Déterminants et prévision des fluctuations de la concentration en polluants dans un environnement intérieur

Rachid Ouaret

UNIVERSITÉ PARIS-EST - Créteil (UPEC)

ÉCOLE DOCTORALE SIE-UPE
SCIENCES, INGÉNIERIE ET ENVIRONNEMENT

T H È S E

présentée en vue de l'obtention du titre de

Docteur en Sciences de l'Ingénieur

l'Université Paris-Est, Créteil
Spécialité : Sciences et Techniques de l'Environnement
Mention : STATISTIQUE APPLIQUÉE
présentée par

Rachid OUARET

Déterminants et prévision des fluctuations de la concentration en polluants dans un environnement intérieur

Thèse préparée au
Centre d'Études et de Recherche en Thermique, Environnement et Systèmes (CERTES)
et au Centre Scientifique et Techniques du Bâtiment (CSTB), Université Paris-Est
soutenue le 19/07/2016

devant le jury composé de :

Rapporteurs :	M. Francis ALLARD	-	Professeur (LaSIE, Université de La Rochelle)
	M. Gilles ROUSSEL	-	Maître de conférences, HDR (LISIC, Université du Littoral -Côte-d'Opale)
Examinateur :	Mme. Nadine LOCOGE	-	Professeur (École des Mines de Douai)
Examinateur :	M. Christian SEIGNEUR	-	Professeur (CEREA, École des Ponts ParisTech)
Examinateur :	M. Viorel PETREHUS	-	Maître de conférences (Département de Mathématiques UTCB)
Examinateur :	M. Olivier RAMALHO	-	Docteur (CSTB, Université Paris-Est)
Directeur :	M. Yves CANDAU	-	Professeur (CERTES, Université Paris-Est)
Co-Directeur :	Mme. Anda IONESCU	-	Maître de conférences (CERTES, Université Paris-Est)

TABLE DES MATIÈRES

Remerciements

"*Me tenant comme je suis, un pied dans un pays et l'autre en un autre, je trouve ma condition très heureuse, en ce qu'elle est libre*".

René Descartes (Lettre à la princesse Élisabeth de Bohême, Paris 1648).

Cette thèse est le fruit d'un travail de plusieurs années et de rencontres. Elle est l'aboutissement d'un projet tant scientifique que personnel, parfois affectif. Jamais en cavalier solitaire, ces années de thèse sont jalonnées de rencontres décisives, d'attention et de présences réconfortantes. Que soient remerciés ici les acteurs -et ils ne manquent pas ! - de ces rencontres.

Je remercie vivement et de tout cœur mes directeurs [1] de thèse, Yves Candau, Anda Ionescu et Olivier Ramalho. Yves pour son dynamisme scientifique, sa qualité d'écoute dont il a fait preuve ainsi que la grande liberté et la confiance qu'il m'a accordé pendant ces années. Son calme inébranlable devant les difficultés, ont beaucoup contribué à l'aboutissement de ce travail de thèse. Qu'il trouve ici toute ma reconnaissance.

La qualité scientifique de ce mémoire doit beaucoup à l'œil de lynx d'Anda, redoutable (re)lectrice de chaque partition de ma symphonie peu synchrone. Merci à toi Anda pour ta patience, ta gentillesse et tes conseils qui m'ont souvent aidé à prendre de la hauteur. Au cours de cette thèse, tu as toujours su garder en tête la vision globale et cohérente de mes travaux, leurs grandes lignes, leurs orientations, et les éclairer pour moi. Ton œil précieux sur les cordons de la bourse, compréhensive dans les moments difficiles, tu m'a offert le temps, la liberté et les conditions matérielles pour finir sereinement cette thèse.

Je tiens à adresser mes plus sincères remerciements à Olivier, toujours disponible pour moi malgré un emploi du temps parfois acrobatique, il a été présent au quotidien, tous les matins avec une tasse de café [2] (bien sûr) pour suivre l'avancement de mes travaux. Je resterai toujours impressionné par sa rigueur et son sens de la critique. Je tiens à lui exprimer mes remerciements pour ses questions (perpétuelles) et ses remarques. Je le remercie aussi pour son ouverture d'esprit, pour avoir su aussi bien me laisser la bride sur le cou qu'être présent, toujours au moment approprié.

Monsieur Francis Allard et Monsieur Gilles Roussel ont accepté la mission presque impossible de rapporter ce long manuscrit en peu de temps et dans des circonstances un peu chaotiques. Mission réussie, je les en remercie vivement.

Pour les mêmes raisons, je tiens aussi à adresser mes plus vifs remerciements aux membres du jury qui m'ont fait l'honneur d'accepter de juger mon travail. Monsieur Christian Seigneur, Madame Nadine Locoge et Monsieur Viorel Petrehus, je vous suis reconnaissant du temps consacré à l'évaluer.

Merci à Viorel, à la fois pour sa participation comme membre du jury et de m'avoir initié à la théorie des systèmes dynamiques. Sa curiosité insatiable non seulement pour les mathématiques mais aussi pour toutes les autres formes de connaissance. Je ne compte plus les discussions que nous avons eues, à Paris ou à Bucarest, qui portaient sur la théorie de Chaos, l'informatique, l'histoire, la littérature (de Kafka), l'art...la vie.

Qu'il me soit en outre permis de rendre hommage à State Luminita, disparue avant l'achèvement de cette thèse. Son dévouement à la recherche a été et demeure à mes yeux un exemple. J'ai une pensée émue pour elle.

1. La direction au sens élargis du terme.
2. Revenir quelques temps après pour récupérer la tasse et rediscuter sur un autre point...

Je ne voudrais surtout pas oublier d'exprimer ma gratitude à Madame Evelyne Gehin, directrice du CERTES, et à mes collègues pour leur aide et leur sincère amitié. Je remercie tout particulièrement Guillaume Da pour son humanisme pur et le « petit » Nicolas [3] pour le goûter des samedi !

J'aimerais adresser mes plus vifs remerciements à Madame Séverine Kirchner pour l'accueil qu'elle m'a accordé au sein de l'équipe OQAI aux débuts de ma thèse, durant lesquels elle a manifesté un grand intérêt pour mes travaux. Je la remercie chaleureusement pour sa présence à la soutenance. Je tiens à remercier Madame Corinne Mandin, pour son engagement à la recherche QAI qui m'a beaucoup inspiré pour mes travaux, ainsi que pour l'amitié qu'elle n'a pas cessé de me témoigner.

Les années de thèse, années « *dark side on the moon* [4] », c'est LE MOMENT où on nous paye pour devenir un peu plus intelligent, mais aussi le moment dans lequel on a l'opportunité d'apprendre à transmettre un savoir grâce au métier d'enseignant. Je veux donc saluer les enseignants qui m'ont aidé dans mes débuts et fait confiance par la suite, Vincent Feuillet, Said Iammarene et Philipe Bunel. Avec le sens de détail de Vincent, j'apprends à prévoir le temps dans les temps !

J'ai eu la chance de bénéficier de l'environnement extrêmement favorable que constitue l'Open Space au CSTB pour préparer ma thèse. Quand on dit cela, on pense tout de suite aux fameuses viennoiseries de Sharmila, à la gentillesse de Rukshalla et d'Isabelle et aux éclats de rire avec Stéphane (plats aux HAP, Herbes Aromatiques qui rend les gens Polyglotte). Chers amis, je vous en remercie du fond de mon cœur.

Je remercie tous les membres de l'équipe de l'OQAI et de l'équipe "Micro-détection" de faire des déjeuners à la cantine un moment sympathique, je pense à Mickaël pour sa gentillesse, sa sensibilité et son écoute (et c'est vraie ! !) et à Jacques pour m'avoir fait découvrir l'huile d'olive de Nyons (le mien est meilleur ☺). Mes sincères remerciements vont également à Claire, Valérie, Bruno, Doriane, Guillaume, Sébastien et Enric.

Ces remerciements ne seraient pas complets sans mentionner mes cobureaux de CSTB : Geoffrey Sampedro Lopez, Lucille Labat et Sarka Langer. Geoffrey et Lucille ont contribué tant à la bonne ambiance de travail qu'à d'autres moments plus conviviaux, incluant apéros et éclats de rire. Mille merci à Lucille pour sa contribution et son aide précieuse au projet TRIBU.

Je tiens à remercier tous les lecteurs quels qu'ils soient qui prendront le temps de lire un passage (ou l'entièreté) de ce manuscrit. À la Tarantino, j'en profite donc pour faire un peu de publicité. Si vous la lisez vous comprendrez qu'il est possible (en dernière instance) de prévoir la QAI (Chapitres 6-7), mais cela tient beaucoup aux structures de variabilité (prédictibilité) et aux aspects statistiques des observations (Chapitres 2-3). Un retour aux sources est nécessaire pour comprendre un peu plus sur les sources de fluctuations (Chapitre 4). Enfin vous trouverez la beauté du Chaos en caressant les pages de la section 7.4.2. Cette thèse a été écrite en LaTeX 2_ε LaTeX et LyX. Je remercie tous ceux qui y contribuent et partagent. Pour compenser ma consommation de papiers imprimés de ce long manuscrit, une plantation d'arbres a été faite ☺, merci aux arbres.

Je n'aurais sans doute jamais acquis les savoirs, les capacités, l'envie et la passion nécessaires pour mener cette thèse à bien sans de cruciales rencontres, tout au long de ma vie. Je pense à D'el-Hamid (*que Dieu ait son âme*), qui avec son courage, sa sensibilité, nous a fait grandir malgré lui par l'immense épreuve qui l'a emportée et qui nous a à tout jamais secoués, marqués, changés.

Je remercie ma famille et en particulier mes frères et sœurs pour m'avoir fait partager leur joie de vivre et m'avoir ainsi soutenu dans mes efforts. Je les remercie de m'avoir supporté tout au long de ces années, leur support moral m'a aidé à prendre de la hauteur. Hafit pour tes discussions en spirale et intelligentes,

3. Tu sera grand quand tu lira ça !
4. Album de Pink Floyd qui débute avec des battements de cœur réguliers enchainant plusieurs bruits.

Zoubir pour ton style anarchique et Dostoïevskien, Nourdine pour ta sensibilité humaine et ton humilité et Djebar pour ton ouverture d'esprit et ton intelligence. Mille mercis à mes belles sœurs, mes nièces, mes neveux et mes cousins et aux diners passés en famille. Sans eux, j'aurais (peut-être) quelques kilos de moins, mais mon manuscrit aussi.

Mon regard tourne tendrement vers ma fiancée Yasmine, la première lectrice et correctrice de cette thèse. Merci pour tes soirées de thèse qui rayonnaient la cantine d'un autre été. Celles-ci, nous a ouvert et révélé le mystère de l'attachement et du soutien affectif...mon amour pour toi est sans limite !

Mes plus profonds remerciements vont en premier lieu à mes parents pour leurs sacrifices ainsi que pour leur amour inconditionnel avec lequel ils m'ont appris la simplicité de vivre...vivre simplement. Tout au long de mon cursus, ils m'ont toujours soutenu, encouragé et aidé. Ils ont su me donner toutes les chances pour réussir. Qu'ils trouvent, dans la réalisation de ce travail, l'aboutissement de leurs efforts ainsi que l'expression de ma plus affectueuse gratitude.

OUARET Rachid

Février - 2017 à Amizour, Bejaia.

INTRODUCTION GÉNÉRALE

L'environnement intérieur dans lequel nous pouvons parfois passer plus de 80 % de notre temps reste très mal connu. Ce micro-environnement concentre une multitude de matériaux et produits susceptibles d'émettre des substances chimiques et particulaires dans l'air. L'occupant est exposé à ce mélange qui peut avoir des répercussions sur sa santé à long terme. Il est lui-même à l'origine de certaines activités très polluantes (fumée de tabac, cuisine, travaux etc.) qui, malgré leur durée souvent courte, peuvent influencer de manière durable la qualité de l'air intérieur (QAI). L'occupant est également à même d'agir sur les ouvrants, modifiant le renouvellement de l'air et les échanges entre l'extérieur et l'intérieur. Les paramètres climatiques et les propriétés du bâtiment (ou de la seule pièce) interviennent également pour définir la concentration de ces substances dans l'air et son évolution. Cette dernière reste souvent peu documentée de par la difficulté d'avoir des instruments de mesure en temps réel sensibles et qui ne viennent pas perturber l'environnement d'étude. Prévoir cette évolution en condition réelle d'occupation n'est pratiquement jamais abordé dans les études. Il existe bien des modèles permettant de restituer la concentration observée dans des conditions expérimentales ou selon des scénarios prédéfinis, mais aucun ne prend en compte le comportement réel des occupants, moteur des variations dans les environnements intérieurs.

La prévision de la concentration des polluants de l'air intérieur permettrait aux occupants ou aux gestionnaires du bâtiment d'être informés des niveaux auxquels ils sont et seront exposés, de vérifier qu'elle ne dépasse pas les valeurs limites établies par les instances sanitaires ou réglementaires (code de l'environnement, OMS, ANSES, HCSP [5]), d'anticiper leur comportement pour contrer ou limiter la hausse prévue de la concentration.

Plusieurs solutions sont disponibles pour développer un modèle de prévision. Mais, l'absence d'inventaire d'émissions des innombrables matériaux et produits utilisés en intérieur rend délicate, voire impossible l'utilisation de modèles classiques déterministes à l'heure actuelle. La mise en place de l'étiquetage obligatoire de certaines caractéristiques d'émissions pour les produits de construction et de décoration ou la base de données PANDORE [6] sont des démarches prometteuses mais encore insuffisantes pour

5. OMS : Organisation mondiale de la santé ; ANSES : Agence nationale de sécurité sanitaire de l'alimentation, de l'environnement et du travail ; HCSP : Haut conseil en santé publique.

6. PANDORE : une comPilAtion des émissioNs Des pOlluants de l'aiR intériEur. Abadie & Blondeau (2011). « PANDORA database : A compilation of indoor air pollutant emissions ». HVAC&R Research 17 (4) : 602-613. (http ://lasie.univ-larochelle.fr/PANDORE-une-comPilAtion-des)

combler ce déficit. Par ailleurs, les paramètres climatiques variables dans le temps peuvent modifier ces caractéristiques d'émissions selon des lois encore méconnues.

Le choix s'est donc porté sur l'analyse de l'évolution temporelle des mesures de concentration préalablement observées dont la structure très peu étudiée renfermerait suffisamment d'information pour prévoir avec un minimum d'erreur les concentrations futures. Cette analyse permet également d'étudier les sources des fluctuations de la concentration de polluants pour pouvoir les identifier et déterminer leur contribution. Cette connaissance peut d'ailleurs aider à améliorer les performances d'un modèle de prévision.

L'objectif de la thèse est d'étudier la structure temporelle de la concentration de deux polluants pris comme exemples (le formaldéhyde, un composé organique volatil et les particules en suspension dans l'air) pour identifier et comprendre les sources de fluctuations et pour adapter un modèle de prévision de leur concentration. A cet égard, différents jeux de données ont été utilisés provenant essentiellement d'environnement de bureaux mais également d'une maison expérimentale.

La thèse est organisée en deux parties. Une première partie, organisée en trois chapitres, recadre l'analyse des structures temporelles spécifiques aux chroniques de la qualité de l'air dans un environnement réel, en s'interrogeant sur les propriétés statistiques des séries temporelles de mesure de la QAI. La deuxième partie, structurée en quatre autres chapitres, concerne l'exploitation de l'information véhiculée dans les données afin de répondre à la problématique d'identification des sources et de prévision.

La **première partie** débute par une introduction générale à la problématique de la qualité de l'air intérieur et des phénomènes qui la régissent (**chapitre 1**).

Le **chapitre 2** commence par la présentation des environnements intérieurs. Les données disponibles pour chacun des environnements sont ensuite décrites et l'influence des facteurs climatiques d'occupation et de gestion des ouvrants mise en avant. La présence de valeurs manquantes nécessite la mise en place de méthodes d'imputation qui sont détaillées dans l'annexe. Cette gestion est nécessaire avant toute analyse spectrale et prévision statistique.

La présentation des fondements théoriques de l'analyse des séries temporelles appliquée aux données de la QAI vient clore cette partie et fait l'objet du **chapitre 3**. Elle permet de s'interroger sur les structures intrinsèques à la variabilité temporelle des mesures de polluants issues de la QAI en abordant quatre points principaux : *(i)* l'analyse de la variabilité par l'analyse spectrale classique, *(ii)* l'analyse des structures temporelles par la quantification des propriétés de dépendance, *(iii)* une réflexion sur la notion de prédictibilité des séries temporelles et enfin *(iv)* la décomposition en composantes latentes de la variabilité.

En l'absence de toute étude traitant ces points pour la QAI, il nous a paru nécessaire de tester plusieurs méthodes, afin de pouvoir déboucher une piste de travail, tant théorique que pratique, concernant la prévision par décomposition des séries chronologique. La première contribution originale de cette thèse réside justement dans ce chapitre (3).

La deuxième partie de la thèse s'inscrit directement dans la continuité du travail engagé pour l'analyse des structures de variabilité. Il s'agit tout d'abord d'améliorer nos connaissances sur les fluctuations des sources de concentration ainsi que leurs contributions grâce aux méthodes de séparations aveugle des sources (NMF, ACI, PMF) [7]. Ces méthodes ont été testées dans le cadre de séparation des sources des polluants dans les environnements intérieurs étudiés, dans le cadre du **quatrième chapitre**.

7. NMF : Non-negative matrix factorization, ACI : analyse par composantes indépendantes, PMF : Positive matrix factorization.

Les modèles de prévision sont ensuite discutés notamment au regard des données de haute fréquence, en mettant en évidence la nécessité de développer des méthodes spécifiques aux fluctuations des mesures issues de la QAI. Le **chapitre 5** expose plus précisément le positionnement de nos choix de modèle dans le un contexte plus général des modèles de prévision.

L'approche développée dans les deux chapitres suivants consiste en l'exploitation des différentes composantes de décomposition des séries pour la prévision. Dans le **chapitre 6**, trois méthodes de prévision par décomposition en composantes latentes sont appliquées aux différentes séries temporelles de polluants. La caractérisation préalable des fluctuations développée dans la première partie de la thèse fournit les premiers jalons pour identifier et tester les modèles stochastiques susceptibles de les reproduire.

Dans le **chapitre 7**, on propose une modélisation hybride s'articulant sur la nécessité de prendre en compte les structures temporelles par bandes spectrales, et les modéliser soit par un modèle à changement de régime (STAR, SETAR, MS-TAR) ou à l'aide de la dynamique du chaos. Cette deuxième partie se conclut par une proposition appelée SBD-(TAR/Chaos) [8], s'articulant sur la nécessité de prendre en compte les structures temporelles par bandes spectrales, et de les modéliser soit avec un modèle à changement de régime ou à l'aide de la dynamique du chaos.

La conclusion rappelle les multiples contributions théoriques introduites dans ce travail de cette thèse et présente plusieurs perspectives de ces travaux.

Désireux de faire un document relativement complet et didactique au regard des efforts déployés pour accumuler ces connaissances, nous avons volontairement conservé un grand nombre d'informations pouvant paraître superflues au lecteur averti. Nous nous sommes toutefois efforcé d'en faciliter la lecture et espérons que le lecteur en appréciera le contenu.

