ACADÉMIE DE MONTPELLIER

UNIVERSITÉ DE MONTPELLIER

MASTER
STATISTIQUE POUR LES **S**CIENCES DE LA **V**IE

MÉMOIRE sur le stage :

Développement d'outils d'analyses pour les données de PTR-ToF-MS

Du 1er mars au 31 août 2021

au Centre d'Écologie Fonctionnelle & Évolutive

sous la direction de N. Barthes et J.-M. Roger

par :

JORIS HUGUENIN

soutenu le 9 septembre 2021, devant la commission d'examen :

J.-N. BACRO	Professeur
C. REYNES	Maître de Conférence
R. SABATIER	Professeur

À travers ce rapport, je dédicace cette année de M2 à Thomas M. Kyle, lauréat du prix Ignobel 1991 pour la découverte de l'élément le plus lourd de l'Univers : l'Administratium. Je salue les efforts cumulés mis en oeuvre pour honorer le trentième anniversaire de ce prix et remercie ceux qui ont su faire preuves de patience.

https://www.cefe.cnrs.fr/fr/

Table des matières

Remerciements

Je tenais à remercier Jean-Michel Roger et Nicolas Barthes qui ont bien voulu accepter d'encadrer ce stage. Une chose me dit qu'on a pas fini de faire de la chimiométrie avec tous nos machin-MS.

Pour continuer sur la chimiométrie, je remercie la team de Chemhouse et en particulier Silvia Mas-Garcia et Douglas Rutledge (et leur très léger accent respectif) pour leurs aides et conseils.

Je salue également les autres membres de la PACE : Anne-Genevieve, Raphaelle, Bruno, Patrick et Benoit ; pour la sympathie au quotidien. Il en va de même pour tous les membres du CEFE[1] avec qui je travaille régulièrement ou avec qui je partage des moments chaleureux. Mes pensées vont plus précisément vers Magali, Candice, Cao Li et Anjélica. Et merci aux relecteurs de ce rapport.

Ce rapport conclu un semestre de stage mais également une année de formation continue. Avant septembre 2020, mes dernières heures de cours en tant qu'étudiant dataient de décembre 2011. Depuis cette date, j'ai moi-même enseigné en tant que vacataire à l'IUT durant mes années de thèses. Puis après d'autres contrats, j'ai été embauché à l'Université de Montpellier comme ingénieur. Ainsi en septembre 2020, j'étais tout à la fois néo et ancien étudiant, camarade et collègue. Pour moi qui n'était que de passage, la difficulté des cours à distance m'a affecté mais ne remettait pas en cause ni mon métier principal ni mon cursus universitaire. Pour autant je félicite l'effort des professeurs qui ont su s'adapter avec le peu de moyens mis à leur disposition. Je salue tout autant les étudiants en formation initiale qui jouent leur avenir professionnel.

Enfin, je remercie les copains qui m'ont soutenu pendant l'année. Je déconseille vivement d'accumuler reprise d'étude + travail + confinement/couvre feu + concours de la fonction publique + travaux de rénovation. C'est trop. En particuliers, merci à Ryad, Jean-Michel et Nicolas (encore eux) pour leur aide lors des concours et à Bruce Dickinson, Steve Harris et les autres pour m'avoir accompagné durant l'écriture des multiples rapports.

Merci aussi à ceux qui m'ont ajouté du travail durant cette année : Ngoc qui fut une brillante élève de conduite supervisée et qui n'a failli nous tuer que deux fois, à Batro pour avoir eu la brillante idée de synchroniser la date de sortie de notre jeu[AB20] avec la date du deuxième confinement, et aux créatures de la nuit qu'il faut nourrir d'idées nouvelles chaque semaine.

Enfin merci à Hélène pour me supporter depuis tout ce temps, pour son soutien et pour son aide.

[1] et de Boris qui doit se contenter d'une note de bas de page.

Glossaire

AUC : Area Under the Curve. Calcul de l'aire sous la courbe.

CEFE (UMR 5175) : Centre d'Écologie Fonctionnelle et Évolutive. Laboratoire d'accueil situé sur le site de la DR13 du CNRS, route de Mende à Montpellier.

COV (ou VOC) : Composé Organique Volatil. Molécule produite par des systèmes biologiques et analysée par le PTR-ToF-MS.

Gaz zéro : gaz purifié grâce à un filtre à charbon.

GC-MS : Gaz Chromatography Mass Spectrometry. Autre méthode d'analyse des COV, moins sensible mais plus discriminante que la PTR-ToF-MS.

ICA : Independent Component Analysis. Méthode de calcul permettant de séparer les sources indépendantes.

JADE : Joint Approximation Diagonalization of Eigen-matrices. Algorithme itératif utilisant le principe de l'ICA

MCR-ALS : Multivariate Curve Resolution - Alternating Least Squares. Méthode de calcul permettant de retrouver les spectres purs.

PACE : Plateforme d'Analyses Chimiques en Écologie. Plateforme située au sein de l'UMR 5175.

PCA : Principal Component Analysis (ou ACP). Algorithme utilisé pour faire ressortir la variabilité entre les échantillons.

PTR-ToF-MS : Proton Transfert Reaction Time of Flight Mass Spectrometry. Acronyme utilisé pour l'instrument de mesure (l**e** PTR-ToF-M**Spectrometer**) et pour la méthode d'analyse (l**a** PTR-ToF-M**Spectrometry**). Par abus de langage, l'acronyme PTR-MS est régulièrement utilisé à l'oral.

Introduction générale

À l'instar de l'année 2021, ce stage de M2 a été un peu particulier. Lors de l'élaboration du sujet en janvier, nous avons proposé un sujet résolument tourné vers l'analyse de données. Il s'est avéré que mon temps de stage s'est réparti différemment. Le sujet du stage a évolué afin de mieux couvrir l'ensemble du travail effectué durant ces mois. Le chapitre 1 rend compte de ces aspects en détaillant le contexte de ce stage, l'ouverture qui a été amorcée afin de rendre compatible le travail de la plateforme avec les orientations publiques vers la science ouverte et de la relation entre les analystes et les praticiens.

Le chapitre 2 est dédié à l'instrument de mesure principalement utilisé lors du stage. Le PTR-ToF-MS analyse les molécules volatiles. Après une présentation de cette technique et de notre appareil, je détaillerai les données récupérées ainsi que le package R que j'ai écrit pour procéder à l'analyse. Nous ferons un comparatif entre cette technique et la GC-MS qui est régulièrement utilisée en parallèle.

Enfin, le chapitre 3 détaille les opérations mathématiques utilisées pour l'analyse des données. Je développerai les choix faits quant aux prétraitements et à l'alignement des spectres. Je présenterai également les algorithmes implémentés dans le package, en particulier celui de la MCR-ALS.

Avant l'introduction se trouve un glossaire des termes techniques et des acronymes. À la fin de ce rapport, je livre une conclusion générale sur ce stage ainsi que sur le travail qu'il reste à accomplir. Un soin particulier a été apporté à la découpe des chapitres. Bien que la mode des rapports et articles scientifiques soit au storytelling, les trois chapitres suivants peuvent se lire dans l'ordre d'intérêt du lecteur.

CHAPITRE 1

Le métier d'ingénieur de plateforme

1.1 Contexte

Depuis le 04 novembre 2019, j'occupe un poste d'ingénieur affecté à la Plateforme d'Analyses Chimiques en Écologie (PACE). La PACE, créée en 2000, est spécialisée dans l'analyse chimique pour la communauté de recherche liée à l'écologie, l'environnement et la biodiversité. La PACE comporte six permanents et un CDD. Cette plateforme est un service mutualisé du LabEx CeMEB depuis 2011. L'unité de rattachement est le Centre d'Ecologie Fonctionnelle et Evolutive de Montpellier (CEFE UMR5175 du CNRS). Cette unité mixte possède quatre tutelles : le CNRS, l'Université de Montpellier (UM), l'École Pratique des Hautes Etudes (EPHE) et l'Institut de Recherche pour le Développement (IRD), ainsi que trois partenaires : l'Université Paul Valéry Montpellier 3 (UPVM3), SupAgro Montpellier et l'INRAE. À ce titre, bien que rattachée au CEFE, la PACE accueille des projets de recherche issus des 12 unités du CeMEB (850 personnels permanents) mais aussi de toute la communauté académique nationale ou internationale.

La PACE, à travers sa tutelle Université de Montpellier, a obtenu un financement GEPETOs dans le cadre du Contrat de Plan Etat-Région (CPER) 2015-2020 pour l'achat d'un spectromètre de masse en temps réel permettant d'atteindre des mesures de cinétique fine, même à l'état de trace, abrégé en PTR-ToF-MS (Proton Transfert Reaction Time of Flight Mass Spectrometer) pour la PACE. Le chapitre 2 revient en détail sur cet instrument. Ce projet GEPETOs comporte une part dédiée au recrutement d'un ingénieur d'étude devant bénéficier d'une formation en double compétence de niveau Master 2. Avec mes responsables hiérarchiques, nous avons choisi le M2 SSV pour me permettre de renforcer mes connaissances statistiques. Le stage du second semestre s'est donc effectué sur la plateforme PACE.

Plusieurs étapes ont été définies pour l'ensemble du contrat :

- mise en place des procédures de recueil et de contrôle des données ;
- adaptation des méthodes d'analyses mathématiques pour répondre aux besoins spécifiques du PTR-ToF-MS ;
- organisation de la mise en forme et du stockage des données ;
- assurer la maintenance des bases de données contenant les data produites par l'instrument et les résultats des analyses.

Il était prévu que le point 2 soit particulièrement mis en avant lors de ce stage. En réalité

l'ensemble des points ci-dessus a été mobilisé. Cette introduction du contexte me permet de décrire l'état d'esprit général lié à ce stage. Effectué sur mon lieu de travail, il y avait une dualité entre la nécessité d'encadrer les utilisateurs de la plateforme et la possibilité offerte d'un temps de réflexion et d'exploration. La première modalité imposait le tempo, du pragmatisme et un calendrier à respecter. Ce printemps-été 2021 était particulièrement attendu pour deux raisons. Il fallait en quelque sorte rattraper le travail qui n'avait pu être effectué lors du confinement général du printemps 2020. De plus, les possibilités de l'appareil commencent à être connues et celui-ci est beaucoup demandé. De plus, l'intitulé initial de ce stage était *"Analyse descriptive des données PTR-ToF-MS des COV émis lors du cycle larvaire de la guêpe de l'amande (*Eurytoma amygdali*)*. Je devais accompagner une doctorante dans son travail d'analyse. Nous avons commencé les expériences dès février lors de la période de floraison des amandiers. Après deux mois d'expériences, la quantité de données recueillies était particulièrement importante. La doctorante avait débuté sa thèse en janvier mais le timing expérimental est imposé par la nature. Après le travail de terrain, elle a pu reprendre un rythme normal de thèse et effectuer le travail bibliographique nécessaire à tout projet scientifique.

De mon coté, j'ai participé à la mise en place d'autres expériences sur lesquelles nous reviendrons dans ce rapport. J'ai également pu prendre du recul sur mon activité et la repenser afin qu'elle cadre avec la politique actuelle de science ouverte. Ce stage m'a ainsi permis de renforcer la partie de science reproductible des outils numériques que je conçois pour mes collègues chercheurs.

1.2 Science ouverte

Durant cette période de stage, le gouvernement français a mis en place le Deuxième Plan pour la science ouverte qui définit les actions mises en place sur la période 2021-2024. Le plan est détaillé dans ce document dont je recommande vivement la lecture. Ce guide de route possède les défauts propres à l'époque[1]. Ceci étant, à l'échelle de la communauté scientifique française, ce plan permet une large diffusion du savoir et de la culture scientifique grâce à des actions regroupées dans quatre axes :

- généraliser l'accès ouvert aux publications ;
  - généraliser l'obligation de publication en accès ouvert des articles et livres issus de recherches financées par des fonds publics ;
  - soutenir des modèles économiques d'édition en accès ouvert sans frais de publication pour les auteurs ;
  - favoriser le multilinguisme[2] et la circulation des savoirs scientifiques par la traduction des publications des chercheurs français ;
- structurer, partager et ouvrir les données de la recherche ;
  - mettre en œuvre l'obligation de diffusion des données de recherche financées sur fonds publics ;
  - créer Recherche Data Gouv, la plateforme nationale fédérée des données de la recherche ;

[1] pêle-mêle l'utilisation abondante de mots valises ou la volonté de récompenser les bons élèves par des badges et des prix.

[2] *«Découvrir l'étrangeté d'une pensée en langue étrangère constitue une expérience herméneutique fondamentale, étant donné qu'une telle rencontre peut déclencher un processus qui contribue à augmenter l'incertitude positive et à remettre en cause les convictions propres pour construire ainsi une barrière contre l'ethnocentrisme dans la pensée scientifique.»* Hamel [Ham13]

  - promouvoir l'adoption d'une politique de données sur l'ensemble du cycle des données de la recherche, pour les rendre faciles à trouver, accessibles, interopérables et réutilisables (FAIR) ;

- ouvrir et promouvoir les codes sources produits par la recherche ;

  - valoriser et soutenir la diffusion sous licence libre des codes sources issus de recherches financées sur fonds publics ;
  - mettre en valeur la production des codes sources de l'enseignement supérieur, de la recherche et de l'innovation ;
  - définir et promouvoir une politique en matière de logiciels libres ;

- transformer les pratiques pour faire de la science ouverte le principe par défaut.

  - développer et valoriser les compétences de la science ouverte tout au long du parcours des étudiants et des personnels de la recherche ;
  - valoriser la science ouverte et la diversité des productions scientifiques dans l'évaluation des chercheurs et enseignants-chercheurs, des projets et des établissements de recherche ;
  - tripler le budget de la science ouverte en s'appuyant sur le Fonds national pour la science ouverte et le Programme d'investissements d'avenir.

Les axes concernant le code et les données me concernent directement. Par mon travail, je produis des données ainsi que le code permettant l'analyse de celles-ci par mes collaborateurs. Le code est écrit en langage R sous la forme d'un package, *proVOC*, déposé sur gitHub® [3] . Il génère un workflow qui sauvegarde les paramètres et les données utilisées lors de l'analyse. Le dépôt sur un git permet l'ouverture de ce code. Il est toutefois prévu d'effectuer une migration sur gitLab afin de favoriser l'utilisation des logiciels libres.

