Year : 2013

Etude de l'approvisionnement d'une banque de données avec les résultats provenant de méthodes analytiques différentes dans le cadre du profilage chimique de produits stupéfiants

Julian Broséus

UNIL | Université de Lausanne

Faculté de Droit et des Sciences Criminelles
Ecole des Sciences Criminelles
Institut de Police Scientifique

Etude de l'approvisionnement d'une banque de données avec les résultats provenant de méthodes analytiques différentes dans le cadre du profilage chimique de produits stupéfiants

THÈSE DE DOCTORAT

présentée à la
Faculté de Droit et des Sciences Criminelles
de l'Université de Lausanne

pour l'obtention du grade de
Docteur ès Sciences en science forensique
par

Julian Broséus

Directeur de thèse
Prof. Pierre Esseiva

LAUSANNE
2013

UNIL | Université de Lausanne
Ecole des sciences criminelles
bâtiment Batochime
CH-1015 Lausanne

IMPRIMATUR

A l'issue de la soutenance de thèse, le Jury autorise l'impression de la thèse de M. Julian Broséus, candidat au doctorat en science forensique, intitulée

« Etude de l'approvisionnement d'une banque de données avec les résultats provenant de méthodes analytiques différentes dans le cadre du profilage chimique de produits stupéfiants »

Le Président du Jury

Professeur Pierre Margot

Lausanne, le 27 septembre 2013

UNIL | Université de Lausanne

Faculté de Droit et des Sciences Criminelles

Ecole des Sciences Criminelles

Institut de Police Scientifique

Etude de l'approvisionnement d'une banque de données avec les résultats provenant de méthodes analytiques différentes dans le cadre du profilage chimique de produits stupéfiants

THÈSE DE DOCTORAT

Julian Broséus

Série Criminalistique LIX

ISBN 2-940098-63-8

Remerciements

Ce travail de thèse a été réalisé à l'Institut de Police Scientifique de l'Université de Lausanne. Le jury était composé de Monsieur le Professeur Pierre MARGOT, directeur de l'Institut de Police Scientifique et président du jury, de Monsieur le Professeur Pierre ESSEIVA, directeur de thèse et professeur à l'Institut de Police Scientifique, de Monsieur le Professeur Olivier DELEMONT, professeur à l'Institut de Police Scientifique, de Monsieur le Professeur Serge RUDAZ, professeur à la Section des Sciences Pharmaceutiques de l'Université de Genève et de Monsieur le Docteur Didier THIEBAUT, Président de l'AFSEP et Chargé de Recherche au CNRS. Je remercie l'ensemble du jury de l'intérêt qu'ils ont porté à mes travaux et leur suis particulièrement reconnaissant pour les précieux conseils et corrections.

Je remercie toutes les personnes qui m'ont soutenu et qui m'ont fait bénéficier de leurs connaissances tout au long de cette recherche. Je tiens à remercier tout particulièrement :

Monsieur Frédéric ANGLADA, que j'ai eu la chance de côtoyer lors de mes débuts en tant que doctorant et dont je retiendrai la passion contagieuse pour l'analyse des produits stupéfiants et l'incroyable positivisme quelles que soient les circonstances.

Le Professeur Pierre MARGOT, qui m'a donné ma chance en m'engageant comme assistant-doctorant et qui m'a fait l'honneur de présider le jury de thèse.

Le Professeur Pierre ESSEIVA, sous la direction duquel j'ai eu beaucoup de plaisir à travailler et que je remercie pour sa constante disponibilité.

Le Docteur Benjamin DEBRUS, à qui je dois énormément dans l'écriture du code informatique dans R et avec qui j'ai eu beaucoup de plaisir à collaborer.

Toutes les personnes du groupe stupéfiants avec lesquelles j'ai pu travailler, tout particulièrement Mesdames Laëtitia GASTE et Natacha GENTILE.

Tous mes collègues, amis, anciens étudiants et étudiants de l'Institut de Police Scientifique qui m'ont suivi et soutenu lors de cette aventure.

Chorizo, Grorwin, La Bûche et Sauterelle pour avoir accepté de relire le manuscrit mais avant tout pour leur soutien inconditionnel et leur amitié sans faille.

La Ritale, pour sa compréhension, son risotto et les nombreux fous rires partagés lors des moments critiques qui ont rendu la dernière ligne droite plus facile.

Mes parents, qui m'ont offert la possibilité de poursuivre mes études en Suisse dans les meilleures conditions et qui depuis le début n'ont eu de cesse de m'encourager.

Summary

The analysis of illicit drug specimens includes the identification and often the quantification of the main active substance. However, several laboratories also propose to include a wider perspective with the profiling of these specimens by extracting a chemical signature. Chemical profiling embraces three main aspects: the separation and the detection of the compounds present in the sample, the determination of the chemical profile by selecting the target compounds, and the build-up of a memory of profiles compiled in a database.

The latter is used to find unsuspected links between different police seizures. Illicit drug databases are maintained continuously and chemical profiles retrospective queries can be performed to identify unsuspected connections during police investigation between samples seized in different cases. These connections are dedicated to support law enforcement investigation when combined with traditional police information thanks to both tactical and strategic intelligence.

The fight against illicit drugs trafficking entails large-scaled investigations that require the exchange of information. One strategy for the exchange of information based on chemical profiles is to build a shared database fed by different laboratories. To ensure the similarity of chemical profiles, the implementation of a strict harmonised analytical methodology is nowadays promoted, from sample preparation to sample analysis and data treatment. This approach is named analytical methods harmonization and implies using the same analytical method which is defined by its analytical technology (i.e. separation and detection technologies), apparatus selected for performing analysis (brand and model) along with consumables and analysis parameters set for technologies of separation and detection.

This approach is restrictive and time-consuming due to the intensive laboratory work required to validate results between different laboratories. Moreover, it is also problematic for the long term use of the database in consequence of the analytical inertia and potential loss of information which may be involved. According to this approach, it is quite impossible to consider a new analytical method (e.g. due to the use of a new brand of apparatus) or to integrate new analytical technologies while feeding the same database due to the different nature of results coming from different analytical methods. It is then compulsory to create a

new database based on the modified analytical method and consequently reset the analytical knowledge previously established during several years.

In the present study, another approach than analytical methods harmonization overcoming the drawbacks of the latter was investigated. This approach, described as analytical results harmonization, was evaluated. In such an approach, implementation of a methodology for adjusting analytical results coming from different analytical methods and thus ensuring their similarity is proposed before using a common database.

Résumé

Dans les laboratoires forensiques, les analyses journalières réalisées sur les produits stupéfiants concernent identification, quantification et détermination de la signature chimique. Cette approche implique la création de banques de données compilant les résultats analytiques obtenus.

Les banques de données de produits stupéfiants sont approvisionnées continuellement et permettent la recherche rétrospective de liens chimiques entre différentes saisies policières, non suspectés a priori lors de l'enquête policière. Ces renseignements soutiennent l'investigation des forces de police et doivent être combinés aux informations policières traditionnelles.

A un niveau international, la stratégie prônée pour l'échange en temps réel d'informations liées aux profils chimiques consiste en la création de banques de données harmonisées et partagées par les laboratoires des pays participants. Pour y parvenir, l'utilisation d'une même méthode analytique est recommandée, celle-ci étant définie par ses technologies d'analyses de séparation et de détection, par l'appareillage sélectionné pour réaliser les analyses (marque et modèle) et par les paramètres analytiques décrivant chacune des technologies d'analyse.

Cette approche s'avère contraignante et longue à mettre en place en raison du travail intensif en laboratoire requis pour obtenir des résultats comparables entre différents laboratoires. De plus, elle est problématique sur le long terme pour un laboratoire en raison de l'inertie analytique et de la perte d'informations qui en découlent. En effet, selon cette approche, il n'est pas possible d'implémenter une nouvelle méthode analytique tout en approvisionnant la même banque de données en raison de la nature différente des résultats. Il faut alors créer une nouvelle banque de données approvisionnée par la nouvelle méthode analytique et en conséquence mettre à zéro la mémoire de notre connaissance, établie durant plusieurs années.

Dans ce travail de recherche, une méthodologie est ainsi proposée permettant la comparaison de résultats provenant de méthodes analytiques différentes dans l'optique de l'approvisionnement d'une banque de données par ces dernières.

Table des matières

Introduction

Les méthodes analytiques chimiques et physiques occupent une place prépondérante dans les domaines d'analyses alimentaires (Ötles, 2008; McGorrin, 2009), environnementales (Philip, 2009), pharmaceutiques (Martino et al., 2010), d'odeurs et de parfums (Marsili, 2002; Herrmann, 2010) ou de substances stupéfiantes (Comparin, 2007).

L'implémentation de ces dernières dans les laboratoires s'est faite dans un premier temps dans un souci de séparation de tous les composés présents dans un échantillon pouvant s'avérer complexe. L'identification de composés inconnus s'est révélée alors être un objectif pour les laboratoires forensiques, d'où le développement de technologies analytiques de détection répondant à cette exigence. Finalement, grâce à l'évolution des méthodes analytiques séparatives dont la performance n'a cessé de croître, est née l'idée consistant à établir un profil chimique de spécimens dans un but de comparaisons.

Ainsi, actuellement, les analyses systématiques réalisées dans les laboratoires forensiques poursuivent des buts d'identification, de quantification des composés actifs principaux, d'établissement d'un profil chimique et de comparaison des produits étudiés (par exemple, dans le cadre d'analyses environnementales avec l'identification de polluants, la constitution du profil chimique de ces derniers et l'établissement de leur provenance (Philip, 2009)). Ces objectifs nécessitent que le laboratoire d'analyse mette en place des banques de données pour compiler les résultats des analyses effectuées au cours du temps.

Les banques de données constituent des outils essentiels dans le domaine des sciences forensiques, qu'elles concernent l'ADN, les empreintes digitales, les traces d'outils et de chaussures ou encore les produits stupéfiants. Toutefois, comme nous le verrons dans cette étude, les banques de données implémentées dans le cadre du profilage chimique de produits stupéfiants, domaine d'examen de cette recherche, possèdent des caractéristiques ainsi qu'un mode de fonctionnement propres qui les font se démarquer de celles construites dans les autres domaines. Elles comprennent les résultats – qualitatifs, quantitatifs, ou encore définissant le profilage chimique des spécimens – issus d'une méthode analytique développée, optimisée, validée puis généralement référencée dans la littérature.

Tel que l'illustre la Figure 1, le profilage chimique englobe les étapes principales suivantes : la séparation, la détection voire l'identification des composés présents dans l'échantillon d'intérêt à l'aide d'une méthode analytique donnée, l'établissement du profil chimique par la sélection de composés cibles puis la création d'une mémoire de profils compilés dans une banque de données. Finalement, suite à une analyse comparative des profils chimiques enregistrés dans cette dernière, des liens chimiques peuvent être déterminés entre les différents spécimens analysés et ces renseignements sont alors transmis aux forces de police.

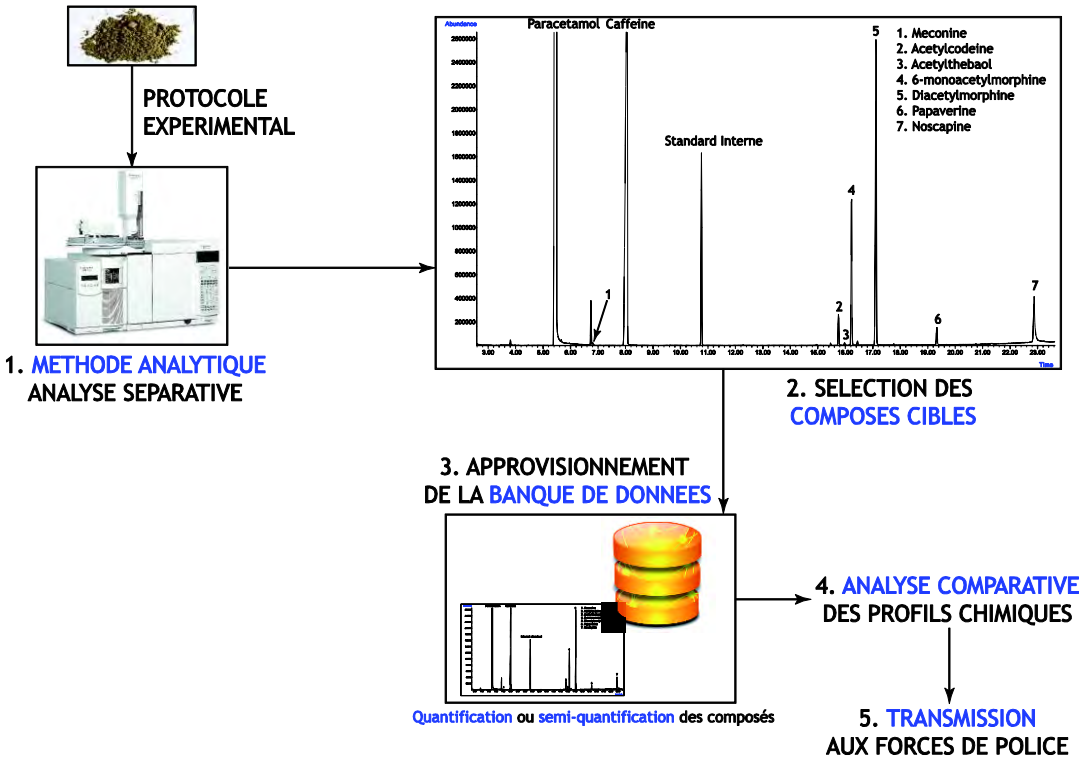


Figure 1. Etapes principales du processus de profilage chimique
(exemple d'un échantillon d'héroïne analysé avec une méthode analytique GC-MS)

Avant de discuter des caractéristiques des banques de données de profils chimiques, il convient de définir leur pourvoyeur, à savoir la « méthode analytique ». Il est effectivement fait mention dans le titre de ce travail de recherche de la notion de différence entre méthodes analytiques, d'où la nécessité d'une telle définition pour déterminer de quelle manière ou à quel point deux méthodes analytiques se distinguent.

Une méthode analytique séparative peut être définie par trois niveaux de paramètres analytiques (définis dans cette étude comme étant les niveaux A, B et C). En effet, une telle méthode se compose de deux technologies d'analyse (niveau de paramètre analytique A), l'une dite de séparation[1] (notée A_{SEP}) et l'autre dite de détection[2] (notée A_{DET}). La première représente la sélectivité de la méthode analytique (temps de rétention et ordre d'élution des composés de l'échantillon analysé) tandis que la seconde représente sa fonction réponse (la réponse analytique obtenue pour chacun des composés, qui est fonction de la concentration de ces derniers dans l'échantillon). Mais une méthode analytique se définit encore par l'équipement utilisé pour son développement et son utilisation en systématique (niveau de paramètre analytique B) et par ce que l'on nomme les paramètres d'analyse ou d'optimisation de la méthode analytique (niveau de paramètre analytique C). Ces derniers sont ajustables d'une analyse à l'autre et décrivent les technologies d'analyse de séparation (C_{SEP}) et de détection (C_{DET}).

Les multiples caractéristiques associées aux méthodes analytiques génèrent une combinaison infinie de ces dernières et des banques de données résultantes qui peuvent être déployées, respectivement, pour l'analyse d'une même substance d'intérêt puis le stockage des résultats analytiques qui en découlent. Pour exemple, le cas de l'analyse de produits stupéfiants en Suisse permet d'observer que plusieurs méthodes de chromatographie en phase gazeuse sont implémentées pour qualifier, quantifier et établir le profil chimique d'un produit stupéfiant dans les laboratoires de sciences forensiques (SGRM, 2009).

Les banques de données se construisent à partir des résultats issus des saisies effectuées par les forces de police et sont ainsi dépendantes et représentatives du travail de ces dernières. Par conséquent, les banques de données de profils chimiques des produits stupéfiants font office de véritable « mémoire » du trafic et de la distribution des produits stupéfiants analysés par le laboratoire en question.

[1] chromatographie en phase gazeuse (GC), chromatographie en phase liquide (LC) ou électrophorèse capillaire (CE) par exemple.

[2] spectrométrie de masse (MS), spectrométrie de masse en tandem (MS/MS), détection à ionisation de flamme (FID), détection par ultraviolets (UV) ou par fluorescence (FL) par exemple.

En effet, l'approvisionnement des banques de données est continu et il est possible de procéder à des analyses rétroactives dans cette mémoire organisée afin de déterminer des liens chimiques éventuels entre différentes saisies policières[3], non soupçonnés a priori lors de l'investigation policière. Ces liens sont dédiés à soutenir l'enquête des autorités de police une fois combinés aux informations d'enquête traditionnelle grâce à des renseignements aussi bien tactiques que stratégiques. Il s'agit ainsi de mettre en lumière les réseaux d'organisation et de distribution du trafic de produits stupéfiants. L'apport primordial de ces renseignements pour la lutte contre le trafic de produits stupéfiants justifie la création et l'utilité de telles banques de données (Esseiva et al., 2003; Esseiva et al., 2007).

Travailler avec une banque de données de profils implique pour tout laboratoire forensique d'être confronté à deux problématiques majeures (cf. Figure 2). La première concerne un seul laboratoire et traite de l'utilisation d'une banque de données sur le long terme (nommée **problématique intra laboratoire** dans cette étude). La mise à jour ou les modifications analytiques de la méthode utilisée initialement pour approvisionner la banque de données illustre cette problématique. La seconde problématique s'avère plus complexe et concerne aussi bien le maintien d'une banque de données par différentes méthodes analytiques que le partage par plusieurs laboratoires d'une banque de données commune (nommée **problématique inter laboratoires** dans cette étude). Pour chacune des problématiques, le laboratoire doit évaluer si les résultats analytiques (c'est-à-dire, les profils chimiques) issus de la méthode modifiée ou provenant de laboratoires différents sont similaires à ceux déjà présents dans la banque de données. En conséquence, le triptyque méthode analytique – profil chimique – banque de données (avec en particulier l'influence de la méthode analytique sur le profil chimique obtenu) représente le cœur de ces deux problématiques. La notion de similarité entre les résultats analytiques dans une perspective transversale et longitudinale est ainsi essentielle dans ce contexte.

[3] La saisie représente la quantité totale de marchandise soustraite par la police. Une saisie peut consister en un ou plusieurs spécimens (ayant des profils chimiques similaires ou non). Au sein des spécimens un ou plusieurs prélèvements sont effectués en laboratoire pour analyses : les échantillons. Les pesées répétées d'un échantillon constituent des réplicats de celui-ci.

Concrètement, les questions à se poser sont les suivantes :

- L'approvisionnement de la banque de données est-il simple avec cette méthode analytique différente?
- Est-il nécessaire d'effectuer un ajustement des profils chimiques obtenus avec cette dernière pour les introduire dans la banque de données et y procéder à des comparaisons de profils?
- Au contraire, la similarité des résultats est-elle suffisamment bonne pour éviter un tel ajustement?
- D'ailleurs, comment évaluer une telle similarité ?

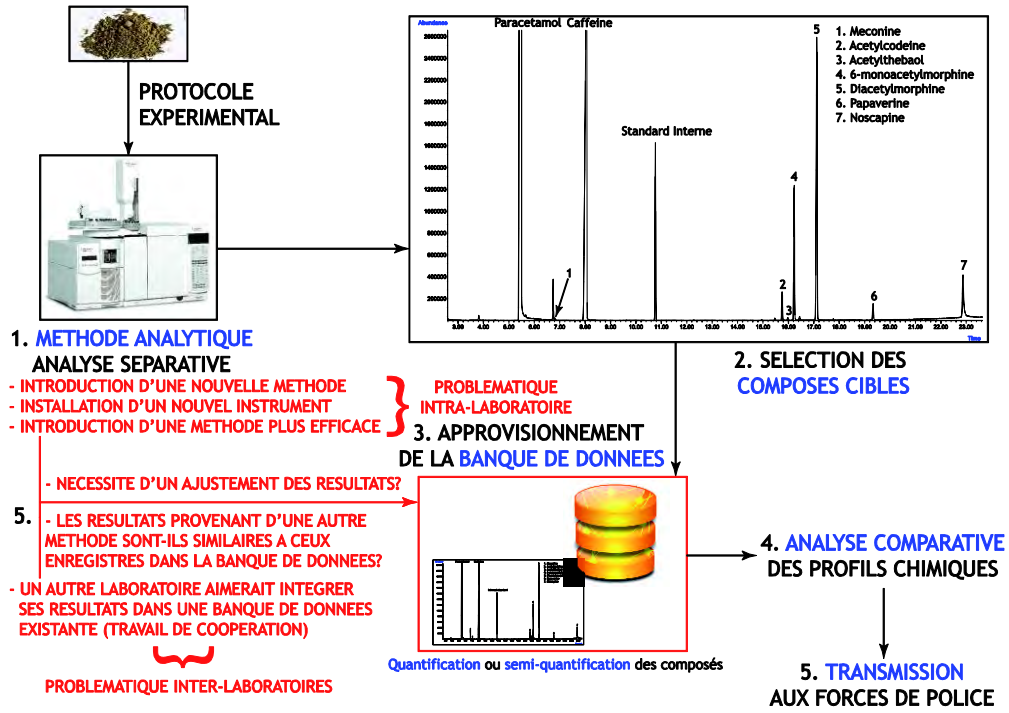


Figure 2. Illustration de la problématique à laquelle tout laboratoire analytique est confronté lors de l'utilisation de banques de données de profils chimiques

La lutte contre le trafic de produits stupéfiants implique des investigations à large échelle qui nécessite l'échange de renseignements. Ce constat est également valable pour les profils chimiques et dans cette optique différentes stratégies ont été identifiées. La première consiste en l'analyse de toutes les saisies de produits stupéfiants dans un seul laboratoire centralisé. Cette stratégie se retrouve dans les pays structurés autour d'un laboratoire forensique central rassemblant toutes les saisies de produits stupéfiants. Cependant, dans le cadre d'une approche internationale systématique, la centralisation des analyses alors que plusieurs pays sont concernés n'est pas vraiment possible pour des raisons politiques évidentes.

En plus de contraintes temporelles et d'une éventuelle surcharge de travail, les inconvénients de cette approche consistent en des problèmes politiques, financiers ou administratifs (tels que l'envoi des saisies) aussi bien que l'intégrité et la stabilité de l'échantillon avant son analyse.

Les deux approches restantes se basent sur une idée commune : l'utilisation d'une banque de données partagée et approvisionnée par différents laboratoires. Il convient de souligner qu'il n'est pas trivial de mettre en commun les résultats analytiques provenant de laboratoires différents pour les comparer entre eux. Les laboratoires doivent s'assurer que les résultats (i.e. les profils chimiques) sont similaires pour les intégrer dans une banque de données commune. Pour atteindre une telle similarité, deux méthodologies différentes se font face : d'une part, **l'harmonisation des méthodes analytiques,** à laquelle le recours est encouragé de nos jours, et, d'autre part, **l'harmonisation des résultats analytiques**, investiguée dans la présente étude. Comme leurs noms l'indiquent, la première se concentre sur les paramètres analytiques définissant la méthode analytique tandis que la seconde se concentre sur les résultats issus de ladite méthode analytique, c'est-à-dire les profils chimiques. Par conséquent, leur niveau d'intégration respectif dans le processus du profil chimique est différent (cf. Figure 3).

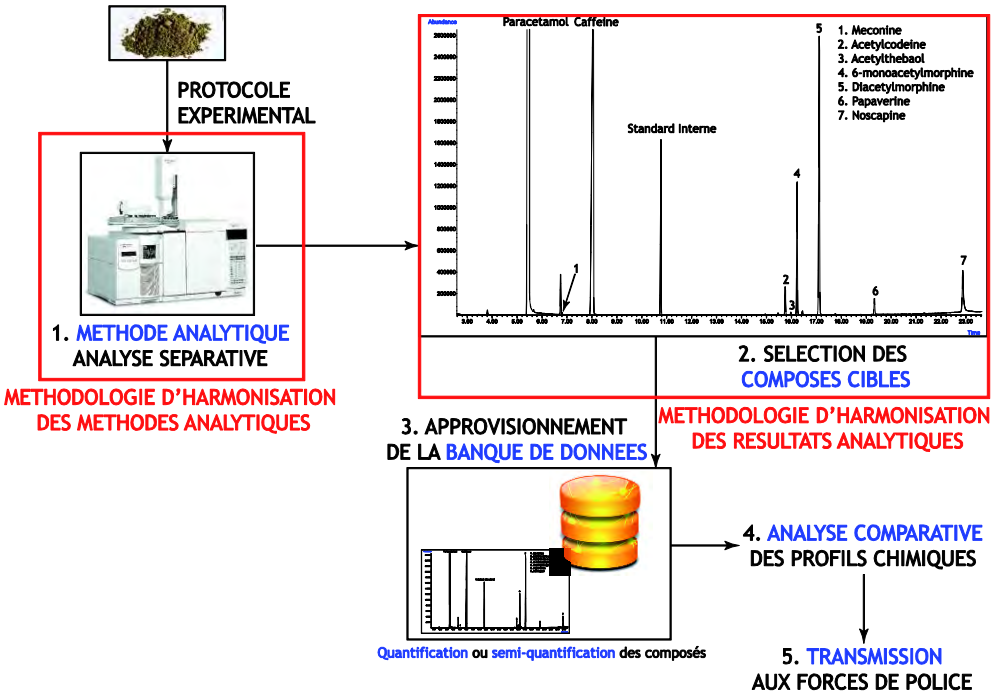


Figure 3. Illustration du niveau d'intégration de chacune des méthodologies d'harmonisation dans les différentes étapes du processus du profilage chimique

L'harmonisation des méthodes analytiques est mise en place de nos jours dans le cadre d'une stratégie prônant la création de banques de données harmonisées, pour comparer les profils chimiques provenant de différents laboratoires analytiques « en temps réel ». En pratique, pour approvisionner une même banque de données, il est recommandé dans le cadre de cette stratégie d'adopter une méthodologie analytique strictement identique de la préparation des échantillons au traitement des données en passant par l'analyse des échantillons (paramètres analytiques A, B et C d'une méthode analytique). Selon cette méthodologie d'harmonisation, les méthodes analytiques en jeu sont considérées comme similaires, bien qu'elles soient implémentées sur des appareillages physiquement différents mais en effet de même marque et modèle.

Bien qu'intéressante en terme de normalisation analytique (p. ex. utilisation en simultanée de paramètres analytiques identiques par plusieurs laboratoires), appliquer la même méthodologie analytique est contraignante et prend un temps considérable en raison du travail en laboratoire intensif requis pour valider les résultats entre les différents laboratoires. De plus, la gamme de méthodes analytiques implémentées dans les laboratoires est vaste : elle est fonction d'exigences propres à chaque laboratoire et de méthodes d'analyses déjà mises en place pour les analyses systématiques ou journalières. Le maintien d'une banque de données par différents laboratoires en suivant les recommandations de l'approche actuelle d'harmonisation des méthodes analytiques n'en est que plus difficile. Cette stratégie s'avère également problématique sur le long terme pour un laboratoire en raison de l'inertie analytique et de la possible perte d'informations qu'elle implique. Selon cette approche, il est en effet impossible d'appliquer une nouvelle technologie d'analyse ou modifier une méthode analytique existante tout en continuant d'approvisionner la même banque de données, en raison de la nature différente des résultats provenant de diverses méthodes analytiques (les réponses analytiques obtenues pour chacun des composés d'un échantillon sont effectivement souvent différentes d'une méthode analytique à une autre). La mise en place d'une nouvelle banque de données approvisionnée par la nouvelle méthode analytique est alors préconisée. Par conséquent, à moins d'un long processus de ré-analyse des spécimens avec la nouvelle méthode analytique pour autant qu'ils soient encore disponibles, la « mémoire » établie durant plusieurs années serait réinitialisée.

Dans la présente étude, une autre approche surmontant ces inconvénients est investiguée. Cette approche, décrite comme étant l'harmonisation des résultats analytiques, va être évaluée dans le cadre du profilage chimique des composés majeurs de produits stupéfiants. Cette stratégie se concentre sur l'ajustement des profils chimiques et postule qu'il n'est pas obligatoire d'appliquer strictement la stratégie d'harmonisation des méthodes analytiques lorsque les méthodes présentent une répétabilité et une reproductibilité acceptables. Il est ainsi considéré que le maintien d'une banque de données à l'aide de résultats issus de méthodes analytiques différentes est, dans certains cas, concevable et applicable.

Une méthodologie consacrée à l'étude des résultats analytiques doit alors être mise en place pour maintenir une banque de données alimentée par diverses méthodes d'analyse et propre à un seul et même laboratoire analytique ou au contraire commune à plusieurs d'entre eux. A la connaissance de l'auteur, aucune étude n'a porté sur le sujet, les recherches se focalisant constamment sur une harmonisation des méthodes d'analyse, avec la mise en place de méthodes analytiques sur de mêmes marques et modèles d'appareillage.

L'étude de cette méthodologie portera principalement sur l'utilisation des méthodes d'analyses dites rapides dans le cadre du profilage chimique de produits stupéfiants. Alors que dans l'industrie pharmaceutique par exemple leurs utilisations sont courantes depuis leurs mises en place, la littérature mentionne très peu d'applications dans les laboratoires de police scientifique dans le cadre du profilage chimique. Pourtant, qualifier un plus grand nombre de spécimens stupéfiants, déterminer leurs profils chimiques et leurs liens éventuels avec d'autres spécimens plus rapidement amélioreraient de manière significative la capacité d'analyse d'un laboratoire, et ce d'autant plus lorsque les laboratoires ont à analyser plusieurs centaines voire plusieurs milliers de spécimens par année. Grâce à de telles méthodes analytiques rapides ces derniers seraient en mesure de fournir les informations requises (type, pureté voire liens chimiques éventuels du spécimen de produit stupéfiant) en quelques minutes, à la différence des plusieurs dizaines de minutes nécessaires actuellement avec les méthodes analytiques conventionnelles[4].

[4] Il est toutefois important de souligner que la méthode analytique n'est pas la seule étape consommatrice de temps dans l'ensemble du processus du profilage chimique (p. ex. l'étape de préparation des échantillons qui précède celle d'analyse des échantillons).

Les informations transmises pourraient alors être exploitées plus rapidement par les services de police et s'avérer plus efficaces dans le cadre de la lutte et la répression du trafic de produits stupéfiants. Un aspect de ce travail de recherche sera donc consacré à la faisabilité du profilage chimique par des méthodes d'analyse rapides et à leur similarité aux méthodes analytiques systématiques.

Ce travail de recherche propose donc l'approche originale d'approvisionner et maintenir une banque de données avec les résultats de différentes méthodes d'extraction d'un profil chimique de produits stupéfiants. Il s'agira ainsi de démontrer que la création d'une banque de données commune à plusieurs méthodes analytiques ou à plusieurs laboratoires ne passe pas nécessairement par une phase contraignante d'harmonisation des méthodes d'analyse. La particularité de ce travail repose sur l'implémentation d'une méthodologie permettant le maintien d'une même banque de données avec des profils chimiques établis par semi-quantification et issus de diverses méthodes analytiques.

Partie A Contexte

La méthodologie développée dans ce travail de recherche se veut applicable à n'importe quel produit d'intérêt et en particulier aux produits stupéfiants dans le cadre du profilage chimique de ces derniers. Ainsi, avant de discuter précisément les problématiques sous-jacentes au maintien d'une banque de données commune à différentes méthodes analytiques, il est dans un premier temps question du profilage de produits stupéfiants (Chapitre 1), de la technologie analytique GC-MS, définie comme la méthode analytique de référence dans ce travail et pourvoyeuse des banques de données de profils (Chapitre 2), puis des évolutions technologiques permettant la mise en place de méthodes dites rapides étudiées dans le cadre de cette recherche (Chapitre 3).

Chapitre 1 Le profilage de produits stupéfiants

NB : La majeure partie des informations fournies dans ce chapitre provient de la combinaison des ouvrages édités par l'ONUDC (UNODC, 2001; UNODC, 2005) et d'articles ou d'ouvrages scientifiques (Esseiva et al., 2007; Esseiva and Margot, 2009).

1.1 Principes et déroulement

Le profilage consiste à extraire de saisies policières de produits stupéfiants des composants nommés « profils ». Lorsque l'on parle de profilage de produits stupéfiants, il est fait référence au processus de détermination des caractéristiques d'un spécimen, qu'elles soient physiques ou chimiques, à l'aide de différentes méthodes analytiques. Tous les moyens possibles pour obtenir des informations et des renseignements visant à établir des liens entre des spécimens doivent être considérés tels que l'utilisation des caractéristiques physiques (couleur, texture, etc.), du packaging (logo par exemple), des composés majeurs, mineurs et en traces (qualification des composés présents, quantification des composés importants) ainsi que des produits de coupage (présence et nature des adultérants et/ou diluants). Les données générées permettent de produire une sélection caractéristique du spécimen de produit stupéfiant : le profil. Selon les données utilisées pour l'établir, celui-ci peut donc être de différentes natures.

Le profilage consiste en une approche scientifique qui vient en appui des renseignements traditionnels obtenus par les services de police lors de leur enquête. Ainsi, pour maximiser la valeur de l'information découlant du profilage, une coopération rapprochée et un mécanisme d'échange d'informations entre tous les acteurs clés du processus d'investigation sont requis, en particulier entre le laboratoire forensique, le magistrat instructeur, les forces de police et les analystes criminels. Le but de l'investigation policière doit être clair pour que le laboratoire détermine l'approche analytique adéquate et interprète les résultats de manière correcte.

Le processus de profilage commence avec l'apport d'une saisie effectuée par les services de police dans le cadre de leur investigation. Généralement, les informations traditionnelles d'enquête accompagnent la saisie. Pour obtenir des informations supplémentaires potentiellement utiles dans le cadre de la lutte contre le trafic de produits stupéfiants[5], une méthodologie analytique consistant en trois phases interconnectées est appliquée sur la saisie.

Dans un premier temps, le produit stupéfiant est caractérisé par un certain nombre de mesures physiques et/ou chimiques (Phase 1). Parmi ces données, celles qui possèdent une valeur comparative intéressante selon les informations que le laboratoire forensique cherche à obtenir sont spécifiquement sélectionnées pour constituer le **profil** du produit stupéfiant (Phase 2). Ensuite, le profil, caractéristique du stupéfiant, peut être comparé aux profils déterminés à l'aide de mesures effectuées dans les mêmes conditions opératoires et regroupés dans une mémoire organisée, la **banque de données** (Phase 3). Dans cette dernière, les spécimens présentant des profils physiques (resp. chimiques) similaires sont regroupés dans des **classes physiques** (resp. **chimiques**)[6]. Dès lors, lorsqu'un nouveau spécimen est analysé, il peut soit former une nouvelle classe physique (resp. chimique) si son profil n'a pas encore été observé, soit être attribué à une classe physique (resp. chimique) préexistante s'il présente un profil physique (resp. chimique) identique. L'étape finale du processus consiste donc en une phase comparative et interprétative visant à déterminer l'existence ou non de **liens physico-chimiques** entre des spécimens saisis dans différentes affaires policières, pour produire du renseignement forensique (c'est-à-dire, l'ensemble des informations extraites des profils des spécimens) par la suite combiné aux renseignements d'enquête (Esseiva and Margot, 2009). Les renseignements qui en découlent sont alors transmis aux forces de police. Les profils physique et chimique peuvent être complémentaires car il pourrait n'y avoir aucune similarité physique entre plusieurs spécimens mais en revanche une similarité chimique, et inversement.

L'enquête policière reste active lors du processus analytique et pourrait ainsi fournir de nouvelles informations. En parallèle, les forces de police pourraient utiliser dans le cadre de leur investigation les renseignements produits par la méthodologie analytique[5].

[5] cf. §1.3.b et 1.3.c pour une présentation des renseignements qu'il est possible d'obtenir.

[6] cf. §1.4.d pour une présentation de la notion de *classe chimique*.

Le profilage étudié dans ce travail de recherche concerne le profilage chimique et la majeure partie de la discussion qui suit le concerne. Le profilage chimique se fait de manière systématique dans les laboratoires spécialisés et les trois phases qui le composent, énoncées dans la partie introductive et au §1.1, sont détaillées à présent. Les deux premières phases concernent respectivement l'étude des spécimens investigués à l'aide d'une méthode analytique et le choix des composés cibles (c'est-à-dire, l'établissement des profils chimiques) tandis que la troisième représente le processus de comparaison de ces derniers dans le but de déterminer des liens chimiques potentiels.

1.2 Phase 1 : Méthode analytique séparative

Le profil chimique se détermine à l'aide d'une méthode d'analyse, quelle qu'elle soit (Nic Daéid and Waddell, 2005).

Toutefois, la sélection des composés cibles influence le choix de l'appareillage analytique ainsi que les informations potentielles qu'il est possible d'obtenir (cf. §1.3). Le choix de la méthodologie analytique s'avère donc crucial et c'est par conséquent au laboratoire forensique de décider de l'approche analytique au regard des informations qu'il souhaite déduire des données collectées.

Les méthodes analytiques conventionnelles (c'est-à-dire, les méthodes analytiques séparatives) mises en place pour établir le profil chimique consistent en une séparation des différents composés, leur identification potentielle, la quantification de la substance illicite (concentration réelle) ainsi qu'en la semi-quantification (réponse analytique) voire la quantification des autres composés présents. Les méthodes utilisées lorsque les composés cibles sont les composés organiques consistent pour la plupart en des techniques de chromatographies en phases gazeuses (GC) sur colonne capillaire – les plus répandues – ou en phases liquides (LC) en phase inverse, couplées à différents modes de détection (Dams et al., 2001).

La chromatographie en phase gazeuse a été largement utilisée depuis la fin des années 1980 et s'avère être la méthode principale et de choix car elle permet aussi bien une grande résolution dans la séparation des composés qu'une bonne sensibilité et reproductibilité (Moore et al., 1984; Neumann, 1984; Neumann, 1994). Alors que la caractérisation des stupéfiants se faisait au départ par GC-FID, la GC-MS lui a été préférée en raison de meilleures sensibilité et spécificité. De plus, la MS possède l'avantage d'assurer l'identification des différents composés ainsi que de pouvoir différencier des composés qui co-éluent. Bien que la GC offre la meilleure capacité de pic (nombre de pics résolus par unité de temps), l'analyse de composés thermiquement dégradables, hautement polaires ou non volatiles peut être problématique et une étape de dérivatisation précédant l'analyse chromatographique devient alors nécessaire. Les techniques de séparation en phase liquide telles que la CE ou la HPLC surmontent de tels problèmes et conviennent plus pour l'analyse de ce type de composés (Dams et al., 2001; Debrus et al., 2010; Tagliaro et al., 2010).

1.3 Phase 2 : Choix des composés cibles

1.3.a Influence du processus de fabrication sur le profil du produit stupéfiant

Qu'ils soient d'origine naturelle (héroïne, cocaïne ou cannabis) ou synthétique (stimulants de type amphétamine), les produits stupéfiants consistent en des mélanges complexes. En revanche, selon la nature du matériel de départ, le processus de fabrication diffère et influence le type d'informations d'ordre forensique qu'il est possible d'obtenir.

Le profil d'un spécimen de produit stupéfiant reflète son histoire et résulte de la contribution de chacune des étapes de fabrication du produit final. Lorsqu'il s'agit d'un produit stupéfiant d'origine naturelle, le profil est premièrement influencé par la graine sélectionnée par le producteur ainsi que les conditions de culture (sol, climat, méthodes de culture). Ces influences peuvent se retrouver dans le cas de stupéfiants synthétiques dans la mesure où les précurseurs seraient d'origine naturelle (Esseiva and Margot, 2009). La qualité et la nature des précurseurs ainsi que la voie de synthèse et ses conditions de réactions contribuent également à la composition du produit final (Stojanovska et al., 2013). C'est pourquoi, déterminer l'origine géographique de ces derniers stupéfiants s'avère peu réaliste.

Dans le processus de production du produit stupéfiant, les méthodes d'extraction et les synthèses qui pourraient suivre représentent des étapes cruciales et influencent grandement la composition chimique du stupéfiant. Les voies de synthèse, la quantité et la nature des produits chimiques, les étapes de purification, etc. ont un impact sur la diversité observée entre lots de production. Chacune de ces variables a une répercussion sur la formation et la proportion des composés incorrectement nommés « impuretés de production » (on parlera plutôt de « produits dérivés ») retrouvés dans le produit final.

Ensuite, le produit stupéfiant est conditionné, emballé, divisé en plusieurs lots par des dealers actifs sur le marché illicite. D'un point de vue forensique, cette phase s'avère intéressante en raison de l'addition de produits de coupage et des caractéristiques de « packaging » qui peuvent fournir des informations supplémentaires sur le réseau de distribution. Finalement, avant son analyse par le laboratoire forensique, les conditions de stockage du spécimen (humidité, température, luminosité) et son vieillissement influencent encore son profil chimique.

Ainsi, bien que les différents spécimens issus du même matériel de départ (par exemple, des feuilles de coca), préparés de la même manière, puissent contenir les mêmes composés (mis à part les produits de coupage), les concentrations relatives dans ces derniers pourraient montrer de grandes variations en raison du nombre important de paramètres affectant le profil chimique ou l'histoire d'un spécimen, en particulier la nature exacte du matériel de départ et les méthodes utilisées pour la transformation, la production, la distribution ou le stockage.

D'après l'historique de chacun des spécimens de produit stupéfiant et l'approche analytique mise en place, un profil chimique peut être établi pour chacun d'eux et ainsi être primordial pour leurs caractérisations et leurs comparaisons aux profils chimiques d'autres spécimens.

1.3.b Influence du processus de fabrication sur les renseignements potentiels

Les composés détectés lors de l'analyse peuvent provenir de plusieurs sources mais appartiennent généralement à l'une des cinq catégories suivantes :

- **les composés naturels** présents dans le matériel brut (feuille de coca, opium) co-extraits lors de la production du produit stupéfiant et qui persistent dans le produit final ;
- **les « impuretés » ou produits dérivés** dus principalement aux procédures appliquées en laboratoire (solvants et/ou réactifs), et lors du transport ou la distribution ;
- **les produits de coupage** ajoutés à n'importe quel point dans la chaîne de distribution ;
- **les artefacts** générés par la procédure analytique ;
- **les contaminations** (ajouts non volontaires).

La Figure 4 ci-dessous résume les différentes étapes de fabrication d'un produit stupéfiant naturel ou synthétique et mentionne les renseignements que peut fournir l'étude des différents composés présents.

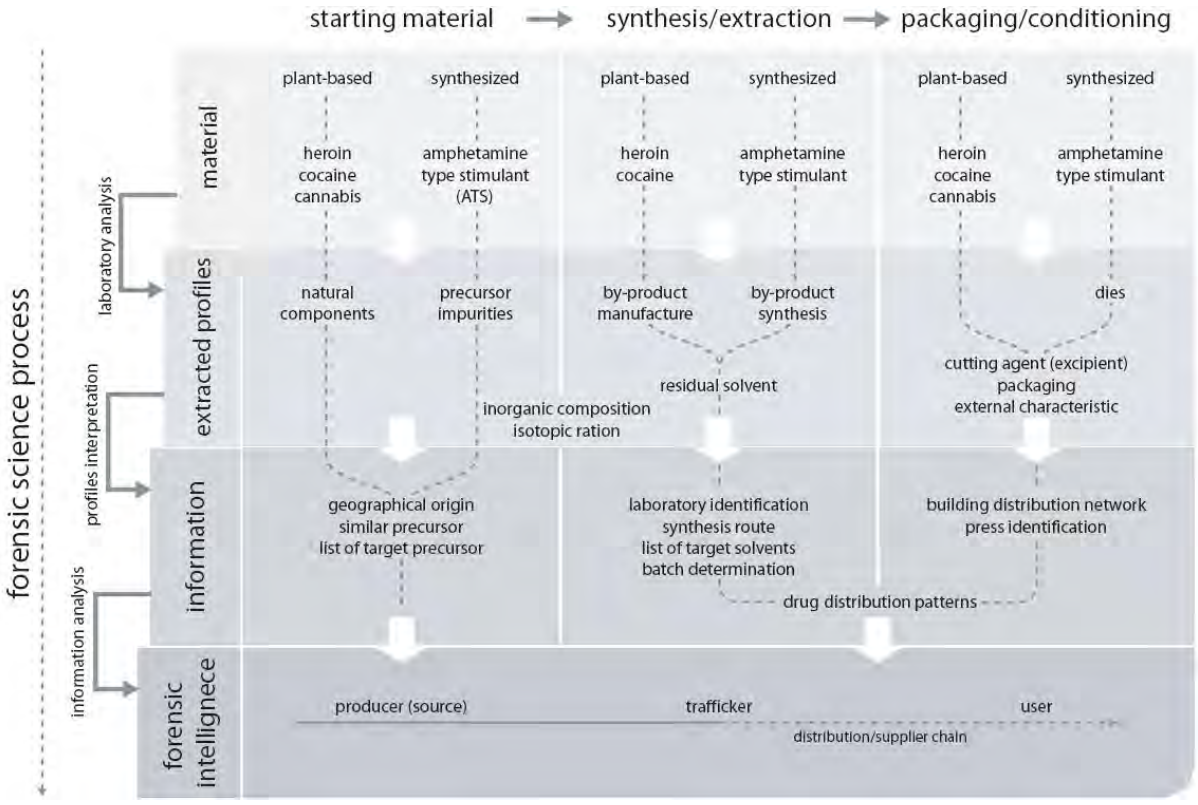


Figure 4. Contribution de chacune des étapes de production au profil du produit stupéfiant et renseignements forensiques potentiels (Esseiva and Margot, 2009)

Parmi l'ensemble des informations utiles pour l'investigation qu'il est possible d'obtenir suite au profilage des produits stupéfiants, on en met en avant généralement quatre. Il s'agit de l'établissement de liens spécifiques entre des échantillons ; de la classification de spécimens de saisies différentes en groupes d'échantillons similaires, permettant ainsi la construction des réseaux de distribution ; de l'identification de la source[7], dont l'origine géographique, d'un spécimen de stupéfiant ; de la surveillance des méthodes de production clandestine et des produits chimiques utilisés.

Pour remplir ces objectifs, le profil chimique peut se baser sur les composés majeurs et mineurs provenant de la plante, les composés dérivés semi-synthétiques spécifiques à la méthode de synthèse et aux conditions expérimentales utilisées, les produits de coupage tels que les adultérants ou les diluants (Collins et al., 2007). La littérature discute également de l'étude des composés en traces (Morello et al., 2010) et des composés inorganiques (Nic Daéid and Waddell, 2005; Waddell-Smith, 2007).

Le profilage peut concerner également les solvants employés durant le processus de fabrication (Dujourdy and Besacier, 2008; Mitrevski et al., 2011; Zacca et al., 2013). A l'aide de ces informations, les nouvelles tendances dans les substances chimiques utilisées pourraient être détectées et les autorités concernées pourraient en être informées.

Concernant la détermination de l'origine géographique, des études récentes ont investigué l'utilisation de l'abondance isotopique de composés particuliers (par exemple, en ^{13}C et ^{15}N) (Carter et al., 2005) pour des spécimens de cannabis (Shibuya et al., 2006; West et al., 2009; Booth et al., 2010; Hurley et al., 2010), d'héroïne et de cocaïne (Ehleringer et al., 1999; Casale et al., 2005; Galimov et al., 2005) ou de type amphétaminique ainsi que récemment pour le GBL, précurseur du GHB (Marclay et al., 2010).

Alors que certaines cibles analytiques peuvent être détectées avec un appareillage conventionnel, d'autres (c'est-à-dire, les éléments en traces ou les rapports isotopiques) nécessitent l'utilisation d'instruments analytiques relativement coûteux que l'on ne retrouve pas dans l'ensemble des laboratoires forensiques.

[7] La notion de *source* est discutée au §1.7.

1.3.c Classification des différentes dimensions de l'information forensique

L'information forensique peut avoir différentes applications selon le destinataire et généralement une classification est effectuée, comme le résume le Tableau 1.

Utilisation	Description	Buts
Soutien politique	Informations centrées sur des préoccupations nationales, internationales et géostratégiques pour le soutien aux initiatives politiques.	Contrôle national et international, identification ou établissement des réseaux de distribution de produits stupéfiants ou les voies de trafic par la classification de spécimens de saisies différentes en groupes d'échantillons similaires, ciblage de la production (contrôle des méthodes de production clandestine, localisation de la production et diminution dans la diversification dans les précurseurs) et identification de la source de précurseurs.
« Drug intelligence »	Informations d'ordre *opérationnel* (*tactique*) et *stratégique* dédiées à la description de la structure et l'organisation du trafic local et régional.	Soutien à l'enquête policière, outil de prévention locale, lutte contre la criminalité en relation au trafic de produits stupéfiants. Déterminer/ comprendre la structure du marché et détecter les nouveaux phénomènes ou les nouvelles tendances dans les réseaux de distribution des stupéfiants.
Elément de preuve	Informations utilisées pour établir un lien spécifique entre deux saisies de produits stupéfiants et ainsi entre les personnes en leur possession dans le cadre d'une affaire particulière.	Démonstration de l'activité criminelle, de son ampleur, de la participation à une organisation criminelle, du lien avec d'autres criminels.

Tableau 1. Classification de l'information forensique obtenue sur la base du profilage (Esseiva and Margot, 2009)

La détermination de l'origine géographique représente une application précise du profilage dans le cadre du **soutien aux initiatives politiques** dont les implications sont discutées au §1.7. La combinaison de plusieurs profils (c'est-à-dire, provenant de l'analyse de différents composés) n'est pas rare lors d'une telle investigation (Esseiva and Margot, 2009). Un autre exemple d'application consiste en le ciblage de la fabrication de produits synthétiques par une étude de leurs compositions chimiques, particulièrement influencées par les précurseurs utilisés et les voies de synthèse implémentées (Gallagher et al., 2012; Stojanovska et al., 2013) (cf. §1.7).

Lorsque le renseignement résultant du profilage vient en appui de l'enquête policière (« **drug intelligence** »), les liens déterminés entre les différentes saisies sont directement inclus puis utilisés dans le cadre de l'investigation, de la même manière que les autres informations collectées par les services de police. Le lien physico-chimique aide alors à formuler des orientations d'enquête utiles et pertinentes. Une fois transmise aux services de police, l'information quant aux liens éventuels peut confirmer/infirmer les hypothèses émises par les enquêteurs voire indiquer d'autres connexions non suspectées lors de l'enquête entre des saisies de produits stupéfiants ou des groupes criminels (Esseiva et al., 2007). Dans un tel cadre, le recours à une banque de données enregistrant les cas précédents devient nécessaire et une méthodologie pour l'identification des échantillons similaires (groupage en classes physico-chimiques) et l'extraction des liens potentiels doit alors être mise en place (cf. §1.4).

Les informations quant aux liens potentiels entre spécimens sont particulièrement recherchées dans le cadre d'affaires spécifiques car elles font office d'**éléments de preuve**. En effet, la question du procureur ou de la police vise la détermination d'un lien spécifique entre deux saisies de produits stupéfiants. Bien qu'un lien chimique déterminé entre des spécimens saisis sur différentes personnes n'implique pas nécessairement qu'une relation existe entre ces dernières, cette information peut être utilisée par les autorités de lutte contre le trafic de produits stupéfiants pour identifier les relations entre les personnes impliquées dans le trafic ou la distribution[8]. Le but de ces comparaisons « case to case » consiste à démontrer que les spécimens en question possèdent une histoire commune (Esseiva and Margot, 2009).

[8] Comme nous le verrons au paragraphe 1.7, des liens peuvent être déterminés aux différentes étapes caractérisant la chaîne d'approvisionnement du stupéfiant.

Pour utiliser les informations du profilage en tant qu'éléments de preuve, les données issues de plusieurs techniques analytiques devraient être utilisées. Ainsi, pour la comparaison de deux échantillons d'héroïne, plusieurs auteurs recommandent la combinaison de l'analyse des composés alcaloïdes majeurs, suivie par celle des composés alcaloïdes mineurs (composés neutres et acides en traces) (Chan et al., 2012), les solvants (Cartier et al., 1997), l'analyse isotopique (Besacier et al., 1997), l'analyse élémentaire (Chan et al., 2013), voire celle des produits de coupage (Terrettaz-Zufferey et al., 2007). Si le profil chimique des spécimens devait correspondre, il devrait être conclu, d'après les méthodologies analytique et statistique mises en place, au partage par les deux spécimens d'une histoire commune de leur origine géographique au réseau de distribution en passant par leur procédé de fabrication. Ce type d'analyse ne nécessite pas le recours à une banque de données recensant l'ensemble des spécimens analysés par le laboratoire. En revanche, une interprétation correcte des résultats de la comparaison requiert une connaissance des influences, sur le profil chimique, de chacune des étapes allant de la fabrication d'un spécimen de produit stupéfiant à son analyse par un laboratoire forensique (cf. §1.7.b). Une telle connaissance s'avère essentielle car elle permet d'interpréter les similarités et les différences observées entre les profils chimiques des spécimens correspondants[9].

Pour rappel, quel que soit son destinataire, le renseignement forensique se combine à celui obtenu lors de l'enquête policière (c'est-à-dire, des informations provenant des données de l'enquête) (Esseiva et al., 2007). L'association de ces deux sources de renseignement permet une meilleure connaissance du trafic de produits stupéfiants (Esseiva et al., 2007).

[9] En particulier, les notions de lots de production doivent être comprises et les variations intra- et inter lots doivent être étudiées.

1.3.d Sélection des composés discriminants pour la phase de comparaison des profils

Pour l'analyse comparative des profils chimiques, une sélection des composés d'intérêt, nommés variables, doit être réalisée pour établir un profil chimique caractéristique et discriminant. Par exemple, si la méthode analytique vise les composés majeurs du produit stupéfiant, la phase de sélection statistique ne va retenir, parmi ces derniers, que ceux qui sont les plus discriminants selon l'objectif de la méthodologie de profilage.

Avant de procéder à la sélection des variables, il convient de procéder au prétraitement statistique des réponses analytiques obtenues pour chacune d'elles. Quel que soit le prétraitement appliqué, l'idée conductrice de cette phase consiste à avoir les variables sur une échelle de valeurs comparable pour pouvoir les sélectionner correctement, les concentrations respectives des composés présents dans un spécimen de produit stupéfiant pouvant être grandement différentes (cf. §6.4.c).

La sélection se fait selon des critères de pouvoir discriminatoire, de répétabilité et de reproductibilité (intra- et inter variabilité au sein des différentes saisies, concentration relative, spécificité). Sur la base de ces variables, la phase d'analyse comparative se déroule par la suite.

1.4 Phase 3 : Analyse comparative des profils chimiques

La troisième et dernière étape consiste à comparer le profil à celui de chacun des spécimens enregistrés dans la mémoire de la banque de données, déterminés avec des mesures effectuées dans les mêmes conditions analytiques comme cela est requis de nos jours (cf. Chapitre 4). Cette étape vise à déterminer l'existence possible ou non de liens chimiques (c'est-à-dire, de profils chimiques similaires) entre des spécimens saisis dans différentes affaires policières, à différentes périodes dans le temps.

1.4.a Approche pour la mise en place d'une banque de données de profils

La finalité d'une méthodologie de profilage consiste en la comparaison des profils respectifs de plusieurs spécimens pour en évaluer leur similarité. Ces comparaisons peuvent déboucher sur plusieurs catégories de renseignements (cf. §1.3.c). Selon la catégorie de renseignements et le type d'informations recherchées (par exemple, lorsque le profilage est utilisé en tant que soutien à l'enquête et vise la détermination de la structure du marché), le recours à une banque de données représentative est requis.

L'approche recommandée pour la collecte de données de profils consiste à :

- générer une banque de données, obtenues suite à la caractérisation du stupéfiant (données physico-chimiques) ; les données collectées doivent être pertinentes et appropriées selon les informations que le laboratoire souhaite fournir (cf. §1.3.a),
- déterminer parmi ces données, lesquelles ont une valeur comparative utile statistiquement (cf. §1.3.d),
- développer un algorithme de comparaison pour identifier les liens potentiels (cf. §1.4.b),
- développer et définir une méthodologie statistique pour évaluer le pouvoir discriminatoire de la méthodologie de profilage (cf. §1.4.b).

La banque de données doit se construire progressivement grâce à l'ajout continu des profils des spécimens analysés par le laboratoire et permet la recherche rétrospective de liens potentiels entre les spécimens de différentes saisies policières.

Les banques de données peuvent être utiles pour des comparaisons aussi bien intra laboratoire qu'inter laboratoires. Toutefois, les recherches rétrospectives de liens potentiels entre profils chimiques ne se font essentiellement qu'au sein d'un laboratoire forensique donné et n'impliquent généralement pas plusieurs laboratoires en raison de difficultés dans l'approvisionnement d'une banque de données par des résultats provenant de plusieurs laboratoires analytiques. Cette problématique du maintien d'une banque de données commune par plusieurs laboratoires est abordée au Chapitre 4.

1.4.b Détermination de la similarité des profils chimiques et évaluation du pouvoir discriminatoire de la méthodologie de profilage

Plusieurs approches peuvent être envisagées pour la phase de comparaison de profils mais l'approche généralement utilisée se fonde sur un ensemble de calculs statistiques de mesure de distance ou de corrélation permettant d'évaluer si des spécimens sont chimiquement liés ou non (Nic Daéid and Waddell, 2005; Hibbert et al., 2010). La littérature fait toutefois mention d'études où la détermination de la similarité se base sur l'utilisation de méthodes statistiques dites supervisées ou non supervisées dans lesquelles des modèles mathématiques plus ou moins complexes sont implémentés.

Approche statistique reposant sur une mesure de l'intra- et l'inter variabilité

Cette approche repose sur la mesure de la variabilité des profils chimiques au sein de populations de spécimens dits liés et non liés ainsi que sur la fixation d'un seuil de décision. Le but de cette méthodologie consiste à comparer objectivement la performance de la méthodologie de profilage dans son ensemble pour la séparation des deux populations précitées (nommées également intra- et inter variabilité, respectivement). Le choix des spécimens s'avère crucial et l'échantillonnage constitué doit ainsi être représentatif.

Une procédure qui peut être adoptée pour estimer le pouvoir discriminatoire de la méthodologie de profilage et en particulier évaluer la valeur comparative du profil chimique correspond à celle proposée par Guéniat et Esseiva qui implique la constitution de deux populations, comme en discute le prochain paragraphe (Gueniat and Esseiva, 2005; Esseiva et al., 2011)

La première population regroupe des spécimens que l'on dit liés, c'est-à-dire, provenant de la même source et partageant des profils chimiques similaires. Par exemple, pour le constituer, une étude et une évaluation de l'homogénéité des spécimens issus d'une même saisie, ou, si disponible, provenant d'un même lot (si un laboratoire clandestin a été démantelé et que les lots prêts pour la distribution peuvent être analysés) pourraient être entreprises (Esseiva and Margot, 2009).

La seconde population consiste elle en des spécimens que l'on dit non liés, c'est-à-dire, provenant de sources supposées différentes (en d'autres termes, provenant de saisies policières différentes et n'ayant aucune relation selon les enquêtes de la police) et pour lesquels on s'attend à ce que leurs profils chimiques soient différents. Concrètement, au sein de chacune des populations, après analyse avec la méthode analytique et établissement du profil chimique, les profils chimiques des spécimens respectifs qui les composent sont comparés 2 à 2 à l'aide d'une mesure de similarité ou de distance. Les distributions des valeurs de similarité calculées pour chaque population sont ainsi obtenues.

Comme cela a été décrit par Lociciro et al. (2008), quand différentes mesures de comparaison sont appliquées sur des profils chimiques, trois situations possibles peuvent être considérées entre les distributions de spécimens liés et non liés (cf. Figure 5). La plus commune, la situation (c) dans la Figure 5, conduit à une superposition entre les populations d'échantillons liés et non liés ce qui implique l'évaluation des taux de vrais positifs (VP), de faux positifs (c'est-à-dire, deux spécimens sont déterminés comme étant chimiquement liés alors qu'ils ne le sont pas ; noté FP) et de faux négatifs (c'est-à-dire, deux spécimens sont déterminés comme n'étant pas chimiquement liés alors qu'ils le sont ; noté FN) obtenus selon une valeur seuil (on parlera également de *seuil de décision*).

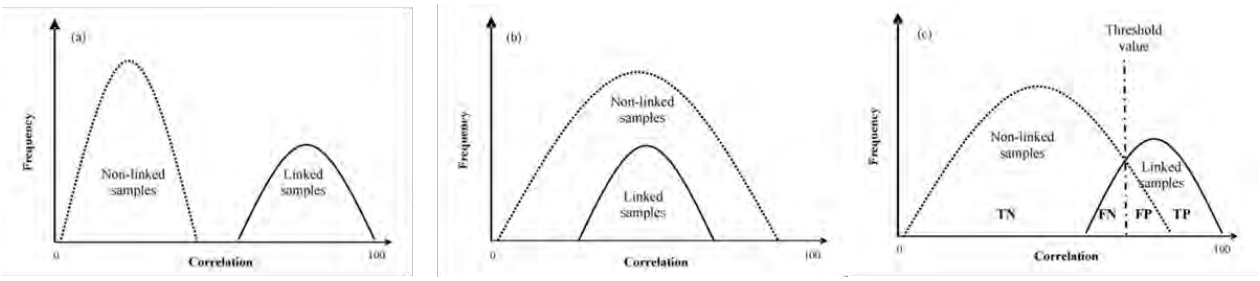


Figure 5. Trois situations possibles lors de l'estimation de la performance de séparation des populations d'échantillons liés et non liés (Lociciro et al., 2008)

Sur la base d'un seuil de décision, la présence ou l'absence d'un lien chimique peut être estimée entre des spécimens de produits stupéfiants. Par exemple, si la mesure de comparaison consiste en une mesure de corrélation, dès que la corrélation de Pearson obtenue lors de la comparaison des profils chimiques des spécimens est supérieure ou égale à la valeur du seuil, alors un lien est confirmé entre ces derniers (c'est-à-dire que les spécimens présentent des composés cibles en proportions similaires). C'est sur cette base que l'on parle alors de spécimens « liés » ou à l'inverse de spécimens « non liés » et par corollaire de lien chimique, ou non.

Le choix d'un seuil de décision particulier représente un compromis. En effet, comme nous l'avons vu au §1.3.c, les résultats du profilage pourraient être utilisés en tant qu'élément de preuve dans une affaire spécifique ou en tant que soutien à l'enquête (utilité opérationnelle). Ainsi, selon la manière dont le laboratoire utilise ces renseignements, le type d'erreurs à minimiser autant que possible n'est pas le même. Si les profils chimiques sont utilisés en tant qu'éléments de preuve, alors le taux de faux positifs devrait être minimisé en choisissant comme seuil une valeur de corrélation plus élevée, conduisant à une diminution dans le taux de vrais positifs. A l'inverse, lorsque les résultats visent à soutenir l'investigation des services de police, alors le taux de faux négatifs devrait être minimisé et le taux de vrais positifs maximisé (en déplaçant le seuil vers les valeurs de corrélation plus faibles). Ainsi, le seuil de décision est choisi selon le recouvrement de l'intra- (qui représente la résultante de l'homogénéité des spécimens et de la variabilité analytique) et l'inter variabilité des spécimens. Finalement, sachant que les taux d'erreur dépendent de la valeur du seuil, ce dernier devrait être choisi pour minimiser l'erreur la moins acceptable (FP vs. FN).

Le motif général de la distribution (par exemple, l'étude du degré de séparation entre les deux populations et la dispersion des valeurs de similarité au sein de chacun d'elles), les taux FP, FN, VP et le seuil de décision établi en fonction de ces derniers sont particulièrement utiles pour juger de la qualité discriminatoire de la méthodologie mise en place. Ainsi, la qualité d'une méthode statistique de comparaison (à l'aide d'une mesure de corrélation ou de distance) se juge par une minimisation des taux de faux positifs et négatifs. De plus, la meilleure méthode statistique de comparaison correspond à celle qui groupe les spécimens liés et qui sépare les spécimens non liés de la meilleure manière.

La Figure 6 ci-dessous illustre le cas où les valeurs de coefficient de corrélation de Pearson qui découlent des comparaisons sont utilisées pour dresser ces deux populations, avec les valeurs de -100 et 100 représentant respectivement une totale différence et une totale similarité des profils chimiques.

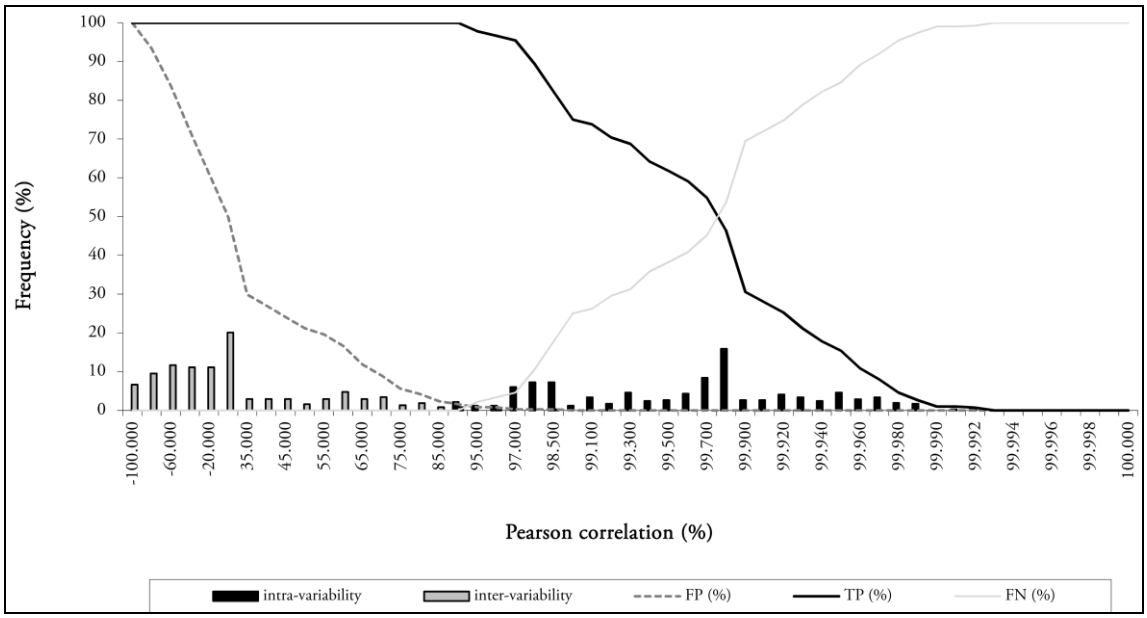


Figure 6. Distribution de l'intra- et l'inter variabilité obtenue pour une méthodologie analytique et statistique

L'intérêt d'une telle démarche, reposant sur l'examen des distributions des populations de spécimens liés et non liés, consiste en la connaissance des taux d'erreur qui accompagnent la méthodologie de profilage dans son ensemble, ces derniers permettant d'évaluer objectivement sa qualité discriminante. Dans ce cadre là, le recours aux courbes ROC (*Receiver Operating Characteristic*) s'avère intéressant car fournissant aussi bien une représentation visuelle qu'une mesure numérique (à l'aide de l'*Area Under the Curve*, AUC) de la qualité discriminante de l'ensemble de la méthodologie implémentée.

Approche reposant sur des méthodes statistiques supervisées et non supervisées

Lorsque les informations du profilage visent à déterminer des groupes d'échantillons similaires, la littérature mentionne l'utilisation de méthodes statistiques que l'on dit *supervisées* (où la connaissance de la distribution dans la banque de données des échantillons au sein des classes physico-chimiques est requise). La littérature présente des applications de modèles mathématiques plus ou moins complexes pour assigner des échantillons inconnus à des classes pré-existantes, selon l'objectif de la méthodologie de profilage et les stupéfiants considérés (Esseiva et al., 2005; Dufey et al., 2007; Broséus et al., 2011). Dans une telle approche, un modèle mathématique doit être créé pour chaque classe physico-chimique. Ce modèle est alors utilisé comme référence pour déterminer si un nouveau candidat appartient à une classe existante. La décision quant à l'appartenance d'un échantillon à telle ou telle classe repose sur des fondements statistiques propres au modèle mathématique implémenté. Comme dans l'approche statistique reposant sur une mesure de l'intra- et de l'inter variabilité, une évaluation précise des taux d'erreurs doit être entreprise à l'aide d'une méthodologie largement discutée dans la littérature spécialisée (Duda et al., 2001; Brereton, 2009; Dixon and Brereton, 2009).

On oppose aux méthodes supervisées les méthodes dites *non supervisées*, c'est-à-dire que la connaissance de la classe des échantillons n'est pas connue a priori. On parle généralement de méthodes de *clustering* ou de regroupement hiérarchique (par exemple, la Classification Ascendante Hiérarchique ou CAH) (Varmuza and Filzmoser, 2009). Bien que ces techniques permettent une visualisation claire de la similarité existant entre les échantillons (pour autant que le nombre d'échantillons ne soit pas trop important), leur utilisation s'avère plus problématique pour déterminer précisément l'appartenance d'échantillons à des classes physico-chimiques similaires en raison de la difficulté de fixer un seuil de décision sur la base des dendrogrammes produits. Ce dernier aspect sera introduit dans le cadre du Chapitre 6.

1.4.c Méthodologie statistique dite *continue*

L'approche statistique pour la mesure de la similarité des profils doit être robuste, fiable et relativement simple à mettre en place pour la comparaison systématique de saisies de produits stupéfiants. Une approche statistique discrète telle qu'une mesure de la similarité entre les profils, approche la plus généralement mise en place, remplit de tels critères. Toutefois, comme en témoigne le précédent paragraphe, une telle approche implique la définition d'un seuil de décision qui détermine la présence ou l'absence d'un lien chimique, et ce de manière catégorique. Par exemple, quelle que soit la valeur de similarité obtenue suite à l'analyse comparative entre deux spécimens, si celle-ci est inférieure au seuil (dans le cas où la mesure de la similarité consiste en un calcul de corrélation), alors on conclut à l'absence de tout lien chimique entre les spécimens. Une telle méthodologie discrète où la réponse consiste en une *absence* ou une *présence* d'un lien peut s'avérer problématique.

En effet, la décision d'absence ou présence d'un lien à l'aide d'un seuil de décision souffre de l'effet « fall of the cliff » (Robertson and Vignaux, 1995), lorsque deux échantillons sont déterminés comme étant similaires ou brusquement différents, selon la valeur de similarité obtenue en comparaison au seuil de décision défini. Par exemple, imaginons que le seuil de décision, basé sur les valeurs de coefficient de corrélation de Pearson des populations des échantillons liés et non liés tel que décrit ci-dessus, soit fixé à 99. L'interprétation quant à la présence ou l'absence d'un lien peut s'avérer délicate lorsque l'on a affaire à des cas que l'on nomme *limites*, c'est-à-dire, où la valeur de corrélation est juste inférieure ou supérieure au seuil défini. En effet, il pourrait sembler aberrant d'accepter un lien entre deux échantillons lorsque la valeur de corrélation est de 99.1 mais de le rejeter si cette dernière est de 98.9. Une méthodologie statistique discrète pose alors un problème pour la continuité de l'établissement des liens. Ainsi, des modèles statistiques continus, tels que l'approche Bayésienne, qui base sa décision sur un rapport de vraisemblance (Aitken, 1995), pourraient s'avérer plus proches de la réalité.

Une telle approche consiste à dire que des valeurs de corrélation très élevées obtenues lors de la comparaison des profils d'échantillons soutiennent fortement l'hypothèse selon laquelle ces derniers proviennent de la même source tandis que des valeurs de corrélation faibles soutiennent au contraire l'hypothèse alternative selon laquelle les échantillons ne proviennent pas de la même source (la source devant être définie). L'utilisation d'un rapport de vraisemblance (Likelihood Ratio, LR) permet d'évaluer la similarité chimique des spécimens. Des valeurs du LR supérieures à 1 soutiennent l'hypothèse principale et ainsi le lien chimique. Plus les valeurs du LR sont élevées plus les résultats supportent l'hypothèse que les spécimens proviennent de la même source. Plus les valeurs du LR diminuent, plus la force du lien diminue jusqu'à obtenir des valeurs inférieures à 1 qui soutiennent alors l'hypothèse alternative consistant à dire que les deux spécimens ne proviennent pas de la même source. Par conséquent, il semblerait que l'implémentation d'une telle approche permette de traiter correctement les valeurs intermédiaires (Lindley, 1977).

Cependant, selon les valeurs du LR, celles-ci « ne soutiennent pas », « soutiennent peu fortement », « soutiennent », « soutiennent fortement » ou « soutiennent très fortement » l'hypothèse selon laquelle un lien chimique existe. Ainsi, par la classification qu'elle engendre, la mise en place d'une échelle verbale pour traiter les valeurs du LR obtenues implique d'être confronté à une relative incertitude quant à la force du lien et produit également l'effet « fall of the cliff ». Ainsi, il n'est pas possible d'affirmer qu'une approche continue représente l'approche idéale pour la comparaison de profils chimiques de produits stupéfiants. Cette approche a toutefois été étudiée et des détails peuvent ainsi être obtenus quant à la procédure à suivre pour sa mise en place (Dujourdy et al., 2003; Bolck et al., 2009).

1.4.d *Classe chimique*

L'issue de la méthodologie statistique, quelle qu'elle soit, consiste en l'identification de profils chimiques similaires entre les spécimens respectifs enregistrés dans la banque de données. Si tel est le cas, alors ces derniers peuvent être regroupés dans une même classe, que l'on nommera chimique en raison des données utilisées dans ce cas pour établir le profil[10].

[10] Rappelons que le profil et la classe peuvent être chimiques, physiques ou les deux selon la nature des caractéristiques considérées par le laboratoire en début de processus.

Ainsi, dans la banque de données, tous les spécimens pour lesquels une similarité de profils a été déterminée sont respectivement regroupés dans une même classe chimique. L'ajout des profils dans la banque de données étant continue, dès qu'un nouveau spécimen est analysé, il peut soit former une nouvelle classe chimique si son profil n'a pas encore été observé, soit être attribué à une classe chimique préexistante s'il présente un profil chimique similaire (Esseiva et al., 2003).

Il découle d'une telle définition que la délimitation des classes chimiques va dépendre intimement, dans un premier temps, de la méthodologie de profilage dans son ensemble et en particulier de son pouvoir discriminatoire et du seuil de décision choisi. Comme cela a été dit ci-dessus, le modèle statistique sélectionné doit permettre le regroupement des spécimens de profils chimiques similaires tout en séparant de manière adéquate les spécimens de profils chimiques différents. Cette classification se veut donc essentiellement statistique.

Dans un second temps, cette délimitation va dépendre également du nombre de spécimens constituant chaque classe chimique ainsi que de la similarité chimique entre ces derniers (cette dernière considération étant abordée ci-dessous dans le troisième et dernier point). Plus le nombre de profils chimiques similaires est grand, plus le risque de superposition des classes est important. En effet, l'alimentation de la banque de données étant continu, ce processus de classification des profils chimiques se caractérise par son dynamisme et la délimitation des classes chimiques n'est par conséquent pas figée et au contraire évolue. Ainsi, lorsque le processus statistique mis en place identifie un spécimen comme appartenant à une classe chimique existante, une modification de la structure de cette dernière ainsi que de celle de l'entier de la banque de données se produit[11].

Finalement, le lien chimique peut se caractériser par une valeur de similarité, après analyse comparative entre les spécimens, égale, juste supérieure, voire largement supérieure au seuil préalablement défini, d'où l'obtention de spécimens plus ou moins similaires, comme en témoigne la Figure 6.

[11] A noter que la banque de données est mise à jour non seulement lors de l'analyse de spécimens mais encore lorsque des informations circonstancielles apportées par les moyens d'enquête conventionnels sont disponibles. Ces informations peuvent, effectivement, en appui aux données existantes, aider à affiner les classes physico-chimiques.

En conséquence, en quelques mots, dans un espace à *n* dimensions (*n* étant le nombre de variables sélectionnées pour définir le profil), deux profils chimiques pour lesquels une valeur de similarité élevée a été obtenue suite à l'analyse comparative seront plus proches l'un de l'autre que deux profils chimiques pour lesquels une valeur de similarité plus faible a été calculée. Alors, au sein d'une classe chimique (c'est-à-dire, un lien chimique a été estimé entre les spécimens qui la constituent), les profils peuvent aussi bien se distribuer dans un espace restreint, si une valeur de similarité élevée a été obtenue entre tous les spécimens qui la composent, que dans un espace plus grand, si une valeur de similarité plus faible a été estimée entre ces derniers (cf. Figure 7).

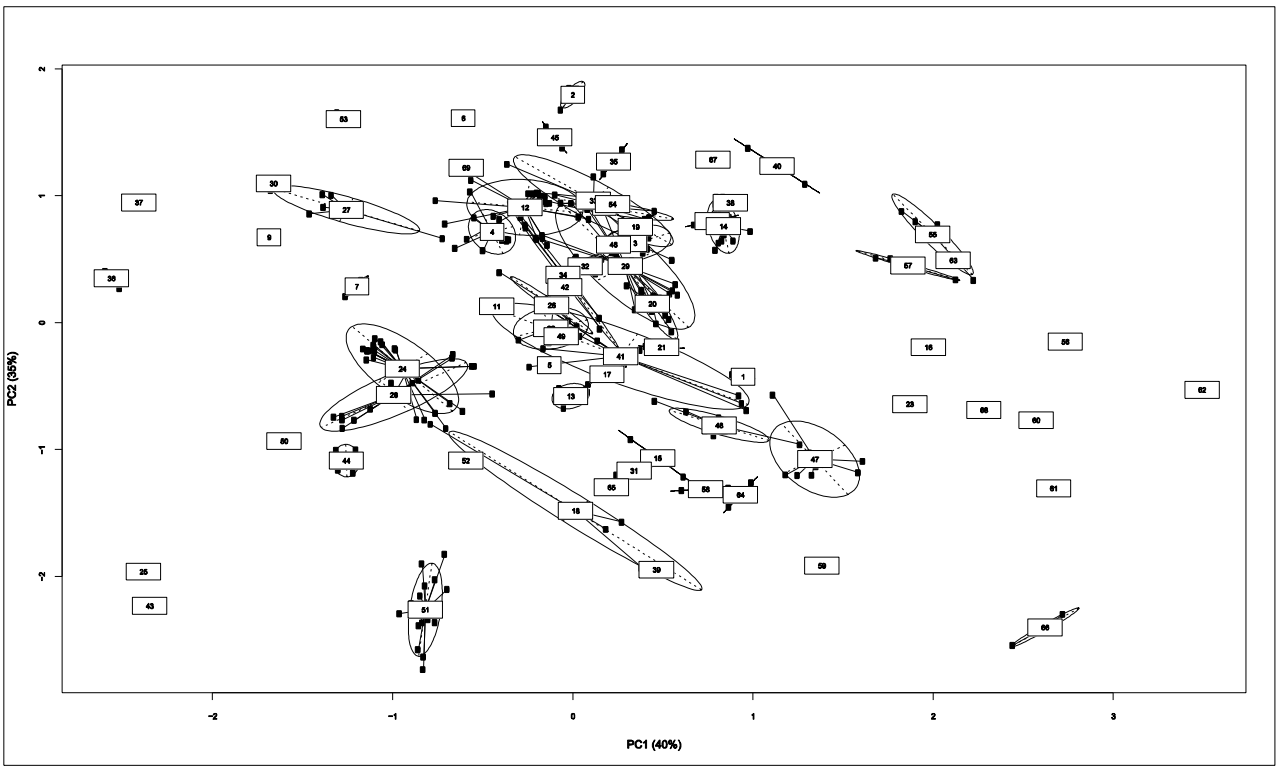


Figure 7. Distribution des scores à l'aide de CP1 et CP2 des 300 spécimens d'héroïne, répartis parmi 69 classes chimiques, saisis durant l'année 2009 en Suisse Romande
(les nombres encadrés et associés à chaque groupe d'échantillons illustrent la classification des profils chimiques déterminés comme étant similaires selon la méthodologie de profilage mise en place à l'IPS)

1.5 Héroïne et profilage chimique

Le profilage chimique abordé dans ce travail de recherche concerne l'héroïne en particulier car il s'agit de l'un des produits stupéfiants les plus rencontrés dans les laboratoires de l'Institut de Police Scientifique de l'Université de Lausanne (IPS).

L'héroïne (3,6-diacetyl ester ou diacétylmorphine) est un composé obtenu après acétylation de la morphine qui est l'alcaloïde le plus actif et le plus abondant de l'opium, suc récolté des capsules du pavot *Papaver somniferum*. Lors du processus de production illégale, plusieurs étapes de précipitations et d'extractions se succèdent. Après ces étapes, un spécimen d'héroïne contient des composés qui proviennent à la fois de l'opium et de l'acétylation des alcaloïdes de l'opium (alcaloïdes opiacés acétylés et non acétylés).

Les constituants majeurs dans un spécimen d'héroïne incluent la diacétylmorphine, la morphine, la codéine, l'acétylcodeine, la 3- et 6- monoacétylmorphine, la thébaïne, la papavérine, la noscapine ou la méconine (Collins et al., 2007). Des détails concernant l'origine et le mode de production de l'héroïne notamment sont présents dans la littérature (Gueniat and Esseiva, 2005). Avant sa mise illégale sur le marché, l'héroïne, essentiellement sous sa forme base ou moins fréquemment sous sa forme acide, est habituellement coupée à l'aide d'adultérants (substances ayant des effets pharmacologiques telles que le paracétamol et la caféine) et/ou de diluants (substances sans effets pharmacologiques, en général des sucres, ajoutés dans une optique de profits). Cette préparation semi-synthétique qu'est l'héroïne constitue alors un mélange complexe de composés organiques et inorganiques.

Ce travail de recherche s'intéresse tout particulièrement au profilage chimique basé sur les composés provenant du pavot et les composés semi-synthétiques, co-extraits avec la morphine et retrouvés en diverses quantités dans l'héroïne, c'est-à-dire, les alcaloïdes opiacés acétylés et non acétylés. Les proportions relatives de ces composés majeurs peuvent varier selon les méthodes de culture et d'extraction – elles-mêmes caractérisées par plusieurs paramètres – mais ne sont pas affectées par l'ajout de produits de coupage : le profil chimique d'un spécimen est alors caractéristique et la comparaison de saisies policières différentes devient possible pour déterminer si un lien chimique entre les spécimens qui les composent existe.

Le profilage chimique de l'héroïne commence par une analyse quantitative de la diacétylmorphine et semi-quantitative voire quantitative des alcaloïdes opiacés principaux précités. Auparavant par GC-FID l'analyse se fait généralement de nos jours par GC-MS (Dams et al., 2001; Dujourdy et al., 2003; Esseiva et al., 2005; Morello et al., 2010). Les techniques de séparation en phase liquide sont également utilisées pour l'établissement du profil chimique de spécimens d'héroïne (Lurie et al., 2004; Lurie and Toske, 2008; Debrus et al., 2010).

1.6 Méthodologie mise en place à l'IPS dans le cadre du profilage chimique de l'héroïne

1.6.a Méthode analytique développée

La méthode analytique qui suit constitue la méthode analytique de référence dans le cadre de ce travail de recherche. L'analyse par GC-MS appliquée en systématique permet d'obtenir un profil chimique d'un spécimen d'héroïne qui répond aux critères chromatographiques conventionnels. Elle permet en une seule analyse la quantification de la diacétylmorphine, la séparation et la semi-quantification des constituants majeurs ainsi que d'un grand nombre d'adultérants et diluants, le tout avec une bonne sensibilité et une bonne résolution des pics (Gueniat and Esseiva, 2005). L'analyse chromatographique d'un échantillon nécessite une étape de dérivatisation au MSTFA pour améliorer la volatilité et la stabilité à haute température de certains constituants très réactifs. Cette étape, d'une durée d'une heure, permet par la suite de réaliser la séparation des composés par GC-MS en un peu moins de 25 minutes (cf. Annexe 2, §1.1).

1.6.b Approche statistique implémentée

A l'IPS, le choix des variables s'est porté sur 6 composés parmi les composés majeurs de l'héroïne. Il s'agit de la méconine (MEC), de l'acétylcodeine (AC), de l'acétylthébaol (AcTB), de la 6 mono-acétylmorphine (6MAM), de la papavérine (PAP) et de la noscapine (NOS)(Esseiva et al., 2011). Une fois les profils chimiques établis pour tous les spécimens composant une saisie, ceux-ci sont comparés à la banque de données, constituée de spécimens provenant de saisies différentes.

Pour l'établissement des liens, la procédure statistique suivante en deux étapes est mise en œuvre. La première étape consiste en une stratégie non-supervisée, réalisée par une Analyse en Composantes Principales (ACP), qui sélectionne les spécimens les plus proches dans la banque de données, c'est-à-dire ceux ayant des profils chimiques similaires. Le processus qui suit correspond à celui détaillé au §1.4. Deux scénarios différents sont possibles. Dans le premier, les nouveaux candidats possèdent des profils chimiques similaires à ceux présents dans une classe chimique préexistante. Alors, une comparaison détaillée à l'aide d'une mesure du coefficient de corrélation de Pearson est effectuée entre les nouveaux candidats et les spécimens présents dans cette classe. Si, selon le seuil de décision défini, le lien chimique est confirmé alors les nouveaux candidats sont insérés dans la classe chimique. La structure de cette classe chimique et de la banque de données de manière générale est alors mise à jour. Dans le second scénario, les nouveaux candidats pourraient présenter un profil similaire à celui d'un spécimen non associé à une classe chimique en particulier. De nouveau, des mesures de similarité permettent d'affirmer la présence ou l'absence d'un lien chimique, selon les critères prédéfinis. Si un lien chimique est confirmé, alors une nouvelle classe chimique est créée et la mémoire mise à jour (Esseiva et al., 2003).

1.6.c Signification opérationnelle et stratégique d'un lien chimique fourni par l'IPS

Si un lien chimique entre des spécimens de saisies différentes est avéré, alors l'information fournie aux autorités policières signifie que ces spécimens appartenaient à une même unité physique avant leur séparation et leur écoulement sur le marché.

La signification des liens chimiques représente un élément essentiel à discuter suite à l'analyse comparative et fait partie intégrante du processus de profilage. Ce processus s'accompagne de limites qu'il est important de connaître et comprendre pour interpréter correctement les résultats obtenus. Comme en témoigne le prochain paragraphe, sans une étape d'interprétation, les informations découlant du profilage ne sauraient être efficaces et utiles pour les autorités de lutte contre le trafic de produits stupéfiants.

1.7 Interprétation des résultats

Pour apprécier le potentiel et les limites du profilage de produits stupéfiants ainsi que les difficultés pour dresser des conclusions de l'analyse comparative, les implications sur les profils chimiques des différentes étapes de production et d'approvisionnement du produit stupéfiant doivent être étudiées.

En effet, les informations découlant du profilage chimique de produit stupéfiant ne peuvent pas être utilisées de la même manière que celles obtenues lors de comparaisons de traces ADN ou digitales. Dans ces derniers domaines, il est possible de déterminer efficacement les sources potentielles à l'origine de ces traces. A l'inverse, en raison de leur nature complexe, le lien entre deux spécimens de produits stupéfiants n'a pas la même signification. Comme en discute ce paragraphe, dans le domaine du profilage chimique de produits stupéfiants, les liens peuvent exister à différents niveaux selon les chaînes de production, de distribution et d'approvisionnement respectives de ces derniers.

1.7.a Lot de production

Lorsqu'un produit stupéfiant est fabriqué, différents lots sont préparés au fur et à mesure. Les conditions de production n'étant jamais exactement les mêmes à chaque fois, des variations se produisent dans les composés constituant les produits finis, issus de la même source. Ainsi, différents lots de production du même laboratoire clandestin présenteront des caractéristiques chimiques différentes (on parle de variabilité inter lots). Les produits stupéfiants n'étant pas nécessairement homogènes, des différences peuvent apparaître au sein d'un seul et même lot de production de produit stupéfiant (on parle de variabilité intra lot). On estime généralement que la variation intra lot est plus faible que la variation inter lots (UNODC, 2001; UNODC, 2005). Déterminer l'appartenance de spécimens au même lot implique que la variabilité dans la composition chimique inter lots soit plus importante que celle intra lot et que la méthodologie analytique fournisse suffisamment d'informations pour saisir l'ampleur de ces deux variabilités (c'est-à-dire, une connaissance précise de ces deux variabilités).

Se prononcer à un tel niveau de source implique une connaissance précise de la composition des lots de production, ce qui est rarement le cas (difficultés d'accès aux laboratoires clandestins).

Dans le cas des produits stupéfiants naturels et semi-synthétiques, la possibilité de variabilité entre les lots est en théorie particulièrement élevée en raison des conditions de laboratoire rudimentaires. Cependant, peu d'informations sont disponibles sur l'importance de la variabilité dans le contenu des spécimens provenant de différents lieux d'un même pays ou de différentes régions. De plus, la taille d'un seul lot de production et la variabilité intra- ou inter lots ne sont pas connus.

La notion de variation entre lots semble être plus définie pour les stupéfiants synthétiques. En effet, plusieurs laboratoires clandestins ont été démantelés en Europe et d'importants lots de production ont été saisis. Il a été alors déterminé que la variation intra lot était plutôt faible, en raison des conditions de fabrication qui sont plus contrôlées qu'avec les produits stupéfiants naturels (UNODC, 2001). Toutefois, il a été démontré qu'une variation dans les profils de stupéfiants synthétiques survenait alors que la même voie de synthèse était strictement suivie par différents chimistes. De plus, des variations dans les profils ont été obtenues lorsque le même chimiste appliquait la même procédure pour produire plusieurs lots. Pour chaque étape de la réaction, le changement dans les conditions de la réaction a conduit à un changement dans le profil (Stojanovska et al., 2013). En conséquence, la variation intra lot peut s'avérer élevée.

1.7.b Interprétation des résultats

Un grand nombre de paramètres doit être pris en compte lors de l'interprétation des résultats que ce soit les implications analytiques de la production et de la distribution illicites ou celles liées aux différences dans les conditions de stockage du produit stupéfiant. Comme cela a été précisé auparavant, une coopération rapprochée entre les autorités de maintien de l'ordre et le laboratoire analytique est nécessaire pour que les résultats soient intégrés au cycle des renseignements du profilage (soutien politique, « drug intelligence » - à but opérationnel ou stratégique - et élément de preuve). Chacune des informations que le processus de profilage permet d'obtenir doit être discutée avec attention.

a) l'identification de la source des spécimens de produit stupéfiant (stupéfiants naturels et semi-synthétiques ou synthétiques)

Le terme « source » peut avoir plusieurs significations et plusieurs niveaux de source peuvent être déterminés. Quand on parle de la détermination de la source d'un spécimen, il peut s'agir de son origine géographique, du laboratoire de production du stupéfiant, de la source d'approvisionnement ou de distribution du spécimen ou encore de l'entité ou unité physique considérée avant la distribution d'un stupéfiant (c'est ce dernier niveau de source qui est considéré à l'IPS dans le cadre des analyses systématiques). Se prononcer à un certain niveau de source implique la mise en place d'une méthodologie de profilage physico-chimique spécifique (spécificité dans le choix des composés cibles ainsi que dans la méthodologie analytique et statistique) ainsi qu'une combinaison des renseignements forensiques aux renseignements d'enquête.

La détermination des régions ou pays d'origine n'est pertinente que dans le cadre d'initiatives politiques internationales (ou bien sûr si cette information est expressément demandée au laboratoire). En effet, les régions d'origine sont généralement bien connues et connaître l'origine précise d'un stupéfiant ne va pas nécessairement être utile pour contrer le trafic de ce stupéfiant dans les pays où il est distribué.

Comme cela a déjà été précisé au §1.3, déterminer l'origine géographique de produits stupéfiants synthétiques n'est pas possible car les précurseurs et méthodes synthétiques utilisés ne sont pas spécifiques à une région (Esseiva and Margot, 2009). Les échantillons ayant un profil chimique similaire pourraient alors plutôt être reliés à un même laboratoire. Mais un tel laboratoire pourrait être localisé n'importe où. En revanche, il est possible de déterminer si deux saisies proviennent du même précurseur et/ou si elles ont été synthétisées selon la même voie de synthèse. En effet, la composition chimique d'un produit stupéfiant synthétique est en partie influencée par les précurseurs (qui pourraient également avoir des impuretés si leur fabrication est clandestine) et par les voies de synthèse (Swist et al., 2005; Stojanovska et al., 2013). Une telle analyse implique de disposer d'une banque de données à jour et représentative de spécimens dits *authentiques* ou *de référence*. Dans le cadre de la lutte contre le trafic de ces substances, de tels renseignements s'avèreraient utiles à un niveau international pour détecter les modifications dans les précurseurs ou les substances chimiques utilisés (selon la voie de synthèse) permettant une meilleure connaissance de la méthodologie synthétique et de l'organisation requises pour la production illicite.