8. Spectral Band Decomposition – (Threshold autoregressive model / Chaos model).

Première partie

DE LA NATURE DES FLUCTUATIONS AUX CONSÉQUENCES DE LEURS TRAITEMENTS STATISTIQUES

What do you see as the greatest challenges facing the profession of statisticians in the coming years ?

Big data, data collected as history, in networks, data collected with many alternative measurements, data collected as photographs and pictures. As always, trying to find the signals in the noise. An interview[9] *with* J. Stuart Hunter *(Professor of Civil and Environnemental Engineering at Princeton)*

Courte introduction de la partie

Quelles sont les propriétés statistiques des séries temporelles de mesure de la Qualité de l'Air Intérieur (QAI) ? Comment les exploiter ? Sont-elles utiles pour le choix d'un modèle de prévision ? Ces questions nous intriguent et sont au cœur de la thèse. D'un côté, comprendre les structures inhérentes aux chroniques de la QAI pourrait nous aider à comprendre les processus générateurs de ces données. D'un autre coté, mesurer le degré de prévisibilité des séries temporelles, au moins d'une certaine manière, pourrait permettre d'appréhender le niveau de difficulté lié à la prévision de certains polluants.

Ces questions sont étudiées tout au long de cette partie. En ligne de mire, on montrera la complexité de la modélisation d'un environnement réel, malgré le dispositif de mesure mis en place. À l'heure actuelle, deux visions s'offrent à nous, une modélisation physique (déterministe) ou une modélisation statistique. Ces deux approches utilisent des sources d'informations et des connaissances qui leurs sont propres. Nous souscrivons pour les différents objectifs de la thèse à l'approche statistique qui permet, d'un point de vue environnemental, d'apporter une nouvelle vision pour appréhender l'analyse de la qualité de l'air intérieur. Les outils statistiques appliqués ou développés constituent dès lors les éléments de ce que nous appelons par la suite, *"l'environnemetrie intérieure"*. L'appellation "*environnemetrie*" est utilisée par Philip Cox en 1972 pour décrire l'ensemble des applications de la statistique en sciences environnementales (Hunter, 1994). Pour nous, il s'agit de l'environnement intérieur et plus précisément de l'air intérieur.

Au cours de cette thèse, on se propose comme objectif non seulement de comprendre la variabilité temporelle inhérente aux séries temporelles issues des mesures de la concentration en polluants dans un environnement intérieur, mais aussi de comprendre les interactions entre les différents paramètres.

Il a fallu pour cela commencer par choisir et appliquer les différents outils disponibles de l'analyse des séries temporelles. Cette étape a été l'occasion de visionner des structures qui, d'un point de vue prévision nous orientent sur des classes de modèles spécifiques ou, au moins, d'éliminer certains, comme les modèles linéaires des séries temporelles, qui ne sont pas adaptés à ces structures.

9. Dans "*Statistics Views*" consulté le 06/04/2016 sur `http://www.statisticsviews.com`.

CHAPITRE 1

QUALITÉ DE L'AIR INTÉRIEUR (QAI) : GÉNÉRALITÉS ET MODÉLISATION

Considéré depuis longtemps comme un abri et une protection de la pollution extérieure, l'environnement intérieur n'est pas exempt de la dégradation de la qualité de l'air pour ses occupants. La qualité de l'air intérieur est aujourd'hui une question de santé publique. Dans ce chapitre sont présentées quelques généralités sur la "Qualité de l'Air Intérieur" (QAI). Tout d'abord, quelques définitions propres à la QAI sont énumérées. Ensuite, les principaux polluants de l'air intérieur et leurs sources sont présentés. On finit par montrer l'intérêt de l'étude de la qualité de l'air intérieur et pourquoi on s'y intéresse, en mettant l'accent sur l'impact sanitaire.

Sommaire

1.1 Introduction

Alors que les niveaux de pollution de l'air extérieur sont aujourd'hui mesurés en continu par les réseaux de surveillance de la qualité de l'air et que leurs effets sanitaires sont relativement bien documentés, on s'interroge depuis peu sur les niveaux et sur l'impact sanitaire des contaminants présents dans les espaces confinés.

En effet, jusque vers le milieu des années 1970, la qualité de l'air intérieur n'a pas été l'objet de grandes préoccupations (Stolwijk, 1992) ; les débats durant cette période se sont orientés vers des questions énergétiques, suite aux chocs pétroliers des années 70. Les études se sont alors focalisées sur les problématiques relevant de l'isolation thermique en vue d'optimiser les performances énergétiques, ce qui, à terme, a eu pour conséquence non seulement d'augmenter le confinement, mais aussi de changer le comportement des occupants.

Tous ces bouleversements au niveau de la conception des bâtiments a soulevé d'autres problématiques liées à la dégradation de la qualité de l'air intérieur, qui peut avoir des effets sur la santé des individus, pouvant se manifester par toute une série de symptômes. C'est à partir de ces manifestations sanitaires que la notion de la qualité l'air intérieur a commencé à être largement débattue par la communauté scientifique, allant des études sur le confinement et sur la ventilation aux études sur les relations de cause à effet entre la mauvaise qualité de l'air et la santé des occupants.

Historiquement, les problèmes liés à la pollution de l'air intérieur étaient incontestablement beaucoup plus apparents que ce qu'ils sont aujourd'hui, la suie trouvée sur les plafonds de grottes préhistoriques fournit de nombreuses preuves des niveaux élevés de pollution suite à une ventilation inadéquate ou à des feux et incendies (Spengler & Sexton, 1983).

Depuis quelques décennies, les relations entre la santé publique et la pollution dans les espaces clos se sont précisées (Hoskins, 2003; Jones, 1999), et le rôle important des émissions des différentes sources a été mis en évidence (Nazaroff & Weschler, 2004; Nazaroff & Singer, 2004; Maroni et al., 1995; Lee et al., 2001; Wallace et al., 2004). Les polluants de l'air intérieur peuvent en effet avoir des effets variés sur la santé des individus, selon les polluants rencontrés : irritations de la peau, nausées, céphalées, pathologies respiratoires, neurologiques, développement de certains cancers etc. (Catelinois et al., 2006; Fisk et al., 2010; Jones, 1999; Sundell, 2004).

1.2 La qualité de l'air intérieur

1.2.1 Définitions

Un environnement intérieur est un volume plus ou moins couvert et plutôt séparé de l'extérieur. On applique généralement le terme d'air intérieur à ces espaces non industriels tels que ceux que l'on trouve dans les habitations privées, les établissements publiques, les immeubles de bureaux, mais aussi les modes de transport, les gares ferroviaires ou aéroportuaires, etc.

La qualité de l'air intérieur (QAI) représente la composition ou l'état de l'air à un instant donné ; elle est jugée bonne ou mauvaise, acceptable ou non. Elle est donc variable selon la perception des individus. En effet, soumis aux mêmes conditions environnementales, les occupants peuvent réagir et interagir avec les différents systèmes du bâti de manière différente. Selon l'ASHRAE (American Society of Heating, Refrigerating and Air-Conditioning Engineers), elle est jugée acceptable lorsque cet air ne comporte

pas de polluants à des concentrations jugées à risque, telles qu'elles ont été fixées par les autorités compétentes, et lorsqu'au moins 80 % des personnes exposées n'expriment pas de mécontentement.

De façon générale, le terme "qualité de l'air intérieur" est caractérisé par la présence de polluants à des niveaux susceptibles d'affecter la santé des occupants, leur confort, leur performance ou l'environnement lui même. Elle dépend de toute une série de variables parmi lesquelles on peut citer : la pollution extérieure, les sources intérieures, les conditions climatiques extérieures et intérieures, la ventilation et la présence des occupants. La contamination intérieure a plusieurs origines : les occupants eux-mêmes et leurs activités, les matériaux de construction, l'accumulation de polluants dans des espaces mal ventilés, la combustion etc. La contamination véhiculée par l'air provenant de l'extérieur s'introduit dans les espaces intérieurs du fait du fonctionnement d'un système de ventilation, par des ouvrants (fenêtres et portes) ou par infiltration à travers les interstices de l'enveloppe du bâtiment.

1.2.2 Spécificités des environnements intérieurs

Le Tableau 1.2.1 compare quelques principales caractéristiques des environnements intérieurs et extérieurs.

TABLE 1.2.1 – Principales spécifications d'un environnement intérieur. D'après Nazaroff et al. (2003).

Paramètre	Atmosphère urbaine	Ambiance intérieure
Ratio surface/volume	$\sim 0.01\,\mathrm{m^2 m^{-3}}$	$\sim 3\,\mathrm{m^2.m^{-3}}$
Précipitation	10–150 cm/an	Absent
Rayonnement/ énergie	$\sim 1000\,\mathrm{Wm^{-2}}$(par jour)	$\sim 1\,\mathrm{Wm^{-2}}$

Les environnements intérieurs sont caractérisés par différents paramètres spécifiques par rapport à l'extérieur :

- *La densité d'occupation* : la présence et le nombre de personnes par surface ou volume. Elle est variable selon les différents espaces clos (habitations, écoles, bureaux, lieux de loisir, etc.) et plus importante à l'intérieur qu'à l'extérieur.
  Une forte densité d'occupation modifie l'ambiance thermique de l'environnement, et le confinement de l'air, ce qui implique des besoins en aération ou en climatisation.

- *La fréquence d'occupation* : les données relatives au temps passé et aux activités menées par les individus, appelés Budget Espace-Temps-Activités (**BETA**) s'accordent sur le fait que **les citadins passent plus de 80% de leurs temps à l'intérieur des espaces clos** (Mandin et al. (2009); Klepeis et al. (2001); Leech et al. (1996)).
  Une étude menée aux États-Unis montre qu'en moyenne, un individu dépense 88% de sa journée à l'intérieur des bâtiments (habitats, bureaux, écoles), 7% dans un véhicule et seulement 5% à l'extérieur (Jones, 1999; Robinson & Nelson, 1995). En France, la campagne représentative au niveau national menée par l'Observatoire de la qualité de l'air intérieur (OQAI) a montré que le temps moyen passé dans son logement est de 16h10 par jour et pour 25% de la population, il est supérieur à 20h (Zeghnoun et al., 2010).

- *Le rapport surface-volume* : les environnements intérieurs sont caractérisés par de nombreuses surfaces disponibles au regard de leur volume restreint. Ces surfaces représentent autant de possibilités d'interactions avec les substances et les particules présentes dans l'air. Ce ratio surface-volume (S/V) varie selon :
  — les dimensions de la pièce ;
  — la proportion de surfaces recouvertes par des produits de construction ou de décoration ;
  — le mobilier présent ;
  — le nombre d'occupants et leur surface corporelle ;
  — les particules en suspension dans l'air.
  Une petite pièce aura un ratio S/V supérieur à une pièce plus grande. De façon générale, le rapport surface-volume dans les locaux est $\geq 2\,\mathrm{m^2 m^{-3}}$(voire $\geq 3\,\mathrm{m^2 m^{-3}}$dans les locaux fortement meublés). Ce rapport est estimé à environ $0,01\,\mathrm{m^2 m^{-3}}$ en air urbain (Nazaroff et al., 2003). Il met en évidence le rôle particulièrement important des surfaces comme sources et puits des polluants de l'air intérieur, leur rôle réservoir de composés organiques semi-volatils et leur rôle dans la réactivité chimique de l'air intérieur.

- *La présence de polluants spécifiques :* la composition de l'air intérieur peut être différente de la composition de l'air extérieur de par la nature des sources mises en oeuvre. Mais surtout, certains contaminants (principalement des composés organiques volatils) se retrouvent à des teneurs plus importantes à l'intérieur comme par exemple le formaldéhyde avec une médiane nationale de $19.6\,\mu\mathrm{g}\cdot\mathrm{m}^{-3}$ à l'intérieur contre $1.9\,\mu\mathrm{g}\cdot\mathrm{m}^{-3}$ à l'extérieur (Kirchner et al., 2007a,b)).

- *Le rayonnement ultraviolet (UV)* : l'atténuation des rayons UV dans un environnement intérieur est beaucoup plus importante par rapport à l'extérieur lorsque les fenêtres sont fermées. L'absorption des UV est variable selon la nature du verre et le type de vitrage, mais elle est de l'ordre de 90 %. Les sources lumineuses intérieures n'émettent que très peu ou pas du tout dans le spectre ultraviolet. En conséquence, la photolyse des substances dans l'air est négligeable à l'intérieur par rapport à l'extérieur. Ce qui explique que certaines substances peuvent plus facilement s'y accumuler comme le dioxyde d'azote ou le formaldéhyde. En situation de fenêtre ouverte, les conditions de rayonnement tendent à se rapprocher des conditions extérieures. L'énergie de rayonnement pénétrant dans l'environnement intérieur est généralement de l'ordre de $\sim 1\,W/m^2(jour)$.

- *Les paramètres climatiques :* l'absence de précipitation et l'amplitude de variation de la température et de l'humidité généralement plus faible dans les environnements intérieurs conduit à des variations relatives des concentrations dans l'air plus faibles qu'à l'extérieur.

1.3 La pollution de l'air intérieur : sources et polluants

La pollution de l'air dans les environnements intérieurs est un phénomène dynamique caractérisé par une variabilité d'émissions de polluants de diverses sources (Seifert & Ullrich, 1987). Celles-ci peuvent être classées en deux grandes catégories :

- *les sources d'émission continues* : elles dépendent généralement des conditions environnementales telles que la température, la vitesse de l'air, et l'humidité relative. Elles dépendent également du comportement des occupants, dont l'action sur les ouvrants va venir modifier les conditions environnementales. Les sources d'émission en continu varient à l'échelle du temps d'un jour,

d'une semaine, ou plus.

- *les sources d'émission intermittentes* : elles dépendent de la présence d'une source ponctuelle comme par exemple la fumée de tabac, la combustion d'un bâtonnet d'encens, l'utilisation d'un produit ménager, etc.) . Ce type d'émission évolue beaucoup plus rapidement dans le temps et peut changer en moins d'une heure, voire en quelques minutes. Généralement, c'est durant ces courtes périodes qu'on observe les teneurs extrêmes en polluants.

Une typologie d'environnements peut être spécifiée par rapport à ces caractéristiques. Ainsi, un immeuble de bureaux peut avoir des sources spécifiques qu'on ne peut trouver dans une résidence privée, et inversement les sources et les polluants d'un immeuble résidentiel sont spécifiques. Par exemple, un bureau est en général caractérisé par l'absence de processus de combustion. En revanche l'activité des photocopieurs et des imprimantes dans ce dernier favorise l'émergence d'un certain type de polluants comme les particules en suspension à des concentrations qui rendent ces environnements spécifiques.

1.3.1 Sources de pollution intérieure

La pollution de l'air intérieur est due à l'interaction complexe de nombreux composés présents à des niveaux très différents selon les lieux et leurs sources d'émission. Chaque type de polluant dépend de nombreuses sources et chaque source peut générer plusieurs polluants.

La pollution de l'air extérieur, transférée à l'intérieur par la ventilation plus la présence de sources intérieures de pollution spécifique liées aux équipements (appareils de chauffage et de combustion, produits de construction, mobilier, etc.) et aux activités humaines (tabagisme, cuisine, bricolage, etc.) vont conditionner la qualité de l'air intérieur.

La concentration en polluant dépend en général de la relation entre le volume d'air contenu dans l'espace clos, le taux de production (ou d'émission) du polluant, son taux d'élimination par réaction ou dépôt sur les surfaces, sa concentration extérieure et les paramètres de transfert de l'air (débit d'air échangé avec l'extérieur) (Maroni et al., 1995). Par ailleurs, les émissions par les matériaux présents dans l'environnement intérieur dépendent aussi de leur âge, des paramètres climatiques et de leur caractéristiques physico-chimiques (porosité, etc.).

La fonction de dilution du renouvellement de l'air dépend des concentrations de chacun des compartiments extérieur et intérieur, et de la stratégie de ventilation . L'impact du taux de renouvellement d'air sur la qualité de l'air intérieur traduit en fait l'ambivalence de sa fonction à la fois source de pollution de l'extérieur vers l'environnement intérieur et puits significatif des polluants de l'air intérieur (Ramalho, 2004).

1.3.2 Sources spécifiques dans les environnements de bureaux

Le bureau est l'endroit dans lequel une partie importante de la population active passe le plus de temps par jour, environ 35 h de présence par semaine. Dans sa récente analyse bibliographique, Wolkoff (2013) recense les travaux menés dans les espaces de bureaux visant deux objectifs : l'impact sanitaire et les performances des occupants. Dans sa conclusion, l'impact des polluants intérieurs dans les bureaux n'est pas à négliger.

Bien que de nombreuses études de la littérature internationale aient rapporté la présence de certains polluants (PM, COV, bio-contaminants etc) aussi bien dans les espaces d'habitats que dans les bureaux,

quelques différences sont quand même à mettre en avant, notamment l'absence des sources de combustion à l'intérieur de ces habitats, qui sont à l'origine de la présence de monoxyde de carbone ou d'oxydes d'azote. En outre, en raison de l'interdiction de fumer dans les lieux publics depuis le 1^{er} février 2007, la présence de certains contaminants est désormais d'importance mineure, voire négligeable dans les espaces de bureaux.

En revanche, d'autres polluants pourraient exister en quantité suffisante pour caractériser ces ambiances, en particulier les COV, l'ozone, les particules (poussière de toner) et le formaldéhyde (Saraga et al., 2011; Salthammer et al., 2010). En effet, ces substances peuvent être libérées dans l'air des locaux par les imprimantes lasers, les photocopieurs et les ordinateurs (Wolkoff et al., 1993; Destaillats et al., 2008; Wensing et al., 2006; Schripp et al., 2009). Récemment, plusieurs travaux ont cherché à quantifier les émissions des imprimantes-photocopieurs en particules, notamment les ultra-fines, et les COVs en général (Kagi et al., 2007; Lee & Hsu, 2007; Lee et al., 2001) Les émissions des particules fines par les imprimantes laser et les photocopieurs dépendent de différents paramètres : l'âge et le modèle de l'imprimante ou de photocopieurs utilisés, l'âge et la charge du toner (Wensing et al., 2006; Lee & Hsu, 2007; Uhde et al., 2006). Plus particulièrement, c'est la mise en service des machines qui serait à l'origine des pics de pollution particulaire dans les environnements type bureau. Autrement dit, le nombre d'impressions et la mise en fonction du copieur est le facteur à prendre en considération pour évaluer l'exposition des occupants à la pollution intérieure dans ce type d'ambiance.

Par ailleurs, l'usage de l'aspirateur est également fréquent dans les bureaux. Ce sont, en partie les produits d'entretien qui génèrent des COV qui seraient à l'origine des *symptômes du syndrome de l'habitat malsain* (Wolkoff, 2013; Wolkoff et al., 2006).

Une revue de littérature sur les polluants émis par les différentes sources spécifiques à l'environnement type bureau a été réalisée par Destaillats et al. (2008).

1.3.3 Types de polluants

Les polluants émis dans les différents environnements intérieurs sont très nombreux et très variés ; on peut les classer en fonction de leurs caractéristiques en :

1. Polluants biologiques (moisissures, bactéries, virus...) ;
2. Polluants physiques (particules fines, amiante, radon, l'humidité, les champs électromagnétiques ...) ;
3. Polluants chimiques (monoxyde de carbone, oxydes d'azote, ozone, métaux lourds, le formaldéhyde et les COV...).