Par ailleurs, je souhaite développer un module permettant de convertir les data brutes dans un format adapté à la future plateforme Recherche Data Gouv. Ce module facilitera la production de *data papers* dont j'ai découvert l'existence au cours des journées casuHAL 2021. Grâce à deux demi-journées de formation sur Rmarkdown et sur la création de package, j'ai pu intégrer la génération de rapports automatisés dans le package[4] qui est un premier pas vers l'écriture automatisée des data papers.[5]

J'ai également découvert et testé les packages et templates permettant de créer sous R un document découpé en chapitre. J'utilisais déjà Rmarkdown depuis un certains temps mais le passage à Bookdown et memoiR a demandé un temps d'adaptation plus long que prévu. Ceci étant, ça n'a pas été en vain puisque j'ai eu plusieurs retours enthousiastes de collègues en vue d'une formation sur ces packages. J'envisage, une fois ce rapport terminé, de développer un clone de memoiR (avec l'accord de l'auteur) sur le style de EcoFoG afin de mettre à disposition aux membres du CEFE un template clé en main. Ceci permettra de favoriser l'écriture de rapports par des logiciels libres et d'améliorer la reproductibilité de la science.

1.3 Plan d'expérience en Écologie chimique

Le chapitre suivant décrit la technique de la PTR-ToF-MS. Pour ce chapitre, nous avons juste besoin de savoir que cette technique analyse la cinétique d'émission des

[3] GitHub a été acheté par Microsoft en juin 2018.
[4] https://daranzolin.github.io/2021-03-03-automated-rmarkdown/
[5] voir également Boettiger et al. [Boe+15]

composés organiques volatils, COV, produits par l'échantillon. L'instrument peut analyser séquentiellement plusieurs échantillons.

Durant la période de stage, j'ai accompagné plusieurs séries d'expériences. Les modèles biologiques analysés varient à chaque expérience. Les plans d'expériences ont été conçus en dialoguant avec les utilisateurs. Bien souvent, ils avaient en tête l'expérience qui répondait le mieux possible à leur question biologique. J'anticipais les incohérences techniques et j'optimisais les paramètres de l'instrument. Avec le recul, je me suis rendu compte que les connaissances acquisent en cours de "recueil planifié de données" n'étaient pas pleinement exploitées.

Sur ce type d'instrument, un des problèmes principal est dû à la contamination de la ligne et de l'instrument par l'échantillon précédent. Nous pouvons analyser ces expériences pour repérer les erreurs commises.

Détection des traces d'un complément alimentaire

Cette expérience avait pour objectif de mesurer la diffusion de l'odeur de nourriture (pour des poules) dans quatre sites d'un mésocosme constitué de deux compartiments. La figure 1.1 illustre cette prise de mesure. Les sites étaient localisés dans la mangeoire (s1, compartiment 1), au centre du compartiment 1 (s2), à la jonction des deux compartiments (s3, côté compartiment 2) et au fond du compartiment 2 (s4). La nourriture au sein de la mangeoire était soit de la nourriture témoin (T) soit dopée par de l'huile essentielle (A et B). Il y avait 5 mésocosmes pour chaque modalité.

Fig. 1.1 : PTR-ToF-MS relié à l'un des quatre sites de mesure d'un mésocosme.

Dans la théorie d'une expérience totalement randomisée, il aurait fallu définir l'ordre

aléatoire de chacune des 60 acquisitions[6]. Dans la pratique, cette expérience nécessitait de déplacer l'instrument d'un mésocosme à l'autre entre chaque acquisition. Cette manipulation aurait drastiquement augmenter la durée de l'expérience. Nous avons donc tiré au hasard l'ordre de passage de chaque mésocosme, effectué un blanc et les acquisitions sur les quatre sites de mesures puis terminé par un nouveau blanc.

Il aurait été juste de tirer au hasard l'ordre de passage des sites. Pour autant, nous avons opté pour une hypothèse forte de commencer par le site le plus éloigné de la mangeoire (s4, s3, s2, s1) et de remonter le gradient de diffusion. Dans les faits, nous laissions assez de temps entre deux mesures pour renouveler plusieurs fois le volume d'analyse à l'intérieur de l'instrument ce qui permettait d'isoler les mesures. Après coup, je pense qu'il aurait été bon de procéder malgré tout à une randomisation de l'ordre des sites au sein d'un mésocosme.

Nous avions également une seule journée pour effectuer l'ensemble des acquistions, ce qui a été juste suffisant. Nous n'avons pas pu effectuer de répétition sur un même échantillon ce qui aurait augmenté la puissance statistique de l'expérience. L'analyse des data a été effectuée par le post-doc en charge de l'expérience. Je n'ai pas de retour sur des potentielles contaminations intra et inter échantillons.

Bien que la réalisation de cette expérience soit statistiquement discutable, elle a été réalisée en tout début de la période de stage et m'a permis la réflexion sur comment intégrer mieux la planification d'expériences.

Caractérisation de COV de l'amandier

Ce travail, qui s'inscrit dans la première année d'un projet de thèse, effectue une analyse temporelle des COV émis par les différents organes (bourgeon, fleur, feuille, fruit) des amandiers. Cette thèse a pour cadre un projet de lutte biologique contre un insecte ravageur des vergers d'amandiers. Les expériences ont durée trois mois à partir de début février. J'ai aidé à la mise en place des expériences, puis j'ai formé la doctorante à l'utilisation de l'instrument. Une fois cette dernière autonome avec l'instrument, j'ai peu assisté au suivi de l'expérience. J'ai toutefois soumis l'idée de suivre le moment spécifique de la floraison avec le PTR-ToF-MS.

Nous avons suivi durant un week-end les COVs émis par un bourgeon que nous avons isolé du reste de la plante. Ce bourgeon avait été sélectionné car il était proche de la floraison. Ainsi, nous avons pu recueillir les odeurs émises avant, pendant et après le débourrage. Cependant, nous n'avons pu obtenir qu'une seule unité expérimentale pour cette expérience. Par ailleurs, nous avions filmé le bourgeons durant l'expérience mais la caméra s'est éteinte après quelques heures. Ainsi, nous n'avons pas de témoin validant l'heure de phénomène. Les données techniques et les résultats de l'analyse sont détaillés à la section 3.3. Nous retenterons probablement cette expérience en février 2022 sur plusieurs unités expérimentales, en privilégiant une série de photos plutôt qu'un film.

Cette expérience de suivi cinétique a permis de mettre en avant les capacités de l'appareil et a fait germer plusieurs idées. Par exemple, nous avons fait des réunions préparatoires pour mettre en œuvre l'une d'elles que nous allons réaliser à l'automne. Nous souhaitons analyser la réponse chimique d'une fleur après la diffusion du son de bourdon, qui est un pollinisateur de cette plante. Une plante ne possède pas d'organe auditif au sens anthropologique du terme mais un son est une vibration caractérisée par des fréquences qui peuvent activer leurs sensilles mecanosensibles.

[6] 5 (mésocosmes) x 3 (modalité) x 4 (site d'acquisition) = 60

Étude de l'émission journalière des lavandes et des figuiers

L'expérience sur les lavandes présentée ici s'inscrit dans un cadre plus vaste s'intéressant aux phénomènes de pollution à l'Ozone O_3 liés au changement climatique global. Je ne reviendrai pas dans ce rapport sur la partie ozonée. Cependant, nous nous sommes intéressés aux variations journalières dans des conditions non ozonées (plantes témoins).

Nous avons mis en place quatre chambres de mesure permettant chacune soit d'isoler un plant de lavande soit d'être un blanc. Ces chambres de mesure sont des cylindres en plastique, ventilés, de volumes équivalents et reliés à un flux d'air zéro d'un débit de 5 litres/min. La longueur et le volume des lignes reliant les chambres à l'instrument de mesure n'étaient pas identiques. Nous avons lancé des acquisitions sur les quatre chambres séquentiellement durant 48h. Nous avons répété ce processus deux fois, pour obtenir neuf plants au total dans notre plan d'expérience. Pour les lavandes, il n'y avait qu'une modalité (plante témoin). Toutefois, un des quatre moteurs permettant la ventilation à l'intérieur de la chambre ne fonctionnait pas. Nous avons décidé que cette chambre serait le blanc. Là encore, cette décision est critiquable. D'un côté, il fallait que le moteur HS soit celui de la chambre "blanc" afin de ne pas créer artificiellement les modalités "plante témoin ventilé" et "plante témoin non-ventilé". D'un autre côté, nous comparons un blanc non-ventilé à des plantes ventilées. Au regard de débit d'air pur de 5L/min, nous avons jugé ce biais négligeable. En revanche, nous aurions pu, au prix de quelques branchements et rebranchements supplémentaires, déplacer le moteur défectueux d'une chambre à l'autre entre les répétitions. Ceci aurait permis de mieux répartir l'erreur expérimentale due aux lignes de mesure. Ceci étant, les résultats présentés dans la section 3.3 montrent que le soucis technique n'a pas engendré de biais notable/significatif.

Quelques jours après l'expérience avec les lavandes, nous avons réutilisé le même dispositif avec des figuiers. Le temps d'analyse a été augmenté pour passer à 5 jours. La chambre "blanc" est restée la même (non ventilée). Cette fois, les plantes étaient soit "pollinisées" soit "non-pollinisées" (témoin). Nous avons décalé la plante témoin d'une chambre à l'autre à chaque répétition. Théorique, la réalisation d'un plan d'expérience en randomisation totale aurait nécessité un tirage au sort à chaque répétition, au risque de tomber trois fois sur la même chambre. Je n'ai encore pas aidé à l'analyse de cette expérience et ne peut pas me prononcer sur la pertinence des choix effectués.

Deux des quatre expériences présentées ici ne seront pas analysées plus en détail dans ce rapport. Pour autant, leur présentation me semblait pertinente pour mettre en avant les difficultés à coller à la théorie des plans d'expériences d'une part et pour mieux comprendre les multiples capacités de l'instrument PTR-ToF-MS.

1.4 Conclusion partielle

J'ai commencé ce métier en novembre 2019. En raison des conditions sanitaires et de la formation continue en M2 SSV, j'ai une vision légèrement biaisée d'une année "standard" sur ce poste. Pour autant, je pense avoir trouvé le bon équilibre entre d'un côté le développement d'outils numériques résolument orientés utilisateurs et qui entrent dans le Deuxième Plan national pour la science ouverte et de l'autre côté l'accompagnement des utilisateurs de la conception de leurs expériences à l'acquisition.

Le temps annuel consacré à chacune de ces facettes est plutôt équilibré bien que les saisons chaudes soient bien évidemment plutôt dédiées au terrain. Après ces quelques descriptions d'expériences, je propose une description plus détaillée de l'instrument PTR-ToF-MS et de son environnement dans le champs de la spectrométrie de masse, chapitre 2, mais un lecteur pressé de connaître les conclusions de ces expériences peut sauter au chapitre 3.

CHAPITRE 2

Présentation de la PTR-ToF-MS

De très nombreux projets de recherche issus du CeMEB nécessitent l'analyse de **C**omposés **O**rganiques **V**olatils (COV ou VOC en anglais) aussi communément appelés "odeurs". Les COVs sont omniprésents dans la nature et permettent, avec les autres sens, une organisation du vivant en agissant comme vecteur d'information de la médiation chimique. Par l'acquisition d'un PTR-ToF-MS, la communauté souhaitait lever trois verrous techniques :

- l'appareil permet un fonctionnement en flux continu avec une résolution temporelle très fine (fréquence d'analyse allant jusqu'à 10 scans par seconde), permettant ainsi le suivi des cinétiques d'émissions de COV. Cet instrument simplifie drastiquement et affine les expériences cinétiques effectuées par des mesures moyennées de GC-MS ;

- le PTR-ToF-MS possède une excellente résolution en masse qui facilite l'identification des molécules. Un spectre est composé de plus de 140 000 mesures de masses couvrant une large gamme des masses des COV biologiques (de 70 à 500 m/z). L'appareil possède donc une résolution environ mille fois supérieure par rapport à un simple quadripôle (PTR-MS ou GC-MS) qui fournit des m/z à l'unité de masse atomique (uma ou Dalton Da) ;

- le seuil de détection extrêmement bas, de l'ordre du ppt (part per trillion), permet une sensibilité adaptée à la mesure de traces.

2.1 Une brève histoire de la PTR-ToF-MS

Durant la décennie 1990, l'*Institut für Ionenphysik der Leopold-Franzens-Universität* d'Innsbruck en Autriche développe un nouvel instrument pour l'analyse chimique des gaz. En travaillant avec la *Universitiitsklinik fiir Innere Medizin*, l'équipe menée par Lindinger développe un spectromètre en phase gazeuse qui ionise les molécules grâce à un gaz neutre ([LLA91] et [LHP93]). Ils comprennent rapidement l'avantage d'utiliser un proton (apporté à la réaction sous la forme H_3O^+) par rapport aux ions Kr^+ et Xe^+. Ils publient une sorte de preuve de concept en 1994 [Lag+94] et détaillent plus finement l'instrumentation et les réactions chimiques [Han+95][1]. En 1998 sort l'article qui fait désormais référence, [LHJ98],

[1] avec des remerciements incroyables

signé par le trio Lindinger, Hansel et Jordan (et sa version courte [LJ98]).

Ionicon commercialise la même année le premier PTR-MS. Cette technique se déploie rapidement dans les sciences de l'atmosphère, médicales et biotech ([BA01]) permettant la détection de traces de COV. Rapide et sensible, la PTR-MS élimine certains désavantages de la GC-MS. Cependant, la PTR-MS détecte de la masse nominale des ions. Un échantillon complexe peut contenir plusieurs ions isobares[2]. Il convient alors de gagner en résolution de masse et séparer ces isobares. [Jor+09] revient sur les tentatives les plus concluantes ([Bla+04], [Enn+05], [Ino+06] et [Tan+07]) avant de proposer son approche : la PTR-ToF-MS. Développé en collaboration avec la précédente équipe, le minutieux article de [GMH10] conclu superbement la vingtaine d'années de développement nécessaire à cette technique. Les années suivantes permettront simplement une amélioration des différentes parties de l'instrument.

Par ailleurs, c'est au tour des mathématiciens d'apporter leur contribution et en particulier à l'équipe de Cappellin avec deux articles. Le premier, [Cap+10], permet de mieux estimer la masse exacte de chaque ion détecté. Le seconde, [Cap+11], offre une méthodologie détaillée de l'utilisation du PTR-ToF-MS de l'acquisition à la fin de l'analyse. Puis en 2015, [Hol15] propose un logiciel pour l'analyse des data mais cette tentative n'a pas été suivie par la communauté d'utilisateurs. Il existe désormais un nombre important d'articles utilisant la PTR-ToF-MS. Bien entendu, des outils communs de chimiométrie ont été utilisés pour renforcer les résultats, [Deu+19]. Mais cela sera abordé dans le chapitre suivant.

Je suppose que la prochaine avancée majeure viendra d'une technologie permettant de discriminer les isomères[3] [Cla+21] [4]. Par exemple, cela permettrait de séparer les différents monoterpènes omniprésents dans les émissions végétales et qui nous concernent particulièrement en écologie chimique. Actuellement, ce désavantage est comblé par le couplage de la PTR-ToF-MS avec une seconde technique d'analyse chimique. Ce court historique montre toutefois que la collaboration entre physiciens, chimistes, biologistes et mathématiciens est fructueuse et promise à un bel avenir.

