination of nicotine in tobacco samples by transmittance measurements after leaching with CHCl3", Analytica Chimica Acta, V. 373, n° 1, (1998), 63-71.

157. Stöbener A., Naefken U., Kleber J., Liese A., "Determination of trace amounts with ATR FTIR spectroscopy and chemometrics: 5-(hydroxymethyl) furfural in honey", Talanta, V. 204, (2019), 1-5.

158. Moreno-Ley C. M., Hernández-Martínez D. M., Osorio-Revilla G., Tapia-Ochoategui A. P.*, et al.*, "Prediction of coumarin and ethyl vanillin in pure vanilla extracts using MID-FTIR spectroscopy and chemometrics", Talanta, V. 197, (2019), 264-9.

159. Koch C., Posch A. E., Goicoechea H. C., Herwig C., Lendl B., "Multi-analyte quantification in bioprocesses by Fourier-transform-infrared spectroscopy by partial least squares regression and multivariate curve resolution", Analytica Chimica Acta, V. 807, (2014), 103-10.

160. Szentirmai V., Wacha A., Németh C., Kitka D.*, et al.*, "Reagent-free total protein quantification of intact extracellular vesicles by attenuated total reflection Fourier transform infrared (ATR-FTIR) spectroscopy", Analytical and Bioanalytical Chemistry, V. 412, (2020), 4619-28.

161. Thermo Fisher Scientific, "Help documentation of the OMNIC^{TM} software version 9.8", Thermo Fisher Scientific Inc., USA, (2017).

162. Kheawfu K., Kaewpinta A., Chanmahasathien W., Rachtanapun P., Jantrawut P., "Extraction of nicotine from tobacco leaves and development of fast dissolving nicotine extract film", Membranes, V. 11, n° 6, (2021), 403.

163. Schmidtke L. M., Smith J. P., Müller M. C., Holzapfel B. P., "Rapid monitoring of grapevine reserves using ATR–FT-IR and chemometrics", Analytica Chimica Acta, V. 732, (2012), 16-25.

164. Visak Z. P., Ilharco L. M., Garcia A. R., Najdanovic-Visak V.*, et al.*, "Volumetric properties and spectroscopic studies of pyridine or nicotine solutions in liquid polyethylene glycols", The Journal of Physical Chemistry B, V. 115, n° 26, (2011), 8481-92.

165. Rijal R., Sah M., Lamichhane H. P., Mallik H. S., "Quantum chemical calculations of nicotine and caffeine molecule in gas phase and solvent using DFT methods", Heliyon, V. 8, n° 12, (2022), e12494.

166. Pongjanyakul T., Khunawattanakul W., Puttipipatkhachorn S., "Physicochemical characterizations and release studies of nicotine–magnesium aluminum silicate complexes", Applied Clay Science, V. 44, n° 3-4, (2009), 242-50.

167. Pongjanyakul T., Khunawattanakul W., Strachan C. J., Gordon K. C.*, et al.*, "Characterization of chitosan–magnesium aluminum silicate nanocomposite films for buccal delivery of nicotine", International journal of biological macromolecules, V. 55, (2013), 24-31.

168. Coates J., “Interpretation of infrared spectra, a practical approach”, Encyclopedia of analytical chemistry, V. 12, (2000), 10815-10837.

169. Thermo Fisher Scientific, "TQ Analyst 9.7 software Help Topics, Thermo Fisher Scientific Inc., USA, (2017).

170. Antonio M., Carneiro R. L., Maggio R. M., "A comparative approach of MIR, NIR and Raman based chemometric strategies for quantification of Form I of Meloxicam in commercial bulk drug", Microchemical Journal, V. 180, (2022), 107575.

171. Iñón F. A., Garrigues J. M., Garrigues S., Molina A., de la Guardia M., "Selection of calibration set samples in determination of olive oil acidity by partial least squares–attenuated total reflectance–Fourier transform infrared spectroscopy", Analytica Chimica Acta, V. 489, n° 1, (2003), 59-75.

172. Cuadros-Rodríguez L., Bagur-González M. G., Sánchez-Vinas M., González-Casado A., Gómez-Sáez A. M., "Principles of analytical calibration/quantification for the separation sciences", Journal of Chromatography A, V. 1158, n° 1-2, (2007), 33-46.

173. Allegrini F., Olivieri A. C., "IUPAC-consistent approach to the limit of detection in partial least-squares calibration", Analytical chemistry, V. 86, n° 15, (2014), 7858-66.

174. Valinger D., Longin L., Grbeš F., Benković M.*, et al.*, "Detection of honey adulteration–The potential of UV-VIS and NIR spectroscopy coupled with multivariate analysis", Lwt, V. 145, (2021), 111316.

175. Tassew Z., Chandravanshi B. S., "Levels of nicotine in Ethiopian tobacco leaves", SpringerPlus, V. 4, (2015), 1-6.

176. Djordjevic M. V., Doran K. A., "Nicotine content and delivery across tobacco products"; Dans: Henningfield J. E., London E. D., Pogun S. (editors), "Nicotine psychopharmacology", Springer Science & Business Media, Heidelberg, (2009), 61-82.

177. Shu R., Ju L., Ni L., Wu S.*, et al.*, "Improving transferability and service life of the calibration model of total plant alkaloids in tobacco leaves on seven NIR spectroscopy devices by multi-step wavelength selection methods", Microchemical Journal, V. 196, (2024), 109522.

178. Cruz-Tirado J., de França R. L., Tumbajulca M., Barraza-Jáuregui G.*, et al.*, "Detection of cumin powder adulteration with allergenic nutshells using FT-IR and portable NIRS coupled with chemometrics", Journal of Food Composition and Analysis, V. 116, (2023), 105044.

179. Bureau S., Quilot-Turion B., Signoret V., Renaud C.*, et al.*, "Determination of the composition in sugars and organic acids in peach using mid infrared spectroscopy: comparison of prediction results according to data sets and different reference methods", Analytical chemistry, V. 85, n° 23, (2013), 11312-8.

180. Materazzi S., Gregori A., Ripani L., Apriceno A., Risoluti R., "Cocaine profiling: Implementation of a predictive model by ATR-FTIR coupled with chemometrics in forensic chemistry", Talanta, V. 166, (2017), 328-35.

181. Lan W., Baeten V., Jaillais B., Renard C. M.*, et al.*, "Comparison of near-infrared, mid-infrared, Raman spectroscopy and near-infrared hyperspectral imaging to determine chemical, structural and rheological properties of apple purees", Journal of Food Engineering, V. 323, (2022), 111002.

182. Idris A. M., Ibrahim S. O., Vasstrand E. N., Johannessen A. C.*, et al.*, "The Swedish snus and the Sudanese toombak: are they different?", Oral oncology, V. 34, n° 6, (1998), 558-66.

183. Ortiz M., Sarabia L., Herrero A., Sánchez M.*, et al.*, "Capability of detection of an analytical method evaluating false positive and false negative (ISO 11843) with partial least squares", Chemometrics and intelligent laboratory systems, V. 69, n° 1-2, (2003), 21-33.

184. ICH, "Validation of analytical procedures: text and methodology", Q2 (R1), V. 1, n° 20, (2005), 05 pages.

185. Largo-Gosens A., Hernández-Altamirano M., García-Calvo L., Alonso-Simón A.*, et al.*, "Fourier transform mid infrared spectroscopy applications for monitoring the structural plasticity of plant cell walls", Frontiers in plant science, V. 5, (2014), 303.

186. dos Santos V. H. J. M., Pontin D., Ponzi G. G. D., e Stepanha A. S. d. G.*, et al.*, "Application of Fourier Transform infrared spectroscopy (FTIR) coupled with multivariate regression for calcium carbonate ($CaCO_3$) quantification in cement", Construction and Building Materials, V. 313, (2021), 125413.

187. Gok S., Severcan M., Goormaghtigh E., Kandemir I., Severcan F., "Differentiation of Anatolian honey samples from different botanical origins by ATR-FTIR spectroscopy using multivariate analysis", Food chemistry, V. 170, (2015), 234-40.

PUBLICATIONS ET COMMUNICATIONS

PUBLICATIONS

1. Fekhar M., Daghbouche Y., Bouzidi N., El Hattab M., "ATR-FTIR spectroscopy combined with chemometrics for quantification of total nicotine in Algerian smokeless tobacco products", Microchemical Journal, V. 193, (2023), 109127. https://doi.org/10.1016/j.microc.2023.109127

2. Fekhar M., Daghbouche Y., Bouzidi N., El Hattab M., "Rapid assessment of smokeless tobacco quality parameters using ATR-FT-MIR spectroscopy: Comparison of analytical/mathematical and machine learning approaches", Microchemical Journal, V. 201, (2024), 110670. https://doi.org/10.1016/j.microc.2024.110670

COMMUNICATION

Fekhar M., Alim S., Bouzidi N., Daghbouche Y., El Hattab M., "Contrôle qualité des compléments sportifs de Lactosérum (la whey) par PLS-ATR-IRTF", 1er Séminaire National de Chimie Appliquée − SNCA2023, USD BLIDA 1, (9-10 Mai 2023). https://www.univ-blida.dz/evenement/seminaire-national-de-chimie-appliquee-snca-2023/

Microchemical Journal 193 (2023) 109127

Contents lists available at ScienceDirect

Microchemical Journal

journal homepage: www.elsevier.com/locate/microc

ATR-FTIR spectroscopy combined with chemometrics for quantification of total nicotine in Algerian smokeless tobacco products

Mohamed Fekhar, Yasmina Daghbouche *, Naima Bouzidi, Mohamed El Hattab

Laboratory of Natural Products Chemistry and of Biomolecules, Faculty of sciences, University Blida 1, Blida, Algeria

ARTICLE INFO

Keywords:
ATR-FTIR spectroscopy
Chemometrics
Nicotine
Smokeless tobacco
Results validation

ABSTRACT

Chemma, an oral moist snuff widespread used in Algeria, is prepared from grounded tobacco and inorganic salts or other plant materials. In addition to its adverse health issues, Chemma contains unknown amounts of nicotine and other harmful constituents. Twenty seven samples of commercial products and tobacco leaves varieties were characterized in the aim to quantitatively determine total nicotine by attenuated total reflectance Fourier transform infrared (ATR-FTIR) spectroscopy coupled with two regression methods (univariate and multivariate). For this purpose, a highly specific-to-analyte extraction method was applied on calibration and test sets of standards for a wide concentration range. The optimal analytical models were evaluated by selecting the appropriate spectral measurement, spectral pre-processing strategies and critical figures of merit estimated according to the latest approaches. Then, the predictive ability of the method was demonstrated with the elliptical joint confidence region test and with comparing to results in literature.

The linear baseline correction and the extended multiplicative scatter correction pre-processing of the combined reduced spectral regions 1333–1299 and 740–690 cm^{-1} followed by partial least squares regression (PLSR) offered the best analytical outcomes. Using this latter, total nicotine content in the manufactured smokeless tobacco (ST) products have been found ranging from 3.9 to 12.6 mg/g of dry weight; while in the ST leaves, it was between 5.9 and 45.0 mg/g, dry weight basis. The developed method, besides to the high predictive performance, provides advanced statistical tests that can qualitatively assess the target analyte in real sample spectra, allowing more reliable result.

With the increasing number of new products containing complex agents and adulterants, the aforementioned method can be successfully adapted to routine analyses for rapid quality control of nicotine in ST and cigarettes.

1. Introduction

In Algeria, smokeless tobacco is most commonly used orally in a form similar to an unflavored loose American-style moist snuff, known locally as Chemma pronounced Shammah. It is a finely ground tobacco, prepared from a mixture of simple cured *Nicotiana rustica* leaves and inorganic salts, alkaline ash or other plant materials sprinkled with treated water. It is taken by placing a pinch, directly or wrapped in a sheet of cigarette paper, between the upper lip and the teeth. Its popularity has been increasing among young generation for recreational purpose, as an alternative to cigarette smoking, or just as a blind imitation to the peers.

Most Algerian tobacco, which is used to make Chemma, is grown in the regions located to the east of the country (Constantinois and oases provinces). Several varieties of this species as well as several commercial sorts of varied qualities named according to the cultivation area can be found [1], and a conventional product could be a mixture of more than one (1) of these varieties.

Abbreviations: Anal. SEN, Analytical sensitivity; ATR, Attenuated total reflectance; AU, Absorbance unit; BC, Simple baseline correction using two points; CDC, Centers for disease control and prevention; $CHCl_3$, Chloroform; ESI, Electrospray ionization; FTIR, Fourier transform infrared; HPLC, High performance liquid chromatography; KBr, Potassium Bromide; LOD, Limit of detection; LOQ, Limit of quantification; MIR, Mid-Infrared; NCT, Nicotine; PLS, Partial least squares; QC, Quality control; R, Correlation coefficient; R^2, Coefficient of determination; REP, Relative error of prediction; RER, Range error ratio; rms, Root mean square; RMSEC, Root mean square error of calibration; RMSECV, Root mean squared error in cross-validation; RMSEP, Root mean square error of prediction; RPD, Ratio of prediction to deviation; SEL, Selectivity; SEN, Sensitivity; ST, Smokeless tobacco; TSNAs, Tobacco specific N'-nitrosamines; UV, Ultraviolet.

* Corresponding author at: University Blida 1, P.O. Box 270 Blida, Algeria.
E-mail addresses: medfr73@gmail.com (M. Fekhar), ydaghbouche@yahoo.fr (Y. Daghbouche), bouzna@yahoo.fr (N. Bouzidi), elhattab@univ-blida.dz (M. El Hattab).

https://doi.org/10.1016/j.microc.2023.109127
Received 24 June 2023; Received in revised form 19 July 2023; Accepted 22 July 2023
Available online 25 July 2023

Nicotine is a toxic addictive agent accounting for 95% of alkaloid content in tobacco [2]. Nicotine itself is not carcinogenic, however, 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone and N′-nitrosonornicotine, which are nitrosated derivatives of nicotine and other minor alkaloid have been classified as group 1 human carcinogens by the International Agency for Research on Cancer (IARC) [3]. These two (2) compounds and at least thirty other carcinogens in ST products [4] have been linked to a wide range of adverse health issues, including cardiovascular diseases, organ weight loss, precancerous oral lesions and receding gums [5–7], and also have been associated to cancers of the oral cavity, lung, pancreas, oesophagus, kidney and bladder [7–9]; whereas, the Chemma has been confirmed to cause oral cavity/oropharynx cancer in a case-control study that compared between Chemma users and non-users of any of tobacco types [10].

For this reason, nicotine level is crucial in understanding the general health effects of using ST and it is, besides TSNAs, the most reported constituent in recent works [11–13]. The nicotine determination in ST products predominantly performed with gas chromatography (GC) coupled with mass spectrometry (MS) in selected ion-monitoring (SIM) mode [11,12] or coupled with flame ionization detection (FID) following the CDC method [14–16]. A novel analysis approach has succeeded to include liquid chromatography-tandem mass spectrometry (LC-ESI-MS/MS) in multiple reaction mode (MRM) [8,17]. However, these techniques are relatively laborious and expensive, time consuming [18], can use large quantities of toxic solvents and non-suitable for routine analyses. Beyond chromatographic techniques, spectrophotometry have not been investigated thoroughly for ST except in few papers for total alkaloids determination by colorimetry [19], NCT release behavior from Snus by UV [20], in the AOAC-International method [21] or for quantifying total nicotine in cigarettes [22]. This latter was a typical work using FT-MIR spectroscopy in transmission mode with a simple linear regression.

Recent studies have shown that near-infrared (NIR) spectroscopy associated with deep learning frameworks performs efficiently for quantifying nicotine levels and multiple chemical components and routine quality parameters in tobacco leaves [23,24], as well as qualitatively in classifying tobaccos according to cultivation regions [25]. Despite their superior performance, both deep and conventional machine learning methods generate predictive models with partial accuracy that were validated on raw samples. These models can still be influenced by uncontrollable environmental and biotic factors affecting the samples as well as any unanticipated additives or adulterants in a final commercial product.

Over the past two decades, ATR-FTIR spectroscopy coupled with chemometrics has emerged as a robust technique in discrimination and quantitation analyses. Plus to its rapidity, easiness, cost-effective and minimum requirement of reagent and sample quantities; it offered results comparable to those of the official methods in quantification of several analytes, such as 5-(hydroxymethyl)furfural in honey [26], coumarin and ethyl vanillin adulterations in pure vanilla extracts [27], Penicillin V and phenoxyacetic acid during Pencillium chrysogenum fermentations [28], and in the detection and quantification of adulteration in milk [29]. In all of the abovementioned cases, MIR spectroscopy has successfully extracted the information from very complex matrices, i.e. natural products, directly by means of simple mathematical multivariate analyses, specifically PLS. On the other hand, and even if it was able to determine a single-component in a matrix, the univariate analysis lacked to the objective validation methods for real samples as it depends on the use of only one variable and the visual assessment of spectra, a thing that becomes more difficult with a large array of samples. On the contrary, the PLS method, depending on the use of multiple variables simultaneously, can evaluate the spectra quantitatively and qualitatively taking advantage of advanced mathematics to provide more reliable quantification or at least detect interferences.

Yet, if the analyte exists in trace amounts and no vibrational bands can be distinguished from the spectra, in such instance, a simple physical separation, chemical extraction and pre-concentration or changing the sampling approach (e.g.: thin dry film technique [30]) can enhance the sensitivity and the selectivity of the analyte.

With the above in mind, the aim of this paper was to develop a simple, quick, highly specific-to-analyte, more efficient method in terms of cost, reagent saving, sampling and suitability as a routine analysis method for accurate determination of total nicotine in commercial Chemma samples using, for the first time, attenuated total reflectance Fourier transform infrared spectroscopy in conjunction with two regression methods (univariate and multivariate).

Questions about the different pre-treatment strategies to improve the model quality, accuracy and precision of the chemometric models and validation of the prediction results were also addressed in this study.

2. Materials and methods

2.1. Standards and chemicals

(−)-Nicotine (≥99%) was of GC-grade and purchased from Sigma-Aldrich (China). Chloroform stabilized with ethanol (≥99%) and 2-propanol (99.8%) were of HPLC-grade from Sigma-Aldrich (France). Sodium hydroxide (NaOH, ≥98%), sodium carbonate anhydrous (Na_2CO_3, 99.5%) and sodium sulfate anhydrous (Na_2SO_4, ≥99%) were of Analytical Reagent grade and purchased from Biochem Chemopharma (France) or Panreac (Spain).

2.2. Smokeless tobacco samples

A convenience sample of seventeen top-selling Algerian ST products were purchased locally from wholesale and retail tobacco stores from six (6) different locations within two provinces (Medea and Blida) located in the north-central of the country between March 2021 and March 2022. These consisted of duplicate samples of one (1) reference product donated by United Tobacco Company QC central laboratory (UTC of MADAR Holding group, Boumerdes), two (2) available commercially certified products of the same manufacturer, their two (2) counterfeited products, ten manufactured illegally and illicit products, one (1) hand-made traditional Chemma and one (1) flavored non-tobacco (mixture of plant materials and species) product destined to aid quit dipping. Samples were chosen to reflect approximately 80% of the market share in Algeria at that time.

Nicotiana rustica species tobacco leaves from six (6) different regions in: Ain Melila, Batna, Biskra and Oued Souf provinces, were provided by QC central laboratory of the UTC. One (1) unidentified tobacco leaf (cultivated in Jijel region) used in making of traditional Chemma was obtained from popular street market. Pure blend cigarettes which include Burley, Oriental and Virginia strips were provided by Algerian-Emirati Tobacco Society (STAEM, Tipaza) were used as references to validate the results of this study.

The content of two (2) sachets of the same product were mixed and homogenized using a coffee grinder, returned to their original packaging, labeled, sealed in plastic bags then stored in freezer at −10 °C until analysis.

2.3. Samples preparation

Acid-base extraction of nicotine is based on the alkaloid property of nicotine and its solubility in different solvents. In this study, extracts were prepared according to the methods described in [31] and [22] with some modifications.

2 g of commercial Chemma product (dried in the dark for 3 days at room temperature) were weighed into a glass vial. 30 ml of distilled water were added, and the capped vial was sonicated for 20 min in ultrasonic bath. Sonication process will generate heat, so the temperature will reach easily 70 °C. Next, a 0.4 g of anhydrous sodium carbonate was appended, and the mixture was shaken then returned for another 10 min into the bath

to repose. After a double filtration and adjustment to pH 12 with sodium hydroxide (1 M), the aqueous filtrate was vortexed twice with 4 ml each of chloroform for 2 min then the decanted extracts were pooled in a glass centrifuge tube and centrifuged at 3000 rpm for 10 min. The obtained chloroformic phase was transferred carefully with a Pasteur pipette, filtering over 0.5 g of anhydrous sodium sulfate to retain water, and subsequently concentrated on a water bath under vacuum at 35 °C to dryness. The crude extract volume was adjusted with chloroform to 3 ml prior to analysis.

To prepare tobacco leaves extracts, only 1 g of material was submitted to the same procedure then diluted to volumes ranged between 1 and 5 ml of chloroform depending on the extraction yields determined gravimetrically.

2.4. Spectra acquisition

ATR-FTIR measurements were carried out using a Nicolet iS10 spectrometer equipped with a single-bounce attenuated total reflectance (Smart iTR accessory) diamond crystal, an XT-KBr beam splitter and a deuterated triglycine sulfate (DTGS-KBr) detector controlled by the OMNIC™ software version 9.8 (Thermo Fisher Scientific, USA). Each spectrum was recorded over the 4000–525 cm^{-1} spectral range through an accumulation of 32 co-added scans and a spectral resolution of 4 cm^{-1} (data spacing of 0.482 cm^{-1}) in absorbance mode.

The Smart iTR has a short pathlength due to the single-bounce ATR design. This feature makes the accessory well suited for quantitative measurements of thin dry films [32].