1.3.3.1 L'ozone (O_3)

L'ozone est présent dans la troposphère comme polluant secondaire. Sous l'effet du rayonnement solaire, les oxydes d'azote, provenant de l'oxydation de l'azote de l'air lors de la combustion du carburant, peuvent réagir avec des composés issus du trafic automobile, des industries, et conduire à la formation de l'ozone. La quantité d'ozone présente dans la troposphère est donc un indicateur d'une pollution importante de l'air ambiant (Finlayson-Pitts & Pitts Jr, 1999). En outre, certains équipements tels que les imprimantes laser ou photocopieurs (lors du fonctionnement) peuvent émettre de l'ozone (He et al., 2010; Morawska et al., 2009; Destaillats et al., 2008; Wensing et al., 2006). A l'extérieur, les pics de pollution à l'ozone interviennent principalement en période estivale et plus particulièrement

en milieu d'après-midi, lorsque les conditions climatiques sont les plus favorables (température élevée, fort rayonnement UV, durée d'insolation importante, vent faible et présence de polluants primaires) (Finlayson-Pitts & Pitts Jr, 1999). Les valeurs guides de l'OMS pour l'ozone sont de 0.076 ppm pour une heure d'exposition et de 0.05 ppm pour 8 heures.

1.3.3.2 Le monoxyde de carbone (CO)

Le monoxyde de carbone est un gaz inodore et incolore, hautement toxique même à des faibles concentrations (Austin et al., 2002). Il provient de la combustion incomplète des combustibles et des carburants, notamment du gaz naturel, des dérivés pétroliers ou du bois (la combustion de tout produit carboné). Dans l'air intérieur, les systèmes de chauffage mal entretenus, les cuisinières à gaz sont les appareils le plus souvent mis en cause comme sources de CO. Les valeurs guides de l'OMS sont de 5 ppm pour 24 heures d'exposition, de 10 ppm pour une heure, et de 90 ppm pour 15 minutes d'exposition. Plusieurs études ont montré qu'à des concentrations suffisantes pour entraîner une concentration de carboxyhémoglobine (molécule d'hémoglobine associée au CO) supérieure à 2 à 3 %, le CO est susceptible de provoquer des effets négatifs sur la santé des malades cardiaques (Brook et al., 2004).

1.3.3.3 Le dioxyde de carbone (CO_2)

Dans les espaces clos, la principale source de CO_2 avec le métabolisme est la combustion. Le métabolisme humain produit du CO_2, qui est libéré dans l'air lors de l'expiration ; la concentration dépend du nombre de personnes présentes, de leurs activités physiques et de la ventilation de l'espace occupé. De façon générale, une personne produit environ $15\,\ell$/h de gaz carbonique au repos, et entre $20 - 40\,\ell$/h en activité. Dans un bureau, cette production est estimée à $20\,\ell$/h (Roulet, 2004). La concentration en CO_2 est utilisée en tant qu'indicateur de confinement et par conséquent, c'est un bon marqueur des bio-effluents. À des concentrations très élevées, le CO_2 peut avoir des effets physiologiques indésirables au niveau des systèmes nerveux central, cardiovasculaire et respiratoire (Institute of Medicine, 2011).

1.3.3.4 Les oxydes d'azote (NO_x)

Il existe diverses variétés d'oxydes d'azote (NO_2, NO, N_2O) parmi lesquels le dioxyde d'azote NO_2 est le plus répandu dans les études de la pollution intérieure (Hänninen et al., 2004; Maroni et al., 1995). Il sont émis lors de la combustion (appareils de chauffage ou de production d'eau chaude, la fumée de tabac, ou par transfert de l'extérieur de la pollution automobile). Ainsi, par exemple, le taux de NO_2 dans une cuisine où fonctionne une cuisinière à gaz peut être 8 à 10 fois plus important qu'à l'extérieur avec des pics supérieures à $1000\,\mu g \cdot m^{-3}$ (Grimaldi & Déoux, 2003). Le NO_2 est un irritant pulmonaire. Le bureau régional pour l'Europe de l'OMS suggère une limite de 200 $\mu g/m^3$(0.11 ppm) pour une heure, de 120 $\mu g/m^3$(0.06 ppm) pour huit heure et au maximum $40\,\mu g/m^3$ pour une exposition annuelle (WHO-Europe, 2000).

1.3.3.5 Les composés organiques volatiles (COVs)

Les COVs correspondent à plusieurs familles chimiques : alcènes, alcanes, aldéhydes, cétones, esters, alcools, etc. Les COVs sont émis par diverses sources : matériaux de construction, colles, nettoyants, produits domestiques, désodorisants, photocopieurs, solvants (Fenech et al., 2010; Destaillats et al.,

2008). Certaines activités des occupants comme le tabagisme, le bricolage, la combustion sont aussi des sources de COVs. Les sources permanentes dominantes sont issues des matériaux de construction et d'isolation (Edwards et al., 2001) ; parmi elles, les aldéhydes, et plus particulièrement le formaldéhyde et l'acétaldéhyde sont souvent majoritaires (Liu et al., 2006). En général, la concentration des COVs est inférieure à $1\,\mathrm{mg}\cdot\mathrm{m}^{-3}$ (Lévesque et al., 2003).

1.3.3.6 Le formaldéhyde (HCHO)

Le formaldéhyde appartient à la famille des COVs et c'est un gaz incolore présentant une odeur caractéristique et qui irrite les voies aériennes supérieures.

Depuis quelques années, la recherche sur l'exposition au formaldéhyde a bénéficié d'une attention considérable dans le domaine de la QAI et ce pour quatre principales raisons (Wolkoff & Nielsen, 2010) : (*i*) la classification de l'IARC (2006) (Centre international de recherche contre le cancer) en tant que substance cancérigène ; (*ii*) les travaux comme ceux menés par Nazaroff & Weschler (2004) et Carslaw (2007) sur les réactions entre l'ozone et les monoterpenes qui forment du formaldéhyde ; (*iii*) les études épidémiologiques sur les effets de l'exposition au HCHO pour les problèmes pulmonaires et (*iv*) les études sur l'exposition des personnes vulnérables : les enfants et les personnes agées. L'OMS recommande une valeur guide de $0.1\,\mathrm{mg}\cdot\mathrm{m}^{-3}$ (0.08 ppm) pour protéger la population contre les effets d'irritation. En France, un décret fixe une valeur guide d'air intérieur réglementaire pour le formaldéhyde à $30\,\mu\mathrm{g}\cdot\mathrm{m}^{-3}$ en moyenne annuelle (JORF, 2011) ;

Les sources de formaldéhyde

Les sources de formaldéhyde dans les espaces clos sont nombreuses et variées car le formaldéhyde est contenu (sous différentes formes) dans plusieurs produits de consommation courante (papier, cosmétiques, détergents, meubles, tapis, bois aggloméré etc.)(Salthammer et al., 2010).

Le HCHO est aussi présent dans l'environnement suite à des processus naturels (produit d'oxydation photochimique, processus de combustion, dans une moindre mesure produit du métabolisme). Il est également un produit chimique industriel majeur et il est largement utilisé pour la production des résines et d'engrais, l'industrie du latex, du traitement des textiles (Elsner et al., 2003; Zhong, 2013), pour la fabrication des colorants, et dans le traitement des fluides pour embaumement (Spengler et al., 2001).

Très répandues dans les années 70 pour l'isolation des bâtiments, les mousses à base d'urée-formol ont été mises en cause pour leurs émissions en formaldéhyde. Bien que cette source ait été interdite par la suite, elle correspond à des sources à émission continue. De façon générale, les principales sources dans un environnement non-fumeur, semblent être les matériaux de construction et les produits de consommation qui émettent du formaldéhyde (Haghighat & De Bellis, 1998; Hodgson et al., 2002; Salthammer et al., 2010). Aussi, les produits dérivés en bois (panneaux contreplaqués, les lamelles minces, les panneaux de fibres de bois, etc) sont souvent cités comme sources potentielles de formaldéhyde (Dassonville et al., 2009; Spengler et al., 2001). Néanmoins, les émissions de ces matériaux tend à diminuer avec leur vieillissement (Grimaldi & Déoux, 2003).

Quant aux activités ponctuelles, elles sont essentiellement liées aux comportement des individus et à leur mode de vie. Les taux d'émission varient alors en fonction des activités, les plus importantes étant la combustion, les cuissons des aliments, le tabagisme et l'utilisation de l'encens. Récemment, Jensen et al. (2015) ont mis en cause la cigarette électronique pour sa potentielle émission de formaldéhyde.

Le formaldéhyde peut également résulter d'un processus physico-chimique (phase gaz entre l'ozone et les alcanes) qui se produit lors de l'utilisation des produits de consommation, tels que les désodorisants, solvants etc. (Atkinson & Arey, 2003; Wisthaler et al., 2005). La Figure 1.3.1 montre les processus physico-chimiques qui interagissent dans la détermination de la concentration du formaldéhyde dans l'air intérieur.

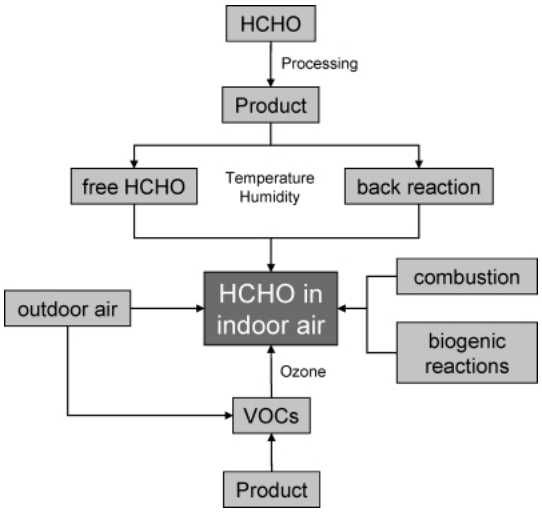


FIGURE 1.3.1 – *Interactions (possibles) des sources et des facteurs de la concentration de HCHO dans l'air intérieur, d'après Salthammer et al. (2010).*

Dans l'étude menée par Funaki et al. (2003), les chercheurs montrent que les équipements électriques ou électroniques composés de matériaux plastiques sont en général des sources de COVs dont le formaldéhyde. Ainsi, lors de leur fonctionnement, le taux d'émission en HCHO par les ordinateurs est environ $9\,\mu g/(\#\ \text{unités} \cdot \text{h})$, soit 9 fois plus que lorsqu'ils sont éteints. Hormis le formaldéhyde qui a une décroissance très faible au cours du temps, les émissions de la plupart des COV diminuent de 5 à 20 % après 4 mois d'utilisation (Malmgren-Hansen et al., 2011).

Par ailleurs, les sources majeurs provenant de l'extérieur sont principalement liées au trafic routier. Les émissions des moteurs à combustion sont variables selon la température, le type de moteur, la composition du carburant ou encore l'âge du véhicule (Alzueta & Glarborg, 2003; Guo, 2011).

Les concentrations de HCHO

En France, les logements présentent en moyenne une concentration en HCHO plus 10 fois plus importante dans l'air intérieur que dans l'air extérieur (Kirchner et al., 2007b).

L'un des premiers articles ayant traité la modélisation des émissions de formaldéhyde en fonction des paramètres climatiques a été publié par ANDERSEN et ses collaborateurs (1975). Selon cette étude, une décroissance hyperbolique des niveaux de formaldéhyde est observée avec une augmentation du taux de renouvellement d'air. Un modèle empirique a été alors ajusté aux mesures observées de formaldéhyde ; ce modèle est basé sur trois sous-systèmes : l'environnement extérieur, l'air ambiant intérieur et un panneau de particules. L'étude suppose néanmoins que l'air extérieur ne contient pas du formaldéhyde. D'autres modèles ont été développés par la suite en se basant essentiellement sur une modélisation

physique et sur des conditions stationnaires (Triebig et al., 1989). Ainsi, dans l'étude de Berge et al. (1980), les auteurs supposent l'absence d'échanges d'air dans la chambre de simulation et reformulèrent les équations d'Andersen et al. (1975). Globalement, les émissions de formaldéhydes sont proportionnelles aux fluctuations de la température et de l'humidité relative (van Netten et al., 1989), mais il semblerait que la dépendance des niveaux du formaldéhyde à la température est plus complexe (Zhang et al., 2007). Salthammer et al. (2010) ont simulé le modèle de Berge en mettant en avant la relation entre l'humidité, la température et le HCHO et les résultats sont présentés dans la Figure 1.3.2.

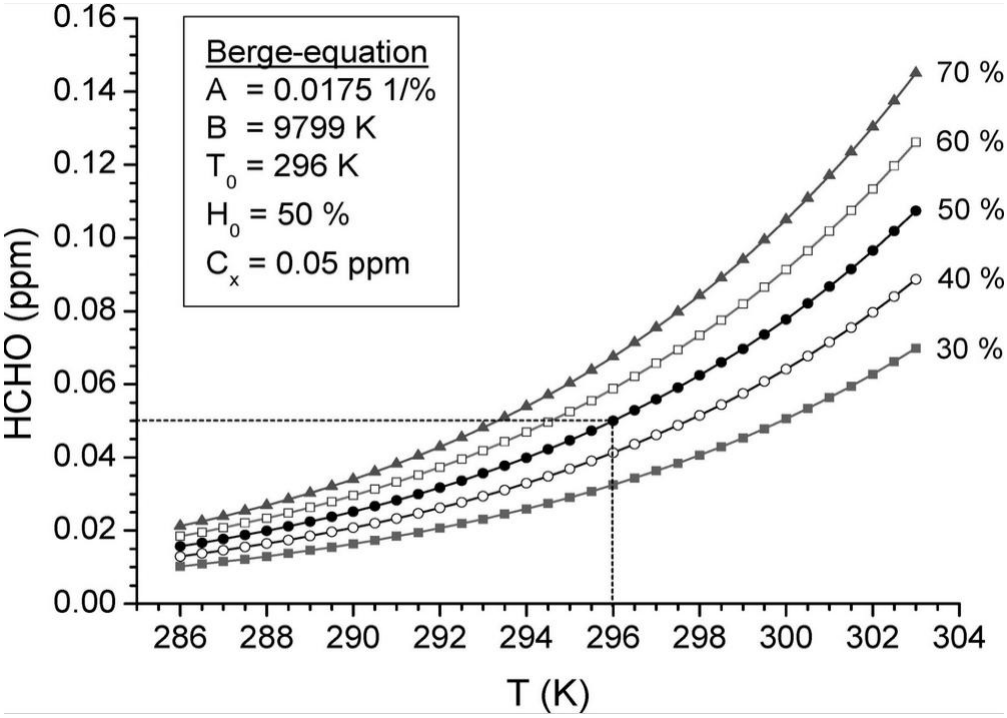


FIGURE 1.3.2 – *Variations des concentrations du formaldéhyde en fonction de la température et de l'humidité relatives, calculées à partir de l'équation de Berge. Les conditions initiales sont : T_0= 296 K, H_0= 50% et C_x =0.05 ppm (d'après Salthammer et al. (2010)).*

Suivant la même logique de modélisation physique, Panzhauser et al. (1993) ont mené une étude sur 100 bâtiments résidentiels. Les valeurs de HCHO les plus élevées ont été observées principalement dans les ménages où il y a des fumeurs, les maisons dotées de gaz naturel ou lors d'une mauvaise ventilation. L'analyse de corrélation ne montre pas de liaison particulière avec aucune de ces composantes ; c'est pour cette raison que les auteurs soupçonnent une relation beaucoup plus complexe et non-linéaire entre ces variables. Un modèle physique a été alors développé afin de mettre en évidence l'importance de certaines variables, notamment la ventilation (le renouvellement de l'air). Dans une de leurs expériences en régulant un débit d'air neuf à $25\,m^3 \cdot h^{-1}$, une variation brusque de la concentration du HCHO a été enregistrée de 0.19 à $0.05\,mg \cdot m^{-3}$ en 20 minutes

La moyenne des concentrations dans les bâtiments résidentiels est en deçà de $0.05\,mg \cdot m^{-3}$ avec des concentrations plus élevées pour les nouveaux bâtiments (Wolkoff & Nielsen, 2010). Ces concentrations

sont obtenues généralement par des mesures agrégées entre 1 jour jusqu'à une semaine. Liu et al. (2006) ont mené une étude sur 234 maisons (353 observations sur 48 h) et ont rapporté une médiane de $0.02\,\mathrm{mg}\cdot\mathrm{m}^{-3}$; une valeur similaire à celle observée dans les maisons québécoises par Gilbert et al. (2006). De façon générale, plus l'environnement est équipé de surfaces en bois et relativement récent avec des températures et niveaux d'humidité suffisants, plus la concentration en formaldéhyde dans l'air est élevée. Une comparaison entre les maisons japonaises (Nagoya) et Suédoises (Uppsala) dans différentes conditions montre cette relation. Dans les établissements publiques en Europe, la concentration moyenne dépasse rarement $0.025\,\mathrm{mg}\cdot\mathrm{m}^{-3}$ (Wolkoff & Nielsen, 2010; De Bruin et al., 2008), tandis qu'elle peut varier de 2 à 10 fois plus en moyenne en Chine (Tang et al., 2009).

Pratiquement, tous ces travaux mettent en évidence l'importance du taux de renouvellement d'air sur les niveaux du HCHO (Gilbert et al., 2006; Salthammer et al., 2010, 1995), ainsi que le fait que la contribution extérieure n'est pas très significative dans la plupart des villes européennes (De Bruin et al., 2008).

En résumé : la concentration dans les bureaux est généralement faible et l'air extérieur permet de diluer la concentration intérieure du HCHO.

La variabilité temporelle de la concentration intérieure en formaldéhyde

Il est très difficile de trouver des études *in situ* traitant l'analyse de la variabilité temporelle des concentrations en formaldéhyde dans l'air intérieur. Les raisons de cette lacune peuvent s'expliquer par le fait que dans la plupart des cas, les méthodes de mesure ne permettent pas d'avoir des séries de données suffisamment longues pour établir des profils de variabilité. En leur absence, la variabilité des sources et de leur contribution est difficile à mettre en évidence. S'ajoute à ces difficultés, le traitement statistique pour de faibles échantillons qui n'est pas adapté pour mettre en évidence les structures de variabilité et encore moins, en présence de valeurs manquantes.

L'étude de Wang et al. (2010) met en évidence l'importance des concentrations mesurées dans deux villes en Chine et une différence entre les observations en hiver et en été. Ainsi, leurs résultats s'accordent avec ceux trouvés par Dingle & Franklin (2002). Les concentrations observées en été étaient beaucoup plus importantes que celles observées en hiver, mais aucune différence significative entre l'hiver et l'automne. Les mêmes conclusions ont été rapportées dans une étude sur les concentrations d'aldéhydes à Paris, dans des centres de nouveaux nés (Dassonville et al., 2009). La concentration intérieure de formaldéhyde varie également en fonction de l'âge du bâtiment et plus particulièrement des matériaux qui tend à diminuer avec le temps (Langer et al., 2016). Par ailleurs, l'émission de formaldéhyde par certains matériaux comme les panneaux de bois varie en fonction de la température et de l'humidité, des paramètres environnementaux qui varient dans le temps (Liang et al., 2016).