2.2 Instrumentation générale

Comme nous l'avons vu précédemment, le PTR-ToF-MS est conçu pour analyser finement la masse moléculaire des échantillons. Pour cela l'instrument aspire un débit d'air constant de 100 ml/min[5]. Cet *air in* sur la figure 2.1 peut être connecté à un gaz zéro pour diluer un échantillon trop concentré (pour éviter la saturation) ou à un gaz étalon pour l'étalonnage de l'instrument. Le flux d'air est ensuite injecté dans le réacteur (*FIMR* pour *focusing ion-molecule reactor*). Dans cette chambre, des ions H^+ issus d'eau ultra pure sont également injectés à débit constant afin d'ioniser les molécules de l'échantillon (*Proton-Transfer-Reaction*).

Il est important de comprendre que, contrairement à une ionisation par électron[6], il n'y a quasiment pas de fragmentation moléculaire. C'est une ionisation douce, comme en ESI (ElectroSpray Ionization) fréquemment utilisée en LC-MS. Les ions ont "une unité de masse molaire de plus"(un proton) que les ions moléculaires. Par exemple, le linalol, un COV fortement produit par la lavande, a une masse molaire de 154,249 g/mol mais sera detecté à 155,256 g/mol. Par ailleurs, le nombre d'ions MH^+ détectés dépend du taux de protonation de la molécule M.

[2] molécules possédant un nombre de nucléons identique.

[3] molécules partageant la même formule brute mais avec un agencement tridimensionnel différent.

[4] si un mémoire de master permettait la science fiction, j'aurais évidemment développé quelques idées à propos de capteurs de champs électromagnétiques ou de spectromètres optiques à la place des capteurs MS actuels.

[5] ou sccm, pour *standard cubic centimeters per minute*

[6] comme c'est le cas en ionisation électronique fréquemment utilisée en GC-MS.

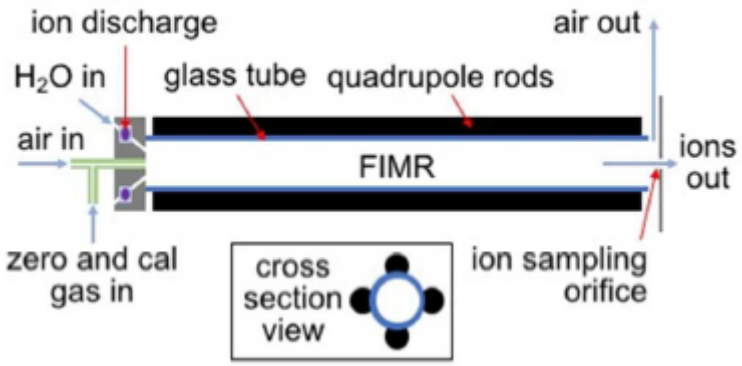

Fig. 2.1 : Schéma de la chambre d'ionisation (Reprinted with permission from Krechmer et al. [Kre+18]. Copyright 2021 American Chemical Society.)

Sous forme d'ions, l'échantillon peut être focalisé par un champ électromagnétique généré par un quadripôle et envoyé dans la partie ToF.

Une fois dans la colonne, l'énergie potentielle d'un pulse électromagnétique à 25kHz est convertie en énergie cinétique par les ions propulsés. À énergie constante, la masse fait la différence lors de la mesure du temps de vol (ToF, figure 2.2). Ainsi les ions les moins lourds arrivent en premiers sur le détecteur. Le comptage du nombre de coups sur le détecteur donne l'intensité par unité de temps. Un étalonnage, avec des ions moléculaires étalons, permet de convertir le *temps de vol* en *masse*, ce qui permet par la suite d'analyser les masses des échantillons inconnus.

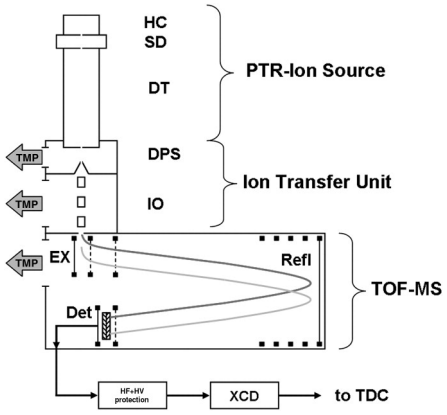

Fig. 2.2 : Schéma de la colonne ToF et du détecteur MS (Reproduit de l'article Graus et al. [GMH10], consulatable ici : https://www.sciencedirect.com/science/article/pii/S1044030510001005#fig1

Pour une compréhension plus exhaustive, je recommande la lecture des articles de Hansel et al. [Han+95] pour la partie PTR-MS, de Graus et al. [GMH10] pour la partie ToF et de Krechmer et al. [Kre+18] spécifique à l'instrument utilisé au CEFE.

2.3 Présentation des spectres

Nous pouvons à présent nous intéresser aux spectres à analyser. J'ai représenté sur la figure 2.3 un spectre de trois plants de lavandes acquis autour de 8h du matin fin juin. Une figure identique mais dynamique peut se télécharger ici[7]. Cette figure dynamique permet d'effectuer

[7]il faut télécharger le fichier puis l'ouvrir avec un navigateur internet.

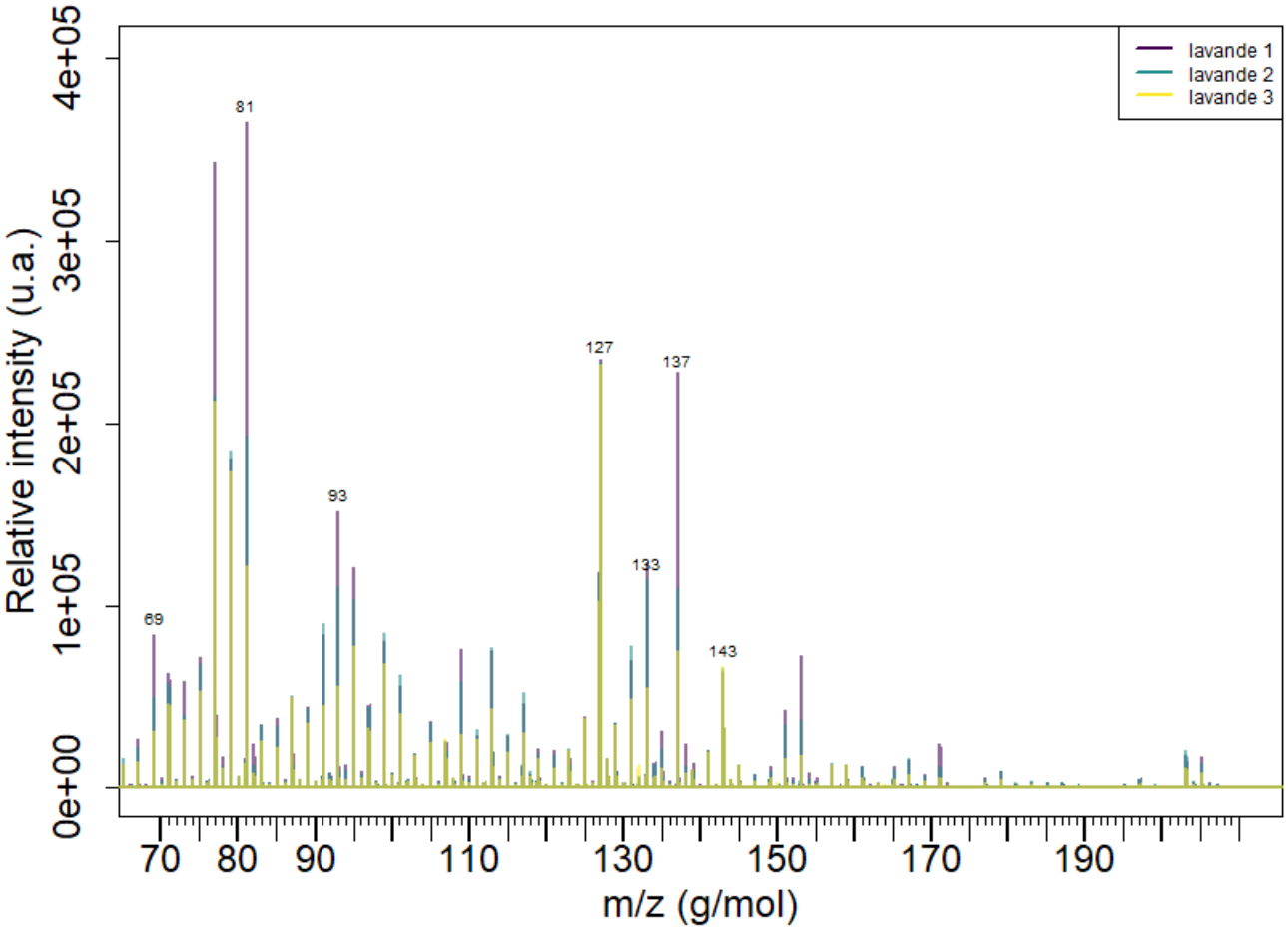

Fig. 2.3 : Spectres de masse de trois lavandes acquis à 8h le 29 juin.

des zooms afin d'avoir une compréhension plus précise qu'avec la simple figure 2.3.

Les pics sont extrêmement fins et centrés sur les masses exactes des ions. Les pics issus des ions isobares se regroupent autour des masses unitaires, soumis à une loi centrale limite locale. Cela s'explique facilement au regard de la masse exacte des atomes composants les COV (H, C, O pour l'essentiel) sur le tableau de Mendeleiev[8]. Il y a plusieurs ordres de grandeur[9] entre les pics très intenses et ceux faibles. Cette intensité ne présage rien quand à l'importance de ce pic dans l'analyse.

La plage spectrale de l'instrument est de 70 à 500 g/mol. Cela est limité pour des raisons techniques alors qu'il y a de nombreuses molécules d'intérêt de masses inférieures à 70g/mol. En revanche la limite, communément acceptée, fixée pour les molécules volatiles et de 300 g/mol. Ces considérations prises en compte, nous obtenons pour chaque spectre un vecteur brut de 158 768 points. Ces spectres sont réduits, après alignement (voir section 2.5 et 3.1) à une taille inférieure à 50 000 points tout en gardant une résolution de 10^{-3} g/mol. De plus, la PTR-ToF-MS a été développée afin d'analyser des cinétiques d'émissions. Dans nos cas, nous enregistrons des spectres de quelques secondes consécutivement durant plusieurs minutes à plusieurs jours. Pour une expérience, nous générons ainsi des dizaines, voir des milliers, de spectres. Ceci étant, nous pouvons dès à présent percevoir les problèmes liés aux temps de calculs que nous allons rencontrer et la nécessité de développer des outils optimisés pour analyser de tels objets.

[8] https://www.youtube.com/watch?v=0lNuTSz6KVM
[9] entre 10 et 10^6

2.4 Technique rivale ou alliée de la GC-MS ?

La métabolomique est le maillon le plus exhaustif de la chaîne -omic. Il se concentre sur l'analyse des molécules, les *métabolites*, produites par un système biologique. Les deux domaines d'analyse sont la résonance magnétique nucléaire (NMR) et la spectrométrie de masse (MS). Pour l'analyse des COV en spectrométrie de masse en écologie chimique, la technique de GC-MS est la plus utilisée à l'heure actuelle. La GC-MS, utilisée depuis des décennies, est largement implantée dans le paysage scientifique international. De plus, de nombreux logiciels performants[10] existent. Il est légitime de se poser la question de la nécessité d'utilisation des deux techniques[11].

Premièrement, la GC-MS accumule les COV dans des pièges chimiques[12] puis les analyse dans un second temps. La durée d'accumulation dans le piège est de l'ordre de la dizaine de minutes jusqu'à 24h. La PTR-ToF-MS a une sensibilité accrue permettant l'acquisition de spectres à une cadence de 25kHz. Afin d'augmenter le ratio signal sur bruit et de drastiquement réduire la quantité de data générée, ces spectres sont additionnés les uns aux autres sur des durées d'acquisitions plus longues[13]. En répétant ces acquisitions, la PTR-ToF-MS permet l'analyse cinétique d'un phénomène grâce à un pas de temps d'accumulation très court. Analyse statique contre analyse quasi-dynamique. Le prix de cette sensibilité est comme bien souvent une sélectivité dégradée. Ainsi la GC-MS, *a contrario* de la PTR-ToF-MS, peut discriminer certains isomères, principalement grâce à sa phase de chromatographie. Cela est permis grâce à la fragmentation des molécules lors de l'étape d'ionisation par un électron qui intervient après la phase de chromatographie.

Bien que les molécules analysées soient les mêmes, les approches et résultats sont donc bien différents. De plus, il est souvent intelligent de croiser ces deux techniques [Maj+18]. De façon empirique, le PTR-ToF-MS analyse séquentiellement plusieurs échantillons. Il est aisé de collecter les COV pour le GC-MS juste avant ou après une séquence dédiée au PTR-ToF-MS. Nous pouvons ainsi, par exemple, obtenir deux ou trois analyses ponctuelles de GC-MS lors d'une journée entière d'analyse PTR-ToF-MS. Nous combinons la sensibilité et la spécificité des deux techniques.

Les données obtenues en GC-MS sont donc plus complexes qu'en PTR-ToF-MS. Chaque échantillon est représenté par une matrice ayant une dimension pour le temps de rétention et une dimension pour le spectre de masse. Il n'est donc pas possible de réutiliser tels quels les outils numériques de la GC-MS. Cependant, il serait dommage de ne pas piquer les idées déjà existantes. J'ai donc créée un package R spécifique à l'analyse des données du PTR-ToF-MS mais en m'inspirant largement de ce qui se fait sur d'autres techniques : workflow (idée piquée à MZmine), bucketing (pratique courante dans la RMN), chimiométrie (largement utilisée pour l'analyse des spectres optiques) ; ou en utilisant des packages destinés à d'autres méthodes (MALDIquant conçu pour le traitement des spectres MALDI-ToF)

2.5 package proVOC

Comme expliqué dans la section sur la science ouverte (1.2), ce package peut se retrouver sur ce dépôt Git.

[10] MZmine pour n'en citer qu'un

[11] ou de façon moins prosaïque : pourquoi diable les physiciens ne restent pas sagement avec leurs bosons et viennent régulièrement embêter leurs collègues avec des idées nouvelles ?