A 0.4 μl of sample, measured using a Hamilton micro syringe with adjustable stop rod, were mounted on the crystal and subsequently left to dry in ambient air (at temperature 25 ± 3 °C and 45–60% of humidity). After evaporation of solvent, within exactly 1 min, the obtained thin dry film was capped with a concave tip of the standard pressure tower enabling stabilization of nicotine during acquisition time (all controlled by real-time monitoring of standards). Before each measurement, the crystal was cleaned, twice, using a soft paper tissue soaked with ethanol 96° then with isopropanol.

The background interferogram was recorded once for each triplicate against evaporated chloroform under identical conditions after this subtracted automatically by the software. Fifteen replicates were obtained for each sample so that the total number was about 700 spectra.

2.5. Data analysis and software

A univariate analysis is a quantitative technique based on Beer's Law allowing the least squares regression using any of the peak height or area of component of interest which must contain minimum absorptions from other components in the sample.

As opposed to univariate approach, mathematical multivariate methods can detect irregular anomalies in the selected spectral regions through the use of a large number of variables, enabling more reliable quantification in complex matrices. For only one component at a time, PLS-1 algorithm performs to find the best correlation between input data matrix of X-variables (absorbances), decomposed into scores (t) and loadings (p), and vector y (concentrations) [33–35]. This allows the mathematical assessment of the weights (w) and the regression coefficients (b), and thus the establishing of a calibration model which can be used later to quantify the analyte:

$$\hat{y} = bX_{New} + e \tag{1}$$

where ŷ is the predicted concentration for the new sample and e are the X-residuals.

First of all, preliminary principal component analysis (PCA) of the standard normal variate (SNV) corrected for baseline and mean centered spectral data in the fingerprint region was conducted using the singular value decomposition (SVD) algorithm with 7 principal components (PCs) to detect and remove outliers using the scores plot and Q-residuals/Hotelling's T^2 values exceeding the 5–25% significance levels [36], then the regression models were built using the remaining spectra divided orderly into calibration and test sets at the ratio of 2:1.

At the regression step, seven (7) bands at 716, 807, 903, 1025, 1189, 1315 and 1428 cm^{-1} were chosen based on their intensity values. Their associated heights or areas corrected for baseline or not were employed in the univariate calibration, whereas, the corresponding pre-processed reduced regions were used in the PLS calibration.

Appropriate pre-treatment methods, including baseline offset and linear baseline corrections (BO and LBC), multiplicative scatter correction (MSC) with its extended options (EMSC), Savitzky-Golay smoothing (SGS) and first- and second-order derivatives (SG 1D' and SG 2D') at 11 points on each side with second polynomial order, and de-trending (Detrend) with second polynomial order; also their various combinations were conceded in the purpose to reduce baseline drifts, light scattering effects, nonlinearity, random errors and noise, overlapping peaks and other uncontrolled external factors.

The models were internally validated using the leave-one-out cross validation method, tested to predict concentrations in an external sample set, and then the performance of each model was checked using: slope, offset, RMSE and R^2 (or R) of calibration, cross-validation and prediction. Slope, R^2 and R values should be close to 1, offset and error values should be close to 0, also the difference between the values should be minimal. Moreover, additional statistical parameters (REP, Bias, RPD and RER) and analytical figures of merit (SEN, Anal. SEN, SEL, LOD and LOQ) were calculated for the best-performing models. In general, when Bias is close to 0, SEL is close to 1, and RPD, RER, SEN, Anal. SEN are much big and REP, LOD, LOQ are minimal, the developed model can be considered to have high accuracy, precision and sensitivity.

RMSE, REP, Bias, RPD and RER were calculated using the following equations [37–40]:

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \tag{2}$$

$$\mathrm{REP} = \frac{\mathrm{RMSEP}}{\bar{y}_{Cal}} 100 \tag{3}$$

$$\mathrm{Bias} = (\bar{\hat{y}} - \bar{y})_{test} \tag{4}$$

$$\mathrm{RPD} = \frac{SD_{y,\,test}}{\mathrm{RMSEP}} \tag{5}$$

$$\mathrm{RER} = \frac{(y_{max} - y_{min})_{test}}{\mathrm{RMSEP}} \tag{6}$$

where n is the number of samples, y_i and $\hat{y}_i$ are the nominal and predicted values for the i^{th} sample, $\bar{y}$ and $\bar{\hat{y}}$ are their corresponding average values, y_{max} and y_{min} are the maximum and minimum values, respectively, and SD_y is the standard deviation of the nominal values in the dataset.

SEN, Anal. SEN, SEL, LOD and LOQ theoretical demonstrations can be found elsewhere [37,41–43].

All data preprocessing and models construction for the univariate analysis were conducted using TQ Analyst 9.7 software (Thermo Fisher Scientific Inc., USA). Multivariate spectral preprocessing, PCA and PLSR were performed with The Unscrambler® X 10.4 (Camo Software AS., Norway). Analytical figures of merit and other statistics were calculated using the UNIVAR and the MVC1 toolboxes v. 2018 [37] and using home-made functions executed in MATLAB R2012a (Mathworks Inc., USA).

3. Results and discussion

3.1. ATR-FTIR spectra and exploratory analysis

The full spectrum of nicotine showed different characteristic absorption bands, their proposed assignments as described in [44–48] are as follow. Weak bands at 3403 (broad) and 1640 cm^{-1}, are due respectively, to the O − H stretching and bending vibrations produced by residual water. The peaks in the region 3000–2725 cm^{-1} are represented to the C − H asymmetric and symmetric stretching of pyrrolidinic CH_2 and pyrrolidinic methylamino CH_3 groups and to other various combined bands. The fingerprint region occurs between 1750 and 600 cm^{-1} and contains about thirty significant signals, the first at 1689 cm^{-1} can be associated with the pyrrolidinic protonated amine group (NH^+) disappeared partially after NaOH adding.

Fig. 1 shows the ATR-FTIR dry film average spectra of nicotine dissolved in chloroform and nicotine submitted to the proposed procedure at the same concentration level over the wavelengths from 600 to 1600 cm^{-1}. Characteristic peaks centered at 1590 and 1577 cm^{-1} probably correlated with C − N, C − C, C = N and/or C = C stretching vibrations of pyridine ring. The major peaks at 1478, 1456, 1447, 1428 and 1417 cm^{-1} are corresponding to asymmetrical/symmetrical bending of pyrrolidinic CH_2 and terminal CH_3 (1447 cm^{-1}), and might be overlapped with H − C − H bending of pyrrolidine at 1478 cm^{-1} and with C − H bending of pyridine at 1428 cm^{-1}. Absorption bands at 1362 and 1343 cm^{-1} are related to asymmetric wagging in CH_2 group of pyrrolidine ring or to C − N stretching of aromatic tertiary amines. A distinct peak at 1315 cm^{-1} corresponded to pyrrolidinic methyl C − N − C wagging and another at 1289 cm^{-1} is for symmetrical wagging of CH_2 group in the same chain. The bands in the range between 1275 and 1167 cm^{-1} are mostly assigned to asymmetric and symmetric torsion vibrations of CH_2 of pyrrolidine ring, while that at 1116 cm^{-1} was resulted from C − N − C asymmetric stretching. Additionally, bands at the wavelength of 1154 and 1086 cm^{-1} are associated with H − C − C − H deformation and C − H symmetric deformation of pyridine ring, respectively. Two dominant peaks, the first at 1045 cm^{-1} is due the pyridinic C − H asymmetric deformation, C = C − C and C = N − C deformations, and the second at 1025 cm^{-1} can be attributed to pyridine and pyrrolidine chains breathing or pyridinic C − N stretching. The band at 971 cm^{-1} is owing to C − H asymmetrical wagging of pyridine and those emerged at 922, 903, 807 cm^{-1} are linked to symmetrical and asymmetrical wagging of CH_3 and CH_2 groups. The most intense peak in the fingerprint region located at 716 cm^{-1} assigned to the out of plane C − H bending of the monosubstituted pyridinic ring, and lastly the absorption band at 616 cm^{-1} attributed to the in-plane C − N − C stretching.

As it can be seen in Fig. 1, the analytical blank does not contain any interference or contamination coming from the protocol, as a result, the spectrum of treated nicotine presents exactly the same characteristic bands of pure nicotine excepting some noise (rms noise = 1.6 10^{-4} AU in the full fingerprint region) and baseline drift, confirming the need only to baseline correction for appropriate quantitative analyses. Therefore, spectra of the treated standards are employed later for model development and quantification as they better represent the nicotine extracted from the real samples.

PCA was utilized as indicated in section 2.5 as an exploratory tool to exclude replicates suspected not belonging to the population of interest. An example for the obtained Q-residuals/Hotelling's T^2 plot and the scores plot with Hotelling's T^2 limit are shown in Supplementary Fig. S1 And S2, respectively. The first and the second principal components explained >80% of the total variance in all cases. In that example, six (6) points were identified from the Q-residuals/Hotelling's T^2 plot at a critical limit of 25% and one (1) point from the scores plot. The use of such high significance level of probability was because there are practically no differences between replicates. Other samples were eliminated from the PLS-scores plot, after the regression was done, keeping only five (5) points at each concentration level.

3.2. Calibration models optimization

The best analytical models for both univariate and multivariate methods was evaluated by selecting the adequate spectral measurement and spectral pre-treatment of the nicotine standards, directly dissolved

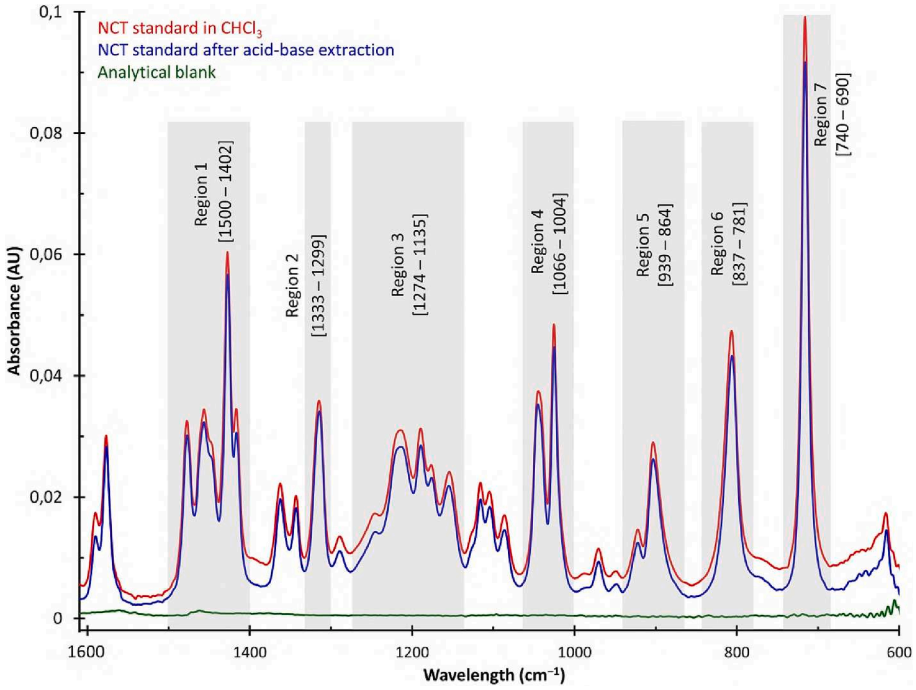


Fig. 1. FTIR-ATR spectra of the nicotine standards (red and blue) at concentrations of 8 $mg.ml^{-1}$ together with the analytical blank spectrum (green). Light grey marked the advantageous spectral ranges for NCT analysis. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

in chloroform and standards extracted following the proposed procedure, relying on trial and error approach strategy. Essentially, MSC was applied on replicates of each concentration stage separately, combined with other pre-processing techniques to remove scatter effects and random variances due to the deposited droplet volume. Since this latter normalizes based on the mean spectrum in a data set, it is preferred instead of SNV which rescale each spectrum roughly from −2 to +2 [49] obtaining the same intensities for correlated spectra. The MSC or EMSC can improve model quality just as the spectral averaging technique [26] leading to a higher quality calibration models.

Supplementary Tables S1 and S2 show the full regression parameters and the followed strategies in univariate and multivariate analyses, respectively. The best models were selected by comparing slopes, offsets, root mean square errors, determination coefficients or correlation coefficients values in calibration and validation sets.

Regarding to the univariate method, TQ Analyst software has an advantage that estimating the RMSE similarly to the approach used in multivariate algorithms; as well a unique distinct parameter so-called performance index (PI) which measures, in percent, the accuracy of a developed model to quantify the test samples [50]. All the selected models have a PI value above 90%. The models corresponding to 716, 903, 1025, 1190 and 1315 cm^{-1} peak heights offered RMSE below 0.500 and coefficients of correlation above 0.9960. Despite of the high values of errors, these models could be accepted for such technique of sampling (thin dry film) at such large concentration range. Fig. 2A displays a visual comparison using the radar chart, inspired from [51] and [52], of the most relevant results (heights at 716 and 1315 cm^{-1}). This result is consistent with that reported in previous study [22] at the same band (1315 cm^{-1}).

On the other hand, since only a single component was considered in the multivariate calibration, thus only one factor was sufficient to describe the data in all resulting PLS models, better characterized then by the used number of variables than latent variables. Overall, when weighting against univariate models, PLS models provided clearly higher quality parameters (Fig. 2B) due to the advanced spectral pretreatments applied on each reduced region minimizing error values by 2 to 3 times, while the best R^2 reached 0.9999.

From all the developed PLS models, that at region 1 (1500–1402 cm^{-1}) offered the worst errors as it contains uneven noise came from the protocol and hence it was excluded. Region 3 (1274–1135 cm^{-1}) includes multiple overlapping bands that may not be differentiated later from other interference agents at the prediction step, so it was also excluded. Although regions 4 (1066–1004 cm^{-1}) and 6 (837–781 cm^{-1}) presented good quality parameters, unfortunately, they assigned to the very common aromatic rings breathing and to symmetrical/asymmetrical wagging of CH_3 and CH_2 groups respectively, so it seem not to be an excellent candidates to predict nicotine in matrix. Region 5 (939–864 cm^{-1}) yielded very good parameters but it has relatively lower intensity comparing with the other bands and could reduce significantly the sensitivity of the method. Subsequently, it remains only the models from regions 2 (1333–1299 cm^{-1}), 7 (740–690 cm^{-1}) and their combinations to be well studied; both regions promoted to be very specific to nicotine. The arrangement of Detrend-BO-SG 2D'-MSC in order, provided the best RMSEP value (0.0865) in region 7, however, the fact that this correction can reduce sensitivity, the number of variables to be taken and conduct to non-correlated values of calibration and prediction errors impose the satisfactoriness with less complicated pre-processing techniques. As a result, regions 2 and 7 were combined together and pre-processed following various pre-processing techniques, whereas the LBC-BO-EMSC-SGS exhibited a slope equals to 0.999, offset equals to 0.006, RMSEC of 0.147 and R^2 of 0.9991, guarantees a good fit of the calibration line to the spectral data, and consequently it were evaluated for further tests.

Pre-processed spectral ranges 2 + 7 and their corresponding calibration lines are visualized in Fig. 3A-D. From the original data regression lines, notice that the variance decreases upon dilution which corroborate the precision of the sampling and the analysis step, and from the pre-processed data, the closeness of regression lines of the calibration and test sets to each other illustrates the high predictive ability of the model.

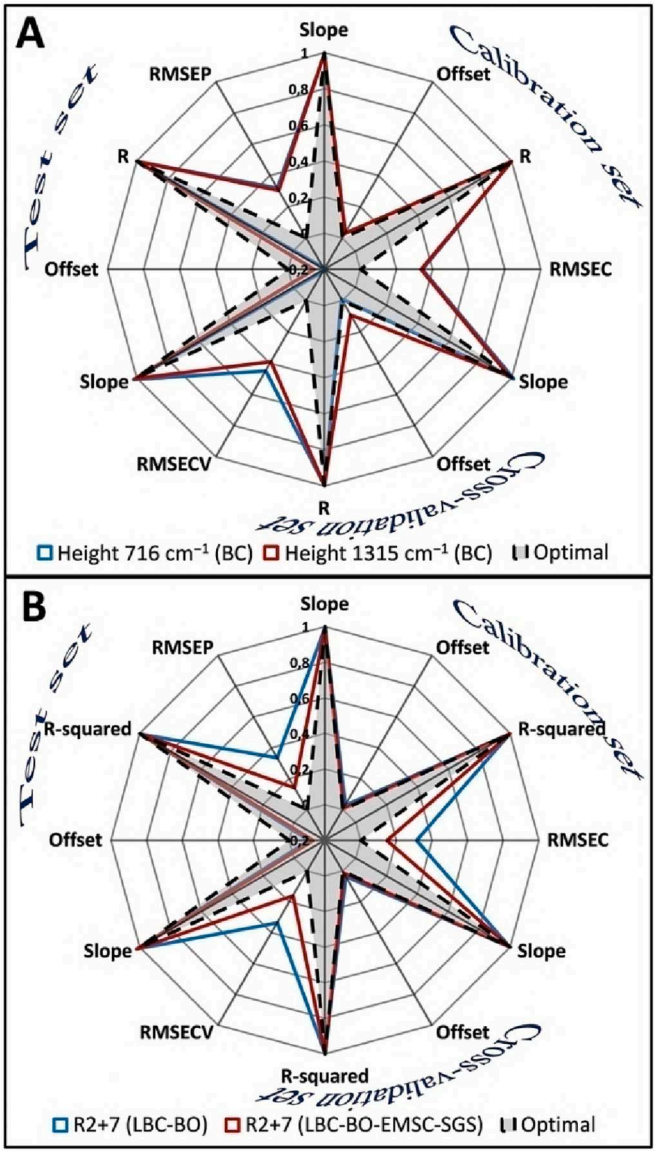


Fig. 2. Radial diagrams for comparison of the optimized univariate (A) and PLS (B) models performance. Optimal pre-established values of Offset and errors RMSE are 0, and of Slope and R-squared (or R) are to be 1. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.).

3.3. Evaluation of the optimal models performance

Before use any of the given models, these were evaluated for additional statistical parameters and figures of merit estimated according to the latest approaches and guidelines proposed in the literature [37,41,42]. Table 1 presents critical statistics assessed during validation for the optimized models for both univariate and multivariate methods.

Starting with corroborating the linearity over the dynamic range, this involves calculating the experimental *F*-value defined as the squared ratio of residual standard deviation to pure error then comparing it with the tabulated value at a significance level α [37]. All the selected models yielded $F_{exp} < F_{critical}$, reconfirming the linearity of the calibration lines. A related parameter, the Haaland's criterion [53], searches for a

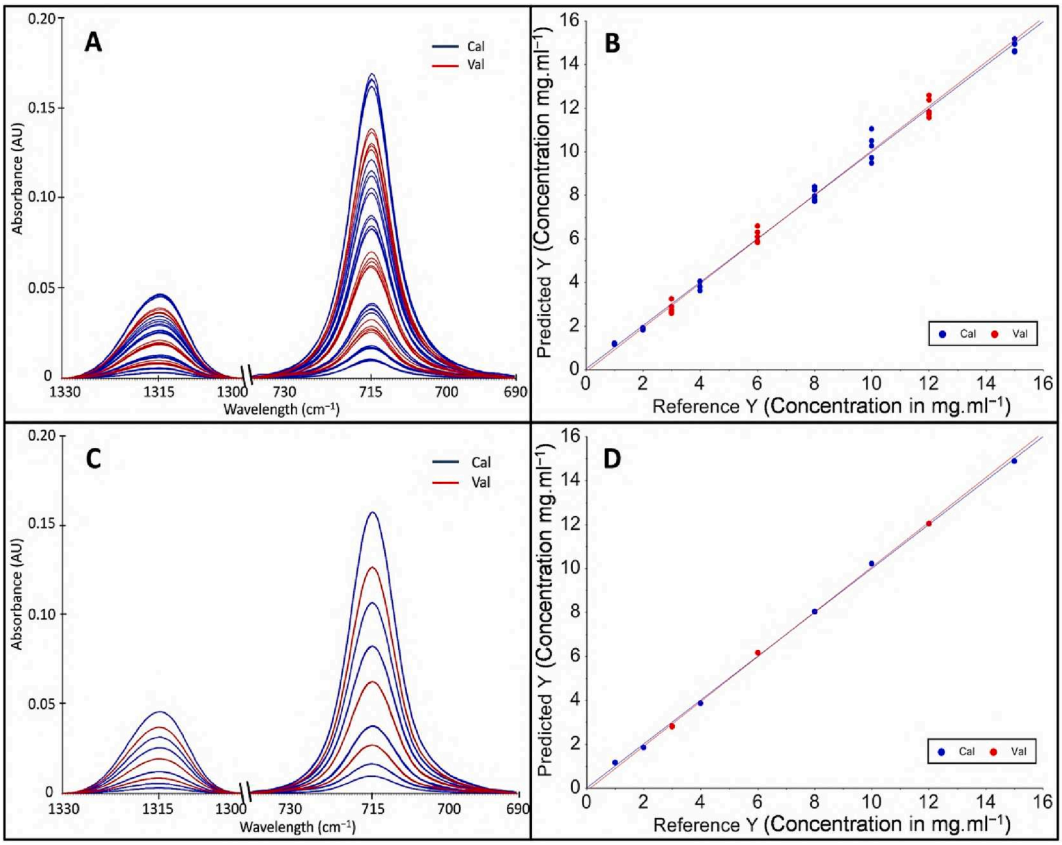


Fig. 3. LBC-BO (A) and LBC-BO-EMSC-SGS (C) pre-processed ranges 2 + 7 and their corresponding predicted vs. actual NCT concentrations (B) and (D) respectively, calculated with PLS-1 algorithm for calibration and validation sets.

probability level $p < 0.75$ for the F ratio. This test was conducted to validate that the models required only one (1) latent variable to predict the unique component in matrix even after its submission to extraction procedure.