1.3.3.7 Les particules en suspension dans l'air (PM)

Le mot aérosol aurait été inventé par Schmauss en 1920 pour désigner toute particule solide ou liquide ayant une vitesse de chute négligeable. Les particules en suspension recouvrent un très large spectre de taille allant de quelques fractions de nanomètres à quelques centaines de micromètres (Seinfeld & Pandis, 2012). Plusieurs classifications ont été élaborées en fonction des effets qu'elles induisent sur la santé ou de leurs caractéristiques physico-chimiques. En effet, on trouve selon la distribution granulométrique : des particules ultrafines ($d_a < 0.1\,\mu$m : où d_a est le diamètre aérodynamique), des particules fines ($0.1 < d_a < 2.5\,\mu$m) et des grosses particules ($d_a > 2.5\,\mu$m) (Chow et al., 1998). Ainsi, les termes PM_{10} et $PM_{2.5}$ représentent la fraction de l'aérosol atmosphérique qui contient les particules ayant un diamètre aérodynamique inférieur ou égal à $10\,\mu$m et $2.5\,\mu$m respectivement.

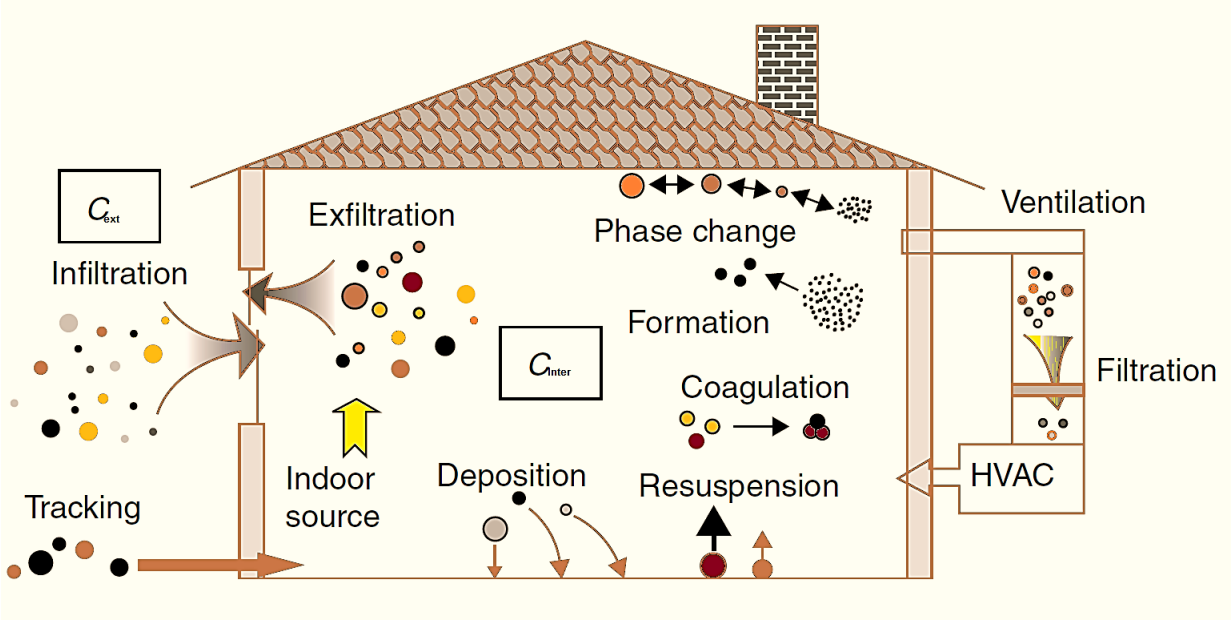


FIGURE 1.3.3 – *Processus affectant les concentrations intérieures des particules (modifié de (Thatcher et al., 2003)). C_{int} et C_{ext} représentent les concentrations intérieures et extérieures, respectivement.*

Processus affectant les concentrations intérieures des particules Le transport, la remise en suspension et le dépôt des particules dans les environnements intérieurs sont fondamentalement influencés par une série de transformations et de différents processus physico-chimiques (Gundel & Sextro, 2005). Les mécanismes de formation et de transformations sont intrinsèquement liés aux différentes sources, aux paramètres climatiques et à l'occupation. Hormis les mécanismes de sorption et de désorption, la Figure 1.3.3 montre les principaux processus qui entrent en jeu dans la détermination de la concentration des particules intérieures.

Ces facteurs pourraient mener à des changements considérables en termes de composition chimique des particules, de leurs caractéristiques physiques et de leurs distributions granulométriques. A ces effets, les concentrations (en masse ou en nombre) des particules et la contribution de leurs sources varieraient de manières différentes en fonction de l'ampleur de ces processus

Sources de particules

Extérieures : À travers les systèmes de ventilations, l'enveloppe du bâtiment et les ouvrants (portes et fenêtres), les particules sont en mouvement continu entre l'air intérieur et l'air extérieur. Au moins la moitié des particules inhalées à l'intérieur sont d'origine extérieure (Wallace et al., 2003). Cette observation montre l'importance de la pollution particulaire extérieure dans la détermination de la concentration en particules dans les espaces intérieurs.

Bien que la capacité d'infiltration des particules soit variable selon la taille des particules (de 0.38 à 0.94 pour 0.02−0.5 μm et de 0.12 à 0.53 pour 0.7−10 μm (Abt et al., 2000)), des études plus ou moins

anciennes ont montré que la part de particules en air intérieur d'origine extérieure peut atteindre 33 % et ceci même avec portes et fenêtres fermées (Alzona et al., 1979).

Dans l'étude de Han et al. (2015), la variation de la concentration intérieure en $PM_{2.5}$peut être expliquée pour environ 81 % à 90 % par les variations extérieures. Un indicateur simple comme le ratio intérieur/extérieur (Indoor/Outdoor-(I/O)) peut être utilisé pour estimer la contribution des niveaux extérieurs aux concentrations intérieures. Cet indicateur dépend d'un très grand nombre de paramètres, mais surtout des sources intérieures et des échanges d'air du bâtiment avec l'extérieur. Notamment, l'ouverture des fenêtres, les équipements de ventilation et la perméabilité liée à la structure du bâti jouent un rôle important dans ces échanges. Il est donc très difficile de faire des extrapolations de mesures observées dans certains types de bâtis sur l'ensemble des environnements intérieurs. Une revue de la littérature sur la relation entre les concentrations intérieures et extérieures par le biais du ratio (I/O), du facteur de pénétration et de l'infiltration a été réalisée par Chen & Zhao (2011).

TABLE 1.3.1 – Ratio intérieur/extérieur (I/O) de la concentration en particules dans les bureaux.

Type de particules	**Ratio**(I/O)	**État d'occupation**	**Type de ventilation**	**Référence**
$PM_{2.5}$	0.12	occupé	Mécanique	Sinclair et al. (1990)
	0.5	NA		Ho et al. (2004)
	0.76	occupé	NA	Tovalin-Ahumada et al. (2007)
	0.51	non occupé	NA	
	0.93	occupé	NA	
	0.88	non occupé	NA	
PM_{10}	0.29	occupé	Mécanique	Chatoutsidou et al. (2015)
	0.19	non occupé		
$0.014 - 0.5\,\mu$m	0.21	occupé		
	0.09	non occupé		
$0.5 - 18\,\mu$m	0.2	occupé		
	0.03	non occupé		
$0.008 - 0.5\,\mu$m	$0.07 - 0.28$	NA	Infiltration	Koponen et al. (2001)
	$0.03 - 0.32$	NA	Mécanique	
$0.01 - 1\,\mu$m	0.77	Fumeurs	NA	
	0.85	Fumeurs	NA	
$0.1 - 10\,\mu$m	$0.02 - 0.50$	NA	Mécanique	Gupta & Cheong (2007)
$2.5 - 15\,\mu$m	0.07	Occupants	Mécanique	Sinclair et al. (1990)

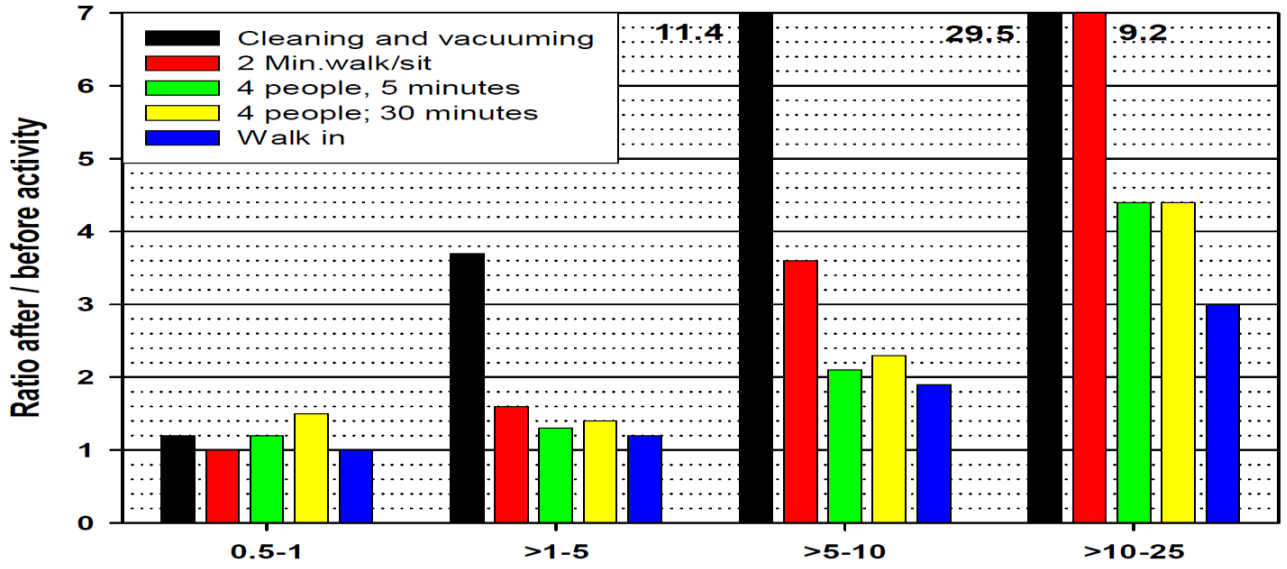


FIGURE 1.3.4 – *Rôle de l'activité des occupants dans la détermination des différents types de particules. Les abscisses correspondent aux diamètres de particules en µ.m. (modifié de (Thatcher & Layton, 1995))*

Le Tableau 1.3.1 donne quelques valeurs du ratio (I/O) observé dans les environnements de type bureaux. Quel que soit l'état d'occupation ou le type de ventilation, le ratio (I/O) est toujours inférieur à 1.

Intérieures : Ces dernières années, de nombreux travaux ont cherché à examiner plus en détail la contribution de l'activité des occupants aux concentrations en particules. L'apport de l'activité du nettoyage par les occupants sur les concentrations des PM a été mis en évidence par plusieurs études (Long et al., 2001; Ulens et al., 2014).

De façon générale, l'utilisation de l'aspirateur et le dépoussiérage contribuent à l'augmentation de la masse des grosses particules $(6-10\,\mu\text{m})$ (Abt et al., 2000). Une analyse par régression linéaire multiple a été appliquée afin d'estimer la contribution des différentes activités domestiques : la cuisine (cuisson, griller, faire sauter etc) ; le nettoyage (aspirateur etc) ; l'occupation (caractérisée par le mouvement des occupants) et le lavage (Abt et al., 2000). Les résultats de cette étude montrent que ces variables contribuent de façon distributive en fonction de la taille des particules. Ainsi, d'après le modèle de régression, la cuisson et le mouvement des occupants ont plus d'impact dans la détermination de la fraction de particules ayant une taille supérieure à $2\,\mu$m, le lavage est la variable la moins importante sauf pour les particules de taille entre $0.02-0.5\,\mu$m.

Les mêmes résultats ont été rapportés dans (Afshari et al., 2005; Thatcher & Layton, 1995) : la contribution de l'activité de nettoyage est infime sur les particules ultrafines. Au contraire, le nettoyage est très significatif pour les particules supérieures à $1\,\mu$m. La Figure 1.3.4 montre les ratios des différentes activités des occupants avant et après l'usage ou les mouvements. Ces observations peuvent être interprétées par la re-suspension des particules déposées sur les surfaces (Thatcher & Layton, 1995).

L'émission de particules fines et ultrafines est variable selon les activités domestiques, (Afshari et al., 2005; Géhin et al., 2008). Les processus de combustion principalement la cuisson des aliments et le

fonctionnement d'un chauffage d'appoint à pétrole représente les sources principales de particules dans les environnements intérieurs avec une forte proportion de particules ultrafines..

Concentrations et variabilité En France, la principale source de données de la pollution particulaire des différents espaces intérieurs reste aujourd'hui les campagnes nationales menées par l'OQAI. Pour la campagne logement (2003−2005) par exemple, les mesures effectuées dans le séjour de 297 logements pour les PM_{10} et de 290 logements pour les $PM_{2.5}$ sont, d'un point de vue répartition spatiale, représentatives de l'ensemble du parc national. Les médianes des concentrations en PM_{10} et en $PM_{2.5}$ sont respectivement $31.3\,\mu g/m^3$ $(\text{max}=523\,\mu g/m^3)$ et $19.1\,\mu g/m^3 (\text{max}=523\,\mu g/m^3)$ (Kirchner et al., 2007a).

Une campagne de mesure en PM_8 dans 133 bureaux parisiens ayant différents systèmes de ventilation montre qu'en moyenne, les concentrations dans les bureaux équipés de ventilation mécanique contrôlée sont les moins élevées, avec $93.5\,\mu g \cdot m^{-3}$, contre $148\,\mu g \cdot m^{-3}$ pour les bureaux équipés d'un climatiseur et $136\,\mu g \cdot m^{-3}$ dans les bureaux ventilés naturellement (Vincent et al., 1997). Les concentrations en $PM_{2.5}$ peuvent atteindre les $265\,\mu g \cdot m^{-3}$ en présence de fumeurs dans les bureaux (Mosqueron et al., 2001) et en moyenne, elles avoisinaient les $100\,\mu g \cdot m^{-3}$ en présence d'au moins deux fumeurs dans un logement (Ramalho et al., 2012).

Récemment, plusieurs études ont cherché à quantifier la contribution des appareils informatiques et des équipements de bureau aux concentrations en particules (Uhde et al., 2006; He et al., 2007; Wensing et al., 2008; Morawska et al., 2008, 2009; Schripp et al., 2009). La plupart de ces études ont été effectuées dans une chambre de simulation, donc l'extrapolation des résultats en environnement réel reste très difficile.

Par ailleurs, l'impact des facteurs météorologiques sur la concentration intérieure est souvent mis en évidence dans la littérature. Chan (2002) a étudié les corrélations entre la variation des concentrations intérieures, extérieures, le ratio intérieur/extérieur et les paramètres climatiques extérieurs. L'étude par régression montre que la température, l'humidité relative et l'irradiance jouent un rôle important dans la détermination des variations du ratio I/O.

Le variabilité spatiale des niveaux de particules pour la campagne "logement" (dont l'analyse des premiers résultats est rapportée dans (Kirchner et al., 2007a; Mandin et al., 2009; Kirchner et al., 2007b)) est homogène sur le parc national : les concentrations en particules ne dépendent pas de la zone urbaine, périurbaine ou rurale du logement (Ramalho et al., 2012). Par contre, une distribution mensuelle des concentrations est observée, avec un maximum à la fin de l'hiver et minimum au cours de l'été (Ramalho et al., 2012). La variation saisonnière de la concentration en particules extérieures et son impact sur l'air intérieur a été largement discutée dans la littérature (Massey et al., 2012; Zhu et al., 2015; Hassanvand et al., 2014; Amato et al., 2014).

1.4 Facteurs influençant la qualité de l'air intérieur

La qualité de l'air intérieur dépend de différents facteurs comme :

- **L'environnement extérieur (macroenvironnement) :** L'environnement extérieur regroupant les sources de pollution extérieure, la nature des sols et leur niveau de contamination, et les conditions climatiques et météorologiques est en perpétuel interaction avec l'environnement intérieur (Institute of Medicine, 2011). L'air circule de l'un à l'autre soit librement (fenêtres ouvertes) ou selon des contraintes (bouches d'entrée d'air, infiltrations). Les conditions extérieures de température, d'humidité et de pression se répercutent sur l'enveloppe du bâti qui restitue tout

ou partie de ces conditions à l'intérieur. Les échanges thermiques entre l'extérieur et l'intérieur du bâtiment jouent un rôle important sur la dispersion des polluants. Par ailleurs, l'ensoleillement provoque un réchauffement des surfaces à l'intérieur et l'extérieur du bâti. L'environnement extérieur influence donc à tout instant la qualité de l'air intérieur. Son action sera modulée par l'éventuelle présence d'un système spécifique de ventilation et l'intervention des occupants sur les ouvrants.

- **Les conditions climatiques intérieures :** Les conditions climatiques intérieures sont la plupart du temps fixées par les occupants ou le gestionnaire du bâtiment au travers de la présence et des paramètres de fonctionnement d'un système de chauffage et parfois un système spécifique de traitement d'air et de ventilation. Ces systèmes vont permettre de contrebalancer ou d'atténuer l'impact des conditions extérieures au profit d'un meilleur confort essentiellement thermique des occupants. Ces conditions vont agir sur les paramètres d'émission des sources et les mouvements d'air entre les différentes volumes de l'espace intérieur et par conséquent sur les niveaux et la distribution des concentrations des substances et particules dans l'air. Ils peuvent également conduire à des conditions favorables à la prolifération des bio-contaminants, des microorganismes susceptibles d'émettre à leur tour des substances et des toxines.

- **Le bâti :** les systèmes et les composants de construction peuvent avoir une influence directe et/ou indirecte sur la qualité de l'air intérieur et en particulier : l'enveloppe du bâti à l'interface extérieur/intérieur, les matériaux de construction et les revêtements intérieurs (sol, murs, plafond, joints, colles), le fonctionnement et la nature du système de chauffage, la ventilation et le conditionnement d'air, les infiltrations et les gaines techniques existantes (tuyauterie, fils électriques) et toutes les relations entre ces éléments.

- **Le mobilier :** La nature des matériaux d'ameublement (essences de bois, bois agglomérés, mousses et tissus), les produits de décoration, d'entretien et de bricolage utilisés, vont venir affecter la qualité de l'air intérieur. En outre, les équipements bureautiques (matériels informatiques, photocopieurs, imprimantes, etc) notamment dans les espaces de bureaux jouent un rôle important (Bakó-Biró et al., 2004).

- **L'occupant :** a un rôle déterminant sur les niveaux de pollution auxquels il est exposé ; en effet, ses activités et son comportement sont des facteurs très influents sur la qualité de l'air intérieur. Il peut activer des sources de pollution (tabagisme, ménage, cuisine, utilisation d'appareils de combustion, présence d'animaux domestiques, produits d'entretien, etc.) et agir sur les ouvrants ou la mise en route d'un système de ventilation ou de traitement d'air (dilution de la concentration des polluants intérieurs, apport de polluants d'origine extérieure en particulier des oxydants (ozone, radicaux libres). La perception individuelle ou collective des occupants en termes de santé et de confort va conditionner son comportement et in fine la qualité de l'air intérieur.