[12] qui se présentent sous différentes formes en fonction des utilisations

[13] typiquement de l'ordre d'une à trente secondes

Explication générale

L'objectif de proVOC est de permettre une vue d'ensemble rapide[14] pour les utilisateurs pressés d'analyser *leurs* données sous R sans avoir besoin d'une grande connaissance de ce langage. La fonction *import_sp()* constitue la première étape de l'analyse. Les données contenues dans des fichiers au format *hdf5* générés par notre instrument sont importés dans l'espace de travail [Fis+21]. Après cette étape, les spectres sont automatiquement alignés puis réduits. Les détails à propos de l'alignement des spectres sont donnés dans la section 3.1. Pour l'étape de réduction, j'ai défini empiriquement de supprimer chaque colonne (masse) qui ne possédait aucune ligne (intensité) supérieure à un seuil. Ce seuil est calculé pour chaque jeu de données comme égal à 20 fois l'écart-type médian de chaque masse[15]. Il est probable que cette étape de réduction puisse être optimisée d'avantage mais je reviens sur les perspectives d'améliorations du package dans la section dédiée, 2.5.

Passées ces étapes, la fonction met en forme les données dans un objet *list*, récupère les metadata (date et heure d'acquisition) puis détecte les pics grâce à la fonction *detectPeaks()* du package **MALDIquant** et calcule les *Area Under the Curve*(AUC) grâce à la function *AUC()* du package **DescTools**. Enfin, cette liste nommée "sp" est sauvegardée en .Rdata. Ce travail préliminaire peut être long.

Une fois effectué, l'import permet à l'utilisateur de lancer facilement une série d'analyse pour explorer rapidement son jeu de données. Pour cela, il lui suffit de modifier des options dans un workflow :

```
# Workflow pour l'analyse :
mt <- list("h5" = mt_h5,                                    # Don't touch
           "acq" = which(mt_h5$use4analysis == TRUE),       # Don't touch
           "wd" = sp$h5$wd,                                 # Don't touch
           #___________________________________#
```

[...]

```
# graphe pour visualiser les spectres de masses de chaque acquisition :
           "view_plot" = FALSE,    # Si TRUE, graphe moyen de chaque acquisition,
                                   # sinon precisez le numero d'une ou plusieurs
                                             # acquistion(s). (ex :
                                             # c(200,300,400)
                                             # mt_h5$end
                                             # c(mt_h5$start, mt_h5$end)
           "view_each.group" = FALSE,    # FALSE ou le nom d'une meta colonne
                                         # ("XXX" de mt_h5$XXX) pour les graphes
                                         # larges ou centres sur un seul pic.
           "view_plot_xmin" = 70,
           "view_plot_xmax" = 250,
           # ou ...
           "view_plot_peak" = FALSE, # c(59,73,81,93,109,127,137,143,153) ou
                                     # False ou c(61,...) Permet de ne regarder
                                     # qu'un seul pic
           #___________________________________#
```

[...]

```
           # ICA
           "ICA" = FALSE,
           "ICA_AUC" = TRUE,
           "nc" = 6
           #___________________________________#
           )

PTR_MS_analysis(sp, mt)       # debute les analyses
```

[14] proVOC : **P**erform a **R**apid **O**verview of the **V**olatil **O**rganic **C**ompound

[15] *cf.* "calculation of threshold" dans la fonction *reduction2()*

L'objet *mt* peut être sauvegardé pour se souvenir des paramètres utilisés lors de l'analyse. Nous pouvons détailler ci-dessous quelques options.

Analyse AUC

```
#__________________________________#

# Aire sous la courbe (AUC) :
"AUC" = TRUE,
"AUC_plot_exp" = FALSE,      # Si TRUE, l'axe 'y' est exponentionnelle.
"AUC_each.group" = FALSE,   # Si l'argument est le nom d'un meta colonne
                             # ( "XXX" de mt_h5$XXX) chaque groupe de cette
                             # colonne aura son graphe. Sinon FALSE.
"AUC_each.Mass" = TRUE,      # Si TRUE, un graphe pour chaque masse
"M.num" = c(81, 137, 153),   # Les masses analysees (utilise pour AUC)
"AUC_plot_dy" = TRUE,        # Si FALSE, plot exporter en .tiff. Si TRUE, plot
                             # exporter en .html (dynamique)
"AUC_format" = "date",       # L'axe du graphe est base sur une duree ("time")
                             # ou sur la date et heure d'acquisition ("date")
#__________________________________#
```

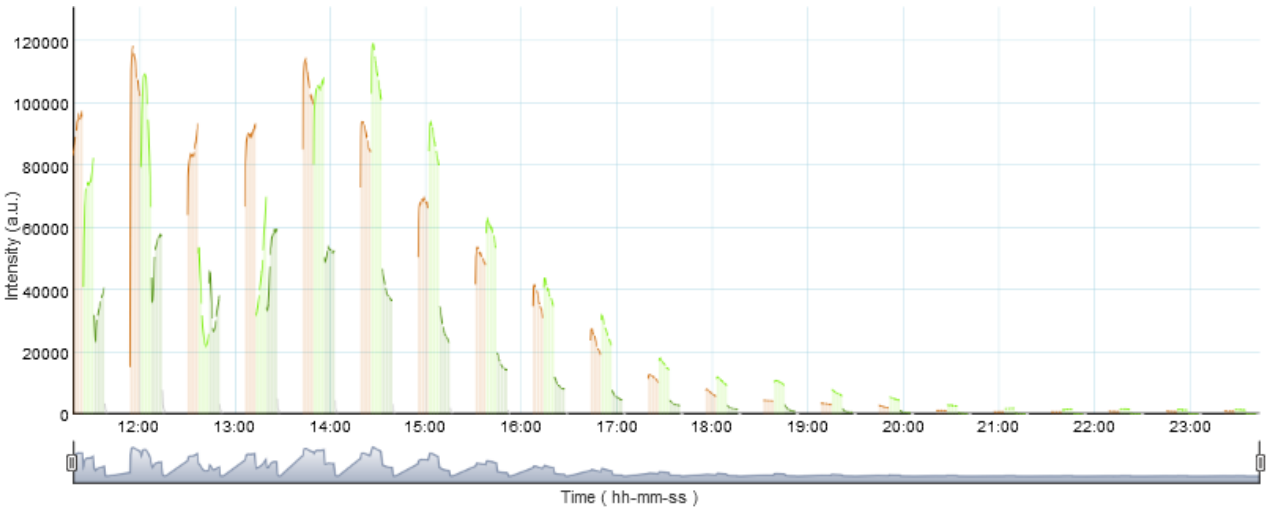

FIG. 2.4 : Cinétique d'émission des AUC à 137 Da de trois figuiers.

La figure 2.4 montre l'évolution quotidienne de l'émission des molécules de masse 137 (de 136.5 à 137.5) de trois plants de figuiers dont le cadre expérimental est décrit section 1.3. Chaque plante est analysée par cycle de 384 secondes composé de 24 acquisitions de 16 secondes (regroupées en trois parties pour des raisons techniques). Vous pouvez télécharger cette figure ici afin de zoomer plus facilement.

Nous pouvons observer que le début de chaque cycle d'enregistrement est systématiquement impacté par la mesure précédente et le volume mort des lignes de mesures. Nous pouvons justifier ce biais expérimental et donc le supprimer en ne prenant pas les 8 premières acquisitions de chaque cycle.

Ce bloc d'options est vraiment très simple d'utilisation. En jouant sur les masses et les modalités incluses dans le graphe, les utilisateurs arrivent facilement à appréhender l'évolution des COV de leur système. Pour autant, nous reviendrons en 2.5 sur l'utilisation de l'AUC et sa méthode de calcul qui impacte grandement ce module.

Étalonnage, régression linéaire et rapport automatisé

```
#_________________________________#
# Calibration
"calibration" = FALSE,
"cal_rapport" = TRUE,     # Redige un rapport automatise pour evaluer la qualite de la calibration
"cal_plot_exp" = TRUE,    # Les axes x et y sont exponentiels ?
"M.conc" = c(1.008, 1.018, 1.013, 1.014, 1.013, 1.013, 1.015, 1.006, 1.007,
             0.994, 1.013),      # La concentration, en ppm, de chaque masse,
                                 # par defaut C = 1.000 ppm. (1 ou FALSE)
#_________________________________#
```

Afin de vérifier l'étalonnage de l'appareil, j'utilise un gaz étalon que je dilue avec un gaz zéro lors des mesures. J'ai développé un script pour analyser cet étalonnage. La sortie de ce script se fait sous forme de rapports automatisés basés sur des régressions linéaires simples. J'ai eu pour projet de développer cela en combinant les différentes masses et en proposant des tests de validation. La base de ce script peut également être reprise pour mesurer la concentration de différentes molécules. Ceci étant, le manque de temps et le côté monovariable ont largement freiné ce projet. D'autant que nous allons le voir, le travail ne manque pas.

Perspectives

Dans cette partie, je n'ai pas développé les analyses plus poussées (alignement et chimiométrie) qui sont expliquées dans le chapitre suivant. Nous avons simplement vu le principe de base de la PTR-ToF-MS ainsi que les modules de base de l'analyse : obtenir le graphe des spectres de masses, le suivi des AUC et le calcul de concentration. Le package proVOC permet d'effectuer ces opérations de façon simple et sans rien coder. De plus, il met à disposition les jeux de données pour que les utilisateurs puissent soit appliquer leur propre traitement soit utiliser des modules avancés.

Le package a été conçu pour assurer une reproductibilité des analyses. Lors de la conception de la structure du package, nous avions peu de recul sur l'utilisation du PTR-ToF-MS. Désormais, je suis convaincu qu'il me faut développer la version 2 très prochainement. Trois grosses modifications vont être effectuées.

La première est assez technique et concerne l'importation de données :

```
# Data importation and reduction ####
f_h5 <- dir("h5") %>%
    grep(".h5", .)  # localise h5 files
ls_h5 <- vector("list", length(f_h5))  # the acquisitions listes
names(ls_h5) <- nm_ls(f_h5)  # and the name
for (i in 1:length(ls_h5)) {
    ls_h5[[i]] <- paste0("h5/", dir("h5")[f_h5[i]]) %>%
        H5Fopen()  # files import
}
mt_h5 <- meta_h5(ls_h5)  # make the meta file
sp <- sp_red(ls_h5, mt_h5)  # spectra reduction
sp$h5 <- list(f_h5 = f_h5, ls_h5 = ls_h5, mt_h5 = mt_h5, wd = getwd())
h5closeAll()  # close the connexion with the h5 files.
# ls_h5 become NULL
```

Les fonctions *H5open()* et *H5close()* du package **rhdf5** permettent d'ouvrir les formats hdf5 afin de les importer puis de les fermer. La fonction *sp_red()*, que j'ai écrite, est une meta fonction qui permet d'importer les spectres, les meta data, d'aligner les spectres, de supprimer les données inutiles et des détecter les peaks. Comme nous pouvons le voir, elle se trouve en sandwich entre l'ouverture de toutes les datas et la fermeture des données. Cette maladresse occupe de la RAM inutilement et ralentit beaucoup cette étape. Pour autant, toucher à la

fonction qui importe les données et les met en forme est une opération risquée qui met en péril la fonctionnalité des fonctions ultérieures.

La seconde modification sera sur le calcul de l'AUC. Initialement, je voulais trouver un moyen simple d'exprimer les 50 000 points des spectres en une valeur plus acceptable pour de l'analyse chimiométrique. J'avais donc piqué l'idée du bucketing des communautés de RMN en calculant l'AUC par masse unitaire (de M-0.5 à M+0.5). Cela permet d'obtenir une matrice de seulement 400 colonnes environ. En réalité, cela masque les cinétiques des molécules isobares qui est un des points forts de la technique PTR-**ToF**-MS. De plus, j'ai découvert après coup une fonction performante pour détecter les pics (*detectPeaks()*). Je peux donc remplacer les AUC unitaires par l'intensité des pics. Il n'y aura alors plus qu'une confusion sur les isomères mais qui correspond aux limites techniques de l'instrument. J'aimerais également travailler sur l'AUC des pics et leur déconvolution afin de savoir si de l'information y est cachée.

La dernière modification se fera sur la structure :

```
PTR_MS_analysis(sp, mt)  # debute les analyses
```

Une seule fonction permet de lancer la série d'analyses précisée dans le workflow *mt* sur les data de l'élément *sp*. La fonction *PTR_MS_analysis()* ne retourne rien, si ce n'est des graphiques, des fichiers *.csv* ou des rapports automatiques. J'aimerais inclure le workflow à l'intérieur de l'objet *sp* et le compléter au fur et à mesure. Ainsi l'utilisateur pourrait lancer des opérations à la volée, par exemple (en reprenant le bloc AUC de la section 2.5) :

```
# Aire sous la courbe (AUC) :
sp <- sp_AUC(sp, AUC_plot_exp = FALSE, AUC_each.group = FALSE,
    AUC_each.Mass = TRUE, M.num = c(81, 137, 153), AUC_plot_dy = TRUE,
    AUC_format = "date")
```

Cette forme franchement plus digeste permettra d'inclure plus simplement une analyse PTR-ToF-MS dans un script ou un rapport Rmarkdown.

J'espère, avec ces modifications importantes, rendre l'analyse plus optimisée en temps et répondre mieux aux attentes des utilisateurs. Ce package est sur GitHub depuis septembre 2020. J'ai fait peu de pub car je connais ses limites et je souhaiterais corriger ses principaux défauts avant de le proposer aux collègues des autres centres de recherche avec qui nous avons des contacts. Les retours sur les quelques démonstrations effectuées sont très positifs et proVOC permettra (je l'espère) de répondre positivement à un réel besoin. De plus proVOC est complètement cohérent avec les principes de l'open science développés au chapitre 1 et intègre des analyses chimiométriques qui sont détaillées au chapitre 3.

Chapitre 3

Méthodes Multivariées

Nous allons à présent nous intéresser aux algorithmes utilisés pour traiter les jeux de données générés par le PTR-ToF-MS. Ce chapitre est divisé en deux parties. Lors de l'écriture du package, j'avais travaillé sur différentes méthodes d'alignement avant de m'arrêter sur l'une d'elles. Le début de ce stage m'a permis de reprendre ce travail avec des idées nouvelles que je présenterai dans une première partie. La deuxième partie présente l'utilisation de la chimiométrie pour l'analyse de données, d'abord avec la PCA puis avec deux méthodes moins connues, l'ICA et la MCR qui ont l'avantage chacune d'essayer de décomposer les données en produits purs.

3.1 Alignement et sélection de variables

Actuellement, le package utilise la fonction *alignSpectra()* du package **MALDIquant**. L'alignement est fortement lié à la détection de pic. Soit le modèle utilisé n'est pas optimum et ajoute ou supprime des pics. Soit il crée un décalage trop important en alignant mal les pics ce qui génére de la confusion lors de l'étape d'identification.

alignSpectra() est performant mais rencontre des limites sur les pics de faible ratio signal sur bruit, comme nous pouvons le voir sur la figure 3.1. Le bruit intensifie l'effet vaguelette et génère de l'ambiguïté sur le nombre de pics présents. Par ailleurs, la fonction créée des pics négatifs qui faussent le calcul de l'AUC.