Concernant les produits stupéfiants naturels et semi-synthétiques, déterminer l'origine géographique est, en théorie, possible. Dans ce cadre là, les composés naturels et/ou dérivés présents dans les différents spécimens sont utilisés. En effet, les caractéristiques chimiques des produits stupéfiants sont relativement spécifiques à des régions géographiques en raison, au sein de chacune des régions, de facteurs qui peuvent influencer significativement la proportion dans ces composés tels que le sol, les conditions climatiques, etc. De plus, les méthodes de production clandestines dans n'importe quelle région sont considérées comme étant très proches. En conséquence, les spécimens du même stupéfiant fabriqués par différents producteurs dans le même pays ou la même région contiennent des proportions similaires des mêmes composés dérivés et pourraient ainsi, en théorie, être classés dans un même groupe. A noter qu'actuellement, les études portant sur la détermination d'une source géographique de produits stupéfiants reposent également sur des mesures de rapports isotopiques. Une stratégie consistant à combiner différents types de profils (composés naturels et dérivés) pour déterminer une origine géographique est souvent recommandée (Esseiva and Margot, 2009). Un spécimen pourrait être assigné à une région d'origine spécifique si son profil chimique présente les caractéristiques d'un spécimen de référence, dont on sait qu'il provient de cette région. Cela implique donc la mise en place d'une banque de données de spécimens d'origine connue maintenue régulièrement et représentative (or obtenir de tels spécimens authentiques s'avère difficile). En revanche, si le matériel de départ (par exemple, les graines) ou les conditions de production étaient significativement modifiés au sein d'une région donnée, alors les profils chimiques de spécimens issus de la même région que les spécimens de référence pourraient finalement ne montrer aucune similarité. L'interprétation des profils chimiques de spécimens dans l'optique d'assigner à ces derniers une région géographique nécessite d'inclure ces modifications dans la banque de données et ainsi éviter l'obtention de résultats inutilisables. L'importance de combiner les résultats du profilage avec les informations d'enquête prend tout son sens lorsqu'une telle interprétation est entreprise.

Pour les raisons évoquées auparavant, il n'est en revanche pas possible de relier une série de lots de production à une seule source de production (c'est-à-dire, le laboratoire clandestin ou le lot spécifique) (cf. 1.7.a).

b) la signification des similarités et différences dans les profils chimiques des spécimens

Lors d'une telle analyse, il faut avoir conscience que la chaîne d'approvisionnement d'un produit stupéfiant peut être longue et complexe, du producteur au consommateur en passant par le trafiquant, le distributeur et le fournisseur (dealer) (cf. Figure 8).

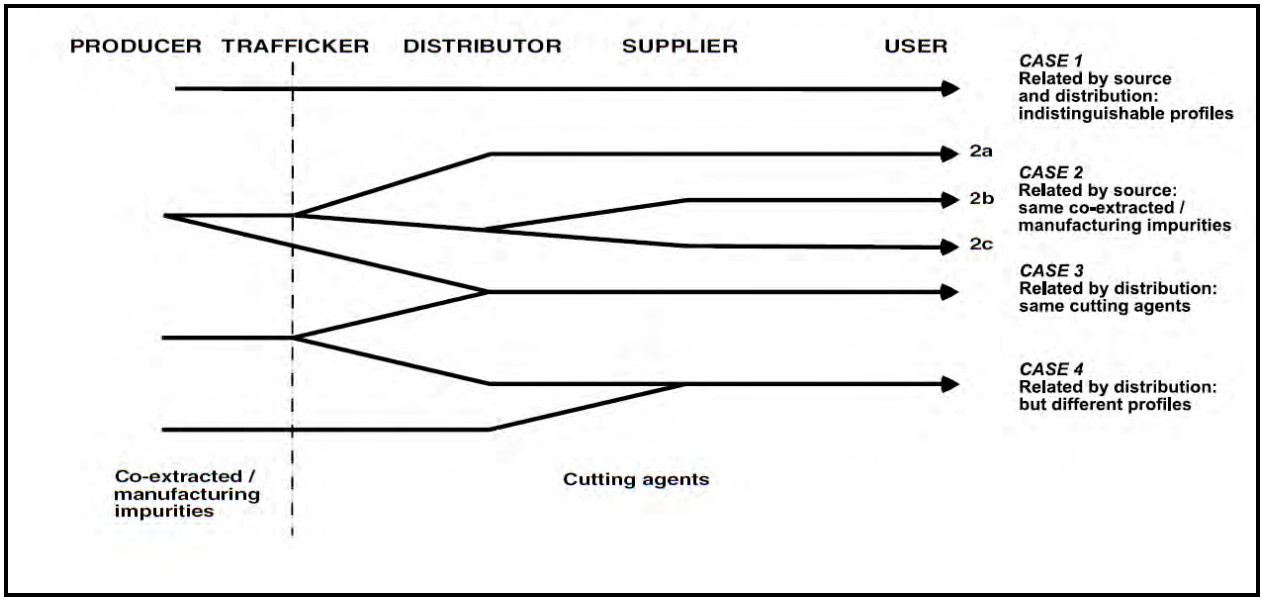


Figure 8. La chaîne de distribution des produits stupéfiants et son impact sur les profils chimiques (UNODC, 2005)

Dès que le produit stupéfiant est préparé ou produit, le producteur pourrait le fournir à un ou plusieurs trafiquants, ce dernier pouvant le fournir à un ou plusieurs distributeurs et ainsi de suite jusqu'au dealer qui peut approvisionner plusieurs consommateurs. A l'inverse, un consommateur pourrait obtenir le produit stupéfiant d'un ou plusieurs dealers, qui lui-même pourrait s'approvisionner chez un ou plusieurs distributeurs. Bien que cela soit moins probable, le distributeur pourrait se fournir chez un ou plusieurs trafiquants et le trafiquant chez un ou plusieurs producteurs. Les spécimens provenant de la même source (c'est-à-dire, provenant du même producteur) peuvent être distribués selon des chaînes d'approvisionnement différentes. Il est important de souligner qu'à chaque étape de l'approvisionnement et de la distribution, des produits de coupage peuvent être ajoutés résultant ainsi en une composition du produit stupéfiant de plus en plus complexe (d'où un profil chimique plus complexe si une méthodologie analytique adéquate est implémentée pour le déterminer).

Dans ce paragraphe, estimons que le profil chimique d'un spécimen consiste dans les composés naturels et issus du processus de fabrication ainsi que dans les produits de coupage. Tel que l'illustre la Figure 8, des liens chimiques entre spécimens pourraient se faire à différents niveaux de la chaîne de distribution :

- ces derniers pourraient avoir une histoire commune, c'est-à-dire qu'ils auraient été produits par le même producteur et distribués dans la même chaîne d'approvisionnement (*CASE 1*).
- ils pourraient être liés par leur source[12] mais non par leur distribution (*CASE 2*).
- finalement, ils pourraient provenir de sources différentes mais être reliés par leur méthode de distribution (*CASE 3* et *CASE 4*).

Dans le *CASE 1* les profils chimiques sont similaires, c'est-à-dire qu'ils contiennent les mêmes proportions relatives des composés co-extraits et/ou de fabrication ainsi que les mêmes produits de coupage, dans les mêmes quantités. Ainsi les spécimens viennent probablement de la même source et appartiennent potentiellement au même réseau de distribution.

Si les profils chimiques sont significativement différents, alors il peut être conclu à l'absence de liens chimiques entre les échantillons. Toutefois, comme le démontre le *CASE 4* il pourrait y avoir une relation entre les spécimens, au niveau du fournisseur.

Il est clair que les profils chimiques de spécimens provenant d'une source commune peuvent être différents si ces derniers sont saisis à différents niveaux de la chaîne de distribution ou s'ils ont été distribués dans différents réseaux. Les produits de coupage pouvant être ajoutés à chaque étape, et différents produits de coupage pouvant être utilisés par différents dealers, les profils chimiques des spécimens présenteront des proportions relatives similaires dans les composés naturels et/ou de fabrication mais des différences dans les produits de coupage (*CASE 2a* à *2c*). Ces spécimens, bien que probablement reliés par la source, ne le sont pas par la distribution et ne possèdent donc pas la même histoire.

Le *CASE 3* fait lui référence à la possibilité d'obtenir des profils chimiques qui diffèrent dans les proportions des composés naturels et/ou de fabrication mais qui présentent des produits de coupage similaires dans leurs natures et leurs proportions.

[12] Ici, le terme « source » fait référence au producteur du produit stupéfiant.

c) l'établissement de liens entre des spécimens

Il s'agit ici de discuter de la validité et de la qualité des résultats et des conclusions qui en découlent. La qualité d'un lien chimique entre deux spécimens dépend de la corrélation entre leurs profils chimiques respectifs et de la fréquence du motif du profil chimique (profil chimique rarement observé auparavant, présence de composés inhabituels, etc.) (Esseiva and Margot, 2009). Lors de l'interprétation des résultats dans un tel cadre, il faut tenir compte du vieillissement des spécimens (cf. §f) ci-dessous).

De plus, il faut avoir conscience que le processus analytique lui-même introduit des différences entre les analyses d'un même spécimen. Il s'agit là des caractéristiques de répétabilité et de reproductibilité de la méthode analytique. Celles-ci doivent être évaluées et l'expert doit les prendre en compte lors de l'estimation de la similarité entre les profils chimiques.

d) l'établissement des réseaux de distribution du produit stupéfiant

L'utilisation seule des profils chimiques rend difficile l'établissement des réseaux de distributions associés à des producteurs ou des organisations criminelles spécifiques. En effet, comme cela a été précisé auparavant, il n'y a pas assez d'informations sur la composition d'un lot de production et sur les variations possibles dans les composés de production entre lots et entre pays ou régions géographiques. Il est donc difficile de lier des spécimens fabriqués par le même producteur mais issus de lots différents tout en discriminant les spécimens fabriqués par un producteur différent dans la même région. L'utilisation de plusieurs méthodes analytiques combinées aux informations d'ordre stratégique, à l'ensemble des renseignements forensiques qu'il est possible de produire ainsi qu'aux données d'enquête est recommandée.

Par exemple, dans le cadre des produits stupéfiants synthétiques, la littérature fait état de travaux concernant la mise en place de systèmes de profilage physique ainsi que l'étude de sa complémentarité au profilage chimique (Milliet et al., 2009; Lopatka and Vallat, 2011; Camargo et al., 2012; Edelman et al., 2013).

La mise en place de systèmes de veille opérationnelle de sites Internet peut également fournir des informations intéressantes quant à la structure des réseaux de distribution de certains précurseurs chimiques en vente libre sur l'Internet (Giannasi et al., 2012). La complémentarité de ce type d'informations à celles résultant des profilages physique et chimique fait d'ailleurs l'objet d'une recherche à l'IPS dont les résultats préliminaires démontrent l'importance de la combinaison de l'ensemble des renseignements forensiques pour décrire l'organisation des réseaux de distribution de certains produits vendus sur Internet (Pazos et al., 2013).

La nature du trafic international de produits stupéfiants complique encore les tentatives de détermination des réseaux de distribution nationale ou internationale. Les spécimens provenant du même lot de production peuvent en effet être distribués dans différentes chaînes d'approvisionnement, comme en témoigneraient les différents produits de coupage, à destination de divers pays. Les produits de coupage identifiés comme étant caractéristiques ou inhabituels pourraient s'avérer utiles pour grouper des spécimens et établir l'ampleur de tels réseaux.

e) l'identification et la caractérisation du matériel de départ spécifique utilisé lors de la production illicite du stupéfiant

Les produits bruts et les précurseurs chimiques (c'est-à-dire, les matières premières) employés dans le processus de production clandestine peuvent également contenir certaines impuretés. La teneur et le type d'impuretés peuvent varier selon la nature des matières premières (opium ou substance chimique, par exemple), si un précurseur chimique provient d'une source légitime ou au contraire s'il a été produit clandestinement. L'identification des composés propres aux précurseurs pourrait ainsi aider à les relier à une source clandestine ou commerciale.

Pour relier des produits stupéfiants aux matières premières correspondantes, une connaissance approfondie des composés présents dans les matières premières est primordiale. De plus, les mécanismes conduisant à la présence de composés dans les produits stupéfiants doivent être compris : ils pourraient être déjà présents dans les matières premières et se retrouver inchangés dans le produit final, provenir de réactions de composés originaux présents dans les matières premières ou être générés en tant que produits dérivés de la production du produit stupéfiant (ceux-ci étant indicatifs des voies de fabrication mais moins intéressants pour identifier les sources des matières premières).

f) stabilité des profils chimiques dans la banque de données

A la connaissance de l'auteur, les informations disponibles concernant la stabilité temporelle des données enregistrées dans la banque de profils sont peu nombreuses. La stabilité temporelle d'un spécimen de produit stupéfiant se définit comme l'évolution naturelle de sa composition, donc de son profil chimique, en fonction du temps et suite à son exposition aux conditions climatiques naturelles (lumière, humidité, température). On parlera plus particulièrement du vieillissement des spécimens. Ce facteur, parmi d'autres, peut péjorer à l'établissement de liens entre spécimens d'un produit stupéfiant et n'est pas spécifique à la source ni à la distribution. La concentration de chacun des composés du profil peut évoluer selon les conditions de stockage auxquelles le spécimen est soumis (lumière, chaleur, humidité, etc.).

Ainsi, des spécimens provenant de la même source, mais exposés à des conditions environnementales différentes dans la chaîne de distribution, pourraient ne pas présenter de lien chimique. De même, certains produits de coupage agressifs tels que l'acide ascorbique pourraient altérer la composition chimique des spécimens et avoir les mêmes conséquences (UNODC, 2001). Alors, le profilage chimique de spécimens dont la composition aurait été altérée pourrait ne pas fournir d'informations fiables et pertinentes quant à leur similarité.

Dans ce cadre là, relevons l'étude de Lociciro et al. (2007) qui aborde l'influence des conditions climatiques sur le profil chimique de spécimens de cocaïne. L'étude, d'une durée de 3 mois, démontre que la température, l'humidité et le temps n'ont pas d'influence significative sur le profil. Par conséquent, les auteurs concluent à la réalisation possible du profilage chimique pendant une période de 3 mois pour des spécimens conservés à température ambiante, sans autre précaution particulière. Au-delà d'une telle période, il n'y a pas d'informations disponibles sur l'évolution du profil chimique.

Or, un tel vieillissement peut avoir en particulier un impact sur la dispersion des valeurs de l'intra variabilité (population des spécimens liés)[13]. En effet, un même spécimen analysé à un temps t_0 puis ré-analysé après plus de 3 mois pourrait, selon le seuil de décision défini, ne plus montrer de similarité chimique. De manière générale, les valeurs définissant l'intra variabilité risquent alors de se déplacer vers de plus faibles valeurs (si la mesure de similarité consiste en une mesure de corrélation) avec en conséquence une diminution du taux de vrais positifs si le même seuil de décision est conservé.

Actuellement, à l'IPS, des recherches traitent de la problématique du vieillissement des produits stupéfiants pour en évaluer son évolution selon les conditions climatiques ainsi que son impact sur la distribution des populations de spécimens liés et non liés (et les taux de faux positifs et négatifs associés) (Nier et al., 2012). D'après les premiers résultats, aucune variation dans le profil chimique n'a été constatée.

Finalement, notons que Strömberg (Strömberg et al., 2000) déconseille de conserver des échantillons plus de 6 mois dans la banque de données pour des questions de gestion de cette dernière (en particulier, pour l'efficacité du processus de recherche de liens) mais également en raison de l'instabilité dans la composition chimique des échantillons. Il s'agit là d'une des rares publications précisant la durée de conservation des profils dans la banque de données. Malheureusement, l'impact de cette conservation relativement courte sur une perte potentielle d'informations n'est pas discuté (par exemple, en regard de la durée de vie d'une classe chimique).

Avant d'aborder la problématique principale de cette étude, touchant à l'approvisionnement et au partage des banques de données, il convient de s'intéresser aux pourvoyeuses de ces dernières, c'est-à-dire, les méthodes analytiques séparatives. En particulier à la GC-MS, car cette méthode représente la méthode analytique de référence dans cette étude. Le chapitre suivant se consacre donc à sa description succincte en centrant le propos sur la partie MS afin d'identifier les paramètres pouvant a priori influencer la similarité de résultats provenant de méthodes analytiques différentes, problématique générale de ce travail. Le Chapitre 3, lui, traite des méthodes chromatographiques rapides qui seront préférentiellement implémentées dans cette recherche.

[13] La dispersion des valeurs qui définissent la population des spécimens non liés devant être a priori moins affectée.

Chapitre 2 Chromatographie Gazeuse – Spectrométrie de Masse[14]

2.1 Généralités

La GC-MS combine un haut pouvoir de séparation des composés avec une détection massique très sélective et sensible. En incorporant une automatisation, une miniaturisation et une simplification dans sa conception, la GC-MS a évolué depuis les années 1970 pour offrir de nos jours un large champ d'applications (Grob and Barry, 2004).

Un nombre restreint de composés peuvent être analysés par GC-MS en comparaison à une technologie d'analyse en LC-MS en raison des volatilités et stabilités thermiques requises pour chacun des composés. Il s'agit là de la limitation principale de cette technique vu que seuls 10% de tous les composés organiques peuvent être analysés par GC-MS. Pour contourner ce problème, des procédures de dérivatisation sont réalisées lors de l'étape de préparation des échantillons pour obtenir des composés volatils et thermiquement stables à haute température, élargissant la gamme d'utilisation de la technique.

Finalement, la GC-MS possède de telles caractéristiques qu'elle représente la technologie analytique la plus largement répandue dans les laboratoires d'analyses. En particulier, l'un de ses avantages consiste en sa capacité d'identification de milliers de composés inconnus au travers de l'utilisation de bibliothèques de spectres de masses conséquentes (par exemple, la banque de données du National Institute of Standards and Technology, NIST).

[14] En fonction des constructeurs, des termes différents existent pour nommer un même élément du MS bien que l'architecture de ce dernier reste similaire. La terminologie mise en place par Agilent® pour nommer les différents éléments du MS s'applique ici.

2.2 Considérations historiques

Ce paragraphe n'a pas pour but de retracer l'historique de la GC-MS mais notons que l'apparition des colonnes capillaires a simplifié le couplage du GC et de la MS, les systèmes de vides des spectromètres de masse s'adaptant facilement aux débits de gaz porteurs faibles (de l'ordre du millilitre par minute) utilisés avec ces colonnes (Grob and Barry, 2004).

En effet, les colonnes remplies utilisées initialement exigeaient des débits de gaz porteurs élevés (de l'ordre de dizaines de millilitres par minute) qui ne pouvaient être directement gérés par les systèmes de pompes des MS. Des interfaces étaient alors nécessaires lors du couplage GC et MS pour séparer (difficilement) le gaz porteur des composés d'intérêt après la chromatographie gazeuse, pour réduire la pression du mélange gazeux avant son entrée dans le MS. En revanche, vu les faibles débits de gaz porteur pratiqués avec les colonnes capillaires, ces dernières peuvent directement être connectées aux sources ioniques et ce sans altérer l'ionisation.

Outre le développement des colonnes capillaires, l'introduction d'ordinateurs pour le traitement des données et l'invention des colonnes GC en verre de silice dans la fin des années 1970 ont entraîné la mise sur le marché d'instruments GC-MS modernes utilisables pour une large gamme d'applications analytiques.

2.3 Principes de la GC-MS

La Figure 9 représente les composants majeurs d'une technologie d'analyse GC-MS.

Les analytes gazeux qui éluent du chromatographe selon leur affinité avec la phase stationnaire sont dirigés à l'aide de la ligne de transfert dans la source ionique du spectromètre de masse où ils sont ionisés puis fragmentés.

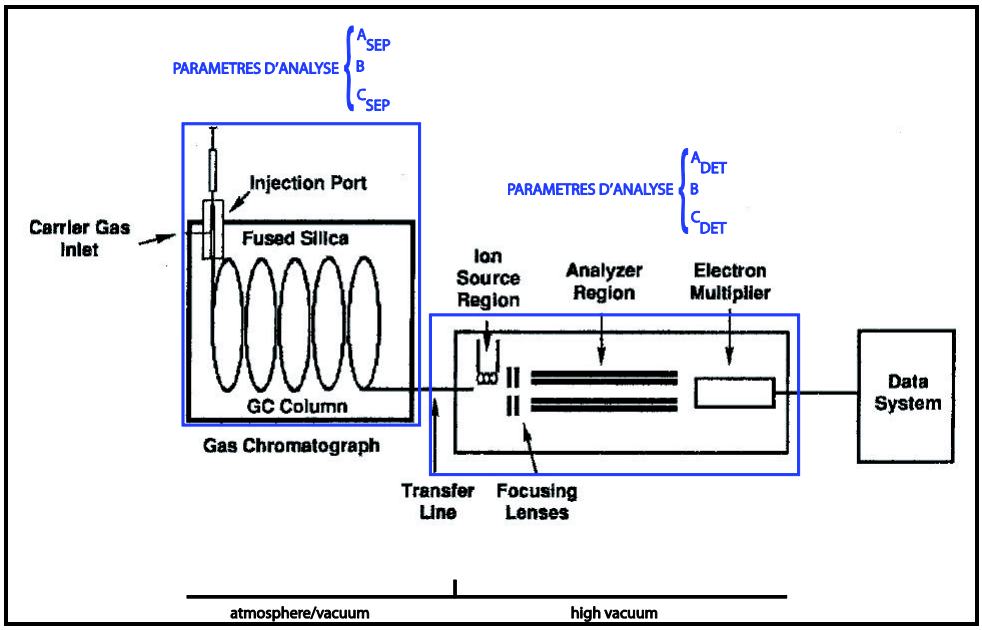


Figure 9. Représentation schématique d'un GC-MS, avec l'identification des paramètres d'analyse au niveau des technologies d'analyse de séparation et de détection (cf. Chapitre 5, Figure 23). Figure réalisée à partir de celle présente dans l'ouvrage correspondant (Grob and Barry, 2004)

Les ions produits sont séparés selon leurs rapports masse sur charge (m/z) par un filtre de masse (ou analyseur) puis détectés par un multiplicateur d'électrons ou de photons qui produit un signal proportionnel au nombre d'ions le frappant. Le spectre de masse qui résulte, pour chacun des composés du mélange, représente l'intensité relative (ou l'abondance) de ces ions en fonction de leurs rapports m/z. La plupart des ions n'ayant qu'une charge, leurs valeurs m/z indiquent leurs masses.

Comme le montre la Figure 10, au fur et à mesure de la séparation chromatographique, l'analyseur de masse effectue un nombre répété de scans sur l'intervalle m/z d'intérêt et le chromatogramme obtenu se compose d'un grand nombre de spectres de masse acquis consécutivement (Gross, 2011).

En chaque point du chromatogramme, les abondances de tous les ions détectés peuvent être sommées pour générer l'abondance ionique totale en ce point dans le temps. Par conséquent, nous obtenons un chromatogramme de tout le courant ionique (ou total-ion current chromatogram, TIC).

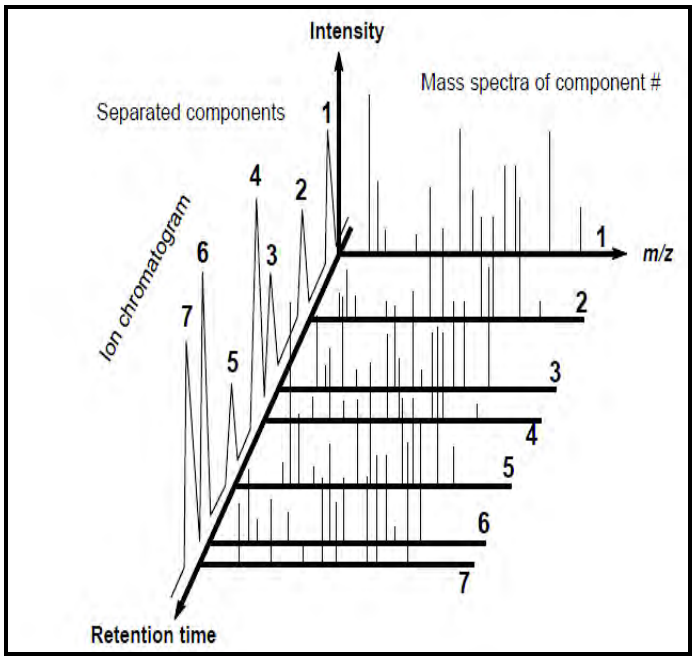


Figure 10. Représentation des trois dimensions en GC-MS : le temps de rétention, l'intensité et le rapport m/z (Gross, 2011)

Les prochains paragraphes détaillent les différents éléments d'un MS, c'est-à-dire, la source ionique, l'analyseur et le détecteur et décrivent ainsi le déroulement d'une analyse au niveau du MS, de l'ionisation à la détection.

2.4 Source ionique

Quelle que soit la source ionique utilisée, son but consiste à fournir une énergie suffisante pour l'ionisation des molécules de l'échantillon, tout en étant maintenue à une température assez haute pour éviter la condensation des composés de l'échantillon (Gross, 2011).

Les deux ionisations généralement rencontrées en GC-MS se nomment l'ionisation électronique (EI) et l'ionisation chimique (CI). L'ionisation électronique (que l'on retrouve également sous les termes d'ionisation à impact électronique ou encore d'impact électronique) est ici exclusivement décrite, la CI n'étant pas utilisée dans cette étude. Le propos sera centré sur l'architecture d'une telle source et les composants qui la constituent.

L'EI représente l'approche classique pour l'ionisation de composés organiques en spectrométrie de masse. Il s'agit d'une technique importante pour l'analyse de composés organiques ayant une polarité faible à moyenne, avec des poids moléculaires jusqu'à environ 1000 u.

La Figure 11 représente l'architecture d'une source EI et le principe de fonctionnement général est décrit par la suite.

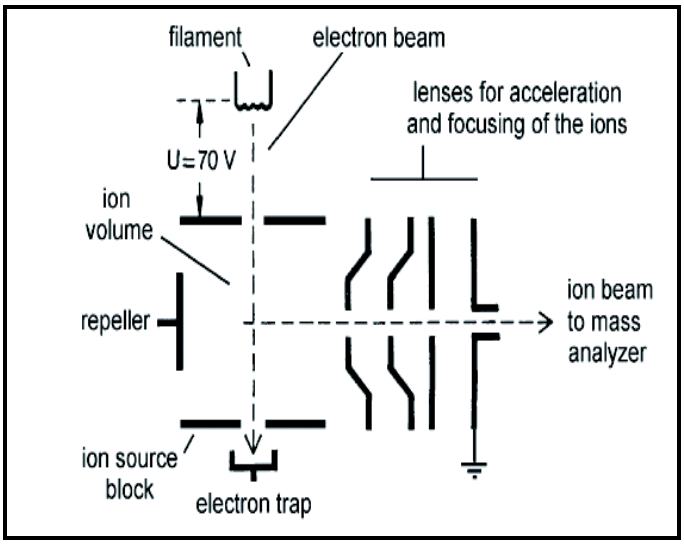


Figure 11. Architecture d'une source à impact électronique (EI) (Gross, 2011)

Création des électrons primaires (Gross, 2011)

Le faisceau d'électrons ionisants est produit par un fil métallique à chauffage résistif ou par un filament en rhénium ou tungstène. Le filament atteint des températures de plus de 2000°C durant l'analyse. Une large gamme de filaments est disponible auprès de différents fabricants et il n'existe pas de grande différence entre eux au niveau de leur qualité de fonctionnement. Le filament peut être un fil raide, un ruban ou une petite bobine.

Une paire d'aimants permanents, avec leur champ aligné en parallèle au faisceau d'électrons, est fixée au-dessus du filament et en-dessous du piège à électrons (electron trap), prévenant ainsi une dispersion des électrons dans tout le volume ionique. Un faisceau étroit est donc obtenu. Les électrons ont une trajectoire hélicoïdale et le faisceau résultant fait environ 1 mm de diamètre.

Les courants appliqués au filament pour le chauffer et obtenir la production d'électrons (nommées courants d'émissions) ont habituellement des valeurs entre 50 et 400 µA. En fonction de l'énergie d'ionisation souhaitée (en eV), le courant d'émission sera différent. L'efficacité d'ionisation et la production d'ions fragments dépendent fortement de la composition chimique de l'analyte et de l'énergie des électrons.

Déroulement de l'ionisation (Gross, 2011)

Le faisceau de molécules gazeuses de l'échantillon, neutres, entre dans la chambre d'ionisation (ou volume ionique) qui correspond à la région d'ionisation à l'intérieur de la source ionique. Cette entrée se fait dans un plan perpendiculaire au plan du papier et coupe le faisceau d'électrons au centre (cf. Figure 11). Pour réduire la perte d'ions il convient d'éviter leurs collisions avec les parois et ainsi les ions sont poussés hors de la chambre d'ionisation immédiatement après leur création grâce à une tension appliquée à l'électrode Repeller. Les ions sont par la suite accélérés et concentrés par un jeu de lentilles (Ion Focus et Entrance Lens) pour les emmener vers le filtre de masse (ou analyseur de masse) (cf. §2.7, Figure 18).

Une ionisation et une extraction ionique efficaces représentent les critères majeurs lors de la construction de sources ioniques.

Efficacité et sensibilité d'une source ionique EI (Gross, 2011)

L'efficacité totale d'une source ionique EI dépend des propriétés intrinsèques du processus d'ionisation, de l'architecture de la source ionique et des valeurs appliquées aux éléments de la source lors de l'analyse.

Il faut noter qu'une toute petite fraction de l'échantillon introduit s'ionise, alors que la grande majorité se retrouve aspirée par les pompes à vide. Cela s'explique par la combinaison de deux paramètres. Tout d'abord, de longs « mean free path » aussi bien pour les électrons que pour les ions.

Le « mean free path » correspond à la distance moyenne qu'un ion parcourt dans une région non close avant qu'il ne percute quelque chose. En spectrométrie de masse, le « mean free path » doit être suffisamment long pour que les ions de l'échantillon puissent traverser de la source ionique au détecteur sans qu'il n'y ait de collisions avec les autres molécules. C'est le système de pompe à vide qui crée un « mean free path » convenable grâce à un vide important[15]. Le second paramètre correspond à la faible section de collision du faisceau d'électron lui-même (comme précisé précédemment, le chemin au travers duquel l'ionisation est efficace est d'environ 1 mm). Ainsi, environ 1% des molécules entrant dans le volume d'ionisation est ionisé et atteindra le détecteur, ce qui fournit une sensibilité élevée en comparaison aux autres méthodes d'ionisation.

2.5 Analyseurs

Au fur et à mesure de leur sortie de la source, les ions entrent dans le filtre de masse (c'est-à-dire, l'analyseur) où leur séparation se fait selon leurs rapports m/z respectifs. L'intervalle de masse d'intérêt est scanné, résultant dans la séparation des ions dans des domaines d'espace ou de temps (Grob and Barry, 2004). Dans le cas d'une ionisation EI aucune restriction n'existe quant au choix du type d'analyseur de masse.

Le « magnetic sector » a été largement utilisé dans les premiers jours de la spectrométrie de masse avant d'être remplacé par le quadrupole en raison de vitesses de scans significativement plus élevées. Le couplage de plusieurs analyseurs (MS^n) a permis une amélioration dans la sélectivité analytique, avec en particulier l'utilisation de double ou triple quadrupole. Récemment, des mesures exactes de masses sont devenues une nécessité augmentant ainsi la demande pour des analyseurs tels que le TOF (time-of-flight) plutôt que des systèmes à quadrupole. On peut également retrouver en EI des analyseurs ICR (Ion Cyclotron Resonance) mais en raison de leurs coûts élevés ils ne sont que rarement utilisés.

[15] L'importance du vide est à souligner. En l'absence d'autres composés (par exemple, dans le vide), les ions moléculaires des composés respectifs se fragmentent en d'autres ions, radicaux ou molécules neutres. Les masses et l'abondance de ces fragments dépendent de la nature des molécules de départ, et c'est ce qui donne à la spectrométrie de masse son pouvoir d'identification.

Le quadrupole représente l'analyseur le plus répandu dans les appareillages GC-MS et par conséquent dans les laboratoires analytiques. De plus, les méthodes analytiques en MS étudiées dans ce travail de recherche utilisent majoritairement le quadrupole comme analyseur. Par conséquent, seul celui-ci est décrit.

Les quadrupoles

Les quadrupoles modernes couvrent un large intervalle de masses, jusqu'à 2000 m/z, avec un haut pouvoir de résolution. Le quadrupole possède un grand nombre d'avantages, en effet :

- Il a une haute transmission.
- Il est léger, très compact et est relativement bon marché.
- Il permet une vitesse de scans élevée.

Comme son nom l'indique, le quadrupole consiste en 4 tubes cylindriques arrangés tel qu'illustré sur la figure suivante.

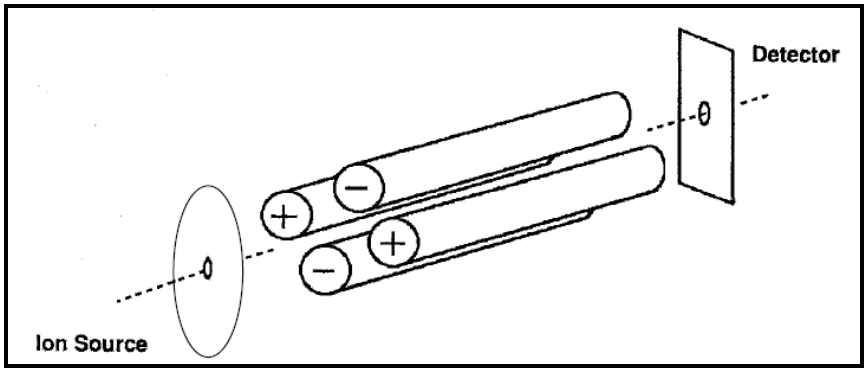


Figure 12. Représentation schématique du quadrupole (Grob and Barry, 2004).

Lors du fonctionnement, les tubes diamétralement opposés travaillent en tandem. Sur la première série de tubes une tension « direct-current » (DC) positive est appliquée tandis que sur la seconde une tension DC négative, de même valeur, est appliquée. De plus, les quatre tubes ont une tension « radiofrequency » (R_f) oscillante à 1 MHz. Les potentiels R_f et DC appliqués aux tubes ne permettent qu'aux ions avec un rapport m/z spécifique d'avoir une trajectoire stable et ainsi de passer jusqu'au détecteur. En augmentant simultanément les potentiels R_f et DC, les ions de m/z croissant vont passer à travers l'analyseur puis être détectés (Grob and Barry, 2004).

Techniques d'acquisition

Les analyseurs possèdent deux modes principaux d'acquisition des masses (ou techniques de scan) (Grob and Barry, 2004; Gross, 2011) :

- le mode d'acquisition SCAN;
- et le mode d'acquisition SIM (Selected Ion Monitoring).

Le choix parmi ces deux modes de scan découle des tâches qui doivent être accomplies telles que par exemple l'identification de composés inconnus, l'analyse de traces ou la quantification de composés cibles, chacune d'elles ayant ses propres exigences instrumentales.

Dans le premier mode d'acquisition, l'analyseur scanne à plusieurs reprises un intervalle m/z prédéfini qui couvre tous les ions moléculaires et ions fragments obtenus durant l'analyse chromatographique d'un échantillon constitué de plusieurs composés. Schématiquement, le quadrupole fonctionne de la manière suivante (cf. Figure 13).

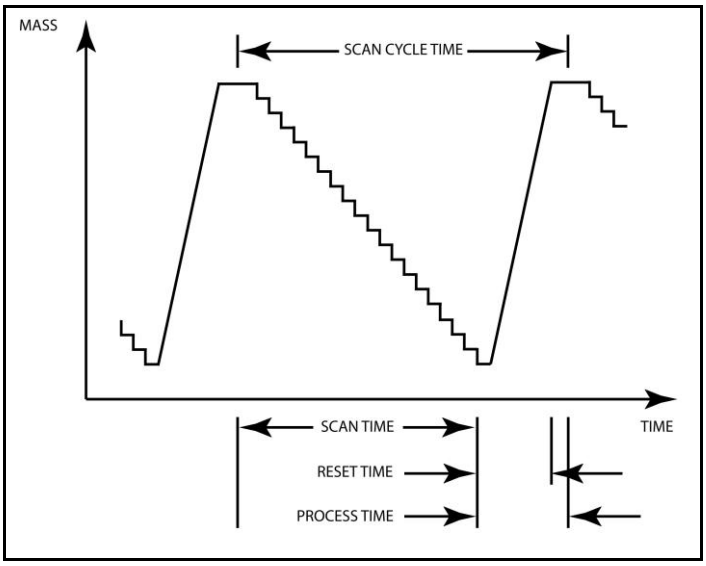


Figure 13. Représentation schématique du mode de fonctionnement d'un quadrupole (Figure adaptée d'une documentation Agilent®)

Au commencement d'un scan, le quadrupole est prêt et attend à la valeur m/z maximale de l'intervalle de scan qui a été spécifié par l'opérateur. Pour acquérir un spectre de masse, l'analyseur avance par paliers de 0.1 AMU jusqu'à la valeur m/z minimale de l'intervalle.

La vitesse que met le quadrupole pour effectuer un scan (c'est-à-dire, la vitesse d'acquisition, exprimée en nombre de scans par seconde) dépend de l'intervalle de scan m/z ajusté et du *sampling rate* ou taux d'échantillonnage. Ce dernier correspond au nombre de fois où l'abondance de chacune des masses est mesurée ou échantillonnée durant un scan. Lorsque le quadrupole atteint la limite inférieure de l'intervalle, il retourne à la limite supérieure en préparation du prochain scan (reset time). Au même moment, les données sont transférées au logiciel de traitement des données (process time).

Le temps nécessaire au quadrupole pour effectuer un scan doit être ajusté judicieusement s'agissant d'un compromis entre la qualité chromatographique des pics des composés (forme du pic, nombre de points de données) et la qualité des spectres de masse respectifs. En effet, si l'on scanne rapidement, chaque pic chromatographique se compose de plusieurs points de données et spectres de masse. La reconstruction du chromatogramme est possible, mais la qualité des spectres peut s'avérer mauvaise en raison de la vitesse à laquelle ils sont obtenus. A l'inverse, si l'on scanne lentement, la qualité des données spectrales peut être bonne mais la forme du pic chromatographique risque d'être mal définie (forme non Gaussienne).

La vitesse d'acquisition peut être ajustée en appliquant la valeur de sampling rate adéquate, une fois l'intervalle de masses m/z déterminé. Plus le sampling rate augmente, plus le quadrupole passe du temps sur chacun des ions de l'intervalle spécifié et ainsi le temps nécessaire pour effectuer un scan sera plus important. Il en résulte une vitesse d'acquisition plus faible et donc moins de points de données sur l'entier du pic chromatographique que si un sampling rate plus faible avait été appliqué[16] (cf. Figure 14). Il faut noter que la sensibilité obtenue sera plus importante pour un sampling rate élevé que pour un sampling rate plus faible, l'échantillonnage de chacun des ions étant plus important.

[16] La problématique du nombre minimum de points par pic à atteindre sera discutée dans le Chapitre 3 ayant trait aux méthodes analytiques rapides, le nombre de points étant une contrainte lors de l'implémentation de ces méthodes d'analyse.

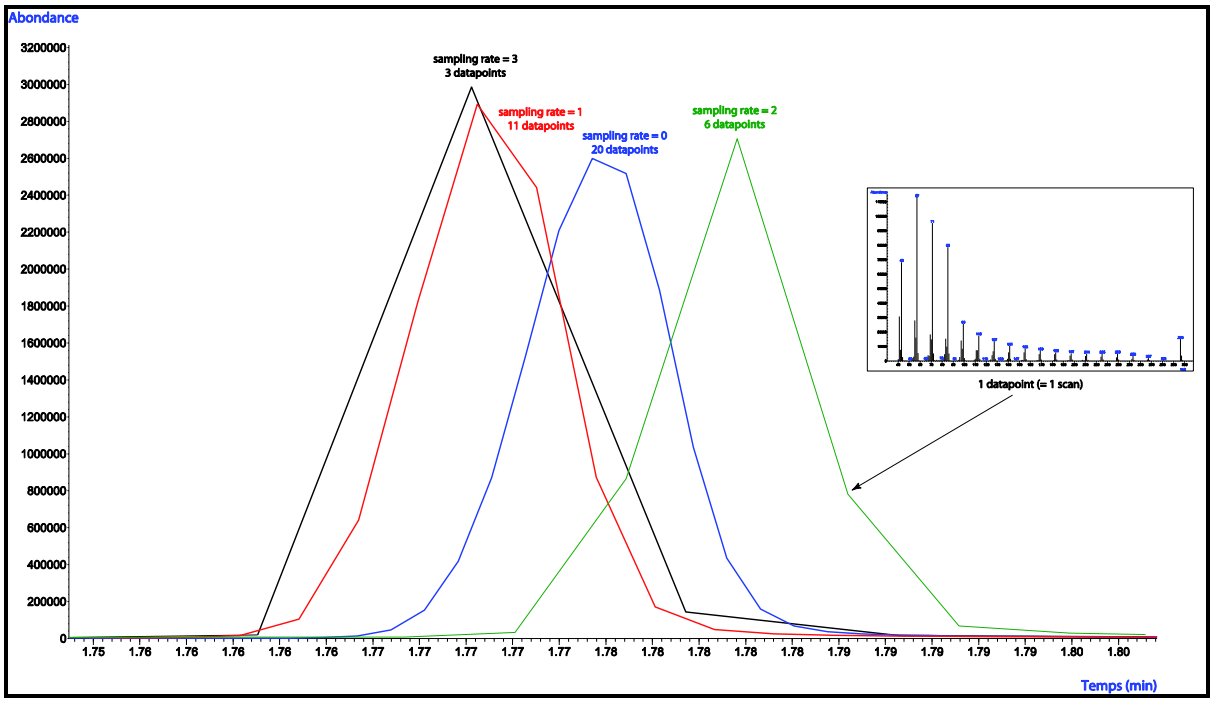


Figure 14. Influence du sampling rate sur la qualité chromatographique des pics. Dans cet exemple, l'intervalle des masses va de 40 à 450 m/z entraînant des vitesses d'acquisition de 20 et 3 scans/sec pour des sampling rate de 0 et 3, respectivement.

Quant à lui, le mode de scan SIM consiste à n'échantillonner que certaines valeurs m/z durant la séparation chromatographique (cf. Figure 15). Par rapport au mode d'acquisition SCAN, il en résulte une sensibilité grandement améliorée, une meilleure définition de la forme du pic, et de meilleures exactitude et précision, l'analyseur passant son temps à n'échantillonner que des masses désirées. Cette technique s'applique en particulier lors de quantification de composés cibles, d'analyses en traces et d'analyses d'échantillons complexes.

En mode d'acquisition SCAN la durée de mesure de chacune des masses de l'intervalle spécifié correspond à environ 100 µsec. En mode d'acquisition SIM, chaque masse est mesurée pendant 100 msec. Sachant que le rapport signal sur bruit est proportionnel à la racine carrée du temps de mesure, il en résulte que le mode SIM est approximativement 30 fois plus sensible que le mode scan. En pratique des améliorations de 20 à 100 fois sont possibles, en fonction de l'instrument, du bruit de fond, de la complexité de l'échantillon, etc.

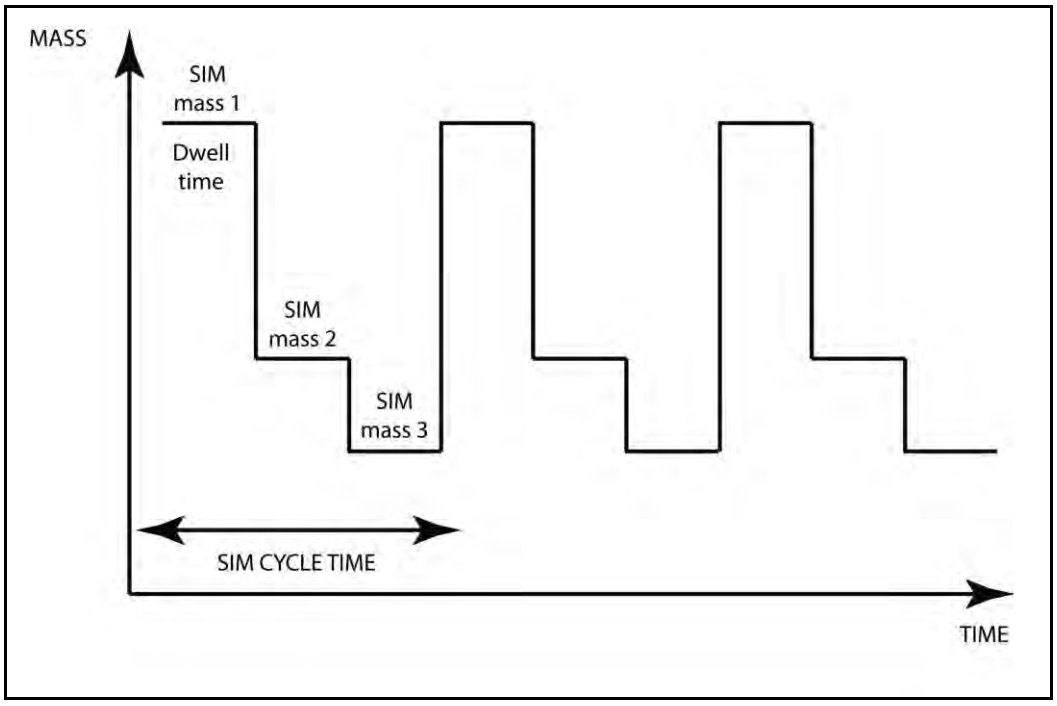


Figure 15. Représentation schématique du mode de fonctionnement de la technique d'acquisition SIM (Figure adaptée d'une documentation Agilent®)

2.6 Multiplicateur d'électrons

Une fois triés selon leurs rapports masse sur charge par le filtre de masse, les ions atteignent le détecteur. Les détecteurs utilisés en MS doivent avoir une réponse rapide et un gain important pour convertir les petits courants ioniques générés en signaux enregistrables.

Le plus populaire d'entre eux est le multiplicateur d'électrons (ou electromultiplier). Celui-ci se compose d'une série de dynodes sur lesquelles la tension appliquée se trouve entre 1 et 3 kV. Lorsque le faisceau d'ions frappe les dynodes il en résulte une émission d'électrons. A chaque impact d'un électron le long du multiplicateur d'électrons, une émission d'électrons se produit. Il en résulte une multiplication du faisceau d'ions pour un gain en courant d'environ 10^5 (cf. Figure 16) (Grob and Barry, 2004; Gross, 2011).

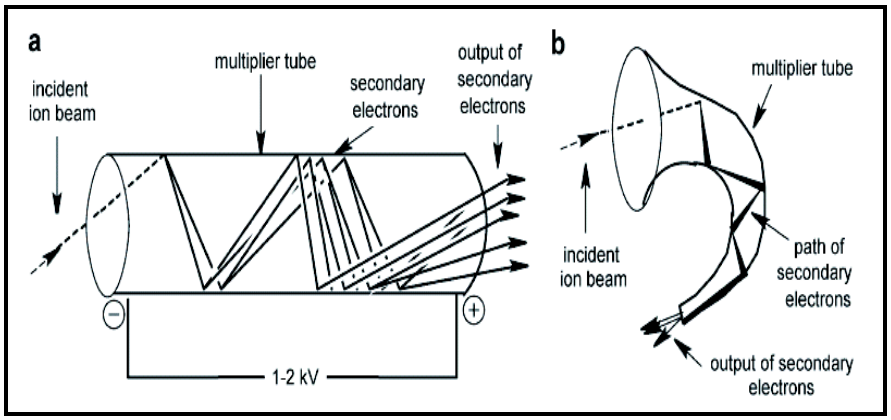


Figure 16. Architectures possibles d'un multiplicateur d'électrons : a) à canal linéaire et b) à canal courbé (Gross, 2011)

Fréquemment, une dynode de conversion avec une tension plus élevée (entre 5 et 20 kV) est insérée avant le multiplicateur. De polarité adéquate elle attire les ions qui sortent de l'analyseur. Leurs impacts sur la dynode de conversion créent des ions secondaires ou des électrons qui peuvent être utilisés pour la détection suivante. Cette dynode à haute énergie (High Energy Dynode, HED) permet ainsi d'augmenter l'énergie du faisceau d'ions (cf. Figure 17). Lors de l'augmentation de la tension appliquée au multiplicateur d'électrons, le signal s'en trouve amplifié.

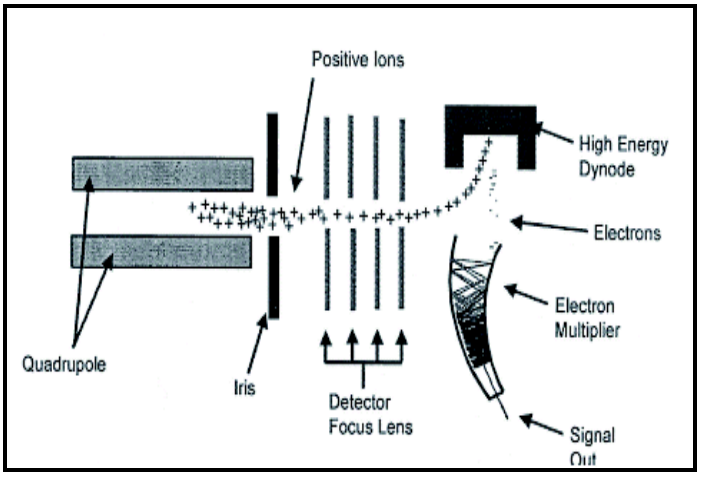


Figure 17. High Energy Dynode et Electromultiplier (Source Agilent®)

Un autre détecteur consiste en le multiplicateur de photons (ou photomultiplier). Le faisceau d'ions frappe une cible recouverte de phosphore qui convertit les ions en photons qui sont ensuite amplifiés et détectés. Les tensions appliquées à ces détecteurs sont plus faibles (400-700 V) et ces derniers ont une durée de vie plus importante que les multiplicateurs d'électrons conventionnels vu qu'ils consistent en des unités fermées non sujettes aux contaminations extérieures.

2.7 « Tune » ou réglage du MS

NB : Les informations discutées dans ce paragraphe sont tirées de la documentation Agilent®, n'étant que très peu étudiées dans les publications scientifiques.

Buts du réglage du MS

Lors de l'utilisation prolongée d'un appareillage, il convient de s'assurer que le MS fonctionne de manière correcte avant de commencer des analyses. En effet, cela pourrait ne pas être le cas en raison des effets exercés par les contaminations sur les éléments de la source par exemple.

Pour ce faire on va procéder au « tune » du MS. Ce procédé fait partie intégrante de la procédure de contrôle qualité d'un appareillage analytique, en particulier de la partie MS. Le but de cette procédure consiste en l'identification de problèmes potentiels et en l'optimisation de la performance du MS. Avec les informations qui en découlent, le nettoyage de la source ionique ou le remplacement du multiplicateur d'électrons peuvent être déterminés par exemple.

Principes de fonctionnement du réglage du MS

Le PFTBA (perfluoroterbutylamine) est le liquide utilisé pour procéder au réglage du MS. Une fois vaporisé et ionisé, ce composé de calibration se fragmente en plusieurs ions parmi lesquels trois d'entre eux (les ions 69, 219 et 502 m/z) sont utilisés lors du « tune » pour vérifier le bon fonctionnement du MS.

La procédure du « tune », automatique, consiste à ajuster un certain nombre de paramètres du spectromètre de masse pour correspondre à des valeurs prédéterminées. Certains paramètres sont purement électroniques et n'affectent que la manière dont l'électronique traite le signal. D'autres paramètres influent sur la qualité des spectres de masse et par conséquent sur la capacité à faire correspondre un spectre de masse donné à ceux des banques de données de spectres de masse. Ces derniers paramètres affectent les éléments de la source ionique, du filtre de masse et du détecteur. Les paramètres de la source ionique affectent le nombre d'ions produits, le nombre d'ions dirigés vers le filtre de masse et la quantité relative d'un ion de masse donnée qui est dirigée dans le filtre de masse. Les paramètres du filtre de masse dans un quadrupole affectent les largeurs de pic, les assignements des masses, la résolution de la masse et la sensibilité. Enfin, les paramètres du détecteur affectent l'intensité du signal et la sensibilité du système.

Plus précisément, le « tune » ajuste pour la source ionique les tensions de l'électrode Repeller et des lentilles Ion Focus et Entrance Lens, pour le quadrupole l'AMU gain/offset et le Mass Axis gain/offset, et pour le détecteur la tension du multiplicateur d'électrons (cf. Figure 18). Ces paramètres sont intimement reliés entre eux et l'optimisation d'un paramètre affecte la valeur optimale d'un autre. Concrètement, les objectifs du « tune » consistent à maximiser la sensibilité tout en maintenant une résolution acceptable, assurer une identification exacte des masses et fournir les abondances relatives désirées le long du spectre.

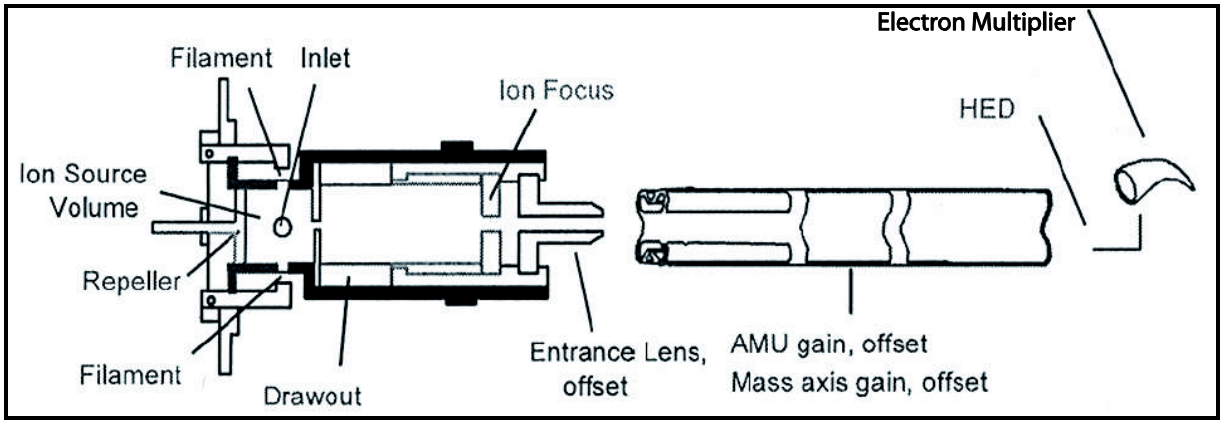


Figure 18. Représentation schématique du MS et des paramètres à ajuster pour la source ionique, le filtre de masse et le détecteur (architecture Agilent®)

Une fois les valeurs optimales déterminées elles sont enregistrées dans un fichier informatique nommé le « tune file » qui fait partie intégrante de la méthode analytique (cf. §5.1, Figure 23). Le Tableau 2 ci-dessous présente l'ensemble des paramètres ajustés lors du « tune » et constituant ce « tune file ».

Elément	Valeurs	Effet
Filament	Énergie d'ionisation 70 eV	Energie du faisceau d'électrons
	Courant d'émission de 300 µA	Nombre d'électrons généré
Repeller	0 – 42.7 volts	Pousse les ions hors de la source
Drawout	Ground potential	Ouverture de l'entrée aux lentilles
Ion Focus	0 – 242.0 volts	Abondance relative
Entrance Lens	0 – 128 mV / amu	Abondance relative
Entrance Lens Offset	0 – 127.5 volts	Abondance relative
AMU Gain	0 – 4095	Affecte la largeur des pics
AMU Offset	0 – 255	Affecte la largeur des pics
Mass Axis Gain	± 2047	Assignement de la masse
Mass Axis Offset	± 499	Assignement de la masse
High Energy Dynode (HED)	- 10,000 volts	Conversion des ions en électrons
Electron Multiplier	0 – 3000 volts	Sensibilité

Tableau 2. Ensemble des paramètres ajustés lors du tune

Bien que le courant d'émission du filament et l'énergie d'ionisation qui en découle puissent être ajustés, les valeurs par défaut sont en général appliquées. L'énergie des électrons (énergie d'ionisation, en eV) représente l'énergie de référence pour le faisceau d'électrons. L'énergie d'ionisation est généralement de 70 eV pour produire des spectres de masse reproductibles pour les molécules organiques, permettant par la suite une recherche dans les banques de données de composés.

Le courant d'émission affecte la production d'électrons par le filament, et influence donc l'ionisation puis la fragmentation. Ainsi, la sensibilité résultante est affectée. Réduire l'énergie des électrons résultera en une ionisation plus douce (moins de fragmentation) des molécules organiques et une diminution notable de la sensibilité. Si le courant augmente, le nombre d'électrons émis augmente, la fragmentation de l'échantillon également mais la durée de vie du filament sera plus courte.

La Figure 19 illustre le rapport généré suite au « tune » du MS. On peut y voir en rouge les paramètres d'analyse C_{DET} au niveau de la source (les tensions du Repeller et des lentilles), de l'analyseur (pour la calibration de l'assignement des masses et de la largeur des pics) et du détecteur (la tension du multiplicateur d'électrons). En bleu figurent les paramètres qui sont ajustés et contrôlés lors du « tune » (points 1 à 8, Figure 19).

Les valeurs obtenues pour les paramètres identifiés sur la figure doivent être contrôlées pour s'assurer du bon fonctionnement du MS (en d'autres termes, si les valeurs obtenues ne correspondent pas à celles attendues, alors une intervention est nécessaire : un nettoyage de source par exemple). Précisément :

- La largeur respective de ces trois pics devrait être de 0.5 +/- 0.1 m/z (*Pw50*, point 1 et point 2).
- Les masses devraient être comprises dans un intervalle de +/- 0.1 m/z pour les ions 69, 219, et 502 (*Mass*, point 5).
- Les abondances relatives devraient montrer que le pic à 69 m/z est le plus important. Relativement à ce pic, celui à 219 m/z et celui à 502 m/z devraient être dans l'intervalle spécifié dans l' « autotune » effectué (*Rel Abund*, point 6).
- Les masses isotopiques devraient être supérieures de 1 m/z aux masses des ions parents (*Iso Mass*, point 7).
- Les rapports isotopiques (indiquant les abondances relatives des isotopes naturellement présents) devraient être proches des valeurs théoriques de 1.08 pour 69 m/z (entre 0.5 et 1.6%), 4.32 pour 219 m/z (entre 3.2 et 5.4%), et 10.09 pour 502 m/z (entre 7.9 et 12.3%) (*Iso Ratio*, point 8).
- Si la masse 28 est supérieure à la masse 18, il pourrait y avoir une fuite d'air quelque part dans le système. De plus, le rapport entre les ions 18 et 69 devrait être inférieur à 20% et celui entre les ions 28 et 69 inférieur à 10%. Si ce n'est pas le cas, alors cela peut s'expliquer si le tune a été effectué moins d'une heure après le vide du MS (*vent*) ou s'il s'agit du premier tune après avoir rempli le vial de calibration.

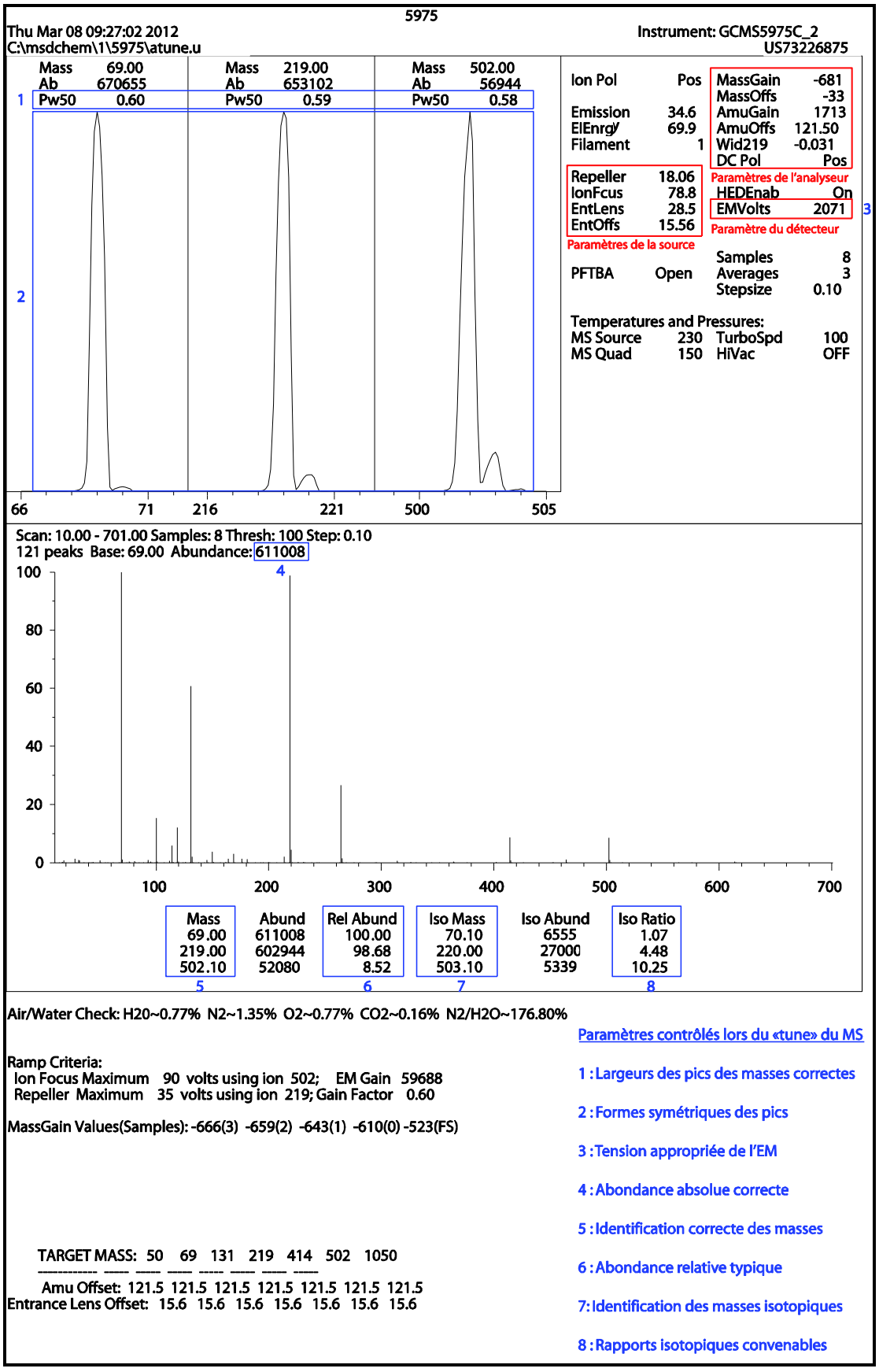


Figure 19. Rapport obtenu suite au « tune » du MS

2.8 Paramètres C_{DET} du MS et similarité des résultats

Le Tableau 3 récapitule l'ensemble des paramètres C_{DET} pour lesquels des valeurs doivent être ajustées lors de l'implémentation d'une méthode analytique.

Niveau	Paramètres
Source ionique	Température de la ligne de transfert*
	Solvent delay*
	Température de la source*
	Energie d'ionisation*
	Courant d'émission*
	Tensions des éléments de la source (Repeller, Ion Focus, Entrance Lens)
Analyseur	Température de l'analyseur *
	Mode d'acquisition (SCAN, SIM)*
	Intervalle de masses en SCAN ou ions en SIM *
	Vitesse d'acquisition*
	Seuil*
	Paramètres définissant l'assignement des masses Mass Gain/Offset et AMU Gain/Offset
Détecteur	*Tension du multiplier*

Tableau 3. Paramètres modifiables d'une analyse à l'autre, qu'ils se trouvent au niveau de la source ionique, du quadrupole ou du multiplicateur d'électrons.

Comme cela a été précisé dans la partie introductive de ce travail, les valeurs appliquées aux paramètres d'analyse C peuvent être modifiées d'une analyse à l'autre, que ce soit les paramètres C_{SEP} ou C_{DET}. Techniquement, ceci est vrai. Toutefois, au sein des paramètres C_{DET}, il est important de souligner que les paramètres identifiés par un astérisque (*) dans le tableau ci-dessus ne sont jamais modifiés une fois la méthode analytique optimisée puis utilisée en systématique. A l'inverse, les trois paramètres restants que l'on nommera *les paramètres du « tune »* (définissant la source ionique, l'analyseur et le détecteur) peuvent être modifiés (cf. Tableau 3).

Cette possibilité de modification des valeurs de ces paramètres se retrouvent dans deux cas de figure majeurs :

- lors du réglage du MS (le « tune »), exécuté régulièrement et où les valeurs optimales déterminées automatiquement varient avec l'usage de l'appareil ;
- lors d'analyses en systématique où, jour après jour, le laboratoire doit certifier une continuité des résultats analytiques, garantir une réponse analytique similaire au fil du temps et assurer une reproductibilité des résultats analytiques. Les valeurs des paramètres du « tune » peuvent alors être modifiées par l'opérateur.

Ainsi, la méthode analytique peut être modifiée au niveau des paramètres d'optimisation de la technologie d'analyse de détection C_{DET}, en particulier les paramètres du « tune », au fur et à mesure de l'utilisation de l'instrument analytique et de la réalisation d'analyses. Les valeurs de ces paramètres peuvent ainsi soit être modifiées par l'opérateur, qui va lui-même décider quelles valeurs appliquer aux éléments de la source, de l'analyseur et du détecteur pour obtenir l'effet souhaité, soit à l'inverse résulter du processus automatique de réglage du MS.

Le Tableau 4 ci-dessous détaille *les paramètres du « tune »* pouvant évoluer d'une analyse à l'autre, sur un même instrument analytique et au fur et à mesure de son utilisation.

Niveau	Paramètres	Influence	Description
Source ionique[17]	Repeller	Sensibilité	Si tension trop faible : trop peu d'ions vont quitter la source, résultant en une mauvaise sensibilité et une faible réponse pour les hautes masses. Si tension trop élevée : trop d'ions à une trop grande vélocité vont quitter la source. Il en résulte un mauvais filtre des masses et une mauvaise résolution des petites masses.
	Ion Focus		Lors du tune du MS la tension offrant la meilleure abondance ionique est choisie.
	Entrance Lens		Ce paramètre fait référence au gain de la lentille d'entrée, une valeur utilisée pour déterminer une tension, fonction de la masse, qui est appliquée à la lentille d'entrée. Lors du tune du MS la valeur offrant la meilleure abondance est choisie. Une augmentation dans la tension augmente les abondances des hautes masses mais diminue celle des faibles masses.
	Entrance Lens Offset		Tension constante appliquée sur la lentille d'entrée. Une augmentation dans l'offset entraîne une augmentation de l'abondance des ions de petites masses sans diminuer substantiellement l'abondance des ions de masses élevées.
Analyseur	Mass Gain / Mass Offset	Assignement des masses et largeur des pics	Facteurs utilisés dans l'équation de calibration de l'axe des masses. Mass Offset affecte toutes les masses de la même manière. Mass Gain affecte les pics de masses élevées plus que ceux de masses faibles.
	AMU Offset / AMU Gain		Affectent le rapport des tensions DC sur R_f. Affectent la largeur des pics de masse. AMU Offset affecte les largeurs de pic de la même manière pour toutes les masses. AMU Gain affecte les pics de masses élevées plus que ceux de masses faibles.
Détecteur	Electron Multiplier	Sensibilité	Augmenter la tension augmente la sensibilité du signal. L'abondance est augmentée en élevant le signal sortant du spectromètre de masse. Augmenter cette tension diminue aussi le temps de vie du multiplicateur. En général, la valeur appliquée est la valeur la plus faible offrant une sensibilité adéquate.

Tableau 4. Détails des paramètres C_{DET} pour lesquels les valeurs sont modifiables d'une analyse à l'autre

[17] Le Repeller, l'Ion Focus et l'Entrance Lens travaillent conjointement pour affecter l'abondance des masses élevées. Le Repeller est particulièrement efficace pour changer l'abondance des hautes masses relativement à celle des faibles masses.

D'après la méthodologie d'harmonisation des méthodes analytiques, il est considéré que si ces dernières présentent des distinctions dans les valeurs des paramètres du « tune », elles ne sont pas considérées comme différentes. Selon cette méthodologie, il ne s'agit pas en effet de paramètres essentiels pour assurer la similarité des résultats analytiques. Parmi ces paramètres, ceux affectant la sensibilité sont particulièrement intéressants dans le contexte de ce travail de recherche, en particulier l'impact de leur modification sur la similarité des résultats analytiques obtenus. Le postulat de la méthodologie d'harmonisation des méthodes analytiques, selon lequel des méthodes analytiques présentant des différences dans les valeurs des paramètres du « tune » ne sont pas considérées comme différentes, pourrait être évalué au regard des résultats obtenus. Il s'agirait concrètement d'estimer si ces différences sont significatives ou non pour l'obtention de résultats analytiques similaires.

Chapitre 3 Les analyses chromatographiques rapides

Dans l'ensemble des domaines d'analyses, assurer la qualité des analyses réalisées et diagnostiquer dans un laps de temps réduit l'apparition de tout problème revêt un intérêt crucial. Une manière d'y parvenir consiste à mettre en place des techniques chromatographiques rapides pour obtenir un contrôle qualité efficace, une fréquence d'analyse et une productivité des laboratoires accrues ainsi qu'un coût par analyse diminué à l'aide de résultats obtenus en quelques minutes. Bien que le but de telles méthodes d'analyses consiste à réduire le temps d'analyse à un minimum, une séparation chromatographique correcte et des résultats qualitatifs et quantitatifs fiables sont conservés. Dans le cadre du profilage chimique de produits stupéfiants, les méthodes analytiques présentent l'avantage de pouvoir procéder à l'établissement et l'exploitation rapides des profils chimiques.

Ces méthodes analytiques présentent par conséquent un intérêt pour le profilage chimique et le prochain paragraphe les concerne donc. Les méthodes analytiques rapides de chromatographie en phase gazeuse ainsi que les éléments théoriques et pratiques qui sous-tendent le passage vers ces méthodes d'analyses rapides sont discutés, la méthode d'analyse de référence étant la GC-MS et les méthodes d'analyses rapides qui seront développées dans cette recherche se basant principalement sur une telle technologie de séparation.

3.1 Des méthodes analytiques conventionnelles aux méthodes analytiques rapides

Les principes et la théorie de la chromatographie en phase gazeuse rapide sont connus depuis les années 1960 suite à l'introduction des colonnes capillaires qui ont ouvert la voie à l'analyse rapide (Sacks, 2004). Mais son utilisation en routine est restée limitée jusqu'au début des années 2000, la durée et la vitesse d'une analyse n'étant alors pas des préoccupations pour les laboratoires d'analyse, qui se concentraient sur les problèmes de séparation et d'identification des composés de mélanges complexes (Cramers et al., 1999).

A la fin des années 90, la diminution du coût par analyse, la nécessité d'augmenter la productivité du laboratoire analytique (grand nombre d'échantillons à analyser, résultats nécessaires plus rapidement) ont débouché sur un intérêt pour les méthodes analytiques rapides. L'utilisation en routine des méthodes analytiques rapides n'a pu se faire que grâce à l'amélioration des équipements, ces derniers devenant alors convenables pour l'implémentation de telles méthodes analytiques (Sacks, 2004).

3.2 Terminologie

Le temps d'analyse en chromatographie en phase gazeuse dépend du type d'échantillon, du nombre de composés à séparer[18] et des conditions expérimentales choisies (Matisová and Dömötörová, 2003). En fonction des approches mises en place pour accélérer la vitesse d'une méthode analytique en GC (cf. § 3.3), les analyses sont rapides à différents degrés. Les termes de « High Speed », « Fast GC », « Very Fast GC » voire « Ultra Fast GC » se retrouvent alors dans la littérature. Une telle classification voit son intérêt lors du choix de l'appareillage – qui peut être conventionnel ou bien spécialisé pour l'analyse rapide – utilisé pour implémenter la méthode analytique.

Plusieurs paramètres caractérisent la vitesse d'une méthode d'analyse en GC selon la terminologie définie par les auteurs correspondants et les prochains paragraphes discutent les divers facteurs employés pour classifier les méthodes analytiques rapides.

Le "speed enhancement factor" (SEF) représente l'un des premiers facteurs défini pour caractériser la rapidité des méthodes analytiques (Dagan and Amirav, 1996). Il correspond au rapport de la vitesse linéaire du gaz porteur et de la longueur de la colonne, en comparaison à la même analyse avec une colonne conventionnelle et des conditions GC normales. « Fast GC », « Very Fast GC » et « Ultra Fast GC » sont définies avec un SEF allant respectivement de 5 à 30, de 30 à 400 et de 400 à 4000.

[18] A noter que lors d'une détection en MS, le nombre de composés à séparer n'est plus une contrainte dans la mesure où la détection et l'identification de composés qui co-éluent restent possibles.

Mastovska et Lehotay (2003) critiquent l'emploi d'un tel facteur car il ne reflète pas forcément la diminution exacte du temps d'analyse, la température de la colonne et le programme de température du four n'étant pas pris en compte alors qu'il s'agit de paramètres pouvant permettre la diminution du temps d'analyse.

Selon Klee and Blumberg (2002), les aspects fondamentaux de séparation de pic et de capacité de pic devraient être pris en compte. Une définition qui tient compte du degré de séparation en fonction du temps, telle qu'une classification à l'aide de la largeur des pics, devrait être privilégiée (Matisová and Dömötörová, 2003). Considérant que toute réduction du temps d'analyse résulte dans une réduction identique de la largeur de la zone chromatographique, en raison du temps de résidence plus court des composés dans la colonne, le temps d'analyse et la largeur des pics à mi-hauteur sont par conséquent recommandés pour définir les méthodes analytiques en GC rapides (Deursen et al., 1999; Korytár et al., 2002) :

Méthode analytique	Temps d'analyse	Largeur de pics
« Fast »	En minutes	1 – 3 s
« Very Fast »	En secondes	0.03 – 0.2 s
« Ultra Fast »	Inférieure à la seconde	0.005 – 0.03 s

Tableau 5. Classification des analyses rapides en GC (Deursen et al., 1999; Korytár et al., 2002)

Le tableau qui suit découle quant à lui d'une définition des analyses rapides plus spécifique et tient compte de manière intéressante des rampes de température et des dimensions de colonne implémentées (Bicchi et al., 2004). Il s'agit en effet de deux approches implémentées pour réduire le temps d'analyse d'une méthode.

Méthode analytique	Temps d'analyse	Largeur de pic	Rampes de température	Longueur de la colonne	Diamètre interne de la colonne
« Fast »	< 10 min	0.5 – 2 s	20 – 60°C / min	5 à 15 m	0.1 – 0.25 mm
« Ultra Fast »	< 1 min	0.05 – 0.2 s	> 1°C / s	2 – 10 m	0.05 – 0.1 mm

Tableau 6. Classification des analyses rapides en GC (Bicchi et al., 2004)

Le tableau suivant résume les caractéristiques des différentes méthodes rapides nommées dans la littérature en comparaison aux méthodes analytiques conventionnelles (Mastovská and Lehotay, 2003). Cette classification sera utilisée dans le cadre de ce travail de recherche.

Méthode analytique	SEF	Temps d'analyse	Largeur de pic (à mi-hauteur)	Fréquence d'acquisition spectrale[19]
Conventionnelle	0.5 – 5 (typiquement 1)	> 10 min	> 1 s	< 2.5 Hz
« Fast »	5 – 30 (autour de 10)	< 10 min	0.2 – 1 s	12.5 – 2.5 Hz
« Very Fast »	30 – 400 (autour de 100)	0.1 – 1 min	0.03 – 0.2 s	83 – 12.5 Hz
« Ultra Fast »	400 – 4000 (autour de 1000)	< 0.1 min	0.005 – 0.03 s	500 – 83 Hz

Tableau 7. Classification des analyses rapides en GC à l'aide des différents paramètres définis dans la littérature (Mastovská and Lehotay, 2003)

L'intérêt d'une classification sur la base de la largeur des pics réside dans la définition des prérequis pour les analyses rapides (appareillage nécessaire dans son ensemble, vitesse d'injection, programme de température, fréquence d'acquisition du détecteur etc.) (Korytár et al., 2002).

Alors que la Fast GC possède une efficacité de séparation (nombre de plateaux théoriques) comparable (voire plus élevée) aux méthodes analytiques conventionnelles et peut être utilisée en routine pour l'analyse de mélanges complexes à l'aide d'équipements conventionnels modernes, la Very Fast GC n'est appliquée que pour les mélanges simples tandis que l'Ultra Fast GC n'est quasiment pas appliquée en raison de sa très faible efficacité (Matisová and Dömötörová, 2003). De plus, la Very Fast GC et l'Ultra Fast GC nécessitent des équipements spécialisés rendant leur application en pratique limitée (Mastovská and Lehotay, 2003). Les limites pratiques des analyses rapides seront discutées dans le paragraphe 3.5.

[19] Fréquence requise pour atteindre 5 points de données le long du pic chromatographique (2 fois la largeur de pic à mi-hauteur), nombre de points suffisants selon Mastovska et Lehotay (2003).

3.3 Approches possibles pour la mise en place

L'influence des principaux paramètres sur la vitesse de n'importe quelle méthode d'analyse (c'est-à-dire, sur le temps d'analyse au niveau de la technologie de séparation) s'exprime à l'aide de l'équation suivante qui définit le temps de rétention du dernier composé cible élué de la colonne (Sacks, 2004) :

$$t_R = N \frac{H}{u}(k' + 1) \qquad (1)$$

L représente la longueur de la colonne (cm), *u* la vitesse linéaire moyenne du gaz porteur (cm/s) et *k'* le facteur de rétention pour le composé cible. Le facteur de rétention *k'* est proportionnel au coefficient de partage K_D dans la mesure où il exprime le rapport de la quantité de soluté dans la phase de stationnaire à la quantité de soluté dans la phase mobile. Ce facteur, sans unité, peut être relié au temps de rétention. Il ne dépend ni du débit, ni de la longueur de la colonne. Pour une substance dont le temps de rétention est t_R, et celui d'une substance non retenue est t_m (temps mort), *k'* peut se calculer par l'expression:

$$k' = \frac{t_R - t_m}{t_m} \qquad (2)$$

Pour introduire l'équation du coefficient de partage, il convient de rappeler que les constituants d'un mélange se partagent entre deux phases non miscibles, l'une mobile et l'autre stationnaire, dans le cadre d'un phénomène dynamique où les molécules passent continuellement d'une phase à l'autre, créant un état d'équilibre entre la phase mobile et la phase stationnaire pour un constituant particulier. À ce moment-là, le rapport des concentrations est égal au rapport des répartitions dans les deux phases ou coefficient de partage K_D.

$$K_D = \frac{C_s}{C_m} \qquad (3)$$

C_s et C_m représentent la concentration du composé dans la phase stationnaire et la phase mobile, respectivement.

K_D peut s'exprimer en fonction des paramètres chromatographiques, ainsi :

$$K_D = k'\beta = k'\frac{r}{2\,d_f} = k'\frac{d_c}{4\,d_f} \qquad (4)$$

Sachant que la hauteur équivalente à un plateau théorique (H) est égale à la longueur de la colonne divisée par le nombre de plateaux de la colonne (N), L peut être remplacée dans l'équation (1) et il en découle l'équation (5) suivante :

$$t_R = N\,\frac{H}{u}(k'+1) \qquad (5)$$

A noter que les facteurs de rétention sont reliés à la température de la colonne T_c par

$$\ln k' = \frac{A}{T_c} + B \qquad (6)$$

où A et B représentent des constantes uniques pour chacun des composés ainsi que pour chaque type de phase stationnaire et rapport de phase volumique.

Ainsi, l'équation (1) démontre que la vitesse d'une méthode analytique peut être augmentée de plusieurs manières qui vont être discutées à présent. Les paramètres de la colonne, modifiés lors de l'implémentation de méthodes analytiques en Fast GC, correspondent à la longueur de la colonne (L), au type de phase stationnaire, à la température ou au programme de température, ou au flux du gaz porteur qui traverse la colonne (u). Finalement, comme l'illustre l'équation (4), le diamètre interne (d_c) et l'épaisseur de film (d_f) influencent k', et leur modification représente donc une approche envisageable pour diminuer le temps d'une analyse.

Agir sur ces paramètres conduit à deux situations différentes, la première consistant en une analyse rapide avec une **diminution de la résolution** à une valeur suffisante, la seconde consistant en une analyse rapide **sans perte de résolution** (Korytár et al., 2002).

La résolution R_S s'exprime par

$$R_S = \frac{2\Delta t'_R}{w_1 + w_2} \qquad (7)$$

Avec w_1 et w_2 les largeurs de pic respectives à la base et $\Delta t'_R = t'_{R2} - t'_{R1}$ (t'_R étant le temps de rétention réduit). Une manière commune d'écrire la résolution est

$$R_S = \frac{1}{4}\sqrt{N}\left(\frac{\alpha - 1}{\alpha}\right)\left(\frac{k'}{1 + k'}\right) \qquad (8)$$

Où N et k' font référence au dernier composé élué de la paire. α représente la sélectivité, sans dimension, égal au rapport des facteurs de rétention de deux solutés dont on veut réaliser la séparation :

$$\alpha = \frac{k'_2}{k'_1} = \frac{(t_{R2} - t_m)}{(t_{R1} - t_m)} \qquad (9)$$

Où k'_1, t_{R1} et k'_2, t_{R2} représentent les facteurs de rétention et les temps de rétention des composés 1 et 2, respectivement.

Etant donné que α et k' sont constants pour une colonne donnée (sous conditions isothermes), la résolution sera dépendante du nombre de plateaux théoriques N. Le terme k' augmente généralement avec une diminution de température de même que α mais dans une moindre mesure. Ainsi, l'on trouve qu'à faibles températures, moins de plateaux théoriques ou une colonne plus courte sont requis pour la même séparation (Grob and Barry, 2004).

Lors d'une perte acceptable de résolution, les différentes approches pour réaliser une analyse rapide, pour les analyses avec programme de température, consistent en l'utilisation de colonnes plus courtes, l'augmentation de la vitesse linéaire du gaz porteur, l'implémentation de températures de colonnes plus élevées, l'utilisation de films plus fins, la mise en place de rampes de température plus élevées, la diminution de la solubilité des solutés dans la phase stationnaire ou l'introduction de phases stationnaires sélectives.

Le maintien de la résolution lors du passage à une méthode analytique rapide requiert quant à elle l'utilisation de colonnes de diamètre interne plus faible. Utiliser l'hydrogène comme gaz vecteur ou utiliser un vide en fin de colonne (détecteur MS) diminuera aussi le temps d'analyse tout en maintenant la résolution originelle. Une revue exhaustive des approches possibles est présente dans la littérature (Klee and Blumberg, 2002; Korytár et al., 2002; Mastovská and Lehotay, 2003; Sacks, 2004) et les plus courantes parmi ces dernières vont à présent être discutées.

Agir sur la **colonne** : **longueur** et **diamètre interne** plus faibles

En combinaison avec d'autres approches, la plupart des méthodes rapides utilisent des colonnes plus courtes. La diminution de L réduit de manière proportionnelle le nombre de plateaux théoriques (N) mais diminue R_S dans une moindre mesure car L est proportionnelle à $\sqrt{R_S}$. Pour obtenir une efficacité similaire voire meilleure en comparaison aux colonnes capillaires conventionnelles, des colonnes de diamètres internes plus faibles, dites microbore[20], sont utilisées car elles possèdent plus de plateaux par mètre (donc une valeur de H plus faible) (cf. Tableau 8). Il est donc possible d'utiliser des colonnes plus courtes pour effectuer la même séparation. En comparaison aux colonnes de diamètres internes plus grands, il est possible de travailler à des vitesses linéaires de gaz porteur plus élevées avec moins de perte dans l'efficacité de séparation en utilisant des colonnes avec un diamètre interne réduit.

Pour illustrer les bénéfices des colonnes plus étroites pour les analyses rapides, Klee et Blumberg ont calculé la largeur de pic en fonction du diamètre interne illustré dans le Tableau 8 (Klee and Blumberg, 2002). D'après leurs résultats, passer d'une colonne de 530 à 100 µm de diamètre interne peut générer approximativement une analyse 9 fois plus rapide à la même résolution et capacité de pic n_p (définie comme étant le nombre de pics parfaitement séparés qui vont s'ajuster dans un chromatogramme avec une résolution R_S spécifiée)[21].

[20] On parle de colonnes megabore, wide bore, narrow bore, microbore ou sub-microbore lorsque le diamètre interne d_c est respectivement supérieur à 0.5, compris entre 0.3 et 0.5, entre 0.2 et 0.3, entre 0.1 et 0.2 et inférieure à 0.1 mm (Mastovská and Lehotay, 2003).

[21] $n_p = 1 + \frac{\sqrt{L/H}}{4R_S} \ln\left(\frac{t_{R1}}{t_m}\right)$

Diamètre interne (µm)	Efficacité[1] (N/m)	Largeurs typiques de pics[2] (s)	Vitesse relative[2]	Capacité d'échantillon relative[2]	Pression à l'injecteur[2] (psi)
530	2060	2.6	1	100	6.7
320	3660	1.2	2.18	22	14.9
250	4630	0.85	3.05	10.5	21.3
200	5830	0.64	4.06	5.4	28.9
100	11580	0.29	8.97	0.67	68.7
50	23160	0.14	18.5	0.084	150.2

[1] : (Wool and Decker, 2002)

[2] : (Klee and Blumberg, 2002)

Tableau 8. Paramètres théoriques et empiriques des colonnes capillaires en fonction du diamètre interne. Tableau élaboré à partir des références précitées.

Comme le montre le tableau précédent (cf. Tableau 8), les problèmes d'une telle approche résident dans la grande pression nécessaire le long de la colonne durant toute l'analyse ainsi que dans la diminution de la capacité d'échantillon de la colonne ($Q_S \propto d^3_c$). Il en résulte que la quantité d'échantillon atteignant la colonne doit être réduite proportionnellement à la diminution de phase stationnaire, dans le but de maintenir une chromatographie similaire à celle obtenue avec la méthode originale utilisant une colonne à plus grand diamètre interne (Klee and Blumberg, 2002). Par conséquent, il convient de se poser la question de la sensibilité des méthodes analytiques reposant sur une telle approche. Ses partisans maintiennent que les pics, plus fins, permettent d'obtenir un rapport signal sur bruit meilleur qui donne une faible LOD même si une quantité plus faible d'échantillon est introduite dans la colonne. Toutefois, l'effet des pics plus fins ne surmonte pas la quantité réduite injectée et la LOD dans son ensemble est plus élevée (Korytár et al., 2002). Finalement, l'injection doit être plus rapide et plus précise et la détection plus rapide pour distinguer les pics plus étroits (Mastovská and Lehotay, 2003). Pour un certain nombre de raisons (par exemple, la capacité d'échantillon, les pressions nécessaires etc.) les colonnes de diamètre interne de 100 µm semblent représenter la limite pour l'usage en routine des méthodes rapides (Klee and Blumberg, 2002; Dömötörová et al., 2006).

Agir sur le **facteur de rétention *k'***

De l'équation (4), il découle que

$$k' = 4K_D \frac{d_f}{d_c} \qquad (10)$$

Ainsi, outre la modification de la température de la colonne et la sélection d'une phase stationnaire différente, l'utilisation d'une colonne de diamètre interne plus élevé et/ou d'une épaisseur de film de la phase stationnaire plus fine permettent de diminuer *k'* (Korytár et al., 2002; Mastovská and Lehotay, 2003).

Des épaisseurs de film réduites diminuent le transfert de masse dans la phase stationnaire et augmentent la résolution. Une autre manière de procéder consiste à augmenter d_c et/ou diminuer d_f, si tous les autres paramètres restent les mêmes. Cela peut avoir plus d'impact sur l'accélération de la séparation que modifier la phase stationnaire. A noter que la diminution de d_f résulte également dans une diminution proportionnelle de la capacité d'échantillon (Q_S). A l'inverse, une Q_S plus élevée se produit par l'augmentation de d_c qui permet d'allonger le temps de vie de la colonne.

Augmenter les rampes de température est une manière simple d'augmenter la vitesse de la séparation, sans nécessairement utiliser d'instrumentation spécialisée. D'après la littérature, des rampes de températures plus élevées conduisent à des températures d'élution des composés plus élevées, une efficacité de séparation diminuée et des temps de refroidissement du four plus longs. Cependant, il semble que ce soit la température initiale qui ait plus d'influence que la température finale sur la durée de refroidissement (Mastovská and Lehotay, 2003).

L'augmentation des rampes de température peut s'effectuer avec les équipements conventionnels modernes ou avec les équipements spécialisés pour lesquels on parle alors de chauffage résistif.

Dans le premier cas, les vitesses des rampes de température sont de l'ordre de 1 à 2°C/s au maximum. Toutefois, les rampes possibles dépendent de l'intervalle de température défini dans le programme du four. 50°C/min correspond à la rampe de température qui est linéaire sur l'entier de l'intervalle de température (Sacks, 2004). La masse thermique du four limite les vitesses de chauffage et de refroidissement et donc la vitesse des méthodes analytiques rapides implémentées sur des équipements conventionnels. De même, les équipements avec des fours plus petits, en raison de difficultés pratiques, ne conviennent pas autant que le chauffage résistif pour la mise en place de méthodes analytiques rapides en GC (Mastovská and Lehotay, 2003).

Dans le second cas, le chauffage résistif s'effectue grâce à un courant électrique utilisé pour chauffer un conducteur qui est enroulé autour de la colonne et la température est contrôlée à l'aide de mesures de résistance. La masse thermique est alors réduite et les vitesses de chauffage et de refroidissement peuvent être très rapides. Un exemple d'un tel équipement est la technologie Agilent Low Thermal Mass (LTM). Des rampes de températures jusqu'à 1800°C/min peuvent être obtenues ouvrant la voie à des séquences d'analyses plus courtes en comparaison aux équipements conventionnels. Bien que la maintenance routinière ne soit pas aisée (colonne difficile d'accès car intégrée au module LTM), cette technologie présente deux avantages principaux en comparaison aux fours conventionnels. Ces derniers consistent en des refroidissements rapides qui permettent l'analyse d'un plus grand nombre d'échantillons pour une même durée et une très bonne répétabilité des temps de rétention et des hauteurs/aires de pics. Une répétabilité comparable avec les équipements conventionnels n'est possible que lorsque les rampes ont une vitesse maximale de 1°C/s (Mastovská and Lehotay, 2003).