1.5 Impact de la pollution intérieure

L'impact de la pollution intérieure sur les occupants est devenu l'un des enjeux majeurs de la santé publique. Certains polluants les plus courants dans l'air intérieur (formaldéhyde, benzène, CO, NO_2, particules, ...), mais aussi des polluants présents dans les produits de consommation courante et dans notre alimentation (phtalates, pesticides, métaux lourds...) sont reconnus comme susceptibles de provo-

quer des effets sanitaires principalement à long terme (Program, 2010). Ces agents chimiques et physiques agissent en effet de façon synergique sur l'organisme, et peuvent se manifester tout d'abord par des symptômes (irritation muqueuse, dyspnée, peau sèche...) liés à la détérioration de la qualité de l'air intérieur (Hoskins, 2003; Menzies & Bourbeau, 1997). Des incidences respiratoires ont été mis en évidence par de nombreuses études récentes (Mendell et al., 2002). Le nombre annuel de décès par cancer du poumon qui serait attribuable à l'exposition domestique au radon en France métropolitaine varie de 1 200 à 2 900 (Kirchner et al., 2011).Le centre international de recherche sur le cancer a récemment réévalué le formaldéhyde, qu'il a classé dans le groupe 1 des substances avérées cancérogènes pour l'homme (Cogliano et al., 2005).

Énormément d'études on été accomplies en vue d'établir les causes des problèmes liés à la QAI, des progrès notables ont été réalisés ces dernières années dans la connaissance des contaminants et les facteurs qui contribuent à l'altérer. On peut trouver un état des connaissances actualisé dans (Brunekreef & Holgate, 2002; Nadadur & Hollingsworth, 2015; Ilacqua et al., 2015) et le rôle des différentes composantes du bâti sur les effets indésirables des expositions aux polluants de différentes sources dans (Spengler et al., 2001).

Le terme "Syndrome de l'Habitat Malsain" (Sick Building Syndrome, SBS en anglais) est contemporain à la crise énergétique des années 70 qui a déclenché des modifications de l'architecture, des matériaux de construction, des équipements du bâti et des systèmes de climatisation afin d'économiser l'énergie. Ainsi, de nombreuses personnes se plaignent d'inconfort, voire de pathologie : la qualité de l'air intérieur est souvent mise en cause (Perdrix et al., 2005). L'étude de Ezzati (2005) a mis l'accent sur le fait que la combinaison des contaminants dans l'air inhalé à l'intérieur a des effets sur les symptômes ou sur une combinaison de symptômes. Dans certaines études, les niveaux de formaldéhyde sont mis en cause dans la survenue des SBS (Brinke et al., 1998), mais la relation entre SBS et qualité de l'air intérieur n'est pas toujours systématique, tant d'autres facteurs psycho-sociologiques peuvent rentrer en ligne de compte (Marchand et al., 2013).

La qualité de l'air intérieur a également un impact sur la performance scolaire et sur le rendement au travail, par conséquent sur la productivité. Ainsi plusieurs études internationales mais surtout américaines se sont penchées sur les coûts induits par une mauvaise qualité de l'air (Fisk & Rosenfeld, 1997; Fisk & Seppanen, 2007; Mendell et al., 2002). Les coûts des effets d'une mauvaise qualité de l'air intérieur en France, calculés selon les indicateurs globaux utilisés par l'OMS, sont estimés entre 12,8 et 38,4 milliards d'euros par an (Kirchner et al., 2011).

1.6 Modéliser la qualité de l'air intérieur : une problématique complexe

Dans cette section, la problématique de la modélisation de la qualité de l'air intérieur est abordée afin d'appréhender le rôle et la complexité du phénomène pour le développement d'un modèle de prévision. Elle donne ainsi des justifications de l'approche statistique proposée qui sera détaillée dans les prochains chapitres.

Un micro-environnement est un espace complexe, du fait de sa caractéristique de forme, de volume, de ses sollicitations nombreuses et fluctuantes dans le temps, de la nature multiple des transferts de polluants et des réactions physico-chimiques dont il peut être le siège. Modéliser un tel système revient à mettre en évidence la transformation des émissions influencées par les divers paramètres (climatiques, spécificité du bâti, l'occupation et l'activité des occupants) en concentrations qui en résultent à un instant donné. Cette transformation peut être abordée soit en suivant le chemin naturel des émissions

jusqu'à la concentration en polluant, ou bien par une approche qui consiste à utiliser les effets comme point de départ (modélisation inverse).

Qu'il s'agisse de la modélisation directe suivant les processus de causes à effets, ou par une modélisation inverse, les deux visions ont pour but de comprendre et/ou d'expliquer la qualité de l'air ambiant des espaces intérieurs.

Face à cette problématique, deux choix s'offrent à nous suivant l'angle d'analyse : une modélisation physique (directe de causes à effets) ou une modélisation statistique. L'utilisation de l'une ou de l'autre nécessite des sources d'information et des connaissances qui leurs sont propres. Alors que dans le premier cas, les modèles requièrent des connaissances sur les causes et sur les effets, donc un déterminisme qui ne connaît que la nécessité ou l'impossibilité, les modèles statistiques dessinent un monde composé d'évènements définis comme des ensembles qui peuvent se réaliser ou non, selon des degrés. La connaissance des causes dans les modèles statistiques n'est pas nécessaire pour fonder une explication des effets, mais ce sont les effets qui sont nécessaires pour inférer sur les causes.

Pour préciser le cadre dans lequel la modélisation de la qualité de l'air des environnements intérieurs peut être construite, on se propose de discuter en premier de l'approche déterministe, puis les "obstacles" auxquels elle est confrontée. Ensuite, nous analysons la "nécessité" d'aborder cette problématique par une approche statistique.

1.6.1 Bref aperçu des modèles physico-chimiques pour la QAI

Il existe de nombreux travaux dont l'approche consiste à modéliser les paramètres de l'environnement intérieur par des modèles physiques déterministes : méthodes zonales ou par maillage. Typiquement, lorsqu'on s'intéresse à l'évolution d'une variable dans le temps, la modélisation se construit généralement par un système d'équations différentielles. Ce dernier a une référence réelle dans le cadre des équations de conservation de masse, ou de façon plus générique les modèles CMB (Chemical Mass Balance).

La modélisation déterministe passe généralement par un découpage du domaine en zones homogènes, ensuite un bilan de masse est appliqué dans chaque zone pour déterminer les différents paramètres. Ce principe a été largement étudié et donne de bons résultats pour des situations expérimentales simples durant lesquelles les conditions ne varient pas ou peu.

Soit $C_{i_{\text{int}}}(t)$ la concentration d'une espèce i d'un polluant intérieur $P_{i_{\text{int}}}$ à instant t mesurée dans la zone z_k de volume v_k, où le mélange de l'air est homogène dans le volume total de la pièce $V\left(m^3\right)$, avec $\cap v_k = \emptyset$ et $\cup v_k = V$. Le modèle déterministe monozone est basé sur l'équation du bilan de la masse décrite suivant la loi de conservation de la masse :

$$V\frac{dC_{i_{\text{int}}}}{dt} = \text{vitesse de la variation de la masse de polluant } P_i \tag{1.6.1}$$

Dans le cas des modèles multizones, cette équation est étendue pour chaque domaine de la zone.

Dans la plupart des cas, plusieurs hypothèses sont formulées afin de rendre l'estimation de la concentration possible, la plus importante étant la situation d'un *état d'équilibre ou état stationnaire* : Pour un facteur de mélange égal à 1, une absence d'infiltration, une absence de désorption des polluants déposés sur les surfaces et l'absence d'un système d'épuration de l'air, l'équation 1.6.1 peut s'écrire :

$$\frac{dC_{i_{\text{int}}}}{dt} = p \cdot \alpha C_{i_{\text{ext}}} - p' \cdot \alpha C_{i_{\text{int}}} + \frac{1}{V}\left(\sum_s Sr_{int}\right) - C_{i_{int}}\sum_l \gamma_l - C_{i_{int}}\sum_j k_j C_{j_{int}} \tag{1.6.2}$$

où

- $C_{i_{\text{int}}}$: la concentration intérieure (µg/m3) mesurée dans un espace ayant un volume $V\left(m^3\right)$;
- $C_{i_{\text{ext}}}$: la concentration extérieure (µg/m3) en polluant $P_{i_{\text{ext}}}$;
- p : facteur de pénétration extérieur-intérieur, grandeur adimensionnelle ;
- p' : facteur de pénétration intérieur-extérieur, grandeur adimensionnelle ;
- α : le taux de renouvellement de l'air $\left(s^{-1}\right)$;
- Sr_{int} : représente le débit d'émission d'une source intérieure particulière de l'espèce i $(\mu g/s)$. La variable Sr_{int} est générique, elle peut être décomposée en plusieurs facteurs en faisant intervenir une constante d'émission $\left(s^{-1}\right)$ qui dépend d'un ou de plusieurs paramètres de diffusion (au sein du matériau, matériau-interface et interface-air), la surface (pour une source surfacique), la charge (ou la taille du réservoir) de l'espèce i, l'âge du matériau et éventuellement des conditions de température et d'humidité propre au matériau/produit ;
- γ_l : constante de sorption sur les surfaces calculée pour chaque nature de surface en tenant compte du rapport surface/volume associé $\left(s^{-1}\right)$. Cette dernière va dépendre des conditions aérauliques à proximité des surfaces, de leur nature ainsi que des conditions de température et humidité de l'air et du matériau ;
- k_j : constante de réaction de second ordre $(m^3/\mu g.s^{-1})$
- $C_{j_{\text{int}}}$: la concentration intérieure du réactif $(m^3/\mu g)$.

Pour la prévision, une des difficultés rencontrées par ce modèle est d'ordre méthodologique : la paramétrisation physico-chimique repose sur une formulation statique. Une difficulté non négligeable réside dans la difficulté pratique de mettre en œuvre ce type de modèle dans un environnement réel. Par exemple, malgré l'importance des informations recensées dans la base de données PANDORE (une comPilAtion des émissioNs Des pOlluants de l'aiR intériEur) (Abadie & Blondeau, 2011), qui regroupe environ 500 sources de polluants représentant près de 7000 données d'émissions de polluants, la prévision reste difficile à mettre en œuvre. En effet, l'étendue des débits d'émission possibles pour un matériau donné n'est pas connue. De plus, les revêtements forment avec la colle et le substrat un matériau composite dont il est difficile de déterminer le débit d'émission final qui ne découle pas d'une hypothèse d'additivité. Par ailleurs, les débits d'émissions évoluent au cours du temps non seulement du fait des conditions climatiques, mais également par l'action de certains oxydants aptes à modifier la nature même des émissions.

Par ailleurs, le modèle ne prend pas en compte la présence et le comportement de l'occupant qui va au travers de ses actions modifier les termes associés à l'émission, au renouvellement d'air et dans une moindre mesure aux surfaces disponibles pour la sorption des espèces présentes dans l'air. Ces paramètres sont cruciaux pour pouvoir prévoir l'évolution des concentrations dans un environnement réel occupé.

1.6.1.1 Réactivité chimique dans le modèle

La prise en compte de la réactivité chimique dans le modèle nécessite des hypothèses sur le nombre d'espèces susceptibles de réagir et sur les réactions chimiques à considérer qu'elles se déroulent en phase homogène ou en phase hétérogène. Elle implique donc un grand nombre de paramètres à introduire dans le modèle qui peuvent varier selon les conditions de température, d'humidité mais également de rayonnement. Différents modèles de réactivité chimique ont été développés pour l'air intérieur : ICEM (Sarwar et al., 2002), MCM (Carslaw, 2007) et INCA-Indoor de Mendez et al. (2015). Ces modèles utilisent plusieurs dizaines d'espèces et de réactions. Pour la prévision, les concentrations initiales de l'ensemble des espèces mises en jeu doivent être renseignées a minima. Ces modèles sont essentiellement

utilisés dans des simulations pour étudier l'impact de différents scénarios et comprendre les mécanismes réactionnels.

1.6.1.2 Absence de mise à jour dynamique dans le modèle

Dans le raisonnement déterministe du modèle 1.6.2, on se place délibérément dans un système statique : la concentration d'un polluant est l'effet direct et instantané de la variation de ses paramètres. Donc, aucun retard ou délai entre l'émission de la source, la formation et la transformation en polluant n'a été envisagé dans ce type de modélisation. Cette "pathologie" dans les modèles physiques, même avec l'introduction d'un certain délai, ne permet pas de répondre aux besoins de prévision.

Il existe plusieurs sources majeures qui conditionnent les limites de cette approche, la plus importante peut être attribuée à l'estimation en temps réel de tous les paramètres qui déterminent la fluctuation d'une concentration, la deuxième, à notre sens, étant l'absence d'une formulation qui permet de projeter, par une certaine application, les données de fluctuations sur un horizon futur, sans faire appel aux prévisions des valeurs de ses paramètres.

1.6.1.3 Difficultés de mise en œuvre et pratique

Le phénomène de pollution intérieure est irréversible et indissociable de la spécificité de son environnement : on ne peut pas assurer la reproductibilité (au sens strict) de l'observation dans les mêmes conditions espacées dans le temps. Cela a des conséquences considérables dans l'interprétation des résultats de calcul. L'échec de ces modèles est généralement attribué à la qualité (pertinence) des données et à leurs incertitudes, ou encore aux difficultés de mises en œuvre.

Par ailleurs, la paramétrisation physico-chimique, telle que représentée dans l'équation 1.6.2, nécessite de recueillir l'ensemble des données relatives à l'environnement étudié. Ces modèles font parfois intervenir plusieurs centaines de variables pour décrire plusieurs phénomènes complexes, ce qui rend la résolution numérique très difficile.

Voyons donc maintenant comment chercher à remplir ces lacunes par une autre approche.

1.6.2 Vers des modèles statistiques (l'environnemetrie intérieure)

L'approche statistique est généralement utilisée lorsque les connaissances *a priori* sur un système sont insuffisantes ou lorsque les paramètres issus des modèles physiques ne peuvent pas être complètement spécifiés. Dans notre cas, seules les sorties (variables temporelles) du système ont pu être recueillies par la mesure. Les modèles statistiques ont alors pour but d'utiliser toute l'information disponible afin de mieux reproduire le comportement du vrai système sur la base de ces données. Notamment, la modélisation inverse permet d'inférer sur la nature du système ou de fournir les prévisions sur l'état futur du système.

La modélisation de la qualité de l'air intérieur peut également être motivée par un objectif pratique : mettre en évidence la variabilité des sources de fluctuation et de leurs contributions. Cette modélisation peut être employée à des fins de prévision : soit par la seule série temporelle des concentrations de polluants prise individuellement, ou *via* un ensemble de variables d'état et de facteurs.

Comme nous l'avons vu précédemment, dans le cadre de la modélisation de la qualité de l'air intérieur à des fins prévisionnelles, nous cherchons à établir un modèle qui fonctionne aussi bien que possible pour différents polluants observés dans différents environnements, et ce pour des questions pratiques

et scientifiques. Autrement dit, il s'agit de vérifier la qualité de prévision d'un polluant sur différentes périodes et d'en dégager les conditions pour juger la stabilité du modèle proposé, au moins à première vue sur un environnement donné.

L'objectif pratique consiste à proposer un modèle qui, dans la plupart des cas, fournit des prévisions d'une qualité acceptable sur la plupart des espaces clos instrumentés. Globalement, il s'agit de confronter un même modèle à différentes données du même polluant afin de déterminer les conditions générales d'une bonne prévision.

1.7 Bilan et conclusion

La problématique de la qualité de l'air intérieur est vaste et en émergence constante. Ce chapitre a passé en revue les différents polluants de l'air intérieur et leurs sources, ainsi que les effets sur la santé de l'occupant. Bien que nombreuses, les données relatives à la concentration des polluants dans l'air sont très disparates ; elles dépendent de plusieurs paramètres et les conditions de mesure sont très différentes selon l'environnement et le type du polluant étudié.

La variabilité temporelle de la concentration des polluants est très peu étudiée. En outre, le pas de temps utilisé dans les mesures est souvent très grand. Par exemple, les mesures les plus pratiques sont réalisées par prélèvement passif et intègrent des durées de l'ordre de la journée ou de la semaine. Ainsi, la concentration de formaldéhyde a été mesurée sur 4,5 jours dans les écoles d'une campagne de surveillance (Michelot et al., 2013) et sur 7 jours dans les logements (Kirchner et al., 2007b). Ces mesures sont interprétées en termes d'exposition malgré des durées d'occupation parfois très variables. Elles ne permettent pas d'extrapoler à des expositions court terme.

À l'heure actuelle, le principe fondamental de la modélisation de la qualité de l'air intérieur est basé sur les équations de la conservation de la masse. Cette dernière traduit le fait que la variation temporelle de la concentration en polluant varie en fonction de la contribution relative des termes sources et puits. Néanmoins, et comme nous l'avons vu à la fin de ce chapitre, les modèles de ce type ne permettent pas de fournir les prévisions de la concentration des polluants de l'air intérieur.

CHAPITRE 2

MESURE DE LA QUALITÉ DE L'AIR DANS UN *MICRO*-ENVIRONNEMENT

La mesure de la qualité de l'air dans un espace intérieur se réduit le plus souvent à évaluer la composition de l'air par un système capable de quantifier les concentrations en polluants cibles dans cet environnement. L'évaluation de la QAI est d'autant plus précise si les mesures sont recueillies avec un pas de temps fin sur une période longue. Pour assurer la surveillance de la QAI, on dispose d'un système de capteurs qui fournit, in situ, la concentration des polluants en temps réel. Dans cette thèse, l'accent a été mis sur la mesure en continu des différents paramètres dans différents environnements.

Sommaire

2.1 Introduction

La mesure des constituants de l'air est nécessaire pour caractériser quantitativement l'air qui nous entoure. Cependant, devant l'immense variété des contaminants émis par différentes sources, la quanti-

fication de tous ces composés est impossible. En pratique, un choix de certains polluants cibles comme indicateurs de la pollution intérieure est toujours effectué. Dans notre cas, nous nous sommes intéressés au formaldéhyde et aux particules fines.

Nous présentons ici le matériel mis en place pour l'évaluation de la variabilité temporelle de la QAI et nos choix de méthodologie pour atteindre les objectifs vers lesquels nous voulons tendre. Tout ce chapitre sera parcouru par une présentation des environnements instrumentés, des campagnes de mesures, des données disponibles et des statistiques sur les concentrations des polluants mesurés.

Bien que descriptives, ces statistiques ont fait émerger des questionnements sur les structures temporelles des séries. Par exemple, existe-il un schéma de variabilité d'un polluant pour tous les environnements étudiés, en toutes circonstances ? On verra plus tard que dans les environnements étudiés, on observe couramment l'existence d'un profil de variation régulier pour le CO_2 et pour certaines fractions de particules, mais rarement pour le formaldéhyde et les particules fines. Ce chapitre est donc dédié à la transcription des premiers traits qui caractérisent la structure inhérente aux fluctuations de la concentration de polluants dans les différents environnements étudiés.

2.2 Environnements intérieurs et campagnes de mesures

Afin d'étudier la variabilité temporelle de la qualité de l'air intérieur, quatre campagnes de mesures ont été réalisées dans les environnements suivants : quatre campagnes dans un environnement réel de type espace paysager réalisées entre 2012 et 2015, une campagne dans un bureau individuel pendant l'année 2010-2011 et une campagne dans une maison expérimentale au cours de l'année 2010. Ces environnements sont situés à Champs-sur-Marne en zone périurbaine sur le site du CSTB. Les deux espaces de type bureaux sont situés dans un bâtiment à deux niveaux comportant : des bureaux, une salle de réunion, des laboratoires et une halle expérimentale.