J'ai trouvé un article très intéressant de Picaud et al. [Pic+18] qui propose un algorithme pour nettoyer la ligne de base, aligner les pics et les détecter. L'article explique les problèmes de faire ces trois étapes successivement et propose donc une alternative. Nous avons discuté avec l'auteur de sa solution. Malheureusement, le script est écrit en C++. Il est possible d'utiliser ce langage sous R, via Rcpp par exemple, mais cela aurait demandé un peu trop de temps. Pour l'instant, cette solution a été mise de côté mais pourra être explorée plus tard.

De plus, Jean-Michel Roger m'a proposé une approche basée sur l'algorithme COVSEL [Rog+11] pour résoudre notre problème. Cette méthode permet de recaler les masses en recalculant l'abscisse pour chaque spectre. On observe cependant dans la théorie (équations 4 et 5) que l'on va devoir diagonaliser une matrice d'environ 150000x150000 points, essentiellement composée de zéro. Cette opération est impossible à effectuer avec un ordinateur de bureau. Pour autant, la solution que j'ai trouvée est d'utiliser l'algorithme COVSEL sur une toute petite partie du spectre et de le faire glisser sur toute la longueur. Cette technique

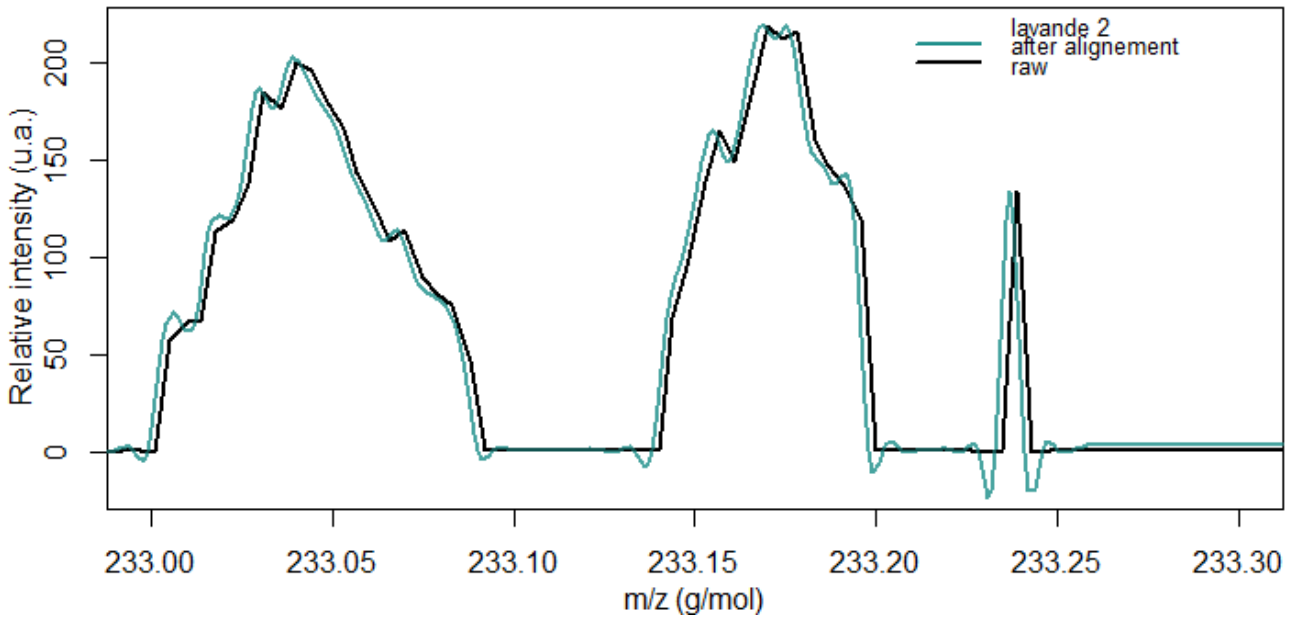

Fig. 3.1 : Zoom sur une série de pics très peu intenses émis par les lavandes. Des vaguelletes et des pics négatifs apparaissent.

fonctionne mais n'est pas du tout optimisée et prend un temps non acceptable (environ une minute par spectre).

Finalement, nous avons fait le choix de garder la fonction *alignSpectra()* de **MALDIquant**. De plus, ce package nous permet d'avoir une liste de tous les pics détectés. Ainsi, nous avons pour un même jeu de données, trois matrices à utiliser pour les analyses chimiométriques :

- la matrice alignée (composée d'environ 50 000 masses) ; *(Mc)*
- les AUC (~340 masses unitaires non nulles) ; *(Ma)*
- les pics (~300 pics, qui peuvent être des isobares). *(Mp)*

Pour une étude chimiométrique, nous pouvons discuter de la pertinence des jeux de données. Avec une matrice alignée *Mc*, on peut estimer que les algorithmes puissants des chimiométriciens vont détecter de petites variations négligées par les méthodes de détection de pics. Les *Mc* peuvent donc détecter les traces de COV. Cependant les deux autres matrices *Ma* et *Mp* (AUC ou pics) possèdent un nombre de colonnes réduit de presque deux ordres de grandeur. De plus, le grand avantage du PTR-ToF-MS est d'obtenir des spectres ayant une excellente résolution de masse (10^{-3}Dalton). Nous pouvons donc utiliser les matrices *Ma* et *Mp* pour les analyses chimiométriques avec une préférence pour les secondes.

Maintenant que les outils pour détecter les pics et les outils d'analyses sont fonctionnels pour les jeux de données du PTR-ToF-MS, il serait très intéressant d'analyser (via une MCR par exemple) comment la modification des paramètres de la fonction *detectPeaks()* (de **MALDIquant**) influe sur les résultats de chimiométrie.

3.2 Analyses Chimiométriques

Analyse en Composante Principale (ACP)

L'ACP[1] est une méthode bien connue et enseignée dans toutes les formations d'analyses statistiques. Elle permet d'obtenir rapidement une vision de la variance des données. ProVOC permet d'effectuer une ACP sur le jeu de donnée de l'utilisateur. Un exemplaire peut être consulté ici. Il décrit le jeu de données utilisé dans la section 1.3 et y sera commenté dans cette même section. L'ACP peut être appliquée à la matrice alignée avec un temps de calcul acceptable.

Cependant, le résultat de l'ACP est un pur objet mathématique qui est souvent très compliqué à interpréter correctement. Je me suis donc intéressé à deux algorithmes que je souhaitais utiliser depuis longtemps[2].

Multivariate Curve Resolution (MCR)

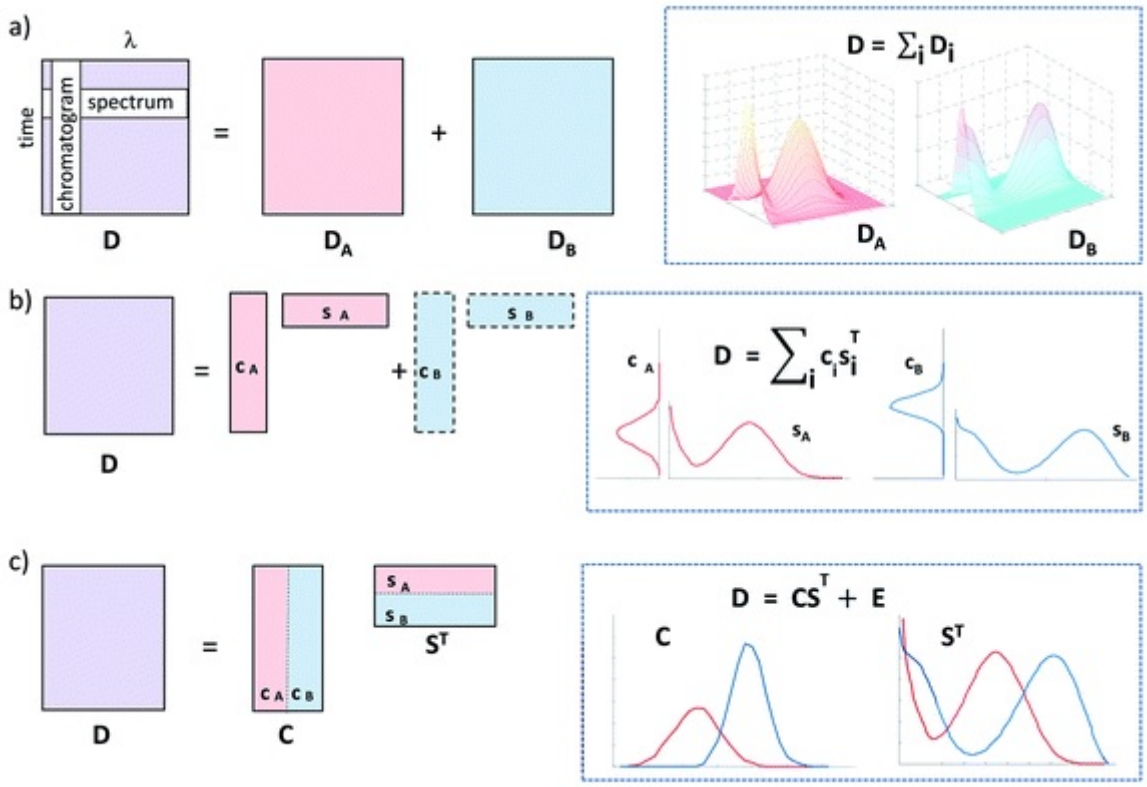

Fig. 3.2 : Bilinear model obtained from MCR for an HPLC-DAD data set. Expressed as (a) sum of pure signal contributions ; (b) sum of the dyads of pure concentration profile and spectra ; and (c) product of matrices of pure concentration profiles and spectra. (Image reproduite de Juan et al. [JJT14] avec l'autorisation de l'éditeur.)

La MCR est une famille de méthodes qui cherche à décomposer l'ensemble des spectres en plusieurs sous-ensembles de produits purs. La figure 3.2 issue de l'article de Juan et al. [JJT14] permet de bien comprendre l'objectif d'une MCR. La matrice D, obtenue par l'expérience, est l'expression de deux produits purs $S_A{}^T$ et $S_B{}^T$ (deux bouquets de COV pour notre PTR-ToF-MS) de concentration C_A et C_B, qui sont les quatre vecteurs que nous cherchons à connaître[3]. Les étapes de l'équation sont décrites dans l'article précédemment cité.

[1] Wikipédia est un haut lieu de la science ouverte
[2] très probablement depuis le GFSV 2014
[3] et dont nous ne connaissons rien.

De plus, la MCR permet l'ajout de contraintes telles que la "non-négativité" qui permet de ne pas avoir de composantes négatives dans les S_i^T (fréquent en ACP) ou la contrainte "d'égalité" qui permet par exemple de renseigner un spectre pur de référence.

Un autre aspect très important de la MCR est de comprendre les trois "ambigüités" liées à l'algorithme :

- l'ambigüité de **permutation**. Il n'y a pas d'ordre dans les spectres purs et leur concentration associée. Le spectre S_i^T ne représente pas plus ou moins l'ensemble du jeu de donnée que le spectre S_j^T.

- l'ambigüité d'**intensité**. L'intensité du spectre pur S_i et sa concentration C_i peuvent être multipliées conjointement par les facteurs k_i et k_i^{-1}. La MCR ne peut donner qu'une concentration relative.

- l'ambigüité de **rotation**. Il est possible d'introduire une matrice inversible **R** dans les équations :

$$D = CS^T + E$$

$$D = (CR)(R^{-1}S^T) + E$$

  Puisque l'algorithme de la MCR cherche à minimiser l'erreur **E**, la matrice **R** ne fait pas varier **E** mais change la composition des résultats.

Les contraintes permettent de lever en partie ces ambigüités.

La MCR permet donc d'initier les matrices **C** et $\mathbf{S^T}$. Cette opération est utilisée dans un ensemble itératif MCR-ALS (alternating least squares) qui permet d'optimiser les résultats en fonction des contraintes utilisées. Par ailleurs, le nombres de composantes est un paramètre essentiel qui correspond véritablement à un nombre de phénomènes naturels. Il ne peut pas être déterminé arbitrairement à l'inverse du nombre de CP dans une ACP. Théoriquement, une MCR peut être effectuée avec le jeu de données de la matrice alignée. Cependant, avec un ordinateur de bureau, les calculs sont assez longs.

Un excellent complément d'information se trouve dans le livre Ruckebusch [Ruc16] et en particulier les chapitres 2 [JT16], 3 [KWB16] et 14 [HDR16]. De plus, le chapitre 7 [RR16] présente l'algorithme de l'ICA. Ces deux algorithmes cherchent tous deux à extraire des spectres purs. Pour autant, les deux approches sont différentes et génèrent parfois des incompréhensions entre les communautés.

Independent Components Analysis (ICA)

L'objectif principal de l'ICA est de retrouver les *signaux sources* mélangés à l'intérieur du jeu de données et la proportion de chacun d'eux. Pour cela, l'ICA s'interresse à la distribution de l'intensité du signal sur un histogramme (*cf.* Fig2 de D. N. Rutledge et al. [RR16]). Selon la loi centrale limite, plus l'histogramme tend vers une répartition gaussienne, plus il y a de chances que ce signal soit du bruit. Inversement, moins la répartition est gaussienne, plus le signal correspond à un signal source. De plus, comme pour une ACP, l'ICA cherche à obtenir des signaux sources indépendants, et donc orthogonaux. Contrairement à la MCR, l'ICA n'impose pas de contrainte. La phrase de [RR16] permet de bien comprendre la différence entre les deux méthodes : *This is because while MCR aims to extract the signals of pure compounds, ICA extracts signals reflecting underlying* ***independent phenomena****, which may in fact be*

combinations of the signals of several pure compounds.[4]
Il existe plusieurs algorithmes itératifs qui utilisent l'ICA, [WDH08] et [Al-15]. Cependant, l'agorithme JADE [CS93] est probablement le plus utilisé pour l'ICA, [RJ13] et [RJ15]. Un package JADE pour R est déposé et maintenu sur le CRAN. Le principe de JADE est de décomposé la matrice $\mathbf{X}$ (le jeu de données expérimentales) en utilisant les loadings d'une ACP. Ces loadings sont centrés et normés puis arrangés pour former un tenseur d'ordre 4. De façon itérative, JADE va ensuite optimiser ces loadings pour les rendre indépendants.

Cette description succincte permet très vite de comprendre que réaliser un tenseur d'ordre 4 à base de vecteur de 55000 points risque d'être problématique. L'ICA ne peut se faire que sur une matrice réduite comme l'AUC ou les pics détectés dans l'optique d'une utilisation avec un ordinateur de bureau.