Agir sur la **vitesse** et le **type** de **gaz porteur**

Le type de gaz porteur et sa vitesse influencent de manière importante la résolution et le temps de rétention. Pour chacun des composés un intervalle de vitesses optimales existe et il n'est pas le même pour chacun d'eux. La vitesse du gaz porteur est donc un compromis entre la résolution requise et le temps d'analyse voulu. Deux paramètres à considérer correspondent à la vitesse linéaire optimale (u_{opt}) et l' « optimum practical gas velocity » (OPGV). u_{opt} représente la vitesse du gaz pour un certain composé qui donne la meilleure efficacité de séparation (le plus faible H, H_{min}). Cette valeur correspond donc à la plus faible vitesse qui devrait toujours être utilisée pour n'importe quelle analyse. Des vitesses plus faibles que u_{opt} conduisent à une mauvaise résolution et un temps d'analyse augmenté. OPGV donne quant à elle l'efficacité maximale par unité de temps et est habituellement un facteur de 1.5 à 2 fois celui de u_{opt}. Cet intervalle d'OPGV est alors celui recommandé car même si une perte (légère) de résolution se produit en comparaison à une vitesse égale à u_{opt}, la réduction dans le temps d'analyse justifie cette faible perte d'efficacité (Wool and Decker, 2002). Selon la théorie, opérer à $u = 2*u_{opt}$ ne produit qu'une perte de 12% dans R_S (Blumberg et al., 1995). Si le détecteur peut gérer les flux plus élevés, alors la voie la plus directe pour augmenter u consiste à augmenter le flux (Mastovská and Lehotay, 2003). De plus, pour augmenter la valeur de u_{opt} il est possible aussi bien d'utiliser une colonne plus courte et plus étroite (diminution de L et d_c) que d'augmenter la diffusion du soluté dans la phase gazeuse en utilisant l'hydrogène (H_2) comme gaz porteur plutôt que l'hélium (He) et/ou diminuer la pression dans la colonne (nommé GC à faible pression). En effet, u_{opt} pour H_2 est à une vitesse linéaire plus élevée en comparaison à He et la courbe de Van Deemter est très plane quand on va vers des valeurs supérieures à u_{opt} indiquant une très faible perte de résolution lors d'une augmentation de la vitesse et donc une diminution dans le temps d'analyse. Au-delà des risques liés à sa haute inflammabilité, l'utilisation en pratique de H_2 est peu fréquente en raison d'inconvénients tels que des changements dans le spectre de masse et des pertes dans l'injecteur GC de certains analytes (en raison de réactions et/ou d'effets de surface)[22] (Mastovská and Lehotay, 2003). Le Tableau 9 résume les approches principales envisageables et leurs avantages/inconvénients pour réaliser des analyses rapides en GC (Korytár et al., 2002; Mastovská and Lehotay, 2003).

[22] Travaux non publiés mais énoncés dans la référence citée.

Terme	Action	Idée	Avantages	Inconvénients
L	↓	Utilisation d'une colonne plus courte (à noter qu'une diminution de L s'accompagne généralement d'une diminution de d_c)	Même capacité que les colonnes conventionnelles ; équipements conventionnels peuvent être utilisés	Peut diminuer la résolution ($R_S \propto \sqrt{L}$) Action irréversible
k	↓	1. Programme de température plus rapide	Gain en temps proportionnel à l'augmentation des rampes	Rampes élevées nécessitent une instrumentation spécialisée Résolution plus faible
		2. Augmenter d_c (pour L identique)	Peut accélérer l'analyse plus qu'en changeant la phase stationnaire Q_S plus élevée	
		3. Modification de la phase stationnaire pour améliorer la sélectivité	Gain significatif dans le temps d'élution possible	Sélection de la phase peut être fastidieuse
		4. Diminuer d_f : film de la phase stationnaire plus fin	Augmentation de la résolution	Diminution de la capacité de la colonne ($Q_S \propto d_f$)
u	1. ↑	Vitesse choisie plus élevée que la vitesse optimale u_{opt}	Pas de nouvel équipement nécessaire	Peut diminuer la résolution Vitesse maximale limitée par les régulateurs de pressions
	2. ≠	Meilleure diffusion du soluté dans la phase gazeuse : H_2 ou GC à faible pression	Même efficacité en moins de temps avec H_2 Coût	Sécurité Effets de surface Modification des spectres MS
d_c	↓	Diamètre interne de la colonne capillaire plus faible	u_{opt} plus élevée Meilleure efficacité (N/m) Résolution identique	Capacité de colonne plus faible ($Q_S \propto d^3_c$) Pressions à l'injecteur élevées et/ou rapport de split élevé ou volume d'injection plus faible

Tableau 9. Approches majeures pour l'implémentation de méthodes analytiques rapides en GC

Notons que des approches jouant sur la sélectivité de la technologie de séparation sont également possibles (Korytár et al., 2002).

3.4 Utilisation en systématique

Le concept de traduction de méthodes analytiques introduit par Klee et Blumberg (Blumberg and Klee, 1998; Klee and Blumberg, 2002) a contribué à l'implémentation en routine des méthodes analytiques rapides. A l'aide de ce concept, les conditions opératoires d'une méthode analytique GC conventionnelle peuvent être ajustées automatiquement pour obtenir une méthode analytique GC rapide, en utilisant un logiciel informatique dédié, disponible sur l'Internet. Comme on a pu le voir, plusieurs paramètres influencent la vitesse d'une méthode analytique. Selon le concept de traduction de méthodes, tous les changements dans les conditions chromatographiques peuvent être classifiés comme étant traduisibles et non traduisibles :

Classe	Paramètres
Traduisible	Longueur de colonne
	Diamètre interne de la colonne
	Epaisseur de film de la colonne
	Gaz porteur
	Vitesse de flux du gaz porteur
	Changements proportionnels dans les taux de chauffage
	Durée des paliers de température
	Détecteur travaillant à pression réduite (MS)
Non traduisible	Phase stationnaire
	Rapport de phase
	Température initiale et températures des paliers

Tableau 10. Classification des paramètres influençant la vitesse d'une analyse selon le concept de traduction des méthodes

En quelques mots, sur la base de cette classification, lors de la traduction de la méthode conventionnelle vers la méthode rapide, les paramètres non traduisibles ne doivent pas être différents (cf. Tableau 10). A l'inverse, les paramètres traduisibles peuvent être ajustés lors de la traduction de la méthode conventionnelle. Ce concept se base sur la notion de temps mort t_m qui peut être considéré comme l'unité de temps fondamental en chromatographie. Ce dernier peut être utilisé pour exprimer les composés reliés temporellement dans toute la métrique chromatographique.

Dans un programme de température normalisée, la durée de chaque palier de température ainsi que les rampes de température sont exprimées en unité t_m, mesurés à la même température. Le logiciel informatique calcule les traductions des programmes de température et des pressions à l'injecteur pour n'importe quel changement dans les dimensions de colonne, de type de gaz porteur, de condition pneumatique (pression ou flux constants) ou une combinaison des trois. Deux méthodes d'analyses mutuellement traduisibles fournissent le même ordre d'élution des composés.

3.5 Contraintes instrumentales

Pour rendre possible l'implémentation en routine des méthodes analytiques rapides, des innovations technologiques ont dû voir le jour : elles concernent entre autres le contrôle électronique de la pression de la phase mobile (EPC), les injecteurs capables d'injecter de très étroites bandes d'échantillon, les colonnes dites narrow- ou microbore, les nouvelles technologies pour le chauffage très rapide de colonne (chauffage direct ou résistif des colonnes pour des analyses Ultra Fast) ainsi que des détecteurs à haute fréquence FID ou MS pour la caractérisation de tous les composés d'un échantillon à haute vitesse (Sacks, 2004).

Sources **extracolonnes** de l'élargissement de la bande

En plus de l'élargissement de la bande du soluté qui se produit dans la colonne, l'élargissement de la bande provient d'autres parties. Majoritairement, l'élargissement de bande d'un pic inclut l'injecteur, les détecteurs, les connexions de l'injecteur à la colonne et de la colonne au détecteur ainsi que le système de traitement des données. En GC conventionnelle, l'élargissement de bande au niveau de la colonne est suffisamment grand pour masquer l'élargissement de bande extracolonne et ce dernier n'est donc pas pris en compte lors des calculs d'efficacité de la colonne. En Fast GC, la colonne étant généralement plus courte, et les vitesses de flux de gaz porteur plus élevées, l'élargissement se produisant dans l'injecteur et le détecteur prend une plus grande proportion dans l'élargissement total de la bande. Or, l'élargissement de la bande du pic influence l'efficacité de la colonne. Plus le pic est large, moins la colonne est efficace et moins nombreux sont les pics qui peuvent être séparés en un temps donné. Par conséquent, l'intérêt en Fast GC est de minimiser l'élargissement de la bande. L'instrumentation doit alors être adéquate, en particulier au niveau de l'injection et de la détection (Sacks, 2004).

Injection

L'injection est l'étape la plus critique en Fast GC car l'élargissement de bande extracolonne est le plus important à ce niveau. Pour minimiser la largeur de bande initiale, la bande de l'échantillon injecté doit être étroite en comparaison à l'élargissement total de la bande chromatographique. L'élargissement se produisant dans l'injecteur est dû à la vitesse d'injection, à la vitesse d'évaporation et la vitesse de transfert de l'échantillon depuis l'injecteur sur la colonne. En raison des flux élevés de gaz porteur dans l'injecteur, le split fournit l'injection la plus rapide de toutes les techniques, conduisant à la largeur de bande initiale la plus étroite dans la colonne, et représente la technique de choix pour les colonnes à faible diamètre et pour des séparations rapides. Le système d'injection le plus simple pour la Fast GC est ainsi l'injection en mode split car elle permet une injection rapide. Opérer à des hauts rapports de split permet d'obtenir des largeurs de bandes très faibles. L'inconvénient majeur d'une telle démarche est alors une mauvaise LOD.

Pour effectuer une injection rapide, les autosampler sont en général utilisés mais des travaux ont expérimenté d'autres types d'injecteur toujours dans l'idée d'obtenir une bande la plus étroite possible (Korytár et al., 2002; Sacks, 2004). Ainsi, des valves mécaniques, des systèmes de préconcentration du mélange gazeux avant son injection dans la colonne et d'autres technologies permettent d'atteindre cet objectif. A noter qu'un liner de diamètre adéquat doit être sélectionné, le liner pouvant contribuer de manière significative à l'élargissement total des pics.

Détecteurs

En Fast GC, plus de pics par unité de temps sont générés, sachant que ces pics sont plus étroits qu'en GC conventionnelle. Des détecteurs rapides ainsi que des systèmes de traitement de données adéquats sont donc requis. Les électromètres/amplificateurs et les systèmes d'acquisition de données pour la plupart des instruments conventionnels sont trop lents et contribuent de manière excessive à l'élargissement de bande extracolonne. Pour minimiser cela, les systèmes électromètres/amplificateurs devraient avoir des constantes de temps de l'ordre de 10 ms et les systèmes d'acquisition de données opérer avec des taux d'acquisition de l'ordre de 100 Hz. Ces besoins sont facilement accomplis avec les outils électroniques modernes. Pour des analyses en Very Fast GC, des constantes de temps plus petites et des taux d'acquisition plus grands pourraient être nécessaires.

Les détecteurs utilisés doivent posséder de hautes fréquences d'acquisition pour qu'un nombre de points de données suffisant le long du pic chromatographique pour qualifier et quantifier un composé soit obtenu. Le nombre de points minimum à atteindre le long d'un pic a été critiqué dans la littérature mais aucun consensus n'a été trouvé, un chiffre allant de 3 à 20 points ayant été proposé (Amirav and Jing, 1995; Dyson, 1999; van Deursen et al., 2000; Dallüge et al., 2002). Même à but quantitatif, de 3 à 4 points au-dessus de la ligne de base semble être adéquat (Mastovská and Lehotay, 2003).

Le Flame-Ionisation Detector ou FID représente le détecteur le plus populaire pour la Fast GC (Korytár et al., 2002). Un « makeup flow rate » élevé est recommandé pour éviter les problèmes entre la sortie de colonne et la flamme. Les FID modernes atteignent des taux d'acquisition de données jusqu'à 200 Hz, suffisant même pour la Very Fast GC. En plus des autres détecteurs disponibles (Flame-Photometric Detector ou FPD, Nitrogen-Phosphorous Detector ou NPD, ElectronCapture Detector ou ECD, Thermal Conductivity Detector ou TCD), un détecteur de choix est le détecteur MS.

La combinaison de l'analyse GC rapide à un MS est nécessaire pour identifier et confirmer correctement des composés, quantifier en sélectionnant des ions spécifiques et permettre la distinction de pics co-élués. Cette possibilité démontre que le MS permet de réduire le temps d'analyse, lors de sa combinaison avec les techniques dites de déconvolution (Mastovská and Lehotay, 2003; Sacks, 2004). Plusieurs analyseurs MS existent et leur mise en place a été critiquée dans la littérature (Mastovská and Lehotay, 2003). Chacun d'eux se définit par l'intervalle de masse couvert, la résolution massique, la vitesse d'acquisition spectrale et le coût. Des analyseurs de type scan tels que le Quadrupole et le Ion trap effectuent entre 15 et 30 spectres/s (en regard de l'intervalle de masses scanné) ce qui pourrait être trop faible pour les analyses rapides. Des analyseurs MS de type non-scan (c'est-à-dire, une acquisition constante des ions) représentent alors une alternative en raison de leurs fréquences d'acquisition particulièrement élevées. En particulier, le TOF-MS représente le détecteur idéal grâce à une fréquence d'acquisition spectrale de 500 spectres/s tout en fournissant une détection exacte des pics et des spectres de masse de haute qualité (paramètre essentiel en détection MS) alors que les largeurs de pics sont de l'ordre de la milliseconde (van Deursen et al., 2000). Toutefois, un frein à l'acquisition de TOF-MS est l'investissement financier qui n'est de loin pas négligeable (~ 500'000 CHF). Ainsi, une étude a montré que l'implémentation des méthodes d'analyse Fast GC pouvait se faire en utilisant des analyseurs Quadrupole qui, en raison de leur fiabilité et de leur coût raisonnable, représentent les détecteurs les plus utilisés en routine quels que soient les domaines analytiques (Kirchner et al., 2005). En jouant sur des paramètres tels que l'intervalle de masses analysé, le mode d'acquisition (SCAN ou SIM) et la fréquence d'acquisition (fréquence d'échantillonnage en SCAN ou dwell time en SIM) le Quadrupole peut permettre d'obtenir un nombre suffisant de points de données ainsi que des spectres de masse de bonne qualité lors d'analyses en Fast GC.

Partie B Problématique et Méthodologie

Dans cette partie, une fois abordées les problématiques et les approches actuelles existant pour le maintien d'une banque de données de profils chimiques dans une perspective longitudinale et transversale (Chapitre 4), l'accent est mis sur les bases théoriques développées et les outils statistiques implémentés dans le cadre de la méthodologie d'harmonisation des résultats analytiques, approche investiguée dans cette étude pour le maintien d'une banque de données par plusieurs méthodes analytiques (Chapitre 5 et Chapitre 6).

Chapitre 4 Tendances actuelles pour le maintien d'une banque de données de profils

4.1 Problématiques liées au maintien d'une banque de données

Tout laboratoire analytique implémentant une banque de données fera face à deux problématiques majeures (cf. Introduction, Figure 2). La première ne concerne que le laboratoire en question et définit l'alimentation de la banque de données sur le long terme. Dans cette étude, cette problématique a été décrite comme étant une **problématique intra laboratoire**. On parle de problématique en raison des questions qui se posent dès lors que la méthode analytique initiale, pourvoyeuse de la banque de données, vient à être modifiée (volonté du laboratoire de mettre en place une méthode analytique plus performante ou nécessité de modifier l'instrument analytique par exemple). Le laboratoire doit dès lors évaluer si une telle modification a un impact sur le maintien de la banque de données. Plus précisément, le laboratoire peut-il poursuivre l'approvisionnement de la banque de données avec les résultats analytiques de la méthode analytique modifiée sans les ajuster au préalable ? Il s'agit, en d'autres termes, de déterminer si les résultats analytiques issus de cette méthode analytique modifiée sont similaires à ceux provenant de la méthode analytique précédente. Cette notion de similarité est critique car elle implique la mise en place d'une méthodologie en mesure d'évaluer puis éventuellement d'ajuster la similarité de résultats issus de méthodes analytiques différentes, comme nous le verrons par la suite. Les questions qui se posent dans la première problématique se retrouvent dans la seconde qui concerne aussi bien l'approvisionnement de la banque de données par différentes méthodes analytiques que son partage avec un ou plusieurs autres laboratoires analytiques. Il s'agit par conséquent d'une **problématique inter laboratoires**. De nouveau, la question de la similarité des résultats provenant des différentes méthodes analytiques, dans l'optique de les centraliser dans une banque de données, se retrouve au premier plan.

Précisons toutefois que la complexité de chacune des problématiques n'est pas la même. Alors que dans le cadre de la problématique intra laboratoire il est avant tout question de la pérennité de la banque de données d'un seul et même laboratoire sur le long terme, la problématique inter laboratoires se consacre quant à elle au partage de ladite banque de données avec un certain nombre de laboratoires. Le nombre de laboratoires analytiques qu'il est possible d'associer ainsi que les variations dans la réponse analytique qui existent entre les laboratoires (variabilité inter laboratoires) sont des éléments à considérer, d'où une problématique plus complexe. Le point commun en revanche entre ces deux problématiques concerne la gestion par le(s) laboratoire(s) d'une banque de données constituée d'analyses effectuées à différentes périodes dans le temps et au sein de laquelle un certain nombre de profils chimiques y est enregistré.

Un cas de figure qui inclut chacune de ces deux problématiques mais qui s'en différencie quelque peu consiste en l'intégration dans une même banque de données des résultats contemporains provenant de plusieurs méthodes analytiques. En d'autres termes, il pourrait s'agir d'un laboratoire ayant plusieurs instruments analytiques sur lesquels une méthode analytique similaire serait implémentée[23]. Le maintien d'une telle banque de données implique des questionnements aussi bien sur son approvisionnement sur le long terme que sur son partage entre les méthodes analytiques en jeu. Il s'agit donc d'un cas de figure regroupant les problématiques intra laboratoire et inter laboratoires.

Sachant qu'un profil chimique donné dépend de la méthode analytique utilisée pour le définir, la similarité des résultats issus de méthodes analytiques différentes (mais qui définissent le profil chimique de manière similaire, c'est-à-dire, avec les mêmes composés cibles ou variables) dépend nécessairement de la similarité analytique[24] existant entre ces dernières. On peut ainsi en déduire que le triptyque méthode analytique – profil chimique – banque de données (et en particulier l'influence de la méthode analytique sur le profil chimique obtenu) se situe au cœur des deux problématiques précitées.

[23] Comme nous le verrons plus loin dans ce chapitre, il s'agit d'un cas particulier de l'implémentation d'une approche appliquée généralement dans le cadre de collaborations internationales : l'harmonisation des méthodes analytiques.

[24] Cette notion de similarité analytique découle de la définition d'une méthode d'analyse établie dans ce travail de recherche (cf. Chapitre 5).

Pour envisager une réponse aux questions qui résultent des problématiques intra- et inter laboratoires, l'étude de la situation actuelle à l'échelle nationale ou internationale pour l'échange de renseignements, sur la base des profils chimiques, s'impose.

4.2 Situation actuelle

La lutte contre le trafic de produits stupéfiants implique des investigations à large échelle qui nécessite l'échange de renseignements. Ainsi, Rand Europe, mandaté par la Cour Européenne, recommande de créer une banque de données à l'échelle européenne constituée des informations détaillées de saisies spécifiques effectuées en Europe[25]. Selon Rand Europe, la création d'une telle banque de données pourrait améliorer la compréhension des flux européens de produits stupéfiants et évaluer leurs évolutions en réponse aux initiatives politiques des Etats Membres, voire permettre la mise en évidence de liens entre les différentes affaires policières.

Dans ce cadre là, les banques de données de profils représentent un outil intéressant offrant des renseignements primordiaux pour proposer des stratégies de luttes contre le trafic national et international de produits stupéfiants (cf. Chapitre 1). De nos jours, différentes approches pour favoriser l'échange de renseignements sur la base des profils chimiques ont été entreprises afin de favoriser l'utilisation de ces informations (Esseiva et al., 2007).

La première démarche consiste à faire analyser les saisies de produits stupéfiants par un seul et même laboratoire analytique. Cette approche représente ainsi la centralisation des analyses. Cette stratégie se retrouve dans les pays structurés autour d'un laboratoire forensique central rassemblant toutes les saisies de produits stupéfiants. Des pays tels que l'Australie, les Etats-Unis, la Finlande ou la Hollande appliquent cette stratégie. En Suisse, une dizaine de laboratoires sont actifs dans les différents cantons et les diverses régions linguistiques et procèdent de manière indépendante les uns des autres aux analyses sur demande des autorités locales.

[25] Understanding illicit drug markets, supply-reduction efforts and drug-related crime in the European Union. Prepared for the European Commission, DG Justice, Freedom and Security. Rand Europe (Contract JLS/2008/C2/001).

Dans le cadre d'une approche de collaboration internationale systématique où plusieurs pays sont concernés, la centralisation des analyses n'est pas vraiment possible pour des raisons politiques évidentes. De plus, un système bâti autour d'un seul laboratoire implique l'envoi des spécimens à ce dernier d'où un certain nombre de barrières administratives et un coût non négligeable. Dans une moindre mesure, la mise en place d'un seul et même laboratoire pour l'analyse de l'ensemble des saisies policières pourrait s'avérer difficile en raison de l'infrastructure nécessaire, en plus d'une éventuelle surcharge de travail. Finalement, les contraintes temporelles sont particulièrement marquées dans une telle stratégie, en raison du temps nécessaire à l'envoi des spécimens venant s'ajouter à celui des analyses soulevant alors la question de l'intégrité et la stabilité de l'échantillon avant son analyse. Par conséquent, une transmission rapide des informations du profilage chimique aux autorités de lutte contre le trafic de produits stupéfiants pourrait être délicate à atteindre.

Ainsi, en particulier pour cette dernière raison, une seconde démarche consiste à utiliser une banque de données partagée et approvisionnée par différents laboratoires. Il convient de souligner qu'il n'est pas trivial de mettre en commun les résultats analytiques provenant de laboratoires différents pour les comparer entre eux. En effet actuellement, il n'est pas possible d'utiliser directement les résultats obtenus par un laboratoire forensique pour les comparer à ceux obtenus par un autre laboratoire analytique, conduisant ainsi à une perte potentielle d'informations concernant les réseaux de distribution des produits stupéfiants.

Lors de l'application d'une telle démarche, la question de cohérence entre les données doit être considérée et les profils chimiques issus des différents laboratoires doivent être similaires pour permettre leur intégration dans la banque de données. C'est là une différence considérable en comparaison aux traces ADN, par exemple. En effet, dans ce dernier domaine divers kits standardisés existent pour l'extraction du profil ADN d'une trace. Ces kits sont développés de telle sorte que lorsqu'ils sont appliqués par différents laboratoires, on peut s'attendre à ce que les résultats soient comparables. De plus, alors qu'un profil chimique se base sur la concentration absolue ou relative respective de chacun des composés cibles, le profil ADN correspond aux séquences non codantes du matériel génétique et la notion d'intensité des marqueurs n'y est pas présente. Par conséquent, la nature de ce type de trace ainsi que l'harmonisation dans la procédure d'extraction de profils facilitent grandement la construction et l'utilisation de banques de données ADN d'ampleur nationale ou internationale (les potentielles difficultés administratives mises à part).

En revanche, dans le cadre du profilage chimique, lorsque l'on envisage le partage d'une banque de données commune, il est crucial de définir la notion de similarité des résultats analytiques et de créer une méthodologie à même de l'estimer puis de l'ajuster le cas échéant. La finalité d'une telle démarche consiste dans la comparaison directe des profils chimiques des spécimens analysés dans les différents laboratoires évitant ainsi l'envoi des saisies à un laboratoire centralisé. Du gain significatif dans le temps de réponse il en découlerait une collaboration améliorée et potentiellement plus efficace entre les différents laboratoires dans le cadre de la lutte contre le trafic de produits stupéfiants.

Pour que les profils chimiques établis dans différents laboratoires d'analyse soient similaires, une harmonisation des méthodes analytiques est prônée de nos jours. Mais le problème majeur de la situation en Suisse et en Europe de manière générale concerne justement le manque d'harmonisation entre ces laboratoires d'analyse. Actuellement, chacun des laboratoires utilise sa propre méthodologie de profilage chimique de produit stupéfiant (pour autant qu'elle existe), impliquant des différences considérables dans toutes les phases du profilage chimique (c'est-à-dire, dans les procédures d'échantillonnage, de préparation des échantillons, d'établissement du profil chimique, dans la structure de la banque de données ainsi que dans la gestion des liens chimiques) et en particulier dans les paramètres des méthodes d'analyse, ces dernières étant les pourvoyeuses des banques de données. L'harmonisation n'est donc pas une phase aisée à mettre en place.

4.3 L'harmonisation des méthodes analytiques

Loin d'être triviale, cette stratégie recommande d'utiliser exactement les mêmes méthodes de préparation et d'analyse des échantillons[26] ainsi que la même méthodologie de traitement des données. Les méthodes analytiques mises en place chez les laboratoires participant à une telle démarche utilisent les mêmes marques et modèles d'appareillage, de consommables et sont définies par les mêmes paramètres d'analyse ou d'optimisation. Par conséquent, les méthodes analytiques sont considérées comme similaires, bien qu'elles soient implémentées sur des instruments analytiques différents.

[26] Comme énoncé dans la partie introductive de ce manuscrit, une méthode analytique séparative se définit par trois niveaux de paramètres analytiques (notés A, B et C dans cette recherche).

En dehors des projets de collaborations internationales, en particulier européennes, disposant de moyens conséquents, l'harmonisation des méthodes analytiques est peu ou pas appliquée pour le partage d'informations entre différents laboratoires.

4.3.a Présentation des projets majeurs de collaborations européennes

Ces projets de collaborations découlent de la stratégie européenne de lutte contre le trafic des produits stupéfiants introduite depuis plusieurs années.

En Novembre 1999, le Conseil de l'Europe a ratifié la Stratégie de l'Union Européenne (UE) sur les Drogues 2000-2004[27]. La Stratégie fut mise en pratique au travers du Plan d'Action de l'UE sur les stupéfiants 2000-2004[28]. Le Plan d'Action avait « une stratégie globale, multidisciplinaire et intégrée pour combattre les produits stupéfiants » avec l'accent mis sur la réduction de la demande, la réduction du trafic de produits stupéfiants, la coopération internationale entre les Etats Membres et avec les partenaires non-UE, et une coordination des diverses initiatives.

En particulier, le Conseil Justice et Affaires Intérieures a décidé en Septembre 2002 que les actions contre la production et le trafic des produits stupéfiants synthétiques devaient être considérées comme une priorité durant les deux dernières années du Plan d'Action de l'UE sur les stupéfiants 2000-2004. Ceci a résulté dans un Plan d'Action séparé sur les produits stupéfiants synthétiques[29]. Le Plan d'Action appelait à la mise en place d'une banque de données de profils des substances chimiques saisies et des actions contre les laboratoires illicites, les chimistes et les réseaux de distribution. De plus, une étude a été lancée sur le profilage des précurseurs chimiques.

[27] Council of the European Union, 13395/99 CORDROGUE 73: European Union Drugs Strategy 2000-2004. Brussels, 26 November 1999.

[28] Council of the European Union, 9183/00, CORDROGUE 32: EU Action Plan on Drugs 2000-2004. Brussels, 7 June 2000.

[29] Council of the European Union, 12452/2/02 CORDROGUE 81: Implementation plan on actions to be taken in regard to the supply of synthetic drugs. Brussels, 26 November 2002.

En 2005, la Stratégie et le Plan d'Action pour 2000-2004 ont été remplacés par la Stratégie de l'UE sur les Drogues 2005-2012[30] et son Plan d'Action 2005-2008. Cette stratégie définissait comme objectif principal l'obtention à la fin de l'année 2012 d'une amélioration mesurable dans l'efficacité, la compétence et la connaissance des actions de lutte par l'UE et ses Etats Membres visant la production, le trafic illicite des produits stupéfiants, les précurseurs, incluant la diversification des précurseurs de produits stupéfiants synthétiques importés dans l'UE.

Le Plan d'Action[31] contient plusieurs initiatives relatives aux produits stupéfiants synthétiques. Elles incluent les actions visant à réduire la fabrication et la distribution des produits stupéfiants synthétiques en développant des projets d'interventions et de renseignements. De plus, le Plan d'Action exige, parmi d'autres éléments, le développement d'une solution à long terme au niveau européen pour l'utilisation des résultats de profilage forensique des produits stupéfiants pour des applications policières tactiques et stratégiques ; la lutte contre l'activité criminelle dans le cadre des précurseurs chimiques et leur diversification ; le renforcement des frontières extérieures et les contrôles intra-Communautaire et un développement supplémentaire de la coopération entre les Etats Membres. A noter que le Plan d'Action est évalué annuellement, permettant l'ajustement des actions ou priorités.

Ainsi, depuis la fin des années 2000 et encore de nos jours, la littérature scientifique fait état de projets d'harmonisation des méthodes analytiques visant principalement les stupéfiants de type synthétique avec pour objectifs la caractérisation de spécimens provenant de saisies différentes ainsi que la mise en évidence de leurs voies de synthèse tels que les produits amphétaminiques (Ballany et al., 2001; Aalberg et al., 2005; Aalberg et al., 2005; Andersson et al., 2007; Andersson et al., 2007; Andersson et al., 2007; Lock et al., 2007; Dujourdy et al., 2008) et les ecstasys (MDMA) (Marquis et al., 2008; Weyermann et al., 2008). Le Tableau 11 et la Figure 20 présentent les projets européens qui se sont succédé.

[30] Council of the European Union, 15074/04 CORDROGUE 77, SAN 187, ENFOPOL 178 RELEX 564: EU Drugs Strategy (2005-2012). Brussels, 22 November 2004.

[31] Council of the European Union, 8652/1/05 REV 1 CORDROGUE 25, SAN 63, ENFOPOL 59 RELEX 240: EU Drugs Action Plan (2005-2008). Brussels, 9 May 2005.

En particulier, le projet EDPS (« European Drug Profiling System ») consiste en 3 axes de travail. Le premier représente le maintien et l'utilisation du système européen de profilage amphétaminique et le deuxième consiste dans la gestion d'un projet voué au développement de méthodes de profilage de l'ecstasy et d'un système de comparaison de pilules illicites. Le troisième concerne l'étude de faisabilité d'un système de profilage forensique de l'héroïne et de la cocaïne, à l'échelle européenne (The European Drug Profiling System (EDPS) / 2010-2013). Ce troisième axe sera présenté au §4.3.b.

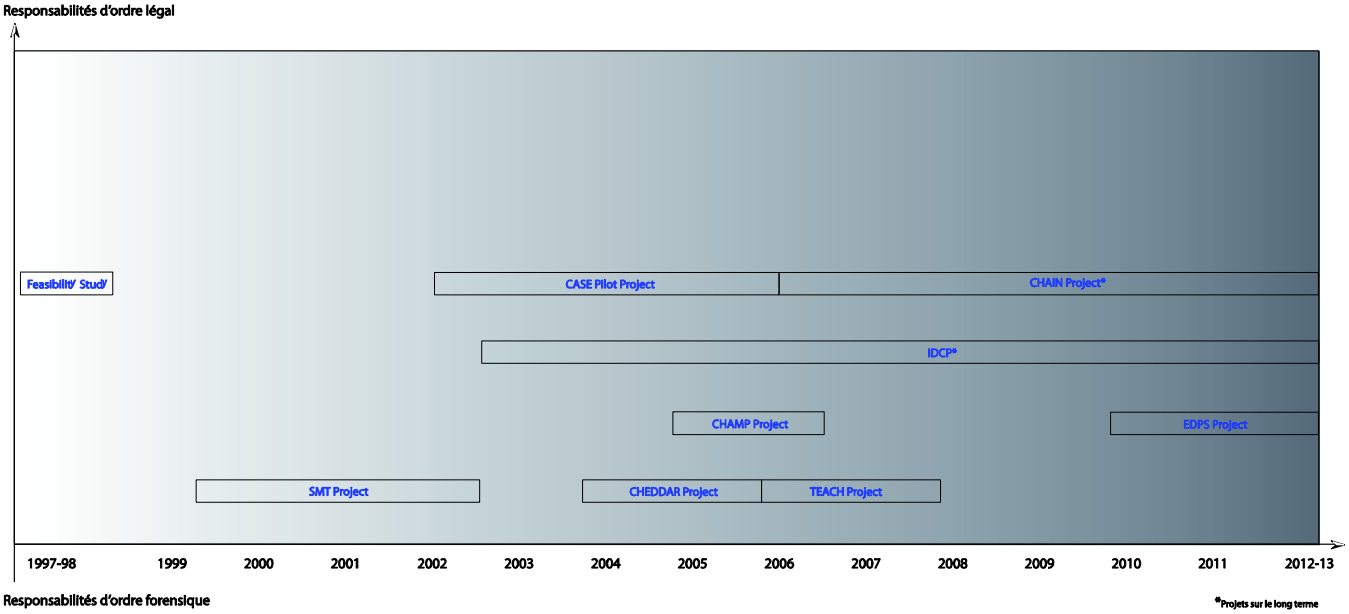


Figure 20. Coopération internationale dans le cadre du profilage de produits stupéfiants (cf. Tableau pour le détail des projets) (Alm, 2006)

Le système européen de profilage de l'amphétamine représente le produit de divers projets européens (CASE, CHEDDAR et CHAIN) et implique l'analyse chimique de produits amphétaminiques à travers l'Europe ainsi que l'insertion des résultats dans une banque de données commune pour y réaliser une comparaison des saisies (The European Drug Profiling System (EDPS) / 2010-2013). L'objectif du profilage chimique des produits stupéfiants synthétiques en Europe consiste à produire des renseignements concernant les procédés de fabrication illicite, par exemple les méthodes adoptées, les précurseurs chimiques employés, etc. et établir les connexions entre les spécimens qui peuvent être utilisés pour des applications tactiques et stratégiques.

Dénomination de l'initiative	Détails de l'initiative	Période
« Feasibility Study »	Projet de coopération entre des laboratoires forensiques et la police ainsi qu'EUROPOL.	1998
« SMT Project »	« Standards, Measurements and Testing » Projet forensique avec pour but de développer une méthode harmonisée pour le profilage des impuretés des saisies d'amphétamine.	1999-2002
« CASE Project »	« Comprehensive Actions against Synthetic drugs in Europe » Projet de coopération entre des laboratoires forensiques et la police ainsi qu'EUROPOL concernant le profilage d'amphétamine. Le projet CHAIN a succédé au projet CASE.	2002-2006
« IDCP* »	Initiative mondiale financée par la DEA avec pour but de promouvoir le développement de banques de données sur l'amphétamine, la méthamphétamine, la MDMA, l'héroïne et la cocaïne.	2002- ?
« CHEDDAR Project »	Projet pour le développement et l'implémentation d'une banque de données commune pour la mise en mémoire et l'évaluation des données provenant des profils des impuretés de l'amphétamine.	2003-2005
« CHAMP Project »	« Collaborative Harmonisation of Methods for Profiling of Amphetamine Type Stimulants » Projet pour le développement de méthodes harmonisées pour l'analyse des impuretés présentes dans les saisies de méthamphétamine et MDMA (ecstasy).	2004-2006
« TEACH Project »	Projet pour entraîner 4 laboratoires forensiques européens à l'utilisation de la méthodologie harmonisée de profilage de l'amphétamine.	2006-2007

Tableau 11. Récapitulatif des initiatives internationales dans le cadre du profilage de produits stupéfiants (* : projets sur le long terme) (Alm, 2006)

« CHAIN Project* »	« Collaborative Harmonised Amphetamine INitiative ». Projet dont l'objectif principal est d'établir un système européen durable de profilage de l'amphétamine (mise en place d'une organisation structurelle et promotion du profilage auprès des autorités légales).	2006- ?
« EDPS Project »	« European Drug Profiling System » Projet faisant suite au projet CHAMP et ayant pour objectifs principaux d'implémenter une méthode de profilage harmonisée de la MDMA dans les laboratoires partenaires ainsi que de promouvoir l'information du profilage forensique à but de renseignements auprès des autorités légales. Mise en place d'une banque de données européenne de profilage des stupéfiants (profilage physique et chimique).	2010-2013

Tableau 11 (suite). Récapitulatif des initiatives internationales dans le cadre du profilage de produits stupéfiants (* : projets sur le long terme) (Alm, 2006)

De tels projets ont permis aux laboratoires forensiques d'accomplir un certain nombre d'objectifs. D'une part, le développement de méthodes analytiques harmonisées pour l'analyse des « composés clés » dans les stupéfiants de type amphétaminique, la caractérisation des voies de synthèse utilisées et la détermination de liens entre des saisies différentes (essentiellement dans le cas de l'amphétamine). D'autre part, et c'est là l'accomplissement majeur de ces projets internationaux, un système d'échange des informations concernant le profilage chimique a été développé à l'aide de banques de données communes partagées et maintenues par les laboratoires de plusieurs pays (Alm, 2006).

4.3.b Etude de la faisabilité d'un système européen de profilage de l'héroïne et de la cocaïne

Une première étude relatée dans la littérature portant sur l'harmonisation des méthodes d'analyse concerne la mise en place d'une méthodologie de profilage chimique de l'héroïne – stupéfiant d'étude de ce travail – commune à différents laboratoires (Strömberg et al., 2000).

Le partage d'une banque de données par trois laboratoires dans le cadre d'une telle stratégie y est investigué avec des méthodes d'analyses en GC-FID. La reproductibilité de la méthode est jugée bonne au sein de chacun des laboratoires, mais pas suffisante entre les laboratoires. Selon les auteurs, les difficultés d'harmonisation de l'intégration des chromatogrammes FID et une mauvaise qualité chromatographique de certaines impuretés considérées comme essentielles pour la discrimination entre les spécimens en sont les raisons principales. Le fait que peu d'optimisation des méthodes de préparation et d'analyse des échantillons – étape nécessaire lors de l'application d'une telle démarche – ait été entreprise avant leurs utilisations par les trois laboratoires peut expliquer leurs difficultés. Les auteurs concluent d'après la méthodologie développée qu'une banque de données commune à plus d'un nombre très limité de laboratoires n'est pas réaliste et recommandent la création d'un laboratoire centralisé.

Selon les auteurs de l'étude de la faisabilité d'un système européen de profilage de la cocaïne et de l'héroïne, dans tous les secteurs d'analyse forensique, les méthodes qui peuvent être utilisées dans différents laboratoires tout en offrant les mêmes résultats sont préférables à celles nécessitant une centralisation des analyses (The European Drug Profiling System (EDPS) / 2010-2013), avec toutes les problématiques que cela entraîne (cf. §4.2). Toutefois, pour associer dans une banque de données commune les résultats provenant de différents laboratoires forensiques, un niveau d'exactitude et de cohérence sur le long terme des données est requis. Or, plus une méthode s'avère complexe plus ceci est difficile à atteindre. La complexité de l'analyse chimique effectuée dans le cadre du profilage chimique, couplée au besoin d'une cohérence des données sur le long terme, nécessite une méthodologie analytique reposant sur la centralisation des analyses ou l'harmonisation des méthodes comme le démontrent les collaborations internationales précédentes (The European Drug Profiling System (EDPS) / 2010-2013).

Par exemple, aux Etats-Unis et en Australie, la *Drug Enforcement Administration* (DEA) et l'*Australian Federal Police* (AFP) ont respectivement adopté une politique de centralisation où un seul laboratoire réalise le profilage chimique de l'ensemble des échantillons provenant de tout le pays. Dans les deux pays, une stratégie d'harmonisation a été mise en place dans la mesure où les mêmes méthodes analytiques sont utilisées sur exactement le même équipement analytique. Les deux pays partagent les résultats mais les données brutes restent séparées.

L'EDPS a adopté une politique qui se situe entre centralisation et harmonisation, où les laboratoires en Australie, Belgique, Grande-Bretagne, Finlande, France, Hollande, Suède et Suisse ont conjointement développé les méthodes, en utilisant un équipement analytique identique et en garantissant respectivement des standards d'assurance qualité de telle sorte que les données brutes puissent être entrées dans une seule banque de données pour l'estimation et l'interprétation de liens chimiques entre les échantillons. Un système européen de profilage des alcaloïdes nécessiterait certainement le même degré de centralisation (The European Drug Profiling System (EDPS) / 2010-2013).

Finalement, les auteurs soulignent la question de la centralisation ou de l'harmonisation qui se posera si un système européen de profilage chimique de l'héroïne et de la cocaïne voit le jour. Ils estiment qu'une harmonisation totale parmi les 27 Etats Membres n'est pas réaliste et serait coûteuse (The European Drug Profiling System (EDPS) / 2010-2013). Etant donné le besoin en cohérence sur le long terme des données et en prenant en compte le climat économique actuel, ils concluent qu'une option plus réaliste qu'une centralisation des analyses serait peut-être l'extraction du plus d'informations possibles par chacun des laboratoires en Europe, après un certain degré d'harmonisation de la méthodologie analytique.

4.3.c Une remise en cause de l'harmonisation des méthodes analytiques

Un travail conséquent et particulièrement long doit être entrepris pour obtenir des méthodes analytiques similaires en vue d'atteindre une similarité des profils chimiques entre les différents laboratoires. Par exemple, dans le cadre des collaborations internationales, plusieurs paramètres analytiques sont investigués pour optimiser la méthode de profilage avant son utilisation tels que la préparation des échantillons, les techniques de détection ou l'intégration des aires de pics.

Bien qu'intéressante en terme de normalisation analytique (par exemple, utilisation en simultanée de paramètres analytiques identiques par plusieurs laboratoires), appliquer la même méthodologie analytique est en revanche peu réaliste en dehors de projets internationaux.

D'une part, il s'agit d'une phase intrinsèquement coûteuse, longue et contraignante à mettre en place (obtention des solvants et substances chimiques auprès des mêmes fournisseurs, acquisition des mêmes marques et modèles de verrerie, d'appareillage et de consommables, phases de validations des résultats entre les différents laboratoires etc.) et, d'autre part, la gamme de méthodes analytiques existantes est si large dans les laboratoires d'analyse que le maintien d'une banque de données par ces derniers n'en est que plus difficile. En effet, plusieurs méthodes analytiques existent pour l'analyse d'un même produit stupéfiant, impliquant alors pour les laboratoires participants un changement de leur méthode analytique, pourtant utilisée en systématique.

Cette stratégie devient également problématique lors du maintien sur le long terme d'une banque de données par un seul et même laboratoire d'analyse pour deux raisons principales. Elle signifie d'une part pour le laboratoire une inertie analytique et d'autre part une perte potentielle des informations de la banque de données. Effectivement, selon cette approche, il n'est plus possible d'approvisionner la banque de données après la modification de la méthode analytique, et ce en raison des réponses analytiques différentes obtenues. La mise en place d'une nouvelle banque de données maintenue par la nouvelle méthode analytique est alors préconisée. Par conséquent, à moins d'un long processus de ré-analyse des spécimens enregistrés dans la banque de données avec la nouvelle méthode analytique – pour autant que ces derniers soient encore disponibles – la « mémoire » établie durant plusieurs années serait mise à zéro.

En conclusion, l'harmonisation des méthodes analytiques offre des résultats concluants mais, d'après la littérature consultée, uniquement lors de projets européens de grande envergure ayant investigué des stupéfiants synthétiques. L'étude de Strömberg et al. (2000) démontre qu'il est difficile d'atteindre une même réussite lors de projets de moindre ampleur et sans un travail en laboratoire intensif. Cette stratégie d'harmonisation n'est donc pas sans inconvénients et il existe ainsi un intérêt à étudier une alternative à cette approche en mesure de surmonter les inconvénients mentionnés.

Une autre stratégie consiste alors à envisager la possibilité d'approvisionner une même banque de données à l'aide de résultats issus de méthodes analytiques différentes. Cette approche définie comme l'harmonisation des résultats analytiques est investiguée dans cette recherche dans le cadre du profilage chimique des composés majeurs de l'héroïne.

4.4 L'harmonisation des résultats analytiques

Cette stratégie postule que si l'on compare des résultats analytiques issus de méthodes répétables et reproductibles, alors le recours à l'harmonisation des méthodes analytiques n'est pas nécessairement requis. Il est en effet considéré que le maintien d'une banque de données à l'aide de résultats issus de méthodes analytiques différentes est, dans certains cas, concevable et applicable.

La mise en place d'une telle démarche signifie concrètement que la méthode analytique utilisée pour l'approvisionnement d'une banque de données pourrait être modifiée et, malgré tout, encore utilisée pour maintenir cette dernière. Cette étude s'inscrivant dans le cadre du profilage chimique de produits stupéfiants et en particulier l'héroïne, les résultats analytiques consistent ainsi en des profils chimiques de spécimens d'héroïne, les profils chimiques étant définis par les réponses analytiques obtenues pour chacun des 6 composés cibles. L'harmonisation des résultats analytiques part du principe que même si les variables définissant le profil chimique d'un spécimen sont analysées avec des méthodes analytiques différentes, les données résultantes pourraient être rendues similaires et, par conséquent, pourraient être stockées au sein d'une même banque de données dans l'optique de déterminer des liens potentiels entre les spécimens (cf. Figure 21).

Dans le cadre d'une telle approche d'harmonisation, pour assurer le maintien d'une même banque de données, une méthodologie à même d'estimer puis éventuellement d'optimiser la similarité des résultats analytiques issus des différentes méthodes d'analyse est proposée.

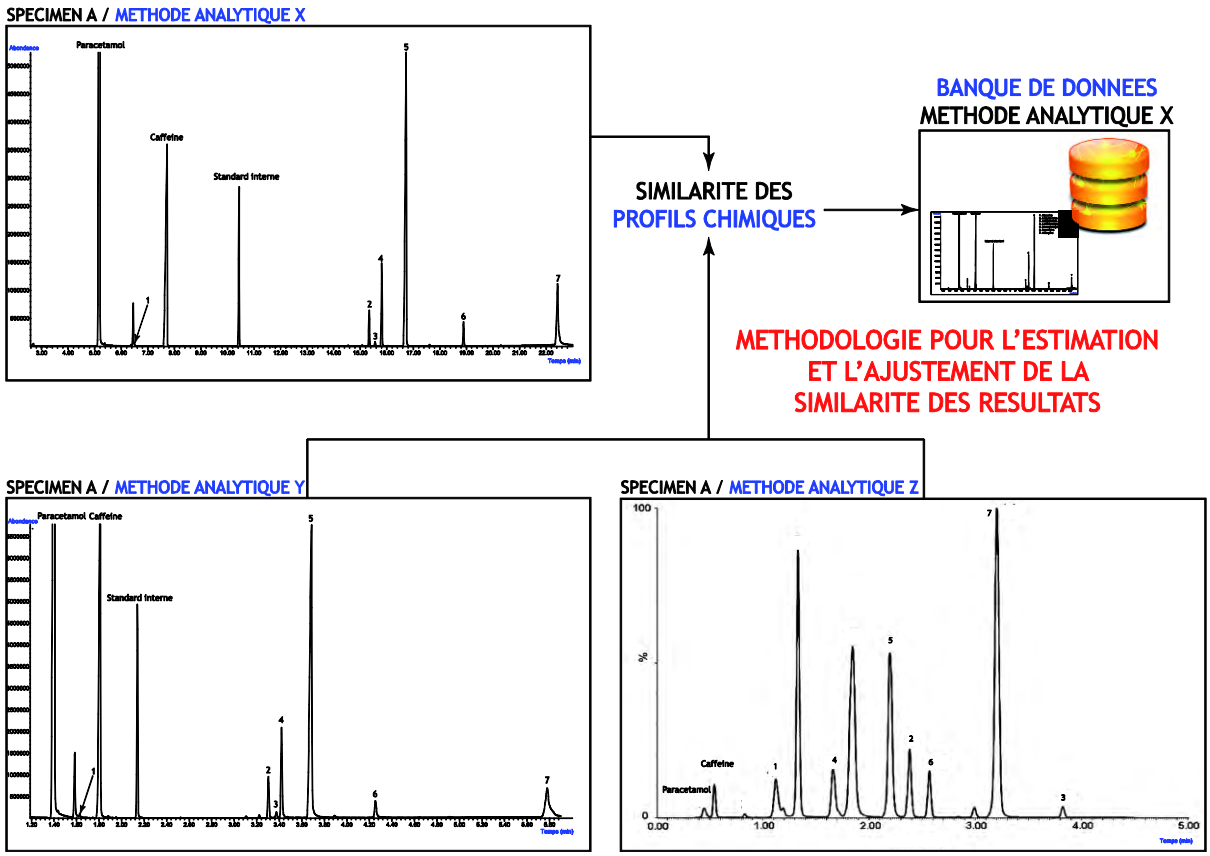


Figure 21. Représentation schématique de la méthodologie d'harmonisation des résultats analytiques

4.4.a Méthodologies d'ajustement

Dans la méthodologie d'harmonisation des résultats analytiques, on distingue en général deux procédés lors de la construction de la méthodologie d'ajustement des résultats analytiques, qui résultent de la manière dont le profil chimique des composés majeurs se définit. Ce dernier peut reposer sur la **quantification** (concentration réelle) ou bien sur la **semi-quantification** (concentration relative) des composés cibles du produit stupéfiant.

Dans le premier cas de figure, si un standard de référence pur pour chacun des composés qui constitue le profil chimique est disponible, la méthodologie d'ajustement peut se baser sur la comparaison des courbes de calibrations de chacun des composés établies avec les diverses méthodes analytiques (c'est-à-dire, sur la relation entre la concentration et la réponse analytique obtenue dans un intervalle de concentration donné pour un certain composé). Il s'agit d'une **méthodologie d'ajustement par calibration** de chacun des composés.

Lorsque la semi-quantification est réalisée ou bien lorsque les standards de référence ne sont pas disponibles (les standards de référence étant chers et difficiles à se procurer lorsqu'ils existent), il est alors nécessaire d'établir, à l'aide d'un échantillonnage, les relations mathématiques existantes, pour chaque composé du profil, entre les aires de pic (c'est-à-dire, les réponses analytiques) provenant des différentes méthodes d'analyse. Cette méthodologie peut être considérée comme une **méthodologie d'ajustement des réponses analytiques** obtenues pour chaque composé.

Dans les deux cas de figure, le profil chimique des produits stupéfiants étant défini par plusieurs composés, respectivement autant de courbes de calibrations ou de relations mathématiques que de composés doivent être établies.

Comme l'illustre la Figure 22 ci-dessous, la connaissance de la concentration réelle de chacune des variables du profilage – et non la concentration relative telle qu'avec la méthodologie d'ajustement des réponses analytiques – évite l'établissement des règles d'ajustements entre chacune des méthodes d'analyses. Toutefois, cet ajustement n'est pas aussi simple qu'il n'y paraît sur le schéma. En effet, tout calcul quantitatif s'accompagne d'une erreur sur la mesure de la concentration. Plusieurs composés constituant le profil chimique, les concentrations plus ou moins élevées de chacun d'eux avec leurs erreurs respectives doivent être combinées et ce sachant que ces concentrations sont calculées par différentes méthodes analytiques. L'intégration des résultats dans une seule et même banque de données dans l'optique d'y déterminer des liens potentiels représente ainsi un enjeu loin d'être trivial. En particulier, la procédure d'estimation de la similarité des profils sur la base des concentrations, en tenant compte des erreurs sur la mesure, devra être définie.

Bien que les règles d'ajustement entre les méthodes analytiques n'aient pas à être définies, il est donc a priori difficile de prétendre que l'ajustement des résultats analytiques soit facilité lorsque la méthodologie d'ajustement par quantification des composés est mise en place. En revanche, l'avantage incontestable de cette méthodologie consiste en l'obtention d'une banque de données stable à long terme et indépendante de la méthode analytique, la concentration réelle de chacun des composés du profil étant déterminée.

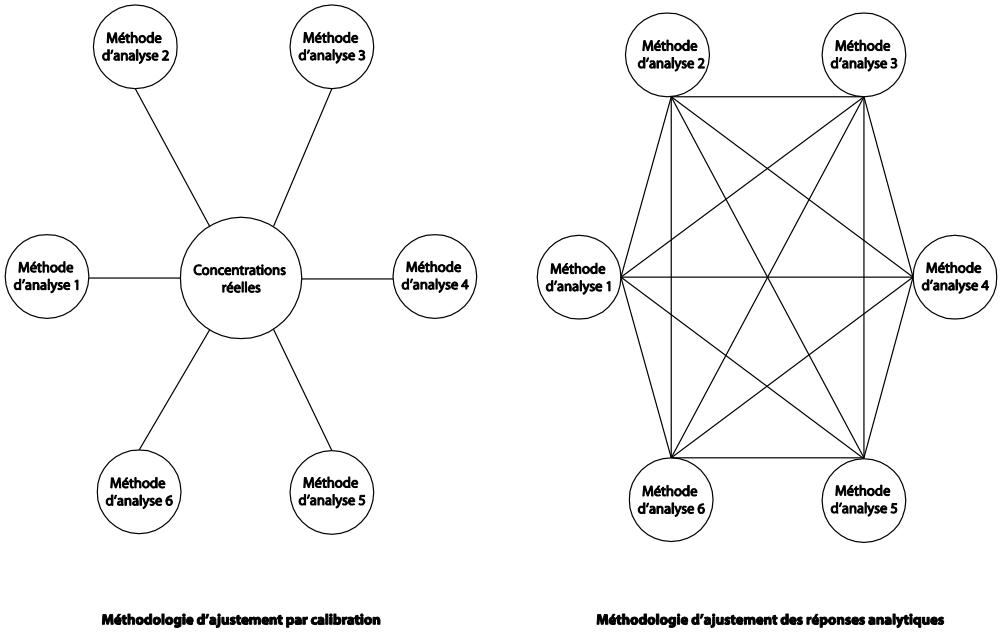


Figure 22. Représentation schématique de l'ajustement entre diverses méthodes analytiques ainsi que des relations entre ces dernières (représentées par des traits) en fonction de la méthodologie mise en place

Le but final des deux méthodologies consiste à assurer la similarité des résultats issus des diverses méthodes analytiques en ajustant dans un premier temps chacun des composés du spécimen puis dans un second temps en convertissant les profils chimiques correspondants. Grâce à ces méthodologies, les réponses analytiques obtenues avec la méthode d'analyse modifiée peuvent être utilisées pour alimenter la banque de données initiale. Il ne serait alors plus nécessaire, d'une part, de suivre les recommandations contraignantes de l'harmonisation des méthodes analytiques et, d'autre part, de créer une nouvelle banque de données avec la méthode d'analyse modifiée tel que recommandé par cette approche (avec en conséquence un long processus de ré-analyse des échantillons ou la mise à zéro de la mémoire de la banque de données). Ces méthodologies sont d'un intérêt majeur pour un laboratoire ayant besoin de renouveler son instrument analytique (nouvelle génération) ou souhaitant mettre en place de nouvelles méthodes d'analyse plus performantes telles que les méthodes analytiques rapides, par exemple.

Même si de telles méthodologies se focalisent sur un même objectif – l'approvisionnement d'une même banque de données à l'aide des résultats issus de diverses méthodes analytiques – elles sont ainsi conceptuellement différentes et leurs mises en pratique dépendent de la manière dont se détermine le profil chimique. Toutefois, ces deux méthodologies pourraient s'avérer complémentaires.

4.4.b Méthodologie d'ajustement mise en place dans ce travail

Comme cela a été précisé au point 4.4.a, la méthodologie d'harmonisation choisie dépend avant tout de la manière dont le profil chimique est établi en systématique par le laboratoire forensique et de la disponibilité des standards de référence des composés d'intérêt. Ce travail de recherche se déroule à l'IPS où la procédure implémentée pour réaliser le profilage chimique implique la semi-quantification des composés majeurs par GC-MS (c'est-à-dire que le profil chimique se compose des aires relatives de ces composés). Par conséquent, dans le cadre de ce travail de recherche, la **méthodologie d'ajustement des réponses analytiques** est investiguée.

4.4.c Etudes portant sur l'harmonisation des résultats analytiques

A la connaissance de l'auteur, une seule étude s'inscrit dans cette démarche et concerne l'approvisionnement d'une banque de données à l'aide de profils chimiques issus de méthodes d'analyses en GC-FID établies sur deux équipements de fabricants différents (Lociciro et al., 2007; Lociciro et al., 2008). Selon leurs résultats, les auteurs de cette étude jugent qu'une banque de données commune peut être approvisionnée par différents laboratoires et que les analyses de toutes les saisies n'ont pas à être effectuées dans un seul laboratoire centralisé. Cependant, selon eux, la possibilité de maintenir une banque de données avec plus de deux laboratoires serait beaucoup plus complexe, comme cela a été souligné dans une étude précédente (Strömberg et al., 2000). Finalement, les auteurs estiment que l'implémentation de l'harmonisation des méthodes analytiques dans chacun des laboratoires participants est requise pour réduire la variabilité entre les profils et par conséquent améliorer leur similarité et l'uniformité de la banque de données.

Malgré cette dernière conclusion, l'étude de Lociciro et al. (2008) démontre que des résultats similaires peuvent être obtenus avec des méthodes d'analyses différentes en termes de marque d'équipement.

En revanche, dans cette dernière étude, aucune autre méthode d'analyse n'a été étudiée et aucune méthodologie n'a été envisagée pour assurer la similarité des résultats analytiques et ainsi permettre le maintien d'une banque de données commune en se démarquant complètement de toute harmonisation des méthodes analytiques.

Chapitre 5 Partage d'une banque de données commune à différentes méthodes analytiques

Cette recherche propose une alternative à l'harmonisation des méthodes analytiques pour l'approvisionnement et le maintien d'une banque de données de profils par différents laboratoires. Une stratégie d'harmonisation des résultats analytiques est par conséquent étudiée. En particulier, cette étude propose une méthodologie pour assurer le maintien d'une banque de données avec des profils chimiques obtenus par semi-quantification et provenant de diverses méthodes analytiques. Cela conduit à la notion novatrice de *différence*, directement reliée à celle de *similarité*, entre les méthodes analytiques.

5.1 Différence analytique

5.1.a Définition

Les techniques chromatographiques implémentées pour le profilage chimique sont investiguées dans cette étude. Il s'agit de les définir de telle sorte que cette définition soit valable pour n'importe quelle méthode analytique. De plus, la définition de la *différence analytique entre méthodes* s'avère essentielle dans le contexte de cette recherche étant donné que la similarité de résultats provenant de méthodes analytiques différentes dépend nécessairement de la similarité intrinsèque entre elles. C'est pourquoi, définir dans quelle mesure des méthodes analytiques pourraient être considérées comme différentes les unes des autres, constitue une étape importante de cette étude.

Toute méthode analytique basée sur un principe de séparation se définit par trois niveaux de paramètres analytiques, nommés A, B et C dans cette étude (cf. Figure 23). Selon ces derniers, il est possible de différencier les méthodes analytiques les unes par rapport aux autres.

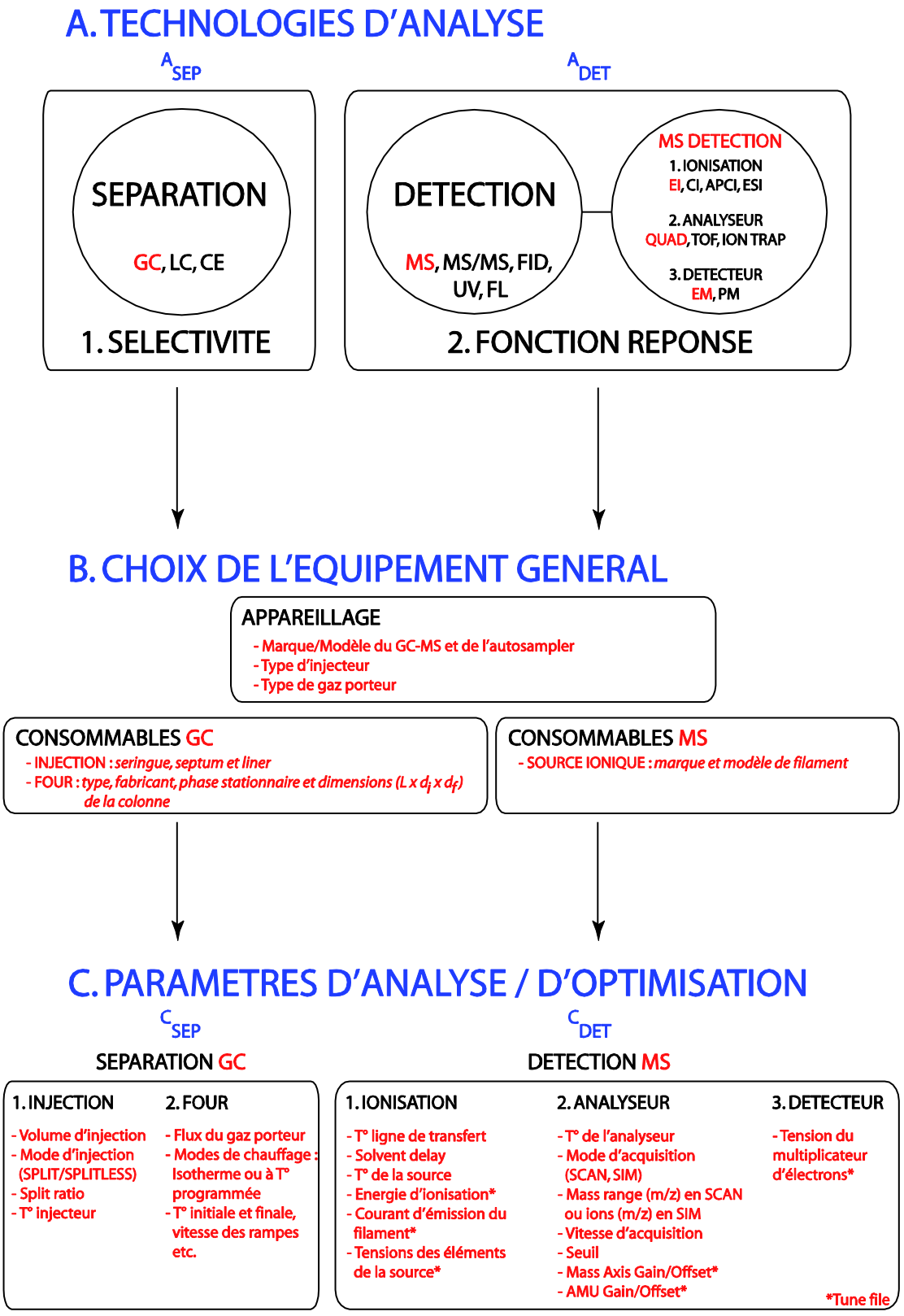


Figure 23. Illustration des paramètres analytiques d'une méthode d'analyse développée avec une technologie d'analyse, un équipement et des paramètres d'analyse donnés

(en rouge pour une méthode GC-MS)

Une méthode analytique séparative se compose de technologies d'analyse de séparation (A_{SEP}) et de détection (A_{DET}). Elle se définit également par l'équipement utilisé pour son développement et son utilisation en systématique (B) et par les paramètres d'analyse ou d'optimisation de la méthode analytique, ajustables d'une analyse à l'autre, et qui décrivent les technologies d'analyse de séparation (C_{SEP}) et de détection (C_{DET}).

Dans le cas où la détection utilisée consiste en une détection MS, alors une méthode analytique se caractérise également par les éléments définissant une telle détection soit la technologie d'ionisation[32], l'analyseur[33] et le détecteur[34] employés (cf. Figure 23, niveau de paramètre analytique A). En GC-MS, les paramètres d'injection (volume d'injection, mode d'injection) ainsi que le programme de température développé pour le four illustrent les paramètres d'analyse C_{SEP}. Les paramètres d'analyse C_{DET} se définissent quant à eux par les paramètres décrivant l'ionisation (énergie d'ionisation), l'analyseur (SCAN ou SIM) et le détecteur (tension appliquée au multiplicateur d'électrons). Il est à noter que pour certains auteurs, la spectrométrie de masse est considérée comme une méthode séparative.

D'une analyse à l'aide d'une méthode chromatographique séparative résulte un chromatogramme qui correspond à une représentation, en fonction du temps, de la séparation qui s'est produite au niveau de la technologie de séparation. Une série de pics se démarquant de la ligne de base est représentée dans le temps où chacun des pics représente la réponse du détecteur pour tous les constituants du mélange analysé (cf. Figure 24).

[32] Ionisation électronique (EI), chimique (CI), chimique à pression atmosphérique (APCI) ou par electrospray (ESI) par exemple.

[33] Quadrupole (QUAD), à temps de vol (TOF), ou à capture d'ions (ION TRAP) par exemple.

[34] Electro-multiplier (EM) ou photo-multiplier (PM) par exemple.

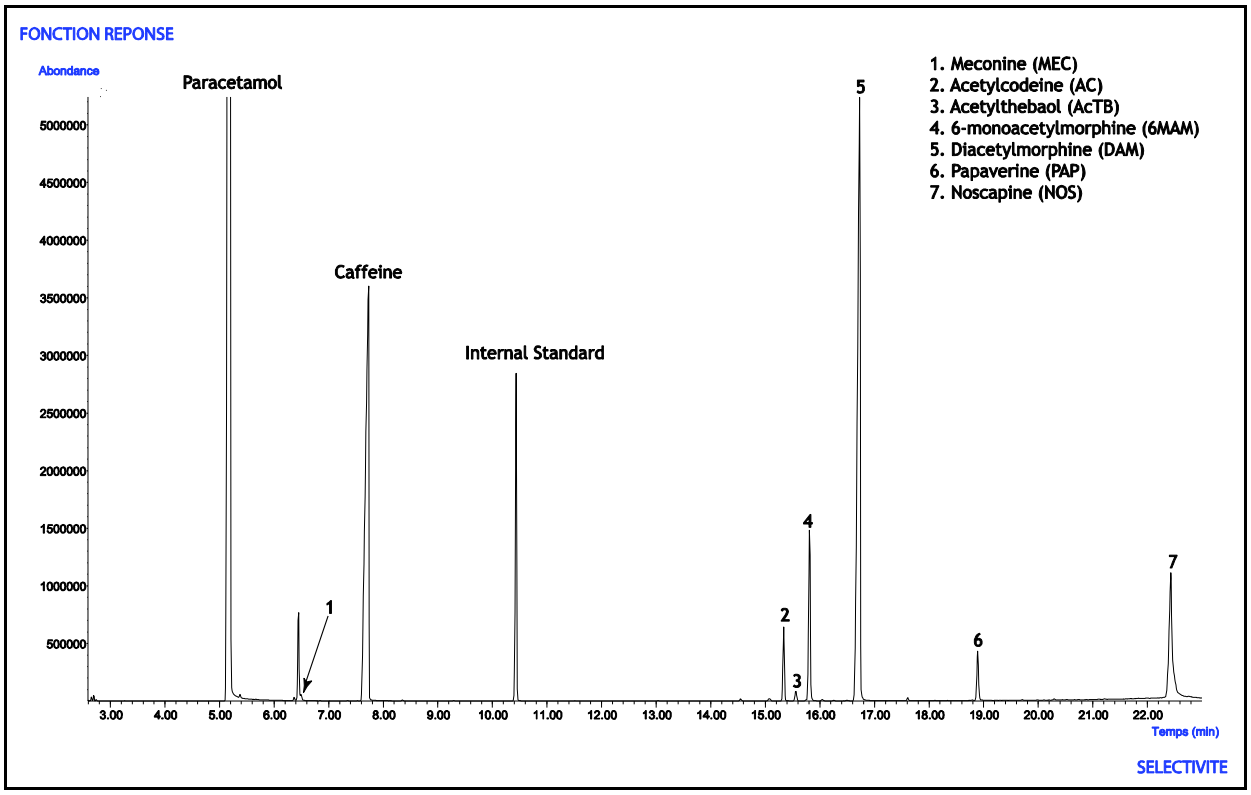


Figure 24. Chromatogramme d'un échantillon d'héroïne obtenu avec la méthode GC-MS de référence

5.1.b Méthodes analytiques similaires ou différentes ?

Les méthodes analytiques peuvent être différentes à trois niveaux de paramètres analytiques (cf. Figure 23). En conséquence, chaque laboratoire visant à comparer ses résultats avec ceux obtenus par d'autres laboratoires se retrouve confronté à plusieurs méthodes analytiques, plus ou moins différentes de sa propre méthode analytique de référence (c'est-à-dire, la méthode analytique validée au sein du laboratoire et utilisée quotidiennement). Avec le formalisme introduit dans cette étude, il devient possible de déterminer le degré de différence entre les méthodes analytiques investiguées voire de prédire la similarité de résultats analytiques obtenus avec des méthodes considérées.

Avec l'introduction d'une telle définition se pose alors la question de savoir si une différence dans un paramètre d'analyse donné, quel qu'il soit, implique d'avoir affaire à des méthodes analytiques dites différentes. A proprement parler, ceci est le cas dès qu'un paramètre analytique diffère. Toutefois, le propos n'est pas toujours aussi catégorique et il arrive que des méthodes analytiques différentes selon la définition établie dans cette étude soient considérées comme similaires.

Les postulats suivants n'ont pas été établis dans le cadre de ce travail de recherche mais découle de la doctrine d'harmonisation des méthodes analytiques, définie dans le §4.3, et sont par conséquent repris dans ce travail. Il est ainsi admis dans cette recherche que deux méthodes d'analyse sans distinction dans les paramètres analytiques mais implémentées sur des instruments analytiques physiquement différents sont similaires. De plus, en MS, même si les valeurs appliquées aux paramètres définissant le « tune file » (paramètre analytique C_{DET}, cf. Figure 23 et Chapitre 2) ne sont pas identiques d'une méthode analytique à l'autre, celles-ci sont également considérées comme étant similaires. En revanche, une différence entre deux méthodes analytiques au niveau de tout autre paramètre analytique (par exemple, en GC, dans les valeurs appliquées pour l'un des paramètres de C_{SEP}) implique que ces dernières sont différentes.

5.2 Les scénarios d'ajustement

La notion de scénarios d'ajustement trouve son origine dans l'existence, pour un laboratoire, de plusieurs *scénarios* ou situations où les méthodes analytiques rencontrées seraient plus ou moins différentes de sa méthode analytique de référence (que ce soit pour la problématique intra laboratoire ou la problématique inter laboratoires). Sachant qu'un *ajustement* des résultats analytiques pourrait alors être nécessaire pour assurer leur similarité, la notion de **scénarios d'ajustement** a été définie dans cette étude. Le recours à la méthodologie d'ajustement des réponses analytiques implique de considérer cette notion avant d'envisager le maintien d'une banque de données par plusieurs méthodes analytiques.

Définir la notion de scénarios d'ajustement et envisager leur classification selon les différences dans les méthodes d'analyses les caractérisant permet d'**évaluer l'influence de chacun des paramètres analytiques sur la similarité des résultats obtenus**. De plus, cela permet d'estimer quelles sont les différences analytiques significatives pour la similarité de résultats provenant de diverses méthodes. L'objectif principal de la définition des scénarios d'ajustement représente sa validité pour n'importe quel laboratoire, quel que soit son domaine analytique et quelle que soit sa méthode analytique de référence. Ainsi, une méthodologie générale d'ajustement des résultats analytiques pourra être mise en place.

Les scénarios d'ajustement ont été définis selon le postulat que, quelle que soit la sélectivité des méthodes d'analyse (ou l'ordre d'élution respectif des composés du spécimen analysé), la similarité des résultats analytiques n'est seulement influencée que par la similarité ou non dans les fonctions réponses respectives (c'est-à-dire, par la réponse analytique de chacun des composés qui est reliée à la technologie de détection de la méthode d'analyse A_{DET}). Par conséquent, bien qu'un grand nombre de scénarios d'ajustement puissent exister, ils peuvent se regrouper en deux classes majeures selon le degré de similarité dans la fonction réponse existant entre les méthodes d'analyse.

Suite à cette première classification, une seconde est réalisée dans chacune des deux classes d'après les différences dans les niveaux de paramètres analytiques B et C (cf. Figure 23). Sept sous-scénarios d'ajustement englobant l'éventail des méthodes analytiques de référence peuvent ainsi être envisagés et le Tableau 12 ci-dessous les récapitule.

D'après A_{DET}, les méthodes d'analyse définies dans le cadre des scénarios 1.1 à 1.5 (cf. Tableau 12) pourraient être fortement reliées à la méthode analytique de référence. A l'inverse, les méthodes d'analyse définies dans le cadre des scénarios 2.1 et 2.2 seraient analytiquement différentes. La classification subséquente réalisée selon les modifications des niveaux de paramètres analytiques B et C permet de classer les méthodes d'analyse par ordre décroissant de similarité analytique à la méthode de référence dans chacune des deux classes (c'est-à-dire, la méthode analytique définie dans le cadre du scénario 1.1 est analytiquement plus similaire à la méthode analytique de référence que ne l'est celle définie dans le scénario 1.2, et ainsi de suite). Selon un tel formalisme, il est attendu une meilleure similarité statistique pour des résultats obtenus avec des méthodes partageant des caractéristiques analytiques proches.

Fonction Réponse A_{DET}	Scénarios d'ajustement	Paramètre(s) analytique modifié(s)	Cas de figure	Instrument Analytique	Equipement
1. Similaire	1.1	C_{DET}	Contrôle qualité (répétabilité/reproductibilité des analyses, autre utilisateur)	Identique	Marque identique Modèle identique
	1.2	C_{DET}	Harmonisation des méthodes analytiques Plusieurs instruments analytiques	*Différent*	Marque identique Modèle identique
	1.3	B C_{DET}	Renouvellement d'équipement Evolution technologique	*Différent*	Marque identique Modèle *différent*
	1.4	B $C_{SEP/DET}$	Méthode analytique plus performante	Identique ou *Différent*	Marque identique Modèle identique ou *différent*
	1.5	B C_{DET}	Introduction d'un nouvel équipement	*Différent*	Marque *différente*
2. Différente	2.1	A_{DET} B $C_{SEP/DET}$	Méthode analytique plus performante	*Différent*	-
	2.2	$A_{SEP/DET}$ B $C_{SEP/DET}$	Méthode analytique plus performante	*Différent*	-

Tableau 12. Définition des scénarios d'ajustement auxquels n'importe quel laboratoire pourrait être confronté

Le scénario d'ajustement 1.1 fait référence à l'utilisation d'une méthode analytique aussi bien sur le long terme que par des opérateurs différents (étude de répétabilité et reproductibilité des analyses) (cf. Tableau 12). Ce scénario est la condition sine qua non à l'utilisation en systématique de toute méthode analytique. L'utilisation en routine d'une méthode analytique n'est envisageable que si et seulement si les résultats analytiques sont répétables et reproductibles. Par conséquent, ce scénario a été défini pour montrer qu'il s'agit là du premier ajustement auquel un laboratoire analytique est confronté : garantir une continuité des résultats analytiques jour après jour et démontrer le bon fonctionnement de la méthode analytique dans son ensemble. Comme cela est décrit dans le Chapitre 2, cette possibilité se retrouve tout particulièrement lorsque la méthode analytique se définit par une technologie d'analyse de détection MS, où les valeurs appliquées aux éléments du « tune file » (par exemple, dans les tensions appliquées aux éléments de la source ionique) peuvent évoluer au fur et à mesure de l'utilisation de l'appareillage. En particulier, ces modifications peuvent survenir suite au « tune » du MS (dont le but consiste en l'optimisation des valeurs attribuées aux paramètres de la source ionique, de l'analyseur de masse et du détecteur) ou être décidées par l'opérateur. Ce dernier cas de figure illustre la possibilité pour l'opérateur de modifier la méthode analytique au niveau des paramètres d'optimisation de la technologie d'analyse de détection C_{DET} pour garantir une réponse analytique similaire au fil du temps et assurer ainsi une reproductibilité des résultats analytiques. Précisons que la modification des paramètres du « tune » du MS n'entraîne pas une méthode analytique différente selon la méthodologie d'harmonisation des méthodes analytiques. Il est toutefois important de mentionner l'existence de telles différences entre les méthodes analytiques et envisager ainsi une éventuelle influence de leur part dans la similarité des résultats.

Il découle de ce qui précède que le scénario d'ajustement 1.1 ne sera pas étudié en tant que tel dans ce travail de recherche. Comme cela a été déjà précisé, ce scénario est en réalité intégré à chacun des autres scénarios d'ajustement vu qu'il s'agit de la condition à remplir pour l'utilisation sur le long terme d'une méthode analytique et renvoie ainsi au processus de validation analytique de toute méthode appliquée en systématique.

Le scénario d'ajustement 1.2 concerne la situation où, en comparaison à la méthode analytique de référence, les analyses sont réalisées sur un instrument analytique différent mais avec une méthode analytique similaire au niveau des technologies d'analyse de séparation et de détection, de la marque et du modèle de l'appareillage, et des paramètres d'optimisation de séparation et de détection (autres que ceux du « tune file », les valeurs attribuées à ces derniers dans le cas d'une détection MS pouvant évoluer au fur et à mesure de l'utilisation d'un même appareillage et entre appareils analytiques). Comme cela a déjà été décrit dans la partie introductive de ce travail, ce scénario fait donc typiquement référence à la méthodologie d'harmonisation des méthodes analytiques ou à la problématique d'un laboratoire possédant plusieurs instruments analytiques, quand dans les deux cas une méthode analytique similaire est implémentée sur plusieurs appareils.

Précisons que pour certains scénarios la méthode analytique, en plus d'être modifiée au niveau de la technologie d'analyse de détection, peut l'être au niveau de celle de séparation ($C_{SEP/DET}$, cf. scénarios d'ajustement 1.4, 2.1 et 2.2). Les modifications analytiques possibles dans le cadre de chacun des scénarios d'ajustement sont ainsi mentionnées dans le tableau ci-dessus.

Tandis que le scénario d'ajustement 1.3 concerne le changement de modèle d'appareillage, le 1.4 consiste en l'implémentation d'une méthode analytique plus performante (telle qu'une méthode analytique rapide). Ainsi, des modifications de l'équipement (colonne chromatographique) et des paramètres d'analyse C_{SEP} (flux du gaz porteur, vitesse des rampes de température) ou C_{DET} (vitesse d'acquisition) définissent ce scénario d'ajustement. A priori, la méthode analytique définie au sein de ce dernier présente une similarité plus élevée que celle mise en place dans le scénario 1.5 – qui se définit quant à elle par un changement de marque d'appareillage –, par rapport à une certaine méthode d'analyse de référence. En effet, le postulat de départ adopté lors de la définition des scénarios d'ajustement veut que les fonctions réponses respectives des méthodes analytiques influencent significativement la similarité des résultats obtenus. Or, dans le scénario 1.4 la technologie d'analyse A_{DET} ne change pas et uniquement la sélectivité se retrouve modifiée en comparaison à la méthode analytique de référence.

Sachant que la stratégie la plus répandue pour le développement d'une méthode d'analyse rapide consiste en l'utilisation de colonnes chromatographiques plus courtes et plus étroites, et que la différence majeure entre les méthodes analytiques se trouve être les durées d'élutions respectives des composés, l'ajustement des résultats obtenus pourrait se définir comme étant un *ajustement géométrique*. En revanche, le changement de marque d'appareillage tel que défini dans le cadre du scénario 1.5 pourrait avoir une influence importante sur la similarité des résultats analytiques. En effet, en raison des différences d'une marque à l'autre dans les composants des instruments analytiques respectifs utilisés pour la technologie A_{DET} ainsi que dans leurs réglages d'usine correspondants, une modification de la réponse analytique de chacun des composés pourrait se produire.

Les deux derniers scénarios 2.1 et 2.2 décrivent la situation d'un laboratoire souhaitant disposer en son sein d'une certaine polyvalence dans les techniques d'analyse de séparation et de détection et décidant ainsi d'implémenter des méthodes offrant une sélectivité et/ou une fonction réponse différentes en comparaison à sa méthode de référence. Selon les mêmes principes théoriques que précédemment, les méthodes d'analyse implémentées au sein des scénarios d'ajustement 2.1 et 2.2 peuvent être a priori déterminées comme étant plus ou moins analytiquement similaires à la méthode de référence. La méthode d'analyse développée au sein du scénario 2.2 présente des modifications aussi bien au niveau de la technologie d'analyse de séparation que de celle de détection. Ainsi, cette dernière présente une différence analytique plus prononcée par rapport à la méthode analytique de référence que celle implémentée dans le cadre du scénario 2.1. Ces deux méthodes d'analyses sont celles pour lesquelles assurer la similarité des résultats analytiques obtenus présente a priori le plus de difficultés, la différence analytique de celles-ci avec la méthode analytique de référence étant accentuée.

Au travers de telles définition et classification des scénarios d'ajustement, l'objectif de ce travail de recherche est de fournir à tout laboratoire analytique travaillant dans une optique de profilage chimique une méthodologie lui offrant un éventail de possibilités dans la composition de son parc analytique et dans l'implémentation des méthodes d'analyse.

A l'aide d'une telle méthodologie, il serait aussi bien possible pour un laboratoire d'utiliser plusieurs instruments analytiques (scénario 1.2), de rendre sa méthode d'analyse plus performante (scénario 1.4), d'implémenter sa méthode d'analyse vers un modèle ou une marque d'équipement différente (scénarios 1.3 et 1.5) voire de changer de technologies d'analyse de séparation (sélectivité) et/ou de détection (fonction réponse) (scénarios 2.1 et 2.2, respectivement) sans qu'il n'y ait d'influence à la fois sur le maintien de la banque de données et sur l'étape suivante de comparaisons des profils chimiques.

Chacun des scénarios d'ajustement doit être séparément investigué pour estimer la similarité des résultats analytiques obtenus dans chaque cas, toujours en comparaison à la méthode analytique de référence et ce à l'aide de la même méthodologie proposée au Chapitre 6. Avant de la décrire, il s'agit de précisément définir les scénarios d'ajustement qui vont être investigués dans cette étude pour démontrer les hypothèses de travail.

5.3 Hypothèses de travail

Hypothèse 1 : Le maintien d'une banque de données commune à diverses méthodes d'analyse dépend des caractéristiques analytiques de ces dernières

En comparaison à la méthode analytique de référence GC-MS, des méthodes d'analyse de plus en plus différentes pourraient être envisagées (cf. Tableau 12), ce qui implique que les relations statistiques existant entre les résultats analytiques seraient alors d'autant plus complexes.

Ainsi, cette hypothèse expose que l'on s'attend à obtenir des résultats jugés similaires s'ils proviennent de méthodes partageant des caractéristiques analytiques similaires (cf. Figure 23). Dans le cadre de cette hypothèse, il s'agira alors d'évaluer si la similarité analytique existant entre les méthodes considérées a effectivement une influence sur la similarité statistique évaluée entre les profils chimiques des spécimens correspondants.

Hypothèse 2 : L'utilisation de méthodes analytiques différentes n'est pas un frein au maintien d'une banque de données commune

Deux aspects majeurs doivent être investigués pour démontrer la faisabilité d'approvisionnement d'une banque de données par plusieurs méthodes analytiques, quelles qu'elles soient. Dans un premier temps, il s'agira de démontrer que malgré la mise en commun de profils chimiques obtenus avec des méthodes analytiques différentes, **une méthodologie de profilage efficace** existe, que les résultats du profilage soient utilisés à des fins de preuve dans une affaire particulière ou en tant que soutien dans le cadre d'une enquête policière (sous-hypothèse 2.1). Les éléments discutés au §1.4 sont alors particulièrement utiles et pertinents pour en discuter (cf. §6.4.f). Dans un second temps, il s'agira de démontrer **la conservation de la structure de l'information fournie par l'analyse de la banque de données** une fois que les résultats sont combinés (sous-hypothèse 2.2). Cette seconde sous-hypothèse fait référence à la notion de classe chimique (cf. Chapitre 1) : dans une banque de données, les spécimens présentant des profils chimiques similaires sont regroupés dans des classes chimiques selon la méthodologie statistique de classification mise en place par le laboratoire. Dès lors, lorsqu'un nouveau spécimen est analysé, il peut soit former une nouvelle classe chimique (si son profil n'a pas encore été observé) soit être attribué à une classe chimique préexistante (s'il présente un profil chimique identique) (Esseiva et al., 2005).

Si la méthodologie d'ajustement des données est efficace, il est attendu que :

- la classification des profils chimiques effectuée avec la méthode d'analyse de référence soit conservée au sein de la ou des autres méthodes analytiques (situation A, cf. Figure 25),
- les profils chimiques des mêmes spécimens analysés par des méthodes analytiques différentes, mis en commun dans la banque de données, aboutissent bien dans les mêmes classes (situation B, cf. Figure 25).

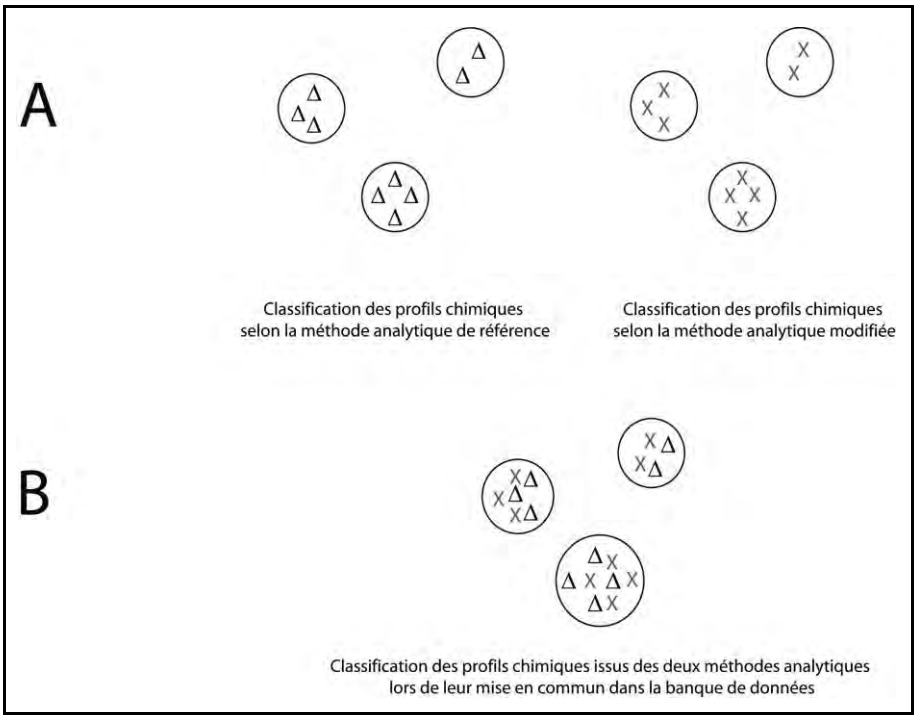


Figure 25. Objectifs poursuivis par la méthodologie d'ajustement quant à la classification des profils chimiques

Pour rappel, la délimitation des classes chimiques consiste en un processus dynamique et repose sur un ensemble de calculs statistiques. Il s'agira ainsi d'estimer dans quelle mesure la différence analytique entre les méthodes influence la classification des profils chimiques et quelles pourraient en être les conséquences. Lors de la mise en commun des résultats dans la banque de données, la potentielle différence statistique existant entre les profils chimiques d'un même spécimen obtenus par deux méthodes analytiques différentes pourrait affecter directement la délimitation des classes chimiques. Un tel cas de figure aurait alors un impact direct sur la qualité discriminatoire de la méthodologie statistique en raison de modifications possibles dans la distribution des populations d'intra- et d'inter variabilité (plus grande dispersion des valeurs d'intra variabilité, risque de superposition de classes chimiques ayant pour conséquence une augmentation des taux de faux par exemple).

Cette hypothèse investigue par conséquent si les règles d'ajustement établies pour l'ensemble des spécimens sont bel et bien adaptées à l'ajustement de chaque spécimen de l'échantillonnage choisi (paramètre méthodologique discuté au Chapitre 6) et ce dans le cadre de chacun des scénarios explicités au préalable. La stratégie d'ajustement que l'on entend développer dans ce travail de recherche doit être valable pour l'ensemble des spécimens de la banque de données et il faut alors s'assurer que malgré leurs différences de concentration et/ou de produits de coupage l'ajustement se fait de la même manière pour chacun d'eux.

En d'autres termes, la méthodologie d'ajustement doit être généralisable à l'ensemble des spécimens de la banque de données en faisant abstraction de l'intensité relative de chacun des constituants de leur profil – les 6 alcaloïdes majeurs dans notre cas – et de leur appartenance à telle ou telle classe chimique.

5.4 Scénarios d'ajustement investigués pour évaluer les hypothèses de travail

La problématique exposée au Chapitre 4 permet de souligner la difficulté de maintenir une banque de données avec différentes méthodes analytiques ainsi que les enjeux d'une telle démarche. Pour réaliser les objectifs de cette recherche, comme cela a été énoncé dans le paragraphe précédent, la notion de scénarios d'ajustement a été définie en fonction des différentes situations auxquelles un laboratoire d'analyse pourrait être confronté. Dans le cadre du profilage chimique de produits stupéfiants, une méthode d'analyse de choix est la GC-MS qui est mise en place en systématique à l'IPS. Elle est donc la méthode d'analyse dite de référence dans ce travail de recherche et les méthodes d'analyses investiguées dans cette étude sont plus ou moins différentes de la GC-MS pour chacun des scénarios investigués. L'évaluation des hypothèses mentionnées ci-dessus se fait à l'aide des cinq scénarios d'ajustement suivants discutés dans l'ordre de similarité analytique illustré par le Tableau 12.

5.4.a Scénario d'ajustement 1.2

Le scénario d'ajustement 1.2 fait référence à la mise en place de *la méthodologie d'harmonisation des méthodes analytiques* dans des situations aussi bien intra- qu'inter laboratoires. Le point commun de ces deux scénarios consiste en l'implémentation sur plusieurs instruments analytiques d'une méthode d'analyse similaire.

Théoriquement, il s'agit là du scénario d'ajustement le plus simple, en raison de la similarité analytique existante entre les méthodes d'analyses y implémentées (cf. Tableau 12). Toutefois, la littérature montre que le maintien d'une même banque de données dans le cadre de l'approche d'harmonisation des méthodes analytiques n'est de loin pas trivial (cf. Chapitre 4).

Il peut paraître surprenant que cette approche soit étudiée dans ce travail de recherche, ce dernier cherchant à montrer que l'harmonisation des résultats analytiques est une approche possible. Mais il s'agit justement ici d'évaluer, à l'aide de la méthodologie d'estimation de la similarité développée dans cette étude, dans quelle mesure les résultats sont similaires lorsque l'on harmonise les méthodes analytiques, cette démarche étant celle largement prônée dans la littérature. Il sera ainsi particulièrement intéressant d'utiliser l'estimation de similarité déterminée dans l'approche d'harmonisation des méthodes analytiques comme référentiel pour évaluer la similarité des résultats obtenus dans le cadre de l'approche d'harmonisation des résultats analytiques, et ce pour l'ensemble des scénarios d'ajustement définis.

Ce scénario d'ajustement présente également un intérêt par le simple fait qu'un laboratoire forensique pourrait effectivement vouloir implémenter en son sein des méthodes analytiques similaires sur un certain nombre d'instruments différents.

5.4.b Scénario d'ajustement 1.4

Ce scénario concerne l'ajustement de *méthodes analytiques plus performantes*, en conservant les technologies d'analyse de séparation et de détection de la méthode de référence (c'est-à-dire, en *Fast GC-MS*). Toutefois, l'ajustement des techniques analytiques rapides en GC-MS ne sera testé qu'une fois démontré la possibilité d'établir le profil chimique de cette manière. Les échantillons d'héroïne possèdent une matrice complexe constituée de plusieurs composés naturels, semi-synthétiques et de produits de coupage qu'il s'agit en effet de séparer, détecter et caractériser. Les composés cibles, c'est-à-dire définissant le profil chimique, devront en particulier présenter un nombre suffisant de points de données le long du pic chromatographique et une étude de la répétabilité et de la reproductibilité devra être également entreprise. Le second aspect concerne la phase d'ajustement proprement dite des résultats obtenus en Fast GC-MS vers la méthode de référence. La technologie d'analyse étant identique à celle de la méthode analytique de référence et les paramètres d'analyse étant modifiés, l'influence de ces derniers pourrait être étudiée.

Ainsi, il se pourrait qu'un compromis doive être trouvé entre la méthode d'analyse, sa durée (c'est-à-dire, la vitesse de l'analyse) et la similarité des profils chimiques. En d'autres termes, une analyse trop rapide pourrait être un frein à cette similarité et une certaine limite en temps d'analyse pourrait devoir être fixée.

5.4.c Scénario d'ajustement 1.5

Dans ce cas de figure, la technologie d'analyse est identique entre les méthodes analytiques c'est-à-dire que les techniques de séparation, d'ionisation et de détection sont les mêmes (cf. Tableau 12). De plus, *les paramètres d'analyse de la méthode d'analyse sont similaires mais l'appareillage est modifié*. La méthode analytique de référence choisie étant la GC-MS, les techniques de séparation et de détection sont donc respectivement la GC et la MS. Dans ce scénario, la méthode d'analyse est mise en place sur l'*appareillage d'un autre fabricant*. Il s'agit alors de déterminer dans quelle mesure la différence de marque d'appareillage a une influence sur la similarité des résultats analytiques et établir si les paramètres d'analyse influent de manière plus significative sur la similarité des méthodes analytiques en regard de l'appareillage utilisé. Une étude précédente sur l'approche inter laboratoires (Lociciro et al., 2007; Lociciro et al., 2008) a montré qu'un échange d'informations sur le profilage chimique de la cocaïne était possible malgré l'implémentation des méthodes analytiques sur deux appareillages de fabricants différents en GC-FID. Toutefois, l'aspect d'optimisation de la similarité des méthodes d'analyse n'a pas été évalué et il conviendra alors d'en étudier tous les paramètres.