Les environnements cibles ont été les deux types de bureaux, mais les mesures effectuées dans la maison expérimentale ont été étudiées pour comparer un espace réel à un espace avec des conditions quasi-contrôlées.

2.2.1 Description des bureaux

2.2.1.1 Bureau paysager

L'espace de bureaux (open-space) étudié, appelé aussi bureau paysager compte un nombre de postes de travail variable selon les années (entre 7 et 9) ainsi que cinq bureaux individuels séparés par des cloisons. Les bureaux individuels font partie du même espace d'une surface de $132\,m^2$ et de volume de $364\,m^3$. Quant à l'occupation, l'espace accueille entre 6 à 12 personnes, variable en fonction de l'heure, des jours et des périodes de l'année. En plus des postes fixes, l'open-space est occupé par des stagiaires à certaines périodes de l'année. Cette période de l'année correspond généralement à l'occupation la plus élevée, et par conséquent l'action sur les ouvrants pourrait être plus fréquente durant cette période. En 2013 par contre, une baisse du niveau d'occupation globale est variable selon les mois.

Pour l'espace de bureaux, une moquette couvre le sol de tout l'espace instrumenté et le mobilier est composé principalement par des bureaux en bois compact mélaminé. Le nombre de postes informatiques actifs varie avec l'occupation, mais de façon générale, au moins sept postes sont actifs en permanence

durant les heures de travail. Deux imprimantes étaient actives en 2012 et une imprimante-copieur multifonctions a été mise en fonctionnement depuis 2013.

En plus d'une ventilation par aération (ouverture des fenêtres et des portes contrôlée par les occupants), un système de ventilation simple flux sans balayage pourvoit l'ensemble de l'open-space et des bureaux individuels. Il assure un débit d'extraction d'air constant de 252 m^3/h en 2012 à 228 m^3/h en 2014. L'espace instrumenté communique à travers le reste du bâtiment *via* une seule porte qui mène à une pièce de circulation menant soit à la halle expérimentale, soit à une petite cour extérieure permettant d'atteindre le toit, soit à l'aile gauche du bâtiment. Le plan et une vue globale de l'espace de bureaux sont présentés en Annexe B.

2.2.1.2 Bureau individuel

Le bureau individuel instrumenté est dans le même bâtiment cible situé au même étage que l'espace de bureaux. La surface du bureau est de 12.5 m^2 et son volume est d'environ 31.5 m^3. Le renouvellement d'air du bureau est assuré par un système de ventilation double flux. En 2009, le taux de renouvellement d'air à l'échelle du bloc où se trouve le bureau a été estimé à 0.17 h^{-1} ($\pm 6\%$) par des mesures de gaz traceur (SF_6) avec un système de ventilation à l'arrêt et les ouvrants étaient fermés. Avec seulement le système double flux en fonctionnement, ce taux atteignait 1.26 h^{-1} ; lorsque les fenêtres étaient ouvertes, le renouvellement d'air est estimé de 4 à 5 h^{-1}. Une seule fenêtre est exposée au nord-ouest et n'est pas munie d'entrée d'air. La porte du bureau donne sur un couloir qui communique avec cinq autres bureaux individuels.

Quant à l'occupation, le bureau accueille un occupant et au plus deux personnes supplémentaires durant de très courtes périodes.

2.2.2 Description de la maison expérimentale (MARIA)

La maison expérimentale "MARIA" : Maison Automatisée pour des Recherches Innovantes sur l'Air (Ribéron & O'Kelly, 2002) est une maison construite sur le site du CSTB à Champs sur Marne (*cf* : Figure B.2.1 situé dans l'annexe). La maison possède trois niveaux (deux niveaux habitables et un sous-sol). La maison possède 5 pièces principales dont quatre chambres et un séjour, ainsi que quatre pièces techniques : cuisine, salle de bain WC, douche et cabinet d'aisance. L'ouverture et la fermeture des portes intérieures et des fenêtres sont automatisées, la perméabilité est contrôlée et le système de ventilation est modulable. La maison est construite sur une base carrée de côté extérieur 9.10 m et de hauteur totale d'environ 10.9 m (toiture comprise). La surface habitable (sans sous-sol) est d'environ 142 m^2 et le volume (sans la toiture) est d'environ de 655 m^3. Une campagne de mesure a été menée dans le séjour située en rez-de-jardin, d'une surface totale de 36 m^2 et d'un volume total de 88.4 m^3 entre mars et avril 2010.

Durant la période de mesure, le système de ventilation a été fixé en mode simple flux avec des entrées d'air au niveau des volets roulants de la partie séjour (les fenêtres sont restées closes) et des sorties à travers les bouches d'extraction situées au niveau des pièces de service. Durant cette campagne de mesure, l'occupation et les usages domestiques sont pratiquement absents. A ce titre, la maison expérimentale reflète plus les conditions d'une maison individuelle inoccupée, sans aucune ouverture ni de portes, ni de fenêtres.

Les données issues de ce type d'environnement sont utilisées ici pour étudier par comparaison, l'impact de l'activité des occupants sur le niveau de fluctuation temporelle de la concentration des polluants par le fait qu'ils sont quasi-inexistants.

TABLE 2.3.1 – Les paramètres mesurés dans différents environnements.

Paramètre	Mesure	Nom variable	Unité	Type
Ouverture des fenêtres	détecteur d'ouverture (CSTBox)	**OF**	ouvert/fermé	binaire ou nominale
Ouverture des portes		**OP**		
Détection de mouvement	capteur passif infrarouge	**Occup**	-	binaire ou nominale
Irradiance	solarimètre	**Irr**	Wm^{-2}	réel
La température	Sonde Q-Track (int) ou Station météo (ext)	**T**	°C	réel
L'humidité relative		**Hr**	%	réel
Le dioxyde de carbone		**CO_2**	ppm	réel
Les Particules	compteur optique	**PM**	$\#L^{-1}$(15 domaines de taille)	entier
Les oxydes d'azote	Analyseur MMS	**NO_2, NO**	ppb	réel
L'ozone	Analyseur MMS	**O_3**	ppb	réel
Le formaldéhyde	réaction de Hantzsch	**HCHO**	ppb	réel
La pluie	Détecteur de pluie	**Pl**	pas de pluie / pluie	binaire
La pression	Station météo (ext)	**Prss**	hPa	réel
Hydrocarbures aromatiques polycycliques totaux	Fluorescence	**HAP**	$ng.m^{-3}$	réel

2.3 Données disponibles dans chaque environnement

L'instrumentation active des environnements décrits précédemment a permis de fournir des informations plus ou moins complètes sur la qualité de l'air ambiant de ces espaces, d'abord, des mesures des concentrations des polluants, ensuite, des mesures relatives aux conditions climatiques et finalement l'influence de l'occupation et de l'état des ouvrants.

Nous disposons pour chaque environnement de séries temporelles au pas de temps plus ou moins fin (1 min à 1 h) observées sur des périodes allant de quelques jours à une année. Les différents paramètres mesurés sont listés dans le Tableau 2.3.1.

Chaque série est enregistrée selon le type de la variable considérée (dernière colonne du Tableau 2.3.2) et par différents instruments ou capteurs.

Les données de formaldéhyde ont été enregistrées par un appareil (Analyseur AL4021, A*erolaser*) capable de fournir les concentrations du formaldéhyde à l'échelle du ppb et au pas de temps d'une minute.

Les concentrations en nombre de particules par litre d'air prélevé ont été mesurées par fraction de taille (diamètre en μm) en continu par un capteur optique de particules (Dust Monitor 1.108, *Grimm*). Les données des concentrations des particules représentées par des séries temporelles pour 15 classes de taille réparties entre 0.3 et > 20 μm).

En complément de la mesure en continu des deux polluants cibles, que sont le formaldéhyde et les particules, plusieurs informations complémentaires on été recueillies en continu sur les différents micro-environnements : en l'occurrence, l'état des ouvrants et de l'occupation. Ces deux paramètres permettent d'appréhender les principales causes des fluctuations réelles de la QAI.

La concentration de dioxyde de carbone (CO_2) est souvent mesurée dans un environnement clos pour déterminer le niveau de confinement de l'air ambiant et en déduire ainsi les conditions d'aération et pour qualifier la ventilation des locaux. La concentration du CO_2 reflète en grande partie l'activité métabolique de l'occupant et elle est considérée comme un bon traceur des bio-effluents humains. Le CO_2 et la mesure de l'état d'occupation constituent dès lors la paire indicative la plus fiable sur la présence.

Néanmoins, on peut distinguer des situations où l'espace clos est vacant, mais la concentration de CO_2 reste élevée du fait des apports des pièces voisines occupées. Par ailleurs, le niveau du CO_2 enregistré dépend non seulement de la production métabolique liée à l'occupation mais également du taux de renouvellement de l'air.

Les modules de détection communiquent avec un dispositif (CSTBox) qui permet de collecter et de contrôler les données d'un bâtiment associé à un réseau de capteurs, d'actionneurs ou d'interacteurs. Lorsqu'aucun mouvement n'est détecté, aucune information n'est renvoyée à la CSTBox et dès qu'un mouvement est localisé par l'un des modules, la quantité enregistrée (pendant 10 secondes) est renvoyée. Les données liées à la quantité de mouvement sont transformées en données binaires avec un pas de temps d'une minute.

Les portes et les fenêtres ont été instrumentées et les données relatives à l'état des ouvrants (fenêtres et portes) ont été enregistrées par la CSTBox. Les modules de détection d'ouverture d'ouvrants et les détecteurs de mouvements sans fil ont été associés à la CSTBox. Les données enregistrées par la CSTBox sont des séries temporelles à pas de temps irrégulier : les modules de détection renvoient des informations dès qu'un changement d'état survient. Un prétraitement a été effectué pour synchroniser toutes les séries chronologiques au même pas de temps.

Notons que seul l'espace paysager a été instrumenté pour le recueil des données relatives à l'état des ouvrants. Quant à la détection de mouvement, elle a été mesurée dans le bureau individuel et dans l'open-space. La conversion de la quantité de mouvement en une variable binaire et l'état des ouvrants en une variable nominale nous permet de constituer deux variables aléatoires exogènes qui pourraient permettre d'expliquer des caractéristiques de la variabilité des paramètres de la QAI.

La température, le CO_2 et l'humidité relative sont mesurées à l'intérieur par la sonde Q-Track (TSI Inc.) toutes les minutes.

Pour les paramètres extérieurs, une station météorologique permanente située sur le toit du bâtiment cible permet d'enregistrer les valeurs de la température, de l'humidité relative, de la pression atmosphérique, de l'irradiance solaire, de la vitesse et de la direction du vent de façon automatique. Elle détecte également les évènements pluvieux. En outre, les concentrations des polluants dans l'air extérieur ont été mesurées sur des phases épisodiques, notamment pour les particules et pour le formaldéhyde en 2015.

Le récapitulatif des données disponibles dans chaque environnement est donné dans le Tableau 2.3.2.

TABLE 2.3.2 – Récapitulatif des données disponibles. Les notations des environnements **BI**, **OS12**, **OS13**, **OS14**, **MARIA** se réfèrent aux campagnes de mesures effectuées dans le Bureau Individuel (**BI**), dans l'Open-Space en 2012 (**OS12**), dans l'Open-Space en 2013 (**OS13**), dans l'Open-Space en 2014 (**OS14**), dans l'Open-space en 2015 (**OS15**) et dans la maison expérimentale (**MARIA**), respectivement. L'état des ouvrants dans l'espace de bureaux en 2012 a été exprimé par le nombre de minutes durant lesquelles au moins une fenêtre $\left(\mathrm{OF}^{\#}\right)$ ou une porte $\left(\mathrm{OP}^{\#}\right)$ est ouverte pendant une heure.

Environnement	Polluant		Paramètre climatique		Occupation-Ouvrant	Période	Pas de temps
	Intérieur	**Extérieur**	**Intérieur**	**Extérieur**			
BI	CO_2, PM, HAPs	-	-	-	Occup	2010-2011	de 1 à 10 min
OS12	CO_2, $PM_{[0.35-20]}$	$PM_{[0.35-8.75]}$	T, Irr, Hr	T, Irr, Dv, Vv, Hr	Occup, $OF^{\#}$, $OP^{\#}$	27/01-30/06	Horaires
OS13	HCHO, CO_2, O_3, CO, NO_x	- -	T, Hr Irr,	T, Hr, Pl Dv, Vv, Pa	Occup, OF, OP	Avril-Aout 2013	1-min
OS14	HCHO,CO_2, $PM_{[0.35-20]}$	HCHO,CO_2, $PM_{[0.35-20]}$	T, Hr, Irr	T, Irr, Dv, Vv, Hr,Pl,Pa	Occup, OF, OP	Janvier-Décembre	
OS15	HCHO, CO, O_3, CO_2 $PM_{[0.35-20]}$	HCHO, CO_2 $PM_{[0.35-20]}$	T, Irr, Hr	T, Hr, Irr, Dv, Vv, Pa, Pl	Occup, OF, OP	Janvier-Juin	
MARIA	HCHO, PM, CO_2	PM, CO_2	T, Hr	T, Hr, Dv, Vv, Pa, Pl	-	Mars-Avril 2010	

Face à ce flux d'informations, plusieurs questions se posent à nous de prime abord. La première s'intéresse à la spécificité de l'ensemble des environnements par rapport aux fluctuations des concentrations en polluants : dans quelle mesure la variabilité temporelle peut relever de la spécificité de chaque environnement ? Inversement, on peut s'intéresser aux facteurs déterminants qui rendent la série temporelle plus au moins *prévisible :* quels sont les facteurs qui influencent le degré de "prédictibilité" de la série temporelle ?

On classe les déterminants en facteurs endogènes au système (source → concentration) d'un côté, et en facteurs exogènes à ce système de l'autre côté, représentant l'occupation, l'activité des occupants (N'ayant pas la possibilité d'obtenir l'activité des occupants, nous supposons que l'occupation implique une certaine activité) et les conditions climatiques. La deuxième question s'intéresse à la manière de quantifier ces interactions : d'une part entre les facteurs et les composantes du système, et, d'autre part entres les composantes elles-même.

2.4 Influence du type des données sur le choix des modèles

L'acquisition des données en temps réel permet de mieux mettre en évidence la variabilité temporelle du phénomène étudié. Néanmoins, les paramètres de l'acquisition d'un suivi réel d'une variable aléatoire déterminent le degré de description du phénomène. Par exemple, est-il nécessaire d'avoir deux ans de mesures des concentrations du CO_2 en pas de temps d'une minute dans un espace occupé par une seule personne pour permettre de caractériser la production métabolique de l'occupant ?

Face à cette question, plusieurs facteurs d'influence doivent être mis en évidence dans le cadre de l'analyse des séries temporelles. Notamment, le pas de temps utilisé, la durée de la campagne de mesures et éventuellement la méthode de lissage utilisée.

2.4.1 La résolution temporelle

Un évènement de forte amplitude se déroulant sur une très courte période nécessite un temps de réponse suffisamment fin pour le mettre en évidence. Le pas de temps pourrait alors ne pas être adapté à de nombreuses situations de sauts abrupts. Une agrégation temporelle sur un pas de temps supérieur à la fluctuation a pour effet de lisser fortement le signal et gommer certaines variations très fortes.

Pour certaines variables, le pas de temps doit être suffisamment fin pour permettre une bonne description temporelle de l'ensemble des phénomènes qui ont généré cette variable. En revanche, pour d'autres variables, choisir un pas de temps très fin augmente le degré de "stochasticité" de la variable d'intérêt, et par conséquent, augmente le niveau d'imprévisibilité.

Quoi qu'il en soit, sans connaissance *a priori* du phénomène étudié, nous optons pour la première approche car on peut toujours ramener au deuxième cas par un lissage ou un prétraitement quelconque. Par exemple, il existe deux phénomènes qui déterminent le niveau du CO_2 : l'occupation et l'aération ; une résolution temporelle fine n'est recommandée que pour synchroniser la base de données sur la même échelle de temps. Un lissage de type moyenne mobile est suffisant pour reproduire un niveau de variabilité acceptable. Outre le fait que l'acquisition des données en continu relative à l'aération est quasi-impossible, le phénomène d'échange d'air intérieur-extérieur est très complexe et dépend de plusieurs facteurs dont principalement l'intervention humaine sur les ouvrants.

Or, pour le formaldéhyde ou les particules fines, les connaissances sur l'ensemble des interactions des phénomènes qui génèrent ces variables nécessitent un pas de temps fin pour espérer mettre en évidence

ces phénomènes et détecter les moindres fluctuations. Un lissage comme la moyenne mobile pour ce type de variable crée un artéfact lié à l'autocorrélation.

2.4.2 La durée et la longueur des séries

La durée de la série temporelle est une donnée très importante pour avoir un aperçu sur l'étendu temporel du phénomène étudié. La longueur de la série permet de déceler les caractéristiques inhérentes au système, telles que la tendance ou la saisonnalité. Ainsi, ce que l'on peut prendre pour une tendance pourrait n'être que le début d'un phénomène périodique. Ces deux éléments sont liés aux choix de la résolution temporelle utilisée.

2.5 Statistiques et analyse de la variabilité temporelle

2.5.1 Fluctuations dans le bureau individuel

2.5.1.1 Variabilité du dioxyde de carbone (CO_2) et de l'occupation

Les niveaux de concentration en CO_2 ont été mesurés pendant un an, du 28 Octobre 2010 au 27 Octobre 2011 avec un pas de temps de 10 minutes pendant ($n = 52\,560$ observations). La série temporelle des concentrations du CO_2 est représentée dans la Figure 2.5.1a. L'état d'occupation du bureau est mesuré simultanément avec les concentrations du CO_2. Cette variable est codée en binaire par 0 (inoccupation) et par 1 (occupation). La Figure 2.5.1b donne la représentation temporelle des occurrences d'occupation au cours du temps.
La concentration de CO_2 mesurée durant l'année 2011-2012 dans le bureau individuel variait entre 331 à 1457 ppm avec un coefficient de variation $c_v = 18.18\%$. Le Tableau 2.5.1 reprend quelques statistiques globales relatives à la concentration du CO_2 et en fonction de l'état d'occupation.

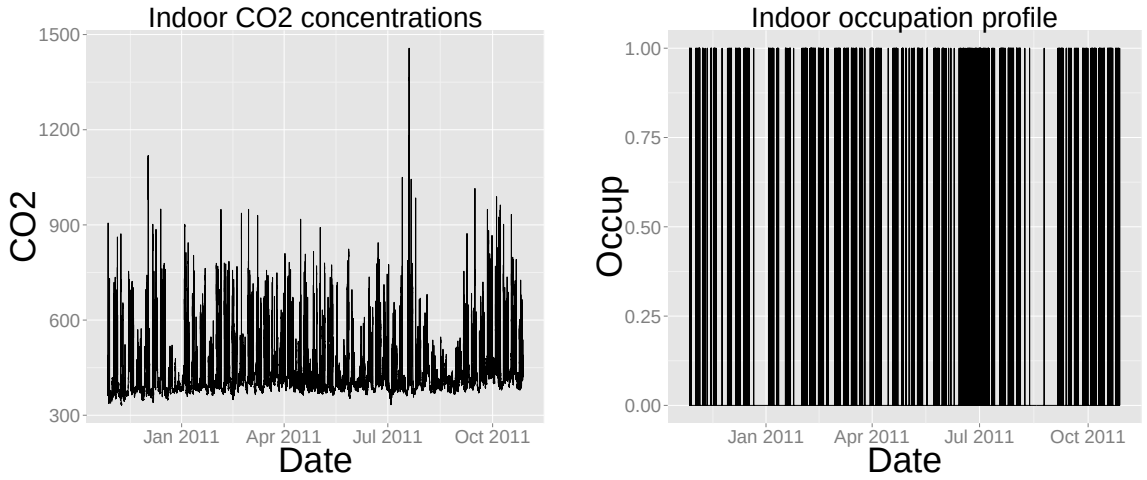
TABLE 2.5.1 – Statistiques globales (quelque soit l'état de l'occupation) et par état d'occupation des niveaux des concentrations en CO_2 dans l'air du bureau individuel.