3.3 Applications

J'ai testé les packages **alsace** [WCF15], **ChemometricsWithR** et **ALS** anfin d'utiliser la MCR-ALS avec R. Les packages **alsace** et **ChemometricsWithR** sont du même auteur mais le premier est encore soutenu. **alsace** utilise le package **ALS** pour le calcul et le rend compatible directement avec le format "liste" donné en sortie de la fonction *detectPeaks()*. Le package **alsace** applique par défaut uniquement la contrainte de non-negativité.

Pour la partie ICA, j'ai utilisé le package **rnirs**[5] dans le package proVOC. Ce module "ICA" est implémenté mais peu fonctionnel et ne travaille qu'avec les AUC. De plus, le package **rnirs** est en cours de transformation et sera bientôt soutenu par la communauté ChemHouse à laquelle je contribue modestement. Toutefois, j'ai échangé avec D.Rutledge qui m'a envoyé un rapport en utilisant les spectres entiers de l'expérience de lavandes (*cf* ci-dessous). Ces résultats sont difficiles à interpréter à cause du format de présentation. Les acquisition des 12 spectres ayant été concatenées. Cependant, on remarque des formes relativement similaires à celles produites par les spectres purs de la MCR que nous allons voir par la suite.

Détermination du nombre de composantes

Je n'ai pas trouvé de méthode théorique pour définir le nombre de composante. Une méthode proposée[6] pour ajuster le nombre de composés est de procéder comme avec une ACP, c'est-à-dire, déduire en fonction du pourcentage exprimé de chaque CP la séparation entre le "signal" et le "bruit". La package **alsace** propose la fonction *smallComps()* pour aider à fixer ce nombre. Je pense qu'on se trouve dans la limite mathématique de l'exercice et qu'il est bon de discuter avec le praticien. J'effectue alors plusieurs cycles en incrémentant le nombre de composantes. En interprétant les résultats, nous arrivons à déterminer une limite entre ce qui reflète un phénomène probable et ce qui est probablement du bruit.

Suivi de la floraison

Cette expérience est expliquée plus en détail dans la section 1.3. Les données ont ensuite été traitées avec proVOC puis analysées par MCR. La figure 3.3 représente les composés purs de l'expérience. Pour qui a l'habitude des analyses ACP, il paraît évident que l'interprétation de ce genre de résultat est plus intuitive.

[4] En effet, alors que la MCR vise à extraire les composés purs, l'ICA extrait les signaux reflétant des **phénomènes indépendants** sous-jacents, qui peuvent en fait être des combinaisons des signaux de plusieurs composés purs.

[5] écrit par Mathieu Lesnoff du Cirad de Montpellier

[6] mais dont je n'ai pas la source.

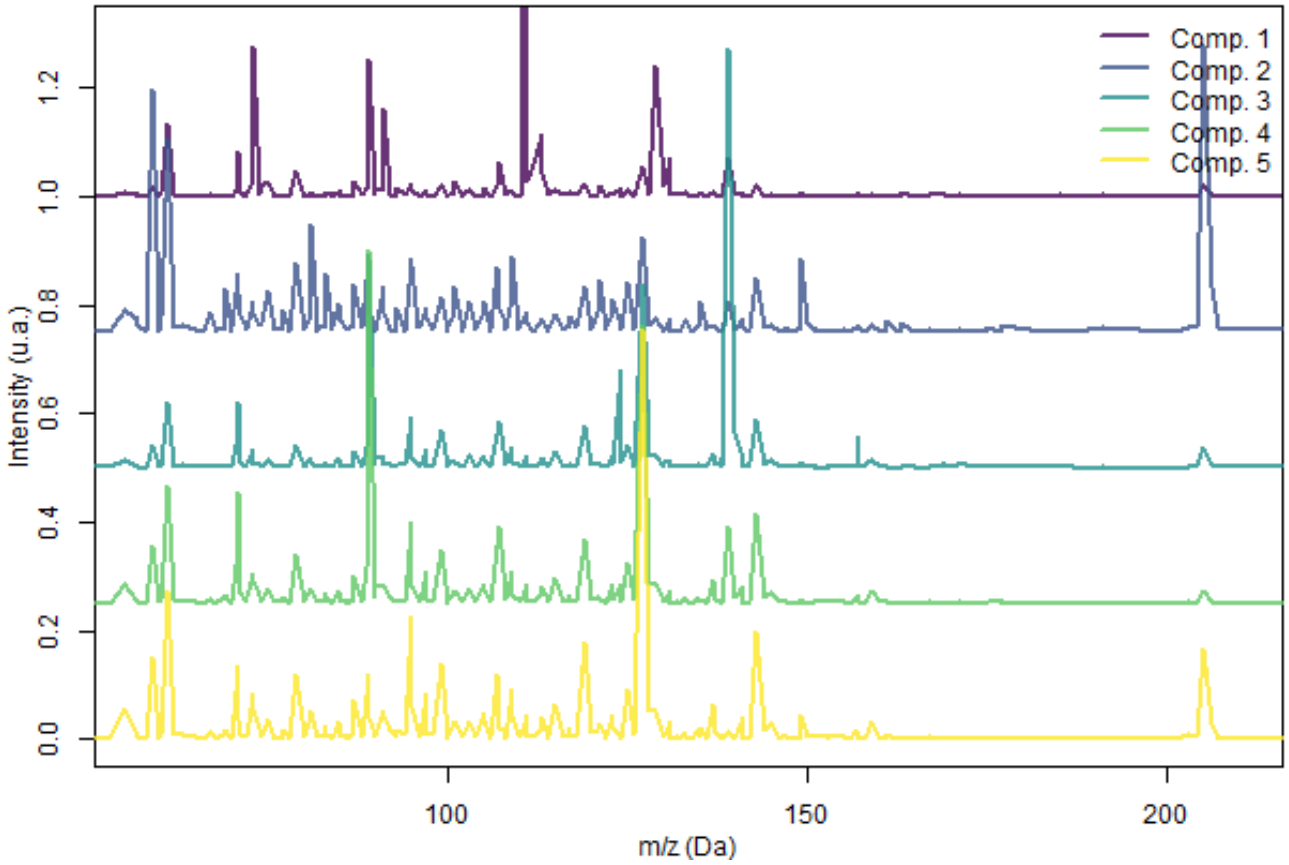

Fig. 3.3 : Spectres pures de la matrice S^i issues de la MCR effectuée sur les données du bourgeon. Les spectres ont été décalés de 0.25 u.a. pour une meilleure compréhension.

La figure 3.4 permet de voir l'évolution des cinq composantes au cours du week-end d'analyse. Il y a deux artefacts sur ces résultats. Premièrement, nous observons des oscillations très régulières. Bien que nous n'en n'ayons pas compris la nature exacte, cela vient du système de climatisation. Secondement, nous avons effectué une mesure "blanche" durant 30 minutes chaque 4h. Cela se remarque très facilement, par exemple sur la composante n°2, nous pouvons voir un pic un peu avant 22h suivi d'un autre un peu avant 2h du matin.

Passés ces artefacts, nous pouvons interpréter les phénomènes. La composante n°2 (en bleu nuit) est monotone décroissante. Ces composés sont ceux de la pollution extérieure à notre système biologique et chassés par le flux d'air zéro en quelques heures. Les composantes n°5 et 4 (jaune et vert) sont celles du bourgeon avec l'alternance phase de nuit/phase de jour (lever du soleil : 07h24). La floraison a très probablement eu lieu aux alentours de 14h. Nous avons deux indices qui laissent penser ça. Premièrement, les cinétiques des composantes 4 et 5 sont perturbées entre 14 et 18h. Secondement, le bouquet d'odeurs de la fleur d'amande, traduit par la composante n°3 augmente à partir de ce moment. En outre, la composante n°1 montre des phénomènes très ponctuels que nous n'avons pas identifiés.

Ces hypothèses d'interprétation nécessitent encore une analyse fine d'identification des composantes mais ces résultats ont été un moteur d'intérêt à la fois pour la PTR-ToF-MS et pour la MCR-ALS.

Analyse de l'émission journalière des lavandes

L'expérience est décrite dans la section 1.3. Pour analyser cette expérience, j'ai procédé de la même façon que pour l'analyse de la floraison de l'amandier. Après discussion avec la doctorante

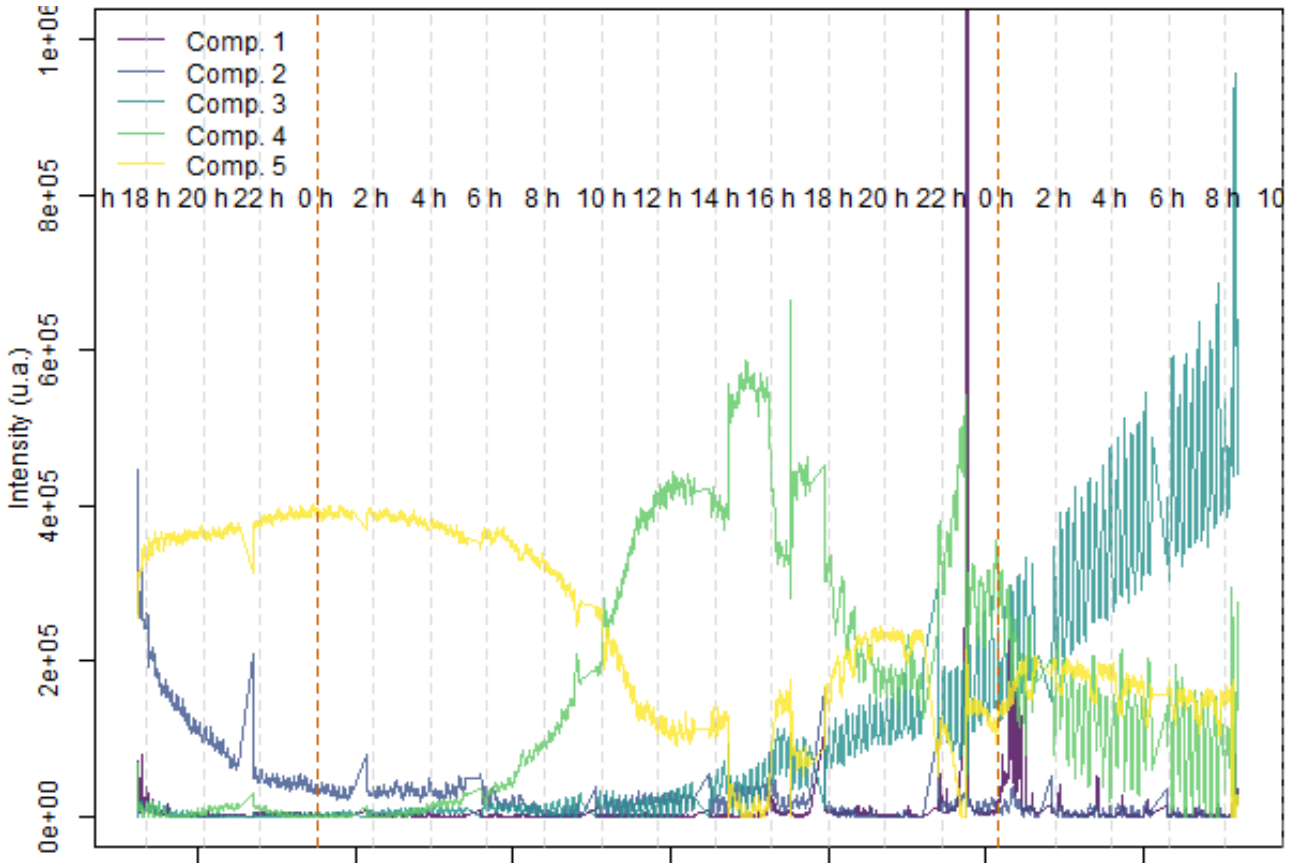

FIG. 3.4 : Cinétique de la matrice des concentrations C^i issues de la MCR effectuée sur les données du bourgeon.

expérimentatrice, nous avons conclu que cinq composantes étaient un nombre optimum pour interpréter les résultats. Ces composantes sont représentées sur la figure 3.5. J'ai rendu plus visible les composantes n°2 (bleu nuit) et 4 (vert) qui sont les plus intéressantes pour notre cas. La figure 3.6 permet l'interprétation de ces résultats. Pour faciliter la lecture, j'ai placé sur une même colonne les unités expérimentales enregistrées le même jour et sur une même ligne les unités expérimentales enregistrées dans une même chambre. Les lignes 1 à 4 correspondent aux chambres 1 à 4. Les colonnes 1 à 3 correspondent aux jours 1, 3 et 5. Les abréviations "lv" et "bl" désignent respectivement les unités expérimentales où se trouvent une lavande et celles servant de blanc.

La composante n°1 (violet) est un artefact qui se produit principalement dans les unités "blanches" avec un fort pic à 11h. Quelques fluctuations se retrouvent dans les autres unités, principalement lors du jour 1. La composante n°3 (turquoise) est un bruit de fond que l'on retrouve dans toutes les analyses[7]. La composante n°5 (jaune) est plus compliquée à interpréter. Elle se retrouve dans les unités "lavandes" mais pas dans les unités "blanches".