The predicted residual error sum of squares (PRESS) of samples that are not used for estimating calibration parameters was carried out. Generally, when the PRESS value is minimal, the predictive ability is higher. The Q^2, known as the cross-validated R^2, is interpreted as the proportion of the variability of samples that are left out during the leave-one-out cross-validation step. It is defined by [54]:

$$Q^2 = 1 - \frac{\text{PRESS}}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (7)$$

where y_i and $\bar{y}$ are the nominal and mean concentration values in the calibration dataset, respectively. A large difference between the Q^2 and the R^2 shows that the model is sensitive to the presence or absence of certain samples [54]. The best values of PRESS (0.75) and Q^2 (0.9990) were obtained for the pre-processed regions 2 + 7 (LBC-BO-EMSC-SGS), showing a significantly higher predictive capacity than that of the simple Beer's law regression.

In classical univariate calibration, the sensitivity is the slope of the calibration graph [43], whereas in PLS, it corresponds to the inverse of length of the vector of regression coefficients [42]. Moreover, the ratio between sensitivities of the analyte in matrix and in its pure form known as selectivity [41,55]; this latter is not reported usually in multivariate methods as no approximation is available to pure analyte spectra in mixture [41]. However, it was calculated here to prove that no contamination was occurred, at a specific spectral measurement, after submission the standards to the recommended procedure of extraction. The sensitivities in PLS models found to be 4 folds, and greater, higher than in univariate ones. Even higher analytical sensitivity values were calculated for multivariate method attaining 131 ml.mg^{-1} due to the minimal noise at these regions. Selectivity values were not considerably different from one (1) in both types of regression techniques confirming the qualitative suitability of the set of standards for nicotine quantification.

A very important figure of merit to be reported is the limit of detection (LOD), recognized as the minimum detectable concentration of an analyte relating a certain risk of false detects and false non-detects (α- and β-errors) [42,56]. Regarding to PLS, the LOD has two limits; minimum and maximum, related to samples situated near to the center and at the edges of the multivariate calibration space, respectively [56]. The limit of quantification (LOQ), in turn, is estimated as the concentration level for which is 3 times the associated LOD [41]. Detection limits in the order of 1.2 and 1.3 mg.ml^{-1} were achieved in univariate models, and 0.15–0.16 mg.ml^{-1} for LBC-BO pre-processing and 0.33–0.35 mg.ml^{-1} for LBC-BO-EMSC-SGS pre-processed PLS models.

Finally, from the statistics obtained at the prediction step, the values of Root Mean Square Error of Prediction (RMSEP, 0.31 mg.ml^{-1}), Relative Error of Prediction (REP, 4.6%), Ratio of Prediction to Deviation (RPD, 12.7) and Range Error Ratio (RER, 29.5) at the height 1315 cm^{-1} were found for univariate models. In the case of PLS models, the best values of RMSEP, REP, RPD and RER equal to 0.14 mg.ml^{-1}, 2.1%, 27.1 and 63.1 respectively, were achieved for LBC-BO-EMSC-SGS spectral pre-treatment of the regions 2 + 7. This indicates an excellent ability (RPD > 8.1 [38] and RER > 20 [57]) of the corresponding model to predict the concentration of NCT under the proposed conditions in the test set unless if it is not biased.

3.4. Accuracy and precision evaluation

The trueness and precision of the best-performing models were inferred using bias, recovery values and the elliptical joint confidence

Table 1
Statistical parameters and figures of merit for the optimal univariate and multivariate models.

Method	Beer's law		PLS-1	
Spectral measurement	Height 716 cm^{-1}	Height 1315 cm^{-1}	Regions 2+7	Regions 2+7
Pre-treatment	BC: 740-690 cm^{-1}	BC: 1333-1299 cm^{-1}	LBC-BO	LBC-BO-EMSC-SGS
Calibration set (N=30)				
Linear concentration range ($mg.ml^{-1}$)	1.0–15.0			
PRESS	4.0	3.9	3.4	0.75
Q^2	0.9944	0.9946	0.9953	0.9990
F-test (F_{exp}) [$F_{crit\ (0.05,\ 28,\ 24)} = 1.95$]	1.10	1.23	1.00	1.24
p-value	0.41	0.30	0.50	0.72
Sensitivity (AU.ml. mg^{-1})	0.0114	0.00314	0.0456	0.0452
Anal. SEN ($ml.mg^{-1}$)	2.81	2.87	103	131
SEL	1.02	1.09	0.986	0.981
LOD[¶] ($mg.ml^{-1}$)	1.3	1.2	0.33–0.35	0.15–0.16
LOQ[§] ($mg.ml^{-1}$)	3.7	3.6	0.98–1.0	0.46–0.47
Test set (n=15)				
Levels	3.0 6.0 12.0			
RMSEP ($mg.ml^{-1}$)	0.32	0.31	0.34	0.14
REP (%)	4.8	4.6	5.0	2.1
RPD	12.1	12.7	11.5	27.1
RER	28.0	29.5	26.8	63.1
Bias ($mg.ml^{-1}$)	−0.063	−0.073	$-1.5\ 10^{-3}$	$-3.0\ 10^{-4}$
Mean recovery ± SD (%)	98.1 ± 5.6	97.8 ± 6.6	98.9 ± 6.8	98.8 ± 3.9
t-test (t_{exp}) [$t_{crit\ (0.025,\ 14)} = 2.15$]	1.34	1.31	0.609	1.15

[¶] The univariate limit of detection (LOD_u) and the multivariate limits of detection [LOD_{min} - LOD_{max}].

[§] The univariate limit of quantitation (LOQ_u) and the multivariate limits of quantitation [LOQ_{min} - LOQ_{max}].

Abbreviation: R2+7, Combined spectral ranges: 1333-1299 + 740-690 cm^{-1}.

region (EJCR) test for the test set.

The average value of the difference between predicted and nominal concentrations for the test samples (Bias), revealed the following values −0.073, −0.063, −0.0015 and −0.0003 $mg.ml^{-1}$, again in favor of multivariate models. These small values suggest a random distribution of points around the calibration line for all four models.

The analyte recovery, usually evaluated in standard addition method [51], nevertheless, it was estimated here for the treated test samples predicted from the calibration line as the ratio between predicted and actual concentrations. For all models, the mean recoveries do not differ statistically from 100% since the calculated t_{exp} is smaller than the tabulated value; in line with the obtained selectivity values in section 3.3. These findings justified, once again, the appropriate choice of the set of treated standards to establish the calibration lines for the purpose of quantifying nicotine in the commercial samples.

Now, in order to check properly the accuracy of the proposed models and assess the presence of constant or proportional bias, the EJCR for the slope and intercept of the regression predicted versus nominal concentrations for the test samples is considered. If the ideal point (slope = 1, intercept = 0) lies inside the confidence region given at a chosen confidence level α, then bias is absent [58–61]. Fig. 4A illustrates the EJCR test conducted to the best-performing models based on ordinary least-squares (OLS). Notice that, the theoretical point is located inside the confidence regions for all models supporting the absence of systematic or bias errors. Nonetheless, the models corrected only for baseline show wider ellipses putting the precision of the sampling technique on the line. Also, the height 1315 cm^{-1} corrected for baseline showed a small shift to the left, away from the ideal point; revealing the efficiency of multivariate pre-processing techniques even in baseline correction. On the other hand, the pre-processed PLS model with EMSC show a

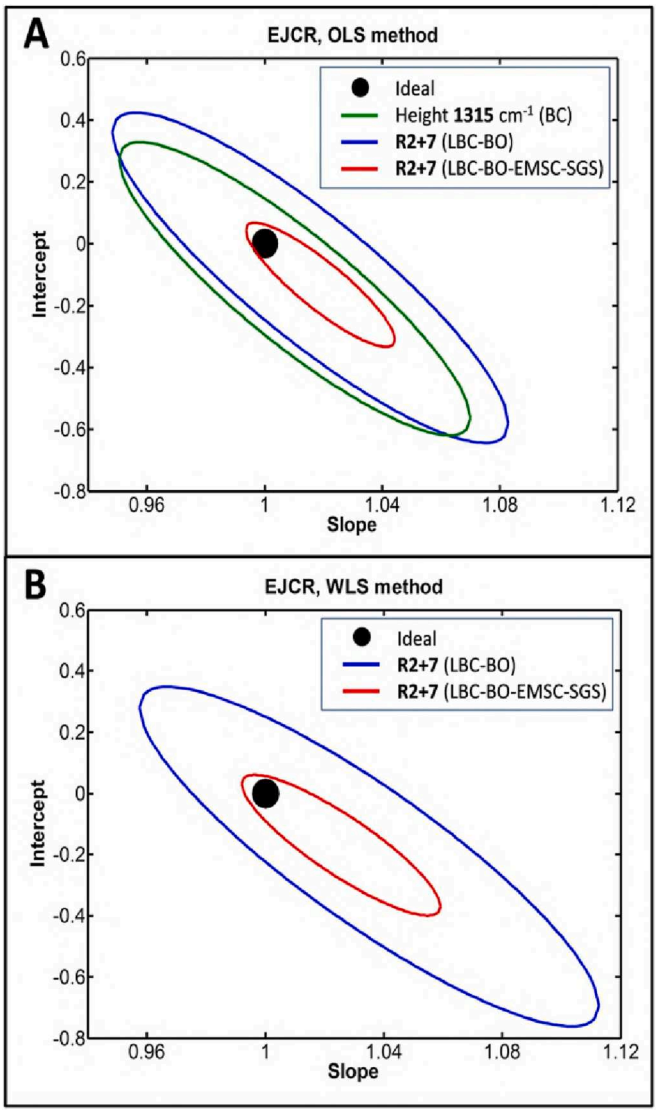


Fig. 4. EJCR in the slope-intercept plane performed for the best models based on ordinary least-squares (A) and weighted least-squares (B). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.).

narrower confidence region confirming the precision, whereas the ideal point lies near to the focus of the ellipse; this keeps unanswered question about the accuracy of the obtained model.

To gain further insight into the accuracy of the developed PLS models, the EJCR test using the weighted least-squares (WLS) was performed. This latter usually done when non-constant variance can be assumed [59,61]. For that, a new methodology has been investigated taking into account the deviation values estimated by The Unscrambler software for weighting each predicted concentration of the test samples individually, allowing more robustness since the mathematical pre-processing techniques can lead to misestimating of standard deviation values (see section 3.5 below). Fig. 4B displays the new EJCR based on WLS regressions. Again, the ideal point was inside but not in the center of the confidence regions which were slightly wider but not different significantly from the ones calculated by OLS; such observation could be explained by the high uncertainty coming from the small volume of sample droplet, a result that can be accepted for such technique of sampling and analysis. In conclusion, the use of EJCR for the slope and intercept, both for OLS and WLS, leads to that no bias exists, and thus the recovery can be considered as 100% for the chosen models.

3.5. Analysis of commercial samples

In this part, the validated model used the corrected height at 1315 cm^{-1} (univariate) and the one used the pre-processed regions 2 + 7 with LBC-BO-EMSC-SGS (PLS-1) were applied to determine the content of nicotine in seventeen commercial samples of ST and ten varieties of tobacco leaves treated in the same way as the calibration standards set and their corresponding spectra. Fig. 5 shows a comparison between ATR-FTIR spectra of tobacco-containing products (from sample 1 to sample 16) and the non-tobacco product (sample 17) extracts. As it can be seen from the spectra, the recommended procedure shows a high specificity toward nicotine by isolating it from all tobacco products, demonstrating the high quality of analyte determination that can be obtained at the prediction step. A detailed description of total nicotine content in mg per g of dry weight of sample is reported in Table 2 following the CDC guidelines for reporting of the quantity of NCT in ST products [16,62].

In contrast to the univariate method, whose role ends when predicting the analyte concentration, the multivariate method, besides to the quantitative analysis, provides advanced statistical tests that can estimate how accurate the prediction is; among these tests we exploited the spectrum fit check, deviation values, the Q-residuals/Hotelling's T^2 statistics and variable contributions to model and to residuals.

Starting with the spectrum fit or measurement region check, calculated by the TQ Analyst software, this feature assigned to compare the specified spectral regions of each quantified sample with the spectra of the calibration standards qualitatively to determine their degree of similarity, expressed as a fit value ranges from 0 (no similarity) to 100 (perfect match) [50]. All the commercial tobacco-containing products show excellent correlations between their own and the calibration spectra (above 97.4%), while for the non-tobacco product (13.4%), correctly discriminating between containing and non-containing tobacco samples. For the pure tobacco blends, cigarette strips present better fit values (above 97.3%) than ST leaves (42.9 to 98.9%), this finding might be because of that some blends were not properly cured.

The second parameter, the deviation, it is a valuation of the prediction uncertainty for each individual unknown sample. It is estimated as a function of the global model error, the sample leverage and the sample residual X-variance without taking the actual concentration into account [49]. Despite the fact that the mathematical pre-processing methods improved the quality of the calibration data, they had a negative impact on the calculated standard deviations (found in the order of 10^{-4} to 10^{-3} mg/g, results not shown) which utterly differ from the reality, making the above-mentioned deviation the best substitute. Most of the deviation values were between 0.2 and 0.6 mg/g, and larger values were observed for higher concentrations, a sign of a good prediction quality; however, it should be noted that few of these values were not correlated with the associated spectral fit.

Another test, the Q-residuals and Hotelling's T^2 statistics for real samples describe the new sample distance to calibration model and how well the sample are explained by the model, respectively [49]. Fig. 6A shows the Q-residuals versus Hotelling's T^2 plot with a p-value of 0.05 as critical limit. On the abscissa axis, we can see that all samples have leverages below the critical limit indicating that they were perfectly modeled. The points lying to the right of the plot corresponded, from left to right, as expected to the traditional Chemma extract followed by the leaves from Berzili-Batna then by the non-tobacco product; the latter two have exhibited poor values of spectrum fit, and consequently they are accurately described in the plot as they were different from the other samples. On the ordinate axis, most of the points were above the critical limit. After rechecking the raw data for errors, it could be concluded that these samples were not different but just have extreme values of concentration, and as long as these samples are not influential, the high residual variance probably due to minor contributions from other components in the matrix or just modeling noise; an object that will be discussed intensely through the next parameter.

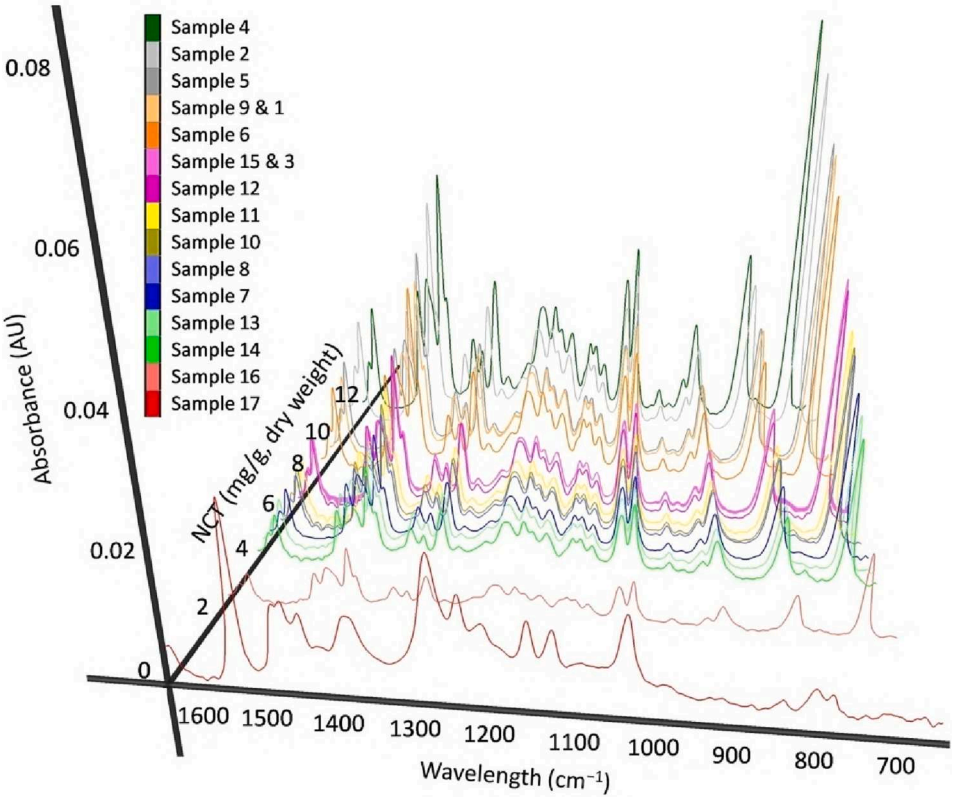


Fig. 5. FTIR-ATR spectra of commercial ST products extracts against total nicotine values. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.).

Table 2
Calculated nicotine content in mg/g of product in real samples performed by the proposed approach.

Method	Beer's law		PLS-1				
Spectral measurement	Height 1315 cm^{-1}		Regions 2+7				
Pre-treatment	BC: 1333-1299 cm^{-1}		LBC-BO-EMSC-SGS		Region fit (%)		
NCT content (mg/g dry weight)	Average[¶]	SD	Average[¶]	Mean deviation	Average[¶]	SD	
Commercial product							
Sample 1	8.5	0.4	8.7	0.2	99.46	0.27	
Sample 2	10.0	0.2	10.4	0.5	98.88	0.28	
Sample 3	6.0	0.3	6.4	0.3	99.06	0.24	
Sample 4	12.0	0.4	11.7	0.6	98.82	0.22	
Sample 5	9.3	0.1	9.0	0.5	98.36	0.21	
Sample 6	8.1	0.6	8.0	0.5	98.14	0.48	
Sample 7	4.5*	0.2	4.6	0.2	98.36	0.58	
Sample 8	5.0*	0.1	5.2	0.2	99.26	0.31	
Sample 9	8.3	0.3	8.7	0.4	98.96	0.27	
Sample 10	5.1*	0.2	5.3	0.2	99.60	0.12	
Sample 11	5.4*	0.5	5.6	0.2	99.50	0.25	
Sample 12	6.6	0.2	6.3	0.4	97.42	0.68	
Sample 13	4.0*	0.4	4.2	0.2	98.58	0.52	
Sample 14	3.7*	0.2	3.9	0.2	98.3	1.1	
Sample 15	6.1	0.4	6.4	0.3	98.64	0.50	
Sample 16 (traditional)	2.2*	0.2	2.4	0.2	97.46	0.89	
Sample 17 (non-tobacco)	< LOD	-	< LOD	-	13.4	4.0	
ST leaves							
Berzili-Ain Melila	5.2	0.3	5.9	0.4	95.5	2.3	
Berzili-Batna	< LOD	-	0.8	0.3	42.9	3.1	
Berzili-Biskra	24.9	0.7	25.3	1.4	98.52	0.37	
Chergui-Ain Melila	9.6	0.7	9.7	0.6	96.7	1.2	
Jijel	45.2	2.1	45.0	1.6	98.88	0.35	
Soufi-Oued Souf	18.1	0.8	18.0	1.1	97.82	0.42	
Zeribet El Oued-Biskra	16.1	0.2	16.4	1.0	98.02	0.33	
Pure blend cigarette							**Reference values**[§]
Burley strips	10.1	0.5	10.2	0.4	99.23	0.25	6.50–47.7 [63,64]
Oriental strips	9.2	0.3	9.3	0.6	97.3	1.2	1.80–12.6 [63]
Virginia strips	8.2	0.4	8.9	0.6	97.76	0.78	6.52–60.4 [63]

[¶] Values reported are the mean of five replicates from two independent measurements of sample.
[*] The computed concentration value was below the $LOQ_{univariate}$ for the considered model.
[§] Values reported in the literature.
Abbreviation: R2+7, Combined spectral ranges: 1333-1299 + 740-690 cm^{-1}.

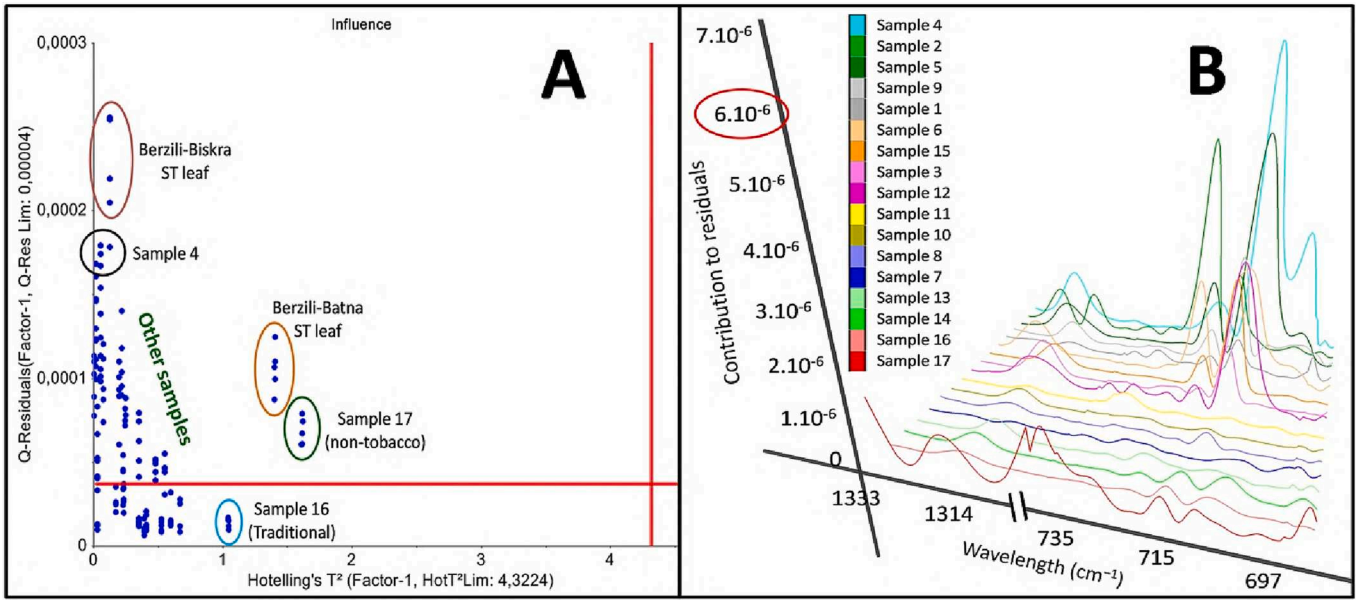


Fig. 6. (A) Q-residuals versus Hotelling's T^2 plot for real samples. Red lines are the associated critical limits with a p-value of 5%. (B) Mean variables contribution to residuals for each commercial sample. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Hotelling's T^2 and Q-Residual contributions, for a specified sample, describe how much each variable contributes to calibration model and to residuals, respectively [49]. When we look to an example in the Supplementary Material (Fig. S3), it is clear that both regions 2 and 7 contributed to model but unequally, where the region 2 shows better performance in predicting that sample. Usually, the smaller the

contribution value, the better the variable is described by the model. From the contributions to residuals plot (Fig. 6B), it can be seen multiple bands around the range of interest which may be related to peak shifts, spectral pre-treatments, instrumental noise, or to unknown interference agents in the matrix (perhaps other minor alkaloids); fortunately, all the samples exhibited contributions to residuals that were from 100 to 2000 times smaller than the corresponding contributions to model, and therefore had practically no influence on the prediction results. Furthermore, no frequent or consistent contribution to residuals between replicates was remarked reconfirming that these variances are due to non-important regions in the spectrum, i.e. instrumental noise. It should be noted that such observations would be impossible to be ascertained with the classical regression method and without chemometrics. Accordingly, it is very recommended that all single component calibrations should be carried out using a multivariate approach in for validating the developed method and the obtained results.