CO_2		n	$\bar{x}$	σ	médiane	Min	Max	$Skew$	$Kurtosis$
	global	52560	433	82	406	331	1457	2.7	10.74
état	0	46232	412	45	400	331	1288	2.62	16.47
	1	6328	590	112	579	333	1457	1.14	3.78

La variabilité diurne dégage un profil horaire marqué par deux pics et une partie stable (Figure 2.5.2a). En effet, on distingue deux phases de fluctuation :

- $[20\,h - 8\,h]$: les box-plots sont quasi-plats avec des concentrations qui varient entre 380 à 550 ppm ;
- $[7\,h - 19\,h]$: très forte variation avec une forme à double-pics (modes à 11 et 15 h) pouvant atteindre les 1450 ppm ; l'accroissement de l'écart type durant cette plage horaire est de 25% par rapport à la variation exprimée sur la journée ;

La variabilité des concentrations dépend en partie de l'état de l'occupation de la pièce. En effet, en période d'occupation, l'écart type de la concentration du CO_2 est de l'ordre de 113 ppm ($c_v = 19\%$), tandis que dans le cas d'inoccupation, l'écart-type est de l'ordre de 45 ppm ($c_v = 10.8\%$). Sur la Figure 2.5.2b, nous essayons de mettre en perspective les propriétés de la distribution de probabilité en fonction de l'occupation. Globalement, la distribution en cas d'occupation est plus aplatie, tandis que la distribution lorsque l'environnement est vacant est très leptokurtique et étalée vers la droite. Les valeurs extrêmes apparaissent plutôt dans le cas d'occupation avec des probabilités très faibles. Ces propriétés exprimées en termes de coefficient d'aplatissement (kurtosis) et de coefficient d'asymétrie (skewness) sont présentées dans le Tableau 2.5.1.



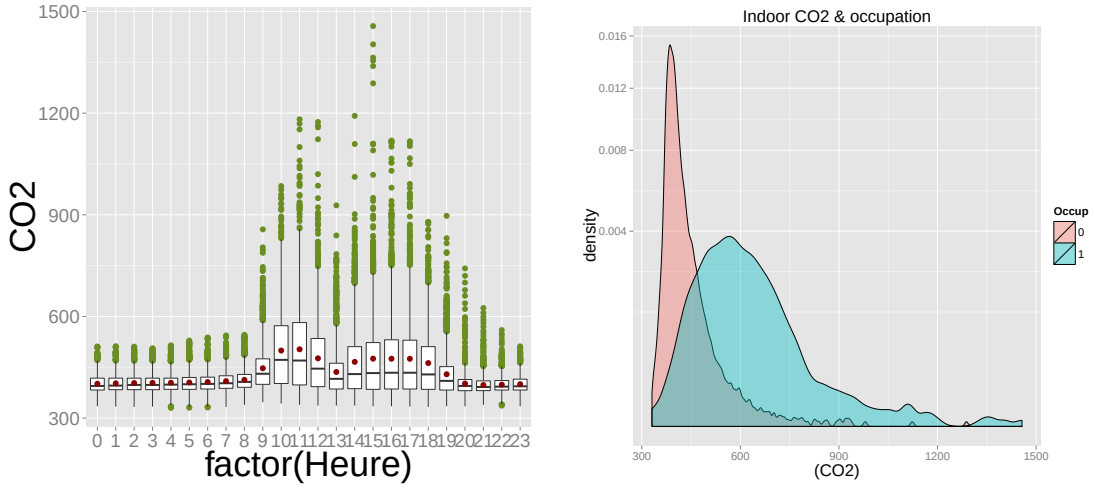
(a) *Concentrations du CO_2 en [ppm] observées sur une période d'une année avec un pas de temps de 10 minutes.*

(b) *État d'occupation du bureau.*

FIGURE 2.5.1 – *Séries temporelles des concentrations du CO_2 (a) et l'état d'occupation (b) observées dans le bureau individuel pendant une année au pas de temps de 10 minutes.*

Les recommandations ayant pour sujet les valeurs limites de la concentration du CO_2 dans l'air intérieur dépendent de plusieurs paramètres, dont le taux de renouvellement d'air et le nombre d'occupants. Néanmoins, si on se réfère aux recommandations de l'ASHRAE, une concentration de CO_2 dépassant 1000 ppm peut entraîner une sensation d'inconfort. Pour un bureau individuel, occupé à 17 % durant toute l'année (week-ends exclus), une valeur de 1000 ppm est signe d'une mauvaise efficacité de la ventilation sur ces périodes. Les valeurs du CO_2 qui dépassent 1000 ppm représentent 0.5% du temps d'occupation, ce qui reste marginal.Dans de rares cas, la concentration dépasse 1000 ppm (1288 ppm à 15 h) alors que le local était vacant. Ce cas atypique laisse suggère la présence d'autres sources (occupants) dans les pièces voisines avec la porte du bureau laissée ouverte.

En ce qui concerne l'occupation, sur les 17% d'occupation totale, le mercredi est le jour le moins représentatif de tous les jours ouvrés, avec 11.5% de temps de présence. Par ailleurs, le mois d'août représente



(a) Profile diurne de la concentration en CO_2en [ppm] .

(b) Densité de probabilité des niveaux de CO_2 par état d'occupation.

FIGURE 2.5.2 – *Niveaux de CO_2 par heure sur une année (a) et densité de probabilité par l'état d'occupation dans un bureau individuel (b). Les points à l'intérieur des boîtes représentent la moyenne de tous les jours par heure et la taille de chaque boîte (coupé par la médiane) renseigne sur le niveau de variabilité diurne (a). L'inoccupation est codée par* 0 *et l'occupation par* 1 *(b).*

que 2.5% du temps d'occupation contre une occupation supérieure à 10% pour les mois de mai, juillet et octobre. Le temps de présence pour le reste des mois variait entre 6% et 9%. Ces taux fournissent une information sur la représentativité temporelle de la source d'occupation : les occupants et leurs activités.

2.5.1.2 Variabilité des particules en suspension (PM)

La concentration en nombre de particules de taille entre 0.3 et 20 μm a été mesurée au pas de temps d'une minute dans le bureau individuel du 21-07-2010 au 04-04-2011. Durant cette période, plusieurs interruptions de l'appareil ont empêché d'avoir une série de mesures complète. Sur 371457 valeurs ($\times$ 15 variables), seules 251843 observations (valides) ont été enregistrées par le capteur, soit environ 32% de valeurs manquantes pour chaque taille de particules. Les valeurs de quelques fractions pour les particules de taille moyenne entre 0.57 et 1.3 μm.sont beaucoup plus touchées.

Pour l'analyse descriptive, nous utilisons 10 fractions (0.35 ; 0.45 ; 1.8 ; 2.5 ; 3.5 ; 4.5 ; 6.25 ; 8.75 ; 12.5 et 17.5 μm) en raison de la validité de ces données. Pour les questions exploratoires : séparations des sources et prévisions, nous utilisons l'ensemble des15 fractions de particules ; la concentration du CO_2 et les paramètres extérieurs au pas de temps de 10 minutes sur la période allant de février 2011 jusqu'à avril 2011. Ces données sont fiables et ne nécessitent aucun prétraitement.

Entre juillet 2010 et début avril 2011, la concentration médiane des particules fines dans l'espace du bureau individuel variait de 8230 #.L^{-1} pour les particules de 0.35 μm à 10 #/L pour les particules de taille moyenne (2.5 μm). Pour les grosses particules (entre 12.5 et 17.5 μm), la concentration moyenne est de l'ordre de 0.4 #/L. Le niveau de fluctuation exprimé par le coefficient de variation (c_v) allait de 150% pour les fines et supérieur à 300% pour les moyennes, ces valeurs ne semblent pas être très pertinentes pour mesurer la variabilité des séries car on note une forte asymétrie ; par conséquent la

moyenne (arithmétique) n'est pas représentative de la distribution. En effet, le coefficient d'asymétrie est positif pour toutes les tailles de particules : les queues de distribution sont étalées vers la droite.

Sur les Figures 2.5.3 et 2.5.4 nous essayons de vérifier l'existence d'un profil horaire et hebdomadaire-type qui se dégagerait pour les particules de diamètre 0.35 et 2.5 μm. En effet, pour la fraction comprise entre 2.5 et 6.25 μm, la distribution diurne montre une tendance journalière marquée par de très fortes variabilités durant les heures du jour et par une moyenne légèrement plus élevée à 11 h. En termes d'écart-type, le profil diurne type présente deux pics de forte variabilité, le premier est vers 11 h et l'autre vers 15 h pouvant atteindre des valeurs maximales de 158×10^3 #/L durant ces heures. Bien que les données soient très influencées par les valeurs extrêmes, la variabilité diurne de la concentration de particules est fonction de leur diamètre.

En outre, les niveaux pour la fraction supérieure à 2.5 μm ont tendance à être plus faibles les week-ends et moins fluctuants que pour les autres jours de la semaine. En revanche, pour les fines particules, il est moins évident de déceler l'effet jour. Remarquons tout de même que pendant les jours de semaine, les concentrations pour les particules de 0.35 μm variaient entre 149×10^3 #/L et 157×10^3 #/L, tandis qu'elles n'atteignaient pas 107×10^3 #/L le samedi et 55×10^3 #/L le dimanche.

La distribution mensuelle ($\simeq$ 8.5 mois) des fluctuations montre que durant le mois de mars, les valeurs enregistrées sont les plus élevées et leur variabilité est la plus prononcée (*cf :* Figure 2.5.5). Bien qu'on n'ait pas à disposition toute l'année des mesures afin de mettre en évidence les différences qu'on aurait pu voir durant l'année 2011, des indications sur la distribution mensuelle observée en 2012 (bureau paysager Section 2.5.3) montrent que les résultats entre les deux campagnes (2011-2012) sont corrélés pour les mois observés : la variation mensuelle des particules de 0.35 μm est moins prononcée que celle des particules de 2.5 μm.

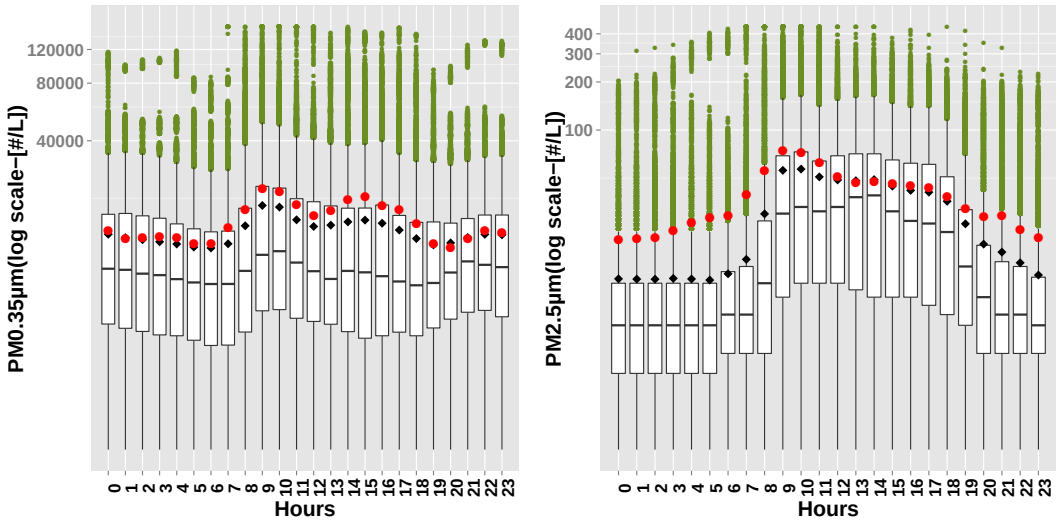


FIGURE 2.5.3 – *Variabilité diurne de la concentration en nombre de particules dans le bureau individuel durant la période du 21-07-2010 au 04-04-2011. Exemple pour les particules de diamètre médian de 0.35 μm (à gauche) et 2.5 μm (à droite). Les valeurs sont exprimées en nombre de particules par litre ($\#L^{-1}$). Le symbole rond rouge représente l'écart-type calculé pour chaque heure durant toute la période et le symbole en losange noir, la moyenne horaire.*

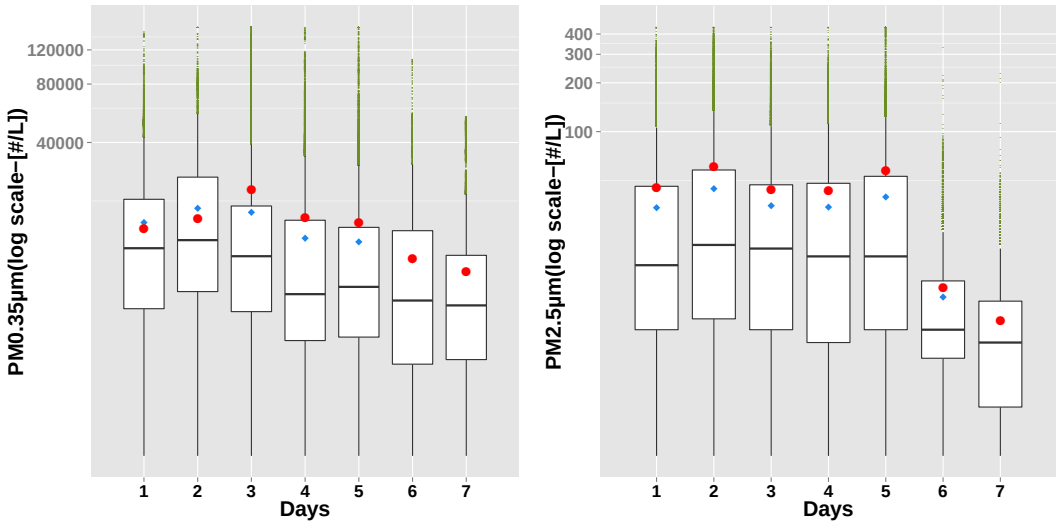


FIGURE 2.5.4 – *Distribution par jour des concentrations en nombre de particules dans le bureau individuel durant la période du 21-07-2010 au 04-04-2011. Exemple pour les particules de diamètre médian de* $0.35\,\mu m$ *(à gauche) et* $2.5\,\mu m$ *(à droite). Les valeurs sont exprimées en nombre de particules par litre* $(\#L^{-1})$*. Le symbole rond rouge représente l'écart-type calculé pour chaque jour durant toute la période et le symbole en losange noir, la moyenne par jour. Les jours sont représentés de 1 à 7 correspondent à la semaine de lundi jusqu'à vendredi.*

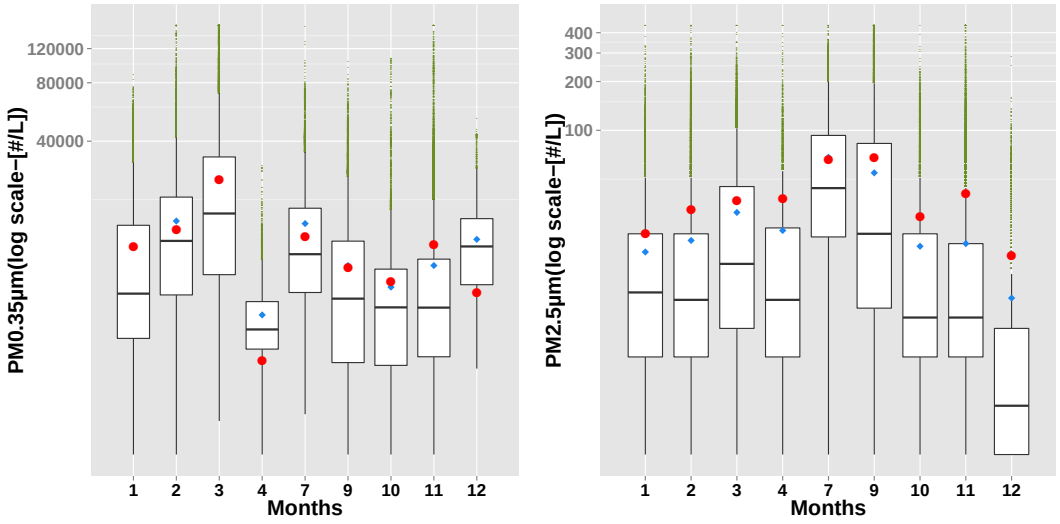


FIGURE 2.5.5 – *Distributions mensuelles des concentrations de particules dans le bureau individuel (20-07-2010 au 04-04-2011). Exemple pour les particules de diamètre médian de* $0.35\,\mu m$ *et* $2.5\,\mu m$*. Les valeurs sont exprimées en nombre de particules par litre* $(\#L^{-1})$*. Les mois sont représentés par 1 (janvier) à 12 (décembre). Note : (janvier,* $n = 44640$*) (février,* $n = 34407$*), (mars,* $n = 44569$*), (avril,* $n = 5760$*), (juillet,* $n = 13966$*), (septembre,* $n = 35208$*), (octobre,* $n = 19422$*), (novembre,* $n = 42625$*), (décembre,* $n = 11246$*).*