Au terme de la procédure de comparaison, une meilleure similarité des résultats obtenus sur des appareillages de même fabricant est attendue (scénario 1.2). Par conséquent, une procédure d'optimisation de la similarité plus aisée que dans le cas où le fabricant diffère est prévisible. En effet, alors qu'aucune différence n'est attendue en ce qui concerne la performance chromatographique, une réponse analytique différente pour un même composé pourrait se produire en raison de la différence au niveau des technologies d'ionisation et de détection, en particulier dans les composants utilisés d'une marque à l'autre pour ces technologies et dans les réglages d'usine respectifs. La capacité à obtenir une réponse analytique similaire pour un même composé entre les instruments analytiques malgré de telles différences pourrait être un enjeu crucial pour la similarité des résultats.

5.4.d Scénarios d'ajustement 2.1 et 2.2

La grande différence avec le scénario précédent se situe au niveau des technologies d'analyse de séparation et/ou de détection entre les deux méthodes analytiques étudiées (cas d'ajustement n°2.1 et 2.2, cf. Tableau 12). L'étude de ce cas de figure se fait dans le cadre des analyses rapides. En effet, l'idée est de profiter au maximum des possibilités offertes par ces technologies pour l'analyse rapide des produits stupéfiants dans un but de profilage chimique. Bien entendu, plusieurs cas d'ajustement sont envisageables, mais dans ce travail de recherche les résultats obtenus par *Fast GC-FID* et *UHPLC-MS/MS* seront comparées à la GC-MS.

Le scénario d'ajustement 2.2 représente sans aucun doute le scénario pour lequel l'ajustement des données est le plus difficile en raison des différences analytiques existant entre la méthode de référence GC-MS et l'UHPLC-MS/MS. Bien que le partage d'une banque de données avec ces deux méthodes soit en pratique peu réaliste, l'intérêt d'investiguer ce scénario repose dans l'évaluation de l'impact sur la similarité des profils d'une différence dans les technologies d'analyse de séparation *et* de détection. Ensuite, cette recherche doit démontrer la capacité de la méthodologie développée à comparer des résultats issus de méthodes analytiques à ce point différentes. L'étude du scénario 2.2 constitue donc une base importante pour estimer des scénarios où les méthodes considérées seraient analytiquement plus similaires, raison pour laquelle il représente ainsi le premier scénario d'ajustement investigué dans cette étude (Debrus et al., 2010).

Chapitre 6 Méthodologie d'estimation et d'optimisation de la similarité des résultats analytiques

6.1 Objectifs

Dans le cadre de la méthodologie d'harmonisation des résultats analytiques abordée dans cette étude, et de la méthodologie d'ajustement des réponses analytiques qui y est investiguée, une méthodologie d'estimation et d'optimisation de la similarité des résultats issus de différentes méthodes est proposée. Bien que cette recherche concerne l'optimisation de la similarité de profils chimiques obtenus à l'aide de différentes méthodes analytiques, la méthodologie d'estimation et d'optimisation est développée de telle sorte qu'elle soit généralisable à n'importe quel autre produit d'intérêt, dont tout autre produit stupéfiant.

Ainsi, les objectifs de cette recherche sont de (Broséus et al., 2013) :

- trouver des **outils** à même **d'estimer** voire **d'optimiser la similarité** de résultats issus de méthodes analytiques différentes ;
- déterminer, finalement, s'il est possible d'introduire des profils chimiques obtenus avec une méthode analytique différente, **quelle qu'elle soit**, dans une **banque de données de référence** et ainsi en envisager son maintien ;
- permettre l'application de cette méthodologie **à d'autres domaines analytiques** confrontés au maintien d'une banque de données par différentes méthodes.

Comme le mentionne le 2ème point, l'idée centrale de cette recherche consiste à proposer une méthodologie permettant d'assurer que, quelles que soient les méthodes analytiques utilisées pour les obtenir (et donc quelle que soit la différence analytique entre elles), les résultats analytiques sont similaires. En effet, comme cela a été évoqué depuis le début de ce travail et en particulier au Chapitre 5 (avec la définition des notions de *différence analytique* et de *scénarios d'ajustement*), un grand nombre de méthodes analytiques différentes (que ce soit en termes de paramètres d'analyse de la méthode, de marque et modèle d'appareillage voire de technologies d'analyse) existent pour extraire les composés d'une même matrice d'intérêt, d'où l'importance de posséder une méthodologie qui s'affranchisse des méthodes analytiques considérées afin d'assurer que les résultats issus de ces dernières soient similaires. Le prochain paragraphe décrit le principe général de la méthodologie d'estimation et d'optimisation de la similarité proposée tandis que les suivants détaillent chacune des étapes la constituant.

6.2 Principe général

Une méthodologie en plusieurs étapes réalisée quelles que soient les méthodes d'analyse en jeu a été définie et rend possible aussi bien l'estimation que l'optimisation de la similarité existante entre des résultats analytiques provenant de méthodes différentes (cf. Figure 26). Deux outils principaux ont été sélectionnés pour cette estimation et sont inclus dans cette **analyse statistique** (cf. §6.4). Le premier combine une Analyse en Composantes Principales (ACP) puis une Classification Ascendante Hiérarchique (CAH) et consiste en deux étapes successives, la première à un niveau dit Global, la seconde à un niveau dit Local (nommée « ACP-CAH Globale et Locale » en français et « Global and Local PCA-HCA » en anglais) (Debrus et al., 2010). Le second outil consiste en l'estimation de la variabilité des profils chimiques de spécimens analysés au sein d'une même méthode et entre différentes méthodes analytiques (Broséus et al., 2013). Cet outil représente l'estimation de l'intra- et l'inter variabilité pour les études intra- et inter méthodes. Les deux outils requièrent l'usage d'un échantillonnage représentatif (cf. §6.4.a). S'il est estimé à l'issue de la de l'étape d'analyse statistique des données que les résultats analytiques provenant des méthodes considérées ne sont pas similaires, alors l'optimisation de la similarité peut se faire en utilisant deux approches différentes, l'ajustement analytique et l'ajustement mathématique, comme en discute le paragraphe 6.3.

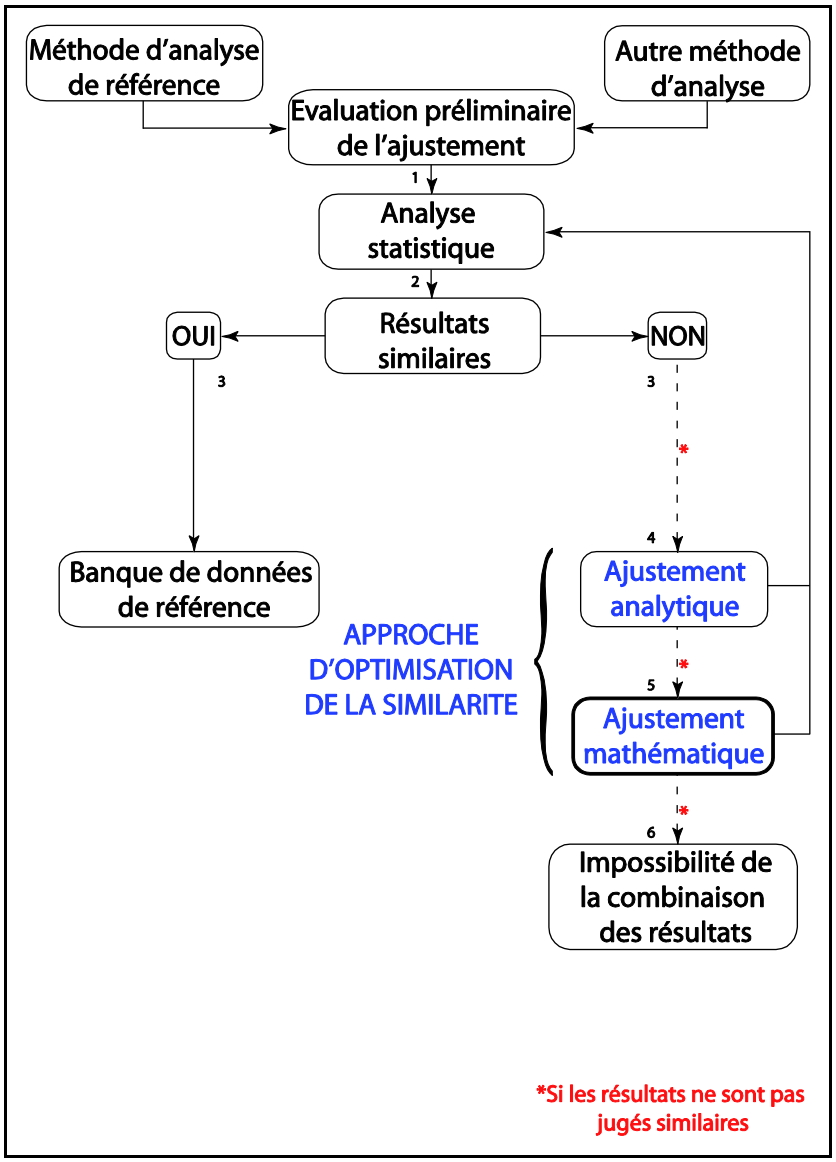


Figure 26. Méthodologie d'estimation et d'optimisation de la similarité de résultats provenant de méthodes analytiques différentes

Si, malgré ces deux étapes, la similarité des résultats analytiques est encore jugée trop différente selon les critères de décision définis, alors il devrait être conclu à l'impossibilité du maintien d'une banque de données commune aux méthodes analytiques investiguées en utilisant la méthodologie d'ajustement des réponses analytiques. Toutefois, l'utilisation de la méthodologie d'ajustement par quantification pourrait être envisagée en dernier ressort, d'où la complémentarité des deux méthodologies mentionnée au Chapitre 4. Cependant, la contrainte principale consiste alors dans le changement de la méthodologie de profilage chimique au sein du laboratoire (dans le cas présent, passage de la semi-quantification à la quantification de chacun des composés du profil chimique) et soulève des questions quant à la gestion future de la banque de données utilisée jusque-là.

6.3 Approche d'optimisation de la similarité

La première étape dans le cadre de l'approche d'optimisation de la similarité des résultats analytiques consiste à déterminer les paramètres analytiques qui l'influencent et la manière de les ajuster en ce sens (étape nommée **ajustement analytique**, cf. §6.3.a). Cette étape fait référence au niveau de paramètre analytique C (caractérisé par les paramètres C_{SEP} et C_{DET}) qui définit les paramètres d'analyse et d'optimisation d'une méthode analytique (cf. Figure 23) et implique par conséquent la comparaison de méthodes analytiques similaires (cf. Tableau 12). Elle prend place lors de l'étape initiale d'**évaluation préliminaire de l'ajustement**.

La seconde étape consiste à établir les relations statistiques (c'est-à-dire, les *règles d'ajustement*), pour chacun des composés du profil chimique, entre les résultats provenant de méthodes analytiques différentes pour en particulier optimiser mathématiquement la similarité existant entre eux (étape nommée **ajustement mathématique**, cf. §6.3.b). Cette approche a été discutée dans une publication précédente (Debrus et al., 2010) dans le cadre de l'étude du scénario d'ajustement 2.2 (cf. Tableau 12).

L'analyse statistique de la similarité des résultats, discutée au §6.4, devrait être effectuée après chaque étape constituant l'approche d'optimisation pour évaluer leur impact respectif sur la similarité des résultats et ainsi estimer l'utilité de chacune d'elles pour le maintien d'une même banque de données par les méthodes considérées.

Dans cette recherche, l'approche de l'ajustement mathématique a été particulièrement investiguée.

6.3.a Evaluation préliminaire de l'ajustement – Ajustement analytique

Comme présenté ci-dessus, deux éléments se côtoient dans cette première étape de la méthodologie proposée (cf. Figure 26). L'une concerne l'obtention d'informations préliminaires sur la similarité des résultats issus des méthodes analytiques considérées ou en d'autres termes consiste en une première évaluation de la complexité de l'ajustement. Elle repose sur une mesure de la similarité des profils chimiques d'échantillons dits *de référence* obtenus avec les différentes méthodes et cherche ainsi à déterminer le degré de similarité entre les composés respectifs. L'autre concerne l'amélioration de la similarité entre les résultats des méthodes considérées grâce à une phase d'ajustement analytique. Ceci implique la détermination des paramètres d'analyse C_{SEP} et C_{DET} qui influencent la similarité analytique ainsi que l'étude précise de la manière dont les ajuster pour accroître cette similarité. Les informations glanées lors de l'évaluation préliminaire pourraient être utilisées pour définir précisément l'approche d'ajustement analytique. Ainsi, ces deux éléments sont étroitement liés, pour autant que l'ajustement analytique soit entrepris.

En effet, l'ajustement analytique est par essence facultatif car il ne concerne que des méthodes analytiquement similaires, d'après la classification réalisée dans cette étude (cf. Tableau 12). De plus, les résultats analytiques pourraient être estimés similaires sur la base de la phase d'évaluation préliminaire[35] ou de l'étape d'analyse statistique, impliquant alors qu'un tel ajustement analytique ne soit pas jugé nécessaire.

[35] En d'autres termes, ceci signifie que l'utilisation de spécimens dits de référence permet une telle conclusion et implique nécessairement l'hypothèse suivante (à démontrer) : si l'ajustement des profils chimiques de ces derniers obtenus avec différentes méthodes est estimé réussi (c'est-à-dire que des profils jugés similaires sont obtenus), alors ceci devrait être le cas pour l'ensemble des échantillons constituant l'échantillonnage sélectionné.

En outre, même si le scénario d'ajustement le permet, il s'agira pour le laboratoire d'analyse de décider s'il souhaite appliquer cette approche ou non, car il pourrait par exemple posséder d'ores et déjà une méthode d'analyse validée qu'il ne souhaiterait pas modifier[36]. En effet, si une amélioration de la similarité était alors observée, une méthode d'analyse ajustée pourrait être utilisée pour la suite du travail de comparaison mais il faudrait auparavant procéder à sa validation analytique. En revanche, si aucune amélioration analytique n'était possible ou si cette approche n'était pas envisagée par le laboratoire en question, alors la méthode analytique initiale serait conservée puis il serait procédé uniquement à l'évaluation préliminaire de l'ajustement avant de passer à l'étape d'analyse statistique.

Dans cette recherche, pour réaliser l'évaluation préliminaire de l'ajustement, 3 spécimens d'héroïne, dits de référence et employés pour le contrôle qualité de l'appareillage, ont été utilisés. Sachant que la pureté des saisies d'héroïne de rue est comprise habituellement entre 5 et 20%, chaque spécimen a été créé pour être représentatif d'un intervalle de concentrations en particulier : de 5 à 10%, de 10 à 15% et finalement de 15 à 20%. Etre en possession de spécimens de référence de concentrations différentes dans les composés cibles s'avère important pour le contrôle qualité mais également pour l'évaluation préliminaire de l'ajustement (pour une présentation de cette importance, cf. §6.4.a).

En pratique, les profils chimiques de ces spécimens obtenus avec la nouvelle méthode analytique peuvent être comparés à ceux de la méthode analytique de référence. La mesure de similarité pourrait être celle mise en place en systématique à l'IPS et présentée au Chapitre 1 : la mesure de la corrélation de Pearson[37]. La similarité des réponses analytiques (c'est-à-dire, les aires de pic) pour chacun des 6 alcaloïdes majeurs[38] composant le profil chimique pourrait être déterminée, respectivement. Ainsi, les composés problématiques (c'est-à-dire, pour un composé donné, la réponse analytique obtenue avec la méthode analytique différente est peu similaire à celle obtenue avec la méthode de référence) pourraient être identifiés.

[36] On parle là d'amélioration de la similarité sur la base de modifications des paramètres analytiques C_{SEP}, telles que le flux de gaz porteur ou du programme de température, et non de celles des paramètres du « tune file » qui n'affectent pas la similarité analytique, selon l'approche d'harmonisation des méthodes analytiques.

[37] Les fondements mathématiques de la corrélation de Pearson sont détaillés dans le §6.4.f.

[38] C'est-à-dire, les composés cibles définis par la méthode de référence : la méconine (MEC), l'acétylcodéine (AC), l'acétylthébaol (AcTB), la 6-monoacétylmorphine (6MAM), la papavérine (PAP) et la noscapine (NOS).

Alors, cela devrait être mis en concordance avec leur qualité chromatographique (mauvaise résolution, « tailing » ou « fronting » par exemple) et, selon le scénario d'ajustement, il pourrait être envisagé de procéder à la phase d'ajustement analytique et effectuer des modifications des paramètres d'analyse de séparation (en GC, le volume d'injection, le split ratio, le flux de gaz porteur, le programme de température du four dans son ensemble) et/ou de détection (en MS, les paramètres du « tune file », cf. Figure 23) pour optimiser la similarité des résultats.

Dans le cadre de l'ajustement analytique, les paramètres définissant le « tune file » en MS devraient être tout particulièrement étudiés, la modification de ces derniers n'impliquant pas une méthode analytique différente selon la méthodologie d'harmonisation des méthodes analytiques (cf. Chapitre 2 et Chapitre 4). Sur la base d'éléments théoriques (cf. Chapitre 2) et d'expérience pratique du groupe expertise des produits stupéfiants de l'IPS, il a été déterminé que les éléments de la source ionique ainsi que le détecteur avaient une certaine influence sur la similarité des résultats. Ainsi, ce sont majoritairement les éléments suivants qui devraient être évalués : le Repeller, l'Ion Focus, l'Entrance Lens, l'Entrance Lens Offset pour les éléments de la source ionique et finalement le voltage de l'electromultiplier pour le détecteur (cf. Figure 18). Une méthodologie a en conséquence été définie dans cette étude pour investiguer leur influence sur la similarité des résultats mais n'a cependant pas été mise en pratique (cf. Annexe 1).

Finalement, il est important de signaler que pour optimiser la similarité entre les résultats, il est possible d'appliquer directement l'étape d'ajustement mathématique à l'aide des règles d'ajustement déterminées (cf. §6.3.b ci-dessous). Toutefois, l'ajustement analytique pourrait avoir une importance non négligeable car il a en théorie une influence sur la qualité de l'ajustement et donc au final sur la faisabilité du maintien d'une même banque de données par les différentes méthodes analytiques considérées. En effet, notons que même si l'approche d'ajustement mathématique pourrait sensiblement améliorer la similarité des résultats, plus cette similarité sera faible lors de l'étape précédente d'évaluation préliminaire de l'ajustement, plus les relations mathématiques entre les résultats analytiques seront complexes (c'est-à-dire que des modèles mathématiques complexes seraient nécessaires pour modéliser les relations entre les résultats).

6.3.b Ajustement mathématique

L'approche de l'ajustement mathématique, discutée dans une publication précédente (Debrus et al., 2010), a été spécifiquement investiguée dans cette étude. Sachant que dans notre cas le profil s'établit par semi-quantification, l'ajustement mathématique consiste, pour chacun des composés du profil, en l'établissement de relations robustes entre les valeurs des aires de pic obtenues avec différentes méthodes analytiques. L'ajustement mathématique nécessite l'utilisation d'un échantillonnage représentatif (c'est-à-dire, des échantillons de diverses concentrations) pour élaborer des modèles mathématiques corrects (cf. §6.4.a). Il faut souligner que plusieurs modèles mathématiques (par exemple, des modèles linéaire, quadratique, cubique, poly-linéaire ou poly-quadratique) peuvent être utilisés pour établir des relations statistiques entre les résultats. En effet, étant donné que certains scénarios d'ajustement impliquent des méthodes d'analyse plus ou moins différentes, des relations mathématiques plus ou moins complexes sont par conséquent attendues. Toutefois, pour l'investigation des scénarios d'ajustement autres que le 2.2 (cf. Tableau 12) – ce dernier ayant requis l'utilisation de modèles complexes (Debrus et al., 2010) –, les modèles mathématiques les plus simples seront privilégiés pour éviter tout risque d'« overfitting » ou de *sur apprentissage* (c'est-à-dire, les modèles linéaire, quadratique et cubique).

Concrètement, les relations statistiques établies pour tous les composés du profil chimique sont utilisées aussi bien pour estimer le degré de similarité entre les résultats analytiques que pour les ajuster mathématiquement. Une fois ajustés mathématiquement, les profils peuvent être introduits dans la banque de données de la méthode de référence malgré le fait qu'ils proviennent d'une autre méthode. Le terme de *données « reference like »* a été défini dans ce travail et correspond aux résultats obtenus avec une autre méthode analytique que celle de référence et ajustés mathématiquement pour être introduits dans la banque de données de référence originelle.

Pratiquement, comme notre profil GC-MS est déterminé par 6 composés (cf. §1.6.b), et étant donné qu'il faut assurer l'ajustement de chacun des composés utilisés pour établir le profil chimique, 6 équations mathématiques doivent être établies. Sachant que le but de la méthodologie consiste à approvisionner la banque de données GC-MS avec les résultats provenant d'une autre méthode analytique, alors les réponses du modèle (c'est-à-dire, le *y* du modèle mathématique) sont les aires de pics GC-MS prétraitées[39] (cf. Figure 27).

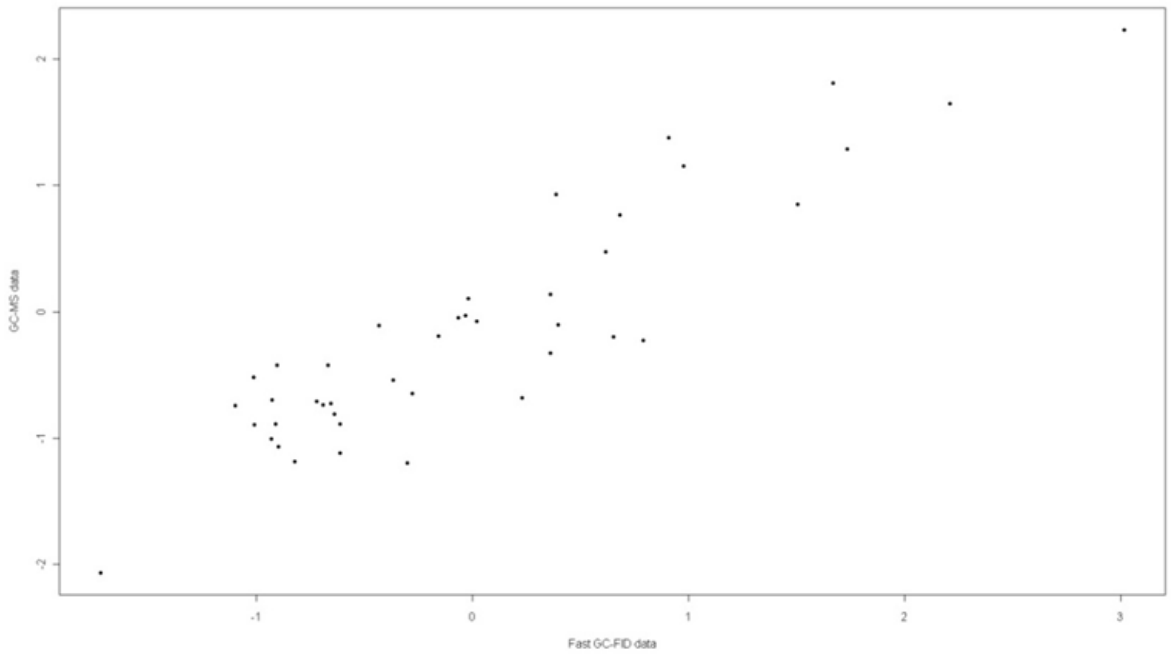


Figure 27. Représentation des aires prétraitées GC-MS du composé MEC pour l'ensemble des spécimens analysés en fonction de celles obtenues en Fast GC-FID

Chaque aire de composé obtenue avec une autre méthode ($COMPOSE_{Autre\ méthode}$) est alors transformée, en utilisant le modèle mathématique approprié, en une aire « GC-MS like » pour ce même composé ($COMPOSE_{GC\text{-}MS\ like}$) qui doit être la plus proche possible de l'aire de ce même composé obtenue en GC-MS ($COMPOSE_{GC\text{-}MS}$). Ainsi, l'ajustement de chacun des composés conduit à un profil chimique « GC-MS like » pour le spécimen correspondant. Le Tableau 14 ci-dessous illustre cette procédure d'établissement des relations statistiques pour chacun des composés du profil, dans le cadre de l'ajustement mathématique de résultats obtenus en Fast GC-FID, par exemple (illustration avec le composé méconine ou MEC).

[39] L'importance du prétraitement est discuté au §6.4.c.

Modèles	MEC
Linéaire	$\mathrm{MEC}_{GC-MS} = \beta_1 . \mathrm{MEC}_{FAST\ GC-FID} + \beta_0$
Quadratique	$\mathrm{MEC}_{GC-MS} = \beta_2 . \mathrm{MEC}^2_{FAST\ GC-FID} + \beta_1 . \mathrm{MEC}_{FAST\ GC-FID} + \beta_0$
Cubique	$\mathrm{MEC}_{GC-MS} = \beta_3 . \mathrm{MEC}^3_{FAST\ GC-FID} + \beta_2 . \mathrm{MEC}^2_{FAST\ GC-FID} + \beta_1 . \mathrm{MEC}_{FAST\ GC-FID} + \beta_0$

Tableau 13. Relations mathématiques établies à l'aide des modèles linéaire, quadratique et cubique (les termes mathématiques sont définis en fonction du composé et de la technique analytique – ici MEC et Fast GC-FID, respectivement)

Il est alors possible d'évaluer la qualité de l'ajustement et donc le degré de similarité des résultats par composé à l'aide des coefficients de détermination ajusté (R^2 ajusté) et de prédiction (Q^2) calculés lors de la comparaison entre les données GC-MS et Fast GC-FID dans cet exemple. Il en découle que le modèle mathématique avec les R^2 et Q^2 les plus élevés peut être déterminé comme étant le plus approprié pour la phase d'ajustement mathématique des résultats obtenus (Debrus et al., 2010).

Dans certains scénarios d'ajustement, les résultats statistiques obtenus pourraient montrer qu'une similarité acceptable est obtenue et donc que le recours à un ajustement mathématique est inutile, d'où l'aspect facultatif de cette étape. Le passage à l'établissement des règles d'ajustement est nécessaire pour apprécier la similarité des résultats mais c'est l'ajustement mathématique qui en découle qui ne l'est pas. De manière à juger du bénéfice de l'ajustement mathématique, l'étape d'analyse statistique qui suit, reposant sur l'ACP-CAH Globale et Locale ainsi que sur les études d'intra- et d'inter variabilité pour des situations intra- et inter méthodes, devrait être réalisée pour les résultats *ajustés* et *non ajustés* mathématiquement de la nouvelle méthode d'analyse.

6.4 Analyse statistique – Estimation de la similarité des résultats

6.4.a Echantillonnage

Pour obtenir une estimation robuste de la similarité des résultats à l'aide des deux outils principaux précités, un échantillonnage représentatif doit être effectué. Dans la banque de données de référence, les échantillons se groupent en classes chimiques en fonction de leurs similarités respectives (qui sont reliées aux concentrations de chacun des composés du profil pour l'échantillon correspondant) (cf. §1.4.d). Il est par conséquent important de démontrer que la méthodologie d'estimation et d'optimisation n'est premièrement pas influencée par l'échantillonnage et deuxièmement est valide pour tous les échantillons, quelles que soient leurs concentrations dans chacun des composés du profil.

Une ACP a été réalisée sur la banque de données de référence GC-MS (304 spécimens d'héroïne saisis en 2009 en Suisse francophone) pour illustrer la distribution des échantillons. Les scores de chacun des spécimens correspondants sont représentés à l'aide des deux premières composantes principales CP1 et CP2 (cf. Figure 28)[40].

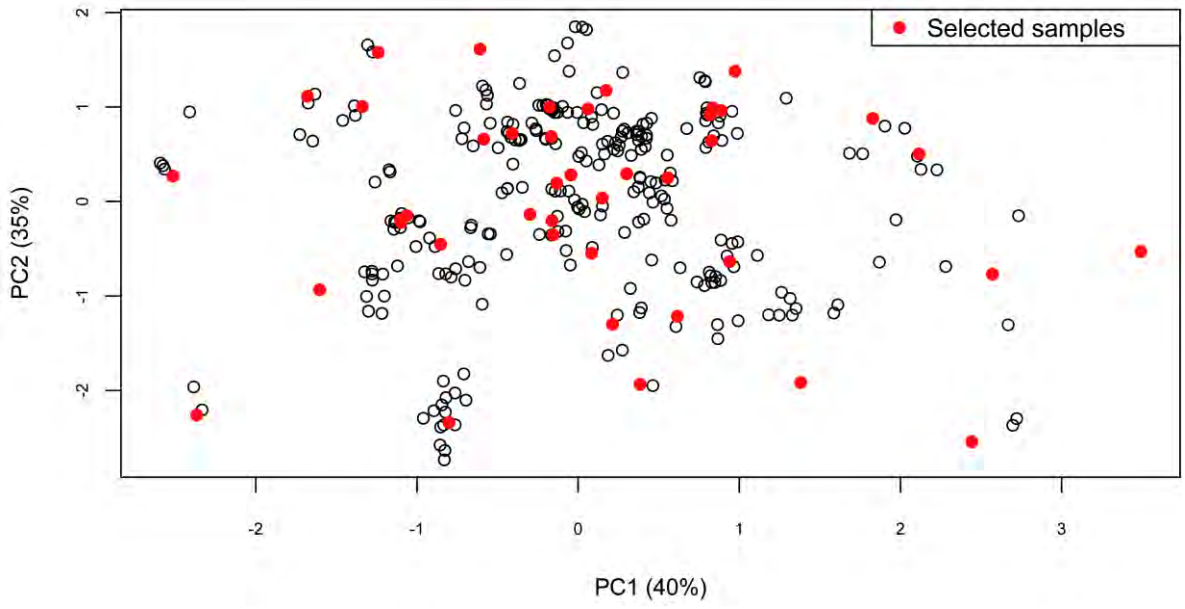


Figure 28. Echantillons sélectionnés dans la banque de données de profils des saisies d'héroïne en 2009

[40] Sachant que plus de 70% de la variance totale est préservée en utilisant les deux premières CPs, la représentation graphique reflète de manière acceptable la structure des données (Varmuza and Filzmoser, 2009).

Pour constituer un échantillonnage représentatif, il s'agit alors de sélectionner des spécimens répartis dans l'entièreté de l'espace d'intérêt sur CP1 et CP2. La disponibilité des spécimens pour constituer l'échantillonnage a également été un facteur à prendre en compte dans la mesure où il fallait que les quantités présentes soient suffisantes pour permettre les analyses dans le cadre de l'investigation de l'ensemble des scénarios d'ajustement. Ainsi, l'échantillonnage choisi comprend 42 spécimens d'héroïne sélectionnés dans la banque de données de référence, qui correspondent aux points rouges dans la Figure 28 ci-dessus.

Se baser sur la répartition des scores des échantillons sur CP1 et CP2 pour atteindre la représentativité de l'échantillonnage n'est pas suffisant et il est important de contrôler que, pour chacun des composés du profil, la distribution des valeurs des aires de l'échantillonnage soit incluse dans celle de la banque données de référence par exemple à l'aide de « boxplots » (cf. Figure 29 puis §6.7.b pour une description théorique des « boxplots »).

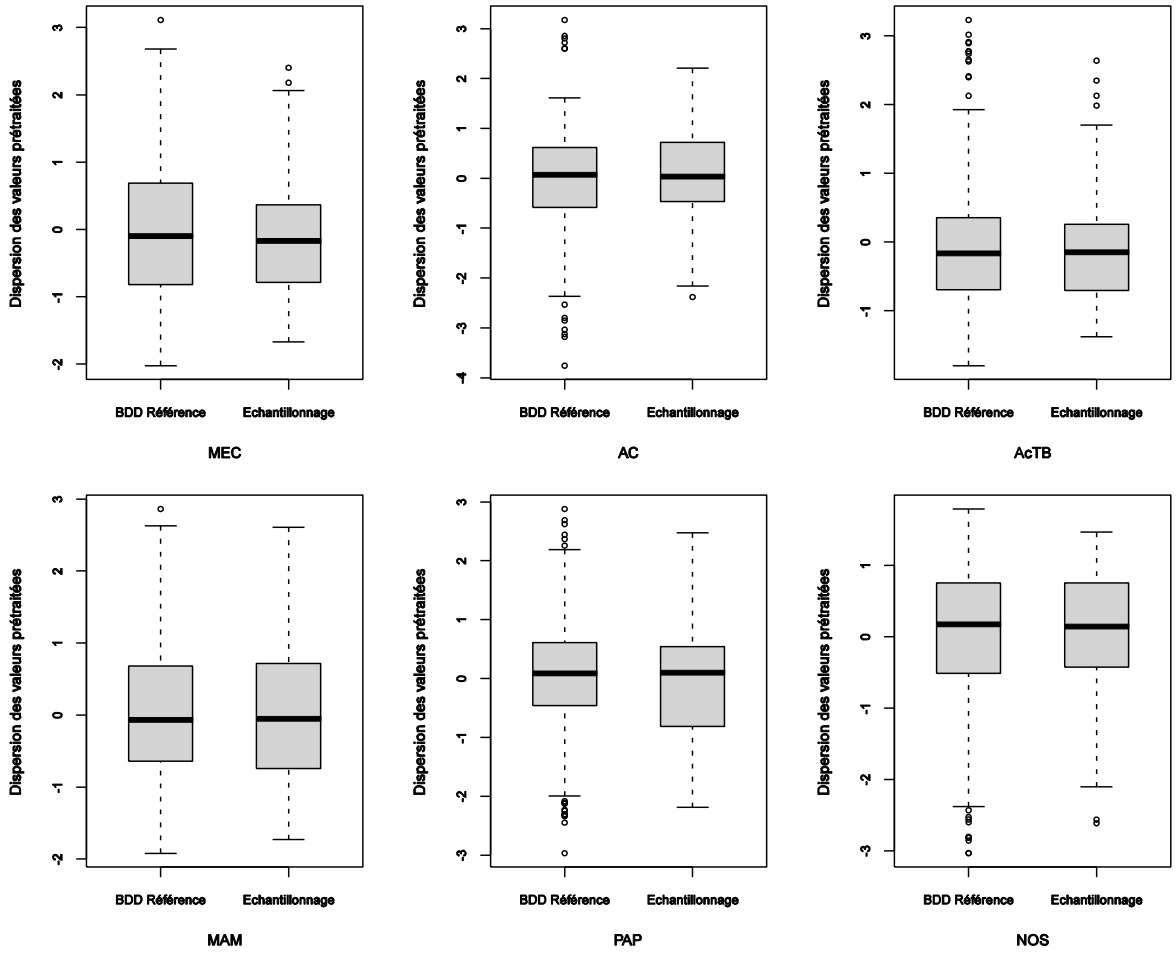


Figure 29. Comparaison de la dispersion des valeurs des composés cibles entre les échantillons présents dans la banque de données de référence (moins les 42 échantillons choisis) et l'échantillonnage sélectionné

La Figure 29 ci-dessus illustre que ceci est le cas sur la base de l'allure générale des « boxplots » et des valeurs médianes respectives. Ainsi, en combinaison à la Figure 28, ces éléments démontrent que la représentativité de l'échantillonnage en termes de distribution des échantillons dans l'entier de la banque de données de référence et de dispersion des valeurs pour chacun des composés est satisfaisante.

Comme cela en a été question au §1.7.b, les spécimens de produits stupéfiants peuvent subir avec le temps une dégradation de leur composition chimique, modifiant ainsi leur profil. Les profils chimiques des échantillons sélectionnés, enregistrés dans la banque de données de référence, ont été réalisés durant l'année 2009, année de leurs saisies respectives. Or, les analyses effectuées dans le cadre des scénarios d'ajustement définis dans cette étude se sont étalées environ de début 2010 à fin 2012 (excepté pour le scénario 2.2, étudié auparavant (Debrus et al., 2010)).

Bien qu'une recherche ait démontré le contraire (Nier et al., 2012), il est envisageable que la composition chimique d'un spécimen ait évolué après l'enregistrement de son profil dans la banque de données de référence (2009) et l'analyse par la suite dans le cadre des scénarios d'ajustement (2010-2012). En conséquence, si les profils chimiques obtenus dans le cadre des scénarios d'ajustement respectifs étaient comparés à ceux enregistrés dans la banque de données de référence (établis de manière contemporaine à leur saisie), alors la mesure de la similarité entre eux pourrait être influencée par cette éventuelle dégradation de la composition chimique. La similarité calculée ne serait alors pas uniquement le reflet de la similarité (ou de la différence) analytique entre les méthodes considérées. Or, cette étude cherche, au travers de la définition des scénarios d'ajustement, à évaluer la part de variabilité due aux différences analytiques dans la similarité des profils chimiques obtenus.

Par conséquent, pour chacun des scénarios d'ajustement – excepté le 2.2 (Debrus et al., 2010) – l'échantillonnage a été analysé simultanément aussi bien avec la méthode d'analyse de référence GC-MS qu'avec la méthode différente considérée et ce sont ces résultats qui sont comparés entre eux. En d'autres termes, les profils chimiques de référence (c'est-à-dire, les profils chimiques des échantillons sélectionnés obtenus par GC-MS et enregistrés dans la banque de données de référence) sont mis à jour. Ainsi, l'effet d'une éventuelle dégradation de la composition chimique des spécimens n'influencera pas la similarité mesurée entre leurs profils chimiques respectifs, obtenus avec des méthodes analytiques différentes (c'est-à-dire, méthode de référence vs. méthode du scénario d'ajustement considéré).

6.4.b Préparation et analyse

Le mode de préparation des échantillons avant leur analyse est le même pour les scénarios d'ajustement 1.2 à 1.5 et 2.1. Pratiquement, trois réplicats de chaque spécimen d'héroïne (environ 8 mg chacun) sont pesés. Dans le cadre de l'étude de l'intra variabilité, ce sont 9 réplicats par spécimen qui ont été préparés puis injectés (cf. §6.4.f) (Broséus et al., 2013). Chaque réplicat est extrait dans 500 µL d'une solution de chloroforme : pyridine (5 :1 v/v) contenant 1 mg/mL d'heneicosane (standard interne). Pour dérivatiser les échantillons, 100 µL de MSTFA sont ajoutés à chaque échantillon et ceux-ci sont ensuite chauffés à 80°C en étuve durant une heure. Finalement, chaque réplicat est individuellement injecté. La procédure détaillant la préparation des échantillons dans le cadre du scénario 2.2 est décrite et discutée dans la publication correspondante (Debrus et al., 2010). Les paramètres analytiques des méthodes développées pour l'analyse des échantillons sont présentés en annexe en fonction des scénarios d'ajustement étudiés (cf. §5.4 et Annexe 2).

6.4.c Prétraitement statistique

Le prétraitement des réponses analytiques est une étape importante dans le processus comparatif des profils chimiques, qui doit être effectué avant la comparaison proprement dite des résultats (c'est-à-dire, la mesure de la similarité) (cf. §1.3.d). Dans le cadre de cette étude et de la problématique générale présentée, le prétraitement des données a une importance encore plus marquée. En effet, comme cela est illustré par la Figure 23 et le Tableau 12, les méthodes analytiques peuvent être différentes selon divers paramètres analytiques. Ainsi, les résultats obtenus sont inévitablement différents dans l'intensité de réponse analytique pour chaque composé du profil, rendant la comparaison directe des profils chimiques impossible. Le prétraitement des données est par conséquent utilisé pour réduire l'influence analytique sur les résultats à comparer, travailler avec des variables sur une échelle de valeurs comparable et trouver des relations statistiques fiables (Massart et al., 1997; Reimann et al., 2008).

Cette recherche ne vise pas à étudier de manière exhaustive l'ensemble des prétraitements existants, même si certaines combinaisons de prétraitements pourraient être plus intéressantes que d'autres en termes de similarité atteinte entre les résultats ou d'amélioration de la séparation des populations d'échantillons liés et non liés, aussi bien pour les études intra- et inter méthodes (Lociciro et al., 2008; Esseiva et al., 2011).

Seul un prétraitement est ainsi investigué, qui se base dans une première étape sur le prétraitement de référence, appliqué en systématique à l'IPS et consistant en l'application de la racine carrée aux données préalablement normalisées à la somme des aires des 6 composés cibles (Esseiva et al., 2011) :

$$N = \sqrt{\frac{x_i^j}{\sum_{i=1}^{6} x_i^j}}$$

Où x_i^j représente l'aire de l'échantillon *j* pour le composé cible *i*, avec *j* allant de 1 à *n* (*n* étant le nombre d'échantillons total) et *i* caractérisant chacun des 6 composés du profil chimique de référence : la méconine (MEC), l'acétylcodéine (AC), l'acétylthébaol (AcTB), la 6-monoacétylmorphine (6MAM), la papavérine (PAP) et la noscapine (NOS).

Dans une seconde étape, deux phases de centrage à la moyenne et de standardisation sont appliquées sur les données normalisées comme l'illustre l'équation suivante.

$$S = \frac{x_{norm_i}^{\ j} - \mu}{SD_i}$$

Avec $x_{norm_i}^{\ j}$ représentant l'aire normalisée de l'échantillon *j* pour le composé cible *i*, μ la moyenne des aires normalisées de *i* pour tous les *j* et SD_i la déviation standard des aires normalisées de *i* pour tous les *j*.

La nécessité de posséder un prétraitement plus avancé que celui de référence se justifie par le fait que, pour certains scénarios d'ajustement, conserver le prétraitement de référence pourrait être satisfaisant tandis que pour d'autres (c'est-à-dire, où les méthodes considérées montreraient une différence analytique initiale importante), il pourrait ne pas atteindre les buts poursuivis par une telle étape et énoncés plus haut.

6.4.d Etude descriptive

Avant l'utilisation des deux outils principaux (cf. §6.4.e et §6.4.f), une étude descriptive des résultats s'impose. Elle concerne les composés du profil chimique dans un premier temps et les profils chimiques en eux-mêmes dans un second temps.

Dans un premier temps, la distribution de la réponse analytique obtenue pour l'échantillonnage, pour chaque composé et pour toutes les méthodes analytiques, est établie à l'aide de *diagrammes en boîtes* ou *boxplots* (cf. §6.7.b).

Dans cette recherche, la distribution de chacun des composés au sein de la méthode de référence peut être respectivement comparée à celles obtenues au sein de la nouvelle méthode d'analyse. Ces éléments préliminaires sur la distribution des variables donneront des informations importantes quant à la similarité des résultats provenant des méthodes d'analyse considérées et permettront une première réflexion quant à la nécessité ou non de recourir à l'étape d'ajustement mathématique. Le bénéfice de cette étape pourra ainsi être rapidement évalué.

La combinaison de ces observations avec les règles d'ajustement calculées permettra d'estimer la qualité de l'ajustement de chacun des composés. Avec cette analyse, les composés dont la similarité, en comparaison à la méthode de référence, est moins bonne par rapport aux autres pourront être identifiés et ces observations pourront être confrontées à celles obtenues lors de la phase d'évaluation préliminaire basée sur l'étude de 3 spécimens de référence. Pour les composés dont l'ajustement serait jugé mauvais, une étude de leur qualité chromatographique devrait être réalisée avant d'envisager une modification des paramètres d'analyse dans le cadre de la phase d'ajustement analytique, permettant potentiellement d'améliorer leur ajustement. Si le passage à l'ajustement mathématique ne permet toujours pas un ajustement satisfaisant du ou des composés problématiques, alors une manière de l'améliorer devrait être investiguée.

En utilisant les résultats d'une ACP appliquée sur les profils chimiques obtenus, une comparaison visuelle, entre la méthode d'analyse de référence et la nouvelle méthode analytique, des facteurs poids (*loadings factor*) de chacun des composés ainsi que de la distribution des scores des échantillons, décrira rapidement la similarité ou non des résultats. En particulier, aussi bien la distribution générale des profils chimiques au sein de chacune des méthodes d'analyse que la distribution de ces derniers les uns par rapport aux autres pourraient être étudiées.

6.4.e ACP-CAH Globale et Locale

Cet outil utilise en séquence l'ACP et le CAH, en deux étapes successives, la première nommée ACP-CAH Globale et la seconde nommée ACP-CAH Locale, pour finalement déterminer si le profil chimique du spécimen correspondant obtenu avec la nouvelle méthode analytique est similaire à celui du même spécimen obtenu avec la méthode d'analyse de référence (cf. Figure 30 et Annexe 3).

La procédure dans laquelle s'insère l'ACP-CAH comprend plusieurs étapes et est résumée de la manière suivante.

1) La banque de données de référence considérée correspond aux analyses GC-MS effectuées sur 304 spécimens d'héroïne saisis en Suisse romande durant l'année 2009.
2) La méthode analytique différente considérée dépend du scénario d'ajustement investigué (cf. §5.2, Tableau 12 et §5.4).
3) L'échantillonnage représente les spécimens sélectionnés parmi la banque de données de référence (cf. §6.4.a).
4) L'échantillonnage sélectionné est analysé simultanément sur chacune des méthodes analytiques pour les raisons explicitées au §6.4.a relatives à l'éventuel vieillissement des échantillons. Pour ces mêmes raisons, les 42 profils chimiques GC-MS des échantillons correspondants sont insérés dans la banque de données de référence, en lieu et place de ces mêmes profils analysés à l'époque, lors de la saisie des spécimens correspondants.
5) Les données analytiques (c'est-à-dire, les profils chimiques obtenus en GC-MS et avec la méthode analytique différente) sont prétraitées selon le prétraitement considéré (cf. §6.4.c).

6) La banque de données GC-MS est alors divisée en trois jeux de données : un set de calibration et un set de validation, aléatoirement composés des 42 échantillons analysés en GC-MS (cf. étape 7) ci-dessous) ainsi qu'un set correspondant à la banque de données GC-MS dite « réduite », car cette dernière contient les profils chimiques de 262 spécimens analysés par GC-MS et qui ne sont ni dans le set de calibration, ni dans le set de validation (304 – 42 = 262).

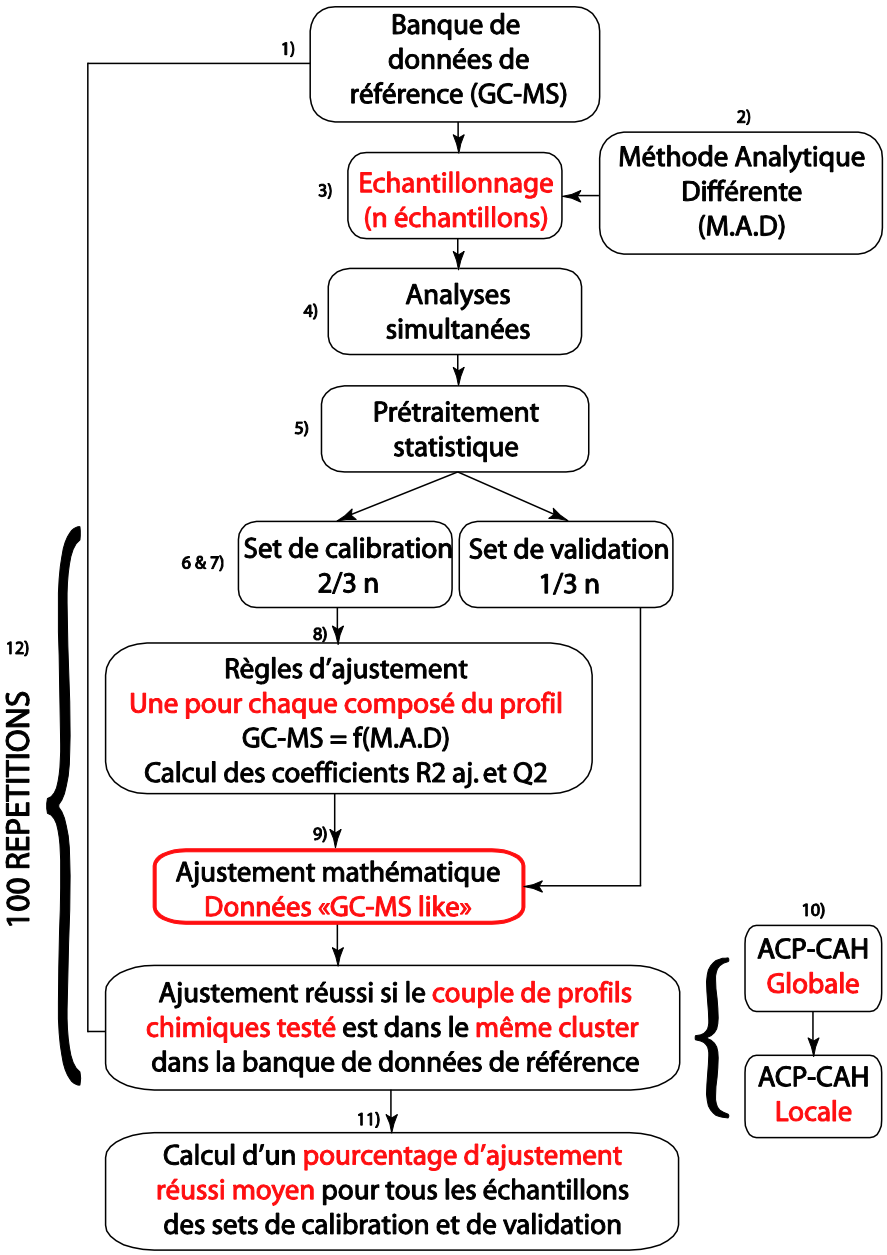


Figure 30. Méthodologie implémentée pour estimer la similarité des profils chimiques de chacun des spécimens provenant des différentes méthodes analytiques

(les chiffres dans la figure font référence aux numéros de chacune des étapes décrites dans ce paragraphe)

7) Les profils chimiques sont aléatoirement choisis pour faire partie d'un set de calibration (2/3 des échantillons) et d'un set de validation (1/3 des échantillons), respectivement (c'est-à-dire, un set de calibration et un set de validation contenant les profils chimiques obtenus avec chacune des méthodes).
8) Sur la base du set de calibration, les règles d'ajustement sont établies pour les modèles mathématiques linéaire, quadratique et cubique, et la similarité, par composé, peut être calculée entre les données obtenues avec la méthode de référence GC-MS et celles obtenues avec la méthode différente considérée, pour chacun des modèles mathématiques (cf. §6.3.b). Pour apprécier la qualité d'ajustement, pour chacun des modèles, les moyennes des R^2 ajustés sont calculées (c'est-à-dire, pour un modèle mathématique donné, la moyenne des R^2 ajustés obtenus pour chaque composé). De plus, les coefficients de prédiction (Q^2) sont estimés à l'aide d'un calcul de « leave-*n*-out cross-validation » (avec *n*=30% de la taille du set de calibration). Ainsi, 9 échantillons ont été aléatoirement enlevés du set de calibration pour créer un set temporaire. Ce set est alors considéré pour ajuster les modèles mathématiques, tandis que les échantillons restants du set de calibration sont utilisés pour estimer Q^2. La procédure a été répétée pour chaque modèle 250 fois pour estimer de manière robuste Q^2.
9) A l'aide des règles d'ajustement préétablies sur le set de calibration, les données du set de validation sont ajustées mathématiquement pour obtenir des données « GC-MS like » (*set de validation « GC-MS like »*), pour chacun des modèles mathématiques (cf. §6.3.b).
10) Pour estimer la similarité des profils chimiques, obtenus avec des méthodes analytiques différentes, la procédure ACP-CAH est mise en place. Il est important de souligner que ce profil peut être ajusté mathématiquement ou non, ce qui permet d'évaluer l'utilité de l'ajustement mathématique selon le scénario d'ajustement étudié en comparant les performances d'ajustement entre des profils ajustés (c'est-à-dire, « GC-MS like ») et non ajustés mathématiquement (Fast GC-FID, par exemple). L'ajustement mathématique peut être fait pour les 3 modèles mathématiques étudiés (linéaire, quadratique et cubique, cf. §6.3.b) et, en fonction du taux d'ajustement atteint, il est possible de déterminer le modèle le plus performant pour l'ajustement des résultats provenant des méthodes considérées.

Comme l'illustrent la Figure 30 ci-dessus et la Figure 31 ci-dessous, l'ACP-CAH est appliquée deux fois de suite. La mesure de la similarité à l'aide de l'ACP-CAH se fait selon le même procédé échantillon par échantillon parmi tous les échantillons des sets de calibration et de validation. Sur la banque de données de référence réduite (262 échantillons), l'ACP est appliquée dans un premier temps pour sélectionner les composantes principales (CPs) expliquant au moins 95% de la variance du jeu de données. Cette étape consiste principalement dans le nettoyage des données dans la mesure où les CPs non sélectionnées, de faibles variances, représentent principalement une variabilité non pertinente (c'est-à-dire, du « bruit ») (Varmuza and Filzmoser, 2009). Alors, le profil chimique GC-MS du spécimen correspondant et le profil chimique du même spécimen analysé avec la méthode différente (mathématiquement ajusté ou non) sont transposés dans l'espace des CPs sélectionnées de la banque de données réduite (à ce moment-là, la banque de données comprend alors 262 + 2 = 264 échantillons). Ensuite, une CAH est effectuée en utilisant les coordonnées des échantillons dans l'espace des CPs considérées comme décrit dans la littérature (Boccard et al., 2007; Stella et al., 2007). En particulier, une mesure de la distance euclidienne existante entre tous les objets (c'est-à-dire, les scores des échantillons sur les CPs choisies) est calculée puis la méthode de mesure entre les clusters est accomplie à l'aide de la méthode de Ward (Ward, 1963) (cf. §6.7.d). Ceci correspond à l'ACP-CAH globale (cf. Figure 31).

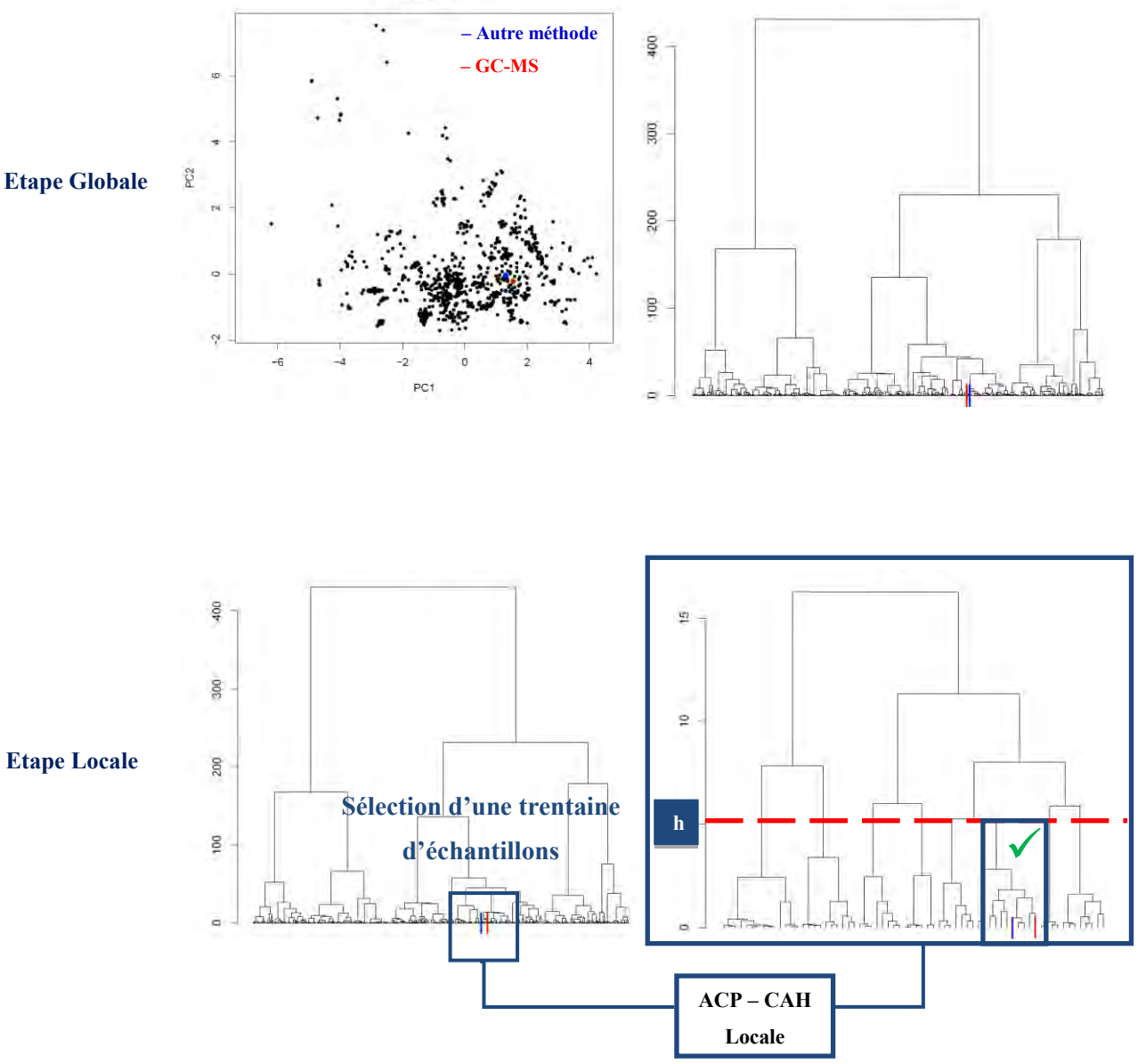


Figure 31. Exemple du processus ACP-CAH pour l'étude de la similarité entre deux profils chimiques d'un même spécimen analysés avec deux méthodes analytiques différentes
(en haut : étape dite globale ; en bas : étape dite locale)

Dans le dendrogramme obtenu, la sélection d'un sous-groupe d'au moins 30 échantillons (choix arbitraire d'environ 10% de la taille totale de la banque de données de référence) présents dans le voisinage immédiat des échantillons testés est réalisée. Sur ce sous-groupe, une étape d'ACP-CAH similaire à celle réalisée ci-dessus est effectuée pour affiner l'évaluation de la similarité des profils considérés dans le cluster sélectionné et ainsi améliorer l'exactitude des prédictions. Concrètement, une ACP est effectuée sur les valeurs d'aires prétraitées de ce sous-groupe d'échantillons (sans les échantillons testés), le nombre approprié de CPs est sélectionné et les échantillons testés sont transposés dans l'espace des CPs sélectionnées pour ce sous-groupe d'échantillons. Ensuite, une CAH basée sur une mesure de distance euclidienne entre les scores des échantillons selon les CPs choisies est effectuée puis la méthode de mesure entre les clusters est accomplie à l'aide de la méthode de Ward. Ceci correspond à l'ACP-CAH locale, appliquée sur les échantillons les plus similaires à ceux testés. Au final, si les deux profils chimiques testés se retrouvent dans le même cluster du dendrogramme local selon la hauteur *h* prédéfinie, alors l'ajustement est considéré comme réussi (cf. Figure 31).

11) Finalement, la performance d'ajustement est estimée en comptant le nombre de fois où les profils d'un même échantillon obtenus avec des méthodes différentes se retrouvent dans le même cluster, selon la hauteur *h* définie dans le dendrogramme local, une fois que le profil chimique obtenu avec la nouvelle méthode est inséré dans la banque de données de référence. Si tel est le cas, alors on considèrera que l'ajustement est réussi pour cet échantillon puis ce procédé est effectué pour tous les échantillons des sets de calibration et de validation. Une performance globale d'ajustement peut alors être calculée d'après la réussite de l'ajustement ou non pour chacun des échantillons.
12) À partir de l'étape 7), la procédure est répétée 100 fois pour s'assurer que chacun des échantillons a au moins une fois fait partie de chacun des deux sets. De plus, une telle répétition permet d'obtenir une estimation correcte de la performance d'ajustement et, lorsque les profils ont été ajustés mathématiquement, pour chacun des modèles mathématiques utilisés.

Le prochain paragraphe présente le second outil statistique principal, l'étude de l'intra- et l'inter variabilité pour les études intra- et inter méthodes.

6.4.f Etudes intra- et inter méthodes

La méthodologie à mettre en place pour l'étude de l'intra- et de l'inter variabilité dans le cadre du profilage chimique ainsi que les notions de seuil de décision et de taux d'erreurs qui en découlent ont été discutées au §1.4 ; ce paragraphe s'attèle donc à démontrer qu'un tel outil peut s'avérer efficace dans le cadre de la problématique investiguée dans cette recherche (Broséus et al., 2013).

Les pourcentages de performance et d'erreurs (VP, FP, FN) ainsi que le seuil de décision calculés permettent dans un premier temps l'évaluation de la performance de la méthodologie implémentée pour séparer les populations d'échantillons liés et non liés pour une seule et même méthode analytique (étude nommée *intra méthode* ou *intra laboratoire* dans cette recherche). Ensuite, par extension, cette performance peut être évaluée dans une perspective plus large, c'est-à-dire lorsque les résultats obtenus avec différentes méthodes sont combinés (étude nommée *inter méthodes* ou *inter laboratoires* dans cette recherche). Cette dernière étude fournit ainsi des renseignements sur la similarité des profils chimiques issus de diverses méthodes et donc sur la possibilité d'approvisionner une banque de données avec différentes méthodes analytiques. En effet, si les résultats issus des diverses méthodes d'analyse sont similaires, alors il n'est pas attendu de grande variation dans les distributions inter méthodes en comparaison de celles intra méthode (en particulier, avec la distribution de la méthode analytique de référence).

La mesure de la similarité employée dans cette recherche est celle mise en place en systématique à l'IPS et repose sur le calcul du coefficient de corrélation de Pearson r_{jk}, défini par (Varmuza and Filzmoser, 2009):

$$r_{jk} = \frac{c_{jk}}{s_j s_k} = \frac{\sum_{i=1}^{n}(x_{ji} - \bar{x}_j)(x_{ki} - \bar{x}_k)}{\sqrt{\sum_{i=1}^{n}(x_{ji} - \bar{x}_j)^2 \sum_{i=1}^{n}(x_{ki} - \bar{x}_k)^2}} \qquad (11)$$

avec c_{jk} la covariance entre les profils chimiques j et k, définis respectivement par i variables (ou composés cibles) et s_j et s_k les déviations standards de j et k. Géométriquement, ce coefficient représente la mesure de l'angle α entre les deux vecteurs $\overrightarrow{x_j}$ et $\overrightarrow{x_k}$.

Plus l'angle est faible, plus les vecteurs sont corrélés. L'intervalle de r_{jk} va de -1 à 1, 1 représentant une relation linéaire parfaite tandis qu'une valeur de -1 indique une parfaite relation linéaire inverse. Dans cette recherche, la valeur du coefficient de corrélation de Pearson est multipliée par 100.

Etude intra méthode

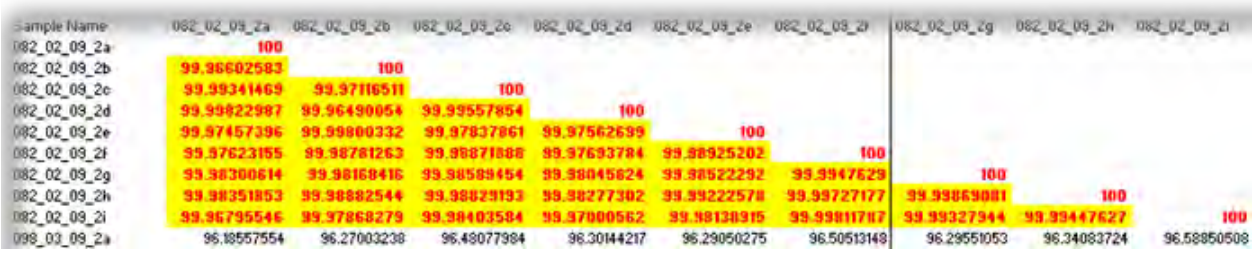
La méthodologie suivante est appliquée pour chacune des méthodes d'analyses investiguées, selon le scénario d'ajustement étudié, pour estimer la performance de séparation entre les populations de liés et non liés.

Pour évaluer l'intra variabilité (échantillons liés) pour chaque méthode un groupe de profils chimiques similaires est créé. Ainsi, 12 spécimens parmi l'échantillonnage complet (c'est-à-dire 42, cf. §6.4.a) sont sélectionnés aléatoirement comme le présente le Tableau 14 ci-dessous. Pour chacun des 12 spécimens, 9 réplicats, ou pesées répétées, sont réalisés et une mesure de similarité est réalisée entre les profils de ces réplicats pour tous les spécimens correspondants. Les valeurs du coefficient de corrélation de Pearson obtenues pour tous les spécimens sont alors sélectionnées (cf. Figure 32). Le nombre de comparaisons à réaliser est : $\left(\frac{9*8}{2}\right)*12 = 432$.

Population	Spécimens sélectionnés
Intra variabilité (12 spécimens)	066_02_09_6_2 ; 082_02_09_2 ; 083_02_09_4 ; 098_03_09_2 ; 098_03_09_3 ; 203_05_09_11 ; 344_09_09_4 ; 362_09_09_2 ; 398_10_09_3 ; 445_11_09_1_4 ; 461_11_09_3 ; 479_12_09_2
Inter variabilité (30 spécimens)	035_01_09_7_1 ; 039_01_09_2 ; 042_01_09_1 ; 098_03_09_4 ; 100_03_09_17 ; 113_03_09_1 ; 166_04_09_3 ; 169_04_09_1 ; 179_04_09_2 ; 201_05_09_2 ; 203_05_09_8 ; 205_05_09_1 ; 210_05_09_2 ; 213_05_09_1 ; 221_05_09_2 ; 226_05_09_1 ; 255_06_09_1 ; 258_06_09_3 ; 267_07_09_1 ; 277_07_09_2 ; 291_07_09_1 ; 341_09_09_3 ; 342_09_09_6 ; 344_09_09_3 ; 362_09_09_1 ; 373_09_09_1 ; 374_09_09_5 ; 401_10_09_2 ; 445_11_09_1_1 ; 455_11_09_3

Tableau 14. Répartition des spécimens sélectionnés puis utilisés pour dresser les distributions d'intra- et d'inter variabilité

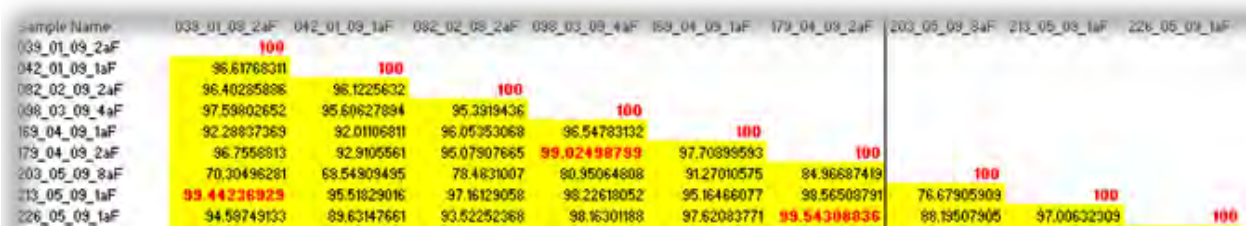
Ainsi, une intra variabilité se distribuant vers des valeurs de coefficient de corrélation de Pearson particulièrement élevées sera obtenue (car constituée de réplicats pour chaque spécimen). Comme il s'agit dans cette recherche de comparer plusieurs méthodes analytiques entre elles, cela permettra d'obtenir une observation valide de l'effet de la répétabilité de la technique analytique sur la distribution d'échantillons de profils similaires (bien que la variabilité due à l'homogénéité des échantillons contribue à l'intra variabilité).



Sample Name	082_02_09_2a	082_02_09_2b	082_02_09_2c	082_02_09_2d	082_02_09_2e	082_02_09_2f	082_02_09_2g	082_02_09_2h	082_02_09_2i
082_02_09_2a	100								
082_02_09_2b	99.86602583	100							
082_02_09_2c	99.99341469	99.97116511	100						
082_02_09_2d	99.99822987	99.96490054	99.99557854	100					
082_02_09_2e	99.97457396	99.99800332	99.97837861	99.97562699	100				
082_02_09_2f	99.97623155	99.98781263	99.98871888	99.97693784	99.98925202	100			
082_02_09_2g	99.98300614	99.98168416	99.98589454	99.98045824	99.98522292	99.9947629	100		
082_02_09_2h	99.98351853	99.98882544	99.98829193	99.98277302	99.99222578	99.99727177	99.99869081	100	
082_02_09_2i	99.96795546	99.97868279	99.98403584	99.97000562	99.98138315	99.99811787	99.99327944	99.99447627	100
098_03_09_2a	96.18557554	96.27003238	96.48077984	96.30144217	96.29050275	96.50513148	96.29551053	96.34083724	96.58850508

Figure 32. Exemple des valeurs de coefficient de corrélation sélectionnées pour un spécimen (cellules surlignées en jaune) pour dresser l'intra variabilité d'une certaine méthode analytique

Pour l'inter variabilité, les 30 spécimens restants provenant de différentes saisies policières sont sélectionnés et 3 réplicats pour chacun d'entre eux sont analysés. Les valeurs moyennes sont calculées pour les 3 réplicats d'un spécimen respectif et sont utilisées pour les mesures de similarité. Les valeurs de coefficient de corrélation de Pearson obtenues entre tous ces profils sont alors sélectionnées (cf. Figure 33). Le nombre de comparaisons à effectuer est : $\left(\frac{30 * 29}{2}\right) = 435.$



Sample Name	039_01_09_2aF	042_01_09_1aF	082_02_09_2aF	098_03_09_4aF	169_04_09_1aF	179_04_09_2aF	203_05_09_8aF	213_05_09_1aF	226_05_09_1aF
039_01_09_2aF	100								
042_01_09_1aF	96.61768311	100							
082_02_09_2aF	96.40285888	96.1225632	100						
098_03_09_4aF	97.59802652	95.60627894	95.3919436	100					
169_04_09_1aF	92.28837369	92.0110681	96.05353068	96.54783132	100				
179_04_09_2aF	96.7556813	92.9105561	95.07907665	99.02498799	97.70899593	100			
203_05_09_8aF	70.30496281	68.54909495	78.4831007	80.95064808	91.27010575	84.96687419	100		
213_05_09_1aF	99.44236929	95.51829016	97.16129058	98.22618052	95.16466077	99.56508791	76.67905909	100	
226_05_09_1aF	94.59749133	89.63147661	93.52252368	98.16301188	97.62083771	99.54308836	88.19507905	97.00632309	100

Figure 33. Exemple des valeurs de coefficient de corrélation sélectionnées (cellules surlignées en jaune) pour dresser l'inter variabilité d'une certaine méthode analytique

Etude inter méthodes

Dans ce cas de figure, les résultats (c'est à dire, les profils chimiques des spécimens correspondants) provenant des différentes méthodes sont comparés entre eux pour estimer leur similarité.

Dans un premier temps, l'ensemble de l'échantillonnage (c'est-à-dire, les 42 spécimens) est utilisé pour déterminer les règles d'ajustement pour chacun des composés entre les méthodes considérées (cf. §6.3.b). L'ajustement mathématique peut ainsi être réalisé pour chaque profil chimique en utilisant le modèle mathématique approprié.

Dans un deuxième temps, les distributions intra- et inter variabilité pour l'étude inter méthodes sont estimées. Pour dresser l'intra variabilité, le profil chimique du spécimen correspondant obtenu avec la méthode de référence GC-MS est comparé à celui obtenu pour le même spécimen avec une autre méthode analytique (en fonction du scénario d'ajustement étudié). Ceci est effectué pour les 9 réplicats de chacun des 12 spécimens représentant la population d'échantillons liés. De même que pour l'étude intra méthode, le coefficient de corrélation de Pearson est utilisé comme mesure de la similarité. Le nombre de comparaisons à effectuer est : (9 * 9 * 12) = 972.

Pour estimer l'inter variabilité, parmi les 30 spécimens formant la population des échantillons non liés, une moitié analysée avec la méthode de référence GC-MS et une autre moitié analysée avec la méthode analytique différente sont sélectionnées aléatoirement. Les valeurs du coefficient de corrélation de Pearson calculées entre les profils chimiques correspondants sont alors utilisées pour établir la distribution (cf. Figure 34). Le nombre de comparaisons à effectuer est : $\left(\frac{30 * 29}{2}\right) = 435.$

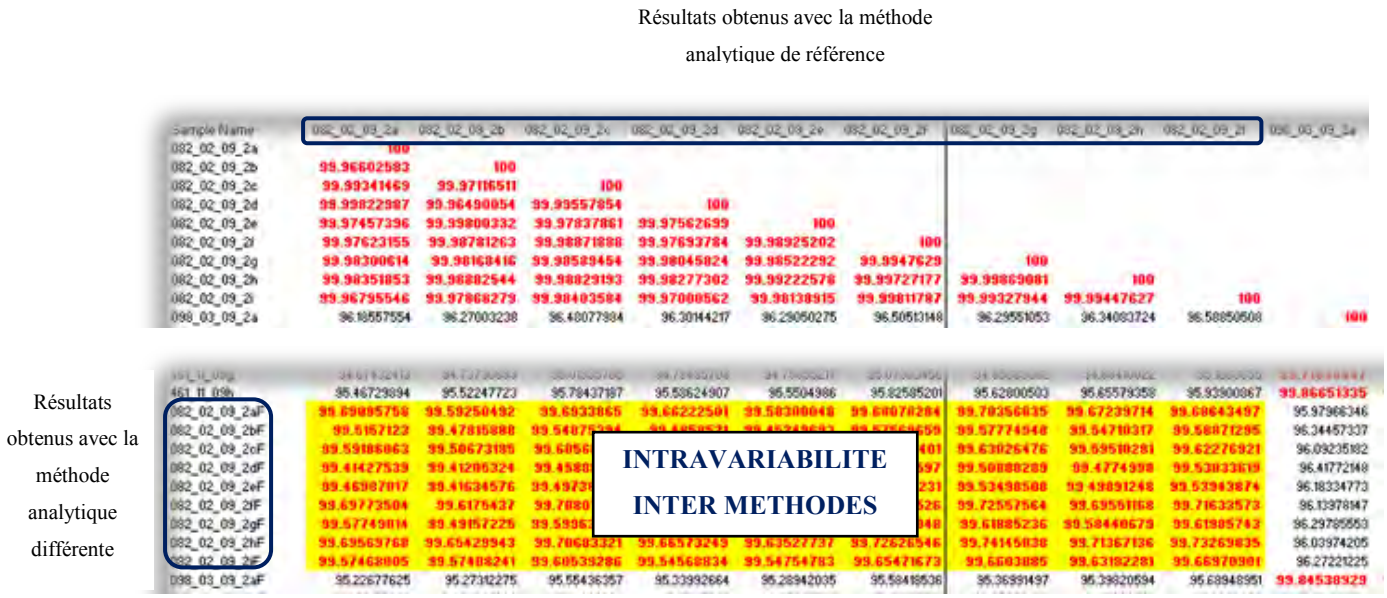


Figure 34. Exemple avec un certain spécimen des valeurs de coefficient de corrélation sélectionnées (cellules surlignées en jaune) pour dresser l'intra variabilité inter méthodes

Le risque de combiner des résultats issus de différentes méthodes consiste à obtenir, en comparaison aux distributions intra méthode, une intra variabilité inter méthodes plus dispersée et se déplaçant vers de plus faibles valeurs de similarité. D'après la manière dont l'intra variabilité inter méthodes est évaluée dans cette recherche, ce déplacement vers de plus faibles valeurs de similarité serait le témoin d'une faible similarité entre les profils chimiques de mêmes spécimens analysés sur des méthodes différentes. Le degré de ce déplacement est particulièrement influencé par la différence analytique initiale entre les méthodes considérées.

Ainsi, les distributions inter méthodes sont estimées avec et sans l'ajustement mathématique des profils pour évaluer l'utilité de ce dernier pour le partage d'une banque de données par différentes méthodes analytiques, selon le scénario d'ajustement. Il s'agira de confirmer l'hypothèse que l'ajustement mathématique des profils chimiques conduit à une amélioration de la performance de la méthodologie de profilage lorsque les résultats provenant de différentes méthodes sont combinés. Si tel est le cas, une amélioration dans la séparation de l'intra- et l'inter variabilité inter méthodes devrait se produire, telle qu'une intra variabilité plus étroite, un déplacement de cette dernière vers de plus faibles valeurs de similarité évité ou, du moins, le degré de ce déplacement diminué.

Finalement, cette étude se focalisant sur le profilage chimique de l'héroïne, les liens chimiques devront être déterminés avec chacune des méthodes d'analyse avec comme résultats référents les liens établis avec la méthode analytique de référence. Les liens seront déterminés en fonction du seuil décisionnel déterminé lors des études d'intra- et d'inter variabilité intra méthode respectives.

6.5 ACP-CAH : influence de *h*

L'intérêt de la démarche ACP-CAH repose sur une étude très fine et très précise de la similarité, car l'évaluation de la réussite ou non de l'ajustement se fait échantillon par échantillon. Toutefois, la difficulté majeure de déterminer la réussite de l'ajustement sur la présence ou non des profils chimiques dans le même cluster est l'estimation d'une hauteur *h* à laquelle couper le dendrogramme local. En effet, dans la présente recherche, la problématique suite à l'utilisation de la CAH ne se situe pas dans la détermination d'un nombre optimal de clusters (cf. §6.7.d pour une présentation de cette dernière).

La méthodologie ACP-CAH implémentée et décrite dans le paragraphe 6.4.e cherche à déterminer si les profils chimiques d'un même spécimen obtenus avec deux méthodes analytiques différentes se retrouvent dans le même cluster une fois que le profil chimique obtenu avec la nouvelle méthode est inséré dans la banque de données de référence. L'idée étant ainsi de fournir des informations quant à la conservation de la structure des données dans la banque des profils de référence (cf. §5.3, sous-hypothèse 2.2). Nécessairement, sur la base des éléments théoriques intrinsèques à la CAH, cette détermination va dépendre de la distance existante entre ces deux profils chimiques, c'est-à-dire la hauteur *h* les séparant dans le dendrogramme à l'étape dite *locale*. Cette hauteur *h* fait ainsi office de seuil de décision quant à l'appartenance au même cluster des deux profils chimiques. En d'autres termes, les profils chimiques sont considérés comme similaires s'ils appartiennent au même cluster, selon la hauteur *h* ayant été définie comme valeur seuil. La hauteur *h* représente ainsi un critère de similarité.

La performance atteinte est directement dépendante de la valeur *h* sélectionnée, pour un scénario d'ajustement considéré. Pour un même scénario d'ajustement, on s'attend à observer une augmentation de la performance d'ajustement corrélée à une augmentation de *h* (la performance d'ajustement sera évidemment plus élevée pour une valeur de *h* plus élevée, le profil chimique obtenu avec la nouvelle méthode d'analyse ayant plus de chances de se retrouver dans le même cluster que le profil de référence une fois inséré dans la banque de données), jusqu'à certainement atteindre un palier (c'est-à-dire qu'à partir d'une certaine valeur de *h*, les profils chimiques testés seraient toujours déterminés comme étant dans le même cluster). Intuitivement, en regard des définitions établies au Chapitre 5, on s'attend à ce que la distance *h* séparant les profils chimiques d'un même spécimen obtenus dans le cadre de scénarios d'ajustements jugés analytiquement proches soit plus faible que lorsque les scénarios d'ajustement investigués sont analytiquement moins similaires.

Ou, en d'autres termes, si la notion de scénarios d'ajustement a été correctement définie, alors la performance d'ajustement obtenue pour des scénarios d'ajustement estimés similaires devrait être plus élevée pour des valeurs de *h* plus faibles que lorsque les scénarios d'ajustement investigués sont analytiquement moins similaires. Par conséquent, sur la base de ces réflexions, une étape nécessaire dans cette étude consiste à investiguer les possibilités existantes pour choisir une valeur de *h* pertinente, ce qui est abordé au Chapitre 7 suivant.

Pour l'investigation de l'ensemble des scénarios d'ajustement et afin de confronter aux résultats les réflexions discutées ci-dessus, il a été décidé d'évaluer l'évolution de la performance d'ajustement en fonction de valeurs croissantes de *h* définies dans un large intervalle de valeurs[41]. Finalement, pour tenter d'estimer une seule et même valeur de *h* pertinente, il pourrait être envisagé de tenir compte de la démarche largement mise en place en police scientifique dans le cadre du profilage chimique de produits stupéfiants : la détermination d'un seuil de décision sur la base des distributions d'intra- et d'inter variabilité, desquelles découlent les taux de VP, FP et FN associés (cf. §1.4.b). Le paragraphe 6.6 discute de cette démarche qui permet en finalité l'évaluation de la complémentarité des deux outils statistiques principaux implémentés dans cette recherche pour estimer la similarité de profils.

[41] De 0.5 à 10 par paliers de 0.5, puis de 11 à 20 par paliers de 1, puis finalement de 22 à 50 par paliers de 2 pour un total de 45 valeurs de *h* testées : 0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5, 5.5, 6, 6.5, 7, 7.5, 8, 8.5, 9, 9.5, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 22, 24, 26, 28, 30, 32, 34, 36, 38, 40, 42, 44, 46, 48 et 50.

6.6 Etude de la complémentarité entre l'ACP-CAH et l'étude des distributions d'intra- et d'inter variabilité inter méthodes

L'évaluation de la complémentarité des deux outils principaux représente une étape cruciale dans cette recherche et pour mener à bien cette démarche, les distributions d'intra- et d'inter variabilité inter méthodes sont utilisées. Concrètement, une mesure du coefficient de corrélation de Pearson est effectuée entre chacun des couples de profils chimiques testés lors de l'étape locale de l'ACP-CAH. Cette mesure se fait que l'ajustement ait été considéré comme réussi ou non d'après la hauteur *h* existant entre ces derniers dans le dendrogramme local.

Ainsi, sachant que pour chacun des scénarios d'ajustement les études inter méthodes sont entreprises et que celles-ci résultent dans la fixation d'un seuil de décision (valeur du coefficient de corrélation de Pearson), alors la validité de *h* en tant que critère de similarité des profils chimiques obtenus avec des méthodes analytiques différentes pourra être évaluée sur la base du seuil. Sachant que du seuil de décision résultent les taux de VP, FP et FN, alors ces valeurs seront particulièrement utiles et informatives pour estimer une valeur de *h* pertinente. Il pourrait par exemple être conclu qu'une certaine valeur de *h* ne convienne pas car le taux de FP associé à une telle valeur, calculé grâce à la mise en relation des mesures de coefficient de corrélation de Pearson pour chacun des couples de profils chimiques testés et du seuil de décision résultant de l'étude d'intra- et d'inter variabilité inter méthodes, serait trop élevé.

En résumé, la mesure du coefficient de corrélation de Pearson a été calculée pour toutes les valeurs de *h* définies dans l'intervalle considéré (cf. §6.4.e), que l'ajustement ait été jugé réussi ou non selon la valeur de *h* dans l'étape locale, et ceci pour tous les échantillons testés, du set de calibration et du set de validation.

Les informations glanées dans le cadre de cette démarche permettront d'estimer si le processus d'ACP-CAH peut être utilisé pour déterminer que deux profils chimiques sont similaires ou bien si son utilité principale reste la description de la similarité des données.

6.7 Outils statistiques utilisés

6.7.a Récapitulatif

Outils	Observations	Jeux de données étudiés
Etude de la dispersion des variables (§6.4.d)	Motif de la dispersion (intervalle de valeurs, outliers, valeurs médianes)	Pour chaque méthode analytique séparément ; avec et sans ajustement mathématique des résultats de la nouvelle méthode analytique ; pour chacun des prétraitements
ACP (§6.4.d)	Loadings des composés Motif de distribution des scores	Pour chaque méthode séparément puis lors de la combinaison des résultats (avec et sans ajustement mathématique)
Règles d'ajustement (§6.3.b et §6.4.e)	Coefficients R^2 ajustés et Q^2 pour chaque composé	Comparaison entre les résultats GC-MS et de l'autre méthode
Méthodologie ACP - CAH (Debrus et al., 2010) (§6.4.e)	Pourcentage d'ajustement réussi	Résultats obtenus avec la nouvelle méthode analytique, avec et sans ajustement
Etudes de l'intra- et de l'inter variabilité intra- et inter méthodes (Broséus et al., 2013) (§6.4.f)	Motif de la distribution Qualité de la séparation des populations Evolution des VP, FP et FN en fonction du seuil décisionnel	Pour chaque méthode séparément Mise en commun des résultats obtenus avec la GC-MS et avec la nouvelle méthode analytique (avec et sans ajustement mathématique des résultats)
Etude des liens chimiques (§6.4.f)	Liens chimiques identifiés avec la nouvelle méthode analytique vs. avec la méthode analytique de référence	Avec et sans ajustement mathématique

Tableau 15. Résumé des différents outils définis pour estimer la similarité des profils chimiques provenant de différentes méthodes analytiques

6.7.b Boxplots

Le « boxplot » consiste en une représentation graphique informative décrivant une distribution de données, basée sur la médiane et les quartiles (Varmuza and Filzmoser, 2009). Ces derniers divisent la distribution des données en 4 parties égales correspondant aux percentiles de 25%, 50% et 75%, aussi appelé le premier (Q_1), le second (Q_2) et le troisième (Q_3) quartile, respectivement. Il s'agit en d'autres termes de chacune des 3 valeurs qui divisent les données triées en 4 parts égales. Le percentile 25% (Q_1) correspond à la valeur pour laquelle 25% des données lui sont inférieures et 75% supérieures. Le second quartile (percentile 50%) est équivalent à la médiane, valeur pour laquelle les données sont partagées en deux parts égales. L'écart interquartile (interquartile range – IQR) correspond à la différence entre le troisième et le premier quartile et représente un critère de dispersion de la série, IQR = $Q_3 - Q_1$. La Figure 35 illustre le boxplot.

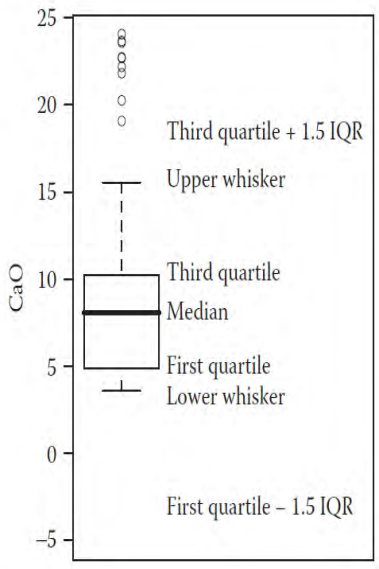


Figure 35. Représentation d'un boxplot avec illustration des définitions et limites (Varmuza and Filzmoser, 2009)

La hauteur de la boîte est donnée par *Q1* et *Q3* et la ligne horizontale dans la boîte correspond à la médiane de la série de données. La largeur, quant à elle, n'a aucune signification. Une moustache s'étend de *Q1* à la plus petite valeur des données dans l'intervalle *Q1* à *Q1* – 1.5*IQR et est nommée moustache inférieure. La moustache supérieure s'étend de *Q3* à la valeur des données la plus grande sur l'intervalle *Q3* à *Q3* + 1.5*IQR. Les données aberrantes (éloignées) ou « outliers » (c'est-à-dire qu'elles ne se trouvent pas dans l'intervalle [*Q1* – 1.5*IQR, *Q3* + 1.5*IQR]) sont représentées graphiquement par des points individuels.

La représentation sous forme de boxplot est valide (car utilisable pour tout type de distribution), robuste (car non influencée par les valeurs éloignées) et permet d'identifier les valeurs éloignées. Ainsi, cette représentation s'avère particulièrement utile pour comparer les caractéristiques de diverses distributions.

6.7.c ACP

La notion de variables linéaires latentes

(Varmuza and Filzmoser, 2009)

Il s'avère nécessaire de présenter un concept fondamental de l'analyse de données multivariées avant de décrire l'ACP : les variables linéaires *latentes* (on parle de *composantes* voire de *facteurs*). Il s'agit de la combinaison mathématique linéaire de plusieurs variables en une seule, cette dernière avec une certaine propriété (par exemple, la séparation de classes, la prédiction d'objets à des classes prédéfinies ou une bonne représentation de la structure des données). La composante résume l'ensemble des variables initiales de manière à obtenir la propriété voulue. La valeur d'une composante est appelée un *score*. L'équation générale des variables latentes linéaires est

$$\boldsymbol{u = b_1 x_1 + b_2 x_2 + \cdots + b_m x_m} \quad (12)$$

Avec u le score (c'est-à-dire, la valeur de composante), b_i les loadings (c'est-à-dire, les coefficients décrivant l'influence des variables sur le score) et x_i les variables du jeu de données initiales.

Les variables forment le vecteur *variable* $\boldsymbol{x} = (x_1, \ldots, x_m)^T$ et les loadings le vecteur *loading* $\boldsymbol{b} = (b_1,\ldots,b_m)^T$ de telle sorte que le score peut être calculé par le produit scalaire de $\boldsymbol{x}^T$ et $\boldsymbol{b}$

$$\boldsymbol{u} = \boldsymbol{x}^T.\boldsymbol{b} \qquad (13)$$

Selon le but visé par l'analyse des données, différents critères mathématiques sont appliqués pour la définition des composantes. Concernant l'ACP, il s'agit d'obtenir la variance maximale pour les scores.

D'un point de vue graphique, le calcul des scores peut être considéré comme une projection orthogonale d'un vecteur $\boldsymbol{x}$ sur une ligne droite définie par le vecteur *loading* $\boldsymbol{b}$ (cf. Figure 36). Un vecteur *score* $\boldsymbol{u}$ est obtenu et contient les scores de chacun des n objets (c'est-à-dire, les valeurs de la composante pour tous les objets).

$$\boldsymbol{u} = \boldsymbol{X}.\boldsymbol{b} \qquad (14)$$

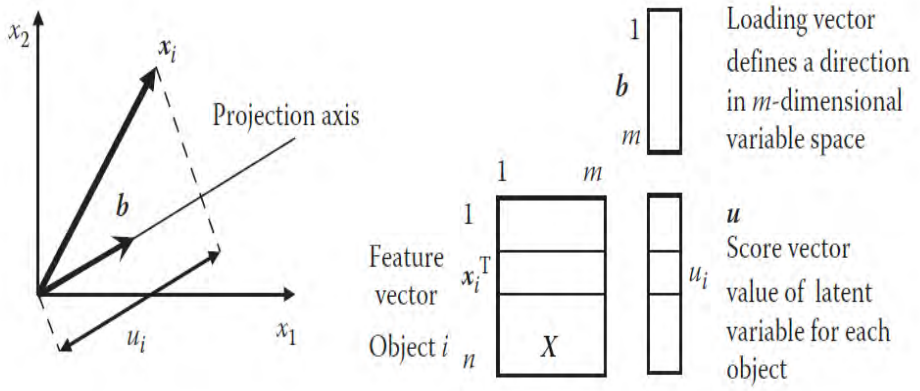


Figure 36. Projection rectangulaire d'un vecteur *variable* $\boldsymbol{x}_i$ sur un axe défini par le vecteur *loading* $\boldsymbol{b}$ résultant dans le score u_i (Varmuza and Filzmoser, 2009)

Si plusieurs composantes sont calculées, les vecteurs *loading* correspondants sont enregistrés dans une matrice des loadings ***B*** et les scores forment une matrice des scores ***U*** (les scores représentent alors différentes combinaisons linéaires des mêmes *m* variables) (cf. Figure 37).

$$\boldsymbol{U} = \boldsymbol{X}.\boldsymbol{B} \quad (15)$$

Le graphique des loadings illustre les similarités des variables et leurs influences sur les scores. Les variables proches de l'origine ont de petits loadings et ont donc généralement moins d'influence sur les scores (par exemple, la variable 4 dans la Figure 37). La position des loadings est à mettre en relation avec la position respective des objets sur le graphique des scores : par exemple, les variables 2 et 5 ont des valeurs élevées pour les objets localisés au même endroit sur le graphique des scores.