Une fois ces trois composantes analysées, nous pouvons nous attarder sur les composantes n°2 et 4. Les expériences ayant été effectuées en juin, j'ai placé deux lignes verticales pour marquer le lever (en jaune) et le coucher (en bleu) du soleil. Le lever du jour est parfaitement synchronisé avec l'émission des COV n°2 et 4. La composante 2 est émise durant toute la journée. La plante arrive en quelques heures à un plateau d'émission puis anticipe une diminution avec le déclin du jour. Il est très intéressant de comparer les profils d'un même jour. Par exemple, nous observons un creux dans les trois profils du jour 1 aux alentours de 15h. La composante

[7] contrairement à l'expérience précédente, ces échantillons sont mal isolés du reste de la serre. De plus la serre sert de stockage des autres plants de lavandes et de figuiers

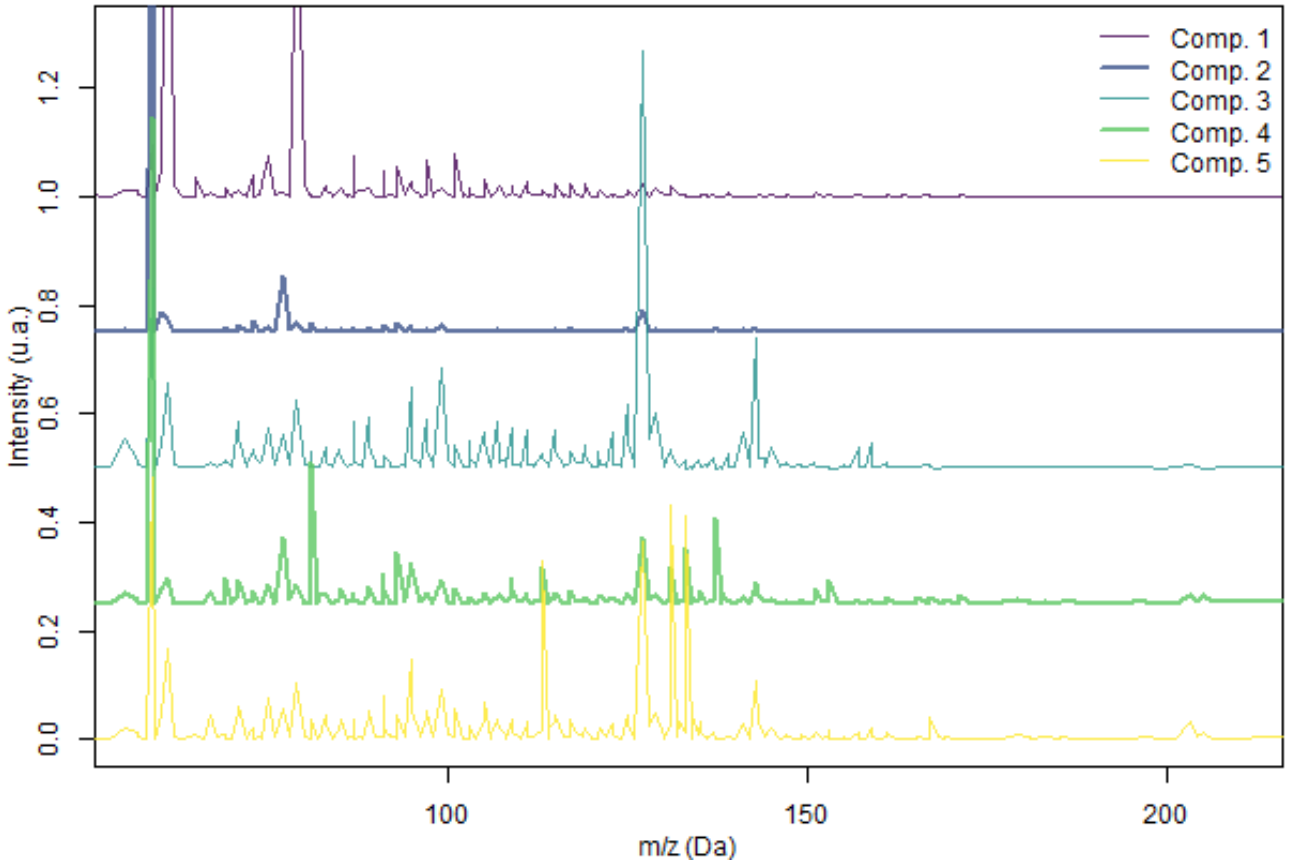

Fig. 3.5 : Spectres pures de la matrice S^i issues de la MCR effectuée sur les données des lavandes. Les spectres ont été décalés de 0.25 u.a. pour une meilleure compréhension.

4 suit un profil d'émission différent de la composante 2. L'émission des COV arrive en 2h à un maximum puis décroit très rapidement par la suite pour retrouver une émission quasi nulle à midi.

Le système circadien de la lavande est bien connu, tout comme le fait que sa production de COV dépend des facteurs extérieurs comme l'ensoleillement [Gui10]. Cependant, le résultat d'une double émission au sein d'une même journée n'a jamais, au meilleur de notre connaissance, été mis si distinctement en valeur. De plus, ce résultat se retrouve sur 9 unités expérimentales réparties en 3 jours et 3 chambres et contrôlées avec un blanc. Si l'expérience de la floraison n'était qu'un test, cette expérience est très concluante.

3.4 Discussions

Les analyses MCR sont satisfaisantes mais nécessitent encore d'être complétées par un travail d'indexations des pics. Cela permettra de consolider d'un point de vue biologique l'hypothèse du double cycle d'émission des lavandes. De mon côté, je peux renforcer l'analyse par plusieurs façon. D'une part, nous avons vu que les rapports automatisés de l'ACP effectués par **proVOC** sont assez sommaires. Le changement de structure du package permettra d'appliquer plus facilement des pré-traitements et des sélections de variables à l'ACP. Cet outil reste un excellent moyen d'exploration à l'aveugle des jeux de données, notamment pour détecter les outliers.
Ensuite, il reste un travail, probablement à effectuer en commun avec les membres de Chemhouse, pour optimiser les analyses MCR et ICA[8] puis en croisant les résultats de

[8]ici le rapport de l'analyse effectuée par Douglas Rutledge.

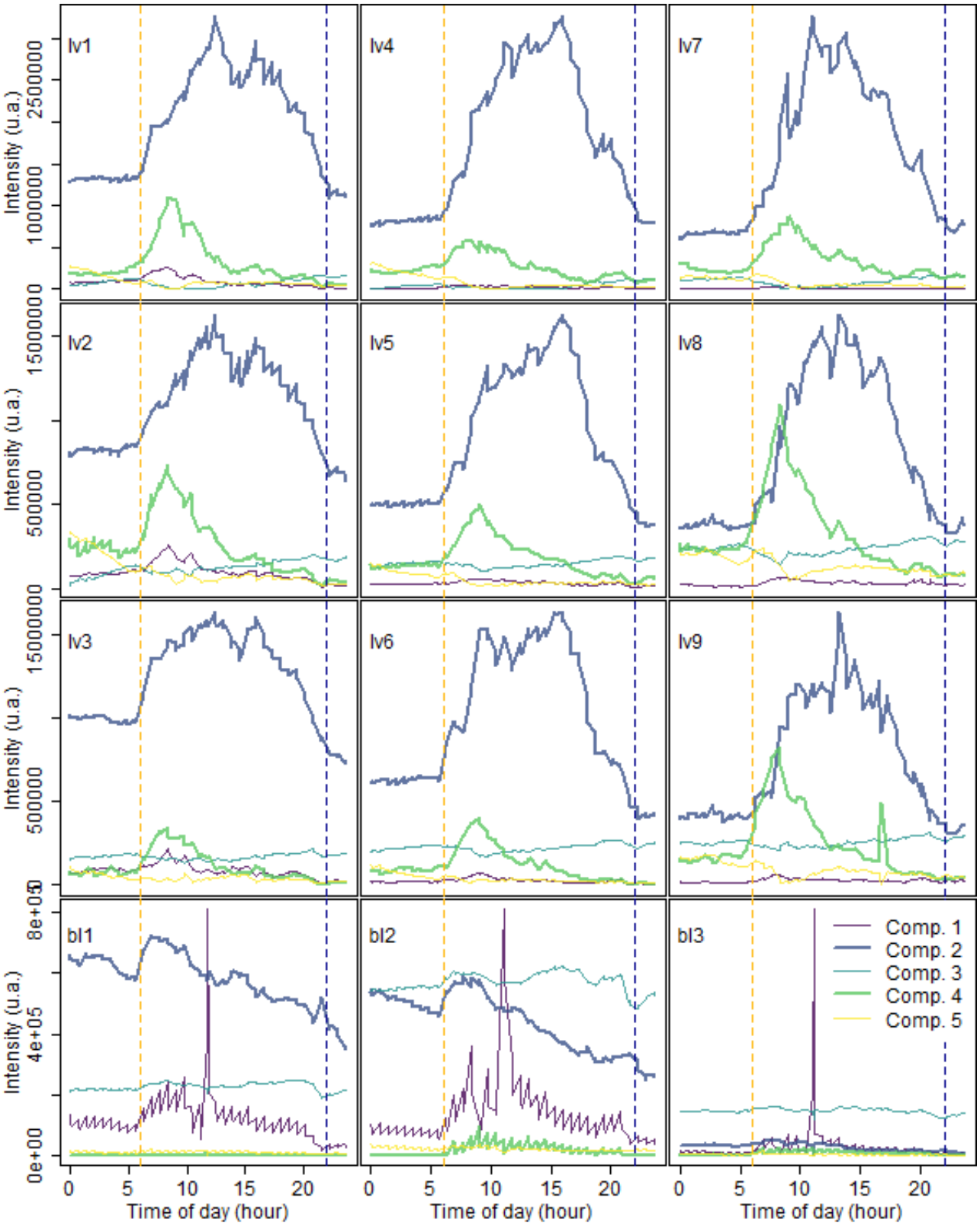

FIG. 3.6 : Cinétique de la matrice des concentrations C^i issues de la MCR effectuée sur les données des lavandes.

l'analyse biologique avec les spectres purs (MCR) et les signaux sources (ICA) afin de définir au plus juste quel algorithme décrit le mieux les lois naturelles.

Chapitre 4

Conclusion

Cette période de ce stage un peu particulier s'achève et ce rapport essaie de rendre compte au plus juste possible du travail effectué durant ces mois de printemps et d'été. Nous avons pu voir qu'en amont du traitement de la data, il y a l'acquisition de celle-ci et que cette tâche était au moins aussi importante. J'ai passé sous silence une grosses partie des problèmes techniques rencontrés et autres pannes[1]. Pour autant, ces contraintes font parties du métier et l'analyste doit savoir s'y plier en ne maudissant que modérément l'expérimentateur, lorsque ce n'est pas lui-même. Il doit être présent dans le projet dés le début et participer à l'élaboration théorique de l'expérience.

Il est aussi crucial de maîtriser la technique de l'instrument de mesure. Par exemple, en spectroscopie optique les pics peuvent avoir un décalage en longueur d'ondes qui traduit un effet de contrainte sur l'échantillon. Ce décalage ne peut pas exister en spectrométrie de masse, ou alors de façon discrète modulo 1,007 Da qui correspond à la masse d'un proton.

Une fois cela acquis et mis en œuvre, les outils numériques ne manquent pas pour décortiquer les données. Je pourrais en ajouter à l'arsenal de proVOC, comme par exemple la PLS (et ses variantes) pour effectuer des régressions. Je me suis attardé pour l'instant à la découverte de la MCR et de l'ICA qui sont moins connues des utilisateurs et qui, de fait, demandent un plus grand travail d'accompagnement pour leur utilisation. Nous avons pu voir les avantages de telles méthodes.

Avant cette année de master, je connaissais très peu de choses sur "les modèles linéaires généralisés", les "séries temporelles et spatiales", les "plans d'expériences". Bien qu'il me faudra quelques contorsions du cortex pour intégrer ces connaissances théoriques obtenues à ma pratique quotidienne, je ne doute pas qu'elles me seront utiles dans un futur proche.

À travers ce rapport, j'espère avoir synthétisé correctement six mois de travaux effectués par un physicien dans un contexte à la frontière entre les mathématiques, l'environnement et la chimie. Cette interdisciplinarité, que je recherche depuis la fin de mon master 2 de photonicien, demande un apprentissage lent mais constant et une volonté permanente de vouloir sortir de sa zone de confort. C'est avant tout un goût pour la découverte et des aventures scientifiques que je recommande à chacun.

[1] mention spéciale à la climatisation des serres, H-S plusieurs semaines à partir de mi-juillet.

Bibliographie

[AB20] J. Artho et R. Batro. *Rotting Christ*. Batro'Games. T. 1. Nov. 2020. URL : http://www.legrog.org/jeux/rotting-christ/rotting-christ-fr (cf. p. vii).

[BA01] L. Babcock et N. Adams. *Advances in Gas Phase Ion Chemistry, Volume 4 - 1st Edition*. 1re éd. T. 4. 2001. URL : https://www-elsevier-com.inee.bib.cnrs.fr/books/advances-in-gas-phase-ion-chemistry/babcock/978-0-444-50929-1 (visité le 12/08/2021) (cf. p. 10).

[Bla+04] R. S. Blake, C. Whyte, C. O. Hughes, A. M. Ellis et P. S. Monks. « Demonstration of proton-transfer reaction time-of-flight mass spectrometry for real-time analysis of trace volatile organic compounds ». eng. In : *Anal Chem* 76.13 (juill. 2004), p. 3841-3845. DOI : 10.1021/ac0498260 (cf. p. 10).

[Boe+15] C. Boettiger, S. Chamberlain, E. Hart et K. Ram. « Building Software, Building Community : Lessons from the rOpenSci Project ». en. In : *Journal of Open Research Software* 3.1 (nov. 2015). Number : 1 Publisher : Ubiquity Press, e8. DOI : 10.5334/jors.bu. URL : http://openresearchsoftware.metajnl.com/articles/10.5334/jors.bu/ (visité le 23/08/2021) (cf. p. 5).

[Cap+10] L. Cappellin, F. Biasioli, A. Fabris, E. Schuhfried, C. Soukoulis, T. D. Märk et F. Gasperi. « Improved mass accuracy in PTR-TOF-MS : Another step towards better compound identification in PTR-MS ». en. In : *International Journal of Mass Spectrometry* 290.1 (fév. 2010), p. 60-63. DOI : 10.1016/j.ijms.2009.11.007. URL : https://linkinghub.elsevier.com/retrieve/pii/S1387380609003571 (visité le 03/05/2021) (cf. p. 10).

[Cap+11] L. Cappellin, F. Biasioli, P. M. Granitto, E. Schuhfried, C. Soukoulis, F. Costa, T. D. Märk et F. Gasperi. « On data analysis in PTR-TOF-MS : From raw spectra to data mining ». en. In : *Sensors and Actuators B : Chemical* 155.1 (juill. 2011), p. 183-190. DOI : 10.1016/j.snb.2010.11.044. URL : https://linkinghub.elsevier.com/retrieve/pii/S0925400510009135 (visité le 03/05/2021) (cf. p. 10).

[CS93] J. Cardoso et A. Souloumiac. « Blind beamforming for non-gaussian signals ». en. In : *IEE Proc. F Radar Signal Process. UK* 140.6 (1993), p. 362. DOI : 10.1049/ip-f-2.1993.0054. URL : https://digital-library.theiet.org/content/journals/10.1049/ip-f-2.1993.0054 (visité le 24/08/2021) (cf. p. 23).

[Cla+21] M. S. Claflin, D. Pagonis, Z. Finewax, A. V. Handschy, D. A. Day, W. L. Brown, J. T. Jayne, D. R. Worsnop, J. L. Jimenez, P. J. Ziemann, J. de Gouw et B. M. Lerner. « An in situ gas chromatograph with automatic detector switching between PTR- and EI-TOF-MS : isomer-resolved measurements of indoor air ». In : *Atmospheric Measurement Techniques* 14.1 (2021), p. 133-152. DOI : 10.5194/amt-14-133-2021. URL : https://amt.copernicus.org/articles/14/133/2021/ (cf. p. 10).

[Deu+19] Z. Deuscher, I. Andriot, E. Sémon, M. Repoux, S. Preys, J.-M. Roger, R. Boulanger, H. Labouré et J.-L. Le Quéré. « Volatile compounds profiling by using proton transfer reaction-time of flight-mass spectrometry (PTR-ToF-MS). The case study of dark chocolates organoleptic differences ». en. In : *J Mass Spectrom* 54.1 (jan. 2019), p. 92-119. doi : 10.1002/jms.4317. url : http://doi.wiley.com/10.1002/jms.4317 (visité le 17/05/2021) (cf. p. 10).