Using the PLS-ATR-FTIR optimal model, total nicotine content in the manufactured ST products have been found ranging from 3.9 to 12.6 mg/g of dry weight; whereas, in the hand-made traditional product, it was only 2.4 mg/g of product due to the high amount of additives among its ingredients. The flavored non-tobacco product was confirmed to contain no detectable NCT. For tobacco leaves, those used in the manufacturing of ST had concentrations varied from 5.9 to 45.0 mg/g, except of one atypical sample (0.8 mg/g), while the pure cigarette blends afforded concentrations between 8.9 and 10.2 mg/g leaf, in line with the reference values reported in literature, and validating the predictive ability of the developed method.

To conclude, the multivariate method using PLS-1 succeeded to provide the best analytical figures of merit and validation parameters, and consequently it was considered the more appropriate for detecting and quantifying nicotine in complex matrices; nevertheless, the univariate method using the corrected height 1315 cm^{-1} could be considered as a simple alternative for nicotine content evaluation with the aforementioned procedure.

4. Conclusions

In this paper, we have shown the feasibility of using a single-bounce ATR-FTIR technique coupled with chemometrics as a simple, low-cost and specific-to-analyte method for the quantitative analysis of nicotine in commercial Algerian ST “Chemma” for the first time.

The calibrated and validated method, for a wide analytical range, showed mainly acceptable outcomes and the obtained results indicate that the LBC and EMSC pre-processing strategies followed by PLS regression with single y-variable offered the best analytical figures of merit.

The novelty of this work being in the use of chemometrics for data analysis and to evaluate the target analyte in real sample spectra, and thus providing robust quantification results, a step that would have been impossible using the classical Beer’s law.

Moreover, the developed method proved a high chemical selectivity toward nicotine, so and with the increasing number of newer ST products containing complex flavoring agents, plant materials and/or adulterants, the ATR-FTIR spectroscopy is very promising not to be alternative for the existing solutions but as a complementary routine analysis method and as a sustainable way to quickly determine nicotine in various tobacco products, which can contribute to a better understanding of the health risks associated with the use of these products and aid in the development of harm reduction strategies.

In addition to variations across geographic region, harvest season, curing and storage conditions of tobaccos, nicotine levels vary mainly depending on the final product composition, which is often unknown and have been set by non-professionals in an unorganized market dominated by counterfeiting. Accordingly, governmental entities are required to prevent this type of illicit trade and to issue laws that impose manufacturers to report nicotine level and the levels of any other constituents that could alter the quality of these products and create a possible threat to the public health.

CRediT authorship contribution statement

Mohamed Fekhar: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Visualization, Writing – original draft, Writing – review & editing. **Yasmina Daghbouche:** Formal analysis, Investigation, Project administration, Supervision, Validation, Visualization, Writing – review & editing. **Naima Bouzidi:** Supervision, Validation, Visualization, Writing – review & editing. **Mohamed El Hattab:** Funding acquisition, Resources, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

Authors acknowledge the financial support of research project with economic and social impact (N° 007 of March 15, 2020) and University-Education Research Projects (PRFU/ B00L01UN090120220001).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.microc.2023.109127.

References

[1] Tobacco in Algeria. http://alger-roi.fr/Alger/documents_algeriens/economique/pages/63_tabac.htm. (Accessed 05 March 2023).
[2] H. Zhang, Y. Pang, Y. Luo, X. Li, H. Chen, S. Han, X. Jiang, F. Zhu, H. Hou, Q. Hu, Enantiomeric composition of nicotine in tobacco leaf, cigarette, smokeless tobacco, and e-liquid by normal phase high-performance liquid chromatography, Chirality 30 (7) (2018) 923–931, https://doi.org/10.1002/chir.22866.
[3] Smokeless Tobacco and Some Tobacco-Specific N-Nitrosamines, World Health Organization International Agency for Research on Cancer, Lyon, France, 2007, pp. 1-592. https://monographs.iarc.fr/ENG/recentpub/mono89.pdf (Accessed 06 January 2018).
[4] S.B. Stanfill, G.N. Connolly, L. Zhang, L.T. Jia, J.E. Henningfield, P. Richter, T. S. Lawler, O.A. Ayo-Yusuf, D.L. Ashley, C.H. Watson, Global surveillance of oral tobacco products: total nicotine, unionised nicotine and tobacco-specific N-nitrosamines, Tob. control 20 (3) (2010) 1–10, https://doi.org/10.1136/tc.2010.037465.
[5] I. Stepanov, L. Biener, A. Knezevich, A.L. Nyman, R. Bliss, J. Jensen, S.S. Hecht, D. K. Hatsukami, Monitoring tobacco-specific N-nitrosamines and nicotine in novel Marlboro and Camel smokeless tobacco products: findings from Round 1 of the New Product Watch, Nicotine & Tob. Res. 14 (3) (2012) 274–281, https://doi.org/10.1093/ntr/ntr209.
[6] H.A. Alhazmi, A. Khalid, S. Sultana, S.I. Abdelwahab, W. Ahsan, M.E. Oraiby, M. Al Bratty, Determination of phytocomponents of twenty-one varieties of smokeless tobacco using gas chromatography-mass spectroscopy (GC-MS), S. Afr. J. of Chem. 72 (2019) 47–54. https://doi.org/10.17159/0379-4350/2019/v72a7.
[7] T.S. Lawler, S.B. Stanfill, L. Zhang, D.L. Ashley, C.H. Watson, Chemical characterization of domestic oral tobacco products: total nicotine, pH, unprotonated nicotine and tobacco-specific N-nitrosamines, Food and chem. toxicol. 57 (2013) 380–386, https://doi.org/10.1016/j.fct.2013.03.011.
[8] S. Nasrin, G. Chen, C.J. Watson, P. Lazarus, Comparison of tobacco-specific nitrosamine levels in smokeless tobacco products: High levels in products from Bangladesh, PloS one 15 (5) (2020) e0233111.
[9] I. Stepanov, J. Jensen, L. Biener, R.L. Bliss, S.S. Hecht, D.K. Hatsukami, Increased pouch sizes and resulting changes in the amounts of nicotine and tobacco-specific N-nitrosamines in single pouches of Camel Snus and Marlboro Snus, Nicotine & Tob. Res. 14 (10) (2012) 1241–1245, https://doi.org/10.1093/ntr/ntr292.
[10] M. Oudjehih, I. Deltour, M.L. Bouhidel, A. Bouhidel, A. Marref, V. Luzon, J. Schüz, H. Bouneceur, M.E. Leon, Smokeless Tobacco Use, Cigarette Smoking, and Upper Aerodigestive Tract Cancers: A Case-Control Study in the Batna Region, Algeria,

2008–2011, Tob. use insights 13 (2020) 1–11, https://doi.org/10.1177/1179173X209022.
[11] T.S. Lawler, S.B. Stanfill, H.T. Tran, G.E. Lee, P.X. Chen, J.B. Kimbrell, J.G. Lisko, C. Fernandez, S.P. Caudill, B.R. deCastro, C.H. Watson, M. Cummings, Chemical analysis of snus products from the United States and northern Europe, PloS one 15 (1) (2020) e0227837.
[12] S.B. Stanfill, R.E. Croucher, P.C. Gupta, J.G. Lisko, T.S. Lawler, P. Kuklenyik, M. Dahiya, B. Duncan, J.B. Kimbrell, E.H. Peuchen, C.H. Watson, Chemical characterization of smokeless tobacco products from South Asia: Nicotine, unprotonated nicotine, tobacco-specific N'-Nitrosamines, and flavor compounds, Food and chem. toxicol. 118 (2018) 626–634, https://doi.org/10.1016/j.fct.2018.05.004.
[13] I. Stepanov, P.C. Gupta, M. Parascandola, K. Yershova, V. Jain, G. Dhumal, D. K. Hatsukami, Constituent variations in smokeless tobacco purchased in Mumbai, India, Tob. Regulatory Sci. 3 (3) (2017) 305–314. https://doi.org/10.18001/TRS.3.3.6.
[14] W. Rickert, P. Joza, A. Trivedi, R. Momin, W. Wagstaff, J. Lauterbach, Chemical and toxicological characterization of commercial smokeless tobacco products available on the Canadian market, Regulatory Toxicol. and Pharmacol. 53 (2) (2009) 121–133, https://doi.org/10.1016/j.yrtph.2008.12.004.
[15] A. McNeill, R. Bedi, S. Islam, M. Alkhatib, R. West, Levels of toxins in oral tobacco products in the UK, Tob. control 15 (1) (2006) 64–67, https://doi.org/10.1136/tc.2005.013011.
[16] Centers for Disease Control and Prevention, Protocol to measure the quantity of nicotine contained in smokeless tobacco products manufactured, imported, or packaged in the United States, Federal Register 62 (85) (1997) 24116–122419.
[17] H. Ji, Y. Wu, F. Fannin, L. Bush, Determination of tobacco alkaloid enantiomers using reversed phase UPLC/MS/MS, Heliyon 5 (5) (2019) e01719.
[18] P. Arena, F. Rigano, P. Guarnaccia, P. Dugo, L. Mondello, E. Trovato, Elucidation of the Lipid Composition of Hemp (Cannabis sativa L.) Products by Means of Gas Chromatography and Ultra-High Performance Liquid Chromatography Coupled to Mass Spectrometry Detection, Molecules 27 (10) (2022) 3358, https://doi.org/10.3390/molecules27103358.
[19] K. McAdam, C. Vas, H. Kimpton, A. Faizi, C. Liu, A. Porter, T. Synnerdahl, P. Karlsson, B. Rodu, Ethyl carbamate in Swedish and American smokeless tobacco products and some factors affecting its concentration, Chem. Central J. 12 (1) (2018) 1–17, https://doi.org/10.1186/s13065-018-0454-x.
[20] P. Li, S. Zeng, J. Zhang, Y. Shen, S. Sun, Y. Zong, J. Xie, D. Wang, J. Yang, Real-Time Monitoring of Nicotine Release Behavior from Smokeless Tobacco (Snus) Based on Fiber Optic Sensing Technology, Dissolution Technol. 26 (4) (2019) 24–30. https://doi.org/10.14227/DT260419P24.
[21] AOAC International, AOAC Official Method 960.07 Alkaloids (Total as Nicotine) in Tobacco Distillation Method, AOAC Official Methods of Analysis of AOAC International, AOAC International, Gaithersburg, MD, 1995, pp. 30-31.
[22] J.M. Garrigues, A. Pérez-Ponce, S. Garrigues, M. de la Guardia, Fourier-transform infrared determination of nicotine in tobacco samples by transmittance measurements after leaching with CHCl3, Anal. Chim. Acta 373 (1) (1998) 63–71, https://doi.org/10.1016/S0003-2670(98)00387-0.
[23] D. Jiang, G. Hu, G. Qi, N. Mazur, A fully convolutional neural network-based regression approach for effective chemical composition analysis using near-infrared spectroscopy in cloud, J. of Artif. Intell. and Technol. 1 (2021) 74–82. https://doi.org/10.37965/jait.2020.0037.
[24] Z. Zhu, G. Qi, Y. Lei, D. Jiang, N. Mazur, Y. Liu, D. Wang, W. Zhu, A long short-term memory neural network based simultaneous quantitative analysis of multiple tobacco chemical components by near-infrared hyperspectroscopy images, Chemosensors 10 (2022) 164, https://doi.org/10.3390/chemosensors10050164.
[25] D. Jiang, G. Qi, G. Hu, N. Mazur, Z. Zhu, D. Wang, A residual neural network based method for the classification of tobacco cultivation regions using near-infrared spectroscopy sensors, Infrared Phys. & Technol. 111 (2020), 103494, https://doi.org/10.1016/j.infrared.2020.103494.
[26] A. Stöbener, U. Naefken, J. Kleber, A. Liese, Determination of trace amounts with ATR FTIR spectroscopy and chemometrics: 5-(hydroxymethyl) furfural in honey, Talanta 204 (2019) 1–5, https://doi.org/10.1016/j.talanta.2019.05.092.
[27] C.M. Moreno-Ley, D.M. Hernández-Martínez, G. Osorio-Revilla, A.P. Tapia-Ochoategui, G. Dávila-Ortiz, T. Gallardo-Velázquez, Prediction of coumarin and ethyl vanillin in pure vanilla extracts using MID-FTIR spectroscopy and chemometrics, Talanta 197 (2019) 264–269, https://doi.org/10.1016/j.talanta.2019.01.033.
[28] C. Koch, A.E. Posch, H.C. Goicoechea, C. Herwig, B. Lendl, Multi-analyte quantification in bioprocesses by Fourier-transform-infrared spectroscopy by partial least squares regression and multivariate curve resolution, Anal. Chim. Acta 807 (2014) 103–110, https://doi.org/10.1016/j.aca.2013.10.042.
[29] S. Sen, Z. Dundar, O. Uncu, B. Ozen, Potential of Fourier-transform infrared spectroscopy in adulteration detection and quality assessment in buffalo and goat milks, Microchem. J. 166 (2021), 106207, https://doi.org/10.1016/j.microc.2021.106207.
[30] V. Szentirmai, A. Wacha, C. Németh, D. Kitka, A. Rácz, K. Héberger, J. Mihály, Z. Varga, Reagent-free total protein quantification of intact extracellular vesicles by attenuated total reflection Fourier transform infrared (ATR-FTIR) spectroscopy, Anal. and Bioanal. Chem. 412 (2020) 4619–4628, https://doi.org/10.1007/s00216-020-02711-8.
[31] K. Kheawfu, A. Kaewpinta, W. Chanmahasathien, P. Rachtanapun, P. Jantrawut, Extraction of nicotine from tobacco leaves and development of fast dissolving nicotine extract film, Membranes 11 (6) (2021) 403, https://doi.org/10.3390/membranes11060403.
[32] USA) (2017).
[33] S. Wold, M. Sjöström, L. Eriksson, PLS-regression: a basic tool of chemometrics, Chemometrics and intell. lab. syst. 58 (2) (2001) 109–130, https://doi.org/10.1016/S0169-7439(01)00155-1.
[34] H. Abdi, Partial least square regression (PLS regression), Encycl. for res. methods for the soc, sci. 6 (4) (2003) 792–795.
[35] H. Abdi, Partial least squares regression and projection on latent structure regression (PLS Regression), Wiley interdiscip, WIREs Comp Stat 2 (1) (2010) 97–106.
[36] L.M. Schmidtke, J.P. Smith, M.C. Müller, B.P. Holzapfel, Rapid monitoring of grapevine reserves using ATR–FT-IR and chemometrics, Anal. Chim. Acta 732 (2012) 16–25, https://doi.org/10.1016/j.aca.2011.10.055.
[37] A.C. Olivieri, Practical guidelines for reporting results in single-and multi-component analytical calibration: A tutorial, Anal. Chim. Acta 868 (2015) 10–22, https://doi.org/10.1016/j.aca.2015.01.017.
[38] P. Williams, The RPD statistic: A tutorial note, NIR news 25 (1) (2014) 22–26, https://doi.org/10.1255/nirn.1419.
[39] T. Fearn, Assessing calibrations: sep, rpd, rer and r 2, NIR news 13 (6) (2002) 12–13, https://doi.org/10.1255/nirn.689.
[40] Z. Tamiji, Z. Habibi, Z. Pourjabbar, M.R. Khoshayand, N. Sadeghi, M. Hajimahmoodi, Detection and quantification of adulteration in turmeric by spectroscopy coupled with chemometrics, J Consum Prot Food Saf 17 (3) (2022) 221–230.
[41] A.C. Olivieri, Analytical figures of merit: from univariate to multiway calibration, Chem. rev. 114 (10) (2014) 5358–5378, https://doi.org/10.1021/cr400455s.
[42] A.C. Olivieri, Introduction to multivariate calibration: A practical approach, Springer, 2018.
[43] A.C. Olivieri, N.M. Faber, J. Ferré, R. Boqué, J.H. Kalivas, H. Mark, Uncertainty estimation and figures of merit for multivariate calibration (IUPAC Technical Report), Pure and Appl. Chem. 78 (3) (2006) 633–661, https://doi.org/10.1351/pac200678030633.
[44] Z.P. Visak, L.M. Ilharco, A.R. Garcia, V. Najdanovic-Visak, J.M. Fareleira, F. J. Caetano, M.L. Kijevcanin, S.P. Serbanovic, Volumetric properties and spectroscopic studies of pyridine or nicotine solutions in liquid polyethylene glycols, The J. of Physical Chem. B 115 (26) (2011) 8481–8492, https://doi.org/10.1021/jp202464h.
[45] R. Rijal, M. Sah, H.P. Lamichhane, H.S. Mallik, Quantum chemical calculations of nicotine and caffeine molecule in gas phase and solvent using DFT methods, Heliyon 8 (12) (2022) e12494.
[46] T. Pongjanyakul, W. Khunawattanakul, C.J. Strachan, K.C. Gordon, S. Puttipipatkhachorn, T. Rades, Characterization of chitosan–magnesium aluminum silicate nanocomposite films for buccal delivery of nicotine, Int. j. of biol. macromol. 55 (2013) 24–31, https://doi.org/10.1016/j.ijbiomac.2012.12.043.
[47] T. Pongjanyakul, W. Khunawattanakul, S. Puttipipatkhachorn, Physicochemical characterizations and release studies of nicotine–magnesium aluminum silicate complexes, Appl. Clay Sci. 44 (3–4) (2009) 242–250, https://doi.org/10.1016/j.clay.2009.03.004.
[48] J. Coates, in: Interpretation of Infrared Spectra, a Practical Approach, Encyclopedia of Analytical Chemistry, John Wiley & Sons Ltd, Chichester, 2000, pp. 10815–10837.
[49] Camo Software AS., Help documentation of The Unscrambler® X 10.4 (Camo Software AS., Norway), 2016.
[50] USA) Help Topics (2017).
[51] M. Antonio, R.L. Carneiro, R.M. Maggio, A comparative approach of MIR, NIR and Raman based chemometric strategies for quantification of Form I of Meloxicam in commercial bulk drug, Microchem. J. 180 (2022), 107575, https://doi.org/10.1016/j.microc.2022.107575.
[52] E. Trovato, F. Vento, D. Creti, P. Dugo, L. Mondello, Elucidation of Analytical-Compositional Fingerprinting of Three Different Species of Chili Pepper by Using Headspace Solid-Phase Microextraction Coupled with Gas Chromatography-Mass Spectrometry Analysis, and Sensory Profile Evaluation, Molecules 27 (7) (2022) 2355, https://doi.org/10.3390/molecules27072355.
[53] F.A. Iñón, J.M. Garrigues, S. Garrigues, A. Molina, M. de la Guardia, Selection of calibration set samples in determination of olive oil acidity by partial least squares–attenuated total reflectance–Fourier transform infrared spectroscopy, Anal. Chim. Acta 489 (1) (2003) 59–75, https://doi.org/10.1016/S0003-2670(03)00711-6.
[54] Addinsoft, XLSTAT version 2016.02 (Addinsoft) Help topics, 2016.
[55] L. Cuadros-Rodríguez, M.G. Bagur-González, M. Sánchez-Vinas, A. González-Casado, A.M. Gómez-Sáez, Principles of analytical calibration/quantification for the separation sciences, J. of Chromatogr. A 1158 (1–2) (2007) 33–46, https://doi.org/10.1016/j.chroma.2007.03.030.
[56] F. Allegrini, A.C. Olivieri, IUPAC-consistent approach to the limit of detection in partial least-squares calibration, Anal. chem. 86 (15) (2014) 7858–7866, https://doi.org/10.1021/ac501786u.
[57] D. Valinger, L. Longin, F. Grbeš, M. Benković, T. Jurina, J.G. Kljusurić, A.J. Tušek, Detection of honey adulteration–The potential of UV-VIS and NIR spectroscopy coupled with multivariate analysis, Lwt 145 (2021), 111316, https://doi.org/10.1016/j.lwt.2021.111316.
[58] A.G. González, M.A. Herrador, A.n.G. Asuero, Intra-laboratory testing of method accuracy from recovery assays, Talanta 48(3) (1999) 729–736, https://doi.org/10.1016/S0039-9140(98)00271-9.
[59] V.G. Franco, V.E. Mantovani, H.C. Goicoechea, A.C. Olivieri, Teaching chemometrics with a bioprocess: analytical methods comparison using bivariate linear regression, The Chem. Educator 7 (2002) 265–269, https://doi.org/10.1007/s00897020596a.