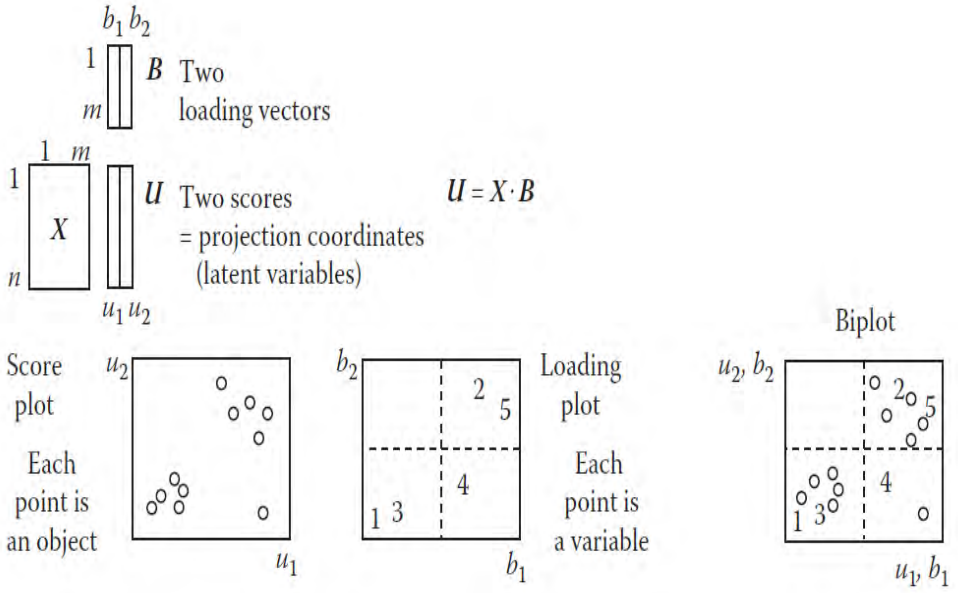


Figure 37. Projection de l'espace à *m* variables ***X*** sur un plan défini par deux vecteurs *loading* de la matrice ***B***. Le graphique résultant contient un point pour chacun des *n* objets, dont les coordonnées sont données par les scores (ici, u_1 et u_2). La figure des loadings illustre l'influence des variables et est à mettre en relation avec la figure des scores. Le biplot représente une combinaison des graphiques des scores et des loadings. (Varmuza and Filzmoser, 2009)

Les graphiques de loadings peuvent être représentés avec des flèches partant de l'origine jusqu'aux points qui représentent les variables correspondantes. Alors, la longueur de la flèche est proportionnelle à la contribution de la variable sur les deux composantes considérées. De plus, l'angle entre deux flèches est une mesure approximative de la corrélation existante entre les deux variables correspondantes (c'est-à-dire, plus l'angle est faible, plus la corrélation est élevée).

Analyse en Composantes Principales

(Varmuza and Filzmoser, 2009)

Le but de l'ACP consiste dans la réduction de la dimension initiale du jeu de données considéré. Il s'agit d'une méthode statistique exploratoire, dite *non supervisée*. Elle peut être considérée comme une méthode de calcul d'un nouveau système de coordonnées formé par les variables latentes (ou composantes), qui est orthogonal, et où uniquement les dimensions (c'est-à-dire, les composantes) les plus informatives sont utilisées.

En plus de permettre la visualisation par représentation graphique de données multivariées, l'ACP transforme des variables hautement corrélées en un set plus petit de variables latentes non corrélées et permet la séparation entre l'information pertinente et le bruit.

Avant d'aborder les relations mathématiques sous-jacentes à l'ACP, les principales caractéristiques de cette dernière sont que :

- la première composante principale (CP1) représente la variable latente linéaire avec le maximum de variance possible ;
- pour les CPs suivantes, la direction de chacune d'elles est respectivement orthogonale à la précédente. De plus, chaque CP explique le maximum de variance possible des scores. Ainsi, les variances des CPs avec un chiffre élevé sont faibles voire nulles ;
- tous les vecteurs *loading* sont orthogonaux les uns par rapport aux autres, ainsi l'ACP représente une rotation du système original de coordonnées orthogonales, avec un nombre plus petit d'axes ;
- les scores sont des variables latentes non corrélées.

Concrètement, l'ACP transforme une matrice de données $\boldsymbol{X}(n \times m)$ – contenant les données de n objets décrits par m variables – en une matrice de plus petite dimension $\boldsymbol{T}(n \times a)$. $\boldsymbol{T}$ représente la matrice des scores pour les n objets et les a composantes principales (CPs) et $\boldsymbol{P}$ correspond à la matrice des loadings (c'est-à-dire, le poids p[42] de chacune des m variables sur chacune des a CPs) (cf. Figure 38).

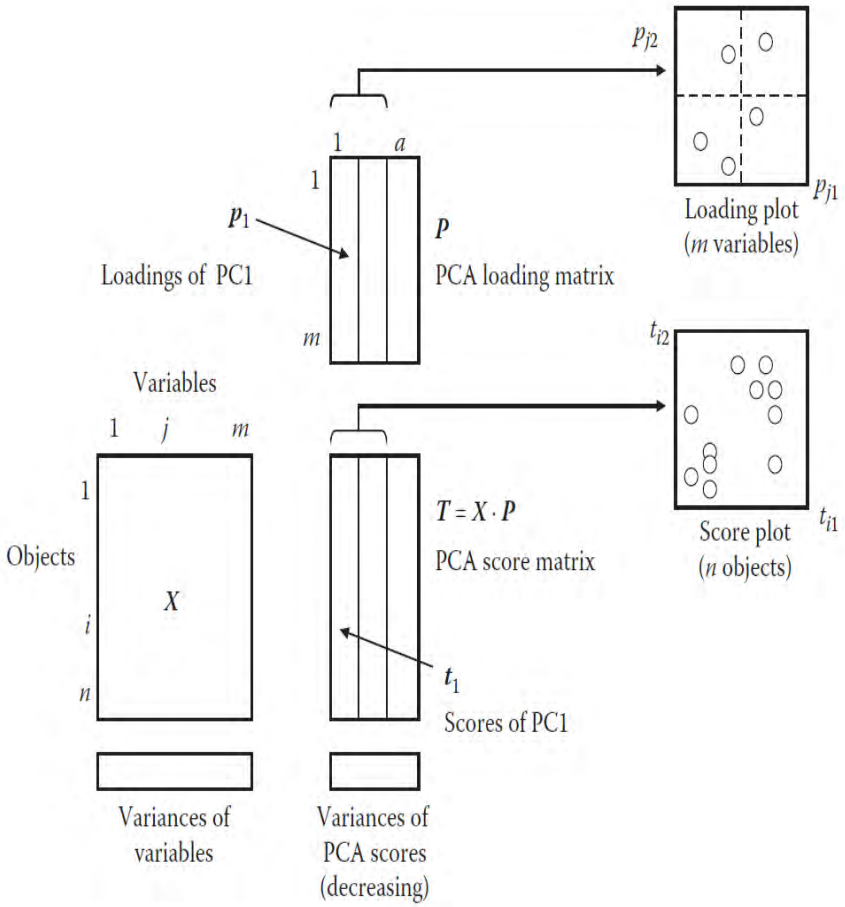


Figure 38. Schéma représentant les matrices de l'ACP (Varmuza and Filzmoser, 2009)

[42] La lettre p (et non b comme précédemment) est généralement utilisée en chimiométrie pour désigner les *loadings* en ACP.

CP1 est une combinaison linéaire des loadings et des variables, tel que

$$\boldsymbol{t_1 = p_{11}x_1 + p_{21}x_2 + \cdots + p_{m1}x_m} \qquad (16)$$

où le score t_1 représente la valeur de CP1 (les coordonnées de projection des n objets dans le nouvel espace de dimension réduit) devant avoir le maximum possible de variance (avec la condition $\boldsymbol{p_1}^{\mathrm{T}}\boldsymbol{p_1} = 1$, car il est commun en chimiométrie de normaliser les longueurs des vecteurs *loading* à 1) et les coefficients p étant les vecteurs *loading*

$$\boldsymbol{p_1 = (p_{11}, \ldots, p_{m1})^{\mathrm{T}}} \qquad (17)$$

Pour la CP2,

$$\boldsymbol{t_2 = p_{12}x_1 + p_{22}x_2 + \cdots + p_{m2}x_m} \qquad (18)$$

avec de nouveau le maximum de variance possible pour t_2 sous la condition $\boldsymbol{p_2}^{\mathrm{T}}\boldsymbol{p_2} = 1$ et la contrainte d'orthogonalité $\boldsymbol{p_1}^{\mathrm{T}}\boldsymbol{p_2} = 0$, où

$$\boldsymbol{p_2 = (p_{12}, \ldots, p_{m2})^{\mathrm{T}}} \qquad (19)$$

Et ainsi de suite pour la création des CPs suivantes, avec la k-ième CP ($3 \leq k \leq m$) définie de telle sorte qu'elle exprime le plus de variance possible sur les scores avec les contraintes que le nouveau vecteur *loading* ait une longueur de 1 et soit orthogonal aux axes précédents (donc, $\boldsymbol{p_j}^{\mathrm{T}}\boldsymbol{p_k} = 0$ avec j, $k = 1,\ldots, m$). Vu que les vecteurs *loading* de toutes les CPs sont orthogonaux les uns par rapport aux autres – comme le sont les axes dans le système de coordonnées original – l'ACP consiste en une rotation du système de coordonnées.

Ainsi, pour un objet i (i allant de 1 à n), défini par un vecteur $\boldsymbol{x_i}$ contenant les variables x_{i1} à x_{im}, le score t_{i1} de CP1 est

$$\boldsymbol{t_{i1} = p_{11}x_{i1} + p_{21}x_{i2} + \cdots + p_{m1}x_{im} = x_i^{\mathrm{T}}.p_1} \qquad (20)$$

Pour tous les *n* objets arrangés en lignes dans la matrice $\boldsymbol{X}$, le vecteur *score* t_1 de CP1 s'obtient par

$$\boldsymbol{t_1 = X.p_1} \quad (21)$$

Tous les vecteurs *loading* sont agencés en colonnes dans la matrice des loadings, $\boldsymbol{P}$, et tous les vecteurs *score* dans la matrice des scores, $\boldsymbol{T}$, tel que

$$\boldsymbol{T = X.P} \quad (22)$$

Finalement, les scores de l'ACP possèdent une propriété mathématique intéressante : ils sont orthogonaux les uns par rapport aux autres, et, sachant que les scores sont habituellement centrés, n'importe quelle paire de vecteurs *score* n'est pas corrélée, résultant en un coefficient de corrélation nul.

$$\boldsymbol{t_j^T . t_k = 0 \quad j,k = 1,\dots,m} \quad (23)$$

Le nombre de CPs à sélectionner est un paramètre important à définir pour la suite des analyses effectuées. Généralement, le choix du nombre de composantes principales à prendre en compte se base sur les variances, exprimées en pourcentage de la variance totale, expliquées par chacune d'elles. Des techniques de validation croisée existent également pour estimer statistiquement le nombre optimal de CPs.

6.7.d Clusterisation

(Varmuza and Filzmoser, 2009)

Principe général

La clusterisation ou le clustering cherche à créer des groupes d'objets similaires (c'est-à-dire, les clusters) alors qu'aucune information a priori sur l'appartenance des objets à tel ou tel groupe n'est disponible et qu'habituellement même pas le nombre de groupes dans les données n'est connu. Il s'agit ainsi d'une méthode exploratoire et *non supervisée* par opposition aux méthodes dites *supervisées* qui nécessitent la connaissance de l'appartenance à un groupe pour chacun des échantillons (au moins pour le jeu de données dit de calibration) (Brereton, 2007; Varmuza and Filzmoser, 2009). Le clustering représente un outil d'investigation puissant, capable de déterminer la structure et d'identifier les groupes sous-jacents de jeux de données alors que cela n'est pas apparent (Adams, 2006). Ainsi, son application, préalable au développement de modèles de classification, est généralement recommandée. Habituellement, des informations a priori sont disponibles quant au nombre de clusters potentiels et à la similarité entre les objets. Alors, le clustering indique si de telles informations sont effectivement reflétées par les données.

Le principe de fonctionnement du clustering est le suivant. Dans un premier temps, des mesures de similarité ou de distance sont calculées entre tous les échantillons du jeu de données caractérisés par n variables (par exemple, les résultats analytiques originaux ou prétraités). Ensuite, à l'aide d'algorithmes de clustering, il s'agit de grouper les échantillons similaires dans un même cluster en s'assurant qu'il y ait une séparation minime entre eux tout en maximisant la séparation entre les différents clusters. Un grand nombre de mesures de similarité ou de distance ainsi que d'algorithmes de clustering existe et leurs choix influencent grandement le groupement final des données, en termes de nombre de clusters et d'appartenance à un cluster.

Les algorithmes de clustering ne reposent pas sur les mêmes principes, ne fonctionnent pas de la même manière et ne grouperont pas nécessairement les objets dans les mêmes clusters. C'est pourquoi, il est généralement recommandé de comparer les résultats obtenus suite à l'utilisation de différentes mesures de similarité ou de distance et d'algorithmes de clustering.

Il est important de souligner qu'en plus des mesures de similarité ou de distance utilisées et des algorithmes de clustering employés, le choix des variables définissant les échantillons ainsi que le prétraitement appliqué sur ces dernières influencent le résultat final.

Métriques de similarité

Généralement, les mesures de distance sont plus utilisées que les mesures de similarité dans le clustering. En particulier, la distance euclidienne représente la mesure la plus utilisée (Varmuza and Filzmoser) et a ainsi été appliquée dans cette recherche. Elle se définit par

$$\boldsymbol{d(Euclid)} = \sqrt{\sum_{j=1}^{m}\left(\boldsymbol{x_{Bj}} - \boldsymbol{x_{Aj}}\right)^2} = \sqrt{\sum_{j=1}^{m}\Delta \boldsymbol{x_j^2}} \qquad (24)$$

avec x_A et x_B les vecteurs définissant les deux objets A et B, et décrits par j variables.

Algorithmes de clustering

Les méthodes de clustering les plus importantes sont :

- les méthodes de partitionnement, où chaque objet est assigné à exactement un groupe ;
- les méthodes hiérarchiques, où les objets sont arrangés de manière hiérarchique sous la forme d'une représentation graphique en arbre, nommée dendrogramme. Ce dernier permet de déterminer visuellement et manuellement le nombre optimal de clusters aussi bien que les relations hiérarchiques entre les différents groupes d'objets ;
- les méthodes de « fuzzy clustering », où chaque objet est assigné à chacun des clusters déterminés à l'aide d'un coefficient d'appartenance. Ces coefficients peuvent être vus comme la probabilité qu'un objet appartienne à n'importe quel cluster ;
- le clustering « model-based », où les différents clusters sont supposés suivre un certain modèle, comme une distribution particulière associée à une moyenne donnée par exemple.

Cette recherche s'intéresse tout particulièrement aux méthodes hiérarchiques qui sont ici implémentées en séquence avec l'ACP.

Le clustering hiérarchique

Les algorithmes standards de clustering hiérarchique produisent une partition des objets ordonnée hiérarchiquement en *k* clusters (*k* allant de 1 à *n*, avec *n* le nombre d'objets). Le clustering hiérarchique peut être ascendant (ou « par agrégation ») ou descendant (ou « par division ») :

- ascendant : au premier niveau de la hiérarchie, chacun des *n* objets forme un cluster séparé, résultant en *n* clusters. Ensuite, les deux clusters les plus proches sont fusionnés et ainsi de suite jusqu'à obtenir tous les objets dans un seul cluster ;
- descendant : au premier niveau de la hiérarchie, tous les *n* objets forment un seul cluster qui est ensuite divisé en deux clusters plus petits et ainsi de suite jusqu'à ce que finalement chaque objet ne forme qu'un cluster séparé.

La Classification Ascendante Hiérarchique (CAH) a été implémentée dans ce travail, celle descendante étant très peu utilisée (Varmuza and Filzmoser, 2009). L'algorithme d'une classification ascendante en utilisant les mesures de distance peut se résumer en 4 étapes (Adams, 2006):

- Etape n°1 : calculer la matrice de distance entre tous les objets (par exemple, calcul de la distance euclidienne) ;
- Etape n°2 : fusionner les objets les plus proches d'après la matrice de distance dans un nouveau cluster ;
- Etape n°3 : calculer une nouvelle matrice de distance entre tous les clusters, en prenant en compte que les clusters produits dans l'étape n°2 auront formé de nouveaux objets et pris la place des données originales ;
- Etape n°4 : procéder de nouveau à l'étape n°2 et ne s'arrêter que lorsqu'il reste un seul cluster.

Comme en témoignent ces étapes, le clustering nécessite aussi bien la mesure de distance ou de similarité entre les objets (dans cette étude, la distance euclidienne) qu'entre les groupes d'objets formant les clusters, pour ainsi déterminer les clusters à fusionner. Ainsi, les nombreuses méthodes d'agrégation diffèrent principalement dans le calcul de la distance entre deux clusters (étape n°3). Les méthodes principales sont le « complete linkage », le « single linkage », l' « average linkage », la « centroïd method » et la « Ward's method ». Dans cette étude, la méthode de mesure implémentée est la méthode de Ward.

La formule générale suivante décrit la mesure de la distance entre clusters, quelle que soit la méthode utilisée pour la calculer (Adams, 2006):

$$\boldsymbol{d_{k(ij)} = \alpha_i d_{ki} + \alpha_j d_{kj} + \beta d_{ij} + \gamma \left| d_{ki} - d_{kj} \right|} \qquad (25)$$

avec $\boldsymbol{d_{ij}}$ la distance entre les objets i et j, $\boldsymbol{d_{k(ij)}}$ la distance entre le cluster k et un nouveau cluster (i,j) formé par la fusion des groupes i et j.

Les valeurs des coefficients α_i, α_j, β et γ sont choisies selon la méthode de mesure inter groupes devant être utilisée, comme l'illustre le Tableau 16 (où n_x représente le nombre d'objets dans le cluster x).

Méthode de mesure	**Coefficients**			
	α_i	α_j	β	γ
Nearest neighbour (single linkage)	0.5	0.5	0	– 0.5
Furthest neighbour (complete linkage)	0.5	0.5	0	0.5
Centroid	$\frac{n_i}{n_i + n_j}$	$\frac{n_i}{n_i + n_j}$	$-\alpha_i . \alpha_j$	0
Median	0.5	0.5	– 0.25	0
Group average	$\frac{n_i}{n_i + n_j}$	$\frac{n_i}{n_i + n_j}$	0	0
Ward's method	$\frac{n_i + n_k}{n_i + n_j + n_k}$	$\frac{n_i + n_k}{n_i + n_j + n_k}$	$\frac{-n_k}{n_i + n_j + n_k}$	0

Tableau 16. Mesures de distance communes utilisées en clustering (Adams, 2006)

Les résultats du clustering hiérarchique sont habituellement illustrés par un dendrogramme (cf. Figure 39). Ce diagramme en arbre présente les objets à la base, sur l'axe des abscisses (les « feuilles ») et les branches qui se rejoignent selon l'ordre donné par l'algorithme d'agrégation. Le dendrogramme montre l'évolution de l'arbre, de la base au sommet, dans l'échelle de la mesure de distance inter clusters (l'axe des ordonnées). L'intérêt d'une telle représentation est de permettre l'identification d'un nombre k de clusters semblant être inhérent à la structure des données. Si tel était le cas, on s'attendrait à avoir k branches clairement visibles sur l'arbre.

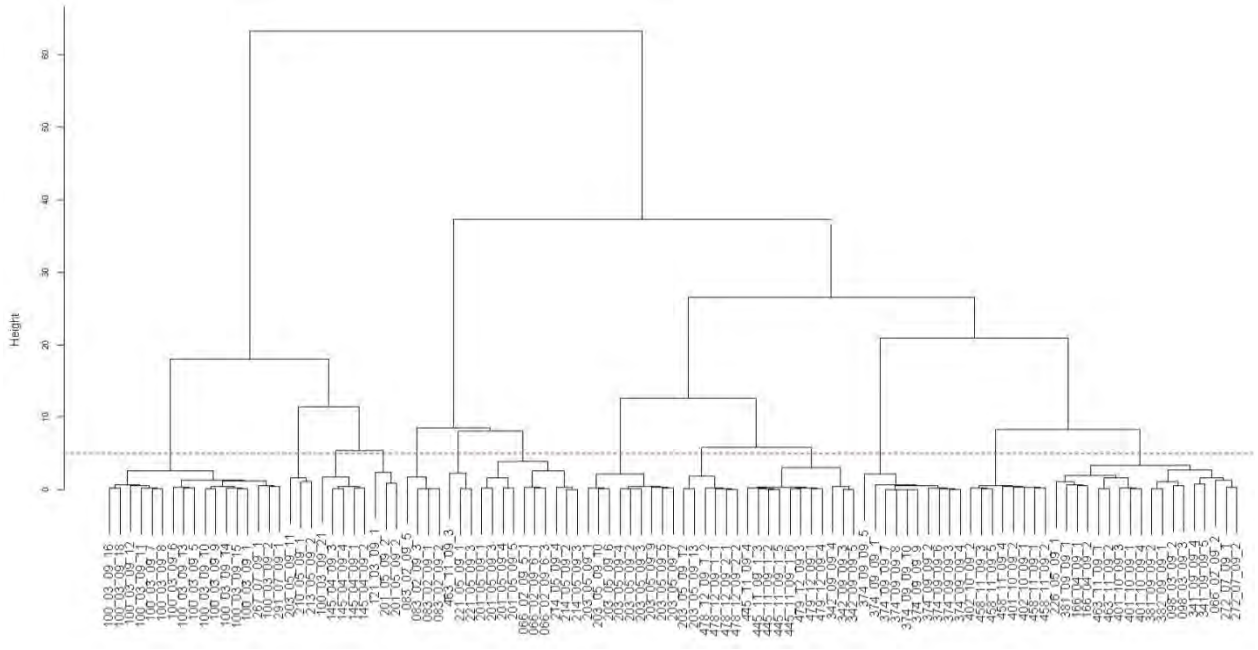


Figure 39. Dendrogramme résultant d'une agrégation avec la méthode Ward appliquée sur des mesures de distance euclidienne entre profils chimiques

Détermination du nombre optimal de clusters

Le point crucial de la plupart des algorithmes de clustering est l'identification correcte du nombre *k* de clusters présents dans le jeu de données (Varmuza and Filzmoser, 2009). En effet, pour déterminer quels objets sont similaires, il faut couper l'arbre à une certaine hauteur et tous les objets connectés sont supposés appartenir au même cluster. Mais l'on voit bien d'après la Figure 39 que selon la hauteur choisie, le nombre de clusters peut être drastiquement différent. Alors, comment choisir ? Comment savoir si les clusters créés correspondent à la « vraie » classe des échantillons ?

La première étape consiste à combiner une information a priori à la visualisation et ainsi d'évaluer la qualité du clustering. Confronter les résultats du clustering aux représentations graphiques des scores et des loadings issus d'une ACP peut ainsi être particulièrement intéressant. Mais quand cela n'est pas possible et étant donné que le nombre exact de clusters est inconnu, une mesure de validité du nombre de clusters doit être calculée afin d'évaluer le groupement obtenu. Il s'agit concrètement de tester différents nombres de clusters et de comparer les résultats par des mesures de validité. Dans la plupart des cas, il n'existe pas un nombre correct mais plutôt un nombre optimal de clusters d'après un certain critère de qualité.

Par exemple, sachant que le but du clustering consiste à trouver des clusters à l'intérieur desquels les objets y seraient autant similaires que possible et des objets entre les différents clusters autant dissimilaires que possible, une mesure de la validité du nombre de clusters peut reposer sur le rapport entre l'homogénéité mesurée au sein des clusters et l'hétérogénéité mesurée entre les clusters. Ainsi, les homogénéités devant être faibles et les hétérogénéités élevées, cette mesure de validité du nombre de clusters devrait être faible. Une représentation graphique de la mesure de validité en fonction du nombre de clusters permet d'indiquer le nombre optimal de clusters. D'autres mesures de validité pourraient suggérer un nombre différent et c'est pourquoi une telle analyse devrait être considérée comme une indication du nombre optimal de clusters tout au plus.

Partie C Résultats et Discussion

Chapitre 7 ACP-CAH : choix d'une valeur de *h* pertinente

L'ACP-CAH représente un outil statistique puissant d'évaluation de l'ajustement de chaque échantillon séparément, une fois celui-ci inséré dans la banque de données de référence. L'ACP-CAH contribue ainsi efficacement à l'étude de la conservation ou non de la structure des classes chimiques au sein de la banque de données après l'ajout de résultats obtenus avec des méthodes différentes (cf. §5.3, sous-hypothèse 2.2). Avec l'ACP-CAH, la réussite de l'ajustement dépend directement de la valeur de *h* fixée. Il est donc primordial de déterminer une valeur *h* d'après laquelle les conclusions tirées quant à la conservation de la structure des données seraient fiables. En d'autres termes, il faut que la valeur de *h* fixée décrive effectivement la conservation ou non de la structure de l'information : c'est-à-dire, la conservation de l'appartenance d'échantillons à une classe chimique précise et ceci pour l'ensemble des profils présents dans la banque de données. Toutefois, de par le mode de calcul de la CAH, la distribution locale des données pourrait avoir une influence sur la hauteur définissant la présence de deux profils dans le même cluster et ainsi dans la réussite de l'ajustement d'un échantillon. Ainsi, la démonstration de l'influence ou non de la distribution locale des données prend tout son sens dans le cadre de la sélection d'une valeur de *h* pertinente.

7.1 Influence de la distribution locale des données dans la réussite de l'ajustement

Pour démontrer l'influence de la distribution locale des données, faisons l'hypothèse qu'une certaine corrélation existe entre une mesure de similarité[43] estimée entre des profils GC-MS et Fast GC-FID ou « GC-MS like » et la réussite de leur ajustement selon une hauteur *h* dans la banque de données de référence. En d'autres termes, on s'attendrait à ce que l'ajustement soit réussi (respectivement « *non réussi* ») pour un profil Fast GC-FID ou « GC-MS like » parce que le coefficient de corrélation de Pearson calculé avec le profil GC-MS est relativement élevé (respectivement *faible*) en regard de la distribution intra- et inter variabilité inter méthodes pour les méthodes considérées (cf. §9.1).

[43] Telle que la mesure du coefficient de corrélation de Pearson, mesure de similarité de référence dans cette étude.

Si le résultat observé ne correspond pas à nos attentes, cela pourrait signifier en particulier que :

1) il n'y a en réalité pas de corrélation entre une valeur de similarité élevée (ou faible) calculée entre un profil GC-MS et un profil obtenu avec une méthode analytique différente (ajusté mathématiquement ou non) et la réussite (ou l'échec) de l'ajustement de ce dernier, pour un spécimen donné ;
2) il y a effectivement une corrélation entre une valeur de similarité faible ou élevée et l'échec ou la réussite de l'ajustement du profil considéré. Toutefois, la réussite de l'ajustement étant estimée dans le dendrogramme local, constitué d'environ les 30 échantillons les plus proches (cf. Chapitre 6), la réussite ou non de l'ajustement pour une hauteur *h* donnée pourrait dépendre de la distribution locale des profils chimiques. En particulier, l'ajustement pourrait dépendre de la présence de profils dont la similarité avec le profil GC-MS est plus élevée que ne l'est celle du profil considéré (Fast GC-FID ou « GC-MS like » dans notre exemple). Par exemple, pour une valeur de *h* relativement faible, plus la région proche du profil GC-MS serait dense, plus l'ajustement du profil du même spécimen mais obtenu avec une méthode différente serait difficile. Ces éléments de réflexion peuvent ainsi expliquer pourquoi l'ajustement d'un profil obtenu en Fast GC-FID par exemple, ajusté mathématiquement ou non, ne soit pas estimé comme réussi pour une certaine valeur de *h* alors que le profil présente une mesure du coefficient de corrélation de Pearson élevée avec le profil GC-MS du spécimen considéré.

L'étude de l'existence ou non d'une certaine corrélation entre une valeur de coefficient de corrélation de Pearson faible ou élevée calculée entre un profil GC-MS et un profil obtenu avec une méthode analytique différente et la réussite de leur ajustement, en regard de la hauteur *h* fixée, est abordée au §7.3.

En partant de l'a priori qu'une certaine corrélation existe, pour illustrer l'influence de la distribution locale des données dans la réussite ou non de l'ajustement il convient dans un premier temps d'étudier les valeurs de coefficient de corrélation de Pearson calculées entre les couples de profils chimiques GC-MS et Fast GC-FID ou « GC-MS like » pour chaque échantillon analysé. Les valeurs médianes calculées d'après l'ensemble des valeurs obtenues pour les 100 itérations à chaque valeur de *h*, pour les comparaisons entre les profils GC-MS et Fast GC-FID ou « GC-MS like » de chaque spécimen, sont présentées au Tableau 33 (cf. Chapitre 10, §10.2) [44].

L'examen de ces valeurs permet de remarquer qu'une fois ajusté mathématiquement, l'ajustement du spécimen 258_06_09_3 est toujours réussi (cf. Tableau 33, page 242). La valeur du coefficient de corrélation de Pearson est particulièrement élevée mais cela n'explique pas à elle seule la réussite de l'ajustement. D'autres spécimens présentent en effet des valeurs de similarité comparables voire plus élevées mais l'ajustement est pourtant considéré comme « non réussi » (par exemple, le spécimen 210_05_09_2, cf. Tableau 33). Comme cela a été discuté ci-dessus, en plus de la hauteur *h* considérée et de l'existence d'une valeur de similarité relativement faible ou élevée entre les profils, la distribution des données dans le dendrogramme local pourrait influencer la réussite ou non de l'ajustement.

Pour illustrer l'influence de la distribution des données dans le dendrogramme local, quatre spécimens sont particulièrement étudiés :

- le spécimen 066_02_09_6_2, pour qui l'ajustement peut être réussi, malgré une valeur de coefficient de corrélation de Pearson relativement faible (cf. Tableau 33);
- le spécimen 258_06_09_3, en raison de son ajustement toujours réussi (cf. Tableau 33);
- le spécimen 210_05_09_2, présentant une valeur de coefficient de corrélation de Pearson plus élevée lors de la comparaison des profils GC-MS et « GC-MS like » que celle du spécimen 258_06_09_3 mais dont l'ajustement n'est pas toujours réussi (cf. Tableau 33);
- le spécimen 267_07_09_1, qui même lorsque des hauteurs élevées sont fixées ne présente pas toujours un ajustement réussi (cf. Tableau 33).

[44] Le gain en similarité statistique après ajustement mathématique illustré par les valeurs du coefficient de corrélation de Pearson ou les hauteurs *h* calculées sera discuté en détail au Chapitre 10.

Le Tableau 17 présente les valeurs de coefficient de corrélation de Pearson calculées lors de la comparaison des profils GC-MS et « GC-MS like » pour les spécimens correspondants, pour un set de validation bien particulier. De plus, le Tableau 17 contient les hauteurs *h* pour lesquelles les profils se retrouvent dans le même cluster dans le dendrogramme local, lors de l'ajustement du profil « GC-MS like » dans la banque de données de référence GC-MS.

Spécimens	Coefficient de corrélation de Pearson	Hauteur h
066_02_09_6_2	75.71	2.08
258_06_09_3	99.59	0.19
210_05_09_2	99.84	1.63
267_07_09_1	98.40	26.93

Tableau 17. Coefficient de corrélation de Pearson et hauteur *h* dans le dendrogramme local estimés entre les profils GC-MS et « GC-MS like » des spécimens 066, 210, 258 et 267 pour un set de validation

L'exemple du spécimen 066_02_09_6_2 illustre clairement la possibilité d'obtenir un ajustement réussi pour de faibles hauteurs dans le cas où, pourtant, une valeur de coefficient de corrélation de Pearson relativement faible est calculée entre les profils GC-MS et « GC-MS like ». La Figure 40 représente la distribution des spécimens proches des profils GC-MS et « GC-MS like » (ici, le profil Fast GC-FID est ajusté mathématiquement selon le modèle linéaire) à l'aide de leurs scores sur CP1 et CP2. La Figure 41 représente le dendrogramme local pour ces mêmes spécimens après l'application de la méthode de groupement *Ward* sur les distances euclidiennes calculées entre les profils.

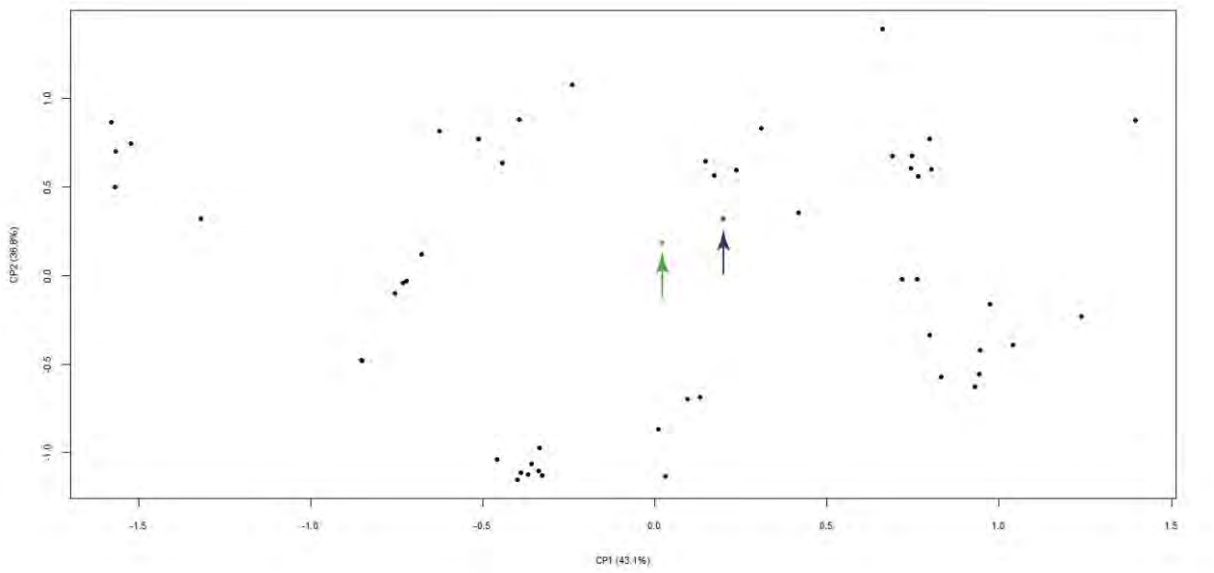


Figure 40. Identification, à l'aide de CP1 et CP2 dans le sous-échantillonnage des spécimens proches, des profils chimiques GC-MS (en vert) et "GC-MS like" (en bleu) pour le spécimen 066_02_09_6_2

D'après leur distribution sur CP1 et CP2, les profils GC-MS et « GC-MS like » sont proches bien que la valeur de coefficient de corrélation de Pearson calculée soit relativement faible. Malgré cette faible valeur de similarité, en raison d'une densité d'échantillons faible aux alentours du profil GC-MS ainsi que de l'absence d'échantillons très similaires au profil GC-MS, les deux profils se retrouvent dans le même cluster lorsque le dendrogramme est coupé à une hauteur de 2.08 (cf. Figure 41).

Cette hauteur est relativement faible en comparaison de celle calculée lors de l'ajustement du profil « GC-MS like » du spécimen 267_07_09_1, où pour un coefficient de corrélation de Pearson de 98.40, une hauteur *h* de 26.93 sépare les profils «GC-MS et « GC-MS like » (l'étude du spécimen 267_07_09_1 est abordée ci-dessous, cf. Tableau 17 ci-dessus et Figure 47 ci-dessous).

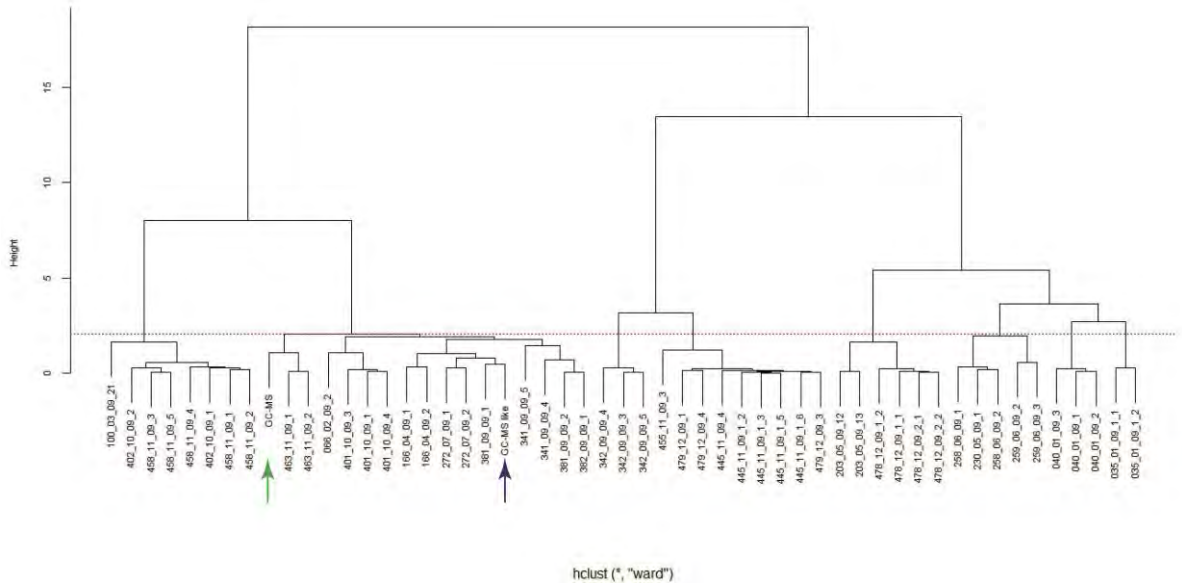


Figure 41. Identification dans le dendrogramme local (c'est-à-dire, dans le sous-échantillonnage des spécimens proches) de la présence dans le même cluster ou non des profils chimiques GC-MS et « GC-MS like » du spécimen 066_02_09_6_2

La Figure 42 représente la distribution des profils GC-MS et « GC-MS like » du spécimen 258_06_09_3 dans le sous-échantillonnage des spécimens proches. D'après les résultats du Tableau 17, sans surprise, les profils GC-MS et « GC-MS like » sont particulièrement proches d'après les scores sur CP1 et CP2.

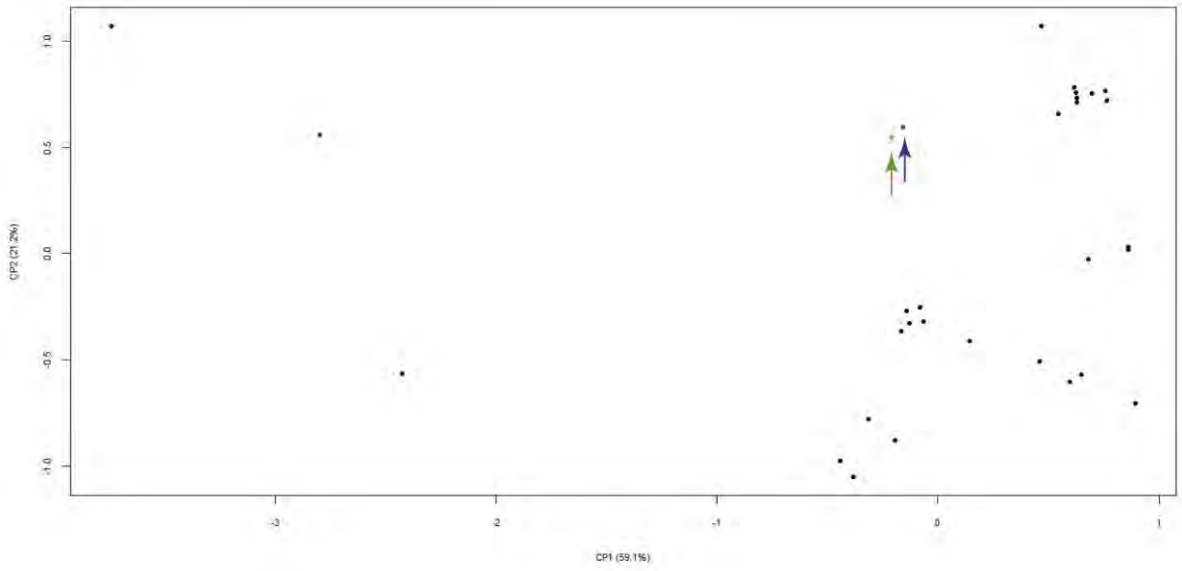


Figure 42. Identification, à l'aide de CP1 et CP2 dans le sous-échantillonnage des spécimens proches, des profils chimiques GC-MS (en vert) et "GC-MS like" (en bleu) pour le spécimen 258_06_09_3

La Figure 43 compare la distance dans le dendrogramme local séparant le profil GC-MS du profil « GC-MS like ». Aucun autre profil ne présentant une similarité avec le profil GC-MS plus élevée que celle du profil « GC-MS like » et la densité d'échantillons à proximité du profil GC-MS étant relativement faible, le profil « GC-MS like » se retrouve effectivement toujours dans le même cluster que le profil GC-MS, et par conséquent son ajustement est toujours réussi même pour une hauteur de 0.5 (cf. Figure 43).

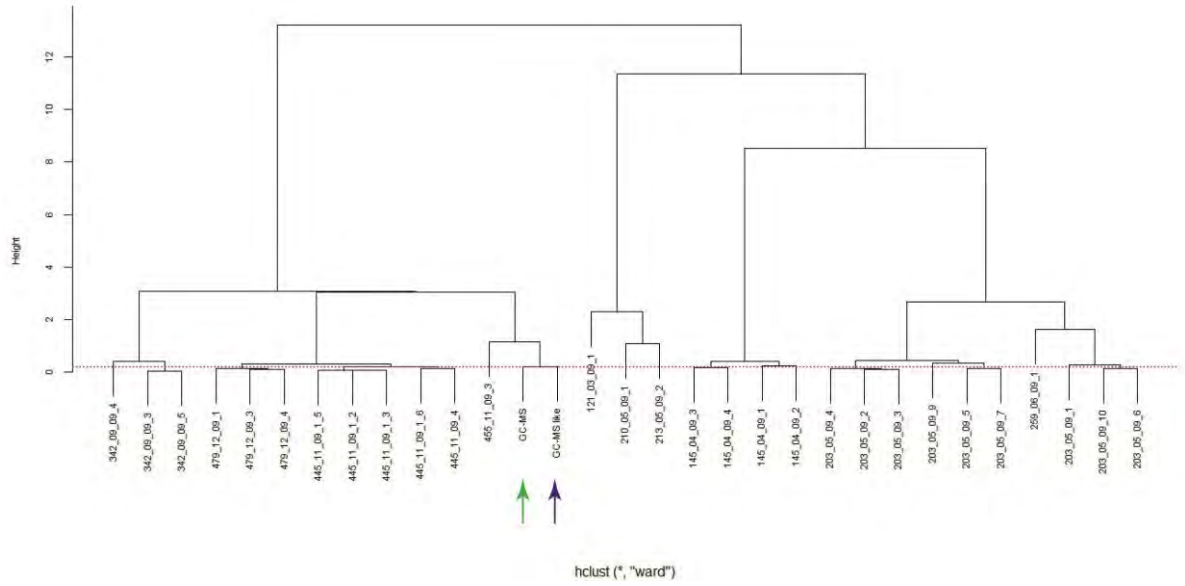


Figure 43. Identification dans le dendrogramme local (c'est-à-dire, dans le sous-échantillonnage des spécimens proches) de la présence dans le même cluster ou non des profils chimiques GC-MS et "GC-MS like" du spécimen 258_06_09_3

La Figure 44 illustre la distribution des profils GC-MS et « GC-MS like » du spécimen 210_05_09_2 tandis que la Figure 45 compare dans le dendrogramme local la distance séparant les profils GC-MS et « GC-MS like ».

Dans cet exemple, il est particulièrement intéressant de relever la présence de deux profils particulièrement similaires au profil GC-MS, d'après leurs scores sur CP1 et CP2. Il s'agit des profils des spécimens 121_03_09_1 et 210_05_09_2.

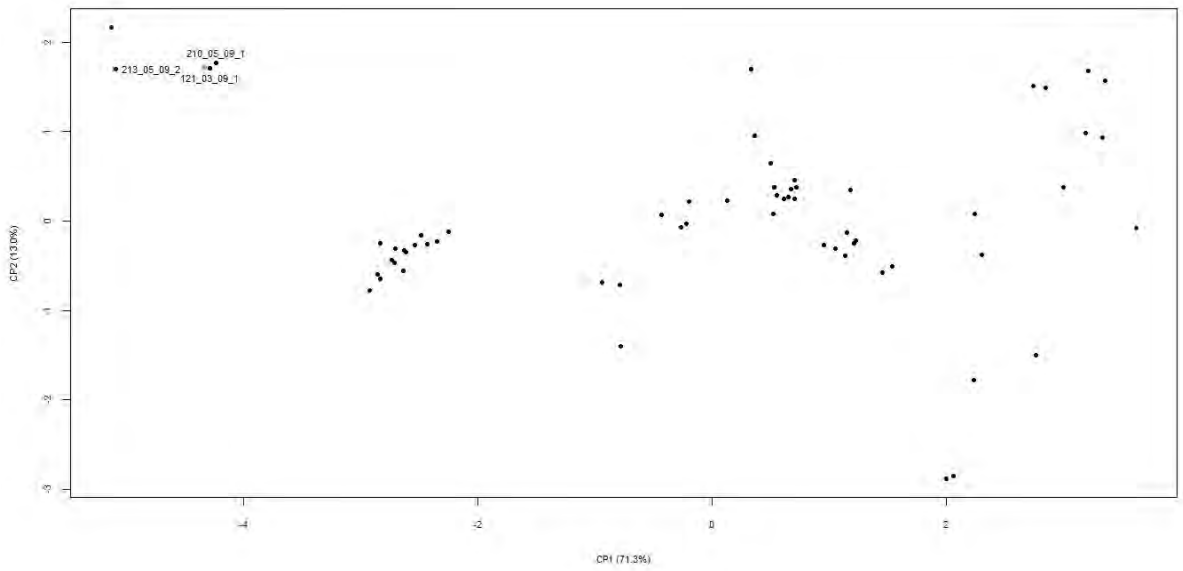


Figure 44. Identification, à l'aide de CP1 et CP2 dans le sous-échantillonnage des spécimens proches, des profils chimiques GC-MS (en vert) et "GC-MS like" (en bleu) pour le spécimen 210_05_09_2

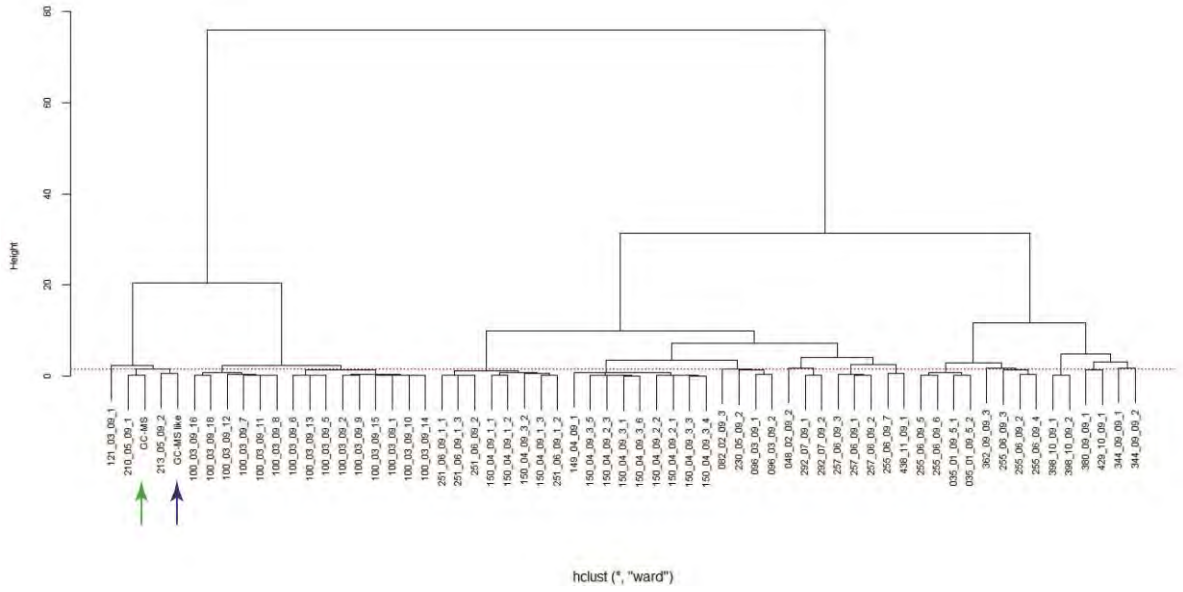


Figure 45. Identification dans le dendrogramme local (c'est-à-dire, dans le sous-échantillonnage des spécimens proches) de la présence dans le même cluster ou non des profils chimiques GC-MS et "GC-MS like" du spécimen 210_05_09_2

Pour les 3 spécimens les plus proches du profil GC-MS, le Tableau 18 présente les distances euclidiennes calculées entre les profils définis par les scores des 4 premières CPs.

	Distance Euclidienne			
Spécimens	GC-MS	« GC-MS like »	121_03_09_1	210_05_09_1
« GC-MS like »	0.96			
121_03_09_1	1.45	1.91		
210_05_09_1	0.14	1.05	1.46	
213_05_09_2	0.99	0.68	2.23	1.08

Tableau 18. Matrice de distance calculée entre les profils des spécimens 210_05_09_2 (GC-MS et « GC-MS like »), 121_03_09_1, 210_05_09_1 et 213_05_09_2

D'après ces valeurs, le profil le plus proche du profil GC-MS obtenu pour le spécimen 210_05_09_2 correspond au premier spécimen de cette saisie, le spécimen 210_05_09_1 (cf. Tableau 18). Ensuite, si l'on retire ces deux spécimens, les deux profils les plus proches correspondent au profil « GC-MS like » du spécimen 210_05_09_2 et au profil GC-MS du spécimen 213_05_09_2 (cf. Tableau 18). Par conséquent, sachant que le profil du 210_05_09_1 est plus proche du profil GC-MS que ne l'est le profil « GC-MS like » et étant donné le mode de fonctionnement des algorithmes de groupement en CAH, le profil « GC-MS like » ne peut dès lors pas se trouver dans le même cluster que le profil GC-MS à la hauteur *h* la plus faible. Ainsi, l'ajustement du profil « GC-MS like » ne sera pas considéré comme réussi pour de telles hauteurs, bien qu'il présente un coefficient de corrélation de Pearson élevé avec le profil GC-MS (valeur d'ailleurs plus élevée que celle existant pour le spécimen 258_06_09_3 dont l'ajustement est toujours réussi, cf. Tableau 17).

Pour ces mêmes spécimens l'examen des valeurs de corrélation calculées entre les profils définis par les valeurs prétraitées des 6 variables, conduit aux mêmes observations que l'examen des mesures de distance sur les profils définis par les scores sur les 4 premières CPs (cf. Tableau 18 et Tableau 19). En particulier, les profils GC-MS et « GC-MS like » sont effectivement similaires, mais la similarité existant entre le profil du spécimen 210_05_09_1 et le profil GC-MS est légèrement plus grande (cf. Tableau 19).

	Coefficient de corrélation de Pearson				
Spécimens	GC-MS	« GC-MS like »	121_03_09_1	210_05_09_1	213_05_09_2
GC-MS	100.00				
« GC-MS like »	99.84	100.00			
121_03_09_1	93.31	93.56	100.00		
210_05_09_1	99.89	99.68	92.38	100.00	
213_05_09_2	99.67	99.57	90.75	99.55	100.00

Tableau 19. Matrice de similarité calculée entre les profils des spécimens 210_05_09_2 (GC-MS et « GC-MS like »), 121_03_09_1, 210_05_09_1 et 213_05_09_2

Par conséquent, l'ajustement du profil « GC-MS like » du spécimen 210_05_09_2 ne sera pas réussi pour des hauteurs de moins de 1.63, dans cet exemple (cf. Tableau 17). Cependant, le calcul du coefficient de corrélation de Pearson entre le profil GC-MS et le profil « GC-MS like » fournira une valeur relativement élevée.

La Figure 46 et la Figure 47 concernent le spécimen 267_07_09_1. Comme le présentent ces graphiques, il s'agit là d'un cas de forte densité d'échantillons dans le voisinage proche du profil GC-MS. L'appartenance des profils GC-MS et « GC-MS like » au même cluster que ce soit à des hauteurs faibles ou élevées s'avère ainsi plus difficile à accomplir, pour les mêmes raisons que celles explicitées pour le spécimen 210_05_09_2 et bien que le coefficient de corrélation de Pearson soit relativement élevé (cf. Tableau 17).

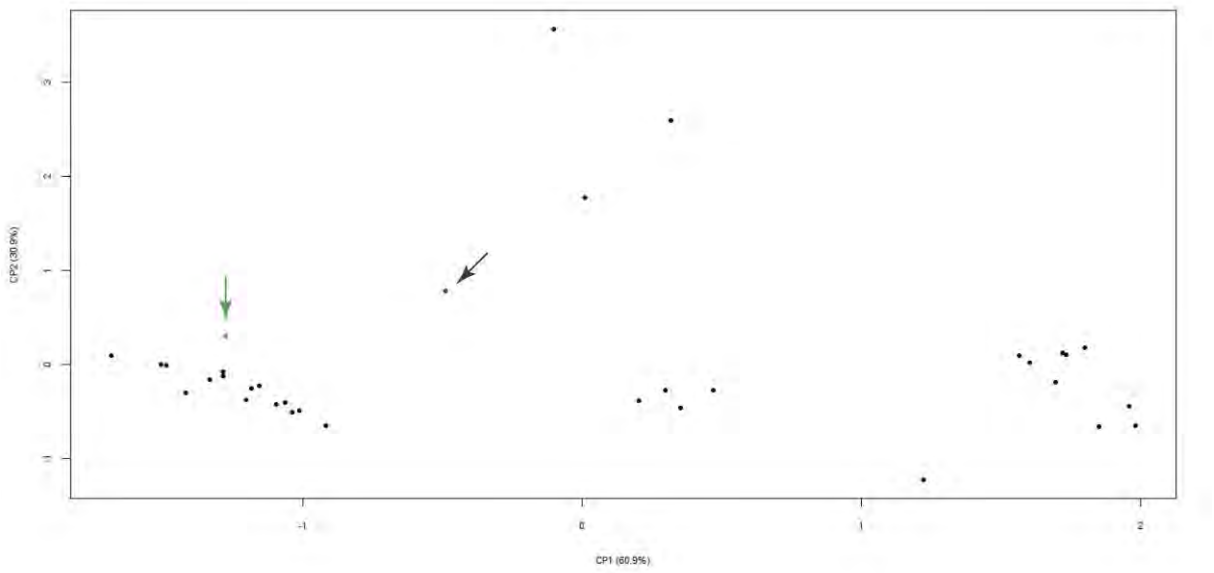


Figure 46. Identification à l'aide de CP1 et CP2 dans le sous-échantillonnage des spécimens proches des profils chimiques GC-MS (en vert) et "GC-MS like" (en bleu) pour le spécimen 267_07_09_1

L'examen du profil Fast GC-FID non ajusté mathématiquement démontre que dans ce cas précis l'ajustement mathématique influe de manière négligeable sur la distance séparant les profils GC-MS et Fast GC-FID ou « GC-MS like ». Alors qu'une distance de 27.5 sépare les profils GC-MS et Fast GC-FID, une distance de 26.93 sépare les profils GC-MS et « GC-MS like » dans le dendrogramme local, illustrant tout l'impact de la densité des données. Quant à lui, le coefficient de corrélation de Pearson passe de 95.16 à 98.40 après ajustement mathématique. En regard de la distribution inter méthodes pour la combinaison de résultats GC-MS et « GC-MS like » (ajustement mathématique selon le modèle linéaire) (cf. §7.3), la valeur de corrélation de 98.40 mesurée entre les deux profils implique certainement une similarité chimique entre ces derniers étant donné que pour un seuil de décision fixé à 98%, il n'y a environ que 0.6% de FP (avec VP égal à environ 30%).

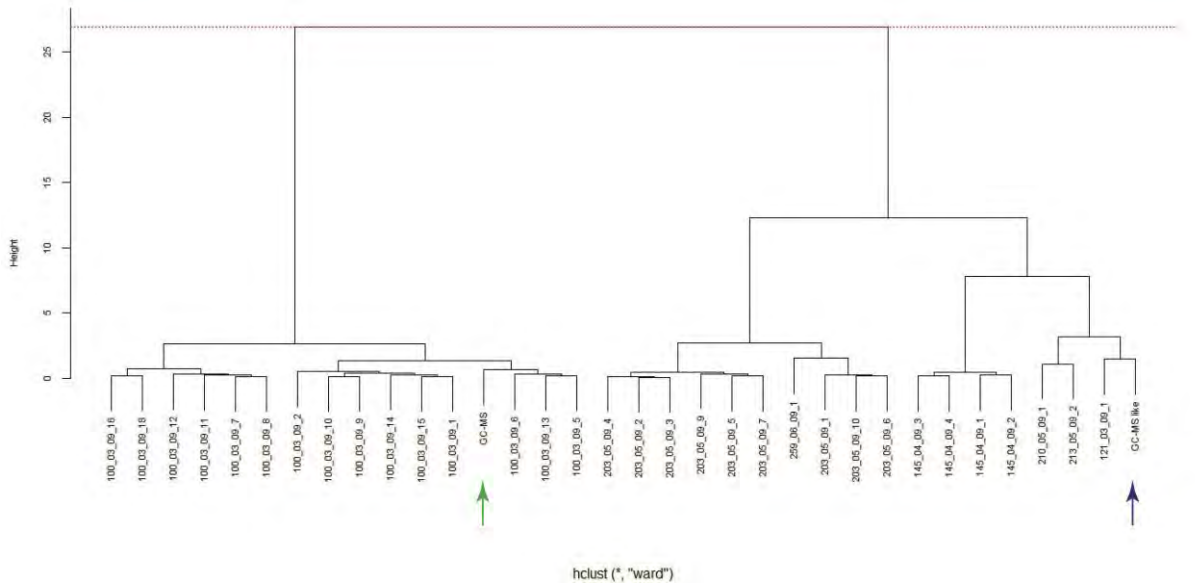


Figure 47. Identification dans le dendrogramme local (c'est-à-dire, dans le sous-échantillonnage des spécimens proches) de la présence dans le même cluster ou non des profils chimiques GC-MS et "GC-MS like" du spécimen 267_07_09_1

Ainsi, au travers de ces quatre exemples, l'influence de la distribution locale des données dans la réussite ou non de l'ajustement de profils a été clairement démontrée. Ce résultat n'a rien de surprenant d'une part en raison du mode de fonctionnement de la CAH et d'autre part dans la mesure où la réussite de l'ajustement dans le dendrogramme local se base sur la présence des deux profils considérés dans le même cluster selon une hauteur *h* (cf. §6.4.e). La présence de profils dont la similarité avec le profil GC-MS est plus élevée que ne l'est celle du profil Fast GC-FID ou « GC-MS like » influencera donc directement la présence ou non de ce dernier dans le même cluster que le profil GC-MS, en regard d'une certaine valeur de *h*. A l'inverse, si la densité d'échantillons autour d'un profil GC-MS est faible, alors le profil Fast GC-FID ou « GC-MS like » du spécimen correspondant aura plus de chances de se retrouver dans le même cluster à des hauteurs *h* faibles, bien qu'une valeur de coefficient de corrélation de Pearson relativement faible soit calculée entre les profils GC-MS et Fast GC-FID ou « GC-MS like ».

L'influence de la distribution locale des données dans l'estimation de la performance d'ajustement n'est pas une surprise considérant la distribution des profils chimiques au sein de la banque de données. Celle-ci est en effet constituée de plusieurs sous-ensembles (les classes chimiques) ayant des distributions et densités de population propres et potentiellement bien différentes les uns des autres. Toutefois, cette observation n'implique pas que l'ACP-CAH ne soit pas utile pour évaluer la réussite de l'ajustement d'un échantillon dans la banque de données de référence. Que des profils GC-MS et « GC-MS like » ne soient pas dans le même cluster d'après une hauteur de *h* faible n'implique pas que la structure des classes chimiques au sein de la banque de données ne soit pas conservée. En effet, tout dépend de la valeur *h* au-delà de laquelle la structure des classes chimiques n'est plus conservée (cf. §7.4). En d'autres termes, fixer une hauteur *h* trop faible n'est pas forcément pertinent pour juger de la conservation de la structure des classes chimiques au sein de la banque de données (mais permet par exemple de décrire efficacement l'influence de la similarité analytique sur la similarité statistique, cf. Chapitre 8, §8.2).

En conséquence, pour se prononcer quant à la conservation de la structure de l'ensemble des classes chimiques au sein de la banque de données, la fixation d'une seule valeur de *h* et le seul recours à l'ACP-CAH ne sont pas souhaitables en raison de l'influence de la distribution locale des données. L'emploi d'une autre mesure de similarité est alors préférable pour garantir l'ajustement correct des résultats – ou l'appartenance à une même classe chimique – dans la banque de données de référence. Rappelons en effet que la pérennité de la structure de l'information au sein de la banque de données, malgré l'ajout de résultats obtenus avec une méthode différente, doit être assurée. Cela implique que, lorsque l'ajustement entre les profils est estimé réussi d'après la valeur de *h*, les profils appartiennent bel et bien à la même classe chimique (c'est-à-dire que dans le même temps les profils sont similaires d'après la mesure du coefficient de corrélation de Pearson en regard du seuil de décision défini lors de l'étude de l'intra- et l'inter variabilité inter méthodes).

7.2 Étude de l'efficacité d'une méthodologie de profilage basée sur l'ACP-CAH

Dans cette recherche, le calcul du coefficient de corrélation de Pearson pour établir les distributions des populations d'intra- et d'inter variabilité représente la mesure de similarité de référence. La performance de discrimination entre ces deux populations permet d'évaluer l'efficacité de la méthodologie de profilage dans son ensemble grâce au calcul de taux d'erreurs et de seuils de décision. Sachant que lors de l'implémentation d'une méthodologie de profilage chimique n'importe quelle mesure de similarité peut être utilisée (cf. Chapitre 1), et que cette recherche propose de déterminer la réussite de l'ajustement à l'aide de l'ACP-CAH, il semble justifié d'investiguer la performance de la discrimination de l'intra- et l'inter variabilité lorsque la mesure de similarité consiste en une CAH. Dans le cadre de l'étude inter méthodes, une observation particulière des hauteurs existant entre les profils d'un même spécimen alors que ceux-ci sont acquis avec des méthodes analytiques différentes a été effectuée.

Pour ce faire, le mode de calcul inhérent au processus ACP-CAH développé dans cette recherche est appliqué. Ainsi, dans un premier temps, une ACP est appliquée sur la BDD de référence réduite (c'est-à-dire, sans l'échantillonnage sélectionné). Ensuite, l'échantillonnage est transposé dans le domaine défini par les 4 premières composantes principales issues de l'ACP expliquant le 95% de variance du jeu de données. Finalement, pour chaque spécimen de l'échantillonnage sélectionné (cf. Chapitre 6, §6.4.a), une CAH sur les scores ACP des profils chimiques obtenus avec des méthodes analytiques différentes a été effectuée (méthode de groupement *Ward* sur les distances euclidiennes). En particulier, la Figure 48 et la Figure 49 représentent les dendrogrammes obtenus pour la comparaison des résultats issus des méthodes analytiques GC-MS et Fast GC-FID dans le cadre du scénario d'ajustement 2.1, selon que l'échantillon fasse partie de l'intra variabilité (cf. Figure 48, 9 réplicats par spécimen) ou de l'inter variabilité (cf. Figure 49, 3 réplicats par spécimen).

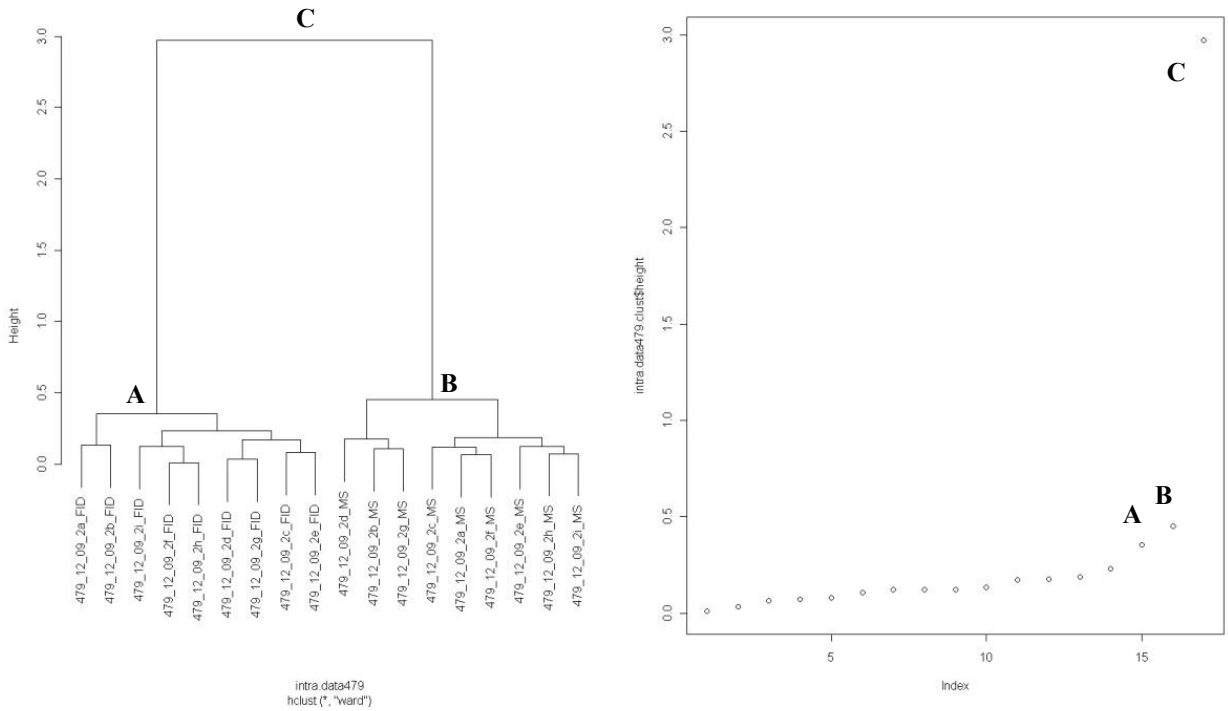


Figure 48. Dendrogramme de la comparaison des profils obtenus avec des méthodes différentes pour un même spécimen de l'intra variabilité

Sur la base d'un tel mode de calcul, trois hauteurs différentes, notées A, B et C peuvent être déterminées dans le dendrogramme produit (cf. Figure 48 et Figure 49) :

- A et B définissent à quel point des profils chimiques obtenus avec une même méthode analytique sont similaires, respectivement ;
- C mesure quant à elle la similarité statistique entre des profils chimiques d'un même spécimen obtenus avec des méthodes analytiques différentes.

Pour dresser les populations d'intra- et d'inter variabilité dans le cadre des études intra et inter méthodes le procédé est alors similaire à celui décrit au Chapitre 6 :

- l'intra variabilité intra méthode se compose des hauteurs calculées lors de la comparaison de tous les réplicats pour chaque spécimen constituant la population de l'intra variabilité (cf. Figure 48, ensemble des hauteurs inférieures ou égales à A ou B, selon la méthode considérée);
- l'inter variabilité intra méthode comprend les hauteurs correspondant à la comparaison de tous les spécimens constituant la population de l'inter variabilité ;
- l'intra variabilité inter méthodes se compose des hauteurs correspondant à la comparaison des profils chimiques GC-MS et « GC-MS like » (c'est-à-dire, les données Fast GC-FID après ajustement mathématique), pour tous les réplicats de chacun des spécimens considérés (cf. Figure 48, hauteur C) ;
- l'inter variabilité inter méthodes correspond aux hauteurs calculées lors de la comparaison des différents spécimens, en choisissant alternativement 1 profil GC-MS et 1 profil « GC-MS like » parmi les 28 spécimens analysés qui constituent la population de l'inter variabilité (cf. Figure 50).

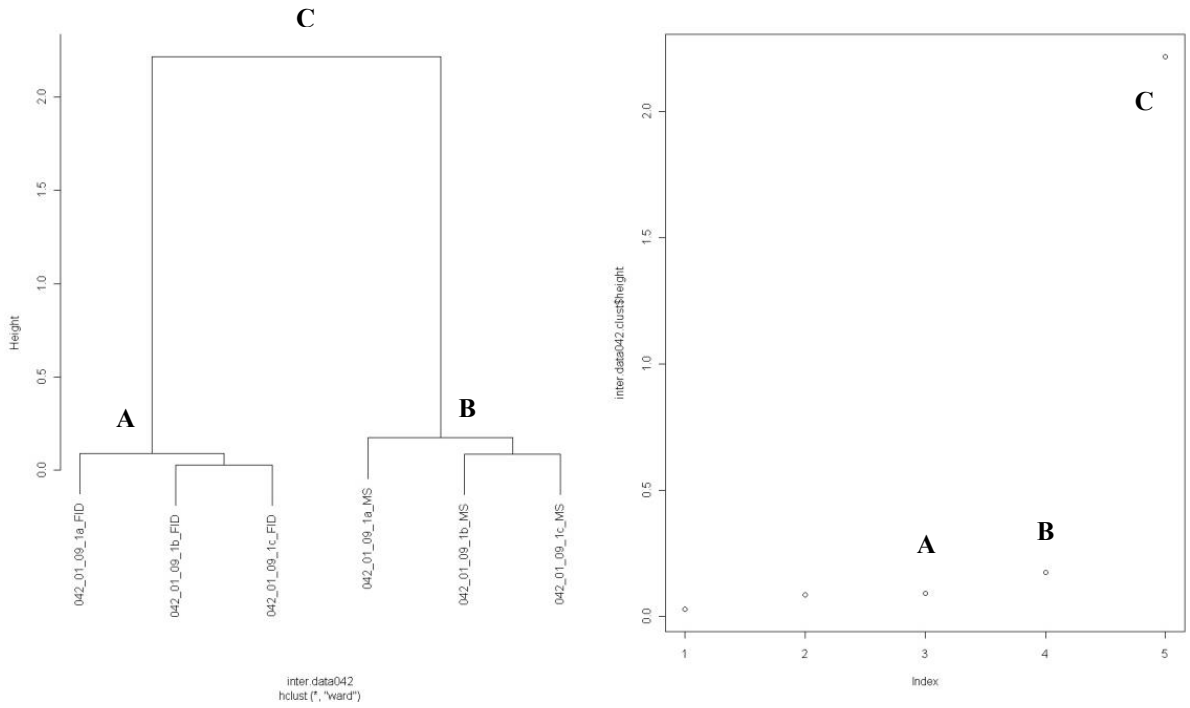


Figure 49. Dendrogramme de la comparaison des profils obtenus avec des méthodes différentes pour un même spécimen de l'inter variabilité

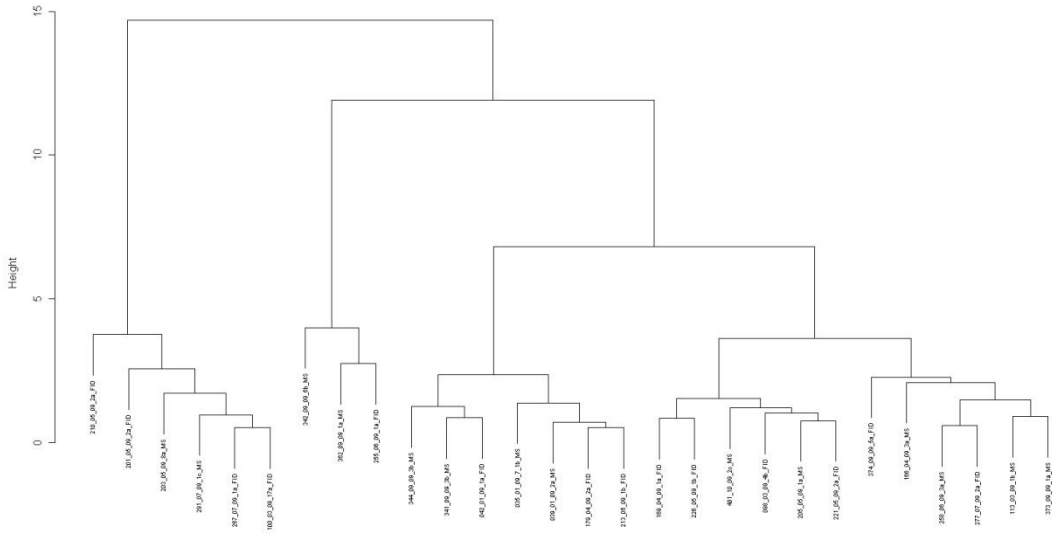


Figure 50. Dendrogramme de la comparaison des profils obtenus avec des méthodes différentes pour dresser l'inter variabilité inter méthodes

La Figure 51 représente la distribution des populations de l'intra et de l'inter variabilité pour l'étude intra méthode de la méthode de référence GC-MS avec une telle mesure de similarité.

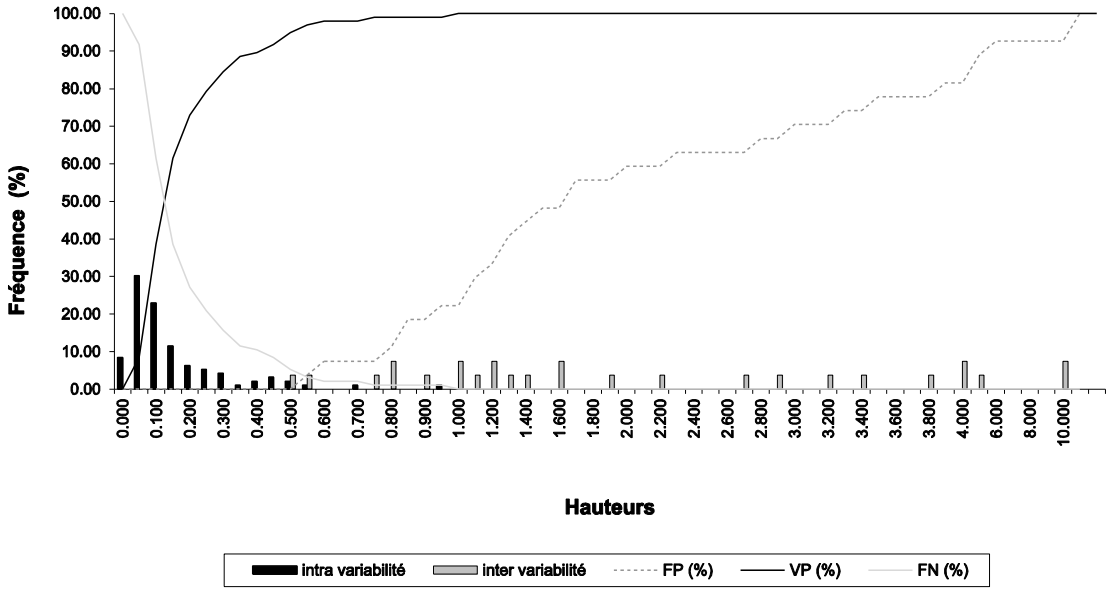


Figure 51. Distribution de l'intra- et l'inter variabilité pour la méthode GC-MS. Pour une question de clarté, l'échelle des valeurs de la hauteur *h* n'est pas linéaire.

Il apparaît important de souligner qu'avec un tel procédé peu de valeurs sont estimées en raison du mode de calcul des méthodes de groupement hiérarchique (pour l'étude intra méthode, 96 et 27 valeurs composent l'intra- et l'inter variabilité, respectivement, tandis que pour l'étude inter méthodes, 11 et 27 valeurs composent l'intra- et l'inter variabilité, respectivement). Se pose alors la question de la pertinence de ce procédé pour déterminer une hauteur *h* faisant office de seuil de décision dans l'évaluation de la similarité d'un couple de profils chimiques et déterminer par conséquent si l'ajustement est réussi dans l'étape locale de l'ACP-CAH.

Au-delà du problème mentionné ci-dessus et relatif au faible nombre de données composant les deux populations, l'approche ACP-CAH présente une efficacité intéressante dans la séparation des populations de l'intra et de l'inter variabilité intra méthode (pour la méthode GC-MS, cf. Tableau 20). En effet, les valeurs composant l'intra variabilité se concentrent vers les faibles valeurs de distance tandis que celles constituant l'inter variabilité se distribuent sur un large intervalle de valeurs de distance plus élevées (cf. Figure 51).

		VP (%)	VN (%)	FP (%)	FN (%)
Seuil	0.5	94.8	100.0	0	5.2
	0.75	99.0	92.6	7.4	1.0

Tableau 20. Performance de la discrimination pour le profilage chimique de l'héroïne par GC-MS lorsque la mesure de la similarité repose sur une ACP-CAH

Les valeurs de seuil déterminées dans le Tableau 20 ne peuvent toutefois pas être utilisées pour déterminer une valeur de *h* sur la base de laquelle juger de la conservation de la structure des classes chimiques dans la banque de données ni être appliquées en tant que telles dans l'interprétation des performances d'ajustement (cf. Chapitre 8, §8.2). En effet, l'influence de la structure locale des données dans l'estimation de la réussite de l'ajustement démontrée auparavant (cf. §7.1) n'est ici pas prise en compte. De plus, avec ces résultats, c'est spécifiquement l'efficacité de l'ACP-CAH pour discriminer les populations de l'intra- et l'inter variabilité qui est estimée, dans le cadre de l'étude intra méthode (alors que lors de l'ajustement, on combine les résultats GC-MS et « GC-MS like », par exemple).

Toutefois, la distribution des valeurs pour l'intra- et l'inter variabilité de la méthode GC-MS pourrait s'avérer utile pour estimer la similarité statistique de résultats obtenus avec des méthodes analytiques différentes. En effet, si deux profils GC-MS et « GC-MS like » par exemple se retrouvent dans le même cluster dans la banque de données à une hauteur de 0.5, alors la structure de l'information dans la banque de données est certainement conservée car la structure locale de la banque de données ne pourrait probablement pas les rapprocher autant. Ainsi, les performances d'ajustement particulièrement élevées calculées pour les scénarios d'ajustement 1.2 et 1.4 sont significatives et révélatrices de la similarité statistique des résultats (cf. §8.2).

La Figure 52 représente la distribution des populations de l'intra et de l'inter variabilité pour l'étude inter méthodes toujours dans le cadre du scénario d'ajustement 2.1. Comme l'illustre la Figure 52, lorsque les données GC-MS et « GC-MS like » sont combinées, le recouvrement entre les deux populations est important. En effet, en comparaison avec les distributions obtenues avec la méthode de référence (cf. Figure 51), les valeurs de l'intra variabilité se déplacent particulièrement vers les hauteurs élevées. Ceci révèle la variabilité existant entre des résultats obtenus avec des méthodes GC-MS et Fast GC-FID qui sont différentes en particulier dans le paramètre analytique A_{DET} (cf. Figure 52).

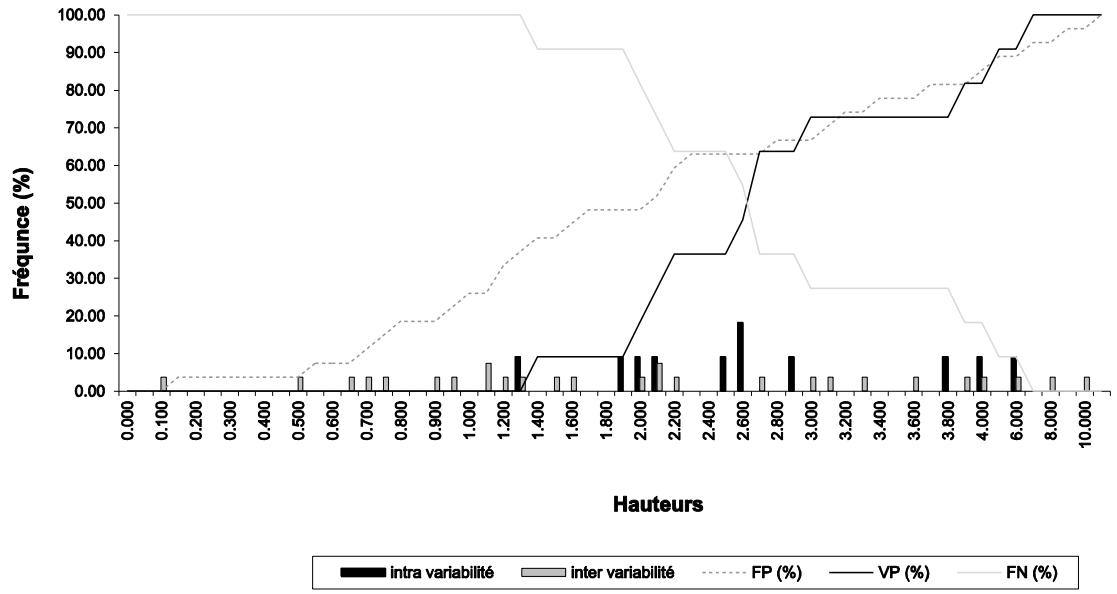


Figure 52. Distribution de l'intra- et l'inter variabilité lorsque les résultats GC-MS et « GC-MS like » (c'est-à-dire, après ajustement mathématique, ici avec le modèle cubique) sont combinés dans le cadre du scénario 2.1. Pour une question de clarté, l'échelle des valeurs de la hauteur *h* n'est pas linéaire.

Bien que pour estimer de manière plus juste l'efficacité réelle d'une telle méthodologie dans la séparation des populations de liés et non liés une augmentation du nombre de données s'impose, sur la base de ces résultats, une mesure de la similarité basée sur l'utilisation combinée de l'ACP et de la CAH démontre des performances de séparation intéressantes dans le cadre de l'étude intra méthode permettant d'envisager l'utilisation des résultats en soutien de l'enquête policière (pour un seuil de 0.75, VP et FN de 99% et 1%, respectivement ; cf. Tableau 20) ou en tant qu'élément de preuve dans une affaire particulière (pour un seuil de 0.5, VP d'environ 95% et FP nul ; cf. Tableau 20).

En revanche, elle ne s'avère pas être une méthodologie de profilage efficace lorsque les données issues de méthodes analytiques différentes sont combinées. En conséquence, définir à l'aide de cette seule mesure de similarité le degré de similarité entre deux profils chimiques ne semble pas approprié.

7.3 Complémentarité entre la mesure du coefficient de corrélation de Pearson et la performance d'ajustement

Etant donné que la classification chimique dans la banque de données de référence découle du calcul du coefficient de corrélation de Pearson, il s'avère intéressant d'examiner les performances d'ajustement en regard des similarités calculées entre les profils chimiques.

Dans ce cadre là, une investigation de la complémentarité entre les mesures du coefficient de corrélation de Pearson et les hauteurs déterminées lors de l'étape locale de l'ACP-CAH doit être entreprise. En quelques mots, il s'agit d'estimer que si les profils sont estimés dans le même cluster d'après une certaine valeur de *h*, alors dans le même temps, la mesure du coefficient de corrélation de Pearson entre eux indique qu'ils sont effectivement liés en regard du seuil de décision fixé et appartiennent dans ce cas à la même classe chimique. Le recours à l'étude des distributions de l'intra- et de l'inter variabilité inter méthodes s'impose.

Les résultats obtenus pour les sets de validation dans le cadre des scénarios d'ajustement 1.5 et 2.1 sont présentés (dans les deux cas, comparaisons de résultats GC-MS et « GC-MS like » après un ajustement selon le modèle linéaire). La Figure 53 illustre la performance d'ajustement réussi et les valeurs médianes du coefficient de corrélation de Pearson en fonction de chaque valeur de *h* pour les comparaisons entre les profils GC-MS et « GC-MS like » des spécimens correspondants, que l'ajustement ait été considéré réussi ou non, dans le cadre du scénario d'ajustement 1.5[45].

[45] Les valeurs médianes sont calculées à partir des mesures du coefficient de corrélation de Pearson estimées lors de la comparaison des profils GC-MS et « GC-MS like » pour les spécimens correspondants des sets de validation, pour chaque scénario d'ajustement, lorsque l'ajustement a été considéré « réussi » ou « non réussi » pour chacune des 100 itérations à chacune des 45 valeurs de *h*.

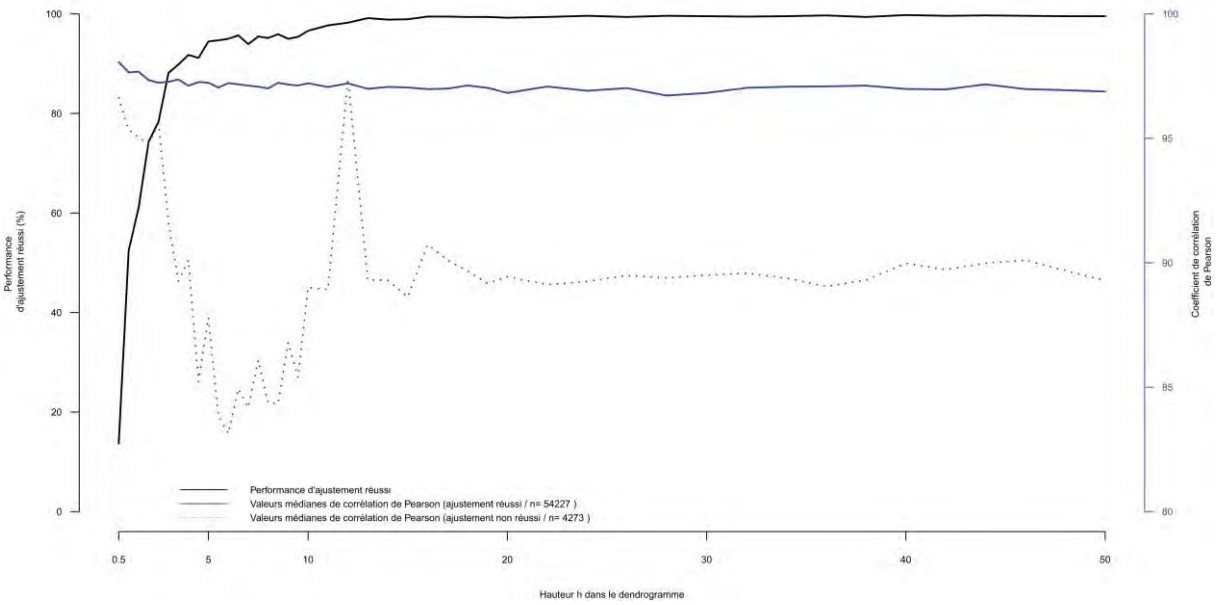


Figure 53. Performance d'ajustement réussi et valeurs médianes de coefficient de corrélation de Pearson en fonction de la hauteur *h* dans le dendrogramme local pour le scénario d'ajustement 1.5 Agilent – Perkin Elmer (résultats issus de l'étude des sets de validation)

Les valeurs médianes obtenues lorsque l'ajustement n'est pas considéré comme réussi sont inférieures à celles calculées lorsque l'ajustement est considéré comme réussi. Ces résultats tendent à démontrer que si une valeur de coefficient de corrélation de Pearson relativement élevée est calculée entre des profils GC-MS et « GC-MS like » (profils définis par les valeurs prétraitées des 6 variables) alors l'ajustement de ces profils devrait être généralement estimé comme réussi dans le dendrogramme local (où les profils sont définis par les scores sur les 4 premières composantes principales). Lorsque des hauteurs faibles sont fixées, la réussite de l'ajustement n'est possible que pour des valeurs de coefficient de corrélation particulièrement hautes (cf. Figure 53).

La Figure 54 quant à elle présente le même type de graphique dans le cadre du scénario d'ajustement 2.1.

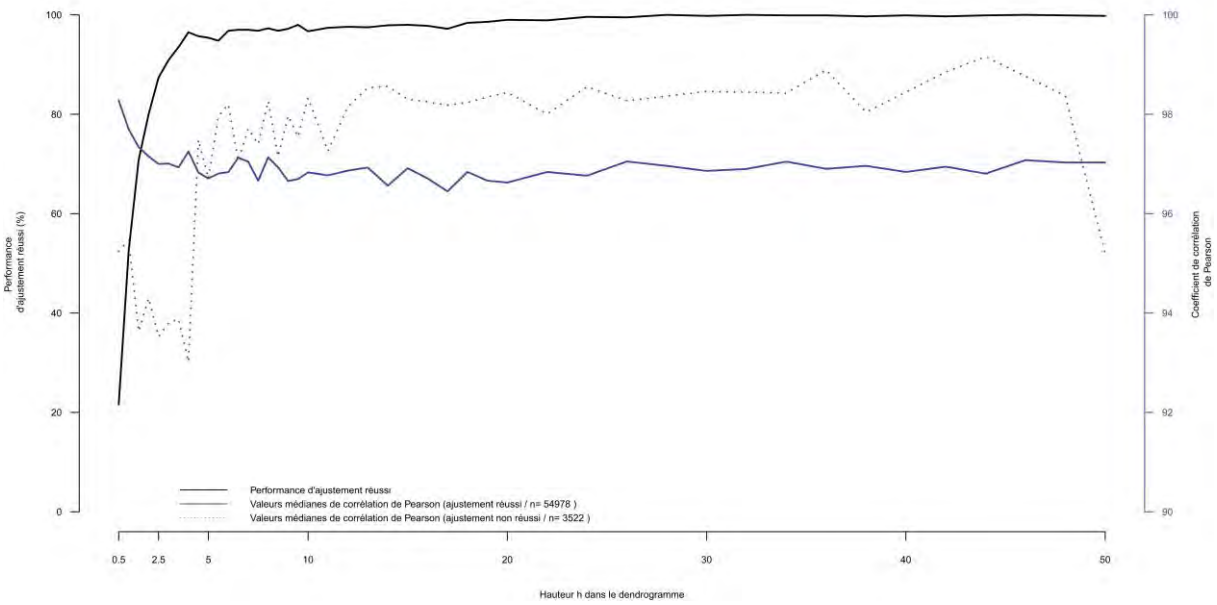


Figure 54. Performance d'ajustement réussi et valeurs médianes de coefficient de corrélation de Pearson en fonction de la hauteur *h* dans le dendrogramme local pour le scénario d'ajustement 2.1 Fast GC-FID (résultats issus de l'étude des sets de validation)

A la différence du scénario 1.5 où la différence entre les valeurs médianes de corrélation est généralement prononcée selon que l'ajustement ait été considéré comme réussi ou non, dans le scénario 2.1, à partir d'une hauteur de 4.5, les valeurs médianes obtenues lorsque l'ajustement est considéré réussi deviennent inférieures à celles calculées lorsque l'ajustement n'est pas réussi (cf. Figure 54). Ainsi, pour des hauteurs particulièrement faibles, de 0.5 à 4.5, dès lors qu'un coefficient de corrélation de Pearson relativement faible est estimé entre les profils GC-MS et « GC-MS like », l'ajustement n'est pas réussi. En d'autres termes, pour de telles hauteurs, il faut que le coefficient de corrélation de Pearson soit particulièrement élevé pour que l'ajustement soit une réussite (et que certainement la structure locale des données soit propice à la réussite de l'ajustement, cf. §7.1).

Le fait que des hautes valeurs de coefficient de corrélation de Pearson soient obtenues alors que l'ajustement n'est pas réussi (valeurs même supérieures que celles obtenues lorsque l'ajustement est réussi après une hauteur de 4.5) peut s'expliquer par l'influence de la structure locale des données et le mode de calcul de la CAH.

En effet, comme démontré au §7.1, il doit exister des échantillons présentant une similarité avec le profil GC-MS plus grande que celle existant entre les profils GC-MS et « GC-MS like », d'où une hauteur plus grande les séparant. Le profil « GC-MS like » ne peut dès lors pas se trouver dans le même cluster malgré une similarité relativement haute. De plus, que des profils GC-MS et « GC-MS like » partagent une similarité élevée, illustrée par un coefficient de corrélation de Pearson élevé, mais qu'ils ne se retrouvent pas dans le même cluster ne représente pas en soi un problème. En effet, l'observation de la Figure 54 illustre que les valeurs médianes de coefficient de Pearson sont effectivement supérieures à une valeur de 95%. Cette valeur se retrouve dans l'intra variabilité estimée pour la combinaison de résultats GC-MS et « GC-MS like » (ajustement selon le modèle mathématique linéaire) (cf. Figure 55 ci-dessous).