[Enn+05] C. J. Ennis, J. C. Reynolds, B. J. Keely et L. J. Carpenter. « A hollow cathode proton transfer reaction time of flight mass spectrometer ». en. In : *International Journal of Mass Spectrometry* 247.1 (déc. 2005), p. 72-80. doi : 10.1016/j.ijms.2005.09.008. url : https://www.sciencedirect.com/science/article/pii/S1387380605002472 (visité le 12/08/2021) (cf. p. 10).

[Fis+21] B. Fischer, M. Smith, G. Pau, M. Morgan et D. v. Twisk. *rhdf5 : R Interface to HDF5*. 2021. doi : 10.18129/B9.bioc.rhdf5. url : https://bioconductor.org/packages/rhdf5/ (visité le 18/08/2021) (cf. p. 14).

[GMH10] M. Graus, M. Müller et A. Hansel. « High resolution PTR-TOF : Quantification and formula confirmation of VOC in real time ». en. In : *J Am Soc Mass Spectrom* 21.6 (juin 2010), p. 1037-1044. doi : 10.1016/j.jasms.2010.02.006. url : https://pubs.acs.org/doi/10.1021/jasms.8b03774 (visité le 03/05/2021) (cf. p. 10, 11).

[Gui10] Y. Guitton. « Diversité des composés terpéniques volatils au sein du genre Lavandula : aspects évolutifs et physiologiques ». fr. Thèse de doct. Université Jean Monnet - Saint-Etienne, déc. 2010. url : https://tel.archives-ouvertes.fr/tel-00675866 (visité le 26/08/2021) (cf. p. 26).

[Ham13] R. Hamel. « L'anglais, langue unique pour les sciences ? Le rôle des modèles plurilingues dans la recherche, la communication scientifique et l'enseignement supérieur ». In : *Synergies Europe* 8 (jan. 2013), p. 53-66 (cf. p. 4).

[Han+95] A. Hansel, A. Jordan, R. Holzinger, P. Prazeller, W. Vogel et W. Lindinger. « Proton transfer reaction mass spectrometry : on-line trace gas analysis at the ppb level ». en. In : *International Journal of Mass Spectrometry and Ion Processes*. Honour Biography David Smith 149-150 (nov. 1995), p. 609-619. doi : 10.1016/0168-1176(95)04294-U. url : https://www.sciencedirect.com/science/article/pii/016811769504294U (visité le 12/08/2021) (cf. p. 9, 11).

[Hol15] R. Holzinger. « PTRwid : A new widget tool for processing PTR-TOF-MS data ». en. In : *Atmos. Meas. Tech.* 8.9 (sept. 2015), p. 3903-3922. doi : 10.5194/amt-8-3903-2015. url : https://amt.copernicus.org/articles/8/3903/2015/ (visité le 17/05/2021) (cf. p. 10).

[HDR16] S. Hugelier, O. Devos et C. Ruckebusch. « Chapter 14 – A smoothness constraint in multivariate curve resolution-alternating least squares of spectroscopy data ». In : *Data Handling in Science and Technology*. T. 30. Resolving spectral mixtures - With applications from ultrafast time-resolved spectroscopy to super-resolution imaging. 2016, p. 453-476. doi : 10.1016/B978-0-444-63638-6.00014-0. url : https://hal.archives-ouvertes.fr/hal-01386913 (visité le 26/05/2021) (cf. p. 22).

[Ino+06] S. Inomata, H. Tanimoto, N. Aoki, J. Hirokawa et Y. Sadanaga. « A novel discharge source of hydronium ions for proton transfer reaction ionization : design, characterization, and performance ». en. In : *Rapid Communications in Mass Spectrometry* 20.6 (2006). _eprint : https ://analyticalsciencejournals.onlinelibrary.wiley.com/d p. 1025-1029. doi : 10.1002/rcm.2405. url : https://analyticalsciencejournals.

onlinelibrary . wiley . com / doi / abs / 10 . 1002 / rcm . 2405 (visité le 12/08/2021) (cf. p. 10).

[Jor+09] A. Jordan, S. Haidacher, G. Hanel, E. Hartungen, L. Märk, H. Seehauser, R. Schottkowsky, P. Sulzer et T. Märk. « A high resolution and high sensitivity proton-transfer-reaction time-of-flight mass spectrometer (PTR-TOF-MS) ». en. In : *International Journal of Mass Spectrometry* 286.2-3 (sept. 2009), p. 122-128. doi : 10 . 1016 / j . ijms . 2009 . 07 . 005. url : https : //linkinghub.elsevier.com/retrieve/pii/S1387380609002371 (visité le 30/04/2021) (cf. p. 10).

[JT16] A. de Juan et R. Tauler. « Chapter 2 - Multivariate Curve Resolution-Alternating Least Squares for Spectroscopic Data ». en. In : *Data Handling in Science and Technology*. Sous la dir. de C. Ruckebusch. T. 30. Resolving Spectral Mixtures. Elsevier, jan. 2016, p. 5-51. doi : 10.1016/B978-0-444-63638-6.00002-4. url : https://www.sciencedirect.com/science/article/pii/B9780444636386000024 (visité le 26/05/2021) (cf. p. 22).

[JJT14] A. de Juan, J. Jaumot et R. Tauler. « Multivariate Curve Resolution (MCR). Solving the mixture analysis problem ». en. In : *Anal. Methods* 6.14 (2014), p. 4964-4976. doi : 10.1039/C4AY00571F. url : http://xlink.rsc.org/?DOI=C4AY00571F (visité le 19/05/2021) (cf. p. 21).

[Kre+18] J. Krechmer, F. Lopez-Hilfiker, A. Koss, M. Hutterli, C. Stoermer, B. Deming, J. Kimmel, C. Warneke, R. Holzinger, J. Jayne, D. Worsnop, K. Fuhrer, M. Gonin et J. de Gouw. « Evaluation of a New Reagent-Ion Source and Focusing Ion–Molecule Reactor for Use in Proton-Transfer-Reaction Mass Spectrometry ». In : *Anal. Chem.* 90.20 (oct. 2018). Publisher : American Chemical Society, p. 12011-12018. doi : 10 . 1021 / acs . analchem . 8b02641. url : https://doi.org/10.1021/acs.analchem.8b02641 (visité le 13/08/2021) (cf. p. 11).

[KWB16] S. Kucheryavskiy, W. Windig et A. Bogomolov. « Chapter 3 - Spectral Unmixing Using the Concept of Pure Variables ». en. In : *Data Handling in Science and Technology*. Sous la dir. de C. Ruckebusch. T. 30. Resolving Spectral Mixtures. Elsevier, jan. 2016, p. 53-99. doi : 10.1016/B978-0-444-63638-6.00003-6. url : https://www.sciencedirect.com/science/article/pii/B9780444636386000036 (visité le 27/05/2021) (cf. p. 22).

[Lag+94] A. Lagg, J. Taucher, A. Hansel et W. Lindinger. « Applications of proton transfer reactions to gas analysis ». en. In : *International Journal of Mass Spectrometry and Ion Processes* 134.1 (juin 1994), p. 55-66. doi : 10.1016/0168-1176(94)03965-8. url : https : / / www . sciencedirect . com / science / article / pii / 0168117694039658 (visité le 12/08/2021) (cf. p. 9).

[LLA91] W. Lindinger, K. Leiter et M. Andriollo. « Industrial Multicomponent Gas Analysis ». In : *Ionen-Technik GmbH* 20.7 (1991), p. 24-25 (cf. p. 9).

[LHJ98] W. Lindinger, A. Hansel et A. Jordan. « On-line monitoring of volatile organic compounds at pptv levels by means of proton-transfer-reaction mass spectrometry (PTR-MS) medical applications, food control and environmental research ». en. In : *International Journal of Mass Spectrometry and Ion Processes* 173.3 (fév. 1998), p. 191-241. doi : 10.1016/S0168-1176(97)00281-4. url : https://www.sciencedirect.com/science/article/pii/S0168117697002814 (visité le 12/08/2021) (cf. p. 9).

[LHP93] W. Lindinger, J. Hirber et H. Paretzke. « An ion/molecule-reaction mass spectrometer used for on-line trace gas analysis ». en. In : *International Journal of Mass Spectrometry and Ion Processes* 129 (nov. 1993), p. 79-88. doi : 10.1016/0168-1176(93)87031-M. url : https://www.sciencedirect.com/science/article/pii/016811769387031M (visité le 12/08/2021) (cf. p. 9).

[LJ98] W. Lindinger et A. Jordan. « Proton-transfer-reaction mass spectrometry (PTR–MS) : on-line monitoring of volatile organic compounds at pptv levels ». en. In : *Chem. Soc. Rev.* 27.5 (jan. 1998). Publisher : The Royal Society of Chemistry, p. 347-375. doi : 10.1039/A827347Z. url : https://pubs.rsc.org/en/content/articlelanding/1998/cs/a827347z (visité le 30/04/2021) (cf. p. 10).

[Maj+18] T. Majchrzak, W. Wojnowski, M. Lubinska-Szczygeł, A. Różańska, J. Namieśnik et T. Dymerski. « PTR-MS and GC-MS as complementary techniques for analysis of volatiles : A tutorial review ». en. In : *Analytica Chimica Acta* 1035 (déc. 2018), p. 1-13. doi : 10.1016/j.aca.2018.06.056. url : https://linkinghub.elsevier.com/retrieve/pii/S0003267018308213 (visité le 17/05/2021) (cf. p. 13).

[Pic+18] V. Picaud, J.-F. Giovannelli, C. Truntzer, J.-P. Charrier, A. Giremus, P. Grangeat et C. Mercier. « Linear MALDI-ToF simultaneous spectrum deconvolution and baseline removal ». In : *BMC Bioinformatics* 19.1 (avr. 2018), p. 123. doi : 10.1186/s12859-018-2116-3. url : https://doi.org/10.1186/s12859-018-2116-3 (visité le 29/04/2021) (cf. p. 19).

[Rog+11] J. Roger, B. Palagos, D. Bertrand et E. Fernandez-Ahumada. « CovSel : Variable selection for highly multivariate and multi-response calibration ». en. In : *Chemometrics and Intelligent Laboratory Systems* 106.2 (avr. 2011), p. 216-223. doi : 10.1016/j.chemolab.2010.10.003. url : https://linkinghub.elsevier.com/retrieve/pii/S0169743910001978 (visité le 03/05/2021) (cf. p. 19).

[Ruc16] C. Ruckebusch. *Resolving Spectral Mixtures, Volume 30 - 1st Edition*. 1re éd. T. 30. Août 2016. url : https://www.sciencedirect.com/bookseries/data-handling-in-science-and-technology/vol/30/suppl/C (visité le 24/08/2021) (cf. p. 22).

[RR16] D. N. Rutledge et D. Rimbaud Bouveresse. « Chapter 7 - Independent Components Analysis : Theory and Applications ». en. In : *Data Handling in Science and Technology*. Sous la dir. de C. Ruckebusch. T. 30. Resolving Spectral Mixtures. Elsevier, jan. 2016, p. 225-277. doi : 10.1016/B978-0-444-63638-6.00007-3. url : https://www.sciencedirect.com/science/article/pii/B9780444636386000073 (visité le 26/05/2021) (cf. p. 22).

[RJ13] D. Rutledge et D. Jouan-Rimbaud Bouveresse. « Independent Components Analysis with the JADE algorithm ». en. In : *TrAC Trends in Analytical Chemistry* 50 (oct. 2013), p. 22-32. doi : 10.1016/j.trac.2013.03.013. url : https://linkinghub.elsevier.com/retrieve/pii/S0165993613001222 (visité le 24/08/2021) (cf. p. 23).

[RJ15] D. Rutledge et D. Jouan-Rimbaud Bouveresse. « Corrigendum to 'Independent Components Analysis with the JADE algorithm' ». en. In : *TrAC Trends in Analytical Chemistry* 67 (avr. 2015), p. 220. doi : 10.1016/j.trac.2015.02.001. url : https://linkinghub.elsevier.com/retrieve/pii/S0165993615000345 (visité le 24/08/2021) (cf. p. 23).

[Al-15] A. Al-Saegh. « Independent Component Analysis for Separation of Speech Mixtures : A Comparison Among Thirty Algorithms ». en. In : *IJEEE* 11.1 (juin 2015), p. 1-9. doi : 10.37917/ijeee.11.1.1. url : http://ijeee.edu.iq/Papers/Vol11-Issue1/102709.pdf (visité le 24/08/2021) (cf. p. 23).

[Tan+07] H. Tanimoto, N. Aoki, S. Inomata, J. Hirokawa et Y. Sadanaga. « Development of a PTR-TOFMS instrument for real-time measurements of volatile organic compounds in air ». en. In : *International Journal of Mass Spectrometry* 263.1 (mai 2007), p. 1-11. doi : 10.1016/j.ijms.2007.01.009. url : https://www.sciencedirect.com/science/article/pii/S1387380607000231 (visité le 12/08/2021) (cf. p. 10).

[WDH08] G. Wang, Q. Ding et Z. Hou. « Independent component analysis and its applications in signal processing for analytical chemistry ». en. In : *TrAC Trends in Analytical Chemistry* 27.4 (avr. 2008), p. 368-376. doi : 10.1016/j.trac.2008.01.009. url : https://linkinghub.elsevier.com/retrieve/pii/S0165993608000101 (visité le 24/08/2021) (cf. p. 23).

[WCF15] R. Wehrens, E. Carvalho et P. D. Fraser. « Metabolite profiling in LC–DAD using multivariate curve resolution : the alsace package for R ». en. In : *Metabolomics* 11.1 (fév. 2015), p. 143-154. doi : 10.1007/s11306-014-0683-5. url : https://doi.org/10.1007/s11306-014-0683-5 (visité le 26/08/2021) (cf. p. 23).

Table des figures

Abstract The PTR-ToF-MS is used to analyze the emission kinetics of VOCs. The CEFE laboratory is equipped with such an instrument since 2019. This memoir deals with the exploitation of datasets in the context of chemical ecology. Several experiments are presented as well as the theory of PTR-ToF-MS. The numerical tools are grouped in the R package: proVOC. A particular focus is made on the use of MCR.

Keywords Chimiometry, Open Science, PTR-ToF-MS, Spectrometry, Environment.

Résumé La PTR-ToF-MS permet d'analyser la cinétique d'émission des COVs. Le laboratoire CEFE s'est doté d'un tel instrument en fin d'année 2019. Ce mémoire traite de l'exploitation de jeu de données dans le cadre de l'écologie chimique. Plusieurs expériences sont présentées ainsi que la théorie de la PTR-ToF-MS. Les outils numériques sont regroupés dans le package R : proVOC. Un focus particulier est fait sur l'utilisation de la MCR.

Mots-Clés Chimiométrie, PTR-ToF-MS, Science ouverte, Spectrométrie, Environnement.