[60] M.V. Amorim, F.S. Costa, C.F. Aragão, K.M. Lima, The use of near infrared spectroscopy and multivariate calibration for determining the active principle of olanzapine in a pharmaceutical formulation, J. of the Braz. Chem. Soc. 28 (2017) 920–926. https://doi.org/10.21577/0103-5053.20160233.
[61] M. Galea-Rojas, M.V. de Castilho, H. Bolfarine, M. de Castro, Detection of analytical bias, Analyst 128 (8) (2003) 1073–1081, https://doi.org/10.1039/b212547a.
[62] Centers for Disease Control and Prevention, Protocol to measure the quantity of nicotine contained in smokeless tobacco products manufactured, imported, or packaged in the United States, Federal Register 64 (55) (1999) 14086–14096.
[63] M.V. Djordjevic, K.A. Doran, in: Handbook of Experimental PharmacologyNicotine PsychoPharmacology, Springer Berlin Heidelberg, Berlin, Heidelberg, 2009, pp. 61–82.
[64] Z. Tassew, B.S. Chandravanshi, Levels of nicotine in Ethiopian tobacco leaves, SpringerPlus 4 (2015) 1–6, https://doi.org/10.1186/s40064-015-1448-y.

Microchemical Journal 201 (2024) 110670

Contents lists available at ScienceDirect

Microchemical Journal

journal homepage: www.elsevier.com/locate/microc

Rapid assessment of smokeless tobacco quality parameters using ATR-FT-MIR spectroscopy: Comparison of analytical/mathematical and machine learning approaches

Mohamed Fekhar, Yasmina Daghbouche *, Naima Bouzidi, Mohamed El Hattab

Laboratory of Natural Products Chemistry and of Biomolecules, Faculty of Sciences, University Blida 1, Blida, Algeria

ARTICLE INFO

Keywords:
ATR-FT-MIR spectroscopy
Smokeless tobacco
Quality parameters
Partial least-squares
Support vector machine

ABSTRACT

The Algerian smokeless tobacco (ST) market is characterized by a significant presence of counterfeit and illicit substandard products. The official methods for reliable ST monitoring are impractical for routine use. This study explores the application of attenuated total reflectance Fourier transform mid-infrared (ATR-FT-MIR) spectroscopy for a comprehensive analysis of two sets of Algerian ST collected in consecutive years (2021 and 2022).

In qualitative assessment, a two-step approach was adopted. Initially, principal component analysis, agglomerative hierarchical clustering, and *k*-means clustering were implemented as complementary unsupervised techniques to group commercial samples into distinct clusters based on reference measurements. These clusters then served as target categories for training supervised models, specifically partial least-squares discriminant analysis (PLS-DA) and support vector machine classification (SVM-C), enabling the classification of new samples solely based on their FT-MIR spectral features. The methodology successfully distinguished four classes, achieving calibration/validation accuracies of 95.8%/84.4% using SVM-C.

To achieve the main objective, partial least-squares regression (PLSR) and support vector machine regression (SVMR) were used to train and validate models for simultaneously predicting moisture, pH, ash, total nicotine, and un-ionized nicotine contents. With the exception of moisture, SVMR showed satisfactory performance, with determination coefficients ranging from 0.82 to 0.93 and prediction-to-deviation ratios from 2.4 to 3.7. This method offers a rapid, easy, and sustainable alternative, providing manufacturers and regulatory bodies with an effective means of controlling ST quality within acceptable prediction error margins for practical applications.

1. Introduction

Until several years ago, MADAR Holding Group (formerly SNTA) held a monopoly on the manufacture and marketing of Algerian moist smokeless tobacco (ST), locally known as “Chemma”. However, due to lax government anti-counterfeiting measures, the informal sector has gradually gained ground in the market. This sector has expanded beyond merely counterfeiting genuine certified products, instead producing and flooding the market with their own low-cost brands, which particularly attract younger consumers. In a ruthless pursuit of maximizing profits from these cheap products, illicit producers often resort to using waste materials of dubious origin, including expired or even rotten ingredients, posing serious health risks to consumers. Faced with the persistent growth of these substandard products, local consumer protection associations and the legal tobacco company find themselves struggling to keep pace, as their testing methods are limited to physical and detectable product characteristics.

In Western countries, the regulation of ST is governed by food legislation, and authorities implement diverse strategies to reduce its accessibility and use. Notably, the U.S. Food and Drug Administration (FDA) rigorously oversees the manufacturing, marketing, and sales of tobacco products and enforces stringent standards to monitor nicotine (NCT) and other ingredient levels [1]. Complementing these efforts, the Centers for Disease Control and Prevention (CDC) provide essential support to the FDA, offering technical assistance and scientific expertise. Specifically addressing ST, the CDC has established a standardized laboratory protocol for the comprehensive analysis of nicotine content, total moisture, and pH in all products manufactured, imported, or

* Corresponding author at: Faculty of Sciences, University Blida 1, P.O. Box 270 Blida, Algeria.
E-mail addresses: medfr73@gmail.com (M. Fekhar), ydaghbouche@yahoo.fr (Y. Daghbouche), bouzna@yahoo.fr (N. Bouzidi), elhattab@univ-blida.dz (M. El Hattab).

https://doi.org/10.1016/j.microc.2024.110670
Received 3 March 2024; Received in revised form 16 April 2024; Accepted 30 April 2024
Available online 1 May 2024

packaged in the U.S. [2].

From a research standpoint, analyzing quality parameters is crucial as these parameters are directly influenced by the product's chemical composition, which in turn depends on the quality of the ingredients and the amount of toxicants present. However, despite their high reliability in monitoring STs, the corresponding procedures are destructive to samples, time-consuming, energy-intensive, and require specialized technical knowledge along with various equipments. These limitations render such procedures unsuitable for routine use.

As an expedient and non-destructive alternative, near-infrared (NIR) spectroscopy is gaining increasing popularity in tobacco assessment. Previous work [3] has reported a successful use of the NIR technique to predict nicotine, physical indices, as well as a number of macromolecules, including cellulose, lignin, pectin, and proteins, in reconstituted tobacco. More recently, Geng *et al.* [4] and Shu *et al.* [5] have introduced new methods based on calibration transfer to determine, respectively, total sugars and total alkaloids in tobacco leaves. Using the same spectroscopic technique, optimum results have been achieved through deep learning approaches for quantifying nicotine [6] and simultaneously predicting nicotine, sugars, total nitrogen, and pH [7]. Nevertheless, these solutions have demonstrated efficiency mainly in predicting the composition of intact raw tobacco, and their performance is not yet thoroughly investigated for commercial intricate matrices such as ST. Moreover, inherent limitations in deep learning-based regressions, such as over-fitting problems [8] and a lack of mathematical frameworks for result validation, can compromise the robustness of the developed models, restraining their broader application.

On the other hand, Fourier transform mid-infrared (FT-MIR), renowned for its high ability to map chemical compounds, offers a wealth of valuable information per spectrum in comparison to NIR spectroscopy [9]. Despite this capability, its application in studying ST has been predominantly restricted to identifying tobacco species, non-tobacco plant components, individual chemicals, and specific additives [10–12]. Within other contexts, the MIR technique, when coupled with attenuated total reflectance (ATR) for sampling and employing partial least-squares (PLS) or support vector machine (SVM) for modeling, has emerged as a robust tool for qualitative and quantitative analyses. Noteworthy applications include predicting nitrogen and starch reserves in grapevines [13], sugars and organic acids in peaches [14], and adulteration profiling of cocaine seizures [15]. In comparative studies, MIR has demonstrated superior performance over NIR and other spectroscopic techniques in determining the physicochemical and rheological properties of apple purees [16] and quantifying adulterants in cumin powder [9].

To the best of our knowledge, no prior works have reported the monitoring of routine parameters in ST using MIR spectroscopy. Consequently, the present study seeks to explore the potential of the ATR-FT-MIR technique, in combination with chemometrics, as a rapid, cost-effective, high-throughput, and user-friendly method. The primary aims are as follows: 1) Identify the major ingredients of Algerian Chemma; 2) Classify the various commercial brands widely available in the markets; 3) Predict five crucial quality parameters, specifically moisture, pH, ashes, total nicotine, and free nicotine. Special attention has been dedicated to evaluating the performance of the developed models, ensuring the reliability and applicability of the proposed method for assessing efficient product control.

2. Materials and methods

2.1. Reagents and chemicals

(−)-Nicotine was of GC-grade and purchased from Sigma-Aldrich (China). Chloroform stabilized with ethanol was of Analytical-grade from Sigma-Aldrich (France). NaOH, anhydrous Na_2CO_3 and Na_2SO_4 were of Analytical Reagent grade and purchased from Biochem Chemopharma (France) or Panreac (Spain). Triple Distilled water (TDW) was used instead of deionized distilled water in pH measurements, and it was prepared in our laboratory according to the methodology described elsewhere [17].

2.2. Collection and preparation of samples

The survey consisted of two parts. The first part involved purchasing 46 samples of native Algerian oral tobaccos between January and June 2021. These samples were obtained from various sources, including wholesale and retail tobacco stores, as well as street vendors, from 10 different locations in two provinces: Medea and Blida. The initial set of samples included one control product provided by United Tobacco Company's quality control central laboratory (UTC of MADAR Group, Boumerdes), two commercially available genuine products bearing the same brand name, two counterfeit analogs of the certified brand, 39 illegally manufactured products, one homemade traditional Chemma variant, and one flavored non-tobacco (plant-based) product intended to aid in quitting tobacco use.

In the second part of the survey, conducted between January and June 2022, an additional 59 samples were collected from the same sources as the first set, with the exception of two new foreign products. This set of samples comprised the two commercial certified products from UTC, 54 shoddy products, two ST products made in Belgium and imported from France, and one other traditional product. It's worth noting that only 33 products from the initial set were resampled, as some samples were no longer available for purchase or had been rebranded. Brandnames are not shown for confidentiality reasons. All of the samples were estimated to collectively represent > 90 % of the ST market share in Algeria.

The content of each product was homogenized using a coffee grinder, returned to its original packaging or tins, labeled, sealed in plastic sleeves, and stored at -10 °C until analysis.

2.3. Reference measurements of ST quality parameters

2.3.1. Total moisture content

Moistures of STs were determined based on the CDC method [2] using a gravimetric oven drying of weighed portion (accurately 5.00 g) at 99.0 ± 1.0 °C for 3 h. This method is referred to as "total moisture" because it measures water and all other product volatile constituents at temperature of 99 °C.

This parameter is linked to overall microbial growth, and more specifically to certain nitrogen-transforming microorganisms that can indirectly promote the formation of tobacco-specific nitrosamines (TSNAs), which are well-known human carcinogens, during the various tobacco production processes [12,18]. Water content can also affect the product's shelf life and other quality parameters, such as pH and NCT levels.

2.3.2. pH determination

Accurately 2.00 g of homogenized product were placed in a 50 ml beaker with 20 ml of TDW and the suspension was magnetically stirred for 30 min as described in the revised protocol of CDC [2]. During filtering the supernatant in the dark, it was allowed to stand for an additional 20 min before analysis.

Besides other factors, the pH plays a crucial role in determining the proportion of NCT in its un-ionized form [19]; at higher pH levels, more NCT is un-ionized. The manipulation of the product's pH can significantly amplify both the pharmacological effects and addictive potential of NCT (see **Subsection 2.3.5** below).

2.3.3. Total ash content

The ash content was determined following previously reported procedure [20] with some modifications. 5.00 g of the whole product were burned in air on a hot plate at 300 °C in a silica dish for 30 min. It was then transferred to a muffle furnace and incinerated at 600 °C for 2 h to

ensure the complete removal of any carbon particles.

This analysis is called "total ash" because it measures the residual ash remaining after tobacco incineration, along with any mineral components. This can provide an estimate of the inorganic additives or adulterants that have been added to the final product.

2.3.4. Total nicotine level

The quantification of total NCT was performed using the method developed by our research group [21]. Briefly, 2.00 g of pre-dried ST were sonicated with 30 ml of distilled water for 20 min then 0.4 g of anhydrous Na_2CO_3 were added to the mixture and subsequently hand-shaken. After filtration and adjustment to pH 12 with NaOH, the aqueous phase was vortexed twice with 4 ml each of chloroform then centrifuged for 10 min. The resulting organic layer was carefully transferred to be concentrated on a water bath under vacuum at 35 °C to dryness. The final volume was adjusted with chloroform. For analysis, a 0.4 µl drop of the solution was mounted via a Hamilton micro syringe to form a thin dry film on the surface of the ATR crystal of an FTIR spectrometer. Multivariate PLSR model was used to quantify nicotine concentration.

Nicotine, the primary alkaloid in tobacco, can undergo demethylation to form other minor alkaloids that can transform via nitrosation to TSNAs during curing or storage at high temperatures [18,22]. Therefore, analyzing NCT is essential for assessing addictiveness and identifying high-TSNA products.

2.3.5. Un-ionized nicotine level

Assuming that nicotine is a weak base: $B + H^+ \rightleftharpoons BH^+$.

The concentration of un-ionized form of NCT can be calculated using the sample pH and the ionization constant ($pK_a = 8.02$) substituted into the Henderson-Hasselbalch equation [2]:

$$pH = pK_a + \log\frac{[B]}{[BH^+]} \quad (1)$$

After rearrangement:

$$FreeNCT(mg/g) = TotalNCT(mg/g) \times \frac{\frac{[B]}{[BH^+]}}{\frac{[B]}{[BH^+]} + 1} \quad (2)$$

Increasing alkalinity converts nicotine salts into the freebase form, which is more easily absorbed through the lining of the mouth. This can lead to a faster rate of nicotine reaching the central nervous system, potentially impacting its effects [19,23].

2.4. Instrumental analysis

All FTIR measurements were conducted using a Nicolet iS10 spectrometer equipped with a Smart iTR accessory diamond crystal (single reflection at incidence angle of 42°) and a deuterated triglycine sulfate (DTGS) detector connected to the in-built OMNIC™ software version 9.8 (Thermo Fisher Scientific, USA). Spectra were recorded over the MIR region ($4000 - 525\ cm^{-1}$) at a resolution of $4\ cm^{-1}$ and an average of 32 scans per spectrum in absorbance mode.

The samples were pressed directly against the diamond crystal using a pointed tip of the standard pressure device. Before each measurement, the crystal was cleaned, twice, with ethanol 96° followed by isopropanol.

2.5. Data analysis

2.5.1. Spectral pre-treatment

Several pre-processing strategies, including baseline offset correction (BO), extended multiplicative scatter correction (EMSC), standard normal variate (SNV), normalization transformations, Savitzky-Golay first-order derivative (SG FD) at 9 points/side with second polynomial order (PO = 2), Savitzky-Golay second-order derivative (SG SD) at 7 points/side with PO = 3, de-trending (DT) with PO = 2, and certain of their combinations were investigated so as to reduce the uncontrolled external effects.

2.5.2. Principal component analysis (PCA)

PCA is an unsupervised pattern recognition technique used for reducing dimensionality, exploring data, visualizing correlations, and detecting outliers [24]. It generates a new set of orthogonal principal axes, which are independent linear combinations of the original variables. The first principal component (PC-1) forms a single axis in space, capturing the maximum variance possible. The second component (PC-2), perpendicular to the first, captures the next highest variance, and so on [24,25].

In this study, mean-centered PCA was performed using the singular value decomposition algorithm with 10 PCs on i) Full spectral replicates, corrected using SNV transformation, to detect outliers stemming from various data collection issues (intra-group outliers). Outlying samples were identified by examining the scores plot and the influence plot (F-residuals vs. Hotelling's T^2 statistics) at a significance level of 0.05 probability; ii) Standardized reference measurements (correlation matrix) of samples, along with their corresponding SNV-transformed spectra (excluding CO_2, ATR crystal, blank and noise regions) as mentioned above, were used to remove influential samples showing significant deviation from the overall trend of the data (inter-group outliers, see supplementary Fig. S1); iii) Standardized reference measurements dataset of the remaining samples from step (ii) to examine distribution trends, revealing any visible clusters could be discriminated based on similarities and dissimilarities between samples.

2.5.3. Agglomerative hierarchical clustering (AHC)

Or Hierarchical Cluster Analysis (HCA) is a partitioning technique consists in constructing a binary clustering tree (dendrogram). It starts by calculating the dissimilarity between the *I* observations stored at the leaves (singletons) then proceeds by merging two by two the "closest" subsets (stored at nodes) based on a measure of dissimilarity between the elements and/or classes through a minimization of a given agglomeration criterion. This process continues until the root of the tree, enclosing all the observations, will be reached [26,27].

In addition to the nature of the data, the clustering outcome depends on the proximity type and the agglomeration strategy. Hence, the simple, flexible, complete, unweighted and weighted pair-group average linkages, as well as Ward's method, were investigated in this work. Furthermore, truncating the dendrogram at a given level was realized automatically depending upon criteria such as entropy or inertia values, or manually by evaluating the relationships between samples at all levels as well as the evolution of within-class variance.

2.5.4. k-means Clustering

It is an unsupervised machine learning technique used to group similar observations into *k* distinct clusters. It starts by selecting *k* data points as initial cluster centers (centroids) then each point is assigned to the nearest centroid based on Euclidean distance. Next, the position of centroids is updated by calculating the new cluster means after each point loss or gain. The assignment and centroid update steps are repeated until convergence, where no point changes clusters [28,29]. This technique has the advantage that the assignment of elements is reversible from iteration to another, a thing that would have been impossible with AHC.

In this study, four classification criteria were used to reach the most acceptable solution, namely the W trace, determinant of W, Wilks lambda, and Trace (W)/Median. The relevant number of classes was determined via the "elbow finding" method from the plot of the within-class variance versus *k* values (varying from 1 to 10).

2.5.5. Partial least squares regression (PLSR)

PLSR is a mathematical-theoretical technique used for modeling re-

lationships in high-dimensional and/or collinear data. It decomposes the predictors (spectra, X-matrix) and the responses (reference measurements, Y-matrix) into scores and loadings [30].

$$X = TP^T + E_X \tag{3}$$

$$Y = UQ^T + E_Y \tag{4}$$

where E_X and E_Y are the residual matrices of X and Y, respectively.

Iteratively, a predictive model is created by maximizing the covariance between T and Y, resulting in weight vectors (W) that used to calculate the regression coefficients (B) used to predict the response variable for new data [30,31].

$$B = W\left(P^T W\right)^{-1} Q^T \tag{5}$$

The estimation of P, Q, T and W may also be achieved by extracting eigenvectors of the smallest size of products of X , X^T, Y and Y^T in the kernel PLS basis algorithm.

Before regression, the remained samples after outlier removal were divided into calibration (67 % of the data) and test (33 % of the data) sets using the Kennard-Stone sample selection method applied to the PCA scores plot of the actual measurement dataset (see step ii in **subsection 2.5.2** and Fig. S2 in supplementary data). To develop robust PLSR models, the appropriate spectral range, spectral pre-processing, and the number of latent variables (LVs) were carefully set up. In addition to considering the full spectral range, two methods were compared for selecting the most relevant wavenumbers: interval-partial least squares (*i*-PLS) and variable importance for the projection (VIP). The optimal number of LVs was determined using two approaches: i) Examining the explained cross-validation (CV) variance plot, seeking the highest value with as few factors as possible; ii) Determining Haaland's criterion, which involves using the first model computing for a p-value < 0.75 of the ratios between the prediction error sum of squares (PRESS) values for each number of LVs and the minimum possible PRESS [31]. All analytical models were constructed using the kernel PLS algorithm with a maximum of 15 factors and mean centering of the data.

2.5.6. Support vector Machine regression (SVMR)

It is a statistical learning technique adapted from the SVM algorithm that was primarily used for classification tasks. SVMR aims to find the optimal function *f(x, w)* that maps predictors (x) to response variables (y). Unlike conventional regression techniques, SVMR can model complex and non-linear relationships using kernel functions ($\phi(x_i, x_i')$) while maintaining a margin of error within a certain tolerance level (ε) to minimize the prediction error. One way to ensure this is by solving the following formulation [32,33]:

$$min_{w,b,\xi,\xi^*} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{I}\left(\xi_i + \xi_i^*\right) \tag{6}$$

Subject to:

$$\begin{cases} y_i - w^T\phi\left(x_i, x_i'\right) - b \le \varepsilon + \xi_i \\ w^T\phi\left(x_i, x_i'\right) + b - y_i \le \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \ge 0; \forall i, i \in (1, 2, \cdots, I) \end{cases} \tag{7}$$

where C is a regularization parameter that determines the trade-off between the model complexity (flatness) and the degree to which deviations larger than ε are tolerated, w is the weight vector, b is the bias term, ξ_i and ξ_i^* are slack variables that contrast the symmetric boundary produced by the "hard margin" loss function. The optimization problem can be solved using standard dualization method utilizing Lagrange multipliers. The solution provides the optimal weights and bias for the support vectors [34].

During the regression step, all quality parameters were optimized for various pre-processing strategies, kernel types (linear, polynomial, and radial basis function), and three hyperparameters (C, ε, and γ) using the ε-SVR algorithm on data scaled to [−1, 1]. The hyperparameter γ, which determines the width of the Gaussian kernel, was only adjusted in custom kernelized SVMR. The initial ε value was set to approximately 5 − 10 % of the corresponding standard deviation value. A comprehensive grid search, logarithmically scaled within the range of [10^{-4}, 10^4], was then conducted to determine the optimal values of C and γ. Each combination of hyperparameters was rigorously evaluated using a 10-fold CV framework, and the combination yielding the lowest RMSECV was chosen. In cases of comparable performance, the model with fewer support vectors (SVs) was preferred due to its reduced sensitivity to minor fluctuations in the training data and lower risk of over-fitting.