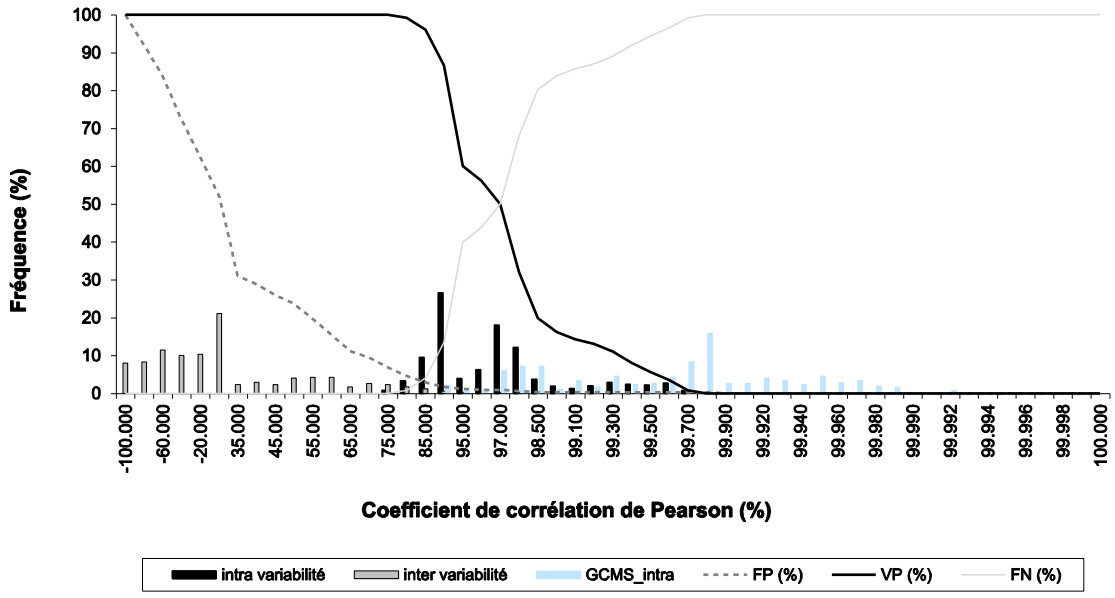


Figure 55. Distribution de l'intra- et l'inter variabilité quand les résultats provenant des méthodes GC-MS et « GC-MS like » (c'est-à-dire, après ajustement mathématique des résultats Fast GC-FID, ici avec le modèle linéaire). Pour une question de clarté, l'échelle des valeurs du coefficient de corrélation de Pearson n'est pas linéaire.

Si un seuil de décision est fixé à 95% dans le cadre de l'étude inter méthodes de ce scénario, alors les taux VP et FP sont égaux à 60.1% et 1.1%, respectivement, démontrant la probabilité élevée que les profils GC-MS et « GC-MS like » soient effectivement liés et appartiennent alors à la même classe chimique bien que l'ajustement soit considéré comme un échec (cf. Figure 54 et Figure 55).

Relevons finalement que la performance d'ajustement étant rapidement élevée (cf. Figure 54), lorsque l'ajustement n'est pas réussi les valeurs de coefficient de corrélation de Pearson élevées ne représentent qu'une part infime de l'ensemble des valeurs calculées.

Sur la base de ces résultats, ce paragraphe démontre qu'une certaine tendance existe dans la relation entre la valeur du coefficient de corrélation de Pearson mesurée entre deux profils chimiques et la réussite ou non de leur ajustement. Finalement, combiner les performances d'ajustement et les mesures du coefficient de corrélation de Pearson en regard des distributions inter méthodes considérées permet une estimation précise de la similarité statistique des profils. Ainsi, il est possible d'apprécier plus objectivement la réussite ou non de l'ajustement de ces derniers, c'est-à-dire leur appartenance à la même classe chimique. Il en découle une estimation plus précise de la conservation ou non de la structure des classes chimiques au sein de la banque de données.

7.4 Estimation de *h* sans prise en compte de la classification chimique

Lors de l'étude du scénario 2.2 (Debrus et al., 2010), la détermination d'une valeur de *h* pertinente s'était faite sans tenir compte des classes chimiques auxquelles appartiennent les profils dans la banque de données de référence. Il s'agissait de ne se baser que sur la distribution statistique des données. Une hauteur de 20 pour couper les dendrogrammes locaux a ainsi été jugée pertinente. En effet, d'après le procédé alors appliqué, les données révélaient l'existence de clusters dans les données. En d'autres termes, les clusters définis à des hauteurs de 20 et moins dans la banque de données de référence sont significativement présents car les profils formant les clusters sont effectivement similaires, d'après le processus de clusterisation (distance euclidienne sur des scores issus d'une ACP puis méthode de groupement *Ward*). Cette valeur de 20 avait donc été utilisée en tant que seuil pour évaluer si l'ajustement entre les profils obtenus avec les méthodes considérées était réussi ou non et par conséquent si la structure des classes chimiques au sein de la banque de données était conservée (Debrus et al., 2010).

7.5 Conclusion

En conclusion, ce chapitre démontre l'influence de la structure locale des données dans l'ajustement de chacun des échantillons dans la banque de données de référence (cf. §7.1). Bien qu'une certaine corrélation existe entre une valeur de coefficient de corrélation de Pearson et la réussite ou non de l'ajustement (cf. §7.3), la structure locale des données influence en partie la réussite de l'ajustement. Le degré de son influence est difficile à juger, cela étant propre à la distribution des échantillons dans l'environnement proche de chacun des spécimens considérés dans la banque de données de référence. En conséquence, trouver une seule hauteur *h* valable pour toutes les classes chimiques s'avère critique dans la mesure où elles ne présentent pas toutes la même dispersion de données et ne contiennent pas toutes le même nombre de profils.

D'après les résultats obtenus, si une valeur de similarité élevée existe entre un profil GC-MS et un profil Fast GC-FID ou « GC-MS like », l'ajustement a de fortes chances d'être réussi. Toutefois, la présence de profils partageant une similarité plus élevée avec le profil GC-MS pourrait ne pas permettre un ajustement réussi pour des hauteurs faibles. A l'inverse, l'ajustement à des hauteurs relativement faibles d'un profil peut être réussi malgré l'existence d'une similarité faible avec le profil GC-MS, selon la densité de spécimens proches existant autour de ce dernier.

Selon les résultats, il est possible d'affirmer que si un ajustement est réussi dès une hauteur de 0.5 alors les profils chimiques sont particulièrement proches dans l'espace défini par les 4 CPs et la structure des classes chimiques est certainement conservée, au sens de la sous-hypothèse 2.2 (cf. §7.2). La hauteur de 20 pourrait représenter la hauteur maximale à laquelle un ajustement pourrait être considéré comme réussi tout en assurant la conservation de la structure de l'information dans la banque de données de référence (cf. §7.4). Cependant, comme l'a montré l'examen en détail du spécimen 267_07_09_1, dans des cas de forte densité locale des données, couper le dendrogramme à une hauteur de 20 pourrait ne pas permettre de détecter l'ensemble des liens chimiques, si l'on prend en référence la mesure du coefficient de corrélation de Pearson. En d'autres termes, alors qu'en coupant le dendrogramme à une hauteur de 20, FP devrait être relativement faible, VP lui risque de ne pas être optimal.

Dans cette recherche et d'après les résultats présentés, en raison de l'influence de la structure locale des données (cf. §7.1) et des faibles performances de l'ACP-CAH pour la discrimination des populations de l'intra- et l'inter variabilité lorsque les résultats provenant de méthodes différentes sont combinés (cf. §7.2), l'utilisation conjointe de l'ACP-CAH et de l'étude de l'intra- et l'inter variabilité inter méthodes est préférée pour évaluer la similarité statistique existant entre des profils obtenus avec des diverses méthodes analytiques. La réussite de l'ajustement dans le dendrogramme local jusqu'à une hauteur de 20 pourrait être vue comme la proximité statistique des profils investigués dans l'espace défini par les CPs considérées. En d'autres termes, que les profils chimiques des échantillons obtenus avec une méthode différente sont transposés à proximité des profils GC-MS correspondants dans la banque de données de référence. Pour évaluer la similarité des profils chimiques et ainsi déterminer leur appartenance ou non à la même classe chimique, la qualité de cette proximité serait alors ensuite évaluée par la mesure du coefficient de corrélation de Pearson entre les profils considérés en regard des distributions de l'intra- et l'inter variabilité inter méthodes.

Chapitre 8 Le maintien d'une banque de données commune à diverses méthodes d'analyse dépend des caractéristiques analytiques de ces dernières (Hypothèse 1)

8.1 Étude des coefficients de détermination ajusté et de prédiction

De manière générale, pour chacun des scénarios et pour chacune des variables, l'application des modèles quadratique et cubique n'apporte pas d'améliorations dans les valeurs des coefficients de détermination ajusté ou de prédiction. De plus, les performances d'ajustement réussi des sets de calibration et de validation ne présentent pas non plus de valeurs plus élevées. Ainsi, d'après ces éléments, l'addition de termes quadratique ou cubique pour établir les règles d'ajustement de chacune des variables n'apporte pas d'améliorations en termes de similarité statistique, une fois l'ajustement des données effectué. C'est donc le modèle linéaire qui a été retenu pour ajuster et prédire les données de chaque composé pour chacun des scénarios d'ajustement.

Notons également que recourir au modèle linéaire s'avère particulièrement intéressant pour réduire le risque de *sur-apprentissage*. Ce dernier se produit lors de l'utilisation de modèles mathématiques complexes pour classifier des données et se caractérise par des performances d'ajustement significativement plus faibles pour le set de validation que pour le set de calibration illustrant ainsi une mauvaise capacité du modèle à généraliser les caractéristiques des données (en d'autres termes, il s'agit là d'une perte de capacité de prédiction pour la classification de nouveaux échantillons). Avant d'aborder les résultats, précisons que dans le cadre du scénario 1.2, c'est l'ajustement des profils chimiques de l'instrument analytique APP 4 vers APP 1 qui a été considéré (cf. §9.1.a).

Le Tableau 21 ci-dessous présente les résultats obtenus pour chacun des scénarios lors de l'application du modèle linéaire. Pour établir le Tableau 21, les moyennes des R^2 ajusté et Q^2 obtenus pour chaque variable d'après les sets de calibration définis sur 100 itérations différentes sont calculées (cf. §6.5.e, alinéa 8). Une moyenne de ces deux coefficients pour chaque scénario est également calculée pour évaluer les capacités prédictives du modèle linéaire pour chaque scénario d'ajustement. Les valeurs des coefficients étant dépendantes du set d'entraînement d'échantillons sélectionnés, la sélection d'un échantillonnage représentatif représente une étape à ne pas négliger.

	R^2 ajusté					Q^2				
Composés	**1.2[46]**	**1.4**	**1.5**	**2.1**	**2.2**	**1.2[46]**	**1.4**	**1.5**	**2.1**	**2.2**
MEC	0.962	0.992	0.895	0.827	0.756	0.965	0.994	0.901	0.828	0.767
AC	0.976	0.964	0.814	0.807	0.776	0.984	0.966	0.816	0.804	0.767
AcTB	0.997	0.995	0.944	0.943	0.785	0.997	0.993	0.942	0.938	0.79
6MAM	0.991	0.996	0.951	0.944	0.618	0.991	0.996	0.953	0.944	0.584
PAP	0.967	0.982	0.838	0.810	0.041	0.968	0.980	0.837	0.797	0.053
NOS	0.987	0.994	0.907	0.943	0.404	0.987	0.994	0.917	0.944	0.412
Moyenne	*0.980*	*0.987*	*0.891*	*0.879*	*0.564*	*0.982*	*0.987*	*0.894*	*0.876*	*0.562*

Tableau 21. Coefficients moyens de détermination ajusté (R^2 ajusté) et coefficients moyens de prédiction (Q^2) pour chaque scénario étudié

D'après ce tableau, on remarque tout d'abord que les valeurs du R^2 ajusté et du Q^2 sont très proches pour chacun des scénarios d'ajustement, impliquant pour le modèle linéaire des capacités similaires d'ajustement puis de prédiction des données. Ensuite, on observe une différence nette dans les valeurs des coefficients selon les scénarios d'ajustement corroborant ainsi la classification des scénarios d'ajustement illustrée au Tableau 12 (§5.2). En effet, les scénarios 1.2 et 1.4 présentent des valeurs de R^2 ajusté et de Q^2 plus élevées que les scénarios 1.5 et 2.1, eux-mêmes avec des valeurs plus élevées que le scénario d'ajustement 2.2 (cf. Tableau 21). Ce dernier scénario, montrant en théorie la similarité analytique avec la méthode de référence GC-MS la plus faible, obtient effectivement les valeurs les plus faibles. En revanche, le modèle linéaire est satisfaisant que ce soit pour l'ajustement ou la prédiction des données obtenues dans le cadre des scénarios 1.2 et 1.4, les valeurs étant relativement élevées pour les deux coefficients.

[46] Concernant le scénario d'ajustement 1.2, pour déterminer ces valeurs, l'ajustement des résultats de l'instrument analytique nommé APP 4 vers l'instrument analytique nommé APP 1 a été considéré (cf. §9.1.a).

Une étude précise du tableau montre que toutes les variables sont bien ajustées pour les scénarios 1.2 et 1.4, tandis que pour les scénarios 1.5 et 2.1 les composés AC, PAP voire MEC présentent des valeurs plus faibles que les trois autres composés (cf. Tableau 21). Ces deux derniers scénarios partageant la même technologie d'analyse de séparation que la méthode de référence GC-MS, il pourrait être envisagé d'améliorer la relation entre les réponses analytiques respectives de ces trois composés en procédant à l'ajustement analytique de la méthode, s'il est estimé que la modification de paramètres analytiques apporte effectivement un gain dans la similarité statistique. Par exemple, il pourrait être envisagé d'agir sur les paramètres analytiques influençant la résolution chromatographique si une mauvaise résolution des composés AC, PAP ou MEC était observée. Le scénario 1.5 partageant la même technologie d'analyse de détection que la méthode de référence, il serait également possible de modifier les paramètres analytiques du MS correspondant au « tune file » pour autant qu'il ait été démontré auparavant que ceci puisse avoir une influence (cf. §2.7).

Finalement, sachant qu'il a été déjà mentionné que les modèles quadratique et cubique ne présentaient pas de performances supérieures au modèle linéaire, il est possible de conclure qu'aucun modèle parmi ces derniers ne parvient à ajuster correctement les réponses analytiques lorsque les données sont obtenues en UHPLC-MS/MS, en particulier celles des composés PAP et dans une moindre mesure NOS, des valeurs proches de 0 étant obtenues. La relation mathématique entre les réponses analytiques obtenues en GC-MS et en UHPLC-MS/MS n'est ainsi pas linéaire, quadratique ou cubique. Ce résultat ne s'avère pas surprenant, les différences analytiques étant importantes entre les deux méthodes, en particulier au niveau des sources d'ionisation respectives des deux techniques analytiques. Alors que l'ionisation en UHPLC-MS/MS se fait en ESI à pression atmosphérique (Electro Spray Ionisation), en GC-MS elle se fait en EI (Electron Impact, cf. Chapitre 2) dans des conditions de vide. Il est certainement probable que cette différence majeure soit responsable du manque de capacité des modèles linéaire, quadratique et cubique à ajuster et prédire les réponses analytiques obtenues en UHPLC-MS/MS. D'après les résultats obtenus dans une précédente étude dans le cadre de cette recherche, un modèle dit *simplifié* est préférable pour ajuster les données lorsque de telles méthodes sont considérées (Debrus et al., 2010).

8.2 Étude des performances d'ajustement et des mesures du coefficient de corrélation de Pearson

La Figure 56 et la Figure 57 illustrent les performances d'ajustement réussi pour chacun des scénarios respectivement, en fonction de la hauteur *h* dans le dendrogramme local (pour rappel, 45 valeurs testées allant de 0.5 à 50), calculées à l'aide du processus ACP-CAH[47]. Les valeurs présentées dans le graphique correspondent aux moyennes des performances obtenues pour chacune des 100 itérations réalisées à chaque valeur de *h* testée, respectivement pour les sets de calibration (cf. Figure 56) et de validation (cf. Figure 57) aléatoirement définis à chaque itération. Le calcul des performances s'est fait pour les données « GC-MS like » des sets de calibration et de validation, c'est-à-dire après ajustement mathématique des données (ici, à l'aide du modèle linéaire). La performance d'ajustement réussi du set de validation représente mieux la capacité prédictive du modèle linéaire pour chaque scénario d'ajustement, à l'inverse de celle calculée pour le set de calibration, en raison de l'utilisation de ce dernier pour établir les règles d'ajustement de chaque variable (cf. Figure 57).

Ces résultats démontrent premièrement qu'une augmentation de la performance d'ajustement va de pair avec une augmentation de la hauteur *h*. Cette observation était attendue dans la mesure où il y a plus de chances que deux profils chimiques d'un même spécimen soient dans le même cluster en fonction d'une hauteur *h* si cette dernière est élevée. Le pourcentage d'ajustement réussi est en d'autres termes directement influencé par la hauteur *h* à laquelle les dendrogrammes locaux sont coupés. Ceci est clairement visible pour les scénarios 1.5, 2.1 et 2.2, les performances obtenues pour les scénarios 1.2 et 1.4 étant déjà particulièrement élevées pour de faibles valeurs de *h* (plus de 90% d'ajustement réussi pour de nouveaux spécimens d'héroïne avec un *h* de 0.5, cf. Figure 57).

[47] Concernant l'estimation de la performance d'ajustement du scénario 1.2, voir la note de bas de page n°46.

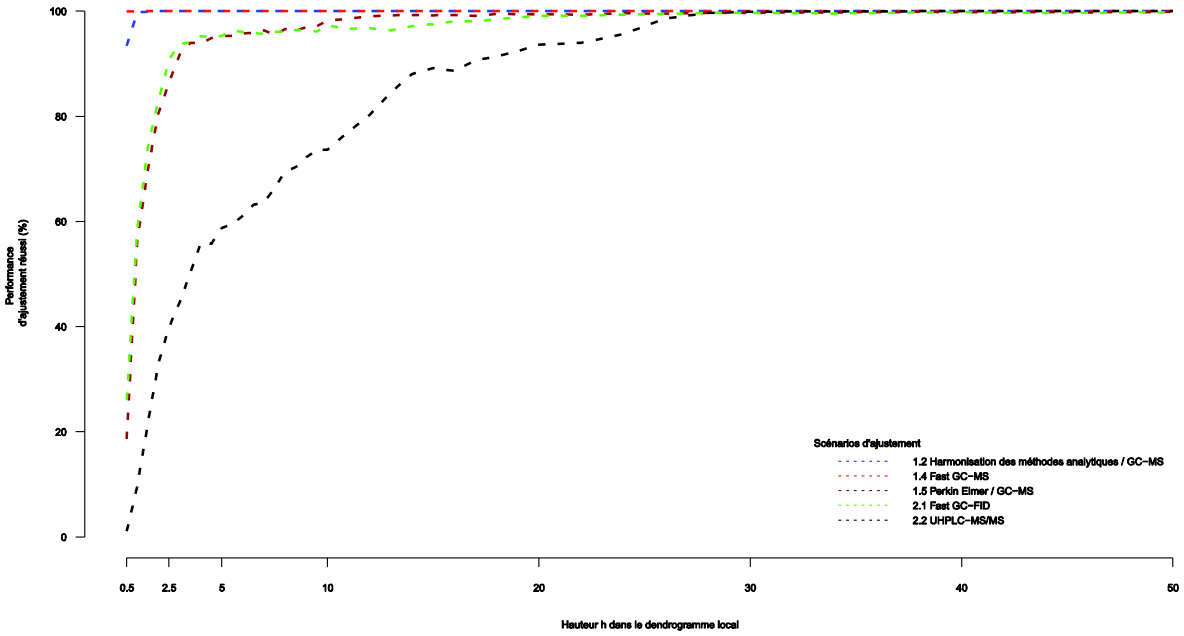


Figure 56. Evolution de la performance d'ajustement réussi pour les *sets de calibration* respectifs en fonction de la hauteur *h* dans le dendrogramme local pour les différents scénarios d'ajustement

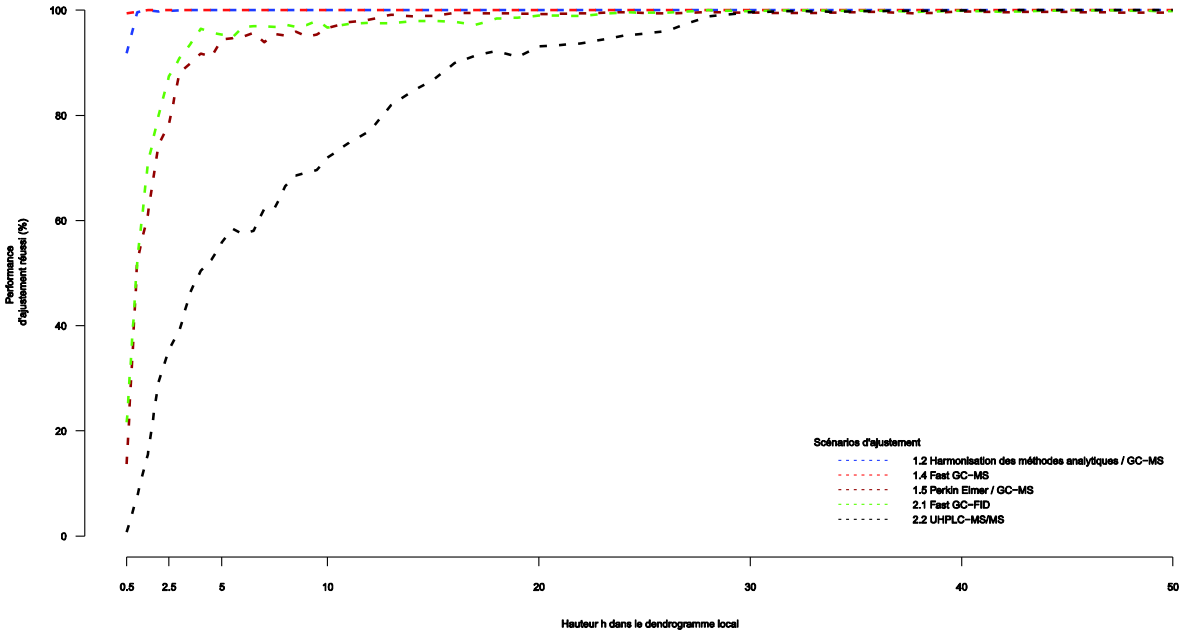


Figure 57. Evolution de la performance d'ajustement réussi pour les *sets de validation* respectifs en fonction de la hauteur *h* dans le dendrogramme local pour les différents scénarios d'ajustement

D'après ces résultats, une corrélation peut être observée entre le coefficient de détermination R^2 ajusté moyen et le pourcentage d'ajustement réussi pour les sets de calibration de chacun des scénarios respectifs. Ceci signifie que si le modèle linéaire ajuste bien les données de calibration (illustré par des R^2 ajusté relativement élevés), alors il peut être attendu que l'ajustement des données « GC-MS like » du set de calibration sera réussi (c'est-à-dire, que les deux profils chimiques d'un même spécimen seront, dans le dendrogramme local, dans le même cluster selon la hauteur *h* fixée, cf. §6.5.e). On observe que plus le R^2 ajusté est élevé, plus le pourcentage d'ajustement réussi du set de calibration est élevé, même à de faibles hauteurs *h* (scénarios 1.2 et 1.4, cf. Figure 56).

De la même manière, une corrélation peut être observée entre le coefficient de prédiction Q^2 de chacun des scénarios et le pourcentage d'ajustement réussi pour les sets de validation respectifs. Ceci signifie que si le modèle linéaire prédit bien les données de calibration (illustré par des Q^2 relativement élevés), alors il peut être attendu que l'ajustement des données « GC-MS like » du set de validation sera réussi. L'utilisation conjointe d'un modèle mathématique relativement simple tel que le modèle linéaire et d'un échantillonnage représentatif contribue à ce résultat.

Finalement, une hiérarchie identique dans la similarité statistique à celle définie ci-dessus à l'aide des R^2 ajusté et des Q^2 (cf. Tableau 21) est également observée d'après les performances d'ajustement réussi respectives des sets de calibration et validation (cf. Figure 56 et Figure 57). Il est ainsi clairement observé que plus la similarité analytique entre la méthode de référence GC-MS et les méthodes considérées est élevée, plus les performances d'ajustement respectives sont élevées : les performances pour les scénarios 1.2 (GC-MS) et 1.4 (Fast GC-MS) sont ainsi plus élevées que celles obtenues pour les scénarios 1.5 (GC-MS / Fabricant différent) et 2.1 (Fast GC-FID), elles-mêmes plus élevées que celles obtenues pour le scénario 2.2 (UHPLC-MS/MS). Ces résultats illustrent ainsi la classification théorique des scénarios d'ajustement effectuée au Chapitre 5 (cf. Tableau 12). Cette tendance est clairement visible lorsque de faibles valeurs de *h* font office de seuil. Alors que les scénarios d'ajustement 1.5 et 2.1 atteignent une performance d'environ 95 % dès une valeur de *h* égale à 5, le scénario 2.2 présentant la différence analytique la plus grande n'atteint une telle performance qu'à partir d'une hauteur de 24 (cf. Figure 57).

Le Tableau 22 présente les valeurs médianes du coefficient de corrélation de Pearson calculées lors de la comparaison des profils GC-MS et « GC-MS like » (ajustement mathématique selon le modèle linéaire) pour les spécimens correspondants des sets de validation, pour chaque scénario d'ajustement, lorsque l'ajustement a été considéré « réussi » pour chacune des 100 itérations à chacune des 45 valeurs de *h*.

Scénario	Coefficient de corrélation de Pearson
1.2	99.86
1.4	99.67
1.5	97.10
2.1	96.93
2.2	78.65

Tableau 22. Valeurs médianes du coefficient de corrélation de Pearson calculées entre les profils GC-MS et « GC-MS like » des spécimens correspondants pour chaque scénario d'ajustement, lorsque l'ajustement a été considéré « réussi » pour chacune des itérations à chaque valeur de *h*

De nouveau, les scénarios peuvent se ranger en trois groupes bien distincts d'après les valeurs médianes calculées (cf. Tableau 22), les scénarios 1.2 et 1.4 formant le premier groupe, les scénarios 1.5 et 2.1 le deuxième, tandis que le scénario 2.2 forme à lui seul le troisième et dernier groupe, celui pour lequel les valeurs du coefficient de corrélation de Pearson sont les plus faibles, révélant la différence analytique la plus grande avec la méthode de référence GC-MS (cf. Tableau 12). Il est intéressant de remarquer que les scénarios d'ajustement peuvent se ranger dans les trois mêmes groupes que ce soit sur la base des performances d'ajustement réussi ou des valeurs médianes du coefficient de corrélation de Pearson (cf. Figure 57 et Tableau 22).

Par conséquent, du moins pour la description générale de la différence analytique existant entre les méthodes considérées, une certaine corrélation existe entre les valeurs moyennes des performances d'ajustement réussi des sets de validation et les valeurs médianes du coefficient de corrélation de Pearson calculées entre les profils GC-MS et « GC-MS like » des sets de validation lorsque l'ajustement a été considéré « réussi ». En d'autres termes, lorsque les performances d'ajustement sont relativement élevées pour un scénario d'ajustement donné, alors les valeurs médianes du coefficient de corrélation de Pearson devraient l'être également, et réciproquement.

Toutefois, cela ne signifie pas bien sûr que tous les échantillons montreront une valeur de coefficient de corrélation de Pearson élevée lors de la comparaison des profils respectifs.

Sur la base de la classification des scénarios d'ajustement (cf. Tableau 12), on aurait pu s'attendre à obtenir une différence plus marquée entre les performances respectives des scénarios 1.5 et 2.1, en particulier des performances plus élevées pour le scénario 1.5 ou plus faibles pour le scénario 2.1 (cf. Figure 57 et Tableau 22). D'après ces résultats et la similarité statistique estimée, une méthode analytique présentant les mêmes technologies d'analyse mais implémentée sur un appareillage de fabricant différent (Agilent vs. Perkin Elmer) présente un niveau comparable de similarité analytique qu'une méthode décrite par une technologie de détection différente (MS vs. FID). Il semblerait ainsi que des conceptions d'appareillages propres à chaque fabricant influencent significativement la similarité des profils chimiques. En particulier, alors qu'il semble pertinent d'exclure le GC comme influençant de manière importante la similarité statistique des profils chimiques, il est en revanche probable qu'une architecture du MS différente, par exemple au niveau de la source ionique, l'influence de manière significative.

La représentation graphique des profils chimiques « GC-MS like » obtenus avec les méthodes considérées à l'aide de CP1 et CP2 permet de visualiser l'influence de la similarité analytique sur la similarité statistique par comparaison avec la position du profil chimique GC-MS du spécimen correspondant pour chacun des scénarios d'ajustements, respectivement (par exemple pour le spécimen 042_01_09_1, cf. le Tableau 23, la Figure 58, la Figure 59, la Figure 60, la Figure 61 et la Figure 62).

Scénario	Coefficient de corrélation de Pearson	Hauteur *h*
1.2	100.00	0.07
1.4	99.84	0.18
1.5	98.14	0.47
2.1	96.94	0.59
2.2	54.59	6.39

Tableau 23. Valeurs du coefficient de corrélation de Pearson et de la hauteur *h* dans le dendrogramme local pour la comparaison des profils GC-MS et « GC-MS like » (ajustement mathématique à l'aide du modèle linéaire) du spécimen 042_01_09_1 pour chaque scénario d'ajustement

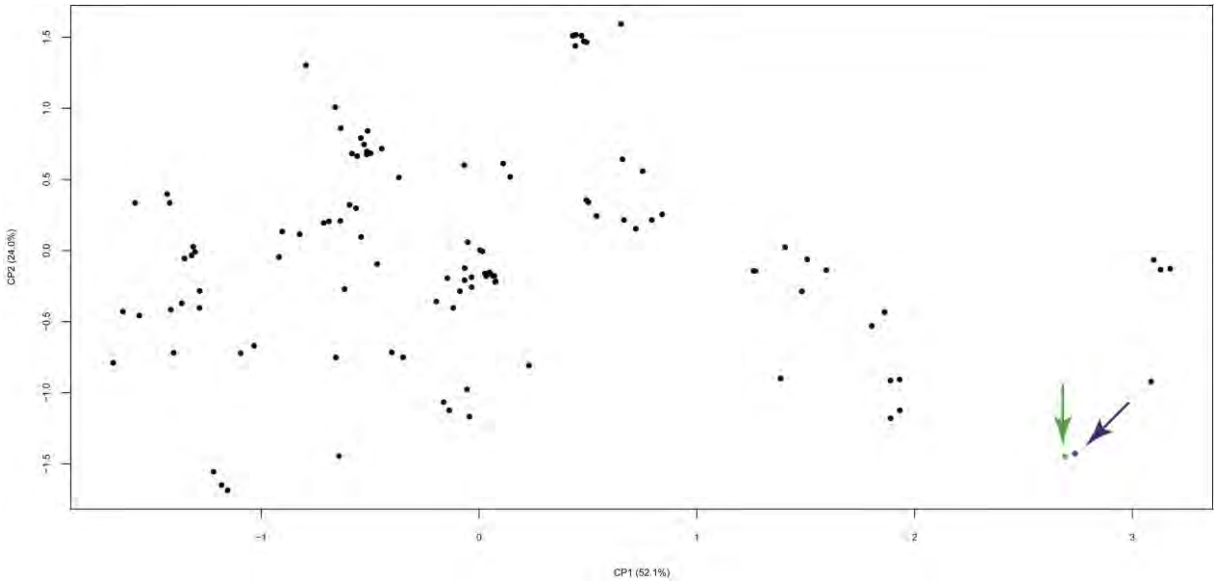


Figure 58. Identification pour le spécimen 042_01_09 des profils GC-MS (en vert) et « GC-MS like » (en bleu, ajustement mathématique à l'aide du modèle linéaire), à l'aide de CP1 et CP2 dans le sous-échantillonnage de leurs spécimens proches, dans le cadre du scénario d'ajustement 1.2

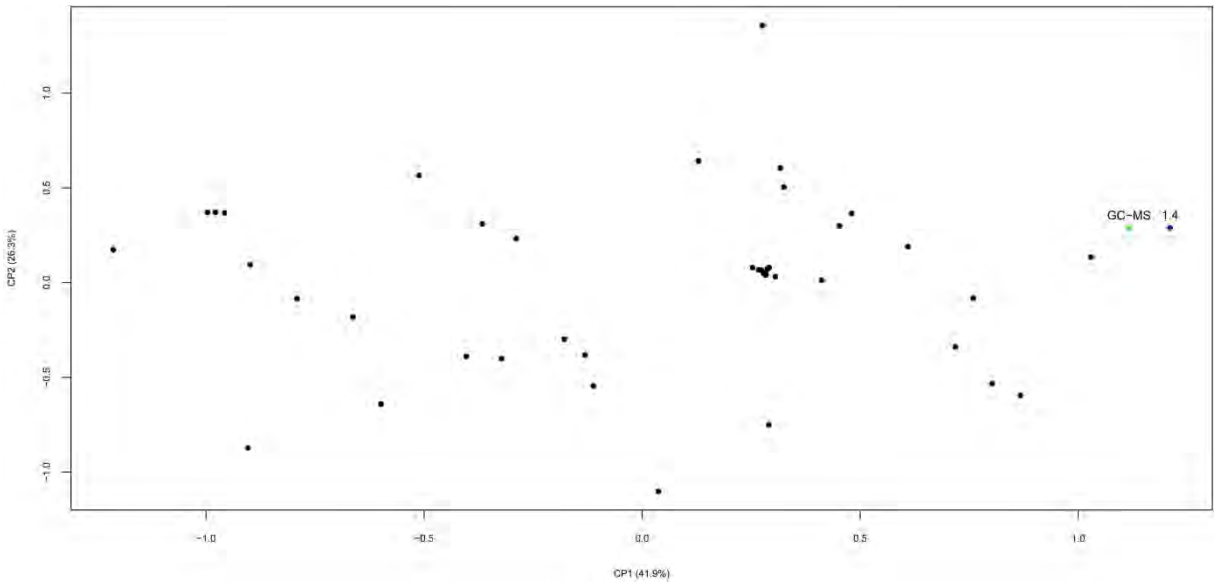


Figure 59. Identification pour le spécimen 042_01_09 des profils GC-MS (en vert) et « GC-MS like » (en bleu, ajustement mathématique à l'aide du modèle linéaire), à l'aide de CP1 et CP2 dans le sous-échantillonnage de leurs spécimens proches, dans le cadre du scénario d'ajustement 1.4

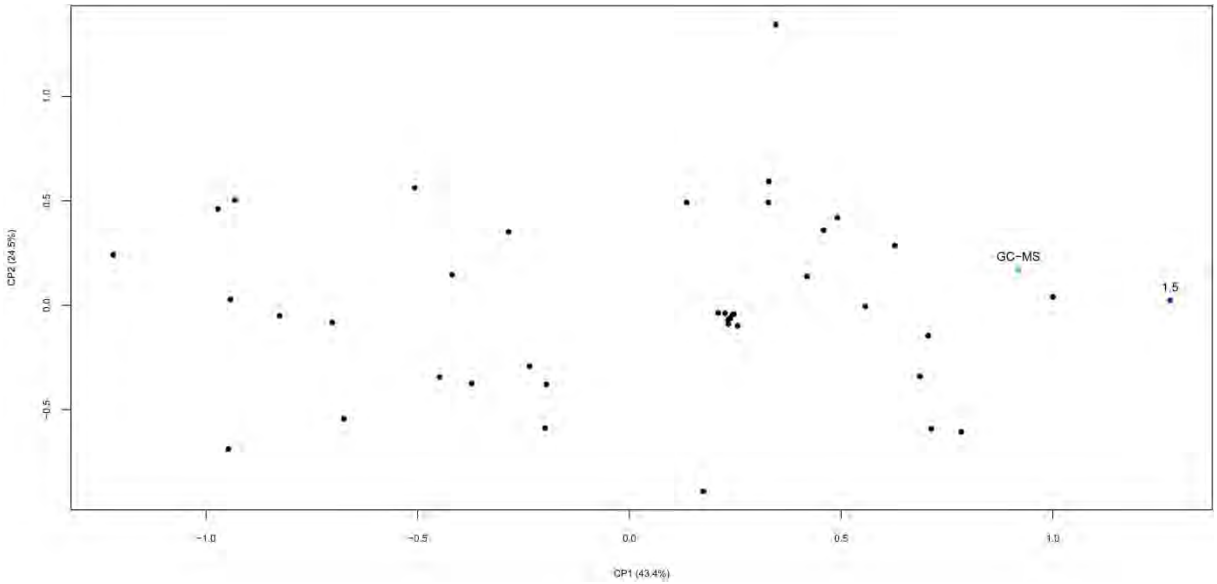


Figure 60. Identification pour le spécimen 042_01_09 des profils GC-MS (en vert) et « GC-MS like » (en bleu, ajustement mathématique à l'aide du modèle linéaire), à l'aide de CP1 et CP2 dans le sous-échantillonnage de leurs spécimens proches, dans le cadre du scénario d'ajustement 1.5

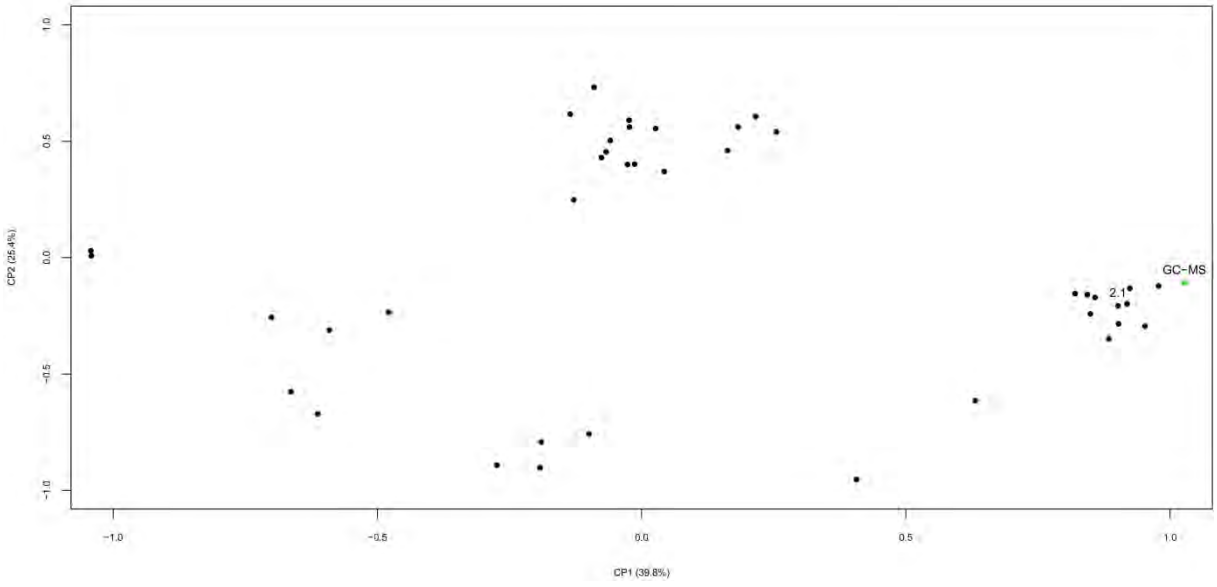


Figure 61. Identification pour le spécimen 042_01_09 des profils GC-MS (en vert) et « GC-MS like » (en bleu, ajustement mathématique à l'aide du modèle linéaire), à l'aide de CP1 et CP2 dans le sous-échantillonnage de leurs spécimens proches, dans le cadre du scénario d'ajustement 2.1

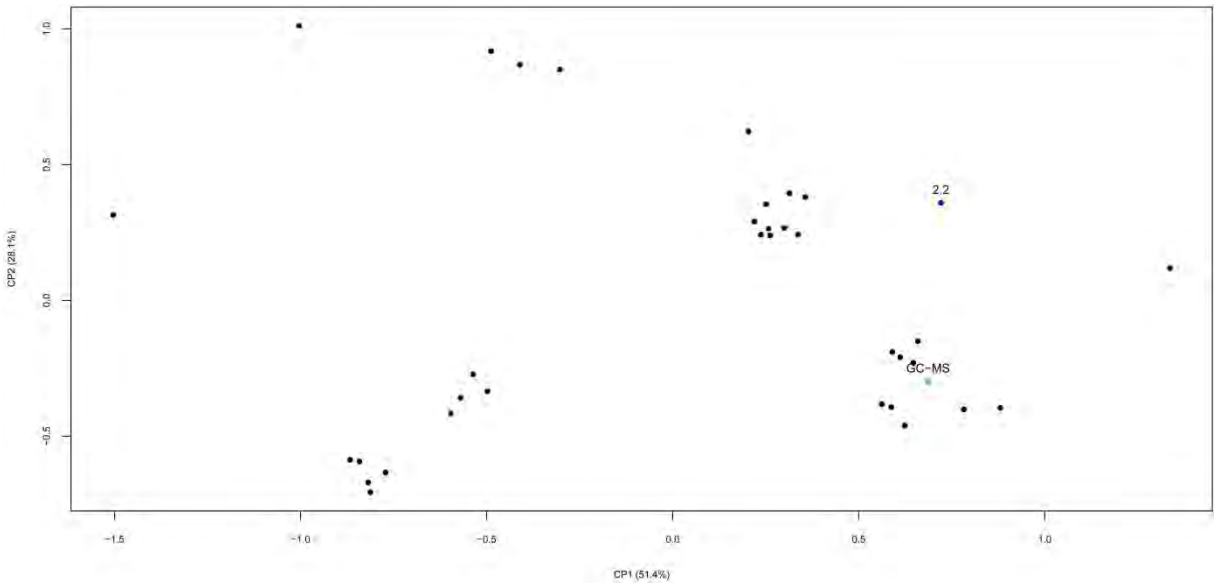


Figure 62. Identification pour le spécimen 042_01_09 des profils GC-MS (en vert) et « GC-MS like » (en bleu, ajustement mathématique à l'aide du modèle linéaire), à l'aide de CP1 et CP2 dans le sous-échantillonnage de leurs spécimens proches, dans le cadre du scénario d'ajustement 2.2

8.3 Conclusion

L'échantillonnage relativement réduit (représentant le 10% des spécimens présents dans la banque de données de référence) permet déjà la création de règles d'ajustement robustes dans la mesure où une prédiction satisfaisante de nouveaux profils chimiques de spécimens d'héroïne est obtenue (cf. Tableau 21 et Figure 57). Sachant que la procédure de calcul des performances d'ajustement réussi se base sur la création aléatoire, 100 fois, de sets de calibration et de validation, augmenter la taille de l'échantillonnage ne devrait pas entraîner une modification importante des performances calculées dans cette recherche, l'estimation des résultats étant robuste. Toutefois, améliorer la représentativité de l'échantillonnage pourrait permettre d'établir des règles d'ajustement d'autant plus valides pour la prédiction de nouveaux profils chimiques, c'est-à-dire améliorer la capacité prédictive des modèles implémentés pour l'ajustement mathématique (bien que la représentativité de l'échantillonnage soit démontrée comme étant satisfaisante, cf. §6.4.a).

Bien que la classification des scénarios d'ajustement se soit faite dans l'idée d'intégrer l'ensemble des scénarios pouvant exister, il est envisageable qu'un laboratoire se trouve dans une situation légèrement différente de celles étudiées dans cette recherche. Ainsi, on se gardera de toute généralisation des résultats si l'on est confronté à d'autres scénarios d'ajustement que ceux investigués. Cependant, l'estimation des résultats étant robuste ceux-ci donnent d'ores et déjà des indications fiables pour l'investigation d'un futur scénario. En effet, en s'aidant de la classification présentée dans le Tableau 12 et des résultats obtenus dans cette recherche, on s'attend à obtenir une certaine similarité statistique pour un scénario donné, bien que celui-ci soit différent de ceux investigués dans cette étude et que la banque de données de référence soit différente.

Finalement, l'investigation à proprement parler de cette première hypothèse principale démontre l'impact de la similarité (ou de la différence) analytique sur la similarité des profils chimiques obtenus par des méthodes différentes. Ainsi, **les résultats expérimentaux obtenus ne permettent pas de réfuter cette première hypothèse de travail**.

En effet, les valeurs médianes du coefficient de corrélation de Pearson calculées entre les profils GC-MS et « GC-MS like » des spécimens correspondants et les performances d'ajustement réussi obtenues à chaque valeur de *h* ainsi que l'évolution des performances en fonction de *h* illustrent l'influence de la différence de niveaux de paramètres analytiques sur la similarité statistique des résultats.

En conséquence, plus les caractéristiques analytiques des méthodes considérées seront différentes, plus la similarité statistique entre les profils chimiques de chacun des spécimens obtenus avec ces dernières sera faible. Il pourrait en découler alors un maintien d'une banque de données commune avec les méthodes considérées d'autant plus difficile à réaliser.

Chapitre 9 La mise en place de méthodes analytiques différentes n'est pas un frein au maintien d'une banque de données commune (Hypothèse 2)

La démonstration de cette seconde hypothèse principale repose sur l'étude des résultats obtenus dans le cadre du scénario d'ajustement 2.1 (analyse des échantillons à l'aide d'une méthode Fast GC-FID). La démonstration qui va suivre devrait être réalisée pour chaque scénario d'ajustement. Ce n'est qu'au travers de la combinaison de tous les résultats que l'hypothèse pourra être objectivement discutée. Pour envisager le maintien d'une banque de données commune à des méthodes analytiques différentes, il s'agit de démontrer qu'une fois les résultats combinés, une méthodologie de profilage efficace est obtenue (cf. §9.1, sous-hypothèse 2.1) et que la structure des classes chimiques au sein de la banque de données est conservée (cf. §9.2, sous-hypothèse 2.2). La même démonstration est effectuée une fois les données ajustées mathématiquement pour évaluer l'intérêt d'une telle approche d'optimisation de la similarité et ainsi estimer si le maintien d'une banque de données commune s'en trouve facilité (cf. Chapitre 10).

9.1 Efficacité de la méthodologie de profilage (sous-hypothèse 2.1)

L'estimation de la performance de la méthodologie de profilage repose sur l'étude de l'intra- et de l'inter variabilité intra- et inter méthodes (Broséus et al., 2013). Avant d'investiguer l'alimentation de la banque de données par le biais de différentes méthodes, il s'agit en effet de démontrer que les deux méthodes peuvent respectivement mettre en place une méthodologie de profilage chimique efficace et ainsi d'étudier les distributions intra méthode. En particulier, l'efficacité s'évalue sur la base des taux VP, FP et FN calculés en fonction d'un seuil de décision (cf. Chapitre 1, §1.4.b et Chapitre 6, §6.4.f). Dans le cas présent, ces indicateurs statistiques sont utilisés pour évaluer la similarité des résultats provenant des méthodes GC-MS et Fast GC-FID.

Comme cela a déjà été mentionné (cf. Chapitre 1), la sélection d'un seuil de décision découle d'un compromis. En effet, les résultats du profilage peuvent être utilisés en tant que soutien à l'enquête policière ou en tant qu'éléments de preuve dans une affaire spécifique (Esseiva et al., 2011). Ainsi, selon la manière dont le laboratoire utilise cette information, le type d'erreurs à minimiser autant que possible n'est pas le même. Si les profils chimiques font office d'éléments de preuve alors le taux de faux positifs (FP) devrait être minimisé en sélectionnant une valeur de coefficient de corrélation de Pearson plus élevée pour le seuil de décision (ce qui conduit à une diminution du taux de vrais positifs, VP). A l'inverse, lorsque les résultats visent à soutenir l'investigation policière, alors le taux de faux négatifs (FN) devrait être minimisé et le taux de vrais positifs maximisé, en déplaçant le seuil vers les valeurs de corrélation faibles. Par conséquent, sachant que les taux d'erreurs dépendent de la valeur du seuil de décision, ce dernier devrait être choisi de telle sorte à minimiser l'erreur la moins acceptable (FP vs. FN).

Ainsi, bien que cette recherche vise à fournir une méthodologie pour l'estimation et l'ajustement des résultats, chaque laboratoire devrait décider si l'alimentation de la banque de données par le biais de différentes méthodes donne des résultats valides selon sa perspective et ses objectifs d'utilisation du profilage chimique.

Avant de présenter l'étude intra- et inter méthodes du scénario d'ajustement 2.1, celle du scénario 1.2 est investiguée. Pour rappel, ce dernier fait référence à l'harmonisation des méthodes analytiques. L'estimation de l'efficacité de la méthodologie de profilage chimique obtenue pour ce scénario peut être utilisée comme indicateur du degré d'efficacité de la méthodologie de profilage chimique obtenue dans le cadre du scénario d'ajustement 2.1, lorsque des résultats MS et FID sont combinés, pour ainsi d'estimer la faisabilité en pratique d'une telle combinaison de méthodes.

9.1.a Scénario d'ajustement 1.2

Dans ce scénario, une méthodologie de profilage chimique de spécimens d'héroïne par une analyse GC-MS a été implémentée sur 4 instruments analytiques différents mais de même marque et modèle d'après les recommandations de l'approche d'harmonisation des méthodes analytiques. Les analyses se déroulent sur des GC Agilent Technologies 7890A couplés à des MS Agilent Technologies 5975C inert XL MSD Triple Axis Detector.

L'échantillonnage présenté au §6.4.a a été analysé sur chaque instrument analytique en méthode rapide dont les paramètres analytiques ainsi que les chromatogrammes correspondants sont présentés en annexe (cf. Annexe 2, §1.2).

Le Tableau 24 présente la terminologie adoptée pour chaque instrument analytique ainsi que les dates auxquelles chacun d'eux est possiblement sorti d'usine d'après la date de la déclaration de conformité.

	ID	Numéro de série GC/MS	Date de la déclaration de conformité
Instrument	APP 1	/US10503701	Février 2010
	APP 2	CN11391114/US11383837	Novembre 2011
	APP 3	CN11211087/US11163723	Avril 2011
	APP 4	CN10949117/US94334025	Juin 2009

Tableau 24. Terminologie, numéros de série du GC et du MS et dates de déclaration de conformité pour chaque instrument analytique

L'harmonisation des méthodes analytiques recommande l'application des mêmes paramètres analytiques au niveau du GC et du MS. Les paramètres définis par le « tune file » (cf. Chapitre 2, §2.7) ne sont en revanche pas concernés dans cette approche d'harmonisation et par conséquent les valeurs optimales définies automatiquement pour chaque instrument analytique sont conservées pour l'ensemble des analyses.

Distribution intra méthode

Le Tableau 25 démontre une similarité des performances de discrimination respectives pour APP 1, APP 2 et APP 4. Même lorsqu'un seuil élevé de 95% est fixé, le taux de vrais positifs reste particulièrement haut tout en produisant un taux de faux positifs de moins de 1%.

		VP (%)				FP (%)			
	GC-MS	**APP 1**	**APP 2**	**APP 3**	**APP 4**	**APP 1**	**APP 2**	**APP 3**	**APP 4**
Seuil	85	100.0	100.0	100.0	100.0	1.6	1.0	2.1	1.6
	90	100.0	100.0	96.8	100.0	1.6	0.5	0.5	1.0
	95	99.0	100.0	94.3	100.0	0.5	0.5	0.5	0.5

Tableau 25. Performance de la discrimination pour le profilage chimique de l'héroïne, d'après les taux d'erreurs et les valeurs de seuil, pour chaque instrument analytique

Les distributions étant également similaires seule celle de APP 4 est présentée (cf. Figure 63). Bien que l'efficacité de la méthodologie de profilage sur APP 3 soit satisfaisante, les performances obtenues sont moins bonnes que celles des trois autres instruments analytiques, comme en témoigne le taux de vrais positifs plus faible à un seuil de 95, par exemple (cf. Tableau 25 et Figure 64). En effet, l'intra variabilité est légèrement plus décalée vers les faibles valeurs de coefficient de corrélation de Pearson, comme l'illustre les taux de vrais positifs et de faux positifs calculés pour une valeur de seuil de 90, en comparaison à ceux estimés pour APP 1, APP 2 et APP 4 (cf. Tableau 25, Figure 63 et Figure 64).

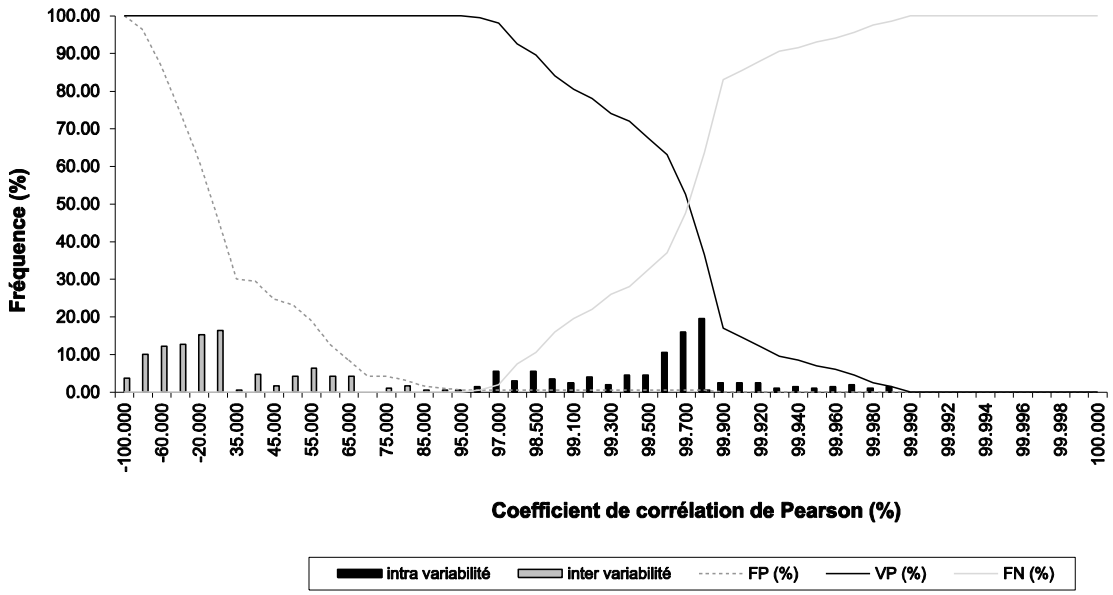


Figure 63. Distribution de l'intra- et l'inter variabilité pour l'instrument analytique APP 4. Pour une question de clarté, l'échelle des valeurs du coefficient de corrélation de Pearson n'est pas linéaire.

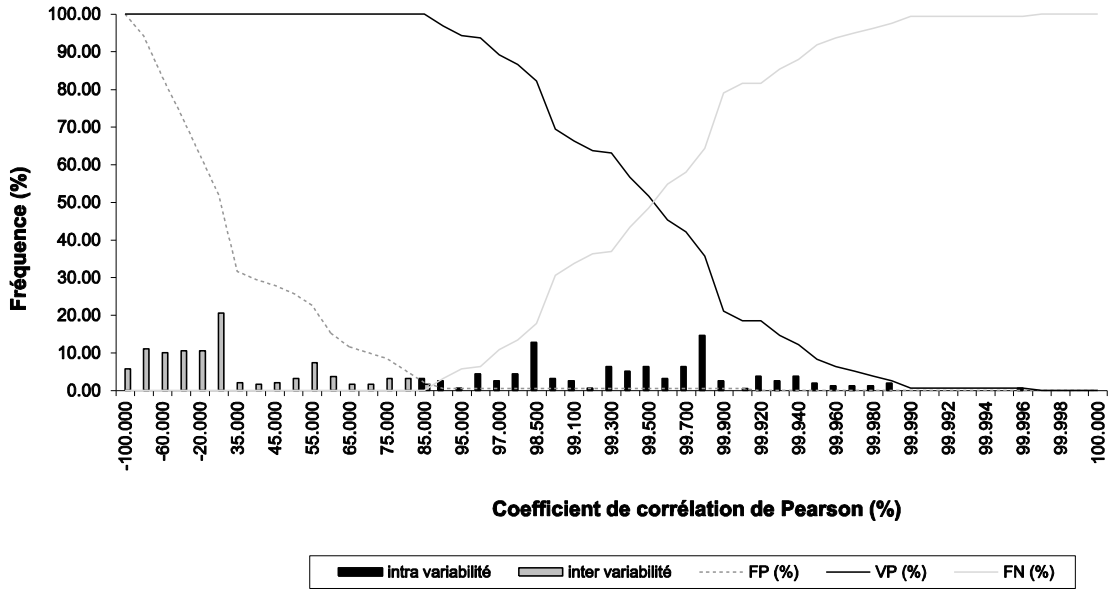


Figure 64. Distribution de l'intra- et l'inter variabilité pour l'instrument analytique APP 3. Pour une question de clarté, l'échelle des valeurs du coefficient de corrélation de Pearson n'est pas linéaire.

D'après ces résultats, on peut conclure qu'une méthodologie de profilage efficace peut être implémentée sur chaque instrument analytique séparément quelle que soit la perspective d'utilisation des profils chimiques. L'évaluation des distributions lorsque les résultats provenant des 4 instruments analytiques sont combinés peut donc être entreprise.

Distribution inter méthodes

Avant d'aborder les résultats, ce scénario correspondant à l'harmonisation des méthodes analytiques, ce sont avant tout les différences dans les variabilités intrinsèques de chaque instrument analytique et, potentiellement, dans les paramètres du « tune » MS qui sont révélées par les distributions inter méthodes.

Pour estimer la distribution des populations d'échantillons liés et non liés lorsque les résultats provenant des 4 instruments sont combinés, un sous-échantillonnage a été pris en compte pour chaque instrument analytique. L'intra- et l'inter variabilité se déterminent ainsi à l'aide de 6 et 20 spécimens, respectivement. De manière similaire à la procédure décrite au Chapitre 6 (§6.4.f), l'intra variabilité se compose des comparaisons entre les réplicats respectifs des 6 spécimens qui définissent l'intra variabilité pour chaque instrument analytique (APP 1 avec APP 2, APP 3 et APP 4, puis APP2 avec APP 3 et APP 4 et finalement APP 3 avec APP 4). Avant d'estimer l'inter variabilité, il faut souligner que chacun des 20 spécimens a été analysé avec chaque instrument analytique. Or, il s'avère important de prendre en compte la variabilité due à chaque instrument analytique dans l'estimation de l'inter variabilité. Ainsi, pour chaque spécimen, un profil parmi les 4 disponibles (un pour chaque instrument analytique) a été aléatoirement sélectionné. Ce processus a été répété 10 fois de telle sorte que tous les profils des spécimens correspondants, obtenus avec les 4 instruments analytiques, auront été sélectionnés au moins une fois dans l'estimation de l'inter variabilité.

La Figure 65 et le Tableau 26 présentent les performances de discrimination calculées lors de la combinaison des résultats provenant des 4 instruments analytiques.

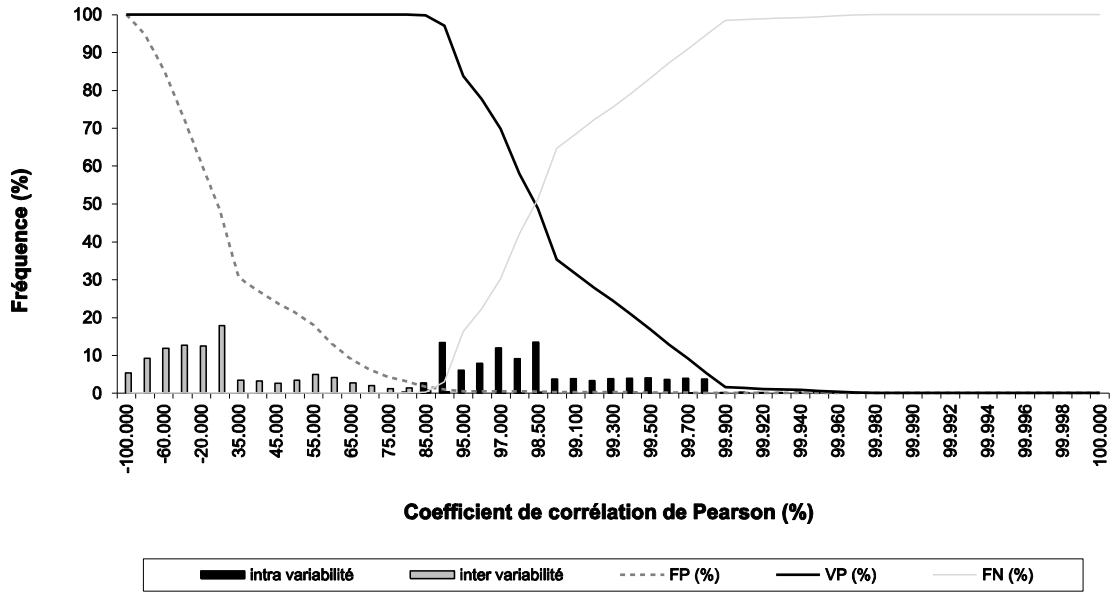


Figure 65. Distribution de l'intra- et l'inter variabilité lorsque les résultats provenant des 4 instruments analytiques sont combinés. Pour une question de clarté, l'échelle des valeurs du coefficient de corrélation de Pearson n'est pas linéaire.

		VP (%)	FP (%)
Seuil	85	99.7	1.7
	90	97.1	1.0

Tableau 26. Performance de la discrimination pour le profilage chimique de l'héroïne lorsque les résultats issus des 4 instruments analytiques sont combinés

En comparaison aux distributions intra méthode, des performances moins bonnes sont obtenues : pour obtenir un taux de vrais positifs élevé, le seuil de décision doit à présent être fixé à une valeur plus faible (85) ce qui produit un taux de faux positifs environ 2 fois supérieur à celui calculé pour APP 2 par exemple (cf. Tableau 26). La Figure 65 illustre clairement le décalage de l'intra variabilité vers des valeurs de coefficient de corrélation plus faibles en comparaison à la situation intra méthode. Ce résultat n'est pas surprenant dans la mesure où les résultats provenant de 4 instruments analytiques sont combinés. La variabilité de ces résultats est donc nécessairement plus élevée que celle existant au sein d'un seul et même instrument analytique, bien que les méthodes analytiques y implémentées ne soient pas différentes d'après la méthodologie d'harmonisation des méthodes analytiques.

Sur la base de ces résultats, les performances calculées sont intéressantes dans une optique de profilage chimique de l'héroïne et par conséquent le maintien d'une banque de données à l'aide de 4 méthodes analytiques identiques implémentées sur des instruments physiquement différentes est envisageable.

9.1.b Scénario d'ajustement 2.1

Les méthodes analytiques GC-MS et Fast GC-FID permettent la séparation et la détection des composés présents dans d'un spécimen d'héroïne en une vingtaine de minutes et en moins de cinq minutes, respectivement (cf. §1.1, Figure 1 et §1.5, Figure 8, Annexe 2). La technologie d'analyse de détection étant différente entre les deux méthodes, l'échelle de valeurs pour chaque variable s'avère bien sûr différente. Dans un tel cas de figure, le prétraitement statistique des données appliqué prend tout son sens et permet la comparaison des réponses analytiques des variables respectives.

Étude descriptive

La Figure 66 représente la corrélation existant entre les résultats GC-MS et Fast GC-FID pour chacune des variables, après prétraitement de celles-ci. La Figure 67 compare la distribution des valeurs obtenues pour chaque variable pour les résultats GC-MS et Fast GC-FID à l'aide de boxplots, après prétraitement des résultats. Tandis que la Figure 66 révèle l'existence d'une certaine corrélation pour les composés AcTB, MAM et NOS, la distribution des valeurs obtenues pour chacune des variables pour les résultats GC-MS et Fast GC-FID illustre que les distributions respectives sont comparables (cf. Figure 67).

Ces résultats sont encourageants dans la mesure où les méthodes considérées présentent une différence dans tous les niveaux de paramètres analytiques, en particulier dans la technologie d'analyse de détection (cf. Chapitre 5, Figure 23 et Tableau 12). La détection en FID est non sélective tandis que la détection en MS est une détection sélective et spécifique. En effet, les données MS correspondent aux aires des ions cibles pour chaque composé constituant le profil alors que les données FID consistent dans les aires de pic pour ces mêmes composés.

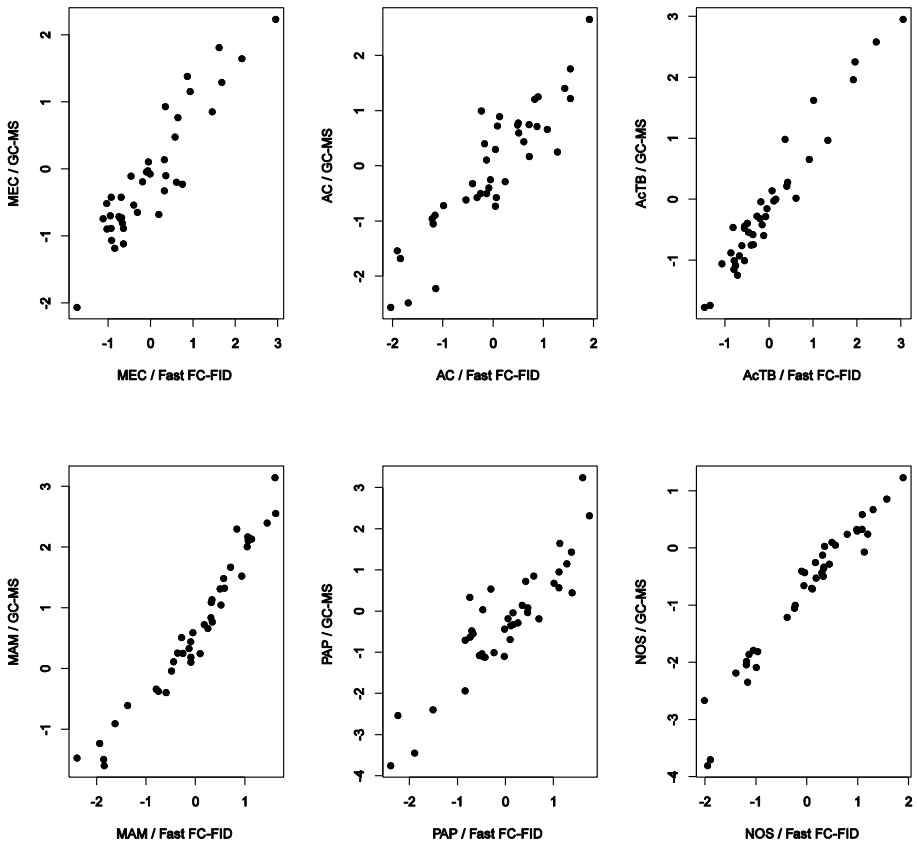


Figure 66. Etude de la corrélation entre les données GC-MS et Fast GC-FID pour chaque variable

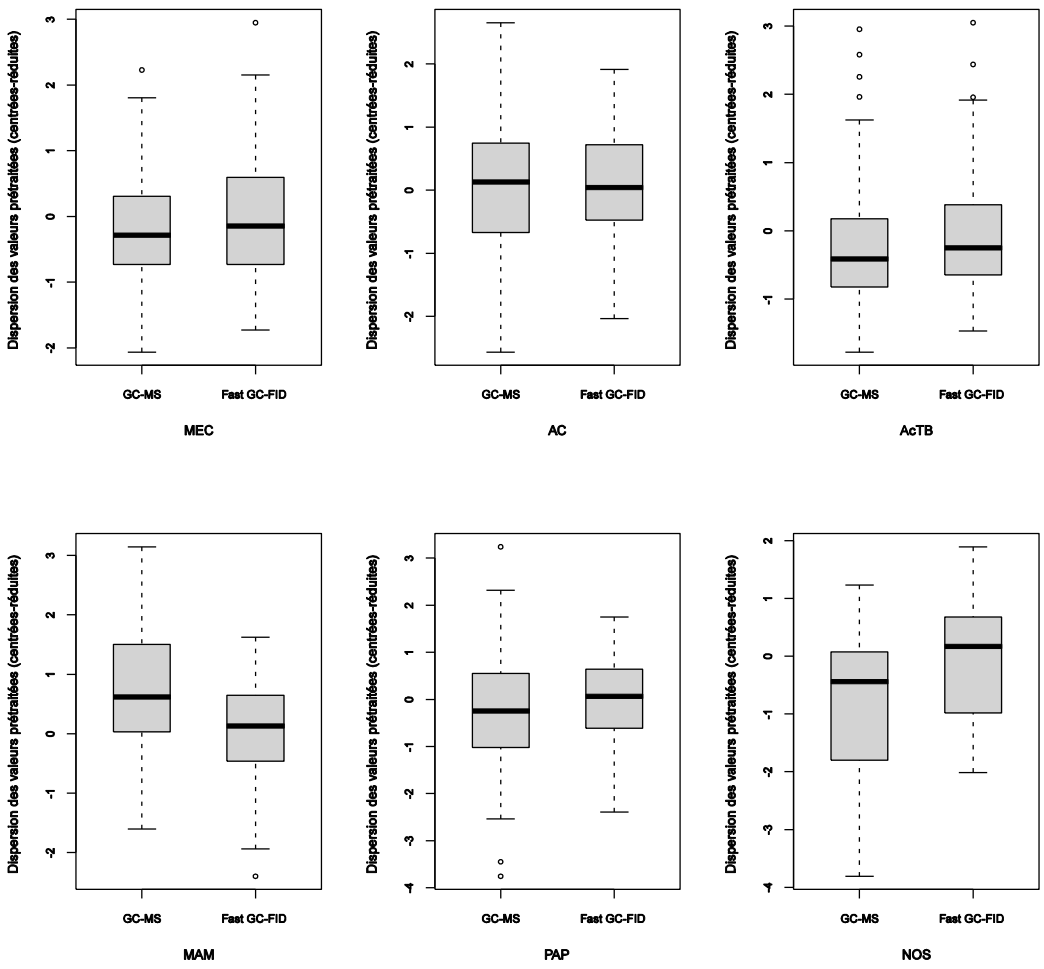


Figure 67. Comparaison de la distribution des valeurs obtenues pour chaque variable pour les résultats GC-MS et Fast GC-FID, après prétraitement des résultats

Distribution intra méthode

Comme l'illustrent la Figure 68 et la Figure 69, les deux méthodes présentent des performances comparables pour la séparation des populations d'échantillons liés et non liés comme en témoignent les motifs des distributions et les taux d'erreurs obtenus en fonction des valeurs de seuil appliquées (cf. Tableau 27). En effet, pour une valeur de seuil identique de 90, le taux de vrais positifs (VP) est à son maximum pour chacune des méthodes. Le taux de faux positifs (FP) pour la méthode Fast GC-FID n'est pas significativement plus élevé que celui obtenu en GC-MS. Lorsque la valeur de seuil est augmentée, on observe une diminution dans le taux de vrais positifs et de faux positifs même si ce dernier pour la Fast GC-FID reste plus élevé.

Ces résultats révèlent des performances de discrimination comparables à celles obtenues dans le cadre du scénario d'ajustement 1.2 pour APP 1, APP2 et APP4 lorsque les analyses sont effectuées en méthode rapide GC-MS (cf. Tableau 25 et Tableau 27).

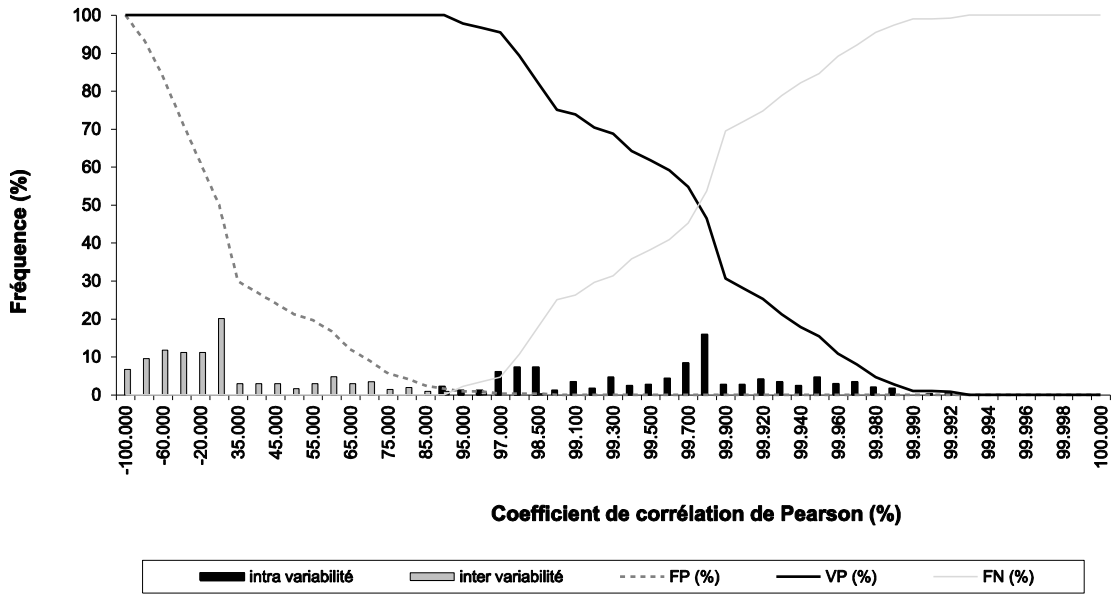


Figure 68. Distribution de l'intra- et l'inter variabilité pour la méthode GC-MS. Pour une question de clarté, l'échelle des valeurs du coefficient de corrélation de Pearson n'est pas linéaire.

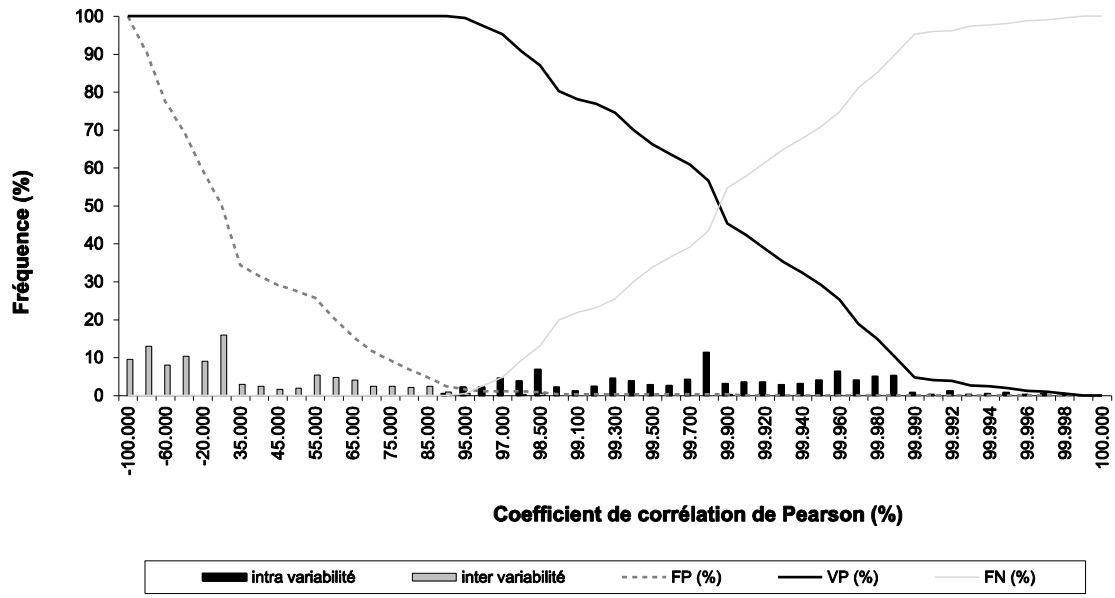


Figure 69. Distribution de l'intra- et l'inter variabilité pour la méthode Fast GC-FID. Pour une question de clarté, l'échelle des valeurs du coefficient de corrélation de Pearson n'est pas linéaire.

		VP (%)		FP (%)	
	Méthodes	**GC-MS**	**Fast GC-FID**	**GC-MS**	**Fast GC-FID**
Seuil	90	100	100	1.6	2.4
	97	95.4	95.3	0.3	1.1

Tableau 27. Performance de la discrimination pour le profilage chimique de l'héroïne, d'après les taux d'erreurs et les valeurs de seuil, pour chaque méthode analytique

D'après ces résultats, une bonne discrimination entre les échantillons dits liés et non liés est obtenue. Ainsi, une méthodologie de profilage chimique peut effectivement être implémentée pour les méthodes GC-MS et Fast GC-FID séparément, quelle que soit la manière dont les résultats du profilage seraient utilisés (en tant que soutien à l'enquête policière ou en tant qu'éléments de preuve). Par conséquent, une banque de données de profils partagée par ces deux méthodes peut être investiguée de la même manière que dans le cadre de l'étude intra méthode.

Distribution inter méthodes

Pour l'évaluation de la distribution inter méthodes, les profils peuvent être ajustés mathématiquement ou non, permettant ainsi d'évaluer l'utilité de cette approche d'optimisation de la similarité (concernant l'ajustement mathématique, cf. Chapitre 10). Sachant que dans le cadre de ce scénario d'ajustement il s'agit d'approvisionner la banque de données de référence GC-MS avec les résultats Fast GC-FID, la distribution de l'intra variabilité en GC-MS (cf. Figure 68) peut être utilisée comme référence pour déterminer si les résultats sont similaires.

D'après la méthodologie appliquée pour établir l'intra variabilité dans le cadre de l'étude inter méthodes, dans le cas où les résultats obtenus en GC-MS et en Fast GC-FID seraient similaires, aucun déplacement des échantillons dits liés vers les valeurs plus faibles de coefficient de corrélation de Pearson ne devrait être observé. En revanche, si les résultats s'avéraient être non similaires, un déplacement vers les valeurs de coefficient de corrélation de Pearson plus faibles devrait se produire.

Comme l'illustre la Figure 70, lorsque l'on compare l'intra variabilité résultant de la combinaison des résultats GC-MS et Fast GC-FID avec celle obtenue en GC-MS (en bleu sur la Figure 70), un déplacement significatif vers des valeurs de coefficient de corrélation de Pearson plus faibles se produit.

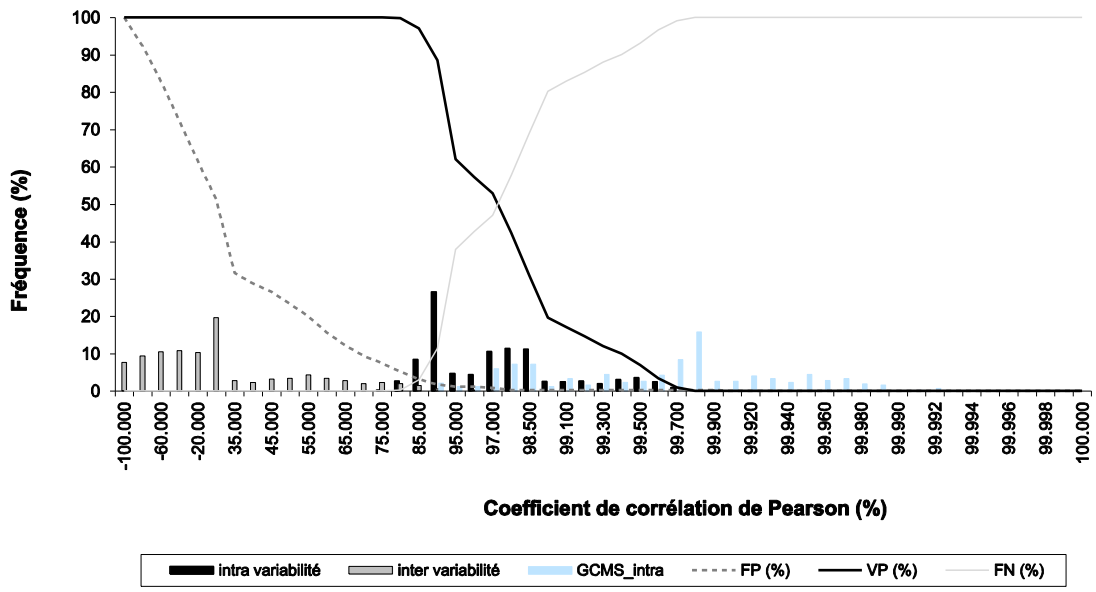


Figure 70. Distribution de l'intra- et l'inter variabilité quand les résultats provenant des méthodes GC-MS et Fast GC-FID sont combinés. Pour une question de clarté, l'échelle des valeurs du coefficient de corrélation de Pearson n'est pas linéaire.

Il s'agit d'une observation attendue dans la mesure où la variabilité entre des résultats provenant de méthodes partageant différentes technologies d'analyse de détection est plus importante que celle atteinte entre des résultats obtenus sur le même instrument analytique. Les taux d'erreurs obtenus ainsi que les valeurs du seuil de décision confirment cette observation (cf. Tableau 28). En effet, s'il est souhaité un taux de vrais positifs maximal, alors le seuil de décision doit être nécessairement diminué (de 90 en GC-MS à 75 lorsque les résultats sont combinés). Bien que le taux de faux positifs obtenu soit 4 fois plus élevé en comparaison de celui atteint en GC-MS, un tel taux pourrait être acceptable pour un laboratoire utilisant les résultats du profilage pour soutenir l'enquête policière.

		VP (%)	FP (%)
Seuil	75	100.0	7.4
	90	88.6	1.7
	95	62.1	1.1

Tableau 28. Performance de la discrimination pour le profilage chimique de l'héroïne lorsque les résultats GC-MS et Fast GC-FID sont combinés

Comme cela a déjà été mentionné, si les résultats du profilage font office d'éléments de preuve dans une affaire particulière, alors le taux de faux positifs (FP) doit être minimisé en sélectionnant une valeur de seuil approprié. Lorsque les données GC-MS et Fast GC-FID sont combinées, cet objectif peut être atteint en choisissant un seuil de coefficient de corrélation de Pearson de 95. En effet, un taux de faux positifs d'environ 1% est en conséquence obtenu. Cependant, le taux de vrais positifs (VP) obtenu n'est alors que de 62.1% impliquant que le laboratoire manque un nombre substantiel des liens chimiques existants (cf. Tableau 28).

La comparaison de ces performances avec celles déterminées pour l'étude inter méthodes du scénario d'ajustement 1.2 montre pour le scénario 2.1 une augmentation par un facteur d'environ 4 du taux de faux positifs lorsque l'on maximise le taux de vrais positifs. Lorsque le taux de faux positifs est minimisé, alors une baisse significative du taux de vrais positifs est constatée (environ 35% de liens ne sont plus détectés). Bien que les résultats d'un nombre relativement important d'instruments analytiques soient combinés dans le scénario 1.2, ce résultat se justifie par le partage par les méthodes des trois mêmes niveaux de paramètres analytiques (et les paramètres du « tune file » mis à part). Ainsi, la variabilité due à la combinaison de plusieurs instruments analytiques s'avère plus faible que la variabilité due à la combinaison de résultats provenant de méthodes GC-MS et Fast GC-FID, tous les niveaux de paramètres analytiques étant alors différents (en particulier la technologie d'analyse de détection).

Rappelons que c'est au laboratoire forensique d'estimer, d'après les performances de discrimination calculées, si le maintien d'une banque de données GC-MS avec des résultats obtenus en Fast GC-FID représente une amélioration significative dans le cadre de la lutte contre le trafic de produits stupéfiants (par exemple, grâce à la réduction du temps d'analyse des spécimens d'héroïne).

Sachant qu'un laboratoire pourrait se satisfaire des performances de séparation atteintes lors de la combinaison de résultats obtenus en GC-MS et Fast GC-FID, il est toutefois intéressant d'investiguer la possibilité d'améliorer l'efficacité de la méthodologie de profilage chimique lorsque de telles méthodes sont combinées grâce au recours à l'ajustement mathématique (cf. §10.1).

Maintenant que l'efficacité de la méthodologie de profilage a été estimée, il s'agit d'évaluer dans quelle mesure la structure des données est conservée lorsque des résultats FID y sont insérés (cf. Chapitre 5, §5.3). Le recours à l'ACP-CAH est particulièrement intéressant dans ce cadre-là car elle permet d'étudier chacun des spécimens séparément.

9.2 Conservation de la structure des données (sous-hypothèse 2.2)

L'ajustement des données Fast GC-FID est investigué dans cette partie, les données « GC-MS like » étant étudiées au §10.2. Avant d'étudier les performances d'ajustement réussi, l'étude des résultats préliminaires découlant d'une ACP appliquée permet de décrire la structure des données. La Figure 71 compare les loadings des échantillons selon la méthode analytique utilisée.

La Figure 71 illustre que pour les deux méthodes, les variables ayant le plus de poids sur CP1 sont les mêmes (MAM et PAP en particulier, AC et AcTB dans une moindre mesure). CP1 n'est quasiment pas influencée par MEC, pour les données GC-MS et Fast GC-FID. Sur CP2, les variables NOS et MEC ont le plus de poids, pour chaque jeu de données. A la différence du jeu de données Fast GC-FID, en GC-MS CP2 n'est quasiment pas influencée par PAP. Alors que sur CP1, AC a environ le même poids pour chacun des jeux de données, CP2 n'est quasiment pas influencée par cette dernière pour les données Fast GC-FID (cf. Figure 71).

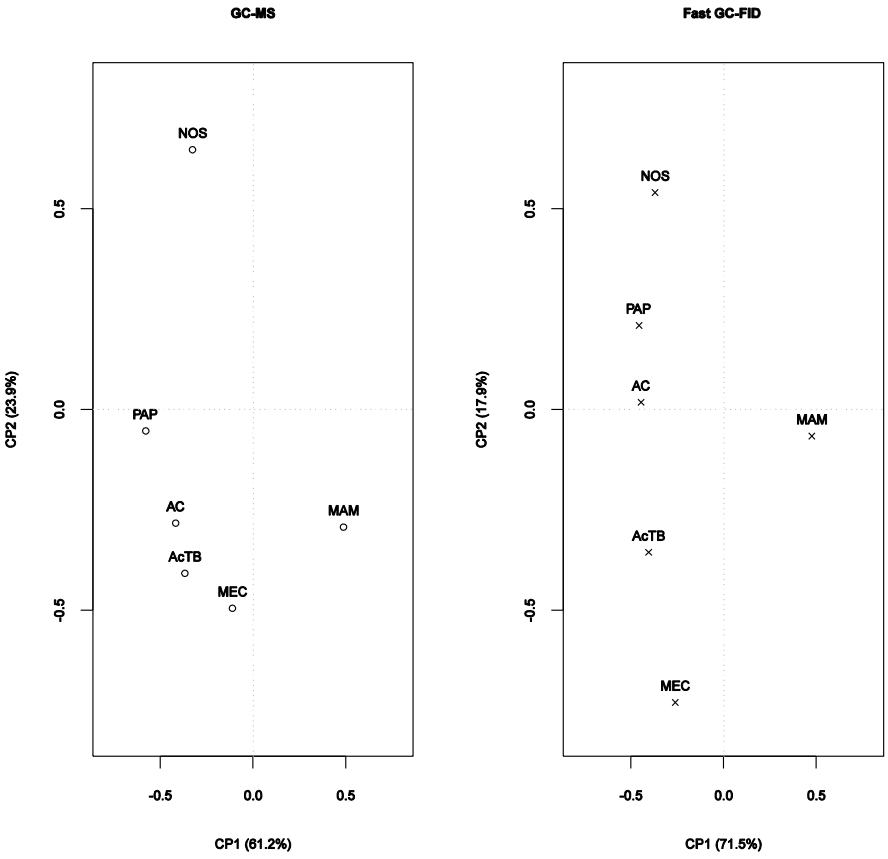


Figure 71. Distribution des loadings pour la GC-MS et la Fast GC-FID, respectivement

9.2.a Liens chimiques respectifs

Cette étude correspond à la situation A définie dans le cadre de la sous-hypothèse 2.2 (cf. §5.3, Figure 26). Il s'agit ici d'estimer si la classification des profils chimiques effectuée avec la méthode d'analyse de référence est similaire à celle réalisée avec la méthode Fast GC-FID pour l'échantillonnage choisi (situation A, cf. Figure 25, §5.3), d'après les distributions respectives de l'intra- et l'inter variabilité intra méthode (cf. §9.1.b). Ces informations peuvent s'avérer intéressantes pour décrire la distribution des profils chimiques pour chaque méthode respectivement avant de les combiner dans une même banque de données. Un grand nombre de discordances dans la classification chimique pourrait par exemple impacter la proximité des profils GC-MS et Fast GC-FID des spécimens correspondants dans la banque de données de référence.

Pour un seuil de décision de 97%, les deux méthodes présentent deux mêmes groupes de spécimens liés, respectivement (cf. Tableau 29).

Spécimens liés
203_05_09_8, 203_05_09_11 et 210_05_09_2
100_03_09_17, 267_07_09_1 et 291_07_09_1

Tableau 29. Spécimens respectivement liés entre eux selon les 2 méthodes analytiques d'après un seuil de décision de 97% (cf. §9.1.b)

Toutefois, à l'inverse de la classification réalisée en GC-MS, en Fast GC-FID un lien chimique est estimé entre les spécimens 098_03_09_3 et 098_03_09_4 tandis qu'aucun lien n'est réalisé entre les spécimens 113_03_09_1 et 277_07_09_2. Ces distinctions dans la classification sont certainement dues aux performances de discriminations différentes estimées au §9.1.b. La distinction dans la fonction réponse des détecteurs MS et FID (sélectivité/spécificité du MS) évoquée plus haut (cf. §9.1.b) engendre une différence dans la réponse analytique pouvant également jouer un rôle dans ces résultats.

9.2.b Conservation des classes chimiques

Ce paragraphe correspond à la situation B définie dans le cadre de la sous-hypothèse 2.2 (cf. §5.3, Figure 25). L'estimation de la similarité statistique des données par comparaison visuelle de la distribution des scores à l'aide de CP1 et CP2 s'avère difficile en raison du nombre important de données. De plus, uniquement les deux premières CPs seraient considérées impliquant que les variances expliquées ne tiennent pas compte des CPs suivantes, d'où une perte d'informations non négligeable quant à la distribution réelle des données et un risque d'élaborer des conclusions erronées. L'étude des performances d'ajustement réussi calculées à l'aide du processus ACP-CAH en sélectionnant les CPs expliquant plus de 95% de la variance des données permet d'évaluer plus objectivement la similarité statistique existant entre les résultats (cf. Chapitre 6).

La Figure 72 présente les performances d'ajustement réussi calculées pour les données Fast GC-FID (c'est-à-dire, sans ajustement mathématique) ainsi que les valeurs médianes du coefficient de corrélation de Pearson calculées pour tous les spécimens à chaque valeur de *h*.

Tandis qu'à une hauteur de 0.5 la performance calculée est d'environ 1%, elle atteint rapidement 90% pour une hauteur de 10 et se porte finalement à environ 97% lorsque le dendrogramme local est coupé à une hauteur de 20 (cf. Figure 72). En regard des différences analytiques que partagent les méthodes considérées, ces résultats s'avèrent positifs. D'après l'interprétation des résultats de l'ACP-CAH réalisée au Chapitre 7 (§7.5), ces résultats démontrent en effet que la majorité des profils Fast GC-FID sont effectivement transposés à proximité des profils GC-MS pour les spécimens correspondants dans l'espace de la banque de données de référence défini par les CPs considérées. Ainsi, les profils GC-MS et Fast GC-FID des spécimens respectifs partagent une certaine similarité statistique.