2.6. Evaluation metrics and software

Both PLSR and SVMR models were validated using the full (leave-one-out) CV method, tested to predict the targets in an independent test set, and the performance of all models was checked using root mean square errors (RMSE) and determination coefficients (R^2) of calibration, cross-validation, and prediction. Critical statistics, including the mean absolute percentage error (MAPE), relative error of prediction (REP), ratio of prediction-to-deviation (RPD), range error ratio (RER), and bias for the test set, were assessed in order to corroborate the robustness of the best-performing models. These were calculated as follows:

$$\text{RMSE} = \sqrt{\frac{1}{I}\sum_{i=1}^{I}\left(y_i - \widehat{y}_i\right)^2} \tag{8}$$

$$\text{R}^2 = 1 - \frac{\sum_{i=1}^{I}\left(y_i - \widehat{y}_i\right)^2}{\sum_{i=1}^{I}\left(y_i - \overline{y}\right)^2} \tag{9}$$

$$\text{MAPE} = 100\sum_{i=1}^{I}\left|\frac{y_i - \widehat{y}_i}{y_i}\right| \tag{10}$$

$$\text{REP} = 100\frac{RMSEP}{\overline{y}_{Cal}} \tag{11}$$

$$\text{RPD} = \frac{SD_y}{SEP} \tag{12}$$

$$\text{RER} = \frac{y_{max} - y_{min}}{SEP} \tag{13}$$

$$\text{Bias} = \frac{1}{I}\left(\sum_{i=1}^{I}\widehat{y}_i - \sum_{i=1}^{I}y_i\right) \tag{14}$$

where I is the number of samples, y_i and $\widehat{y}_i$ are actual and predicted values for the i^{th} sample, y_{max} and y_{min} are the maximum and minimum values in the reference datasets, respectively, $\overline{y}$ is the average of actual values, SD_y is the standard deviation of the reference values in the test dataset, and *SEP* is the RMSEP corrected for bias.

Among analytical figures of merit, the limit of detection (LOD) holds paramount importance as it reveals the lowest detectable concentration of an analyte using the corresponding model, enabling comparison with other methods. According to Allegrini & Olivieri [35], the LOD in PLSR encompasses two limits: lower and upper, considering both the risks of false detects (α-error) and false non-detects (β-error). As discussed in [31], it can be readily calculated from the sensitivity and uncertainties in signals and concentrations. However, this is not the case with SVMR which provides more complex models with potentially non-linear relationships between input features, implicitly transformed into a higher-dimensional space, and the response. Therefore, simpler methods for calculating the LOD are desirable. Several approaches exist in the literature, but the methods presented in Ortiz *et al.* [36] (Eq. (15)) and ICH guidelines [37] (Eq. (16)) appear most suitable for this work.

$$\mathrm{LOD_{pu}} = \frac{2t_{(0.05,I-2)}}{b'}\sqrt{\left(\frac{1}{K}+\frac{1}{I}+\frac{\bar{y}^2}{\sum_{i=1}^{I}(y_i-\bar{y})^2}\right)\frac{\sum_{i=1}^{I}(\hat{y}_i-\hat{y}_i')^2}{I-2}} \quad (15)$$

$$\mathrm{LOD_{ICH}} = \frac{3.3\times SD_a'}{b'} \quad (16)$$

where $t_{(0.05,\ I-2)}$ represents the Student's *t*-distribution for the given level of significance (α and β) and degrees of freedom ($I-2$), b is the slope, SD_a is the standard deviation of the intercept, K is the number of replicates, the prime symbol (′) indicates values estimated from the pseudo-univariate (pu) calibration line, and the remaining terms have been defined previously. The LOQ was estimated as 3 times the corresponding LOD.

In optimal cases, errors and bias should be close to 0, R^2 should be close to 1, and the difference between their respective values should be minimal for the three sets. Overall, the model is optimal when MAPE, REP, LOD and LOQ are minimal, and RPD and RER are much big. Moreover, the prediction accuracy was inferred by conducting the elliptical joint confidence region (EJCR) analysis at a 95 % level for the slope and intercept of the regression predicted versus actual for the testing set, based on ordinary least-squares (OLS). The inclusion of the ideal point (slope = 1, intercept = 0) within the confidence region indicates the absence of significant bias in the estimated values [38].

PCA, AHC, and *k*-means clustering were performed using XLSTAT v.2016 (Addinsoft SARL, France) executed in Microsoft Office (MS) Excel 2010 (Microsoft, USA). Spectral pre-processing, PLS-DA, PLSR, SVM-C, and SVMR were implemented with The Unscrambler® X 10.4 (Camo Software AS., Norway). The *i*-PLS and EJCR tests were performed using the free MATLAB codes referred to in [39]. The remaining statistical parameters were calculated with MS Excel or by costume-made functions executed in MATLAB R2012a (Mathworks Inc., USA).

3. Result and discussion

3.1. Reference measurements of ST quality parameters

Before performing calibration, it is crucial to identify and exclude outliers – samples suspected of not belonging to the target population. This serves two key purposes in spectral analysis. First, it ensures that the model learns from the most representative data during development, enhancing its generalization ability. Second, it guarantees that unknown samples to be tested fall within the model's coverage, leading to more reliable predictions [30]. Analysis of the PCA influence plots (supplementary Fig. S1) revealed nine suspected outliers, including the non-tobacco (plant-based snuff), the traditional Chemma, and the two foreign ST products. Indeed, these STs exhibited extreme values likely due to their abnormal chemical components, rendering them influential and significantly deviating from the overall trend of the data. Despite constituting only 8.6 % of the data, the inclusion of these samples in model development could compromise the accuracy and robustness. Their removal led to an improvement in data uniformity within each set.

The application of the Kennard-Stone−PCA method on combined subsets (inclusive of samples collected in both 2021 and 2022) was conducted to ensure the formation of a balanced dataset. This approach facilitates the inclusion of representative samples in the calibration process, effectively circumventing issues related to the variability in chemical composition among tobacco plants. Such variability typically arises from various factors, including plant genetics, growth conditions (climate, soil), and agricultural practices. By combining samples from different years, a wider range of this natural variability is captured by the model. This, in turn, improves the model's generalization ability to unseen data. The Kennard-Stone method further enhances this advantage by guaranteeing a uniform distribution of selected samples across the feature space, fostering a well-balanced calibration set.

The statistical process control of the quality (physicochemical) parameters before and after outlier removal is summarized in Table 1, whereas the data frequency distributions for the calibration and prediction sets are shown in Fig. S3. As evident from Table 1, the calibration subsets exhibit equivalent or marginally wider value ranges than the testing sets for each parameter. Comparable mean, median, and SD values were obtained for both subsets, reconfirming the effectiveness of the partitioning approach in maintaining data uniformity and ensuring that the selected samples represent the anticipated variations in commercial products. For normality testing, a normal probability plot and two-tailed Kolmogorov-Smirnov test at a significance level of 5 % with Lilliefors' correction were exploited. The results indicate that all quality parameters exhibited a normal distribution after outlier removal and application of Kennard-Stone method, with p-values > 0.1.

On a different point, the observed increase in pH and ash contents among products is attributed most probably to the addition of alkaline agents and inorganic fillers, respectively. While the increase in total nicotine is directly related to the quality of tobacco blend and inversely related to the proportion of other ingredients. Although moisture content is less relevant as it can be manipulated to alter the final weight, it can sometimes serve as an indicator of the product's shelf life and storage conditions.

A preliminary comparison of our findings on Algerian Chemma with oral tobacco products from various global regions studied elsewhere [10], suggests a potential similarity with Sudanese "Toombak". However, additional analyses, such as TSNA levels, are necessary to affirm this conclusion.

3.2. Spectra profile

Like many global moist snuff products, the Chemma may contain non-tobacco plant materials and other additives. Within plants, complex networks of cellulose microfibrils, hemicelluloses, lignin, and structural proteins are present [40]. All of these components can be detected in MIR spectroscopy and collectively contribute to the complexity of the spectrum, making its qualitative interpretation a challenging task. However, through the application of simple mathematical treatments such as subtraction or SD spectrum on adequate samples, the interpretation of the results can be significantly enhanced. Unprocessed spectra of the 96 selected STs are shown in supplementary Fig. S4, as shown, the regions carrying the most important variations between samples occurred between 3700 − 2430 and 1860 − 615 cm^{-1}. Fig. 1 illustrates the average (A) and SD (B) spectra in a normalized scale compared with the most likely used ingredients. Several prominent absorption bands can be identified from the mean spectrum, and their respective assignments [40–42] are detailed as follows: A strong and broad absorption band ranging from 3700 to 3000 cm^{-1}, peaking at 3343 cm^{-1}, can be attributed to O − H stretching of water, O − H stretching of cellulose and carbohydrates, and N − H stretching of proteins in plants. The two well-known peaks at 2918 and 2850 cm^{-1} arise from the C − H asymmetric and symmetric stretching of methyl and methylene groups found in various organic compounds. Tiny peaks at 2513 and 1795 cm^{-1} are specific indicators of the presence of calcium carbonate. A covered, faint band observed at 1732 cm^{-1}, discerned from the pure tobacco blend spectrum, is likely associated with hemicelluloses. The slightly broad band spanning from 1700 to 1600 cm^{-1}, centered at 1615 cm^{-1}, is primarily attributed to C = C stretching in lignin, and possibly in pectin, certainly overlapped with O − H bending of water (1640 cm^{-1}).

By examining the variance of peaks in the SD spectrum, characteristic bands at 1568 and 1539 cm^{-1}, can be overlapped with absorption vibrations from lignin, are linked to the N − H bending and C − N stretching in amide II. This connection is supported by the observation that the older plants/products exhibited, generally, lower protein content. Within the spectral range from 1500 to 1340 cm^{-1}, multiple overlapped bands are discernible. These bands are attributable to the

Table 1
Sample profile of the Algerian ST based on five quality parameters before and after outlier removal.

Sample allocation	Parameter (unit)	Subsets	No. of samples	Min.	Max.	Mean	Median	SD	p-value *
Before outlier removal	Moisture (%)	All samples	105	39.9	52.6	48.2	48.0	1.9	0.52
	pH		105	4.75	11.64	10.22	10.24	0.87	0.071
	Ash (%)		105	4.8	30.4	24.3	24.4	2.7	0.024
	Total NCT (mg/g, dwb)		105	0.0	16.5	8.1	8.1	2.9	0.37
	Free NCT (mg/g, dwb)		105	0.0	15.6	7.9	8.0	2.8	0.58
After outlier removal	Moisture (%)	Calibration	64	44.4	51.9	48.4	48.5	1.7	0.85
		Test	32	44.7	50.8	48.2	48.0	1.5	0.55
	pH	Calibration	64	9.00	11.64	10.34	10.25	0.61	0.68
		Test	32	9.10	11.62	10.36	10.28	0.63	0.45
	Ash (%)	Calibration	64	19.8	28.6	24.7	24.8	1.8	0.97
		Test	32	20.2	28.1	24.4	24.3	1.7	0.99
	Total NCT (mg/g, dwb)	Calibration	64	3.9	15.9	8.0	8.1	2.6	0.70
		Test	32	3.9	15.9	8.6	8.2	2.7	0.97
	Free NCT (mg/g, dwb)	Calibration	64	3.7	15.6	7.9	8.0	2.6	0.70
		Test	32	3.7	15.6	8.5	8.1	2.7	0.96

Abbreviations: dwb, Dry weight basis; NCT, Nicotine; No., Number; SD, Standard deviation
* Probability values are from a two-tailed Kolmogorov-Smirnov test for normality. A $p > 0.1$ indicates that the data distribution cannot be rejected from normality at a significance level of 0.05.

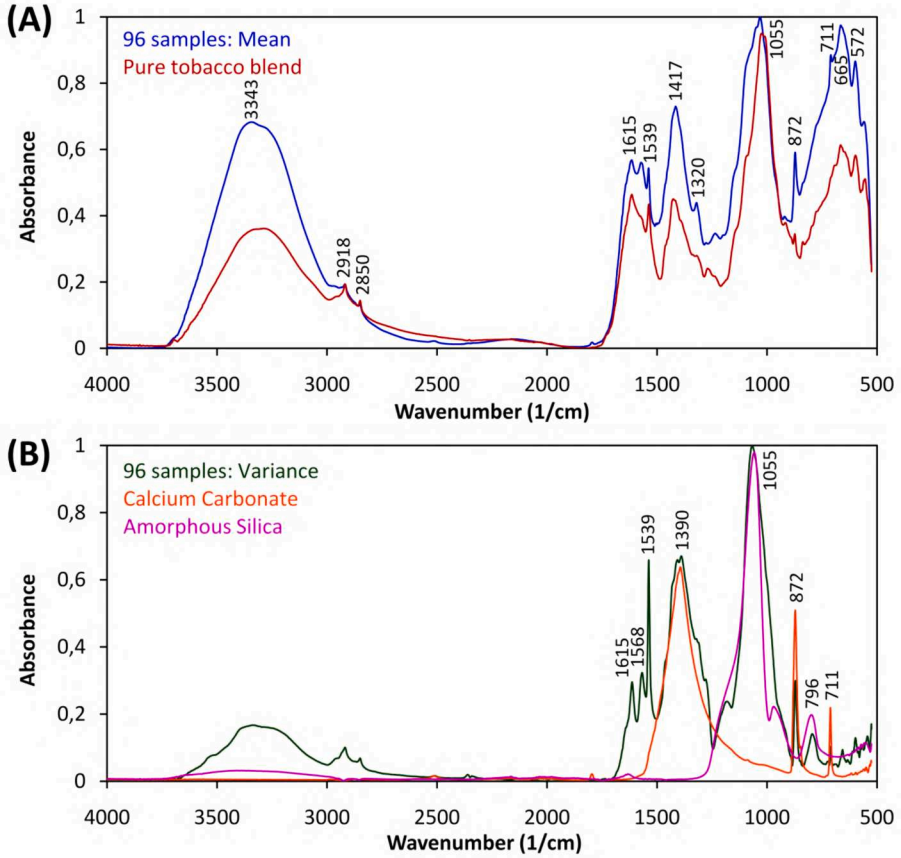


Fig. 1. Representative ATR-FT-MIR spectra in a normalized scale. The average (A) and SD (B) of the selected oral tobaccos compared with the most likely used ingredients.

bending vibrations of C − O − H and C − H, as well as the asymmetrical stretching of C − N − C in cellulose and lignin. Additionally, they coincide with the asymmetric C − O stretching of the carbonate ion (CO_3^{2-}) at approximately 1390 cm^{-1}. The medium-intensity peak at 1320 cm^{-1} is likely correlated with C − H bending of cellulose, C − N stretching in amide I of proteins or various plant components, including lignin, xyloglucan, or wax-like aliphatic compounds. A strong broad-band (1270 − 900 cm^{-1}) with a central peak at 1055 cm^{-1}, confirmed by comparison with the SD spectrum to be assigned to the Si − O asymmetric stretching of amorphous silica. Despite the observed variability in this region, the peak location and overall spectral features, including the presence of an additional band at 796 cm^{-1}, closely resemble the reference spectrum of pure silica and provide further support for our assignment. This range also encompasses heavily overlaps from interfering peaks at 1147, 1093, and 1031 cm^{-1} corresponding to C − C, C − O, C − O − C stretching, and C − O − H bending vibrations in cellulose, lignin, and other carbohydrates. The region between 1000 and 525 cm^{-1} is notable for its baseline shift in all samples containing water, as well as for the over-sizing effects of lower wavenumbers due to the ATR crystal reflection. The presence of a substantial number of weak

bands from common plant minor constituents, including carbohydrates (915 − 840 cm^{-1}), proteins (780 cm^{-1}), and inorganic compounds (615 − 550 cm^{-1}), alongside major constituents, complicates the visual interpretation of this region. Exceptionally, two distinct narrow peaks at 872 and 711 cm^{-1} are respectively identified to be assigned to the out-of-plane and in-plane vibrations of the CO_3^{2-} ion.

3.3. Classification analyses

When dealing with illicit products, traditional quantifiable metrics like popularity or market share are often difficult to obtain due to the clandestine nature of these markets. Instead, factors like perceived quality or reputation within underground markets become more significant. Classifying samples based on these factors can be crucial in conformity control. However, directly incorporating qualitative assessments or occasionally subjective perceptions into a supervised learning model can result in less reliable classifications compared to models built on objective physicochemical measurements. Yet, conducting unsupervised methods on spectral differences of anonymous type of products takes the risk to cluster samples based on irrelevant information in the spectra.

To tackle these challenges, we proposed a two-step approach that combines unsupervised and supervised learning. This involves applying unsupervised methods to reference measurements to extract informative features, which are then used as input for the supervised learning model. This methodology allows for the creation of a model that is both grounded in empirical data and less susceptible to biases and uncertainties of human-labeled categories. Furthermore, it opens up the possibility of achieving future classification directly from spectral measurements, even when traditional quality metrics are unavailable.

At first, PCA was applied on the entire 96 samples to investigate the general trend among local STs with respect to the control sample (see supplementary Fig. S5). A two-dimensional (2-d) scores plot of the first two PCs, which explained 71 % of the total variance, was primarily generated. At least two groups can be differentiated on the positive and negative parts of PC-1, and at most, three groups can be barely distinguished by splitting the PC-1 positive part into two sub-clusters. Overall, no clear sample groupings or neat separation of data was perceived between these clusters, especially on the axes.

Consequently, in order to gain an idea of the suitable number of classes, an AHC analysis should be performed. The AHC dendrogram was created using the same dataset previously used in the PCA. After several tests, a satisfactory partitioning of observations was achieved using the Ward's method (Fig. 2A). This latter was reported to afford the best outcomes in previous studies [43]. The automatic truncating of the dendrogram depending on the calculated entropy value shows a high dissimilarity value (up to 60) dividing all samples into three major classes. This finding agrees with the previous PCA results in many aspects. Nonetheless, examination of the within-class variances (Fig. S6 in supplementary material), searching for "the elbow point" suggests better truncating the data into four classes. Class-1, composed of 30 observations, was the first discriminated on the far right arm, and it includes the cheapest unpopular shoddy products. Merged with Class-1 at the first level, Class-3 contained the most popular and low-cost illegal trademarks, while comprising 25 observations. The second level of the dendrogram, incorporating samples showed close similarity to the two certified brands of the UTC, was divided into two sub-clusters. The first one on the right arm, denoted as Class-2, composed of 27 observations and represents more pricey and abundant illegal products analogous to the certified commercial product N°2. The last separated samples, designated as Class-4, included the control sample, the certified product N°1, counterfeit analogs and other substandard products sharing similar physicochemical characteristics. These outcomes were unexpected since the two genuine brands are anticipated to group together due to their fundamentally identical constituents predetermined by the UTC Company.

In a third step, *k*-means clustering was performed to further evaluate the results obtained from AHC. The commonly used criterion in *k*-means, the W trace (pooled sums of squares and cross-products matrix), provided the most favorable outcome in this context. Minimizing the W trace minimizes the total within-class variance for a given number of classes, as consequence the heterogeneity of the groups. Indeed, a lower inertia value was obtained using this technique (2.26) compared to AHC (2.35), further recommending four classes for the studied dataset. The sample agreement rate, which refers to the percentage of samples assigned to the same category by both methods, was 83.3 %. Meanwhile, the number of observations in each cluster changed to 21, 22, 29, and 24 for Classes 1, 2, 3, and 4, respectively. Interestingly, the *k*-means outcome closely mirrors the underlying patterns associated with the quality distinctions among various brands. This is further confirmed by a 90.6 % agreement rate between the *k*-means clusters and the reference patterns based on three key qualitative attributes of products: price, reputation, and perceived quality. We believe that the remaining 9.4 % discrepancy could be due to confusion or misleading attempts by sellers. The *k*-means model, which relies on objective physicochemical measurements, was likely able to detect these discrepancies.

However, while *k*-means demonstrated superior performance, it lacks the ability to visualize the proximity between classes or observations, a capability that AHC and PCA possess. Accordingly, the results from *k*-means were exploited qualitatively to display observations in different colors, locate centroids, and exhibit confidence ellipses around each class. Yet, relying exclusively on the first two PCs proved insufficient for fully capturing the complexity of the data. The incorporation of PC-3, on the other hand, further enhanced its representation, accounting for > 91 % of the total variance. A deep examination of the reference categorization and outputs (Fig. 2B − D) revealed results that, while not drastically different, offered more meaningful insights compared to AHC. Notably, the genuine commercial products are logically regrouped within Class-4, as well as their counterfeit analogs, and only a small number of samples exhibited class switching, leading to improved homogeneity within groups. Essentially, the products grouped together with authentic brands within the same cluster are those that most successfully emulated the genuine product's characteristics. This suggests that these products effectively replicated the original formula, potentially posing a challenge in distinguishing them from authentic brands based on the assessed quality parameters. Excluding a few samples that can be considered atypical, these findings reinforce the conclusions drawn from AHC analysis. Moreover, Fig. 2C and D depicting PC-2 vs. PC-1 and PC-3 vs. PC-1 scores successfully distinguished Class-1 from Class-3 and Class-4 from Class-2, respectively. The data distribution exhibits a uniform pattern resembling a regular tetrahedron (see Fig. 2E).

Now, for a comprehensive understanding of the physicochemical variations underlying the observed group separation, PCA loading coefficients assume the pivotal role of quantifying the individual influence of each variable (parameter) on the designated PCs. The corresponding PCA loadings were plotted with class centroids in 3-d plot as shown in Fig. 2F. On this chart, the type of correlation can be inferred from angles, either between the variables or between the variables and the PCA axes. As it can be seen, nicotine was mainly responsible for the sample grouping across PC-1. Notice that total NCT and free NCT coincided; this is explainable since the majority of products had pH values > 9.4, at this level, 96 % of nicotine is on its un-ionized form. The moisture and ash were positively correlated and both contributed to PC-2, along with independent variations towards nicotine. This suggests that total ash is mostly linked to the inorganic "kieselguhr" used as a hygroscopic agent and filler in making Chemma. Kieselguhr, composed of 90 % amorphous silica, is consistent with the distinct Si − O stretching band observed in FTIR spectra. The pH was highly loaded onto PC-3, and it was perpendicular on both nicotine and moisture/ash vectors, which means that they are unrelated to each other.

Aiming to interpret each class characteristics, samples located in the

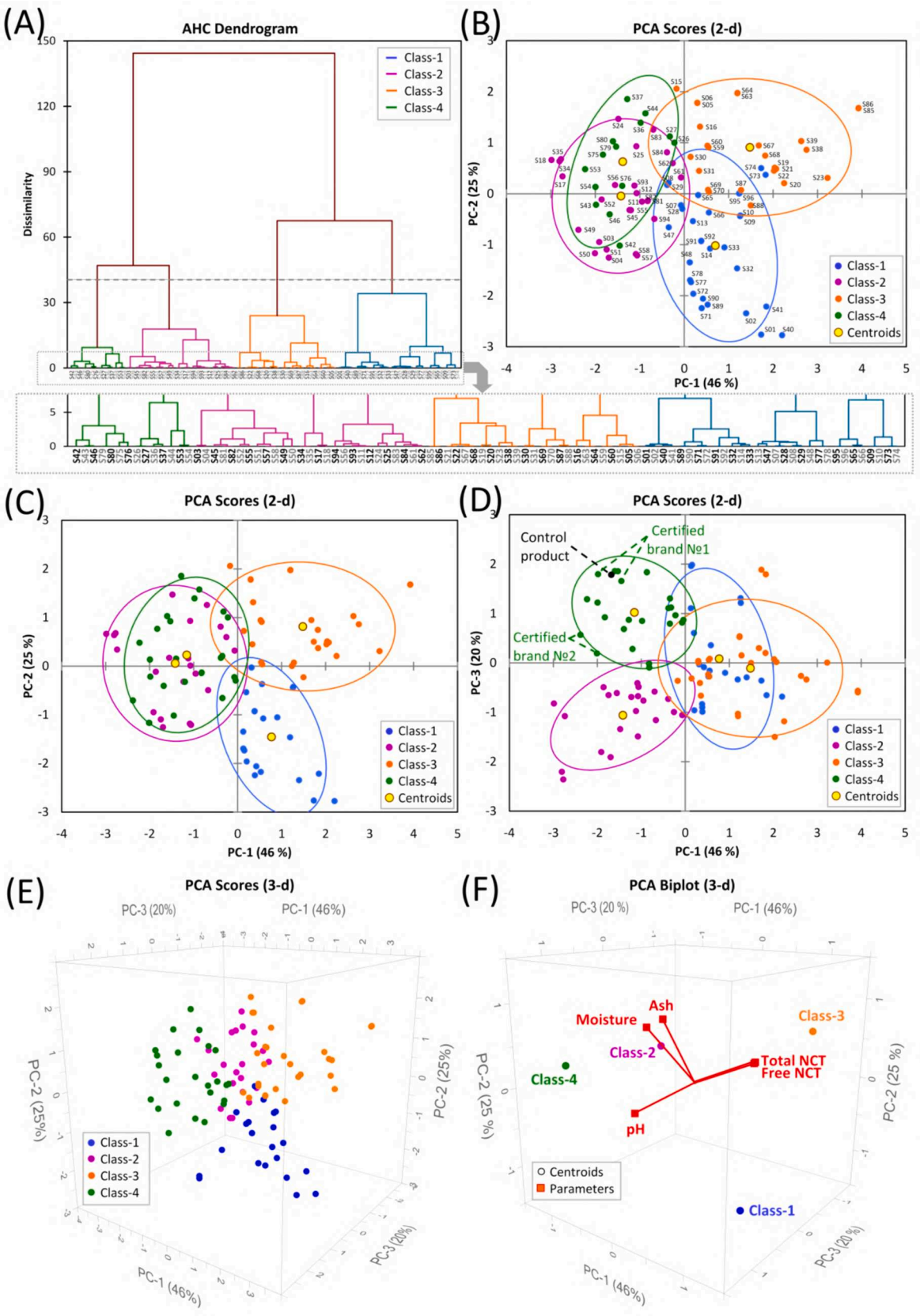


Fig. 2. AHC and PCA analyses based on reference measurements. A) AHC dendrogram. B) PCA scores displaying class according to AHC outcomes. C), D), and E) PCA scores displaying class according to *k*-means outcomes. F) PCA biplot of loadings and class centroids. (Ellipses are at an 80% confidence interval. For interpretation of the references to color and text in this figure legend, the reader is referred to the web version of this article.).

same direction as a given parameter have large values for that parameter. The samples of Class-3 have extreme NCT concentrations and relatively lower pH levels. Class-1 shows medium concentration of NCT and the lowest moisture and ash contents. Given their alignment with PC-3 (highly loaded with pH), samples of Class-4 exhibit the highest pH values, indicating their alkaline nature. In other aspects, Class-4 shares similarities with Class-2, such as low NCT and high moisture/ash contents. These findings affirm again that not only the proportion of other ingredients affects the nicotine level in a final product but also the used tobacco variety.

As a last step, the outputs from *k*-means were used to train two supervised methods, specifically partial least squares discriminant analysis (PLS-DA) and support vector machine classification (SVM-C). This approach allows leveraging the unsupervised clustering based on physicochemical measurements and translating it into a supervised learning classification task. Despite the potential limitation that *k*-means clusters might not perfectly reflect the exact reference categories, this method provides a reliable starting point for a more objective characterization of the samples.

PLS-DA had basically the same principles of PLSR, although the task (categorical, in the case of PLS-DA) differs. The key difference between support vector machine regression (SVMR) and support vector machine classification (SVM-C) lies in their core tasks and objectives. While SVMR tries to find a hyperplane that minimizes the prediction error along with maintaining a certain margin, SVM-C seeks to identify a hyperplane within the feature space that maximizes the margin between distinct classes, facilitating effective separation.

In similar way to a regression, various spectral pre-treatments were tested (supplementary Table S1). The discriminatory power of models was evaluated by calculating the accuracy (ratio of the samples correctly classified to the total samples) of calibration and cross-validation. The confusion matrices (Fig. 3) reveal that both predictive models encountered difficulties in clustering Class-1 samples, possibly due to high sample similarity or intricate product composition. Notably, despite the use of a high number of LVs (n = 14) in PLS-DA to capture the data's complexity, SVM-C outperformed, achieving calibration/validation accuracy percentages of 95.8/84.4 compared to 85.4/67.7 for PLS-DA. This signifies that SVM-C has the potential to be a valuable tool for future classification of unknown samples based on their MIR spectra, without relying on subjective judgments.

In summary, the use of *k*-means clustering and SVM-C has facilitated the classification of commercial products based on measured physicochemical parameters into four categories reflecting the samples' actual qualitative attributes (comprising price, reputation, and perceived quality). The proposed approach, which depends on simple spectral measurements, offers a promising solution for quickly monitoring and identifying substandard products that could pose health risks. A distinct drawback, however, is this method's insufficiency in differentiating between genuine products and their counterfeit analogs relying solely on conventional quality parameters. This is likely due to meticulous imitation efforts. To address this issue, future research will focus on exploring methods like data-driven soft independent modeling of class analogy (DD-SIMCA) to develop robust models built on datasets from FTIR spectrometry, or combined FTIR and other analytical technique.

3.4. Prediction of quality parameters

In the main part of this study, PLSR and SVMR models were trained and validated to characterize moisture, pH, ash, total nicotine, and free nicotine contents in commercial Chemma products. Determination of the first three parameters can be more directly established from the FT-MIR spectra since multiple corresponding bands were already distinguished (**Part 3.2**). On the contrary, due to its low abundance in a high complex matrix, nicotine determination from the whole sample spectra is mostly indirect. The detailed results of optimization process are presented in Tables S2 and S3 in supplementary material for PLSR and SVMR, respectively, whereas only the key models were summarized in Table 2.

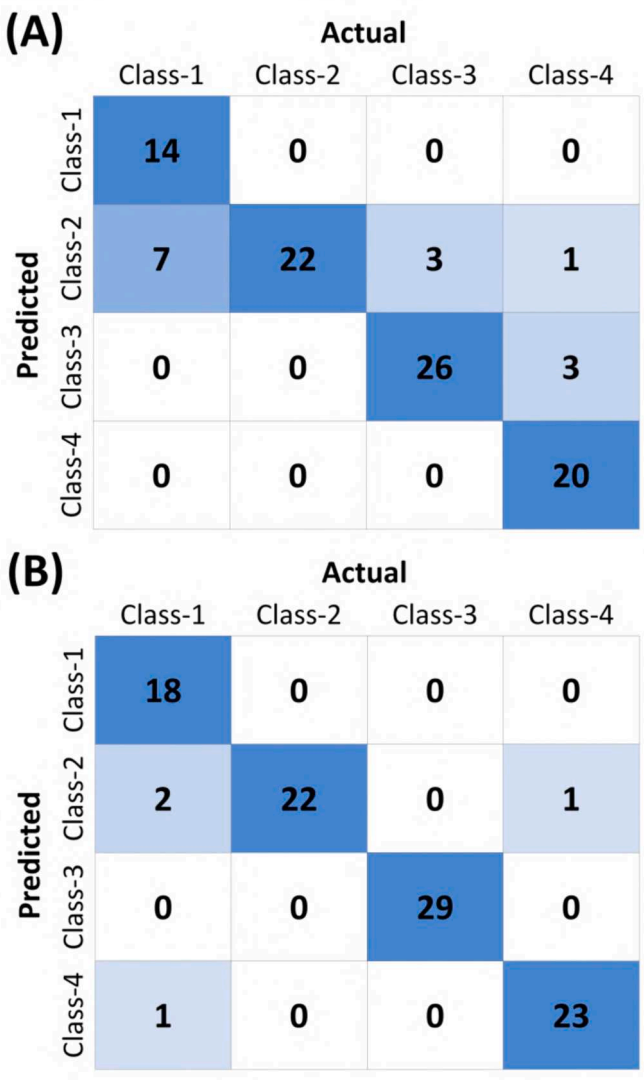


Fig. 3. Confusion matrices of the training set of (A) PLS-DA and (B) SVM-C models used for the classification of commercial products.

Regardless of the regression method employed, the selection of optimal models hinged on criteria such as RMSEP, R^2_P, and RMSEP/RMSEC ratio, associated to models with the lowest number of LVs/SVs. Notably, the application of SNV, EMSC, and SG FD individually demonstrated an enhancement in the performance of both regression methods. On the other hand, while *i*-PLS and VIP exhibited marginal improvements in PLSR outputs, they adversely affected the predictive ability of SVMR (data not shown). This finding was anticipated, given that the two variable selection methods are rooted in PLS principles and are thus compatible only with PLS. In addition to *i*-PLS and VIP, the regression coefficients (refer to Figs. S7 and S8) were employed to assess the contribution of specific wavenumbers in the regression model for the designated number of LVs. In this context, Haaland's criterion yielded highly satisfactory results, achieving this with the fewest number of factors compared to the commonly used CV variance method. All these mathematical considerations underscore the robustness of PLSR as a potent and reliable tool in linear regression, safeguarding its models from over-fitting or under-fitting issues.

Unlike PLSR, no valid criteria exist to assess data fitting in SVMR. In spite of that, maintaining the RMSECV/RMSEC ratio, R^2_{CV}, C and γ values the closest possible to one (1) can ensure good overall generalization

Table 2
Regression parameters of the calibration, cross-validation and prediction procedures for the optimal analytical models calculated by PLSR and SVMR.

Method	Parameter (unit)	Spectral pre-treatment	Selected spectral range (cm^{-1})	Kernel type Hyper-parameters (C; ε; γ)	No. of LVs/ SVs	Training		Cross-validation		Testing	
						RMSEC	R^2_C	RMSECV	R^2_{CV}	RMSEP	R^2_P
PLSR	Moisture (%)	SG FD	VIP 1703–841 721–615	–	5 *	1.04	0.6300	1.39	0.3608	1.14	0.4358
	pH	SG FD	*i*-PLS 1700–680	–	3 §	0.248	0.8311	0.297	0.7663	0.251	0.8362
	Ash (%)	EMSC	3700–2430 1860–615	–	2 §	1.33	0.4611	1.44	0.3921	1.02	0.6582
	Total NCT (mg/g, dwb)	SNV	*i*-PLS 2997–2789 1754–1502 1339–701	–	4 *§	1.27	0.7586	1.46	0.6891	1.37	0.7499
	Un-ionized NCT (mg/g, dwb)	SNV	*i*-PLS 2997–2789 1754–1502 1339–701	–	4 *§	1.26	0.7568	1.46	0.6842	1.35	0.7521
SVMR	Moisture (%)	SNV-DT	3700–2430 1860–615	Linear 0.03; 0.15	43	0.819	0.7816	1.34	0.4367	0.916	0.6258
	pH	SNV	3700–2430 1860–615	Linear 0.03; 0.08	41	0.171	0.9220	0.297	0.7606	0.168	0.9266
	Ash (%)	SNV	3700–2430 1860–615	RBF 0.1; 0.1; 0.001	43	0.710	0.8587	1.23	0.5451	0.728	0.8197
	Total NCT (mg/g, dwb)	EMSC	3700–2430 1860–615	RBF 10; 0.1; 0.001	39	0.526	0.9685	1.67	0.5898	0.835	0.9031
	Un-ionized NCT (mg/g, dwb)	EMSC	3700–2430 1860–615	RBF 10; 0.1; 0.001	42	0.521	0.9682	1.67	0.5785	0.831	0.9024

Abbreviations: C, Cost function; ε, Error margin; γ, Radial diameter, specific to SVMR; dwb, Dry weight basis; DT, De-trending; EMSC, Extended multiplicative scatter correction; i-PLS, interval-Partial least-squares; No. of LVs/ SVs, Number of latent variables or support vectors in PLSR or SVMR models, respectively; RBF, Radial basis function; SG FD, Savitzky-Golay first-order derivative; SNV, Standard normal variate; VIP, Variable importance for the projection.
* Value estimated according to the explained cross-validation variances.
§ Value estimated according to the Haaland's criterion (first model with $p < 0.75$ for the ratio $PRESS_k/\min(PRESS)$).

ability. Another important parameter that must be not ignored is the number of SVs. In this study, the best-performing models showed that 61 − 67 % of the training samples were used as SVs. Values seem relatively high when comparing with the results from Schmidtke *et al.* [13] which disclosed a maximum percentage of 54 %; however, that can be accepted for such a small and complex dataset. Subsequent critical statistics (Table 3) and plots of relationship between the predicted and reference values (Fig. 4A − J) were implemented for the purpose to evaluate the prediction accuracy in the independent testing set. Details for each parameter are as follows.

Concerning total moisture, both PLSR and SVMR did not offer satisfactory estimations of ST moistures, resulting in RMSEP > 0.91, R^2_P < 0.63, and RPD < 1.7. According to the RPD scale referred to in [44], the developed models are only suitable for rough screening purposes. From 'the predicted versus reference' plots (Figs. 4A and B), most of the points were dispersed further away from the regression lines, and the severe divergence between calibration and test fits indicate a poor predictive ability. An examination of the regression coefficient plot (Fig. S8A), searching for the spectral variations responsible to this lack of fit in PLSR, revealed uneven noise characteristics and the use of irrelevant wavenumbers in constructing the model instead of water absorption bands. One explanation for this finding can be attributed to that water bands among the spectra were arbitrarily overlapped with different interfering agents, or probably because of the fact that the reference procedure measures not only water but a sum of volatile compounds which can be relatively influential. If the latter is the case, hence investigation of alternative methods for measuring water/moisture in the ST, like those described in McAdam *et al.* [20] may give values which can better correlate with the spectra of our samples. Unusually, small values of MAPE (1.4 %) and REP (1.9 %) were obtained.

Regarding pH, the optimum result was achieved for linear SVMR with RMSEP = 0.17, $R^2_P \approx 0.93$, and RPD = 3.7. PLSR in turn also

Table 3
Evaluation metrics and figures of merit for the relevant PLSR and SVMR models calculated for the testing set.

Method	Parameter (unit)	MAPE (%)	REP (%)	RPD *	RER *	Bias	LOD_{ICH} §	LOD_{pu} ¶	LOQ_{ICH} §	LOQ_{pu} ¶
PLSR	Moisture (%)	1.8	2.3	1.3	5.4	0.22	15.5	16.3	47.1	49.0
	pH	1.86	2.43	2.5	9.9	$1.4\ 10^{-3}$	1.96	2.19	5.93	6.5
	Ash (%)	3.4	4.1	1.8	8.2	0.36	11.2	13.1	34.0	39.4
	Total NCT (mg/g, dwb)	13.4	17.1	2.0	8.9	−0.36	2.0	5.3	6.0	16.0
	Un-ionized NCT (mg/g, dwb)	13.6	12.1	2.0	9.0	−0.36	2.0	5.3	6.0	15.9
SVMR	Moisture (%)	1.4	1.9	1.7	6.7	0.16	10.7	11.3	32.5	33.8
	pH	1.29	1.63	3.7	14.8	−0.018	1.26	1.41	3.83	4.22
	Ash (%)	2.5	2.9	2.4	11.0	0.17	4.2	4.9	12.7	14.8
	Total NCT (mg/g, dwb)	9.5	10.4	3.2	14.2	−0.067	0.6	1.7	1.9	5.1
	Un-ionized NCT (mg/g, dwb)	9.5	10.5	3.2	14.2	−0.084	0.6	1.7	1.9	5.1

Abbreviations: dwb, Dry weight basis; NCT, Nicotine.
* Unitless metric.
§ Value determined according to the ICH guidelines.
¶ Value determined from the pseudo-univariate regression line.

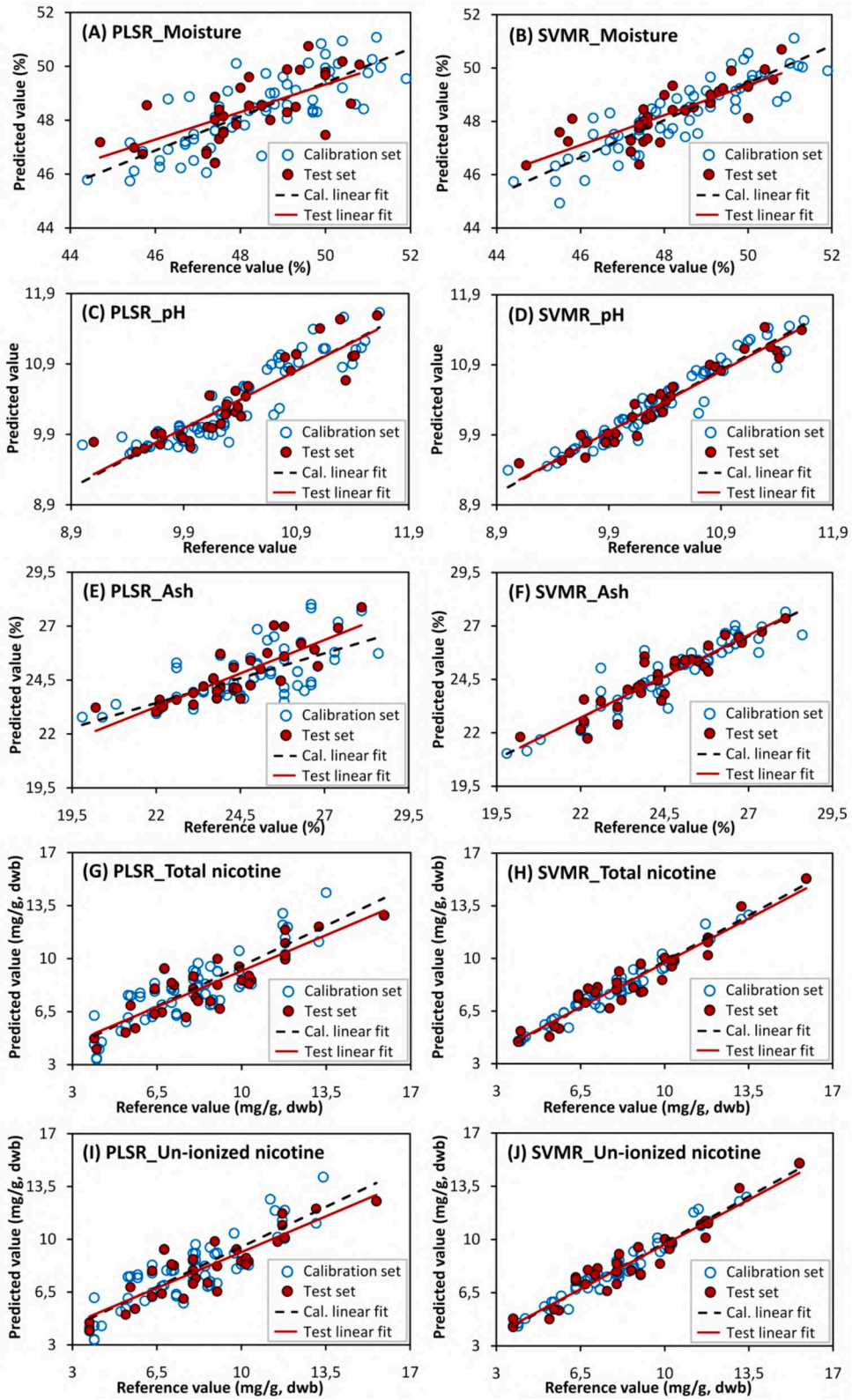


Fig. 4. Regression lines of actual versus predicted values by PLSR (A, C, E, G and I) and SVMR (B, D, F, H and J) of moisture, pH, ashes, total nicotine and free nicotine contents, respectively.

presented very good outputs in the SG FD-processed *i*-PLS range 1700 − 680 cm^{-1} giving RMSEP = 0.25, $R_p^2 \approx 0.84$, and RPD = 2.5 using three factors. The perfect overlay of the two fitting lines to each other (Figs. 4C and D) with MAPE and REP values of the same low order of magnitude imply high prediction capability. The RPD scale mentioned above indicates that the SVMR model can be successfully applied in process control to predict pH level. From the PLS regression coefficients, it can be seen positive correlations to the CO_3^{2-} ion bands (at 1315, 854, and 717 cm^{-1}) and negative correlations to the amide II in plant (1540 cm^{-1}) as well as to the SiO_2 vibrations in kieselguhr (1168 and 1063

cm^{-1}) in the derivative spectra, logically linking between the measured pH and the proportion of $CaCO_3$ (alkaline modifier) within the product ingredients.