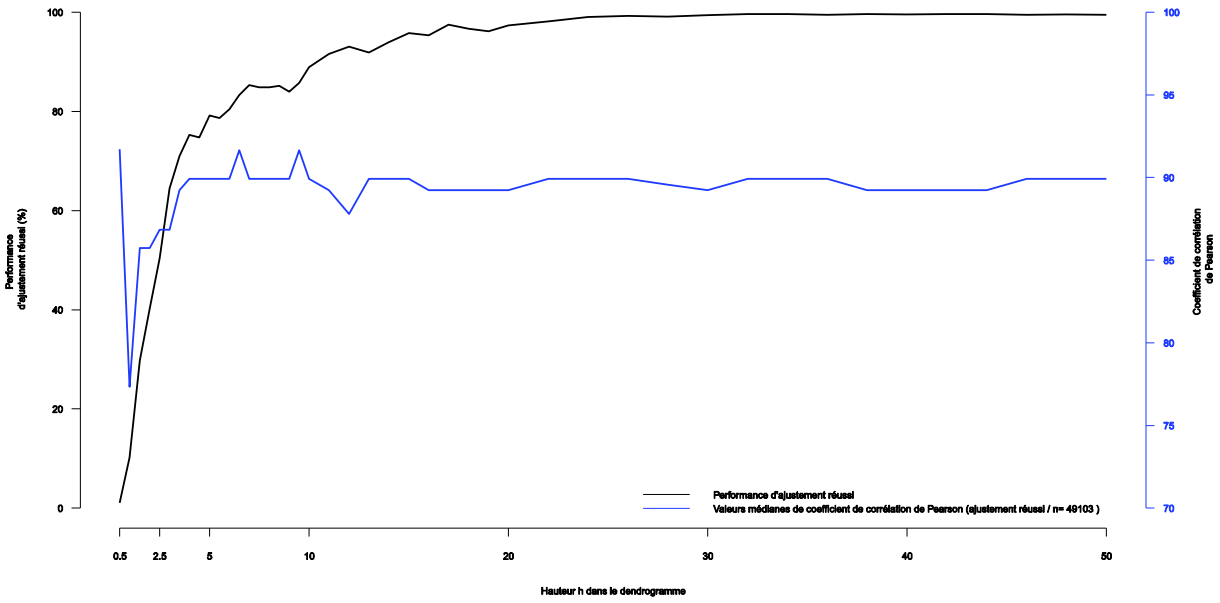


Figure 72. Performances d'ajustement réussi et valeurs médianes de coefficient de corrélation de Pearson en fonction de la hauteur *h* dans le dendrogramme local pour les données Fast GC-FID

Pour estimer ce degré de similarité statistique, il est proposé de se référer aux valeurs du coefficient de corrélation de Pearson obtenues lors de la comparaison des couples de profils GC-MS et Fast GC-FID pour chaque spécimen (cf. §6.5), dont les valeurs médianes calculées à chaque *h* sont présentées sur la Figure 72. Le recours aux distributions de l'intra- et l'inter variabilité inter méthodes est alors particulièrement utile pour juger du degré de similarité (cf. §9.1.b). Une fois une hauteur de 3.5 – 4 fixée, la courbe des valeurs médianes tourne autour des 90% de coefficient de corrélation de Pearson témoignant ainsi d'une similarité relativement élevée en regard de la distribution inter méthodes (lorsque les résultats MS et FID non ajustés mathématiquement sont combinés, VP atteint 88.6% pour un FP de 1.7% lorsqu'un seuil de 90% est fixé, cf. §9.1.b).

Pour estimer précisément si l'ajustement de chacun des spécimens est réussi et ainsi discuter de l'appartenance des profils respectifs à la même classe chimique (c'est-à-dire, dans le cas où ils seraient liés d'après le seuil de décision fixé dans le cadre de l'étude inter méthodes, situation B de la sous-hypothèse 2.2, cf. Figure 25, §5.3), il n'y a d'autre choix que d'investiguer les valeurs de coefficient de corrélation de Pearson estimées pour chaque couple de profils GC-MS et Fast GC-FID pour chaque spécimen correspondant (cf. §7.5). Les résultats devraient alors être confrontés à l'étude des distributions inter méthodes, selon la perspective d'utilisation des résultats du profilage (utilisation en tant que soutien de l'enquête policière ou en tant qu'élément de preuve dans une affaire particulière).

Ces résultats ne sont pas présentés ici dans la mesure où le but de cette recherche n'est pas de démontrer la faisabilité de l'approvisionnement d'une banque de données de résultats MS et FID. Il s'agit en effet de fournir une méthodologie permettant d'estimer cette faisabilité quel que soit le scénario d'ajustement considéré. Les résultats sont discutés en revanche au §10.2 pour démontrer l'apport de l'ajustement mathématique dans l'amélioration de la similarité statistique des profils chimiques.

9.3 Conclusion

La méthodologie proposée dans cette étude a permis l'estimation de résultats issus de méthodes différentes dans tous les niveaux de paramètres analytiques et en particulier dans la technologie d'analyse de détection (cf. Chapitre 5, Figure 23 et Tableau 12). De plus, le profil chimique n'est pas déterminé de la même manière pour chaque composé cible (aires des ions cibles en MS vs. aires de pics en FID). Malgré ces différences, l'étude des distributions de l'intra- et l'inter variabilité ainsi que l'ACP-CAH représentent des outils statistiques efficaces pour estimer la similarité des résultats et évaluer la possibilité du maintien d'une banque de données avec les méthodes considérées.

L'utilisation d'un échantillonnage limité mais représentatif donne rapidement des informations pertinentes quant à la faisabilité pratique du partage d'une banque de données par différentes méthodes. Alors que l'ACP-CAH permet de constater dans quelle mesure la structure des données est conservée, l'étude de l'intra- et l'inter variabilité permet d'estimer l'efficacité de la méthodologie de profilage une fois les résultats combinés selon l'objectif poursuivi par cette dernière (utilisation des résultats en tant que soutien de l'enquête policière ou en tant qu'éléments de preuve dans une affaire spécifique).

Les résultats expérimentaux obtenus dans le cadre du scénario d'ajustement 2.1 et investiguant les sous-hypothèses relatives à la conservation de la structure des données et à l'efficacité de la méthodologie de profilage **ne permettent pas de réfuter la seconde hypothèse de travail**. En effet, malgré la combinaison de résultats provenant de méthodes différentes selon les paramètres analytiques A_{DET}, B et $C_{SEP/DET}$, les performances obtenues s'avèrent satisfaisantes.

Chapitre 10 Etude de l'ajustement mathématique des résultats analytiques

Il s'agit dans ce chapitre d'évaluer l'efficacité de l'approche d'ajustement mathématique pour améliorer la similarité statistique des profils chimiques d'un même spécimen obtenus avec des méthodes analytiques différentes. De par l'assurance de la conservation de la structure des données qui en découlerait, le maintien d'une banque de données commune à diverses méthodes en serait facilité. L'étude des résultats de l'intra- et l'inter variabilité inter méthodes combinés à ceux de l'ACP-CAH permet d'évaluer le gain dans la similarité statistique des profils chimiques, en comparaison à la situation où les profils provenant de la méthode différente ne seraient pas ajustés mathématiquement (cf. Chapitre 9).

10.1 Efficacité de la méthodologie de profilage (inter méthodes)

Dans le cadre de l'étude de l'intra- et l'inter variabilité inter méthodes, en ajustant mathématiquement les données, l'idée principale consiste à obtenir une intra variabilité aussi fine que possible et plus proche des hautes valeurs de coefficient de corrélation de Pearson. En effet, en comparaison à la distribution de l'intra variabilité de la méthode de référence GC-MS, un décalage se produit vers les faibles valeurs de similarité en raison de la variabilité existant entre des résultats MS et FID. En quelques mots, l'ajustement mathématique vise à éviter le déplacement de la population des échantillons liés vers les valeurs de corrélation plus faibles ou au moins de diminuer le degré de ce déplacement.

Comme cela a été discuté dans la publication traitant de l'ajustement mathématique (Debrus et al., 2010) ainsi que dans le Chapitre 6 de cet ouvrage, plusieurs modèles mathématiques peuvent être testés et utilisés pour ajuster les données Fast GC-FID et obtenir des données « GC-MS like » (c'est-à-dire des données Fast GC-FID ajustés mathématiquement). Pour rappel, sachant que le but de la méthodologie consiste à approvisionner la banque de données de référence GC-MS avec les résultats obtenus en Fast GC-FID, alors les réponses du modèle (c'est-à-dire, le y du modèle mathématique) correspondent aux aires de pics GC-MS normalisées puis centrées-réduites (cf. §6.3.b).

Comme le §8.1 en a discuté, grâce aux relations mathématiques établies pour chaque composé du profil (c'est-à-dire, les règles d'ajustement), le degré de similarité entre les valeurs GC-MS et Fast GC-FID peut être évalué en calculant les coefficients R^2 ajusté et Q^2 respectifs selon le modèle mathématique appliqué. Les valeurs moyennes de ces coefficients calculées d'après les sets de calibration obtenus pour 100 itérations sont présentées dans le Tableau 30. De plus, une valeur moyenne est calculée pour estimer les capacités prédictives de chaque modèle mathématique.

	$R^2_{ajusté}$			Q^2		
Composés	**Linéaire**	**Quadratique**	**Cubique**	**Linéaire**	**Quadratique**	**Cubique**
MEC	0.827	0.829	0.820	0.826	0.829	0.808
AC	0.807	0.796	0.802	0.813	0.801	0.805
AcTB	0.943	0.936	0.936	0.942	0.933	0.930
6MAM	0.944	0.964	0.963	0.945	0.965	0.964
PAP	0.810	0.791	0.819	0.810	0.786	0.811
NOS	0.943	0.956	0.954	0.945	0.957	0.956
Moyenne	*0.879*	*0.878*	*0.883*	*0.880*	*0.878*	*0.879*

Tableau 30. Valeurs moyennes des R^2 ajusté et Q^2 entre les données GC-MS et Fast GC-FID en fonction des modèles mathématiques utilisés

D'après l'étude de ces coefficients, et tel que mentionné au §8.1, les trois modèles mathématiques appliqués ajustent et prédisent avec une efficacité similaire les données. L'observation des valeurs obtenues pour chaque variable démontre que la similarité des réponses analytiques des composés AcTB, 6MAM et NOS est élevée. Sachant qu'une valeur proche de 0.8 peut être considérée comme étant révélatrice d'un degré de similarité acceptable dans le cadre de ce scénario d'ajustement, chacun des modèles ajuste et prédit convenablement la relation existant entre les données GC-MS et Fast GC-FID.

Ces trois modèles mathématiques ont ainsi été utilisés pour ajuster mathématiquement chaque profil chimique dans le cadre de l'étude de l'intra- et de l'inter variabilité inter méthodes. Le modèle cubique, dont les équations mathématiques pour chaque composé sont présentées dans le Tableau 31 ci-dessous, a démontré de meilleurs résultats en termes de taux de vrais positifs atteint pour une même valeur de seuil de décision en comparaison aux modèles linéaire et quadratique. Ce sont donc les résultats obtenus pour celui-ci qui sont présentés dans ce paragraphe, d'autant que sa complexité relativement plus importante n'en reste pas moins acceptable.

Composés	Modèle cubique
MEC	$\mathrm{MEC}_{GC-MS} = \beta_3.\mathrm{MEC}^3_{FAST\ GC-FID} + \beta_2.\mathrm{MEC}^2_{FAST\ GC-FID} + \beta_1.\mathrm{MEC}_{FAST\ GC-FID} + \beta_0$
AC	$\mathrm{AC}_{GC-MS} = \beta_3.\mathrm{AC}^3_{FAST\ GC-FID} + \beta_2.\mathrm{AC}^2_{FAST\ GC-FID} + \beta_1.\mathrm{AC}_{FAST\ GC-FID} + \beta_0$
AcTB	$\mathrm{AcTB}_{GC-MS} = \beta_3.\mathrm{AcTB}^3_{FAST\ GC-FID} + \beta_2.\mathrm{AcTB}^2_{FAST\ GC-FID} + \beta_1.\mathrm{AcTB}_{FAST\ GC-FID} + \beta_0$
6MAM	$6\mathrm{MAM}_{GC-MS} = \beta_3.6\mathrm{MAM}^3_{FAST\ GC-FID} + \beta_2.6\mathrm{MAM}^2_{FAST\ GC-FID} + \beta_1.6\mathrm{MAM}_{FAST\ GC-FID} + \beta_0$
PAP	$\mathrm{PAP}_{GC-MS} = \beta_3.\mathrm{PAP}^3_{FAST\ GC-FID} + \beta_2.\mathrm{PAP}^2_{FAST\ GC-FID} + \beta_1.\mathrm{PAP}_{FAST\ GC-FID} + \beta_0$
NOS	$\mathrm{NOS}_{GC-MS} = \beta_3.\mathrm{NOS}^3_{FAST\ GC-FID} + \beta_2.\mathrm{NOS}^2_{FAST\ GC-FID} + \beta_1.\mathrm{NOS}_{FAST\ GC-FID} + \beta_0$

Tableau 31. Modèles cubiques utilisés pour l'ajustement de tous les composés (les termes mathématiques sont identifiés en fonction du composé et les indices correspondent à la méthode analytique)

En comparant visuellement la distribution des valeurs prétraitées pour chacune des variables selon que les données aient été ajustées mathématiquement (cf. Figure 73) ou non (cf. Figure 67), on observe une nette amélioration de la similarité des données (valeurs médianes respectives plus proches pour les composés MAM et NOS, par exemple).

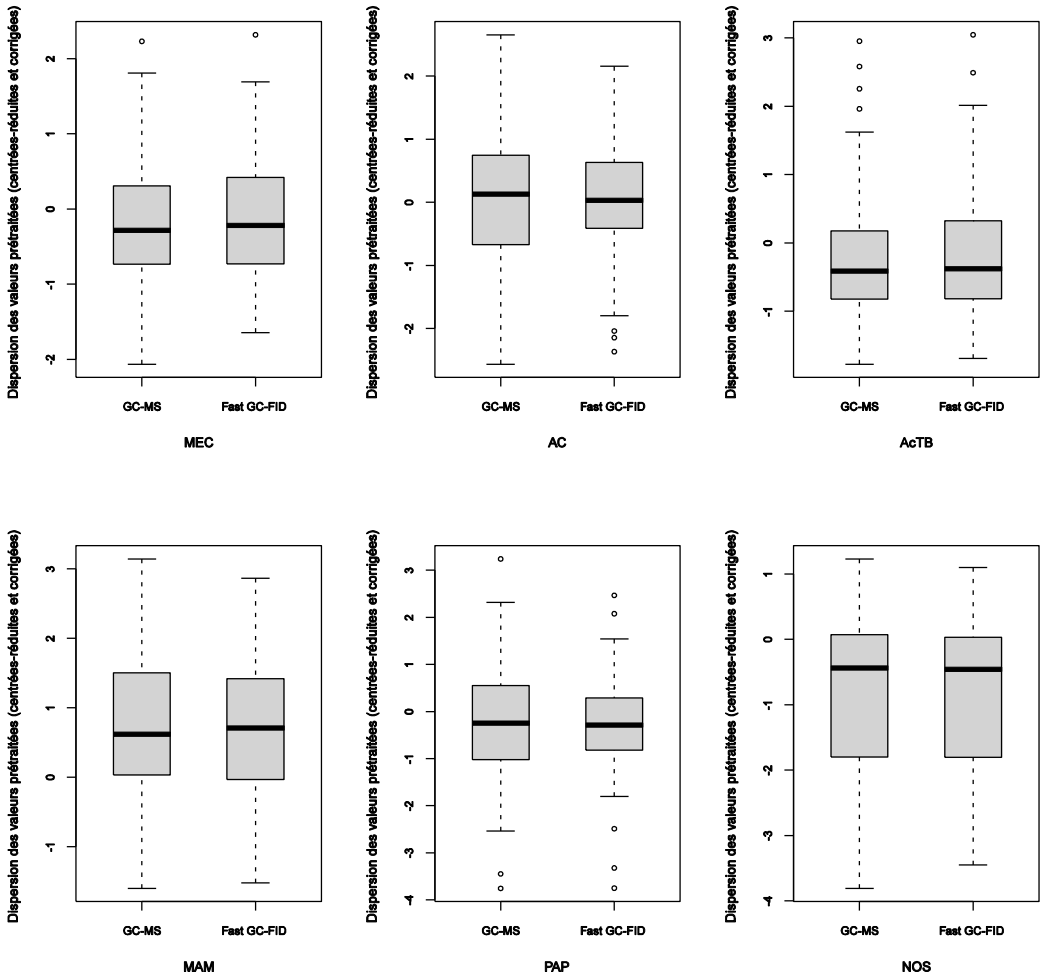


Figure 73. Comparaison de la distribution des variables prétraitées pour les données GC-MS et « GC-MS like » (c'est-à-dire, après ajustement mathématique des résultats Fast GC-FID, ici avec le modèle cubique).

La Figure 74 présente quant à elle la distribution de l'intra- et l'inter variabilité obtenue lorsque les résultats GC-MS et « GC-MS like » sont combinés. Le Tableau 32 illustre les taux d'erreurs calculés selon les seuils de décision définis dans un premier temps dans une optique de maximisation du taux de vrais positifs (utilisation des résultats du profilage pour soutenir l'enquête policière) puis dans un second temps dans une optique de minimisation du taux de faux positifs (résultats du profilage faisant office d'élément de preuve). Après comparaison de ces résultats à ceux obtenus sans ajustement mathématique (cf. Figure 70 et Tableau 28), il est possible d'évaluer l'impact d'une telle approche d'optimisation de la similarité statistique des résultats.

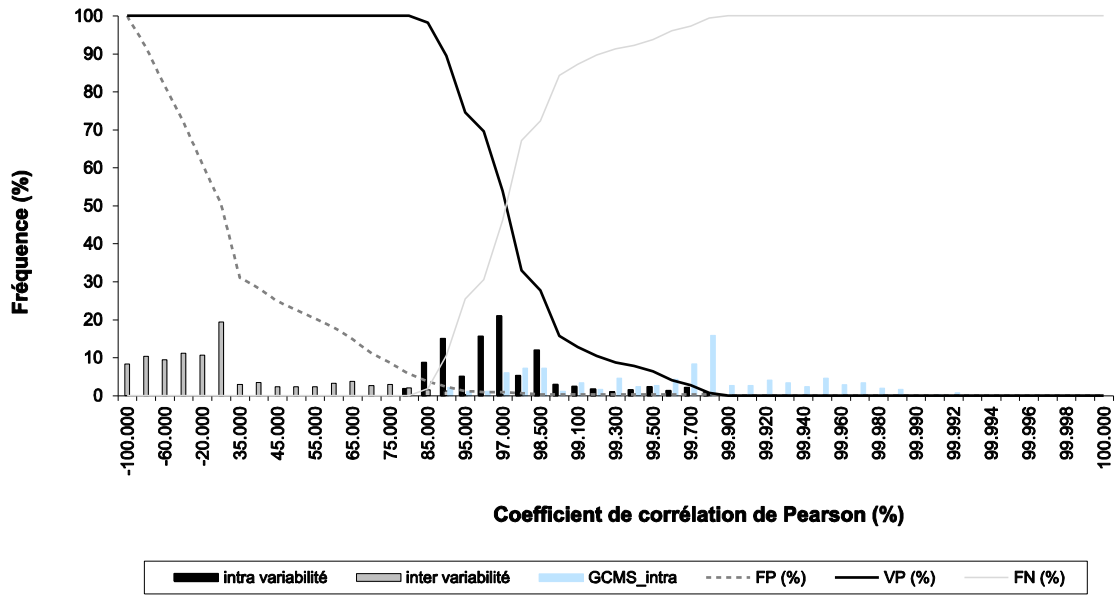


Figure 74. Distribution de l'intra- et l'inter variabilité pour la combinaison de résultats GC-MS et « GC-MS like » (c'est-à-dire, après ajustement mathématique des résultats Fast GC-FID, ici avec le modèle cubique). Pour une question de clarté, l'échelle des valeurs du coefficient de corrélation de Pearson n'est pas linéaire.

Une nette amélioration de la similarité des résultats est clairement observée. Bien qu'un déplacement de l'intra variabilité vers les valeurs de corrélation plus faibles se produise toujours en comparaison à la distribution obtenue en GC-MS (cf. Figure 74), celui-ci s'avère à présent moins important que sans ajustement mathématique (cf. Figure 70).

		VP (%)	FP (%)
Seuil	80	100	5.7
	95	74.5	1.1

Tableau 32. Performance de la discrimination pour le profilage chimique de l'héroïne lorsque les résultats GC-MS et « GC-MS like » sont combinés

Les taux d'erreurs obtenus confirment cette tendance (cf. Tableau 32). Grâce à l'ajustement des résultats Fast GC-FID avec le modèle cubique, la valeur de coefficient de corrélation de Pearson faisant office de seuil de décision peut être plus élevée (80) tout en ayant VP à son maximum et en obtenant même une diminution dans FP. L'amélioration de la similarité est démontrée également quand un seuil de coefficient de corrélation de Pearson de 95 est sélectionné, VP étant plus élevé que sans ajustement mathématique. Par conséquent, le laboratoire pourrait détecter plus de liens chimiques qui existent effectivement en appliquant l'ajustement mathématique que si ce dernier n'était pas réalisé (cf. Figure 74 et Tableau 32).

10.2 Conservation de la structure des données

A la différence de l'étude de l'intra- et l'inter variabilité, l'ACP-CAH permet une investigation de chacun des spécimens séparément. Avant d'étudier les performances d'ajustement réussi, l'étude des résultats préliminaires de l'ACP appliquée permet de décrire la structure des données. La Figure 75 compare les loadings des échantillons selon la méthode analytique utilisée et une fois les données Fast GC-FID mathématiquement ajustées (c'est-à-dire, les données « GC-MS like »).

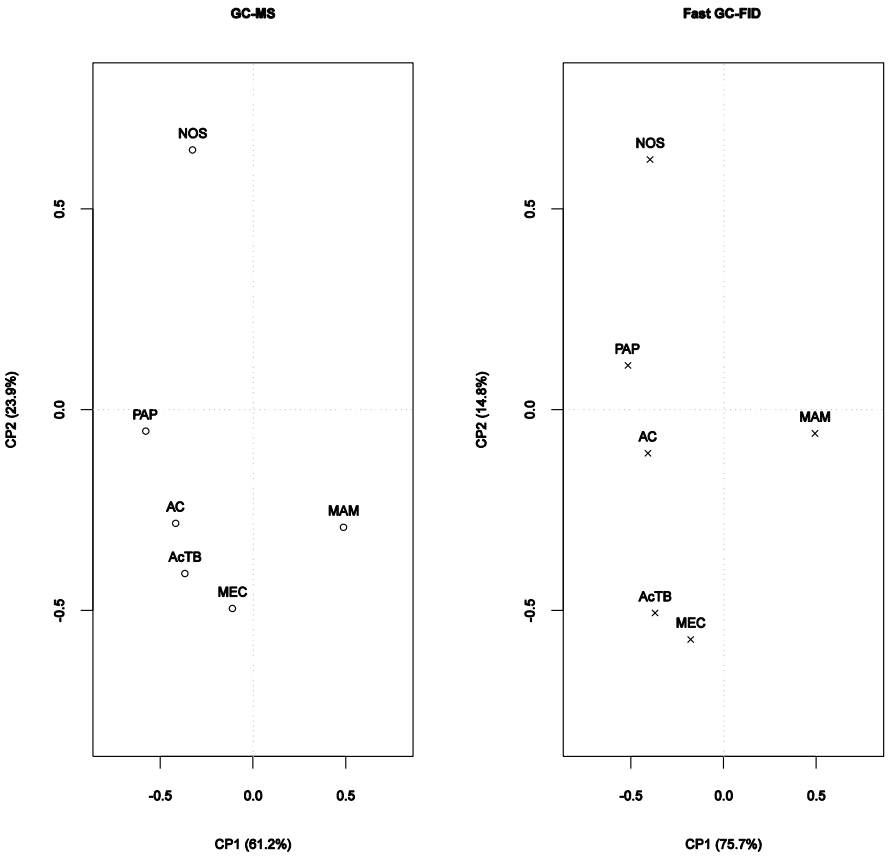


Figure 75. Distribution des loadings pour les données GC-MS et « GC-MS like » (c'est-à-dire les données Fast GC-FID ajustées mathématiquement, ici avec le modèle cubique), respectivement

L'idée principale dans cette section étant de démontrer l'utilité de l'ajustement mathématique, cette figure se compare à celle obtenue sans ajustement mathématique des données Fast GC-FID (cf. Figure 71). Sur la base de la distribution des données sur CP1 et CP2, les différences principales dans la distribution de ces derniers consistent dans le rapprochement des variables MEC et AcTB (valeurs d'autant plus négatives sur CP1), le rapprochement de PAP vers une valeur nulle sur CP2 et le passage de AC en valeur négative sur CP2, devenant alors plus similaire à la distribution des loadings pour les données GC-MS.

10.2.a Performances d'ajustement

La Figure 76 présente les performances d'ajustement réussi calculées pour les sets de validation sur 100 itérations lorsque les données sont ajustées mathématiquement avec les modèles linéaire, quadratique ou cubique, en comparaison aux performances obtenues sans aucun ajustement mathématique.

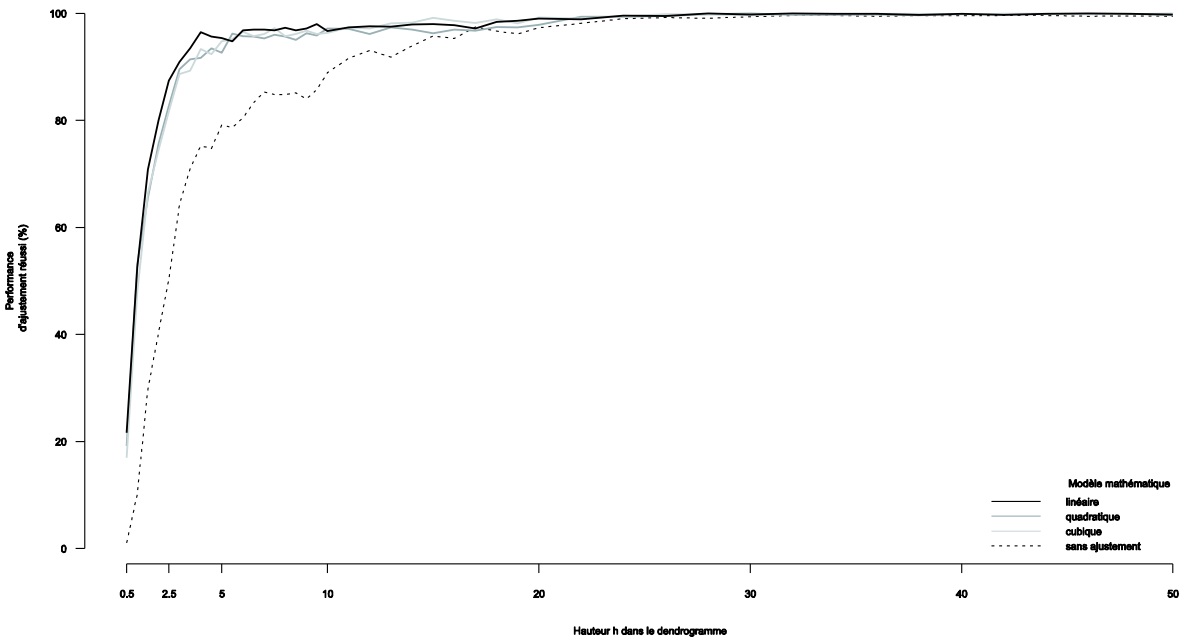


Figure 76. Illustration de l'influence de l'ajustement mathématique selon le modèle mathématique appliqué à l'aide des performances d'ajustement réussi en fonction de la hauteur *h* dans le dendrogramme local pour le scénario d'ajustement 2.1

Il est avant tout intéressant de remarquer que les trois modèles performent de manière similaire, ce qui est en adéquation avec les résultats présentés au Tableau 30, d'où l'existence d'une certaine corrélation entre les valeurs de Q^2 obtenues à partir du set de calibration et les performances d'ajustement réussi calculées sur le set de validation (cf. Figure 76). La relation de proportionnalité existant entre la réponse analytique et la concentration en GC-MS ou en GC-FID peut expliquer les performances d'ajustement particulièrement élevées à des hauteurs relativement faibles lorsque le modèle linéaire est appliqué et ceci malgré la différence dans la nature des données brutes évoquée à plusieurs reprises (sélectivité/spécificité du MS).

Sur la base des moyennes des performances d'ajustement calculées dans chacun des cas de figure, l'amélioration de la similarité statistique grâce à l'ajustement mathématique (en particulier avec le modèle mathématique le plus simple, le modèle linéaire) est clairement établie. Alors que la performance s'élevait à 1% sans ajustement mathématique des données (cf. Figure 72), à présent elle atteint environ 20% (cf. Figure 76). Non seulement les valeurs obtenues à de faibles valeurs de *h* sont plus élevées mais surtout les performances d'ajustement réussi atteignent plus rapidement des valeurs proches des 100% (cf. Figure 76). Sachant que le mode de calcul des performances dans l'ACP-CAH se fait d'après la distance existant entre les profils chimiques de chacun des spécimens dans le dendrogramme local (cf. Chapitre 6), cela implique nécessairement une plus grande proximité entre ces derniers dans l'espace de la banque de données de référence défini par les CPs considérées. Cette distance plus faible entre eux résulte sans aucun doute d'une similarité statistique plus élevée une fois que les données Fast GC-FID ont été mathématiquement ajustées.

L'observation du nombre de fois où l'ajustement est considéré comme étant réussi (c'est-à-dire, des profils dans le même cluster en fonction de la hauteur *h* dans le dendrogramme local) selon que l'on compare des données GC-MS et Fast GC-FID ou GC-MS et « GC-MS like » démontre également à quel point l'ajustement mathématique améliore la similarité statistique des profils chimiques (cf. Figure 77).

En effet, on observe une augmentation significative du nombre de profils pour lesquels l'ajustement est considéré comme réussi à de faibles valeurs de *h* (cf. Figure 77), alors que la présence de deux profils dans un même cluster défini par de telles hauteurs s'avère logiquement plus difficile (cf. Chapitre 7). De plus, la distribution des profils dont l'ajustement n'est pas réussi est plus étroite et moins intense : au-delà d'une hauteur de 4 il n'y a quasiment plus de profils dont l'ajustement n'est pas réussi (ce qui implique que dans le même temps, l'ajustement d'environ tous les profils est réussi). Finalement, l'ajustement est réussi pour tous les profils « GC-MS like » à des valeurs de *h* plus faibles que pour les profils Fast GC-FID (cf. Figure 77).

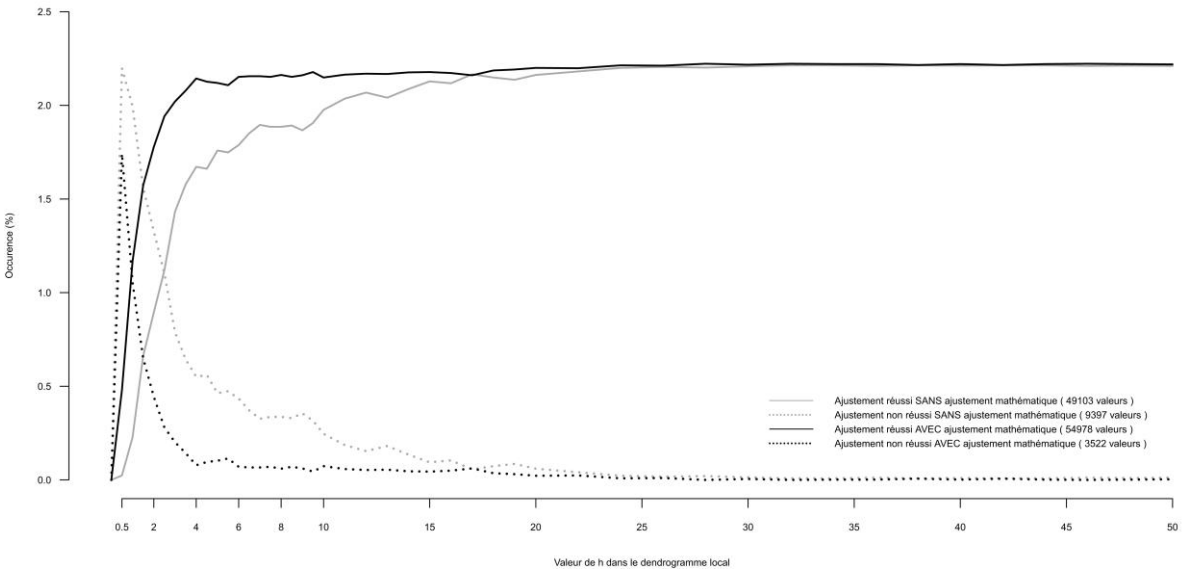


Figure 77. Evolution du nombre de valeurs par valeur de *h* que l'ajustement soit réussi ou non, après ajustement mathématique des données selon le modèle linéaire

10.2.b Valeurs du coefficient de corrélation de Pearson

Après comparaison avec la Figure 72, l'étude des valeurs médianes de coefficient de corrélation de Pearson estimées lors des comparaisons des profils GC-MS et « GC-MS like » des spécimens correspondants, pour l'ensemble des sets de validation et à chaque valeur de *h*, illustre le gain en similarité statistique (cf. Figure 78).

La courbe correspondant aux valeurs médianes s'élève à présent constamment au-delà de 95% de coefficient de corrélation de Pearson, valeur médiane jamais atteinte sans ajustement mathématique des données (cf. Figure 72 et Figure 78). Comme cela a déjà été discuté au §7.3, d'après les distributions de l'intra- et l'inter variabilité inter méthodes lorsque les résultats GC-MS et Fast GC-FID (après ajustement linéaire) sont combinés, une telle valeur médiane témoigne d'une probabilité élevée que les profils GC-MS et « GC-MS like » soient effectivement liés (cf. §7.3, Figure 55).

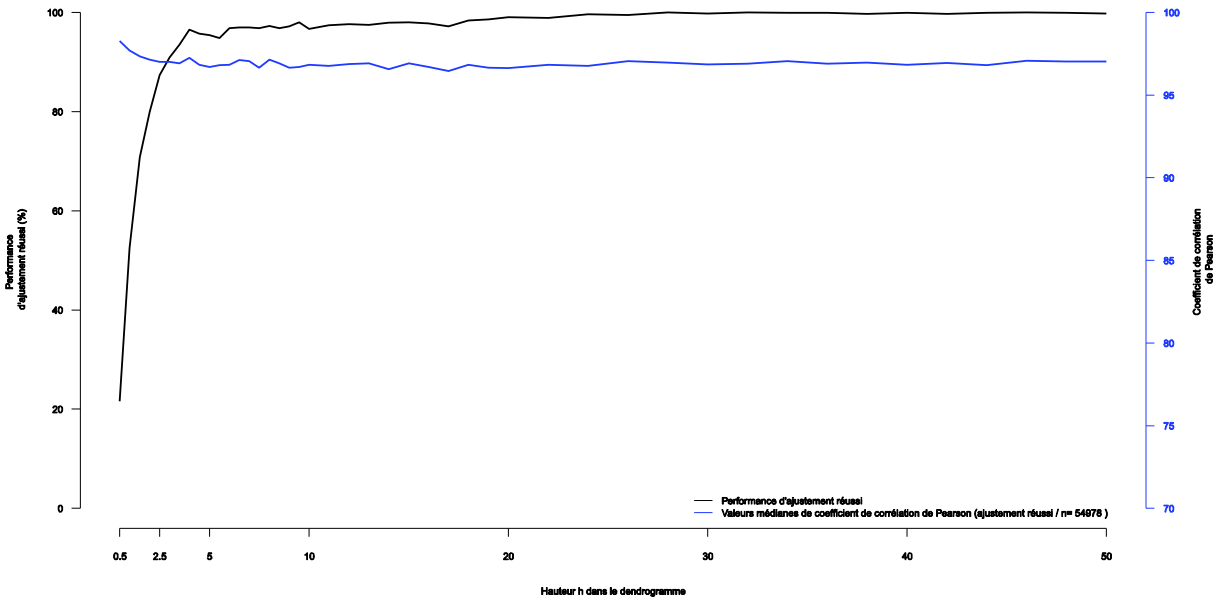


Figure 78. Performances d'ajustement réussi et valeurs médianes de coefficient de corrélation de Pearson en fonction de la hauteur *h* dans le dendrogramme local après ajustement mathématique selon le modèle linéaire des données Fast GC-FID (c'est-à-dire, des données « GC-MS like »)

Dans la lignée des observations précédentes, l'étude des valeurs de coefficient de corrélation de Pearson calculées lors de la mesure de la similarité pour chacun des couples GC-MS vs. Fast GC-FID ou GC-MS vs. « GC-MS like » pour les spécimens considérés témoigne de l'intérêt d'appliquer l'ajustement mathématique (cf. Figure 79).

La distribution des valeurs de similarité lorsque les profils sont ajustés se concentre vers les hautes valeurs de coefficient de corrélation de Pearson, comme l'illustre la médiane des boxplots respectifs (cf. Figure 79). Toutefois, les distributions des profils « non ajustés » et « ajustés » se recouvrent, démontrant l'obtention de hautes valeurs de similarité sans nécessairement effectuer l'ajustement mathématique (ce qui, en soi, représente un résultat encourageant dans l'optique de maintenir une banque de données commune malgré les différences analytiques existantes).

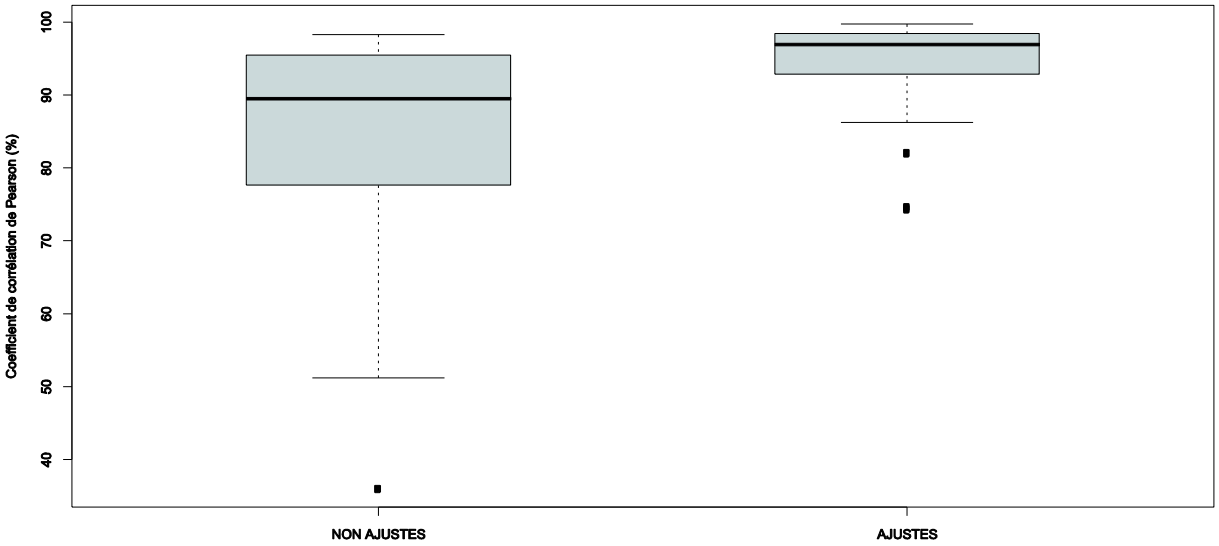


Figure 79. Comparaison de la distribution respective des valeurs de coefficient de corrélation de Pearson lorsque les profils chimiques sont ajustés à l'aide du modèle linéaire ou non (scénario d'ajustement 2.1)

Une observation plus précise de la répartition des valeurs du coefficient de corrélation de Pearson s'impose alors pour évaluer la part que représentent les valeurs élevées parmi la population des « non ajustés » et la comparer à celle de la population des « ajustés ». Ainsi, la Figure 80 et la Figure 81 comparent la répartition des valeurs de coefficient de corrélation de Pearson, que les données Fast GC-FID aient été ajustées avec le modèle linéaire (cf. Figure 81) ou non (cf. Figure 80).

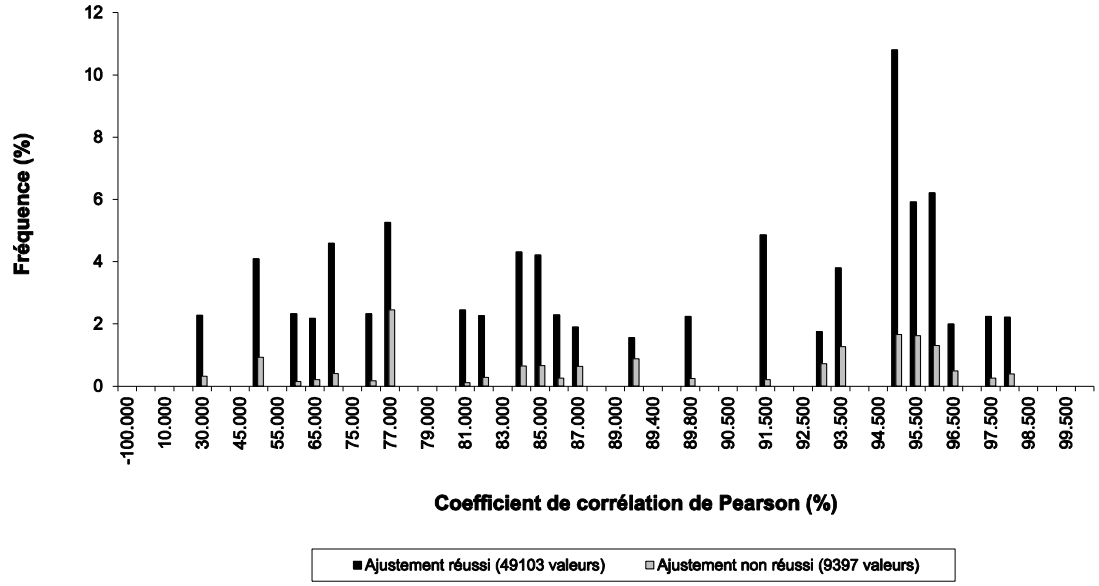


Figure 80. Fréquence d'apparition des valeurs du coefficient de corrélation de Pearson calculé pour les profils GC-MS et Fast GC-FID pour chaque spécimen correspondant, sans ajustement mathématique. Pour une question de clarté, l'échelle des valeurs du coefficient de corrélation de Pearson n'est pas linéaire.

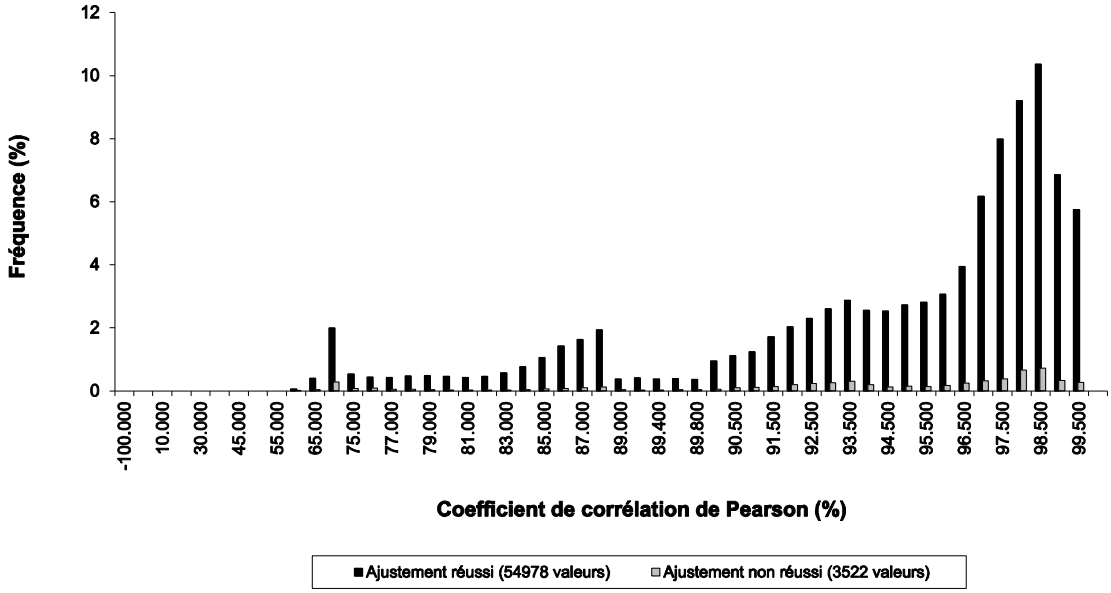


Figure 81. Fréquence d'apparition des valeurs du coefficient de corrélation de Pearson calculé pour les profils GC-MS et « GC-MS like » (modèle linéaire) pour chaque spécimen correspondant. Pour une question de clarté, l'échelle des valeurs du coefficient de corrélation de Pearson n'est pas linéaire.

Alors que parmi les profils « non ajustés » mathématiquement ceux dont l'ajustement a été jugé « non réussi » représentent environ 16% de l'ensemble des valeurs, une fois ajustés mathématiquement ils ne représentent plus qu'environ 6% de l'ensemble des valeurs (au nombre total de 58500).

Que l'ajustement ait été considéré comme réussi ou non, les valeurs de coefficient de corrélation de Pearson de 90 à 100% représentent environ 48% et 84% de l'ensemble des valeurs pour des profils non ajustés et ajustés mathématiquement, respectivement (cf. Figure 80 et Figure 81). Dans l'intervalle de 95 à 100%, les proportions sont d'environ 35% et 62% pour des profils non ajustés et ajustés mathématiquement, respectivement. Ainsi, alors que pour les profils ajustés mathématiquement les hautes valeurs de coefficient de corrélation de Pearson représentent la majeure partie des valeurs de similarité calculées, lorsque les profils ne sont pas ajustés mathématiquement, les hautes valeurs de coefficient de corrélation de Pearson représentent moins de la moitié de l'ensemble des valeurs de similarité.

L'ajustement mathématique influence de manière importante les mesures de similarité calculées entre les profils chimiques comme en témoigne le déplacement de la distribution vers les hautes valeurs de coefficient de corrélation de Pearson (cf. Figure 81), en comparaison à la distribution obtenue sans ajustement mathématique des profils (cf. Figure 80). Il est d'ailleurs intéressant de remarquer qu'également pour des profils dont l'ajustement a été jugé « non réussi », les valeurs de similarité sont plus élevées une fois les profils ajustés mathématiquement (apparition de valeurs au-delà de 98.5%, cf. Figure 81). Que l'ajustement ne soit dès lors pas réussi s'explique par l'influence de la distribution locale des données (cf. §7.1). L'examen en détail des coefficients de corrélation de Pearson estimés pour tous les spécimens des sets de validation pour toutes les hauteurs *h* démontre que ce cas de figure concerne particulièrement le spécimen 066_02_09_6_2 (cf. §7.1).

Le Tableau 33 présente pour toutes les itérations et les hauteurs *h* considérées les valeurs moyennes du coefficient de corrélation de Pearson calculées pour tous les spécimens des sets de validation, entre un profil GC-MS et un profil Fast GC-FID après ajustement mathématique (selon le modèle linéaire, colonne nommée « Ajustés ») ou non (colonne nommée « Non ajustés »), et que l'ajustement ait été considéré comme étant réussi ou non lors de l'étape locale de l'ACP-CAH.

	Non ajustés	**Ajustés**	
Spécimens	**Echec / Réussite**	**Echec**	**Réussite**
035_01_09_7_1	84.79	94.11 ± 1.78	94.51 ± 1.67
039_01_09_2	77.25	81.81 ± 3.07	82.06 ± 2.63
042_01_09_1	93.60	96.42 ± 0.80	96.65 ± 0.84
066_02_09_6_2	62.12	71.52 ± 3.83	74.39 ± 4.28
082_02_09_2	85.73	96.39 ± 0.82	96.48 ± 0.92
083_02_09_4	95.98	97.64 ± 0.63	97.74 ± 0.79
098_03_09_2	76.33	94.17 ± 0.96	94.68 ± 1.26
098_03_09_3	91.71	85.83 ± 1.55	86.22 ± 1.62
098_03_09_4	69.68	96.91	97.56 ± 0.60
100_03_09_17	96.72	99.10 ± 0.31	99.35 ± 0.27
113_03_09_1	77.94	95.85 ± 0.98	96.14 ± 0.88
166_04_09_3	95.47	91.77 ± 0.82	91.74 ± 0.90
169_04_09_1	36.02	87.67 ± 2.30	89.16 ± 2.40
179_04_09_2	81.23	97.73 ± 0.17	98.48 ± 0.57
201_05_09_2	95.45	97.91 ± 0.50	97.95 ± 0.62
203_05_09_11	96.33	98.71 ± 0.33	98.73 ± 0.37
203_05_09_8	95.78	98.84 ± 0.36	98.77 ± 0.42
205_05_09_1	51.17	75.09 ± 2.42	74.65 ± 2.66
210_05_09_2	98.29	99.77 ± 0.16	99.73 ± 0.21
213_05_09_1	72.34	97.13 ± 0.36	98.63 ± 0.58
221_05_09_2	95.42	97.61 ± 0.47	97.58 ± 0.55
226_05_09_1	53.60	97.64	97.77 ± 0.69
255_06_09_1	89.91	94.31 ± 0.92	94.39 ± 1.01
258_06_09_3	87.80	-	99.47 ± 0.36

Tableau 33. Valeurs moyennes du coefficient de corrélation de Pearson estimées, pour tous les spécimens, entre un profil GC-MS et un profil Fast GC-FID après ajustement mathématique (modèle linéaire, colonne nommée « Ajustés ») ou non (colonne nommée « Non ajustés »), que l'ajustement ait été considéré réussi ou non (valeurs issues des résultats obtenus pour les sets de validation).

	Non ajustés	Ajustés	
Spécimens	**Echec / Réussite**	**Echec**	**Réussite**
267_07_09_1	95.16	98.09 ± 0.47	98.17 ± 0.41
277_07_09_2	84.01	97.86 ± 0.82	98.53 ± 0.50
291_07_09_1	96.05	98.68 ± 0.37	98.78 ± 0.33
341_09_09_3	85.94	88.23 ± 1.33	88.40 ± 1.50
342_09_09_6	93.32	97.18 ± 0.45	97.26 ± 0.47
344_09_09_3	96.15	92.64 ± 1.14	92.49 ± 1.14
344_09_09_4	91.65	98.24 ± 0.47	98.31 ± 0.71
362_09_09_1	86.85	93.31 ± 1.10	93.25 ± 1.32
362_09_09_2	95.55	99.27 ± 0.34	99.45 ± 0.45
373_09_09_1	97.79	98.33 ± 0.57	98.35 ± 0.49
374_09_09_5	95.12	95.23 ± 0.70	95.33 ± 0.74
398_10_09_3	93.91	93.49 ± 0.58	93.62 ± 0.63
401_10_09_2	89.22	93.62 ± 1.13	93.34 ± 0.99
445_11_09_1_4	82.02	96.59 ± 0.53	97.41 ± 0.71
461_11_09_3	74.46	88.07 ± 1.75	88.28 ± 1.79
479_12_09_2	77.33	91.49 ± 1.57	91.73 ± 1.50

Tableau 33 (suite). Valeurs moyennes du coefficient de corrélation de Pearson estimées, pour tous les spécimens, entre un profil GC-MS et un profil Fast GC-FID après ajustement mathématique (colonne nommée « Ajustés ») ou non (colonne nommée « Non ajustés »), que l'ajustement ait été considéré réussi ou non (valeurs issues des résultats obtenus pour les sets de validation).

La présence d'un écart-type pour les valeurs moyennes du coefficient de corrélation de Pearson des profils ajustés se justifie par la création pour une certaine valeur de *h* d'autant de règles d'ajustement que de sets de calibration. En conséquence, pour chaque spécimen des sets de validation correspondants, le profil « GC-MS like » n'est pas toujours décrit par les mêmes valeurs pour chacune des 6 variables et donc une mesure du coefficient de corrélation de Pearson différente est obtenue lors de la comparaison des profils GC-MS et « GC-MS like » (cf. Tableau 33). Concernant les profils « non ajustés », lors de la mesure de la similarité entre un profil GC-MS et un profil Fast GC-FID (c'est-à-dire, sans ajustement mathématique), la même valeur du coefficient de corrélation de Pearson est toujours obtenue, d'où l'absence d'écart-type (cf. Tableau 33).

De manière générale, on observe une amélioration de la similarité statistique une fois les profils ajustés (par exemple, pour le spécimen 169_04_09_1 on passe d'une valeur d'environ 36 à une valeur de 87 ou 89 selon que l'ajustement soit estimé réussi ou non par la suite, cf. Tableau 33). Toutefois, pour certains spécimens, le coefficient de corrélation de Pearson pour le profil non ajusté est plus élevé que pour le profil ajusté mathématiquement (par exemple, les spécimens 098_03_09_3, 166_04_09_3 et 344_09_09_3).

Une telle observation démontre que, dans une relative faible mesure, les règles d'ajustement ne sont pas adaptées à tous les échantillons. Sachant que la création des sets de calibration est aléatoire pour chacune des 100 itérations et que les règles d'ajustement sont établies d'après ces sets (représentant les 2/3 de l'échantillonnage), les valeurs de coefficient de corrélation de Pearson calculées entre les profils GC-MS et « GC-MS like » peuvent varier selon l'itération considérée à chaque valeur de *h*. De plus, il est possible que ces spécimens présentent un profil bien particulier et différent de ceux des autres spécimens de l'échantillonnage impliquant que des relations mathématiques linéaires élaborées à l'aide de l'échantillonnage choisi ne permettent pas de les ajuster efficacement. Pour investiguer cette hypothèse, la Figure 82 représente la distribution des profils chimiques par variable avec les trois échantillons mentionnés ci-dessus mis en avant.

Sur la seule base de cette représentation graphique il s'avère pourtant difficile de démontrer une éventuelle particularité des profils (cf. Figure 82). Etant donné les différences dans les paramètres analytiques que partagent les méthodes MS et FID et dans le type de données brutes utilisées (aires de ions cibles vs. aires de pics), ne pas obtenir d'amélioration dans la mesure du coefficient de corrélation de Pearson après ajustement mathématique linéaire pour seulement 3 échantillons représente un résultat très satisfaisant, le nombre d'échantillons composant l'échantillonnage étant relativement faible en comparaison à la taille totale de la banque de données de référence.

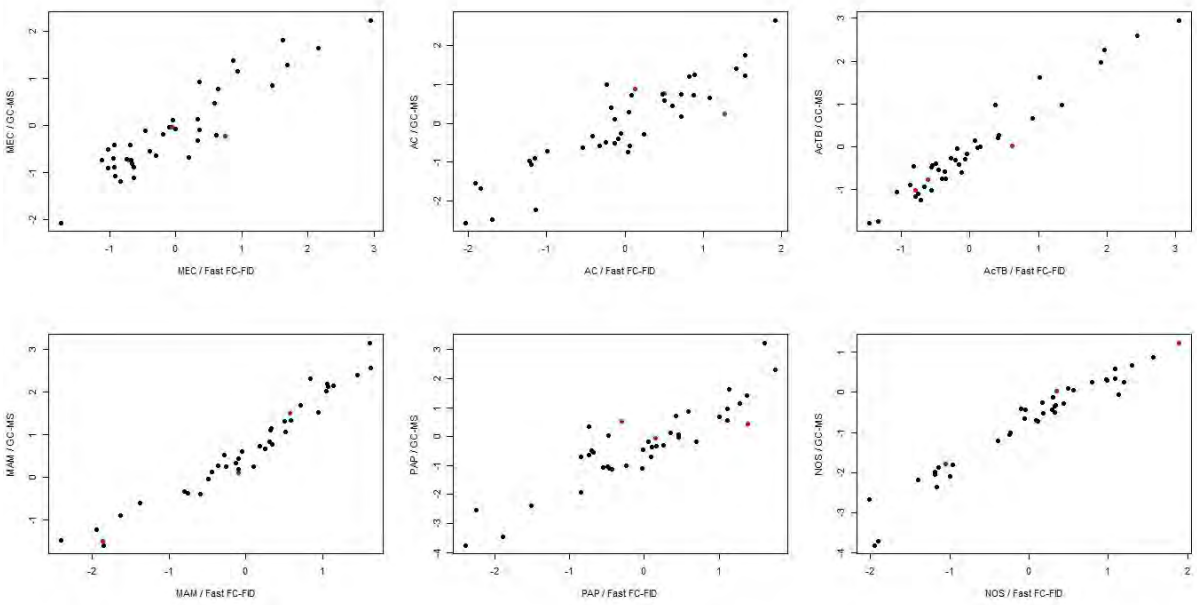


Figure 82. Représentation graphique de la distribution des profils chimiques, par variable, avec en rouge les trois spécimens pour lesquels l'ajustement mathématique conduit à une valeur du coefficient de corrélation de Pearson plus faible.

10.3 Conclusion

Ce chapitre démontre aussi bien la possibilité d'estimer que d'ajuster la similarité statistique de résultats obtenus avec des méthodes différentes et ce en considérant tous les niveaux de paramètres d'analyse. Il a été prouvé que l'ajustement mathématique améliore la similarité statistique existant entre les profils d'un même spécimen obtenus avec des méthodes différentes.

En comparaison aux profils non ajustés mathématiquement, l'utilisation de modèles mathématiques simples pour établir les règles d'ajustement fournit une méthodologie de profilage plus efficace (cf. §10.1). En effet, grâce à la production d'une meilleure séparation entre les populations d'échantillons liés (c'est-à-dire l'intra variabilité) et non liés (c'est-à-dire l'inter variabilité), l'efficacité de la méthodologie s'avère meilleure en termes de taux des vrais positifs et faux positifs obtenus en regard d'un seuil de décision fixé que les résultats du profilage soient utilisés en soutien à l'enquête ou qu'ils fassent office d'éléments de preuve dans une affaire particulière.

De plus, l'ACP-CAH démontre une plus grande performance dans la conservation de la structure des données une fois les profils mathématiquement ajustés (cf. §10.2.a). Dès que les profils sont ajustés mathématiquement, même à des hauteurs *h* faibles dans le dendrogramme local, la performance d'ajustement réussi est significativement plus élevée.

L'obtention de valeurs de coefficient de corrélation de Pearson plus élevées lors de la comparaison des profils GC-MS et « GC-MS like » que pour les profils Fast GC-FID des spécimens correspondants contribue à démontrer l'intérêt d'appliquer une telle approche d'optimisation de la similarité dans le cadre de l'approche d'harmonisation des résultats analytiques (cf. §10.2.b).

Chapitre 11 Discussion générale

11.1 Introduction

Le but principal de cette étude était de proposer et investiguer une approche dite d'harmonisation des résultats analytiques permettant l'estimation voire l'ajustement de la similarité statistique de profils chimiques de produits stupéfiants obtenus avec des méthodes analytiques différentes, quelles qu'elles soient. L'objectif consistant alors à rendre possible l'échange de résultats entre différents laboratoires appliquant des méthodologies analytiques distinctes et fournir ainsi potentiellement un outil plus efficace pour lutter contre le trafic de produits stupéfiants. En permettant le maintien d'une banque de données commune avec diverses méthodes analytiques, il s'agissait ainsi de proposer une alternative à la contraignante et coûteuse harmonisation des méthodes qui représente la méthodologie largement prônée actuellement. Ainsi, en cas de modification de la méthode analytique appliquée en routine par un laboratoire ou de comparaison de résultats analytiques entre différents laboratoires, la création d'une nouvelle banque de données ne serait pas nécessaire.

Avant de discuter les résultats, rappelons que quelle que soit la qualité des performances obtenues lors de l'ajustement de résultats provenant de méthodes différentes, il appartient aux laboratoires participants de décider s'ils peuvent s'en satisfaire selon leurs perspectives d'utilisation des profils chimiques. Les différents scénarios d'ajustement identifiés dans cette recherche ont été analysés afin d'évaluer les hypothèses définies dans le cadre de cette recherche et non dans l'optique de prouver que chaque scénario était faisable en pratique. L'étude des scénarios d'ajustement a permis de proposer une méthodologie globale à même d'estimer et d'optimiser la similarité de résultats quelles que soient les différences analytiques entre les méthodes utilisées pour, finalement, harmoniser les résultats, c'est-à-dire les profils chimiques dans cette recherche.

11.2 Approche d'harmonisation des résultats analytiques

11.2.a Le maintien d'une banque de données commune par différentes méthodes analytiques

Les résultats obtenus démontrent que si l'on compare des données issues de méthodes d'analyse répétables et reproductibles alors on peut douter de la nécessité de recourir impérativement à l'harmonisation des méthodes analytiques, du moins pour une certaine gamme de méthodes différentes (c'est-à-dire, selon la différence dans les niveaux de paramètres analytiques existant entre les méthodes considérées, cf. §11.3). Ainsi, le maintien d'une banque de données à l'aide de résultats issus de méthodes analytiques différentes est, dans certains cas, concevable et applicable.

La méthodologie d'ajustement des réponses analytiques appliquée dans cette étude implique l'établissement des relations mathématiques entre les réponses analytiques respectives de chacun des composés obtenues avec les différentes méthodes. La constitution d'un échantillonnage représentatif des spécimens répertoriés dans la banque de données représente ainsi une étape cruciale pour tous les laboratoires désirant comparer leurs résultats respectifs. Une estimation précise de la différence analytique entre les méthodes et de la similarité statistique existant entre les résultats en dépend.

L'ACP-CAH et l'étude de l'intra- et l'inter variabilité appliqués conjointement dans le cadre de la méthodologie développée se sont révélés être des outils statistiques judicieux pour estimer la similarité existant entre les profils, que ces derniers soient ajustés mathématiquement ou non. Sur la base des résultats obtenus avec ces deux outils, une réponse objective aux deux hypothèses principales a pu être donnée. Effectivement, le maintien d'une banque de données commune à diverses méthodes d'analyse dépend des caractéristiques analytiques de ces dernières (Hypothèse 1). L'impact de la différence analytique sur la similarité des profils chimiques obtenus par des méthodes différentes est démontré à l'aide des valeurs du coefficient de corrélation de Pearson calculées entre les profils des spécimens correspondants, des performances d'ajustement réussi obtenues à chaque valeur de *h* ainsi que de l'évolution des performances en fonction de *h*.

L'étude de l'intra- et l'inter variabilité inter méthodes illustre également clairement l'influence de la différence analytique. En effet, comme l'a présenté une étude précédente (Lociciro et al., 2008), le déplacement des mesures composant l'intra variabilité vers de plus faibles valeurs de similarité peut être constaté et son degré évalué. Ces deux outils statistiques s'avèrent nécessaires et complémentaires pour démontrer que la mise en place de méthodes analytiques différentes n'est pas un frein au maintien d'une banque de données commune (Hypothèse 2). Une fois les résultats provenant des diverses méthodes combinés, l'étude de l'intra- et l'inter variabilité permet d'apprécier l'efficacité de la méthodologie profilage (sous-hypothèse 2.1), tandis que l'ACP-CAH contribue à évaluer la conservation de la structure des données (sous-hypothèse 2.2).

11.2.b Signification d'un ajustement réussi

Comme discuté dans le chapitre précédent, la réussite de l'ajustement dans la banque de données de référence de deux profils chimiques d'un même spécimen mais déterminés par deux méthodes différentes doit signifier la présence de ces derniers dans la même classe chimique[48]. Il pourrait être rajouté que la réussite de l'ajustement doit *seulement* signifier la présence de ces derniers dans la même classe chimique. En effet, il n'est pas nécessaire que le profil « GC-MS like » soit transposé exactement au même endroit[49] dans l'espace de la banque de données de référence que le profil GC-MS du spécimen correspondant, en raison d'une caractéristique inhérente aux banques de données de profils. Ces dernières sont en constante évolution dans le sens où jour après jour et d'analyses en analyses, de nouveaux profils y sont ajoutés. La classification chimique des profils représente ainsi un processus dynamique se caractérisant par une modification constante de la distribution des classes. En conséquence, il est possible d'estimer que, dans une certaine mesure, la variabilité engendrée par la combinaison de résultats issus de méthodes analytiques différentes s'insère dans la variabilité engendrée par la constante évolution des classes chimiques dans la banque de données de référence.

[48] Comme cela a déjà été souligné, c'est la raison pour laquelle la seule détermination de la réussite de l'ajustement à l'aide d'une méthode de clustering n'est pas recommandée en raison de l'influence de la structure locale des données et des différences de distribution des données entre les différentes classes chimiques.

[49] En d'autres termes, une transposition au même endroit signifierait une présence dans le même cluster à une hauteur *h* relativement faible.

Démontrer l'appartenance des profils à la même classe chimique devrait alors représenter une preuve suffisante de la réussite de l'ajustement. Toutefois, les résultats obtenus ont démontré que l'utilisation seule de l'ACP-CAH pour apprécier la conservation de la classification chimique lors de la combinaison des résultats n'était pas suffisante. En particulier, déterminer l'appartenance de deux profils à la même classe chimique ne s'avère pas judicieux. Le recours à une autre mesure de similarité telle que la mesure du coefficient de corrélation de Pearson ne peut être que recommandé, que ce soit pour décrire la similarité statistique des méthodes sur la base des mesures estimées pour l'ensemble des profils chimiques ou pour estimer le degré de similarité entre les profils chimiques d'un spécimen en particulier.

Dans ce dernier cas de figure, il s'agirait dans un premier temps de mesurer dans le dendrogramme *local*[50] la distance séparant les deux profils considérés (où ces derniers sont définis par les scores respectifs sur les CPs sélectionnées), de représenter graphiquement la distribution des deux profils pour visualiser la distribution locale des données à l'aide de CP1 et CP2 par exemple, puis de mesurer le coefficient de corrélation de Pearson entre ces deux profils chimiques (définis par les aires de pic prétraitées des composés cibles). L'estimation du degré de similarité entre les profils d'après la mesure de la corrélation se ferait alors en regard de la distribution de l'intra- et l'inter variabilité inter méthodes, c'est-à-dire de la valeur du seuil de décision ainsi que des taux d'erreurs relatifs, selon la perspective d'utilisation des résultats du profilage (en tant que soutien à l'enquête policière ou en tant qu'élément de preuve dans une affaire particulière). Cette procédure pourrait bien sûr s'appliquer qu'il ait été décidé ou non de procéder à l'ajustement mathématique des données.

11.2.c Optimisation de la similarité des profils chimiques

L'approche d'optimisation de la similarité basée sur un ajustement mathématique des données a pu être testée pour chacun des scénarios d'ajustement et son utilité évaluée. Dans cette étude, un échantillonnage relativement limité (représentant un peu plus que les 10% des échantillons contenus dans la banque de données de référence) et conçu de telle sorte à être représentatif s'est avéré convenable pour aussi bien estimer qu'éventuellement ajuster la similarité des résultats analytiques issus des différentes méthodes.

[50] Dendrogramme local pouvant être constitué d'environ les 30 échantillons les plus proches d'après le dendrogramme global ou de l'ensemble des échantillons sélectionnés lorsque l'on « coupe » ce dernier à une hauteur de 20 par exemple (cf. §7.4).

A l'aide d'un tel échantillonnage et de modèles mathématiques relativement simples (linéaire, quadratique et cubique), il a été prouvé que des relations mathématiques robustes pouvaient être établies et qu'elles possédaient une qualité prédictive élevée pour l'ajustement de profils chimiques d'héroïne inconnus, selon les scénarios d'ajustement. L'ajustement mathématique permet de réduire efficacement la variabilité entre les résultats obtenus et les rendre plus similaires comme en témoignent une meilleure efficacité de la méthodologie de profilage (taux de vrais positifs et faux positifs obtenus en regard d'un seuil de décision), une meilleure conservation de la structure des données (plus grande performance d'ajustement réussi pour de faibles hauteurs *h*) et une obtention de valeurs de coefficient de corrélation de Pearson plus élevées lors de la comparaison respective des profils GC-MS et « GC-MS like » pour le spécimen correspondant. Ces résultats contribuent à démontrer l'intérêt d'appliquer une telle approche d'optimisation de la similarité dans le cadre de l'approche d'harmonisation des résultats analytiques. La combinaison de profils chimiques issus de méthodes différentes dans l'optique d'alimenter une seule et même banque de données n'en est que plus envisageable.

11.2.d Limites de l'étude

Finalement, bien que les résultats prouvent qu'appliquer la stratégie d'harmonisation des résultats analytiques peut réduire efficacement la variabilité entre les résultats obtenus avec différentes méthodes, il est important de souligner que la combinaison de résultats provenant de plus que deux méthodes n'est pas étudiée dans cette recherche.

Rappelons alors que la méthodologie d'ajustement des réponses analytiques implique l'établissement des relations statistiques entre toutes les méthodes analytiques, ou du moins entre la méthode de référence et les méthodes considérées s'il s'agit de maintenir la banque de données de référence avec ces dernières. L'ACP-CAH et l'étude de l'intra- et l'inter variabilité représentent des outils à même d'évaluer la faisabilité pratique de l'association de résultats quel que soit le nombre de méthodes analytiques combinées. Dans ce cas de figure et d'après ces deux outils statistiques, une augmentation de la variabilité peut être attendue, comme en ont discuté des études précédentes (Strömberg et al., 2000; Lociciro et al., 2007; Lociciro et al., 2008).

En effet, comme le démontre la présente étude, lorsque des résultats provenant de méthodes analytiques similaires sont combinés, la variabilité peut déjà être plus importante sur la base de l'étude de l'intra- et l'inter variabilité inter méthodes (scénario d'ajustement 1.2 correspondant à l'harmonisation des méthodes analytiques), d'où l'obtention probable d'une variabilité d'autant plus grande lorsque des résultats provenant de méthodes différentes seront combinés. Alors, l'application de modèles mathématiques plus complexes pourrait s'avérer par exemple nécessaire pour assurer une similarité statistique satisfaisante malgré la combinaison de plusieurs méthodes. Sans aucun doute, le degré de variabilité alors estimé dépendra avant tout de la différence analytique que partagent les méthodes d'analyses considérées.

Comme en discute le prochain paragraphe, cette étude démontre effectivement l'impact de la différence analytique des méthodes sur la similarité statistique des résultats qui en sont issus mais révèle également que le degré de différence statistique dépendra des paramètres analytiques modifiés. En d'autres termes, toute modification d'un paramètre analytique n'entraîne pas forcément d'effet significatif sur la similarité des profils chimiques.

11.3 Influence de la similarité analytique sur la similarité statistique

11.3.a Scénarios d'ajustement

Les résultats obtenus pour les données « GC-MS like » respectives (ajustement d'après le modèle linéaire), démontrent que la similarité analytique partagée par les méthodes considérées influence la similarité statistique calculée entre les profils chimiques correspondants, en termes de performance d'ajustement dans la banque de données de référence ou de coefficient de corrélation de Pearson mesuré. Ils illustrent et confirment en conséquence la classification réalisée dans cette recherche au Tableau 12.

En quelques mots, plus les caractéristiques analytiques des méthodes considérées seront différentes, plus la similarité statistique existant entre les profils chimiques de chacun des spécimens obtenus avec ces dernières – en termes de coefficient de corrélation de Pearson mesuré ou de hauteur *h* les séparant dans la banque de données de référence[51] – sera faible. Il en découlera alors un maintien d'une banque de données commune avec les méthodes considérées d'autant plus difficile à réaliser.

11.3.b Paramètres analytiques d'influence significative sur la similarité statistique

D'après la méthodologie développée, des différences dans les niveaux de paramètres analytiques C ne représentent pas un frein au maintien d'une banque de données avec les résultats obtenus, à en juger par la similarité statistique atteinte avec les résultats issus de la méthode analytique de référence. De plus, si la marque d'instrument analytique reste similaire, alors une différence de modèle (scénario d'ajustement 1.4) ou d'instrument analytique (scénario d'ajustement 1.2) n'impacte pas significativement la similarité statistique. En revanche, comme l'illustre la comparaison du scénario d'ajustement 1.5, une distinction dans la marque du fabricant (niveau de paramètre analytique B) influence de manière non négligeable la similarité statistique des résultats bien que les niveaux de paramètres analytiques A et C (les paramètres du « tune » MS exceptés) restent similaires. En effet, une telle différence engendre une similarité statistique similaire à celle estimée lorsque le paramètre analytique A_{DET} est différent (pour autant que A_{SEP} soit similaire, scénario 2.1). Finalement, le maintien d'une banque de données commune par des méthodes partageant des différences dans tous les niveaux de paramètres analytiques pourrait s'avérer difficile. Toutefois, bien que ceci soit particulièrement vrai pour le scénario d'ajustement 2.2, les résultats démontrent que si au minimum la technologie d'analyse de séparation est similaire (scénario d'ajustement 2.1) alors l'approvisionnement d'une banque de données avec les résultats obtenus demeure faisable.

[51] Pour autant que la structure locale des données ait peu d'influence sur la hauteur *h* séparant les deux profils.

La relativement faible similarité estimée pour les résultats du scénario 2.2 s'explique par le fait qu'une différence dans A_{SEP} engendre nécessairement une différence dans A_{DET}, en particulier dans le mode d'ionisation (alors qu'une différence dans A_{DET} n'implique pas une différence dans A_{SEP}). Si l'on prend l'exemple du scénario d'ajustement 2.2, tandis que l'ionisation en UHPLC-MS/MS se fait en ESI (Electro Spray Ionisation) à pression atmosphérique, en GC-MS elle se fait en EI (Electron Impact, cf. Chapitre 2) dans des conditions de vide. Une telle différence impacte directement la similarité statistique que partagent les résultats obtenus et limite forcément la possibilité de combiner dans une banque de données commune les profils chimiques provenant respectivement de ces deux méthodes.

La différence analytique initiale impactera également les relations mathématiques établies entre les réponses analytiques des composés respectifs. Bien que l'ajustement mathématique puisse améliorer significativement la similarité des profils chimiques, plus la similarité statistique sera faible au départ (en raison d'une différence analytique importante), plus des modèles mathématiques complexes devront alors être mis en place pour garantir l'ajustement des profils dans la banque de données de référence. Un parfait exemple en est l'incapacité des modèles linéaire, quadratique ou cubique d'ajuster puis de prédire les résultats obtenus en UHPLC-MS/MS alors que l'application de modèles plus complexes révèlent des performances satisfaisantes (Debrus et al., 2010). A l'inverse, un modèle linéaire est en mesure d'ajuster puis de prédire de manière satisfaisante les résultats obtenus avec la plupart des autres méthodes analytiques testées (en particulier les méthodes investiguées dans le cadre des scénarios 1.2 et 1.4). Ainsi, l'ajustement mathématique à l'aide du modèle linéaire représente un bon point de départ pour l'ajustement de résultats provenant de méthodes différentes mais partageant au moins une technologie d'analyse de séparation similaire.

Finalement, en raison des performances d'ajustement obtenues relativement faibles ainsi que de la trop grande complexité des modèles mathématiques devant être appliqués pour surmonter ce problème, la comparaison de méthodes différentes dans le paramètre analytique A_{SEP} n'est pas recommandée dans la mesure où une différence dans ce dernier entraîne également une différence dans A_{DET}. D'après les résultats expérimentaux, cette combinaison de différence analytique (A_{SEP} + A_{DET}) influence de manière significative la similarité statistique des profils chimiques. En conséquence, une telle combinaison de différences analytiques représente un frein pour l'approvisionnement d'une banque de données commune avec des résultats provenant de méthodes partageant une telle distinction.

Pour évaluer l'influence sur la similarité statistique des résultats d'une différence analytique dans A_{SEP}, il s'agirait de comparer les résultats analytiques obtenus avec des méthodes différentes selon A_{SEP} mais similaires selon A_{DET}, si cela était techniquement envisageable.

11.4 Perspectives

11.4.a Prétraitement statistique des données

Une étude approfondie des différents prétraitements et mesures de similarité qu'il serait possible d'appliquer sur les données pourrait conduire à de meilleures résultats de similarité. Par exemple, lorsque les résultats sont combinés, une meilleure séparation des populations d'échantillons liés et non liés ainsi qu'une absence de décalage (ou du moins, une diminution de son degré) de l'intra variabilité inter méthodes vers les plus faibles valeurs de similarité pourraient en théorie être atteintes. La méthodologie développée devrait permettre sur la base des résultats obtenus de définir le prétraitement le plus intéressant pour le maintien d'une banque de données commune par différentes méthodes.

11.4.b Utilisation de l'ACP-CAH

Les performances d'ajustement calculées pour les différents scénarios d'ajustement ont permis de décrire précisément la similarité analytique existant entre les méthodes investiguées. En effet, les performances obtenues à des hauteurs *h* faibles ainsi que l'évolution des performances en fonction de l'augmentation de *h* permettent d'estimer les similarités statistiques entre des profils provenant de méthodes différentes. En prenant du recul par rapport au but premier de l'ACP-CAH tel que recherché dans cette étude, cet outil peut être considéré comme étant à même de comparer la performance de différentes méthodes analytiques. En d'autres termes, le processus ACP-CAH tel qu'implémenté dans cette étude pourrait aider un laboratoire à décider de l'approche analytique à mettre en place en présentant clairement des résultats robustes, selon la problématique investiguée (qui, dans notre cas, consiste en l'étude de la présence de profils chimiques d'un même spécimen dans un même cluster). La méthode analytique permettant d'atteindre rapidement (c'est-à-dire, pour des hauteurs faibles) des performances élevées pourrait être en effet aisément identifiée.

Les performances pourraient alors être étudiées en regard des caractéristiques intrinsèques des méthodes analytiques telles que la facilité pratique de mise en œuvre ou le coût et ainsi participer au choix le plus convenable pour le laboratoire forensique. De plus, selon la problématique sous investigation, la performance estimée pourrait faire office de critère de choix d'une valeur de *h*. En effet, si pour un laboratoire l'objectif consiste à assurer une performance donnée, alors il lui serait aisé de déterminer la hauteur minimale devant être fixée pour l'atteindre.

Les résultats présentés dans cette recherche démontrent clairement l'influence de la distribution locale des données sur la présence ou non de deux profils chimiques dans le même cluster d'après une certaine hauteur *h* en raison de l'algorithme de groupement inhérent à la CAH. Cette observation implique qu'une estimation précise de la similarité chimique entre deux profils ne peut se baser uniquement sur une CAH. Une autre mesure de similarité telle qu'une mesure du coefficient de corrélation de Pearson est recommandée car révélant la relation directe existant entre les deux profils.

De la même manière que cela a été implémenté dans cette étude, la CAH pourrait être utilisée en routine pour trier les données présentes dans la banque de données. Dans le cadre d'une telle implémentation, une utilisation conjointe de l'ACP et de la CAH comme cela est le cas dans cette recherche s'avère pertinente. Cette phase de tri pourrait reposer sur une étape globale puis une étape locale d'ACP-CAH (en sélectionnant également un nombre adéquat de CPs) avant d'apprécier plus finement la similarité des profils présents dans le dendrogramme local. Ainsi, lorsqu'un nouveau profil serait inséré dans la banque de données, l'ACP-CAH pourrait permettre de sélectionner les profils les plus proches. Le sous-échantillonnage des profils les plus proches dans le dendrogramme local pourrait soit représenter le 10% de la totalité des échantillons présents dans la banque de données ou soit être constitué de tous les échantillons présents dans le même cluster selon la hauteur *h* prédéfinie à laquelle le dendrogramme global serait « coupé ». Il est important de mentionner que la manière de procéder pour trier les échantillons ainsi que la hauteur *h* utilisée auront une incidence directe sur la possibilité ou non de détecter les liens existant éventuellement d'après la mesure du coefficient de corrélation de Pearson (en d'autres termes, si le tri est trop restrictif alors le risque existerait de manquer des liens potentiels). Ensuite, la méthode de référence pour la mesure de la similarité entre deux profils, telle que la mesure du coefficient de corrélation de Pearson, serait appliquée.

La similarité des profils, et donc leur appartenance ou non à la même classe chimique, serait estimée sur la base de la valeur obtenue lors de la comparaison 2 à 2 des profils tout en se référant au seuil de décision défini lors de l'étude de l'intra- et de l'inter variabilité intra méthode.

Uniquement les profils les plus proches seraient ainsi comparés plus en détails, d'où un gain de temps important et une interprétation des résultats plus aisée que s'il fallait procéder à la comparaison de tous les profils enregistrés dans la banque de données. Cette approche pourrait parfaitement trouver sa place au sein du processus analytique appliqué par les laboratoires des domaines alimentaire (McIntyre et al., 2011) et pharmaceutique (Been et al., 2011; Sacré et al., 2011) lorsque ces derniers sont confrontés à des produits suspectés d'être contrefaits. Après une identification du produit, s'il est confirmé qu'il s'agit d'une contrefaçon, alors il convient de le caractériser en ayant recours à des outils chimiométriques (Dégardin et al., 2011; Deconinck et al., 2012).

Le but de ces derniers consiste en particulier à classifier dans la banque de données recensant les précédentes contrefaçons rencontrées le nouveau produit contrefait et ainsi à déterminer s'il appartient ou non à une classe physico-chimique déjà déterminée. L'ACP-CAH Globale et Locale permettrait alors d'estimer la similarité du nouvel échantillon testé avec ceux enregistrés dans la banque de données dans un premier temps puis de sélectionner les échantillons les plus proches pour affiner l'estimation de la similarité à l'aide des mesures de similarité obtenues dans un second temps (d'une part, grâce à la distance entre les profils dans le dendrogramme local et, d'autre part, à la mesure de la corrélation par exemple entre ces derniers). Ces domaines travaillant avec des banques de données « vivantes », en raison de l'ajout potentiellement régulier de nouveaux profils de contrefaçons, les classes déterminées sont en constante évolution d'où l'intérêt de recourir à des méthodes statistiques non supervisées, telles que l'ACP et la CAH. De plus, vu la manière dont elles sont combinées à une mesure de coefficient de corrélation de Pearson dans cette recherche, il en résulte une estimation précise de la similarité existant entre des échantillons enregistrés dans une banque de données et un nouvel échantillon à prédire.

Outre l'emploi d'outils chimiométriques, un ensemble de méthodes spectroscopiques ou chromatographiques peut être utilisé pour caractériser au maximum un produit contrefait dont la classe n'a jamais été rencontrée auparavant ou pour confirmer les relations estimées entre le nouvel échantillon et ceux enregistrés dans la banque de données. Or, l'approche ACP-CAH développée dans cette étude peut également permettre la combinaison de données issues de méthodes différentes et comparer les informations qu'elles fournissent, comme ceci est discuté dans les prochains paragraphes dans le contexte d'analyse des produits stupéfiants.

L'ACP-CAH s'insère parfaitement dans les nouvelles tendances d'analyse des produits stupéfiants par exemple, où une première méthode analytique relativement rapide et non destructive est implémentée pour opérer un premier tri au sein de la banque de données (et ainsi enlever du processus de comparaison future les échantillons ne présentant aucune similarité avec le nouvel échantillon testé) (Lopatka and Vallat, 2011; Camargo et al., 2012) tandis qu'une seconde méthode, plus précise dans l'estimation de la similarité existant entre chaque paire d'échantillons, est ensuite envisagée (Esseiva et al., 2011).

En effet, le but final de l'étape globale de l'ACP-CAH consiste justement à permettre la sélection du voisinage immédiat du nouvel échantillon à prédire en utilisant le dendrogramme produit. Alors, sur ce sous-groupe d'échantillons proches, une étape locale d'ACP-CAH est effectuée pour affiner l'évaluation de la similarité des profils et ainsi améliorer l'exactitude des prédictions. Dans notre recherche, bien que des méthodes différentes dans tous les niveaux de paramètres analytiques aient été utilisées, l'établissement du profil chimique des spécimens de produit stupéfiant reste similaire, c'est-à-dire la combinaison de la réponse analytique de chacun des composés majeurs jugés discriminants. C'est ensuite ce profil qui est comparé dans les étapes globale et locale de l'ACP-CAH. Mais, étant donné la manière dont le code informatique sous-jacent a été conçu, il est tout à fait envisageable de combiner plusieurs données utilisées ensuite pour les étapes globale et locale. En clair, l'étape globale pourrait reposer sur la comparaison de données issues d'une méthode rapide et non destructive et l'étape locale reposer quant à elle sur la comparaison de résultats provenant d'une méthode analytique plus efficace pour l'estimation de liens chimiques et permettant ainsi d'affiner l'évaluation de la similarité existant dans le sous-échantillonnage des profils les plus proches. Et par conséquent, différentes informations forensiques (profilage physico-chimique) pourraient être combinées.

Une telle approche implique alors pour les deux méthodes analytiques une certaine aptitude dans l'évaluation d'une similarité. En effet, les deux méthodes devraient fournir une efficacité acceptable (mais pas nécessairement similaire) pour différencier l'intra- de l'inter variabilité. Ou, à l'inverse, il ne serait pas envisageable que la méthode utilisée pour le tri initial ne soit pas apte à le faire et amène à la comparaison future des profils ne partageant aucune similarité entre eux. La première méthode analytique utilisée devrait être performante avant tout pour déterminer l'absence de toute similarité entre les profils tandis que la seconde devrait en revanche être performante pour assurer la présence d'une similarité entre les profils. Une étude précise pour chacune des deux méthodes de l'efficacité de la méthodologie de profilage devrait alors être envisagée en préalable à leur utilisation en séquence, par exemple par une étude de la variabilité des populations d'échantillons « liés » et « non liés » tel qu'appliqué dans cette recherche.

Quelle que soit la métrique employée pour cette estimation, elle devrait alors être utilisée pour évaluer la similarité de tous les profils présents dans le dendrogramme local, en raison de la démonstration d'influence de la structure locale des données sur les groupements opérés par la CAH (tel que fait dans cette recherche à l'aide d'une mesure du coefficient de corrélation de Pearson). La conclusion quant à la présence d'une relation chimique serait basée sur le seuil de décision et les taux d'erreurs relatifs déterminés lors de l'étude de l'intra- et l'inter variabilité pour des profils définis par les données obtenues avec la seconde méthode analytique.

En résumé, une telle approche serait intéressante à plusieurs niveaux. Premièrement, l'utilisation en séquence de profils établis à l'aide de différentes méthodologies analytiques dans l'optique de tirer le meilleur parti de toutes les données forensiques disponibles. Deuxièmement, l'utilisation à deux reprises de l'ACP-CAH pour enlever du processus de comparaison future les échantillons ne présentant aucune similarité avec ceux présents dans la banque de données puis ensuite assurer la présence ou non de relations entre les échantillons les plus proches du nouvel échantillon introduit. Du point de vue de la routine quotidienne d'un laboratoire forensique, ces deux premiers points s'avèrent particulièrement intéressants.

En effet, d'après les caractéristiques intrinsèques des méthodes qui devraient être implémentées pour atteindre les buts poursuivis dans le cadre d'une telle approche, là où la première méthode analytique serait peu coûteuse, rapide et facile à en mettre en œuvre pratiquement (grâce à une préparation des échantillons limitée par exemple), la seconde méthode demanderait un investissement pratique plus important et l'analyse serait plus longue. Par conséquent, d'un point de vue pratique et économique (ressources financières et humaines), l'intérêt d'analyser avec la seconde méthode uniquement les échantillons ayant été auparavant triés est clairement établi. Troisièmement, tel qu'investigué dans cette recherche, les mesures de similarité reposant d'une part sur une mesure de distance dans le dendrogramme local et d'autre part sur une mesure de corrélation entre les profils testés pourraient être combinées et comparées pour offrir un outil à même d'étudier précisément la similarité de profils en particulier ainsi que la structure des données dans l'environnement proche de ces derniers.

Il est important de noter que l'approche ACP-CAH peut s'avérer particulièrement intéressante pour comparer les données issues de méthodes différentes. En pratique, l'ACP-CAH pourrait être appliqué sur chacun des jeux de données séparément et il s'agirait alors de comparer les dendrogrammes locaux produits (présence ou non des mêmes profils, distance entre ces derniers, architecture du dendrogramme etc.). Visuellement, la contribution des données physico-chimiques voire informatiques (issues de la surveillance de sites Internet par exemple) pourrait être ainsi évaluée et par-dessus tout l'apport de ces données pour une meilleure compréhension des réseaux de distribution des produits illicites considérés estimé.

Dans le contexte de l'approvisionnement d'une banque de données commune et d'après la signification de la réussite de l'ajustement (cf. §11.2.b), la nécessité de disposer d'outils à même de représenter graphiquement l'architecture de la classification chimique ne peut être que souligné. Il pourrait être ainsi intéressant d'utiliser une caractéristique importante de la CAH. En effet, la CAH permet de visualiser graphiquement à l'aide du dendrogramme les similarités existant entre différents objets lorsque ceux-ci sont décrits par plusieurs variables, comme le sont les profils chimiques. Bien que la visualisation des similarités existantes atteigne ses limites pour de grands jeux de données, la CAH pourrait aider à visualiser l'architecture des banques de données de profils chimiques et en particulier la distribution des classes chimiques prédéterminées. Dans une telle optique, il s'agirait uniquement de décrire les similarités existant entre les profils et non d'estimer la présence ou non de liens chimiques.

L'investigation d'autres outils statistiques dédiés à la visualisation de grands jeux de données devrait être encouragée.

Finalement, pour s'assurer dans le temps de la qualité des analyses réalisées sur un même instrument analytique, il serait possible d'avoir recours à l'ACP-CAH. En effet, il s'agirait d'assurer que jour après jour les profils chimiques des échantillons de contrôle soient toujours proches les uns des autres dans l'espace défini. Une continuité dans la définition du profil chimique serait alors assurée, paramètre essentiel à l'utilisation sur le long terme de la méthodologie de profilage chimique.

11.4.c Méthodologie d'ajustement par calibration de chacun des composés

Grâce à l'ajustement des données à l'aide de modèles mathématiques relativement simples, la méthodologie se concentrant sur les réponses analytiques a démontré toute son efficacité pour optimiser la similarité des résultats dans la majorité des scénarios d'ajustement investigués et rendre ainsi plausible le maintien d'une banque de données commune avec différentes méthodes. La trop faible similarité estimée entre des résultats GC-MS et UHPLC-MS/MS malgré l'ajustement mathématique opéré rend inenvisageable un partage d'une même banque de données avec lesdites méthodes. Quelles solutions s'offrent alors à un laboratoire désirant absolument persister dans une telle entreprise ?

La première idée serait d'appliquer des modèles mathématiques plus complexes qui ont prouvé leur capacité à ajuster efficacement de telles données (Debrus et al., 2010). Toutefois, le risque de sur-apprentissage est important avec de tels modèles et leur emploi devrait idéalement être limité. La seconde solution consiste alors à investiguer la mise en œuvre de la méthodologie d'ajustement par calibration de chacun des composés du profil. Cette méthodologie pourrait permettre de réduire encore plus significativement la variabilité existant entre les résultats obtenus, tout en appliquant des modèles mathématiques relativement simples. Cependant, l'inconvénient majeur de cette démarche représente la nécessaire modification de la méthodologie de profilage chimique de référence : il s'agirait en effet de quantifier chacun des composés cibles et non plus seulement de les semi-quantifier.

Au-delà de la difficulté potentielle pour se procurer les standards nécessaires (couplée à leur coût élevé), une nouvelle banque de données devrait alors être construite, avec toutes les conséquences pratiques que cela suppose voire les pertes d'informations potentielles. A noter que la méthodologie d'ajustement par calibration de chacun des composés du profil pourrait s'avérer utile si la méthodologie d'ajustement des réponses analytiques ne permet pas d'atteindre une similarité satisfaisante lors du maintien d'une banque de données par plus de deux méthodes analytiques, cas de figure évoqué au §11.2.d.

11.4.d Maintien à long terme d'une banque de données commune par des méthodes analytiques différentes

La méthodologie présentée a permis de démontrer l'influence du degré de différence analytique sur les profils obtenus et d'envisager le maintien d'une banque de données commune à l'aide de résultats reflétant objectivement la similarité statistique existant entre les profils chimiques. Mais, d'une certaine manière, ces résultats reflètent la similarité analytique entre les méthodes à un moment donné (correspondant en particulier à la période d'analyse de l'échantillonnage).

Imaginons que sur la base de résultats considérés similaires deux laboratoires alimentent la même banque de données. Qu'en est-il de cette similarité après quelques semaines ou quelques mois d'analyses effectuées sur chaque instrument analytique respectif ? Assurer la répétabilité et la reproductibilité de chaque méthode respectivement assure-t-il la possibilité de maintenir sur le long terme la banque de données commune ? Quels critères appliquer pour considérer que la répétabilité et la reproductibilité des méthodes conviennent pour alimenter une même banque de données ? Est-il nécessaire de contrôler périodiquement la similarité statistique des profils analysés sur chaque méthode séparément ?