For total ash content, PLSR offered somewhat higher RMSEP (1.0), $R_P^2 \approx 0.66$, and RPD = 1.8 using only two LVs. Herein, SVMR based on Radial basis function (RBF) obviously enhanced the performance with RMSEP (0.73) decreased by 28.6 % and R_P^2 (0.82) increased by 24.5 %, while the RPD became 2.4 which means fair precision and suitability of the model for screening purposes. The quality of the regression lines 'predicted vs. actual' demonstrated a satisfactory goodness of fit for SVMR compared to PLSR that showed residual divergence of the testing set line away from the calibration line as depicted in Fig. 4E and F. Notice too that lower values of MAPE and REP were obtained for the SVMR model. Fig. S8C in supplementary data shows that the broad peak centered at 1120 cm^{-1}, which overlaps and obscures the Si − O stretching of silica at nearly 1055 cm^{-1}, is a significant contributor to predicting ash content; reconfirming the keen relationship between the current parameter and kieselguhr amount in the samples. It is noteworthy to observe slight shifts in peak positions attributed to matrix effects in that plot and others.

Since total and un-ionized nicotine presented comparable outputs for our samples, they were collectively discussed as a single parameter. Although exhibited a relatively higher RMSEP (0.84), SVMR RBF-based models demonstrated acceptable predictive capabilities with R_P^2 of approximately 0.90 and RPD of 3.2. In comparison, PLSR yielded RMSEP of 1.4, R_P^2 of 0.75, and RPD of 2.0. In contrast to ashes, the nicotine concentrations in the examined samples disclosed greater variability (SD = 2.6). This variability supports the suitability of SVMR models for quality control applications, considering the aforementioned RPD scale. Interestingly, smaller RMSECV values were obtained for PLSR. This suggests plausible slight over-fitting issue in SVMR, nevertheless, excellent correlations between the predictions of SVMR and the wet method values could be seen in Fig. 4G − J. An examination of the regression coefficient plots (supplementary Figs. S8D and E) show that the band around 1540 cm^{-1}, attributable to amide II, emerged as an important variable for identifying the specific tobacco variety used in the samples. Smaller contributions were observed from the bands at 1150 and 1080 cm^{-1}, which correspond to cellulose in tobacco leaves. This finding is consistent with the known composition of tobacco, which inherently contains nicotine, meaning that the presence and amount of tobacco plant material in the sample would directly impact the concentration of nicotine. The peak heights at 1640 and 1320 cm^{-1}, already linked to the presence of water and amide I/lignin vibrations, respectively, were negatively loaded across the used factors. MAPEs decreased from 13.6 % to 9.5 %, and REPs decreased from 17.1 % to 10.4 % when employing SVMR. However, it is important to note that these values, although improved, still exceed the desired looking-forward values.

For a more in-depth assessment of the accuracy and precision of the optimal models, EJCR test was recommended to check for the presence of bias rather than the other individual tests such as analyte recovery values [38]. As shown in Fig. 5A and B, moisture bears the wider and furthest confidence regions from the ideal point, thereby implying their non-suitability for estimation of moistures in the products. Regarding the remaining four parameters, SVMR showed confidence regions that were both narrower and closer to the (1, 0) point. Nevertheless, none of the ellipses encompassed the ideal point. In such situation and to assume that the predicted values from these models are significantly biased, subsequent EJCRs based on weighted least-squares (WLS) or bilinear least-squares (BLS) should be performed [45]. Unfortunately, due to the limited quantity of some samples, certain parameters were analyzed just once resulting in missing SD values which prevent from conducting the latter methods.

In addition, Table 3 shows the LOD and LOQ estimated according to the two analytical approaches introduced in **Subsection 2.6**. Calculated

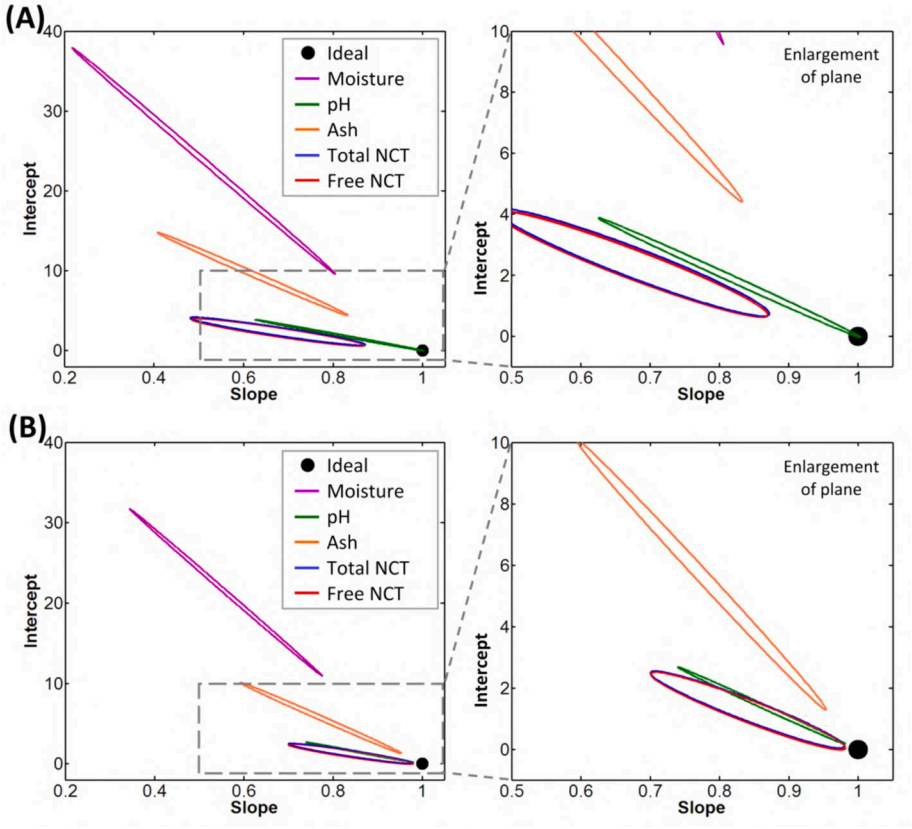


Fig. 5. EJCRs in the slope-intercept plane performed for the optimal models of (A) PLSR and (B) SVMR based on ordinary least-squares. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article).

using the approach proposed by Allegrini & Olivieri [35], the LOD ranges [min – max] for the best PLSR models were 12.3 – 12.4 %, 1.77 – 1.80, 7.6 – 7.7 %, 1.7 – 2.5 mg/g dry weight basis (dwb), and 1.7 – 2.4 mg/g dwb, respectively, for moisture, pH, ashes, total nicotine, and free nicotine. These results compare favorably with the reported LOD_{ICH}, whereas the LOD_{pu} was noticeably over-estimated. This overestimation was anticipated given the finite uncertainties in the predicted values. However, this does not invalidate the latter approach for other instances. Logically, lower LODs and LOQs were achieved in SVMR models. The obtained outcomes from EJCR analysis and LOD/LOQ values were overall in alignment with the other evaluation metrics, which demonstrate the stability and robustness of the proposed methodology.

Lastly, the constructed SVMR models accurately predicted four quality parameters, specifically pH, ashes, total nicotine, and free nicotine in oral tobaccos from Algerian markets. Nonetheless, it is worth noting the challenges posed by the complex and diverse ingredients in Chemma formulations, significant moisture interference, and the relatively small dataset, which may have impacted the performances of models. Therefore, future studies will be conducted with larger and potentially more diverse dataset, employing effective calibration transfer technique to address moisture interference, and exploring more advanced machine learning or deep learning frameworks to improve the models' performance and encompass additional quality parameters.

4. Conclusions

This paper introduces, for the first time, a novel methodology for the rapid survey of ST products using mid-infrared spectroscopy coupled with mathematical-theoretical and machine learning frameworks. By leveraging spectral features, we successfully characterized the presence of different qualities of tobacco leaves within samples and identified the presence of calcium carbonate and kieselguhr as major ingredients in the Algerian Chemma.

PCA, AHC, and *k*-means clustering were implemented as complementary unsupervised learning techniques to group commercial samples into four distinct clusters based on physicochemical reference measurements. Subsequent application of supervised SVM-C, using the outputs from *k*-means clustering, achieved 95.8 %/84.4 % calibration/validation accuracies, effectively discriminated between commercial products based on their SNV-processed spectra. This qualitative information equips regulatory authorities to identify and track suspected products that may pose health risks to consumers.

The combination of ATR-FTIR spectroscopy with SVMR facilitated the simultaneous determination of pH, ashes, total nicotine, and unionized nicotine contents through simple measurements. The achieved outcomes, with $R^2_P \geq 0.82$ and RPD ≥ 2.4, allowed the optimal models to be successfully integrated into the quality control of commercial products. Additionally, the consistency across all evaluation metrics underscores high stability, reliability, and robustness.

The proposed methodology significantly reduced the required amount and time for sample analysis compared to reference procedures. However, a drawback of the method is that it generates predictive models with partial accuracy, emphasizing the necessity for further development to achieve optimal performance.

CRediT authorship contribution statement

Mohamed Fekhar: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Yasmina Dagbouche:** Project administration, Supervision, Writing – review & editing. **Naima Bouzidi:** Funding acquisition, Supervision, Writing – review & editing. **Mohamed El Hattab:** Resources, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data that has been used is confidential.

Acknowledgements

The authors acknowledge the financial support of research project with economic and social impact (N° 007 of March 15, 2020) and University-Education Research Projects (PRFU/ B00L01UN0901 20220001).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.microc.2024.110670.

References

[1] U.S. Food and Drug Administration, Family Smoking Prevention and Tobacco Control Act - An Overview, 2020. https://www.fda.gov/tobacco-products/rules-regulations-and-guidance/family-smoking-prevention-and-tobacco-control-act-overview/. (Accessed 01 October 2023).
[2] Centers for Disease Control and Prevention, Notice regarding revisions to the laboratory protocol to measure the quantity of nicotine contained in smokeless tobacco products manufactured, imported, or packaged in the United States, Federal Register, 74 (2009) 712–719.
[3] L. Wu, B. Wang, L. Zhang, R. Duan, R. Gao, Y. Yin, X. Liu, X. Bai, Determination of routine chemicals, physical indices and macromolecular substances in reconstituted tobacco using near infrared spectroscopy combined with sample set partitioning, J. of NIR Spectrosc. 28 (2020) 153–162, https://doi.org/10.1177/0967033520905371.
[4] Y. Geng, H. Shen, H. Ni, Y. Tian, Z. Zhao, Y. Chen, X. Liu, Non-destructive determination of total sugar content in tobacco filament based on calibration transfer with parameter free adjustment, Microchem. J. 181 (2022) 107797, https://doi.org/10.1016/j.microc.2022.107797.
[5] R. Shu, L. Ju, L. Ni, S. Wu, L. Zhang, J. Ge, S. Ye, S. Luan, Improving transferability and service life of the calibration model of total plant alkaloids in tobacco leaves on seven NIR spectroscopy devices by multi-step wavelength selection methods, Microchem. J. 196 (2024) 109522, https://doi.org/10.1016/j.microc.2023.109522.
[6] D. Jiang, G. Hu, G. Qi, N. Mazur, A fully convolutional neural network-based regression approach for effective chemical composition analysis using near-infrared spectroscopy in cloud, J. of Artif. Intell. and Technol. 1 (2021) 74–82, https://doi.org/10.37965/jait.2020.0037.
[7] Z. Zhu, G. Qi, Y. Lei, D. Jiang, N. Mazur, Y. Liu, D. Wang, W. Zhu, A long short-term memory neural network based simultaneous quantitative analysis of multiple tobacco chemical components by near-infrared hyperspectroscopy images, Chemosensors 10 (2022) 164, https://doi.org/10.3390/chemosensors10050164.
[8] Y. Zhang, Q. Cong, Y. Xie, B. Zhao, Quantitative analysis of routine chemical constituents in tobacco by near-infrared spectroscopy and support vector machine, Spectrochimica Acta Part A: Mol. and Biomol Spectrosc. 71 (2008) 1408–1413, https://doi.org/10.1016/j.saa.2008.04.020.
[9] J. Cruz-Tirado, R.L. de França, M. Tumbajulca, G. Barraza-Jáuregui, D.F. Barbin, R. Siche, Detection of cumin powder adulteration with allergenic nutshells using FT-IR and portable NIRS coupled with chemometrics, J. of Food Compos. and Anal. 116 (2023) 105044, https://doi.org/10.1016/j.jfca.2022.105044.
[10] S.B. Stanfill, G.N. Connolly, L. Zhang, L.T. Jia, J.E. Henningfield, P. Richter, T. S. Lawler, O.A. Ayo-Yusuf, D.L. Ashley, C.H. Watson, Global surveillance of oral tobacco products: total nicotine, unionised nicotine and tobacco-specific N-nitrosamines, Tob. Control 20 (2010) 1–10, https://doi.org/10.1136/tc.2010.037465.
[11] S.B. Stanfill, A.L.O. da Silva, J.G. Lisko, T.S. Lawler, P. Kuklenyik, R.E. Tyx, E. H. Peuchen, P. Richter, C.H. Watson, Comprehensive chemical characterization of Rapé tobacco products: Nicotine, un-ionized nicotine, tobacco-specific N′-nitrosamines, polycyclic aromatic hydrocarbons, and flavor constituents, Food and Chem. Toxicol. 82 (2015) 50–58, https://doi.org/10.1016/j.fct.2015.04.016.
[12] S.B. Stanfill, R.E. Croucher, P.C. Gupta, J.G. Lisko, T.S. Lawler, P. Kuklenyik, M. Dahiya, B. Duncan, J.B. Kimbrell, E.H. Peuchen, C.H. Watson, Chemical characterization of smokeless tobacco products from South Asia: Nicotine, unprotonated nicotine, tobacco-specific N'-Nitrosamines, and flavor compounds, Food and Chem. Toxicol. 118 (2018) 626–634, https://doi.org/10.1016/j.fct.2018.05.004.

[13] L.M. Schmidtke, J.P. Smith, M.C. Müller, B.P. Holzapfel, Rapid monitoring of grapevine reserves using ATR–FT-IR and chemometrics, Analytica Chimica Acta 732 (2012) 16–25, https://doi.org/10.1016/j.aca.2011.10.055.
[14] S. Bureau, B. Quilot-Turion, V. Signoret, C. Renaud, M. Maucourt, D. Bancel, C. M. Renard, Determination of the composition in sugars and organic acids in peach using mid infrared spectroscopy: comparison of prediction results according to data sets and different reference methods, Anal. Chem. 85 (2013) 11312–11318, https://doi.org/10.1021/ac402428s.
[15] S. Materazzi, A. Gregori, L. Ripani, A. Apriceno, R. Risoluti, Cocaine profiling: Implementation of a predictive model by ATR-FTIR coupled with chemometrics in forensic chemistry, Talanta 166 (2017) 328–335, https://doi.org/10.1016/j.talanta.2017.01.045.
[16] W. Lan, V. Baeten, B. Jaillais, C.M. Renard, Q. Arnould, S. Chen, A. Leca, S. Bureau, Comparison of near-infrared, mid-infrared, Raman spectroscopy and near-infrared hyperspectral imaging to determine chemical, structural and rheological properties of apple purees, J. of Food Eng. 323 (2022) 111002, https://doi.org/10.1016/j.jfoodeng.2022.111002.
[17] M. Ali, Re: What is the exact preparation method of double distilled water for electrochemical experiments?, 2016. https://www.researchgate.net/post/What_is_the_exact_preparation_method_of_double_distilled_water_for_electrochemical_experiments/. (Accessed 15 June 2022).
[18] J. Wang, H. Yang, H. Shi, J. Zhou, R. Bai, M. Zhang, T. Jin, Nitrate and nitrite promote formation of tobacco-specific nitrosamines via nitrogen oxides intermediates during postcured storage under warm temperature, J. of Chem. 2017 (2017) 6135215, https://doi.org/10.1155/2017/6135215.
[19] S.L. Tomar, J.E. Henningfield, Review of the evidence that pH is a determinant of nicotine dosage from oral use of smokeless tobacco, Tob. Control 6 (1997) 219–225, https://doi.org/10.1136/tc.6.3.219.
[20] K.G. McAdam, H. Kimpton, A. Faizi, A. Porter, B. Rodu, The composition of contemporary American and Swedish smokeless tobacco products, BMC Chem. 13 (2019) 1–15, https://doi.org/10.1186/s13065-019-0548-0.
[21] M. Fekhar, Y. Daghbouche, N. Bouzidi, M. El Hattab, ATR-FTIR spectroscopy combined with chemometrics for quantification of total nicotine in Algerian smokeless tobacco products, Microchem. J. 193 (2023) 109127, https://doi.org/10.1016/j.microc.2023.109127.
[22] A. Kumar, D. Bhartiya, J. Kaur, S. Kumari, H. Singh, D. Saraf, D.N. Sinha, R. Mehrotra, Regulation of toxic contents of smokeless tobacco products, The Indian J. of Med. Res. 148 (2018) 14, https://doi.org/10.4103/ijmr.IJMR_2025_17.
[23] A.M. Idris, S.O. Ibrahim, E.N. Vasstrand, A.C. Johannessen, J. Lillehaug, B. Magnusson, M. Wallström, J.-M. Hirsch, R. Nilsen, The Swedish snus and the Sudanese toombak: are they different? Oral Oncol. 34 (1998) 558–566.
[24] A. Tharwat, Principal component analysis-a tutorial, Int. J. of App. Pattern Recognit. 3 (2016) 197–240, https://doi.org/10.1504/IJAPR.2016.079733.
[25] J. Shlens, A tutorial on principal component analysis, arXiv preprint, arXiv: 1404.1100 (2014).
[26] F. Nielsen, Hierarchical clustering, Introduction to HPC with MPI for Data, Science (2016) 195–211, https://doi.org/10.1007/978-3-319-21903-5_8.
[27] F. Murtagh, P. Contreras, Algorithms for hierarchical clustering: an overview, Wires Data Mining Knowl. Discov. 2 (2011) 86–97, https://doi.org/10.1002/widm.53.
[28] J. Miller, J.C. Miller, Statistics and chemometrics for analytical chemistry, 6 ed., Pearson education2018.
[29] D. Sisodia, L. Singh, S. Sisodia, K. Saxena, Clustering techniques: a brief survey of different clustering algorithms, Int. J. of Latest Trends in Eng. and Technol. 1 (2012) 82–87.
[30] X. Chu, Y. Huang, Y.-H. Yun, X. Bian, Chemometric methods in analytical spectroscopy technology, Springer (2022), https://doi.org/10.1007/978-981-19-1625-0.
[31] A.C. Olivieri, Introduction to multivariate calibration: a practical approach, Springer (2018), https://doi.org/10.1007/978-3-319-97097-4.
[32] K. Cheng, Z. Lu, Y. Zhou, Y. Shi, Y. Wei, Global sensitivity analysis using support vector regression, App. Math. Model. 49 (2017) 587–598, https://doi.org/10.1016/j.apm.2017.05.026.
[33] R. Ortaç-kabaoğlu, A support vector regression method for reducing the high-order systems to first-order plus time-delay forms, IU-J. of Electr. & Electron. Eng. 11 (2012) 1305–1309.
[34] A.J. Smola, B. Schölkopf, A tutorial on support vector regression, Stat. and Comput. 14 (2004) 199–222, https://doi.org/10.1023/B:STCO.0000035301.49549.88.
[35] F. Allegrini, A.C. Olivieri, IUPAC-consistent approach to the limit of detection in partial least-squares calibration, Anal. Chem. 86 (2014) 7858–7866, https://doi.org/10.1021/ac501786u.
[36] M. Ortiz, L. Sarabia, A. Herrero, M. Sánchez, M. Sanz, M. Rueda, D. Giménez, M. Meléndez, Capability of detection of an analytical method evaluating false positive and false negative (ISO 11843) with partial least squares, Chemom. and Intell. Lab. Syst. 69 (2003) 21–33, https://doi.org/10.1016/S0169-7439(03)00110-2.
[37] International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use, Validation of analytical procedures: text and methodology, Q2 (R1), 1 (2005) 05.
[38] A.G. González, M.A. Herrador, A.n.G. Asuero, , Intra-laboratory testing of method accuracy from recovery assays, Talanta 48 (1999) 729–736, https://doi.org/10.1016/S0039-9140(98)00271-9.
[39] A.C. Olivieri, Practical guidelines for reporting results in single-and multi-component analytical calibration: a tutorial, Analytica Chimica Acta 868 (2015) 10–22, https://doi.org/10.1016/j.aca.2015.01.017.
[40] A. Largo-Gosens, M. Hernández-Altamirano, L. García-Calvo, A. Alonso-Simón, J. Álvarez, J.L. Acebes, Fourier transform mid infrared spectroscopy applications for monitoring the structural plasticity of plant cell walls, Frontiers in Plant Sci. 5 (2014) 303, https://doi.org/10.3389/fpls.2014.00303.
[41] B.H. Stuart, Infrared spectroscopy: fundamentals and applications, John Wiley & Sons (2004), https://doi.org/10.1002/0470011149.
[42] V.H.J.M. dos Santos, D. Pontin, G.G.D. Ponzi, A.S.d.G. e Stepanha, R.B. Martel, M. K. Schütz, S.M.O. Einloft, F. Dalla Vecchia, Application of Fourier Transform infrared spectroscopy (FTIR) coupled with multivariate regression for calcium carbonate (CaCO3) quantification in cement, Constr. and Build. Mater. 313 (2021) 125413, https://doi.org/10.1016/j.conbuildmat.2021.125413.
[43] S. Gok, M. Severcan, E. Goormaghtigh, I. Kandemir, F. Severcan, Differentiation of Anatolian honey samples from different botanical origins by ATR-FTIR spectroscopy using multivariate analysis, Food Chem. 170 (2015) 234–240, https://doi.org/10.1016/j.foodchem.2014.08.040.
[44] P. Williams, The RPD statistic: a tutorial note, NIR News 25 (2014) 22–26, https://doi.org/10.1255/nirn.1419.
[45] V.G. Franco, V.E. Mantovani, H.C. Goicoechea, A.C. Olivieri, Teaching chemometrics with a bioprocess: analytical methods comparison using bivariate linear regression, The Chem. Educator 7 (2002) 265–269, https://doi.org/10.1007/s00897020596a.