Bien que les profils de nouveaux échantillons puissent être ajustés dans la banque de données de référence sur la base des relations mathématiques prédéterminées, il semblerait judicieux d'effectivement investiguer qu'après un certain temps la similarité statistique est toujours satisfaisante.

En analysant un échantillonnage moins conséquent, les relations mathématiques déterminées à l'origine pourraient être examinées (équation mathématique établie ainsi que les coefficients de détermination ajusté et de prédiction) et ainsi estimer si elles sont toujours « à jour » ou si leur modification s'avère judicieuse. L'échantillonnage réduit à analyser pourrait consister dans les échantillons dits de référence, représentatifs de la composition et de la concentration des saisies de rue, et utilisés jour après jour pour contrôler le bon fonctionnement de la méthodologie analytique. Il suffirait alors que les laboratoires participants partagent les mêmes échantillons de référence. Un minimum de trois échantillons de référence de concentrations respectives faible, moyenne et élevée (en regard de l'intervalle de concentration des saisies habituellement estimées) devrait alors être envisagé. Avant de procéder de cette manière, il devrait être prouvé qu'un nombre réduit de spécimens de référence est à même de décrire efficacement les relations existant entre les résultats obtenus avec les différentes méthodes. De plus, pour diminuer le plus possible la variabilité due à l'inhomogénéité des échantillons, une homogénéisation soigneuse de chacun des spécimens devrait être entreprise, comme cela a été entrepris dans cette recherche.

Finalement, une manière de surmonter les diverses questions mentionnées ci-dessus consisterait à envisager le problème différemment. Plutôt que d'ajuster les résultats d'une méthode analytique différente pour les introduire dans la banque de données de référence, les profils contenus dans la banque de données de référence pourraient au contraire être ajustés de telle sorte à être similaires à ceux obtenus avec la méthode analytique différente. Bien que développée dans une optique opposée, la méthodologie appliquée dans cette recherche devrait être applicable, en particulier la nécessité de procéder à un échantillonnage représentatif et d'établir les relations mathématiques entre les résultats. Cette démarche serait particulièrement intéressante dans le cadre de la problématique intra laboratoire.

L'idée d'ajuster la banque de donnée de référence se justifie d'autant plus que la gestion des liens chimiques entre les échantillons devient difficile lors de l'augmentation croissante du nombre de profils enregistrés dans la banque, particulièrement en raison du recouvrement de classes chimiques pouvant rendre l'interprétation des résultats du profilage problématique. Ainsi, s'il est décidé d'ajuster les résultats de la banque de données de référence, uniquement les profils les plus récents pourraient être concernés, impliquant que l'ensemble de la mémoire conçue durant plusieurs années ne serait plus considéré. Mais se pose alors la question de la durée pendant laquelle les profils pourraient, ou devraient, être conservés.

Rappelons que l'objectif de construire une telle mémoire consiste à déterminer des liens entre des saisies policières effectuées à différentes périodes dans le temps et en divers lieux (et donc, n'ayant aucun lien entre elles selon les investigations des forces de police) pour décrire les réseaux de distribution des produits stupéfiants une fois les résultats du profilage combinés aux informations d'enquête traditionnelle. Alors, dans ce cadre là, enlever les échantillons enregistrés dans la banque de données pourrait provoquer une perte potentielle d'informations. A la connaissance de l'auteur, peu d'auteurs discutent de cette problématique. Strömberg et al. (2000) représente une rare publication à l'aborder et les auteurs y proposent de ne conserver les profils que 6 mois dans la banque de données, essentiellement pour des considérations de gestion de celle-ci, sans prendre en compte l'utilité finale des profils chimiques et les renseignements qu'ils peuvent apporter.

Au contraire, la durée de conservation des profils chimiques dans la banque de données devrait plutôt reposer sur une étude de la durée de vie des classes chimiques que sur une « simple » question de facilité dans la gestion des données. Une recherche précédente a démontré que des saisies ayant un profil chimique estimé similaire ont été observées sur le marché pendant presque 3 ans (Esseiva, 2004). Il est clair qu'une telle information, importante pour une meilleure compréhension du dynamisme d'écoulement de produits stupéfiants, ne peut être obtenue que grâce à la conservation des profils pendant une durée significative. La question de conservation des données dans la banque étant délicate, une discussion devrait avoir lieu entre les différentes parties impliquées dans le cycle du renseignement du profilage, particulièrement le laboratoire forensique, les analystes criminels et les forces de police, pour juger de la perte potentielle d'informations si les profils les plus « vieux » n'étaient pas conservés et décider conjointement de la durée de conservation. En particulier, les forces de police étant les bénéficiaires principaux de ces informations, une durée de conservation appropriée pourrait être définie selon l'usage qu'ils font des résultats du profilage. Par exemple, les forces de police pourraient n'être intéressées que dans les liens chimiques contemporains avec la saisie venant d'être réalisée. Alors, dans un tel cas de figure, la question de l'opportunité de conserver les profils chimiques durant une période de temps de l'ordre de plusieurs années pourrait effectivement être soulevée.

Conclusion

La criminalité organisée concerne une grande variété d'activités criminelles, parmi lesquelles la principale se concentre sur la fabrication et le trafic de produits stupéfiants. En conséquence, pour influencer l'activité criminelle, une meilleure compréhension des voies de trafic et des réseaux de distribution s'avère essentielle. Nécessairement, cela implique un échange de données sur la base des résultats du profilage entre les laboratoires participant à la lutte nationale ou internationale contre le trafic de produits stupéfiants.

Dans ce contexte, le défi de cette recherche consistait alors à démontrer la faisabilité pratique de l'approvisionnement avec différentes méthodes analytiques ou du partage entre plusieurs laboratoires d'une banque de données commune. En effet, il a été décidé de s'écarter complètement de toute harmonisation des méthodes analytiques, qui représente la stratégie largement prônée de nos jours mais remise en cause par les laboratoires en raison de difficultés pratiques lors de sa mise en place (car contraignante, coûteuse et soulevant des problématiques de maintien sur le long terme et de partage de la banque de données qui impliquent une inertie analytique voire une perte d'informations pour le laboratoire).

Par conséquent, une stratégie avantageuse décrite comme étant l'harmonisation des résultats analytiques a été investiguée pour surmonter ces difficultés. Sachant que la comparaison de résultats obtenus avec des méthodes différentes n'est pas directement possible, une méthodologie à même d'estimer voire d'optimiser la similarité des résultats a été développée. Dans ce cadre là, les notions essentielles de *différence analytique* et de *scénario d'ajustement* jamais abordées auparavant dans la littérature ont été définies.

Une fois combinées aux résultats obtenus dans cette recherche, ces notions, valables pour tout domaine analytique, ont permis d'identifier les paramètres analytiques influençant significativement la similarité statistique des résultats et donnent à présent le moyen de prédire la similarité à attendre pour des méthodes considérées. De plus, l'obtention d'une méthodologie de profilage efficace ainsi que la conservation de la structure des données ont été identifiées comme étant les conditions sine qua non à l'approvisionnement d'une banque de données commune par des résultats issus de diverses méthodes.

La particularité de la méthodologie d'estimation et d'optimisation de la similarité développée repose sur l'implémentation d'outils statistiques permettant le maintien d'une même banque de données malgré l'établissement de profils chimiques par semi-quantification et issus de diverses méthodes analytiques.

À l'aide d'un échantillonnage relativement limité mais représentatif, la détermination des relations mathématiques existant entre les résultats offre la possibilité d'ajuster mathématiquement les profils obtenus avec différentes méthodes pour les rendre similaires à ceux enregistrés dans la banque de données de référence. L'ajustement mathématique des données constitue sans aucun doute une étape fondamentale dans une telle stratégie et l'objectif final de combinaison de ces profils dans une seule et même banque de données n'en devient que plus envisageable.

L'étude de l'intra- et l'inter variabilité et l'ACP-CAH Globale et Locale se sont révélés être des outils statistiques judicieux et complémentaires pour estimer la similarité de résultats issus de méthodes plus ou moins différentes et ainsi évaluer la faisabilité pratique de l'association de ces derniers au sein d'une banque de données commune.

Grâce à ces derniers, il a été démontré que la création d'une banque de données commune à plusieurs méthodes analytiques ou à plusieurs laboratoires ne passait pas nécessairement par une phase contraignante d'harmonisation des méthodes d'analyse. En effet, le maintien d'une banque de données commune à diverses méthodes répétables et reproductibles dépend avant tout des caractéristiques analytiques de ces dernières. D'après les résultats obtenus, plus les caractéristiques analytiques des méthodes considérées seront différentes, plus la similarité statistique existant entre les profils chimiques sera faible. Ainsi, le maintien d'une banque de données à l'aide de profils issus de méthodes analytiques différentes est, dans certains cas (c'est-à-dire, selon la différence dans les niveaux de paramètres analytiques existant entre les méthodes considérées), concevable et applicable. Les résultats expérimentaux ont démontré que seul le paramètre analytique A_{SEP} influençait de manière significative la similarité statistique des profils chimiques et qu'une différence dans ce paramètre partagée par les méthodes considérées représentait un frein au maintien d'une banque de données commune.

Par conséquent, à l'aide d'une harmonisation des résultats analytiques, il serait aussi bien possible pour un laboratoire d'utiliser plusieurs instruments analytiques, de rendre sa méthode plus performante, d'implémenter sa méthode vers un modèle ou une marque d'équipement différente voire de changer de technologies d'analyse de détection sans qu'il n'y ait d'influence à la fois sur le maintien de la banque de données et sur l'étape suivante de comparaisons des résultats analytiques.

D'un point de vue purement statistique, il a été clairement démontré que la similarité entre deux profils chimiques ne pouvait être uniquement évaluée par une CAH en raison de l'influence certaine de la structure des données sur le dendrogramme final. Cette démonstration peut avoir son importance tant la seule utilisation de la CAH est récurrente dans la communauté scientifique pour classifier des données. D'après les résultats obtenus, l'association d'une mesure de similarité basée sur une CAH à une mesure de corrélation par exemple entre les profils est recommandée. L'intégration de telles mesures à la méthodologie développée dans cette recherche offre une combinaison de résultats statistiques révélant précisément la relation existant entre des profils chimiques.

Finalement, cette démarche, appliquée avec succès dans le cadre du profilage chimique de produits stupéfiants, a été développée de telle sorte à être applicable à tout domaine analytique confronté à une problématique d'extension spatio-temporelle d'une banque de données de résultats analytiques. En particulier, la méthodologie d'estimation et d'optimisation de la similarité pourrait non seulement être appliquée à d'autres produits stupéfiants que celui d'étude mais également à d'autres domaines analytiques. Dans ce cadre là, l'ACP-CAH Globale et Locale possède un large champ d'applications envisageables dans le cadre du profilage de produits stupéfiants ou lors d'analyses de produits alimentaires et médicamenteux contrefaits que ce soit pour trier les échantillons de la banque données et affiner par la suite l'évaluation de la similarité ou pour permettre la comparaison de différentes données forensiques (profilage physico-chimique par exemple).

Dans le contexte de la lutte internationale contre le trafic de produits stupéfiants, cette recherche ouvre la perspective à un échange de résultats entre des laboratoires partageant une méthodologie analytique différente pouvant améliorer la lutte contre le trafic de ces substances grâce à l'apport potentiel en renseignements qui en découlerait.

Bibliographie

Aalberg, L., K. Andersson, C. Bertler, H. Borén, M. D. Cole, J. Dahlén, Y. Finnon, H. Huizer, K. Jalava, E. Kaa, E. Lock, A. Lopes, A. Poortman-Van Der Meer and E. Sippola (2005). "Development of a harmonised method for the profiling of amphetamines: I. Synthesis of standards and compilation of analytical data." Forensic Science International **149**(2-3): 219-229.

Aalberg, L., K. Andersson, C. Bertler, M. D. Cole, Y. Finnon, H. Huizer, K. Jalava, E. Kaa, E. Lock, A. Lopes, A. Poortman-Van Der Meer, E. Sippola and J. Dahlén (2005). "Development of a harmonised method for the profiling of amphetamines: II. Stability of impurities in organic solvents." Forensic Science International **149**(2-3): 231-241.

Adams, M. J. (2006). Pattern recognition I: Unsupervised analysis. Chemometrics in Analytical Spectroscopy, The Royal Society of Chemistry.

Aitken, C. G. G. (1995). Statistics and the Evaluation of Evidence for Forensic Scientists. Chichester, Wiley.

Alm, S. (2006). Forensic Intelligence/Impurity profiling; an historical review, Collaborative Harmonised Amphetamine INitiative (CHAIN), AGIS 2006.

Amirav, A. and H. Jing (1995). "Pulsed flame photometer detector for gas chromatography." Analytical Chemistry **67**(18): 3305-3318.

Andersson, K., K. Jalava, E. Lock, Y. Finnon, H. Huizer, E. Kaa, A. Lopes, A. Poortman-van der Meer, M. D. Cole, J. Dahlén and E. Sippola (2007). "Development of a harmonised method for the profiling of amphetamines. III. Development of the gas chromatographic method." Forensic Science International **169**(1): 50-63.

Andersson, K., K. Jalava, E. Lock, H. Huizer, E. Kaa, A. Lopes, A. Poortman-van der Meer, M. D. Cole, J. Dahlén and E. Sippola (2007). "Development of a harmonised method for the profiling of amphetamines. IV. Optimisation of sample preparation." Forensic Science International **169**(1): 64-76.

Andersson, K., E. Lock, K. Jalava, H. Huizer, S. Jonson, E. Kaa, A. Lopes, A. Poortman-van der Meer, E. Sippola, L. Dujourdy and J. Dahlén (2007). "Development of a harmonised method for the profiling of amphetamines VI. Evaluation of methods for comparison of amphetamine." Forensic Science International **169**(1): 86-99.

Ballany, J., B. Caddy, M. Cole, Y. Finnon, L. Aalberg, K. Janhunen, E. Sippola, K. Andersson, C. Bertler, J. Dahlén, I. Kopp, L. Dujourdy, E. Lock, P. Margot, H. Huizer, A. Poortman, E. Kaa and A. Lopes (2001). "Development of a harmonised pan-European method for the profiling of amphetamines." Science and Justice - Journal of the Forensic Science Society **41**(3): 193-196.

Béen, F., Y. Roggo, K. Degardin, P. Esseiva and P. Margot (2011). "Profiling of counterfeit medicines by vibrational spectroscopy." Forensic Science International **211**(1-3): 83-100.

Besacier, F., H. Chaudron-Thozet, M. Rousseau-Tsangaris, J. Girard and A. Lamotte (1997). "Comparative chemical analyses of drug samples: General approach and application to heroin." Forensic Science International **85**(2): 113-125.

Bicchi, C., C. Brunelli, C. Cordero, P. Rubiolo, M. Galli and A. Sironi (2004). "Direct resistively heated column gas chromatography (Ultrafast module-GC) for high-speed analysis of essential oils of differing complexities." Journal of Chromatography A **1024**(1-2): 195-207.

Blumberg, L. M., T. A. Berger and M. Klee (1995). "Constant flow versus constant pressure in a temperature-programmed gas chromatograph." Journal of High Resolution Chromatography **18**(6): 378-380.

Blumberg, L. M. and M. S. Klee (1998). "Method Translation and Retention Time Locking in Partition GC." Analytical Chemistry **70**(18): 3828-3839.

Boccard, J., E. Grata, A. Thiocone, J. Y. Gauvrit, P. Lantéri, P. A. Carrupt, J. L. Wolfender and S. Rudaz (2007). "Multivariate data analysis of rapid LC-TOF/MS experiments from Arabidopsis thaliana stressed by wounding." Chemometrics and Intelligent Laboratory Systems **86**(2 SPEC. ISS.): 189-197.

Bolck, A., C. Weyermann, L. Dujourdy, P. Esseiva and J. van den Berg (2009). "Different likelihood ratio approaches to evaluate the strength of evidence of MDMA tablet comparisons." Forensic Science International **191**(1-3): 42-51.

Booth, A. L., M. J. Wooller, T. Howe and N. Haubenstock (2010). "Tracing geographic and temporal trafficking patterns for marijuana in Alaska using stable isotopes (C, N, O and H)." Forensic Science International **202**(1-3): 45-53.

Brereton, R. G. (2007). Applied Chemometrics for Scientists, John Wiley & Sons, Ltd.

Brereton, R. G. (2009). Chemometrics for Pattern Recognition, John Wiley & Sons, Ltd.

Broséus, J., B. Debrus, O. Delémont, S. Rudaz and P. Esseiva (2013). "Study of common database feeding with results coming from different analytical methods in the framework of illicit drugs chemical profiling." Forensic Science International **In press**.

Broséus, J., M. Vallat and P. Esseiva (2011). "Multi-class differentiation of cannabis seedlings in a forensic context." Chemometrics and Intelligent Laboratory Systems **107**(2): 343-350.

Camargo, J., P. Esseiva, F. Gonzalez, J. Wist and L. Patiny (2012). "Monitoring of illicit pill distribution networks using an image collection exploration framework." Forensic Science International **223**(1-3): 298-305.

Carter, J. F., R. Sleeman, J. C. Hill, F. Idoine and E. L. Titterton (2005). "Isotope ratio mass spectrometry as a tool for forensic investigation (examples from recent studies)." Science and Justice - Journal of the Forensic Science Society **45**(3): 141-149.

Cartier, J., O. Gueniat and M. D. Cole (1997). "Headspace analysis of solvents in cocaine and heroin samples." Science & Justice **37**(3): 175-181.

Casale, J. F., J. R. Ehleringer, D. R. Morello and M. J. Lott (2005). "Isotopic fractionation of carbon and nitrogen during the illicit processing of cocaine and heroin in South America." Journal of Forensic Sciences **50**(6): 1315-1321.

Chan, K. W., G. H. Tan and R. C. S. Wong (2012). "Gas chromatographic method validation for the analysis of major components in illicit heroin seized in Malaysia." Science and Justice **52**(1): 9-16.

Chan, K. W., G. H. Tan and R. C. S. Wong (2013). "Investigation of trace inorganic elements in street doses of heroin." Science and Justice **53**(1): 73-80.

Collins, M., J. Huttunen, I. Evans and J. Robertson (2007). "Illicit drug profiling: the Australian experience." Australian Journal of Forensic Sciences **39**(1): 25 - 32.

Comparin, J. (2007). Drugs and Toxicology. Interpol Forensic Science Symposium Reports 2007, Interpol.

Cramers, C. A., H.-G. Janssen, M. M. van Deursen and P. A. Leclercq (1999). "High-speed gas chromatography: an overview of various concepts." Journal of Chromatography A **856**(1-2): 315-329.

Dagan, S. and A. Amirav (1996). "Fast, very fast, and ultra-fast gas chromatography-mass spectrometry of thermally labile steroids, carbamates, and drugs in supersonic molecular beams." Journal of the American Society for Mass Spectrometry **7**(8): 737-752.

Dallüge, J., R. J. J. Vreuls, D. J. Van Iperen, M. Van Rijn and U. A. T. Brinkman (2002). "Resistively heated gas chromatography coupled to quadrupole mass spectrometry." Journal of Separation Science **25**(9): 608-614.

Dams, R., T. Benijts, W. E. Lambert, D. L. Massart and A. P. De Leenheer (2001). "Heroin impurity profiling: trends throughout a decade of experimenting." Forensic Science International **123**(2-3): 81-88.

Debrus, B., J. Broséus, D. Guillarme, P. Lebrun, P. Hubert, J. L. Veuthey, P. Esseiva and S. Rudaz (2010). "Innovative methodology to transfer conventional GC-MS heroin profiling to UHPLC-MS/MS." Analytical and Bioanalytical Chemistry: 1-12.

Deconinck, E., P. Y. Sacré, D. Coomans and J. De Beer (2012). "Classification trees based on infrared spectroscopic data to discriminate between genuine and counterfeit medicines." Journal of Pharmaceutical and Biomedical Analysis **57**(1): 68-75.

Dégardin, K., Y. Roggo, F. Been and P. Margot (2011). "Detection and chemical profiling of medicine counterfeits by Raman spectroscopy and chemometrics." Analytica Chimica Acta **705**(1-2): 334-341.

Deursen, M. v., J. Beens, C. A. Cramers and H.-G. Janssen (1999). "Possibilities and Limitations of Fast Temperature Programming as a Route towards Fast GC." Journal of High Resolution Chromatography **22**(9): 509-513.

Dixon, S. J. and R. G. Brereton (2009). "Comparison of performance of five common classifiers represented as boundary methods: Euclidean Distance to Centroids, Linear Discriminant Analysis, Quadratic Discriminant Analysis, Learning Vector Quantization and Support Vector Machines, as dependent on data structure." Chemometrics and Intelligent Laboratory Systems **95**(1): 1-17.

Dömötörová, M., M. Kirchner, E. Matisová and J. d. Zeeuw (2006). "Possibilities and limitations of fast GC with narrow-bore columns." Journal of Separation Science **29**(8): 1051-1063.

Duda, R. O., P. E. Hart and D. G. Stork (2001). Pattern classification. New York, Wiley.

Dufey, V., L. Dujourdy, F. Besacier and H. Chaudron (2007). "A quick and automated method for profiling heroin samples for tactical intelligence purposes." Forensic Science International **169**(2-3): 108-117.

Dujourdy, L., G. Barbati, F. Taroni, O. Guéniat, P. Esseiva, F. Anglada and P. Margot (2003). "Evaluation of links in heroin seizures." Forensic Science International **131**(2-3): 171-183.

Dujourdy, L. and F. Besacier (2008). "Headspace profiling of cocaine samples for intelligence purposes." Forensic Science International **179**(2-3): 111-122.

Dujourdy, L., V. Dufey, F. Besacier, N. Miano, R. Marquis, E. Lock, L. Aalberg, S. Dieckmann, F. Zrcek and J. S. Bozenko Jr (2008). "Drug intelligence based on organic impurities in illicit MA samples." Forensic Science International **177**(2-3): 153-161.

Dyson, N. (1999). "Peak distortion, data sampling errors and the integrator in the measurement of very narrow chromatographic peaks." Journal of Chromatography A **842**(1-2): 321-340.

Edelman, G., M. Lopatka and M. Aalders (2013). "Objective color classification of ecstasy tablets by hyperspectral imaging." Journal of Forensic Sciences.

Ehleringer, J. R., D. A. Cooper, M. J. Lott and C. S. Cook (1999). "Geo-location of heroin and cocaine by stable isotope ratios." Forensic Science International **106**(1): 27-35.

Esseiva, P. (2004). Le profilage de l'héroïne et de la cocaïne : Mise en place d'une systématique permettant une utilisation opérationnelle des liens chimiques. Faculté de Droit, Ecole des Sciences Criminelles, Institut de Police Scientifique. Lausanne, Université de Lausanne.

Esseiva, P., F. Anglada, L. Dujourdy, F. Taroni, P. Margot, E. Du Pasquier, M. Dawson, C. Roux and P. Doble (2005). "Chemical profiling and classification of illicit heroin by principal component analysis, calculation of inter sample correlation and artificial neural networks." Talanta **67**(2): 360-367.

Esseiva, P., L. Dujourdy, F. Anglada, F. Taroni and P. Margot (2003). "A methodology for illicit heroin seizures comparison in a drug intelligence perspective using large databases." Forensic Science International **132**(2): 139-152.

Esseiva, P., L. Gaste, D. Alvarez and F. Anglada (2011). "Illicit drug profiling, reflection on statistical comparisons." Forensic Science International **207**(1-3): 27-34.

Esseiva, P., S. Ioset, F. Anglada, L. Gaste, O. Ribaux, P. Margot, A. Gallusser, A. Biedermann, Y. Specht and E. Ottinger (2007). "Forensic drug Intelligence: An important tool in law enforcement." Forensic Science International **167**(2-3): 247-254.

Esseiva, P. and P. Margot (2009). Drug Profiling. Wiley Encyclopedia of Forensic Science, John Wiley & Sons, Ltd.

Galimov, E. M., V. S. Sevastyanov, E. V. Kulbachevskaya and A. A. Golyavin (2005). "Isotope ratio mass spectrometry: d13C and d15 N analysis for tracing the origin of illicit drugs." Rapid Communications in Mass Spectrometry **19**(10): 1213-1216.

Gallagher, R., R. Shimmon and A. M. McDonagh (2012). "Synthesis and impurity profiling of MDMA prepared from commonly available starting materials." Forensic Science International **223**(1-3): 306-313.

Giannasi, P., D. Pazos, P. Esseiva and Q. Rossy (2012). "Detection and analysis of websites selling GBL on the Internet: Prospects for criminal intelligence." Revue Internationale de Criminologie et de Police Technique et Scientifique (RICTPS) **65**(4): 468-479.

Grob, R. L. and E. F. Barry (2004). Modern Practice of Gas Chromatography, Wiley Interscience.

Gross, J. H. (2011). Mass Spectrometry : A Textbook, Springer.

Gueniat, O. and P. Esseiva, Eds. (2005). Le profilage de l'héroïne et de la cocaïne - Une méthodologie moderne de lutte contre le trafic illicite. Collections Sciences Forensiques. Lausanne, Presses polytechniques et universitaires romandes.

Herrmann, A., Ed. (2010). The Chemistry and Biology of Volatiles, John Wiley & Sons, Ltd.

Hibbert, D. B., D. Blackmore, J. Li, D. Ebrahimi, M. Collins, S. Vujic and P. Gavoyannis (2010). "A probabilistic approach to heroin signatures." Analytical and Bioanalytical Chemistry **396**(2): 765-773.

Hurley, J. M., J. B. West and J. R. Ehleringer (2010). "Stable isotope models to predict geographic origin and cultivation conditions of marijuana." Science and Justice **50**(2): 86-93.

Kirchner, M., E. Matisová, S. Hrouzková and J. d. Zeeuw (2005). "Possibilities and limitations of quadrupole mass spectrometric detector in fast gas chromatography." Journal of Chromatography A **1090**(1-2): 126-132.

Klee, M. and L. Blumberg (2002). Theoretical and practical aspects of fast gas chromatography and method translation : Fast gas chromatography. Niles, IL, ETATS-UNIS, Preston Publications.

Korytár, P., H.-G. Janssen, E. Matisová and U. A. T. Brinkman (2002). "Practical fast gas chromatography: methods, instrumentation and applications." TrAC Trends in Analytical Chemistry **21**(9-10): 558-572.

Lindley, D. V. (1977). "A problem in forensic science." Biometrika **64**(2): 207-213.

Lociciro, S., P. Esseiva, P. Hayoz, L. Dujourdy, F. Besacier and P. Margot (2008). "Cocaine profiling for strategic intelligence, a cross-border project between France and Switzerland: Part II. Validation of the statistical methodology for the profiling of cocaine." Forensic Science International **177**(2-3): 199-206.

Lociciro, S., P. Hayoz, P. Esseiva, L. Dujourdy, F. Besacier and P. Margot (2007). "Cocaine profiling for strategic intelligence purposes, a cross-border project between France and Switzerland: Part I. Optimisation and harmonisation of the profiling method." Forensic Science International **167**(2-3): 220-228.

Lock, E., L. Aalberg, K. Andersson, J. Dahlén, M. D. Cole, Y. Finnon, H. Huizer, K. Jalava, E. Kaa, A. Lopes, A. Poortman-van der Meer and E. Sippola (2007). "Development of a harmonised method for the profiling of amphetamines V. Determination of the variability of the optimised method." Forensic Science International **169**(1): 77-85.

Lopatka, M. and M. Vallat (2011). "Surface granularity as a discriminating feature of illicit tablets." Forensic Science International **210**(1-3): 188-194.

Lurie, I. S., P. A. Hays, A. E. Garcia and S. Panicker (2004). "Use of dynamically coated capillaries for the determination of heroin, basic impurities and adulterants with capillary electrophoresis." Journal of Chromatography A **1034**(1-2): 227-235.

Lurie, I. S. and S. G. Toske (2008). "Applicability of ultra-performance liquid chromatography-tandem mass spectrometry for heroin profiling." Journal of Chromatography A **1188**(2): 322-326.

Marclay, F., D. Pazos, O. Delémont, P. Esseiva and C. Saudan (2010). "Potential of IRMS technology for tracing gamma-butyrolactone (GBL)." Forensic Science International **198**(1-3): 46-52.

Marquis, R., C. Weyermann, C. Delaporte, P. Esseiva, L. Aalberg, F. Besacier, J. S. Bozenko Jr, R. Dahlenburg, C. Kopper and F. Zrcek (2008). "Drug intelligence based on MDMA tablets data. 2. Physical characteristics profiling." Forensic Science International **178**(1): 34-39.

Marsili, R., Ed. (2002). Flavor, Fragrance, and Odor Analysis, Second Edition, CRC Press.

Martino, R., M. Malet-Martino, V. Gilard and S. Balayssac (2010). "Counterfeit drugs: Analytical techniques for their identification." Analytical and Bioanalytical Chemistry **398**(1): 77-92.

Massart, D. L., B. G. M. Vandeginste, L. M. C. Buydens, S. De Jong, P. J. Lewi and J. Smeyers-Verbeke (1997). Handbook of Chemometrics and Qualimetrics : Part A. Amsterdam, Elsevier.

Mastovská, K. and S. J. Lehotay (2003). "Practical approaches to fast gas chromatography-mass spectrometry." Journal of Chromatography A **1000**(1-2): 153-180.

Matisová, E. and M. Dömötörová (2003). "Fast gas chromatography and its use in trace analysis." Journal of Chromatography A **1000**(1-2): 199-221.

McGorrin, R. J. (2009). "One Hundred Years of Progress in Food Analysis." Journal of Agricultural and Food Chemistry **57**(18): 8076-8088.

McIntyre, A. C., M. L. Bilyk, A. Nordon, G. Colquhoun and D. Littlejohn (2011). "Detection of counterfeit Scotch whisky samples using mid-infrared spectrometry with an attenuated total reflectance probe incorporating polycrystalline silver halide fibres." Analytica Chimica Acta **690**(2): 228-233.

Milliet, Q., C. Weyermann and P. Esseiva (2009). "The profiling of MDMA tablets: A study of the combination of physical characteristics and organic impurities as sources of information." Forensic Science International **187**(1-3): 58-65.

Mitrevski, B., B. Veleska, E. Engel, P. Wynne, S. M. Song and P. J. Marriott (2011). "Chemical signature of ecstasy volatiles by comprehensive two-dimensional gas chromatography." Forensic Science International **209**(1-3): 11-20.

Moore, J. M., A. C. Allen and D. A. Cooper (1984). "Determination of manufacturing impurities in heroin by capillary gas chromatography with electron capture detection after derivatization with heptafluorobutyric anhydride." Analytical Chemistry **56**(4): 642-646.

Morello, D. R., S. D. Cooper, S. Panicker and J. F. Casale (2010). "Signature profiling and classification of illicit heroin by GC-MS analysis of acidic and neutral manufacturing impurities." Journal of Forensic Sciences **55**(1): 42-49.

Neumann, H. (1984). "Analysis of opium and crude morphine samples by capillary gas chromatography : Comparison of impurity profiles." Journal of Chromatography A **315**: 404-411.

Neumann, H. (1994). "Comparison of heroin by capillary gas chromatography in Germany." Forensic Science International **69**(1): 7-16.

Nic Daéid, N. and R. J. H. Waddell (2005). "The analytical and chemometric procedures used to profile illicit drug seizures." Talanta **67**(2): 280-285.

Nier, C., N. Gentile and P. Esseiva (2012). "Etude de l'impact du vieillissement et de la pureté sur le profil chimique de l'héroïne." Revue Internationale de Criminologie et de Police Technique et Scientifique (RICTPS) **65**(4): 480-491.

Ötles, S., Ed. (2008). Handbook of Food Analysis Instruments, CRC Press.

Pazos, D., P. Giannasi, Q. Rossy and P. Esseiva (2013). "Combining Internet monitoring processes, packaging and isotopic analyses to determine the market structure: Example of Gamma Butyrolactone." Forensic Science International **In press**.

Philip, R. P. (2009). Environmental Forensics: An Evolutionary Perspective. Proceedings of the 2009 International Network of Environmental Forensics (INEF) Annual Conference R. D. Morrison, Royal Society of Chemistry.

Reimann, C., P. Filzmoser, R. Garrett and R. Dutter (2008). Statistical data analysis explained : applied environmental statistics with R. Chichester, England ; Hoboken, NJ, John Wiley & Sons.

Robertson, B. and G. A. Vignaux (1995). Interpreting Evidence : Evaluating Forensic Science in the Courtroom. Chichester, Wiley.

Sacks, R. D. (2004). High-Speed Gas Chromatography. Modern Practice of Gas Chromatography (Fourth Edition). E. F. B. Robert L. Grob**:** 229-274.

Sacré, P. Y., E. Deconinck, M. Daszykowski, P. Courselle, R. Vancauwenberghe, P. Chiap, J. Crommen and J. O. De Beer (2011). "Impurity fingerprints for the identification of counterfeit medicines-A feasibility study." Analytica Chimica Acta **701**(2): 224-231.

Shibuya, E. K., J. E. Souza Sarkis, O. N. Neto, M. Z. Moreira and R. L. Victoria (2006). "Sourcing Brazilian marijuana by applying IRMS analysis to seized samples." Forensic Science International **160**(1): 35-43.

Stella, C., S. Rudaz, J. Y. Gauvrit, P. Lantéri, A. Huteau, A. Tchapla and J. L. Veuthey (2007). "Characterization and comparison of the chromatographic performance of different types of reversed-phase stationary phases." Journal of Pharmaceutical and Biomedical Analysis **43**(1): 89-98.

Stojanovska, N., S. Fu, M. Tahtouh, T. Kelly, A. Beavis and K. P. Kirkbride (2013). "A review of impurity profiling and synthetic route of manufacture of methylamphetamine, 3,4-methylenedioxymethylamphetamine, amphetamine, dimethylamphetamine and p-methoxyamphetamine." Forensic Science International **224**(1-3): 8-26.

Strömberg, L., L. Lundberg, H. Neumann, B. Bobon, H. Huizer and N. W. van der Stelt (2000). "Heroin impurity profiling: A harmonization study for retrospective comparisons." Forensic Science International **114**(2): 67-88.

Swist, M., J. Wilamowski and A. Parczewski (2005). "Determination of synthesis method of ecstasy based on the basic impurities." Forensic Science International **152**(2-3): 175-184.

Tagliaro, F., J. Pascali, A. Fanigliulo and F. Bortolotti (2010). "Recent advances in the application of CE to forensic sciences: A update over years 2007-2009." Electrophoresis **31**(1): 251-259.

Terrettaz-Zufferey, A. L., F. Ratle, O. Ribaux, P. Esseiva and M. Kanevski (2007). "Pattern detection in forensic case data using graph theory: Application to heroin cutting agents." Forensic Science International **167**(2-3): 242-246.

UNODC (2001). Drug characterization/Impurity profiling, *Background and Concepts*. *S. Section*. Vienna, **United Nations**.

UNODC (2005). Methods for impurity profiling of heroin and cocaine. L. a. S. Section. Vienna, UNITED NATIONS.

van Deursen, M. M., J. Beens, H. G. Janssen, P. A. Leclercq and C. A. Cramers (2000). "Evaluation of time-of-flight mass spectrometric detection for fast gas chromatography." Journal of Chromatography A **878**(2): 205-213.

Varmuza, K. and P. Filzmoser (2009). Introduction to Multivariate Statistical Analysis in Chemometrics, CRC Press.

Waddell-Smith, R. J. H. (2007). "A review of recent advances in impurity profiling of illicit MDMA samples." Journal of Forensic Sciences **52**(6): 1297-1304.

Ward, J. H., Jr. (1963). "Hierarchical Grouping to Optimize an Objective Function." Journal of the American Statistical Association **58**(301): 236-244.

West, J. B., J. M. Hurley and J. R. Ehleringer (2009). "Stable isotope ratios of marijuana. I. Carbon and nitrogen stable isotopes describe growth conditions." Journal of Forensic Sciences **54**(1): 84-89.

Weyermann, C., R. Marquis, C. Delaporte, P. Esseiva, E. Lock, L. Aalberg, J. S. Bozenko Jr, S. Dieckmann, L. Dujourdy and F. Zrcek (2008). "Drug intelligence based on MDMA tablets data. I. Organic impurities profiling." Forensic Science International **177**(1): 11-16.

Wool, L. and D. Decker (2002). "Practical fast gas chromatography for contract laboratory program pesticide analyses." Journal of Chromatographic Science **40**(8): 434-440.

Zacca, J. J., T. S. Grobério, A. O. Maldaner, M. L. Vieira and J. W. B. Braga (2013). "Correlation of cocaine hydrochloride samples seized in Brazil based on determination of residual solvents: An innovative chemometric method for determination of linkage thresholds." Analytical Chemistry **85**(4): 2457-2464.

Liste des figures

INTRODUCTION

CHAPITRE 1

CHAPITRE 2

CHAPITRE 4

CHAPITRE 5

CHAPITRE 6

CHAPITRE 7

CHAPITRE 8

CHAPITRE 9

CHAPITRE 10

Liste des tableaux

CHAPITRE 1

CHAPITRE 2

CHAPITRE 3

CHAPITRE 4

CHAPITRE 5

CHAPITRE 6

CHAPITRE 7

CHAPITRE 8

CHAPITRE 9

CHAPITRE 10

UNIL | Université de Lausanne

Faculté de Droit et des Sciences Criminelles

Ecole des Sciences Criminelles

Institut de Police Scientifique

Etude de l'approvisionnement d'une banque de données avec les résultats provenant de méthodes analytiques différentes dans le cadre du profilage chimique de produits stupéfiants

CAHIER DES ANNEXES

Julian Broséus

LAUSANNE

2013

Annexe 1 Ajustement analytique

La méthodologie suivante devrait permettre d'obtenir une bonne connaissance des paramètres du « tune » sélectionnés pour leur influence potentielle sur la similarité statistique des résultats et ainsi pertinents dans le cadre de l'approche d'ajustement analytique (cf. Chapitre 6, §6.4.a). Une telle méthodologie (cf. Tableau 1) fournira des informations sur la manière de d'agir sur ces paramètres pour assurer par exemple la répétabilité/reproductibilité des analyses dans le cadre du profilage chimique. Ainsi, dans le cadre du contrôle du bon fonctionnement de l'appareillage, si une valeur de similarité sortait du « control chart », il s'agirait alors d'estimer si cela peut être attribué à la modification d'un paramètre du « tune » en particulier. Il s'agirait également, sur la base des résultats obtenus, d'évaluer a priori si une modification dans les valeurs du « tune » pour les paramètres susmentionnés pourrait avoir une influence sur le profil chimique et donc sur la valeur de similarité calculée. Finalement, par extension, seuls les paramètres influençant la similarité des profils chimiques pourraient être déterminés (pour une future application dans le cadre de l'ajustement analytique, par exemple). Dans le cadre de la problématique générale de cette recherche, une telle investigation pourrait permettre de déterminer, si l'on vise à combiner les résultats de différentes méthodes analytiques, s'il est pertinent de mettre les mêmes valeurs à ces paramètres du « tune » MS, ou si, comme l'estime la méthodologie d'harmonisation des méthodes analytiques, ceci n'a pas d'influence sur la similarité des résultats analytiques.

Etape n°	Détails
1	Examen approfondi de l'évolution des valeurs de ces paramètres en regard de l'évolution du control chart (établi à l'aide de mesures de similarité d'échantillons de référence) sur l'instrument analytique utilisé en systématique à l'IPS (NB : l'obtention d'une valeur extrême ou hors limites n'est pas nécessairement due à une modification dans les paramètres du « tune », mais peut être le signe d'une maintenance à effectuer).
2	Création d'un control chart, sur un instrument analytique dédié à cette étude : - Analyses et construction du control chart en prenant en compte les variabilités dues aux maintenances d'appareillage (plus ou moins importantes, par exemple, changement du liner, de la colonne, nettoyage de la source, changement du filament etc.) et examen de leurs influences sur la similarité obtenue ; - Mesures de la similarité sur la base des composés présents dans les solutions « Grob » ; - Mesures de la similarité sur la base des composés cibles du profil chimique présents dans les 3 spécimens de référence, de concentrations différentes, définis plus haut.
3	Les réponses dans le modèle seraient les surfaces nominales de chacun des composés (aires brutes et prétraitées des ions cibles des composés respectifs), selon le type d'échantillons utilisé (« Grob » ou spécimens d'héroïne) ainsi que les 8 réponses obtenues suite au « tune » du MS sur le calibrant, le PFTBA (cf. Chapitre 2).
4	Détermination des bornes pour chacun des critères a) « one factor at a time » (c'est-à-dire, un critère pour lequel on modifie la valeur nominale (médiane) et les autres que l'on laisse à leurs valeurs nominales respectives) ; b) dans un premier temps, regarder si les paramètres d'acceptabilité du tune d'après la calibration sur le PFTBA sont toujours atteints selon la diminution/augmentation des valeurs d'un paramètre ; c) en regard des aires brutes (p. ex. le signal obtenu est-il toujours acceptable malgré cette nouvelle valeur, en particulier pour les composés en plus faible concentration ?).
5	« Design of Experiment » (Brereton, 2007) a) 35 expérimentations (« Full Factorial » à 2 niveaux sur 5 facteurs (32) + 3 expérimentations aux valeurs médianes de tous les paramètres) b) Y = 7 rép. PFTBA et les aires brutes/pré-traitées pour chacun des composés du profil
6	Etude des distributions des populations d'intra- et d'inter variabilité, et leurs variations lorsque les analyses sont faites avec différentes combinaisons du « tune » et en variant les valeurs de ces paramètres

Tableau 1. Définition des différentes étapes envisagées pour l'investigation des paramètres du "tune" MS

Annexe 2 Méthodes analytiques implémentées

Les lecteurs intéressés par les paramètres d'analyses implémentés dans le cadre du scénario d'ajustement 2.2 sont renvoyés à la publication correspondante (Debrus et al., 2010).

1.1 Méthode de référence GC-MS

Pour les analyses un GC Agilent 6890 couplé à un spectromètre de masse Agilent 5975 a été utilisé. La séparation des analytes est accomplie sur une colonne capillaire HP5-ms (30 m longueur x 0.25 mm diamètre interne x 0.25 µm épaisseur de film, J&W Scientific). Les injections sont réalisées en mode split avec un liner rempli de laine de verre (Agilent Technologies No. 5183-4711). Le programme de température commence à 150°C, augmente successivement à 250°C (8°C/min) puis 320°C (6°C/min) pour une durée totale de 24 minutes environ. 2 µL de chaque échantillon ont été injectés avec l'hélium comme gaz porteur (flux constant, 1 mL/min) avec un rapport de split de 1 :50. Les températures appliquées sont de 250°C pour l'injecteur, 280°C pour la ligne de transfert, 230°C pour la source ionique et 150°C pour le quadrupole. Les données ont été acquises en mode scan (30-450 m/z) avec un sampling rate de 3 (1.77 scans/s) et analysées avec le logiciel MSD Enhanced ChemStation v. D.02.00.275 (Agilent Technologies).

Le chromatogramme typique obtenu pour l'analyse d'un spécimen de référence est illustré par la Figure 1.

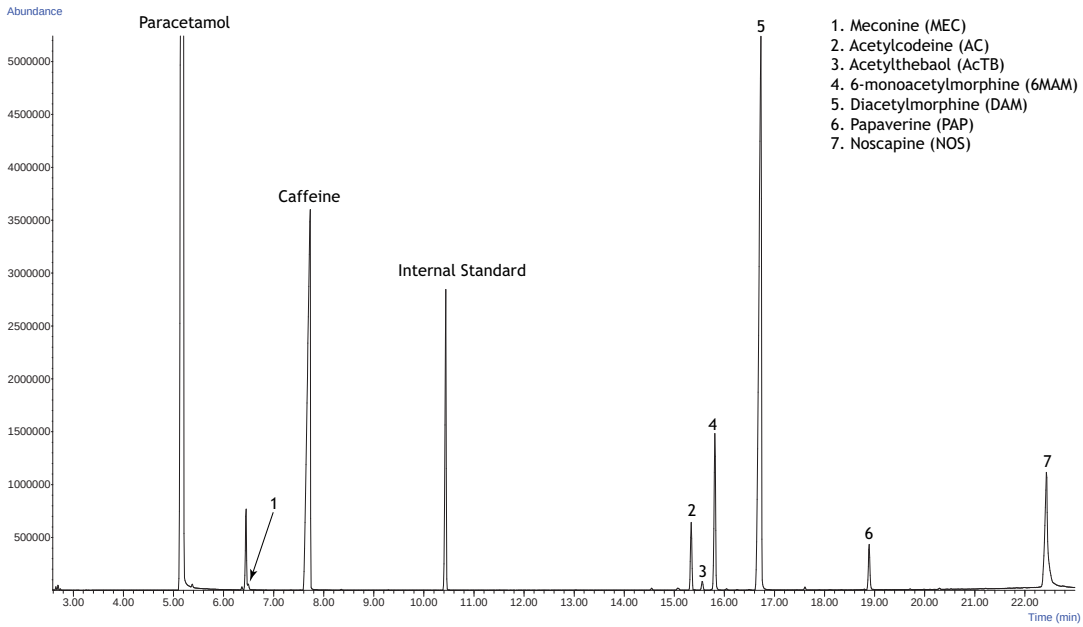


Figure 1. Chromatogramme typique d'un échantillon d'héroïne obtenu avec la méthode de référence GC-MS

1.2 Scénario 1.2

Pour les analyses un GC Agilent 7890A couplé à un spectromètre de masse Agilent Technologies 5975C inert XL MSD Triple Axis Detector a été utilisé. La séparation des analytes est accomplie sur une colonne capillaire HP5-ms (12 m longueur x 0.20 mm diamètre interne x 0.33 µm épaisseur de film, J&W Scientific). Les injections sont réalisées en mode split avec un liner rempli de laine de verre (Agilent Technologies No. 5183-4711). Le programme de température commence à 180°C, augmente successivement à 260°C (40°C/min, palier tenu 0.5 min), 280°C (60°C/min, palier tenu 0.5 min) puis 300°C (40°C/min, palier tenu 2 min) pour une durée totale de 6 minutes environ. 1 µL de chaque échantillon a été injecté avec l'hélium comme gaz porteur (flux constant, 1 mL/min) avec un rapport de split de 1 :25. Les températures appliquées sont de 270°C pour l'injecteur, 250°C pour la ligne de transfert, 230°C pour la source ionique et 150°C pour le quadrupole. Les données ont été acquises en mode scan (50-450 m/z) avec un sampling rate de 1 et analysées avec le logiciel MSD Enhanced ChemStation v. D.02.00.275 (Agilent Technologies).

Les chromatogrammes obtenus sur chaque instrument pour l'analyse d'un même spécimen de référence sont illustrés par la Figure 2, la Figure 3, la Figure 4 et la Figure 5, respectivement.

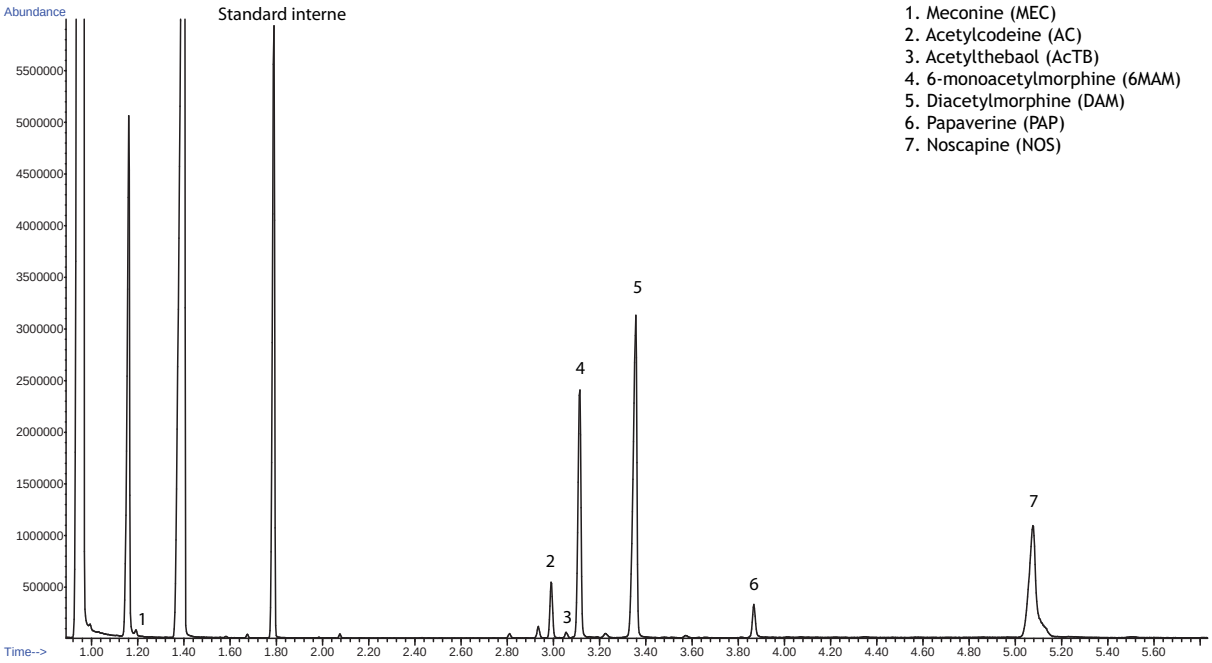


Figure 2. Chromatogramme typique d'un échantillon d'héroïne obtenu avec APP 1

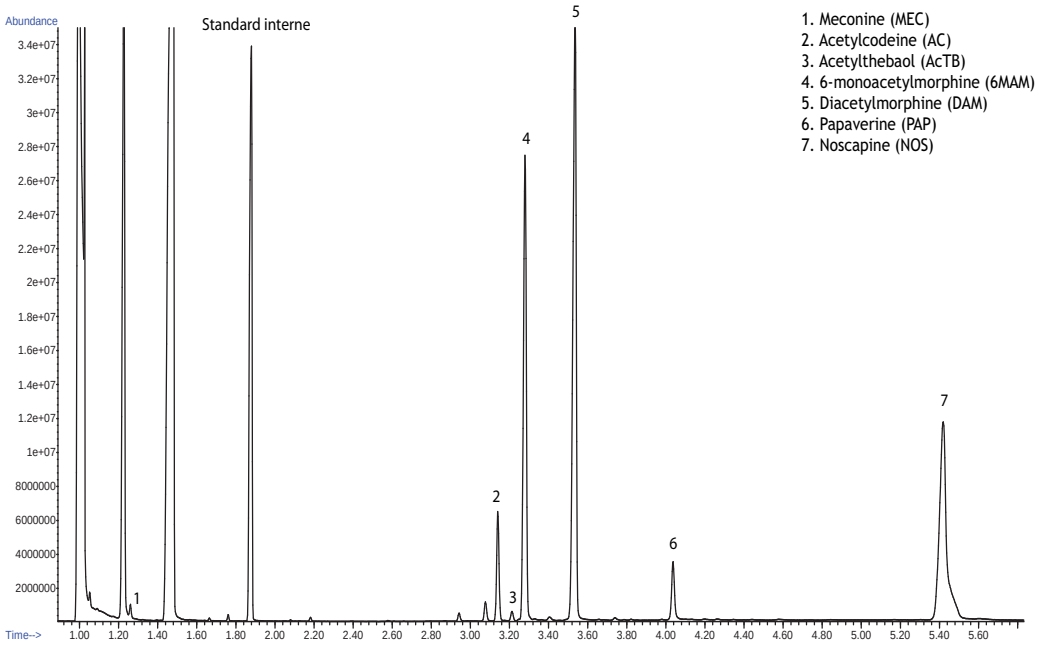


Figure 3. Chromatogramme typique d'un échantillon d'héroïne obtenu avec APP 2

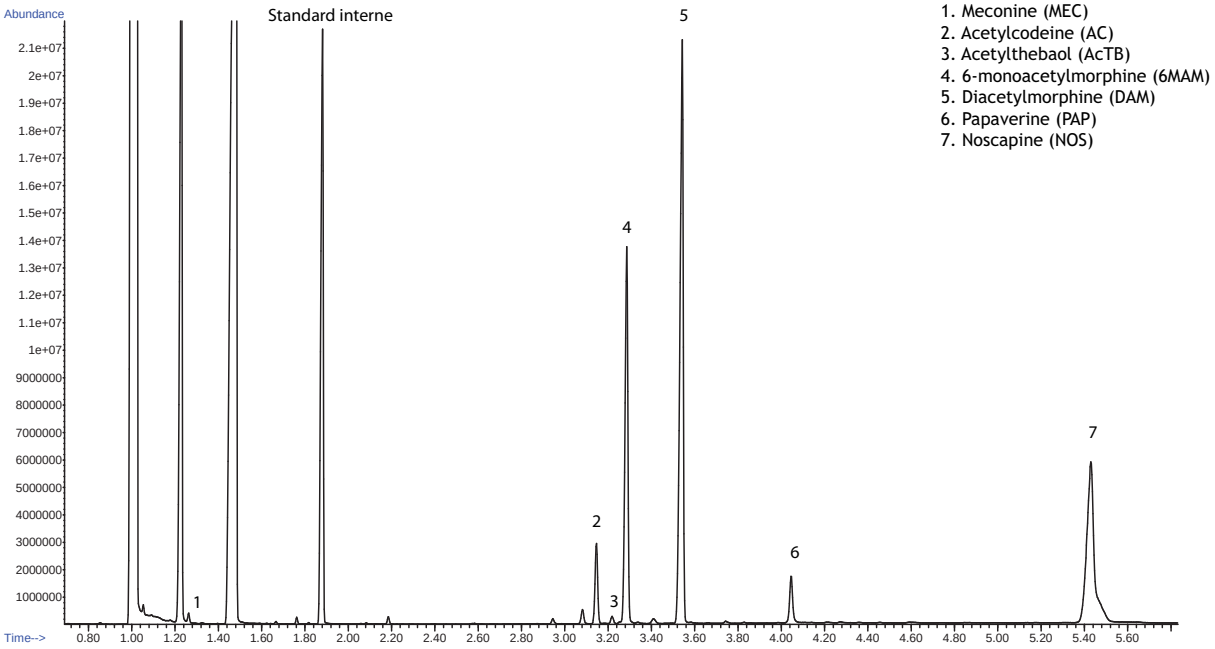


Figure 4. Chromatogramme typique d'un échantillon d'héroïne obtenu avec APP 3

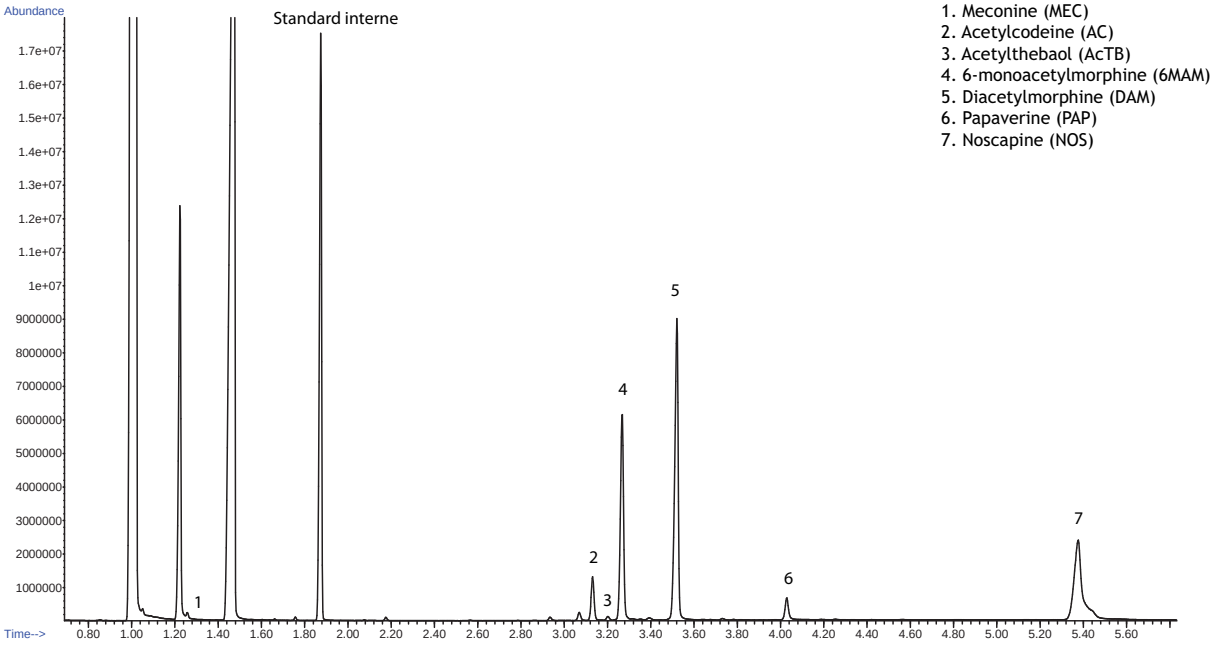


Figure 5. Chromatogramme typique d'un échantillon d'héroïne obtenu avec APP 4

1.3 Scénario 1.4

Pour les analyses un GC Agilent 6890A couplé à un spectromètre de masse Agilent 5975C a été utilisé. La séparation des analytes est accomplie sur une colonne capillaire HP5-ms (20 m longueur x 0.18 mm diamètre interne x 0.18 µm épaisseur de film, J&W Scientific). Les injections sont réalisées en mode split avec un liner rempli de laine de verre (Agilent Technologies No. 5183-4711). Le programme de température commence à 180°C, augmente successivement à 290°C (60°C/min, palier tenu 0.7 min) puis 320°C (8°C/min, palier tenu 0.1 min) pour une durée totale de 6 minutes environ. 1 µL de chaque échantillon a été injecté avec l'hélium comme gaz porteur (flux constant, 0.72 mL/min) avec un rapport de split de 1 :70. Les températures appliquées sont de 250°C pour l'injecteur, 280°C pour la ligne de transfert, 230°C pour la source ionique et 150°C pour le quadrupole. Les données ont été acquises en mode scan (40-450 m/z) avec un sampling rate de 1 et analysées avec le logiciel MSD Enhanced ChemStation v. D.02.00.275 (Agilent Technologies).

Le chromatogramme typique obtenu pour l'analyse d'un spécimen de référence est illustré par la Figure 6.

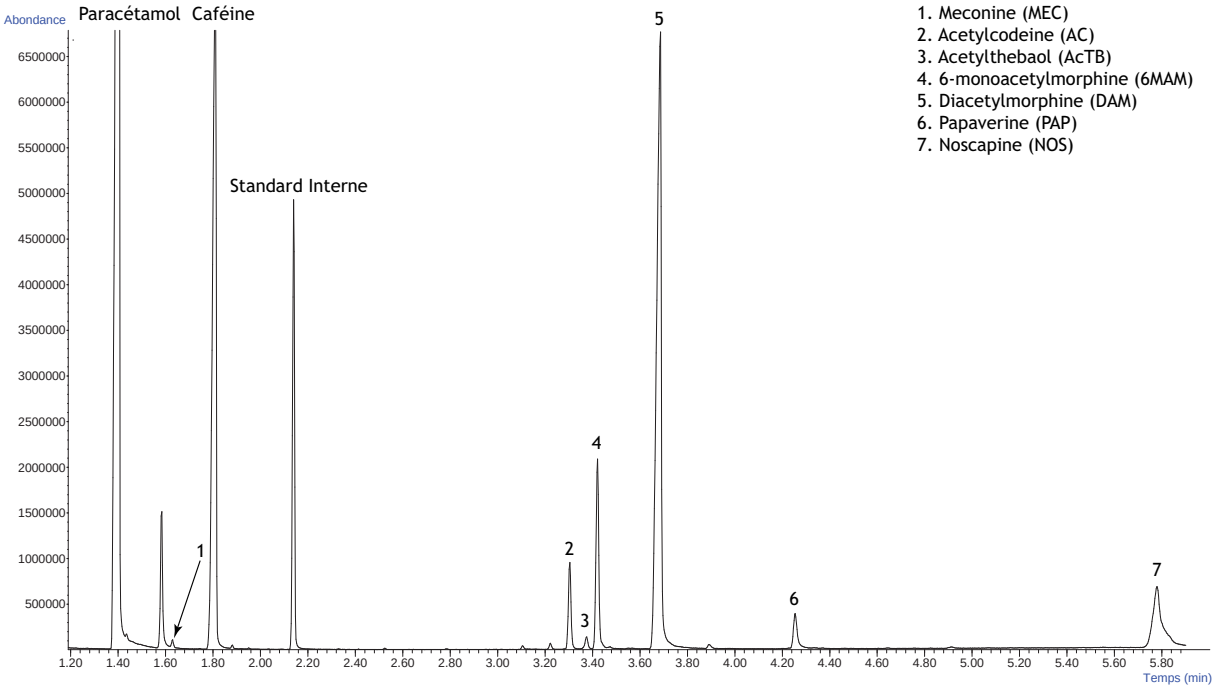


Figure 6. Chromatogramme typique d'un échantillon d'héroïne obtenu en Fast GC-MS

1.4 Scénario 1.5

Pour les analyses un GC Perkin Elmer Clarus 500 couplé à un spectromètre de masse Perkin Elmer Clarus 560D/S a été utilisé. La séparation des analytes est accomplie sur une colonne capillaire HP5-ms (30 m longueur x 0.25 mm diamètre interne x 0.25 µm épaisseur de film, J&W Scientific). Les injections sont réalisées en mode split avec un liner rempli de laine de verre (Perkin Elmer N6502009). Le programme de température commence à 150°C, augmente successivement à 250°C (8°C/min) puis 320°C (6°C/min) pour une durée totale de 24 minutes environ. 2 µL de chaque échantillon a été injecté avec l'hélium comme gaz porteur (flux constant, 1 mL/min) avec un rapport de split de 1 :50. Les températures appliquées sont de 250°C pour l'injecteur, 280°C pour la ligne de transfert, 230°C pour la source ionique et 150°C pour le quadrupole. Les données ont été acquises en mode scan (10-450 m/z) et analysées avec le logiciel TurboMass 5.4.2 GC/MS Software.

Le chromatogramme typique obtenu pour l'analyse d'un spécimen de référence est illustré par la Figure 7.

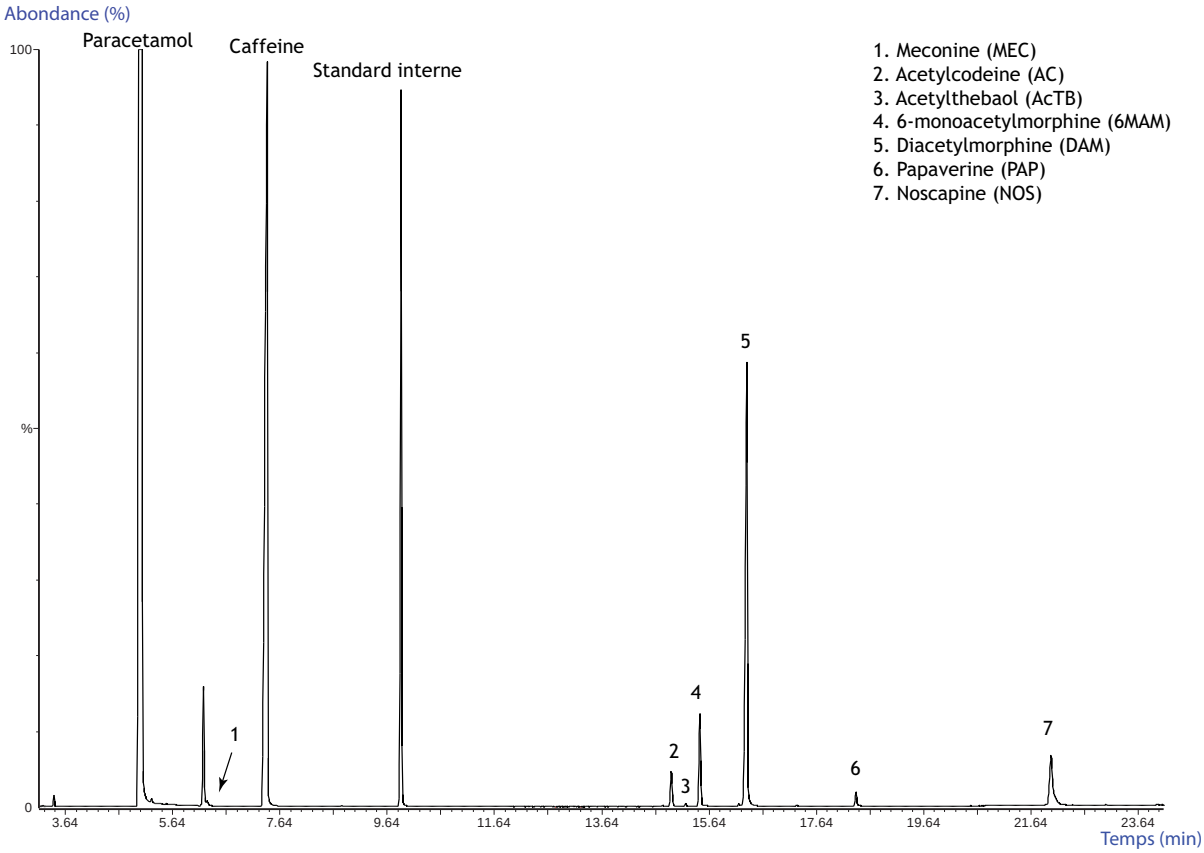


Figure 7. Chromatogramme typique d'un échantillon d'héroïne obtenu en GC-MS avec un appareillage Perkin Elmer

1.5 Scénario 2.1

Pour les analyses un GC Agilent 6850 Series II couplé à un détecteur à ionisation de flamme (FID). La séparation des analytes est accomplie sur une colonne capillaire HP5-ms (20 m longueur x 0.18 mm diamètre interne x 0.18 µm épaisseur de film, J&W Scientific). Les injections sont réalisées en mode split avec un liner rempli de laine de verre (Agilent Technologies No. 5183-4711). Le programme de température commence à 180°C, augmente successivement à 295°C (60°C/min) puis 320°C (10°C/min, palier tenu 0.3 min) pour une durée totale de 5 minutes environ. 1 µL de chaque échantillon a été injecté avec l'hélium comme gaz porteur (flux constant, 1.3 mL/min) avec un rapport de split de 1 :70. Les températures appliquées sont de 250°C pour l'injecteur et 300°C pour le détecteur. La fréquence d'acquisition a été fixée à 20 Hz. et analysées avec le logiciel MSD Enhanced ChemStation v. D.02.00.275 (Agilent Technologies).

Le chromatogramme typique obtenu pour l'analyse d'un spécimen de référence est illustré par la Figure 8.

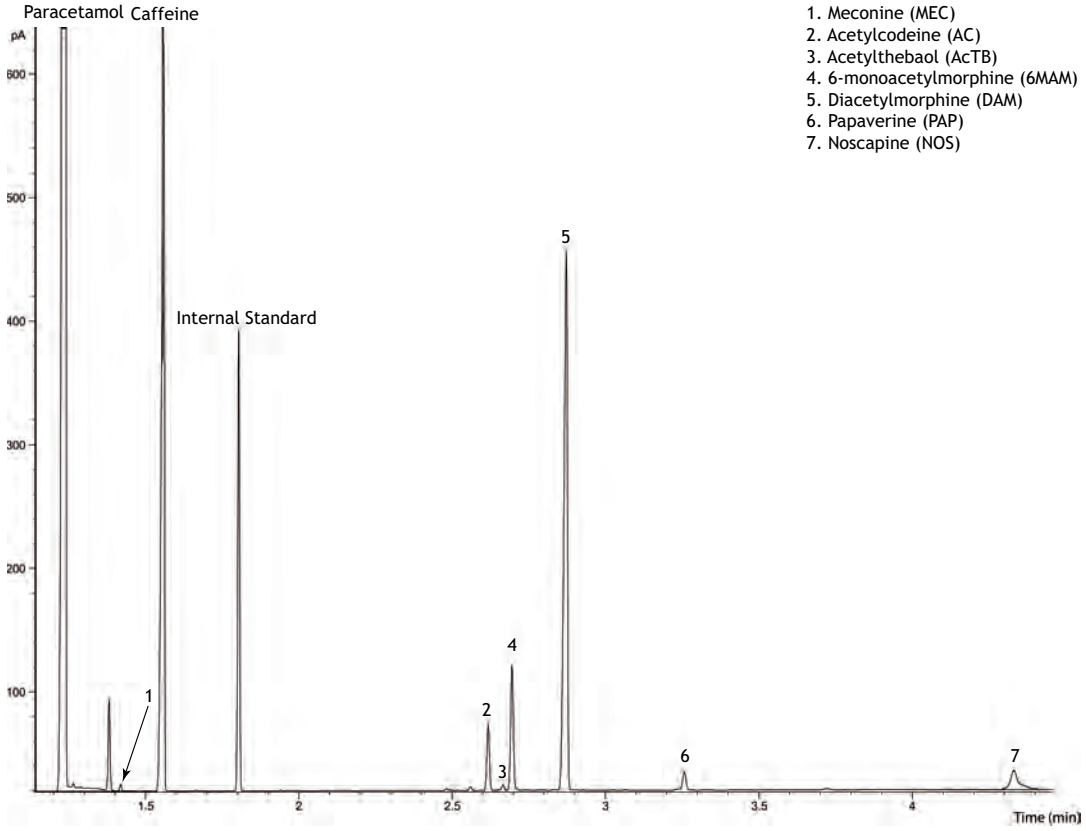


Figure 8. Chromatogramme typique d'un échantillon d'héroïne de référence obtenu en Fast GC-FID

Annexe 3 ACP-CAH Globale et Locale

Ci-dessous se trouve le code source développé avec le logiciel Revolution R Enterprise version 6.1.0 utilisant le logiciel R version 2.14.2.

Packages utiles

```
library(ade4)
library(car)
library(MASS)
library(sampling)
```

Chemin des codes sources

```
source("D:/Données/Boulot/Résultats/Scripts/pareto-plot.r")
source("D:/Données/Boulot/Résultats/Scripts/crossvalstepAIC.r")
source("D:/Données/Boulot/Résultats/Scripts/criteria.r")
setwd("D:/Données/Boulot/Résultats/GC-FID/Tests")
```

1. Données GC/MS

Importation

```
o.data<-read.table(file="D:/Données/Boulot/Résultats/GC-
FID/BDD2009_SIMULdata.csv",header=T,sep=";")
rownames(o.data)<-o.data[,1]
gc.base<-o.data[,-1]
```

Groupage des échantillons (replicats -> specimen)

```
myseq<-vector()
char.tst<-rownames(gc.base)
char.grp<-char.tst
for (i in 1:nrow(gc.base)) {
nc<-nchar(char.tst[i])
char.grp[i]<-substring(char.tst[i],1,nc-1)
}

dup<-unique(char.grp)
group<-vector()
ind<-1
for (i in 1:length(dup)){
tmp<-which(char.grp==dup[i])
group[tmp]<-ind
ind<-ind+1
}
gc.base<-cbind(group,gc.base)
```

Calcul de la moyenne des réplicats pour chaque échantillon

```
ngroup<-max(gc.base[,1])
gc.tmp<-vector()
tmp<-as.data.frame(matrix(ncol=8,nrow=1))
for (i in 1:ngroup){
w<-which((gc.base[,1]==i)==TRUE)
tmp[1,2:8]<-mean(gc.base[w,2:8])
rownames(tmp)<-dup[i]
tmp[1,1]<-gc.base[w,1:2][1,1]
gc.tmp<-rbind(gc.tmp,tmp)
}
colnames(gc.tmp)<-colnames(gc.base)
gc.data<-gc.tmp[,-c(1)]
```

Modification des rownames

```
for (i in 1:length(rownames(gc.data))) {
if (substring(rownames(gc.data)[i],nchar(rownames(gc.data)[i]),nchar(rownames(gc.data)[i]))=="_")
rownames(gc.data)[i]<- substring(rownames(gc.data)[i],1,nchar(rownames(gc.data)[i])-1)
}
```

1.1. Prétraitement des données (Normalisation + UV scaling)

```
gc<-gc.data[,-5]
gc[,7]<-apply(gc,1,sum)
hero<-t(apply(as.matrix(gc),1,function(x)sqrt((x/x[7]))))
gc.data<-as.data.frame(hero)[,-7]
gc.data<-as.data.frame(scale(gc.data))
```

2. Importation des données de l'autre méthode

```
u.data<-read.table(file="D:/Données/Boulot/Résultats/GC-FID/FIDRawData_cleaned_wo
outliers.csv",header=T,sep=";")
rownames(u.data)<-u.data[,1]
u.data <- u.data[,-1]
```

Groupage des échantillons (réplicats -> spécimen)

```
myseq<-vector()
char.tst<-rownames(u.data)
char.grp<-char.tst
for (i in 1:nrow(u.data)) {
nc<-nchar(char.tst[i])
char.grp[i]<-substring(char.tst[i],1,nc-1)
}

dup<-unique(char.grp)
group<-vector()
ind<-1
for (i in 1:length(dup)){
tmp<-which(char.grp==dup[i])
group[tmp]<-ind
ind<-ind+1
```

```
}
u.data2<-cbind(group,u.data)
```

Calcul de la moyenne des réplicats pour chaque échantillon

```
ngroup<-max(u.data2[,1])
uplc.base<-vector()
tmp<-as.data.frame(matrix(ncol=8,nrow=1))
for (i in 1:ngroup){
w<-which((u.data2[,1]==i)==TRUE)
tmp[1,2:8]<-mean(u.data2[w,2:8])
rownames(tmp)<-dup[i]
tmp[1,1]<-u.data2[w,1:2][1,1]
uplc.base<-rbind(uplc.base,tmp)
}
```

Noms des colonnes

```
uplc.data<-uplc.base[,-1]
colnames(uplc.data)<-colnames(u.data)
```

2.1 Prétraitement des données

```
up<-uplc.data[,-5]
up[,7]<-apply(up,1,sum)
hero<-t(apply(as.matrix(up),1,function(x)sqrt((x/x[7]))))
up<-as.data.frame(hero)[,-7]
uplc.data_a<-up
uplc.data_a<-as.data.frame(scale(uplc.data_a))
```

3. Approche ACP-CAH Globale et Locale

Création des objets nécessaires (stockage des données)

```
resultats.finaux_cal<-vector()
resultats.finaux_val<-vector()
succes.adj2 <- vector()
nosucces.adj2 <- vector()
succes.adj <- vector()
nosucces.adj <- vector()
tmp.m <- vector()
tmp.ns.m <- vector()
tmp.m2 <- vector()
tmp.ns.m2 <- vector()
cor.tmp <- vector()
cor.ns.tmp <- vector()
cor.ns.tmp2 <- vector()
cor.tmp2 <- vector()
```

Création des 45 valeurs de h testées

```
my.h.increment <-seq(0.5,10,0.5)
my.h.increment <-c(my.h.increment,seq(11,20,1))
my.h.increment <-c(my.h.increment,seq(22,50,2))
total <- length(my.h.increment)
```

Création de la barre d'avancement du code

```
pb <- winProgressBar(title = "Vive la pêche", min = 0,max = total, width = 300)
```

!!! Première boucle : Tests de chacune des 45 valeurs de h !!!

```
for(h.ju in my.h.increment){
```

!!! Deuxième boucle : 100 itérations à chaque valeur de h !!!

```
for (iter in 1:100){
```

3.1 Création des sets de calibration et validation pour la méthode différente

```
st <- strata(data=uplc.data_a,size=round((2/3)*nrow(uplc.data_a)), method=c("srswor"))
testindex <- st$ID_unit
uplc.data <- uplc.data_a[testindex, ]
uplc2.data <- uplc.data_a[-testindex, ]
```

3.2 Création set de Cal GC-MS (extraction depuis la BDD de référence)

```
gc.tested<-vector()
for (i in 1:length(rownames(uplc.data))){
w<-which(rownames(gc.data)==rownames(uplc.data)[i])
if (length(w)==0) cat(i,"\n")
tmp<-gc.data[w,]
gc.tested<-rbind(gc.tested,tmp)
}
colnames(gc.tested)<-colnames(gc.data)
```

3.3 Création set de Val GC-MS (extraction depuis la BDD de référence)

```
gc2.tested<-vector()
for (i in 1:length(rownames(uplc2.data))){
w<-which(rownames(gc.data)==rownames(uplc2.data)[i])
if (length(w)==0) cat(i,"\n")
tmp<-gc.data[w,]
gc2.tested<-rbind(gc2.tested,tmp)
}
colnames(gc2.tested)<-colnames(gc.data)
```

3.4 Modèles mathématiques / Règles d'ajustement

```
cat("MLR started...","\n")
source("C:/Données/Thèse/Résultats_Thèse/le belge/Q2-cross-validation.r")
source("C:/Données/Thèse/Résultats_Thèse/le belge/Q2-cross-validation_group.r")
def.par <- par(no.readonly = TRUE)
```

!!! Définition des x du modèle !!!

```
attach(uplc.data)
```

Modèle

```
fit<-list()
for (i in 1:6){ # 6 variables
```

!!! Définition des y du modèle !!!

```
y<-gc.tested[,i]
```

Modèles (linéaire, quadratique, cubique / à choisir en enlevant le « # »)

```
f.upper<-"y~"
f.upper<-paste(f.upper,"I(",colnames(uplc.data)[i],")",sep="")
#f.upper<-paste(f.upper,"I(",colnames(uplc.data)[i],")","+I(",colnames(uplc.data)[i],"^2)",sep="")
#f.upper<-
paste(f.upper,"I(",colnames(uplc.data)[i],")","+I(",colnames(uplc.data)[i],"^2)","+I(",colnames(uplc.data)[i],"^3)",sep="")
```

Régression

```
fit[[i]]<- lm(f.upper)
```

Calcul du Q^2

```
q2<-Q2_CVg(fit[[i]],uplc.data,y,30,250)
```

Projection des données par variable avec R^2

```
par(mfrow=c(1,2))
plot(predict.lm(fit[[i]],uplc.data),gc.tested[,i],pch=21,bg=rgb(0,0,0,0.15),xlab="predicted
values",ylab="observed
values",main=colnames(uplc.data)[i],sub=paste("R²adjusted=",round(summary(fit[[i]])$adj,4)))
abline(0,1,col="blue",lty=2)
plot(predict.lm(fit[[i]],uplc.data),gc.tested[,i]-
predict.lm(fit[[i]],uplc.data),pch=21,bg=rgb(0,0,0,0.15),xlab="predicted
values",ylab="residuals",sub=paste("Q²=",round(q2,4)))
text(predict.lm(fit[[i]],uplc.data),gc.tested[,i]-
predict.lm(fit[[i]],uplc.data),names(predict.lm(fit[[i]],uplc.data)),pos=3,cex=0.5)
title(colnames(uplc.data)[i])
dev.copy2pdf(file=paste(i,"-",colnames(uplc.data)[i],"-model.pdf",sep=""))
}
detach(uplc.data)

graphics.off()

par(def.par)
for (i in 1:6) {
cat(" ----------","\n",colnames(uplc.data)[i],"\n","----------","\n")
pareto.plot(fit[[i]],option="squared",new.window=FALSE,main=colnames(uplc.data)[i])
dev.copy2pdf(file=paste(i,"-",colnames(uplc.data)[i],"-pareto-plot.pdf",sep=""))

print(summary(fit[[i]]))
}
graphics.off()
```

4. Ajustement Mathématique

Prédiction : Set de Cal "GC-MS like"

```
uplc.pred<-vector()
for (i in 1:6) {
tmp<-predict.lm(fit[[i]],uplc.data)
uplc.pred<-cbind(uplc.pred,tmp)
}
colnames(uplc.pred)<-colnames(gc.tested)
uplc.pred<-as.data.frame(uplc.pred)
```

Prédiction : Set de Val "GC-MS like"

```
uplc2.pred<-vector()
for (i in 1:6) {
tmp<-predict.lm(fit[[i]],uplc2.data)
uplc2.pred<-cbind(uplc2.pred,tmp)
}
colnames(uplc2.pred)<-colnames(gc2.tested)
uplc2.pred<-as.data.frame(uplc2.pred)
```

5. Calcul de la performance d'ajustement : Set de Cal

Création de la BDD réduite

```
rem<-vector()
for (i in 1:nrow(uplc.data)){
rem[i]<-which(rownames(gc.data)==rownames(uplc.data)[i])
}
gc.m36<-gc.data[-rem,]
nt<-1

mydist<-vector()
myheight<-vector()
myheightst<-vector()
mylogical<-vector()
mylogicalst<-vector()
```

Analyse des profils chimiques de l'échantillonnage les uns après les autres

```
for (nt in 1:length(rem)){
cat("First step",nt,"/",length(rem),"- working on",rownames(uplc.data)[nt],"...")
```

5.1 ACP-CAH globale : ACP

ACP sur la BDD réduite

```
pca<-prcomp(gc.m36,scale=FALSE)
```

Transposition des profils GC dans l'espace ACP de la BDD réduite

```
gc<-gc.tested
tmp<-t(apply(gc,1,function(x)x-pca$center))
tmp<-as.matrix(tmp)
tmp<-t(apply(tmp,1,function(x) x%*%pca$rotation))
```

Transposition des profils de l'autre méthode dans l'espace ACP de la BDD réduite

```
up<-uplc.pred
tmp2<-t(apply(up,1,function(x)x-pca$center))
tmp2<-as.matrix(tmp2)
tmp2<-t(apply(tmp2,1,function(x) x%*%pca$rotation))
```

Sélection du nombre de CPs adéquat

```
npc<-4
```

Ajout à la BDD réduite du couple de profils chimiques pour le specimen correspondant

```
colnames(tmp)<-colnames(pca$x)
colnames(tmp2)<-colnames(pca$x)
global<-rbind(pca$x[,1:npc],tmp[nt,1:npc],tmp2[nt,1:npc])
rownames(global)[nrow(global)-1]<-rownames(tmp)[nt]
rownames(global)[nrow(global)]<-rownames(tmp)[nt]
```

5.1 ACP-CAH globale : CAH

Calcul de la distance euclidienne sur les scores des CPs correspondantes

```
tmp.dist<-dist(rbind(tmp[nt,1:npc],tmp2[nt,1:npc]))
mydist<-rbind(mydist,tmp.dist)

pca.dist<-dist(global) # calcul des distances euclidiennes
hc<-hclust(pca.dist,method="ward") # groupement méthode Ward
```

Coupe du dendrogramme pour voir à quelle hauteur le cluster contient les données GC-MS et « GC-MS like »

```
.calch<-TRUE #FALSE
if (.calch) {
c.tmp<-vector()
clust<-vector()
myseq<-seq(0,5,0.001)
myseq<-c(myseq,seq(5.1,50,0.1))
myseq<-c(myseq,seq(51,max(hc$height),1))
for (i in myseq){
mycut<-cutree(hc,h=i)
myclust<-mycut[which(names(mycut)==rownames(tmp)[nt])]
if (length(myclust)>2) stop("more than 2 samples GC-UPLC")
if (myclust[1]==myclust[2]) break
}
```

```
tmp.h<-i
myheight<-rbind(myheight,tmp.h)
}
```

Contrôle de la présence des deux profils dans le même cluster (dendrogramme globale) pour une hauteur donnée

```
mycut<-cutree(hc,h=5)
myclust<-mycut[which(names(mycut)==rownames(tmp)[nt])]
if (length(myclust)>2) stop("more than 2 samples GC-UPLC")
tmp.h<-0
if (myclust[1]==myclust[2]) tmp.h<-1
mylogical<-rbind(mylogical,tmp.h)

cat("Done","\n")
```

5.2 ACP-CAH locale

Sélection d'environ les 30 profils les plus proches

```
c.tmp<-vector()
clust<-vector()
myseq<-seq(0.1,max(hc$height),1)
for (i in myseq){
mycut<-cutree(hc,h=i)
myclust<-mycut[which(names(mycut)==rownames(tmp)[nt])[2]]
c.tmp[1]<-i
c.tmp[2]<-length(which(mycut==myclust))
clust<-rbind(clust,c.tmp)
}

myh<-clust[min(which((clust[,2]-30)>0,arr.ind=TRUE)),1]
myn<-clust[min(which((clust[,2]-30)>0,arr.ind=TRUE)),2]

mycut<-cutree(hc,h=myh)
myclust<-mycut[which(names(mycut)==rownames(tmp)[nt])[2]]
mysamples<-names(which(mycut==myclust))

gc.sub<-vector()
for (i in 1:myn){
g.tmp<-gc.m36[which(rownames(gc.m36)==mysamples[i]),]
gc.sub<-rbind(gc.sub,g.tmp)
}
```

ACP

```
cat("Second step",nt,"/",length(rem),"- working on",rownames(uplc.data)[nt],"...")
pca<-prcomp(gc.sub,scale=FALSE)
```

Transposition des données GC-MS du set de Cal dans l'espace ACP des profils les plus proches

```
gc<-gc.tested
tmp<-t(apply(gc,1,function(x)x-pca$center))
tmp<-as.matrix(tmp)
tmp<-t(apply(tmp,1,function(x) x%*%pca$rotation))
```

Transposition des données « GC-MS like » du set de Cal dans l'espace ACP des profils les plus proches

```
up<-uplc.pred
tmp2<-t(apply(up,1,function(x)x-pca$center))
tmp2<-as.matrix(tmp2)
tmp2<-t(apply(tmp2,1,function(x) x%*%pca$rotation))
```

Ajout au sous-échantillonnage du couple de profils chimiques pour le spécimen correspondant (profils définis par les scores sur les CPs correspondantes)

```
## dendrogram GC-sub
colnames(tmp)<-colnames(pca$x)
colnames(tmp2)<-colnames(pca$x)
global<-rbind(pca$x[,1:npc],tmp[nt,1:npc],tmp2[nt,1:npc])
rownames(global)[nrow(global)-1]<-rownames(tmp)[nt]
rownames(global)[nrow(global)]<-rownames(tmp)[nt]
```

CAH

Calcul des distances euclidiennes entre les scores puis groupement Ward

```
pca.dist<-dist(global)
hc<-hclust(pca.dist,method="ward") # groupement méthode Ward
```

Coupe du dendrogramme pour voir à quelle hauteur se trouvent les deux profils testés

```
if (.calch) {
c.tmp<-vector()
clust<-vector()
myseq<-seq(0,max(hc$height),0.01)
for (i in myseq){
mycut<-cutree(hc,h=i)
myclust<-mycut[which(names(mycut)==rownames(tmp)[nt])]
if (length(myclust)>2) stop("more than 2 samples GC-UPLC")
if (myclust[1]==myclust[2]) break
}
tmp.h<-i
myheightst<-rbind(myheightst,tmp.h)
}
```

Vérification de la présence des deux profils pour une hauteur donnée = h.ju (45 valeurs à tester)

```
mycut<-cutree(hc,h=h.ju)
myclust<-mycut[which(names(mycut)==rownames(tmp)[nt])]
if (length(myclust)>2) stop("more than 2 samples GC-UPLC")
tmp.h<-0
if (myclust[1]==myclust[2]) tmp.h<-1
mylogicalst<-rbind(mylogicalst,tmp.h)
```

Si ajustement réussi (profils GC-MS et "GC-MS like" dans le même cluster d'après h.ju) : calcul coefficient de corrélation de Pearson pour les profils testés

```
if (tmp.h==1){
tmp.m <- rbind(gc.tested[nt,],uplc.pred[nt,])
cor.tmp <- 100*cor(t(tmp.m))
cor.tmp <- as.data.frame(cor.tmp[2,1])
cor.tmp <- cbind(cor.tmp,h.ju)
rownames(cor.tmp) <- rownames(gc.tested[nt,])
succes.adj <- rbind(succes.adj,cor.tmp)
```

Si ajustement non réussi (profils GC-MS et "GC-MS like" pas dans le même cluster d'après h.ju) : calcul coefficient de corrélation de Pearson pour les profils testés

```
} else {

tmp.ns.m <- rbind(gc.tested[nt,],uplc.pred[nt,])
cor.ns.tmp <- 100*cor(t(tmp.ns.m))
cor.ns.tmp <- as.data.frame(cor.ns.tmp[2,1])
cor.ns.tmp <- cbind(cor.ns.tmp,h.ju)
rownames(cor.ns.tmp) <- rownames(gc.tested[nt,])
nosucces.adj <- rbind(nosucces.adj,cor.ns.tmp)
}

cat("Done","\n")

}
```

Fin de la boucle pour tous les échantillons pour les 100 itérations à chaque h.ju

Calcul de la performance d'ajustement réussi

```
rownames(mydist)<-rownames(uplc.data)
if (.calch) rownames(myheight)<-rownames(uplc.data)
if (.calch) rownames(myheightst)<-rownames(uplc.data)
rownames(mylogical)<-rownames(uplc.data)
pred.pct<-sum(mylogical)/length(mylogical)*100
cat("First step",pred.pct,"% of good prediction","\n")
rownames(mylogicalst)<-rownames(uplc.data)
pred.pctst<-sum(mylogicalst)/length(mylogicalst)*100
cat("Second step",pred.pctst,"% of good prediction","\n")
```

6. Calcul de la performance d'ajustement : Set de Val

Processus similaire à celui appliqué au set de Cal (donc, non présenté ici)

.
.
.
.
.
.
.
.

```
tmp1 <- cbind(pred.pctst,h.ju)
tmp2 <- cbind(pred.pctst2,h.ju)
resultats.finaux_cal<-rbind(resultats.finaux_cal,tmp1)
resultats.finaux_val<-rbind(resultats.finaux_val,tmp2)
```

Avancement de la barre d'avancement

```
setWinProgressBar(pb, h.ju, title=paste("Height = ", h.ju, "/ Itération n°", iter, "sur 100"),"h done")

 }
```

7. Sortie des résultats

```
write.table(resultats.finaux_cal,"results_cal.csv")

}
close(pb)

colnames(succes.adj)[1] <- c("Pearson value")
colnames(nosucces.adj)[1] <- c("Pearson value")
write.table(succes.adj,"Pearson_succes.adj.csv")
write.table(nosucces.adj,"Pearson_nosucces.adj.csv")
```

Calcul de la moyenne et l'écart-type des performances à chaque valeur de h (set de Cal)

```
dup.h<-unique(resultats.finaux_cal[,2])
gc.tmp.h<-vector()
tmp.h<-as.data.frame(matrix(ncol=2,nrow=1))
for (i in dup.h){
w<-which((resultats.finaux_cal[,2]==i)==TRUE)
tmp.h[1,1]<-mean(resultats.finaux_cal[w,1])
tmp.h[1,2]<-sd(resultats.finaux_cal[w,1])
rownames(tmp.h)<-i
gc.tmp.h<-rbind(gc.tmp.h,tmp.h)
}
colnames(gc.tmp.h)<-c("mean","sd")

write.table(gc.tmp.h,"results_cal_mean_sd.csv")
```

Calcul de la valeur médiane et l'écart-type des valeurs de coefficients de corrélation de Pearson pour un ajustement réussi à chaque valeur de h
(set de Cal)

```
dup.h<-unique(succes.adj[,2])
gc.tmp.h<-vector()
tmp.h<-as.data.frame(matrix(ncol=2,nrow=1))
for (i in dup.h){
w<-which((succes.adj[,2]==i)==TRUE)
tmp.h[1,1]<-median(succes.adj[w,1])
tmp.h[1,2]<-sd(succes.adj[w,1])
rownames(tmp.h)<-i
gc.tmp.h<-rbind(gc.tmp.h,tmp.h)
}
colnames(gc.tmp.h)<-c("median","sd")

write.table(gc.tmp.h,"Pearson_succes.adj_median_sd.csv")
```

Calcul de la valeur médiane et l'écart-type des valeurs de coefficients de corrélation de Pearson pour un ajustement NON réussi à chaque valeur de h
(set de Cal)

```
dup.h<-unique(nosucces.adj[,2])
gc.tmp.h<-vector()
tmp.h<-as.data.frame(matrix(ncol=2,nrow=1))
for (i in dup.h){
w<-which((nosucces.adj[,2]==i)==TRUE)
tmp.h[1,1]<-median(nosucces.adj[w,1])
tmp.h[1,2]<-sd(nosucces.adj[w,1])
rownames(tmp.h)<-i
gc.tmp.h<-rbind(gc.tmp.h,tmp.h)
}
colnames(gc.tmp.h)<-c("median","sd")

write.table(gc.tmp.h,"Pearson_nosucces.adj_median_sd.csv")
```

!!! FIN !!!

Série Criminalistique LIX
ISBN 2-940098-63-8