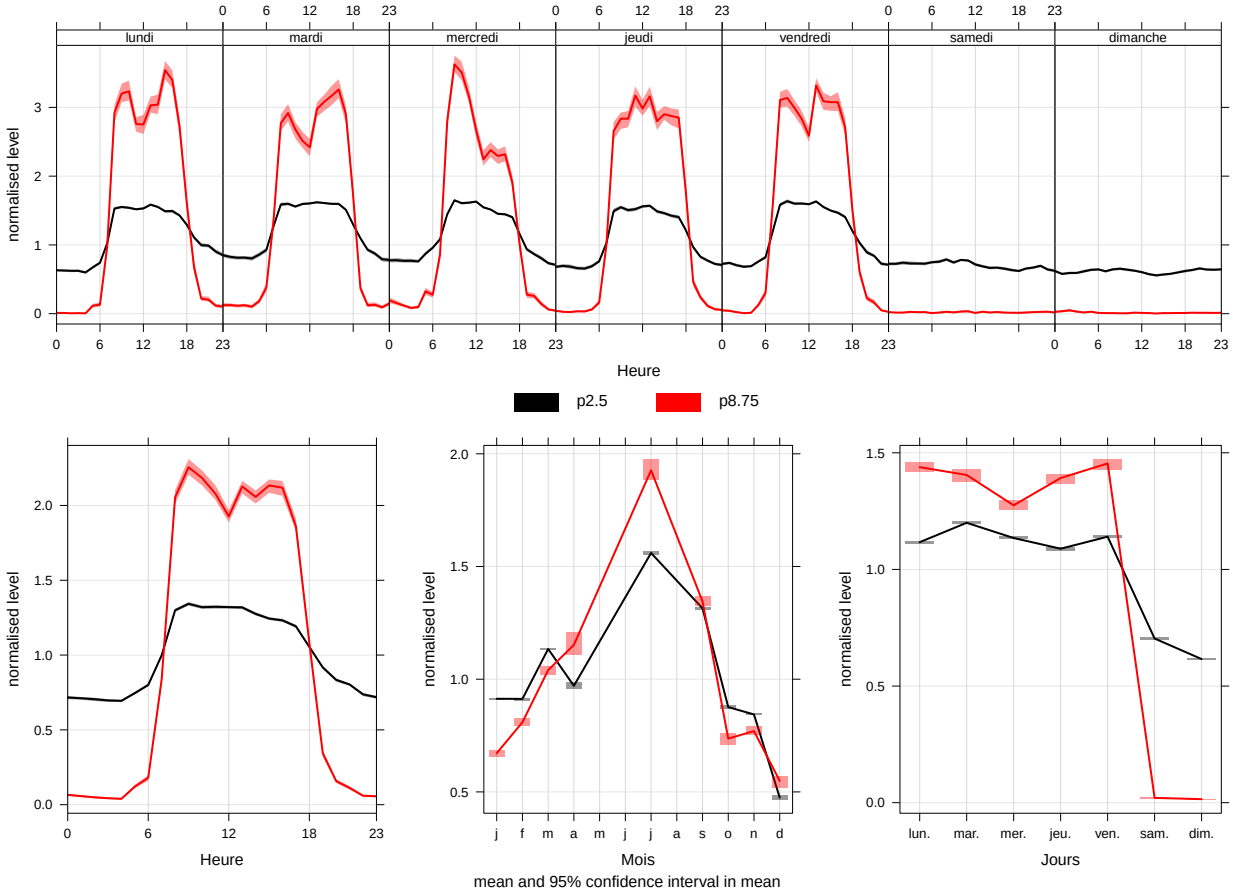


FIGURE 2.5.6 – *Récapitulatif de la variabilité de certaines fractions de particules moyennes : p2.5=2.5 μm et p8.75=8.75 μm. Les données sont exprimées en logarithmes et translatées par une particule/L. La normalisation est effectuée par la moyenne de la concentration sur la période.*

De façon générale, la contribution du mois de mars pour la détermination de la variabilité des particules fines intérieures est très comparable avec les contributions de ce mois pour les concentrations massiques des $PM_{10}(\mu g.m^{-3})$ extérieures. Pour cette comparaison, nous avons récupéré les données horaires de PM_{10} pour la période allant de 2009 à 2013 enregistrées à la station de Lognes (*cf.* Annexe C.2). *Le profil saisonnier de la concentration en PM_{10} extérieure est similaire avec la distribution mensuelle de la concentration des fines particules intérieures.*

Le récapitulatif de la variabilité horaire, journalière, hebdomadaire et mensuelle des fractions de particules de taille 2.5 μ.m et 8.75 μ.m est donné dans la Figure 2.5.6. Les profils de variabilités de ces fractions sont mis en évidence sur les différentes échelles temporelles : horaire, diurne, hebdomadaire et mensuel. Cette Figure (*cf.* 2.5.6) montre en outre une particularité remarquable : la variabilité des particules de

taille 2.5 μ.m est plus élevée que la variabilité des particules de tailles 8.75 μ.m durant les heures creuses (de 18 jusqu'à 6) et les week-ends ; tandis que durant les heures d'occupations, cette relation est inversée par un facteur de 2. On peut expliquer ces observations par le fait que durant la période d'inoccupation, les particules de taille 8.75 μ.m se déposent sur les surfaces et uniquement durant les heures d'occupation que ces particules sont remises en suspension par l'activité des occupants.

Enfin, la distribution diurne par mois montrée dans la Figure 2.5.7 dévoile l'aspect sinusoïdal de la variation des particules de taille moyenne. En effet, pour les mois de septembre à décembre : une forme en cloche est plus prononcée durant les heures de la journée. Cette caractéristique commence à disparaître au mois de janvier jusqu'à l'inversion totale de la courbe au mois de mars. Autrement dit, les valeurs les plus élevées sont observées durant les heures de nuit avec une médiane stable durant les heures de la journée. Ceci pourrait être attribué au fait que durant la fin du trimestre, la remise en suspension de particules déposées sur les surfaces est déterminée par le comportement de l'occupant durant les heures de travail. Mais pour le mois de mars, il très difficile d'avancer des hypothèses physiques sur ce type de variation diurne, d'autres informations sont nécessaires pour identifier l'origine de cette variabilité.

Par ailleurs, pour les particules de taille $< 0.35\,\mu$m et hormis le mois de décembre, le profil moyen diurne par mois montre très peu de régularité horaire par rapport à la distribution mensuelle. Cette irrégularité renseigne sur l'importance du caractère aléatoire porté par les concentrations de particules fines. Nous allons confirmer cette hypothèse dans le chapitre dédié à la mesure de prévisibilité et la dimension fractale des séries temporelles. Les implications de cette irrégularité dans le contexte de prévision sont de grande importance.

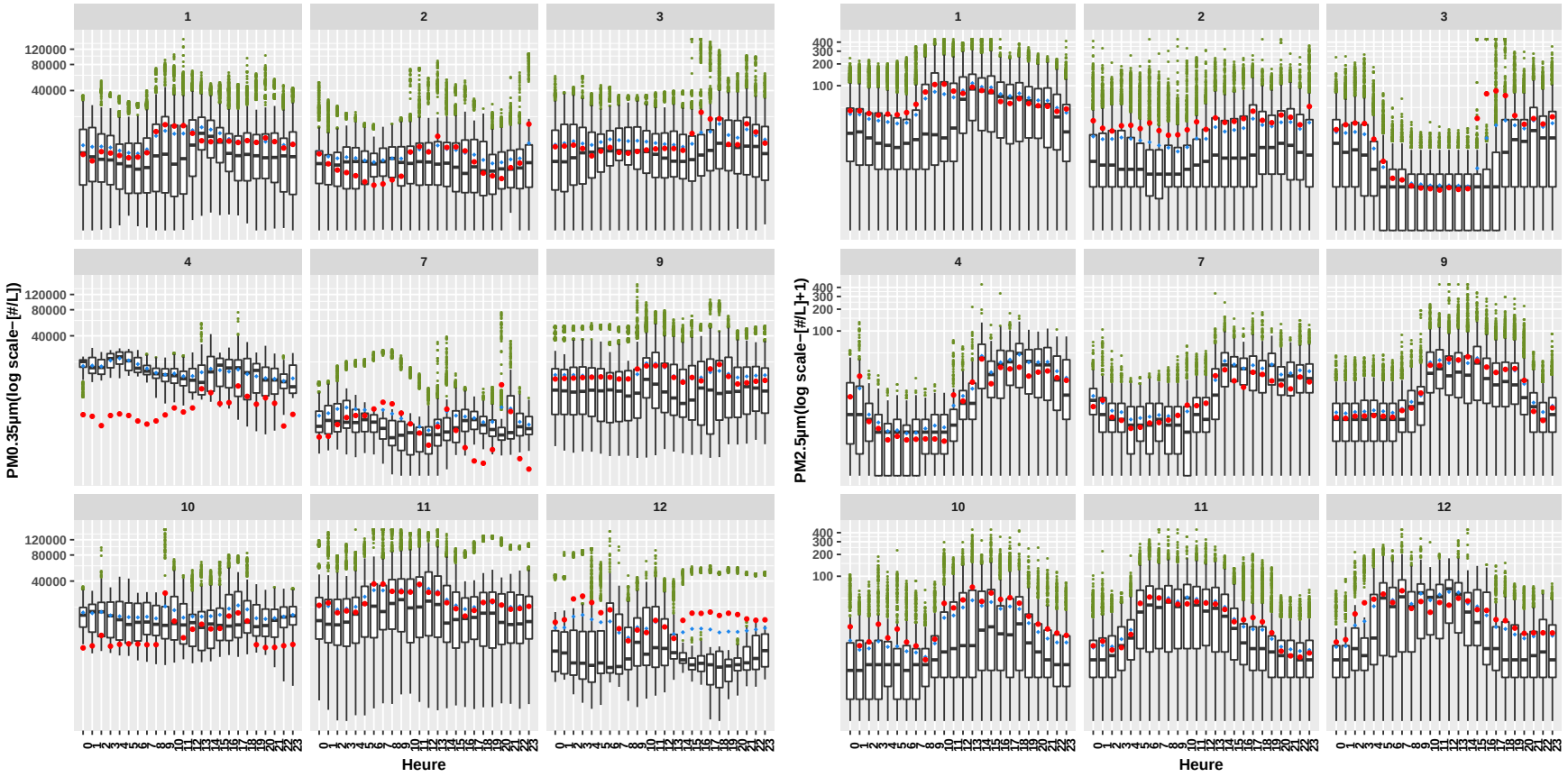


FIGURE 2.5.7 – *Distribution diurne par mois de la concentration en nombre de particules (1 pour Janvier, 2 pour Février, etc.) dans le bureau individuel. Exemples de particules de diamètre médian de* $0.35\,\mu m$ *et* $2.5\,\mu m$ *. Les valeurs sont exprimées en nombre de particules par litre* ($\#L^{-1}$). *Le symbole rond rouge représente l'écart-type calculé pour chaque heure de chaque mois, le symbole losange bleu, la moyenne horaire de chaque mois.*

2.5.2 Fluctuations dans la maison expérimentale

2.5.2.1 Variabilité du Formaldéhyde (HCHO)

Les niveaux de la concentration en HCHO dans la maison expérimentale ont été mesurés durant la période du 14/04/2010 au 03/05/2010 au pas de temps d'une minute ($n = 23809$ observations). Nous rappelons ici que ces données sont issues d'une expérimentation étendue sur plusieurs phases dans le cadre d'une évaluation de la stratégie de chauffage (convecteur électrique et chauffage d'appoint) sur la qualité de l'air intérieur [rapport interne CSTB]. La période que nous avons choisi d'étudier correspond à la phase où aucune de ces deux stratégies de chauffages n'a été mise en fonctionnement : la phase de référence.

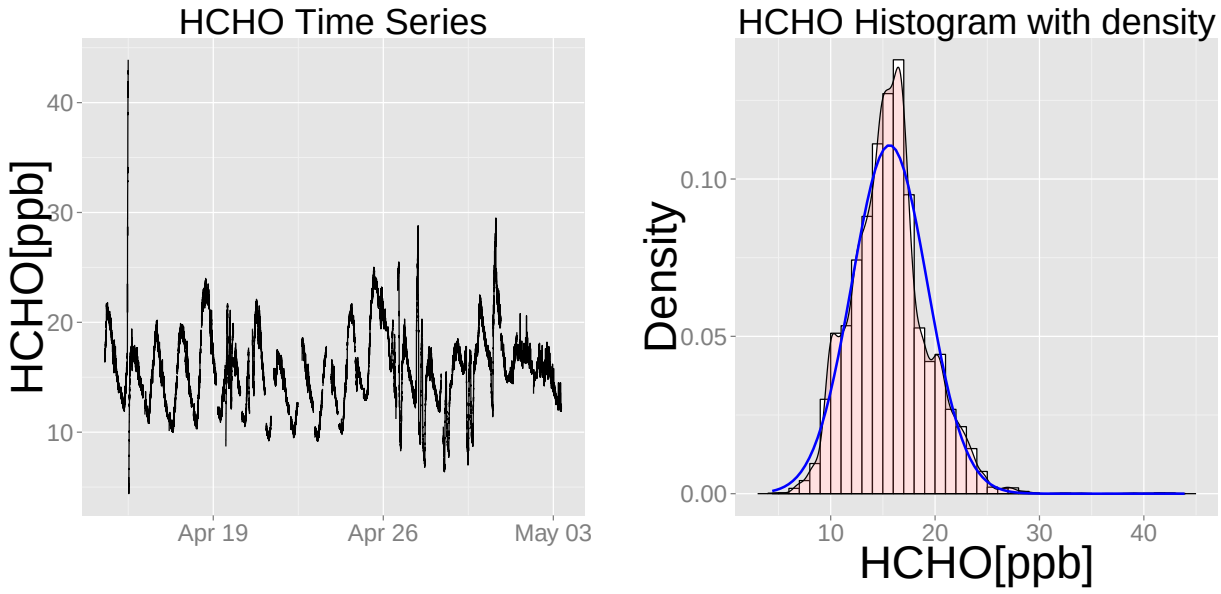


FIGURE 2.5.8 – *Fluctuation de la concentration du HCHO dans la maison expérimentale. Les mesures couvrent la période du 14/04/2010 au 03/05/2010 au pas de temps d'une minute (à gauche) ; l'histogramme associé et les densités de probabilité estimées : la partie pleine par la méthode du noyau (kernel) et la courbe bleu est l'estimation par la loi normale (à droite).*

La Figure 2.5.8 montre la série temporelle des concentrations du formaldéhyde ainsi que la densité de probabilité associée. La concentration moyenne observée est de 15.51 ppb, le coefficient de variation estimé est de 23% ; quant aux valeurs de centiles p_1 et p_{95} ils sont respectivement 8.4 ppb et 24.6 ppb. L'histogramme estimé montre une distribution presque symétrique (*skewness*= 0.63) et similaire à la loi sécante hyperbolique [1] (*kurtosis* = 2.31). Cette loi partage plusieurs propriétés avec la loi normale ; néanmoins comme nous le montre la figure 2.5.8, l'histogramme estimé par la méthode de noyau est plus leptokurtique que la loi normale. Ceci peut être traduit par le fait que le processus générateur de ces

1. La fonction de masse de la loi sécante hyperbolique est $f(x) = \frac{1}{2}\text{sech}\left(\frac{\pi}{2}x\right)$ et la fonction de répartition est donnée par $F(x) = \frac{\pi}{2}\arctan\left[\exp\left(\frac{\pi}{2}x\right)\right]$.

données est un groupement de facteurs engendrant 95% des données autour de leur moyenne. Dans ce cas et contrairement aux particules dans le bureau individuel, la moyenne et le coefficient de variation sont des indicateurs descriptifs "fiables".

Durant les treize premiers jours, la dynamique du HCHO semble suivre un comportement sinusoïdal avec des amplitudes de variation plus au moins grandes. La dernière semaine (du 27/04 au 03/05), la dynamique est perturbée par de fortes variations durant lesquelles la concentration en HCHO passe de 15 ppb à 8 ppb, puis remonte jusqu'à atteindre les 30 ppb. On note par ailleurs une hausse accrue de la concentration atteignant un maximum de 44 ppb suivie par une chute brutale pour descendre à 4.4 ppb. Ce mouvement est observé pendant seulement quelques minutes. Cette hausse (de 65% par rapport à la moyenne) est due à la combustion d'une bâtonnet d'encens dans le séjour sur une courte période suivi par une ouverture des fenêtres.

La Figure 2.5.9 montre le profil diurne et la distribution hebdomadaire des concentrations de HCHO durant les 19 jours de mesure. De façon générale, le profil diurne pour l'ensemble de la période présente une allure quasi-périodique avec une variabilité plus importante durant les heures de la journée. Quant à la distribution hebdomadaire, aucun profil apparent ne peut être distingué.

Au premier abord, ces observations nous renseignent sur quelques facteurs mis en jeu dans la détermination de la variabilité du formaldéhyde en l'absence d'occupant. En effet, dans la maison expérimentale, les paramètres environnementaux sont les seuls facteurs fluctuants permettant d'expliquer les fluctuations de la concentrationdu formaldéhyde dans l'air intérieur. Du fait de l'absence de modification des conditions opératoires (aération et occupation) , très peu de fluctuations abruptes ont été observées et la dynamique temporelle est quasi-déterministe.

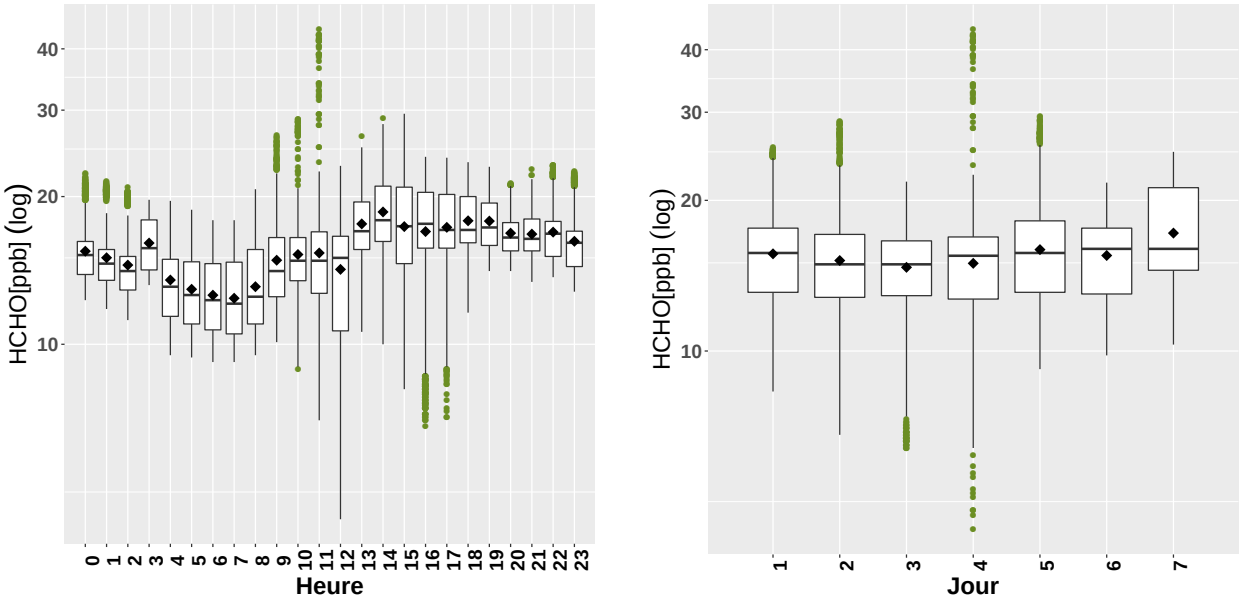


FIGURE 2.5.9 – *Profil diurne et distribution hebdomadaire de la concentration en formaldéhyde dans la maison expérimentale (MARIA). Les mesures couvrent une période allant de 14/04/2010 au 03/05/2010*

2.5.2.2 Variabilité des particules en suspension (PM)

Les concentrations en nombre de particules intérieures de taille entre 0.3 et 20 μm (15 gammes) et extérieures de taille entre 0.3 et 2.5μm (8 gammes) ont été mesurées au pas de temps d'une minute dans la maison expérimentale (MARIA) de 29/03/2010 au 26/04/2010. Durant cette période, un chauffage d'appoint a été actionné plusieurs fois pendant une semaine, du 06/04/2010 au 12/04/2010.

En outre, pour 8 gammes de particules, la différence entre la concentration extérieure et intérieure a été calculée pour chaque instant : $\Delta \mathrm{pm}_i\,(t) = \mathrm{C}_{i(ext)}(t) - \mathrm{C}_{i(int)}(t)$, $i = 0.35, \ldots, 2.5\,\mu$m.

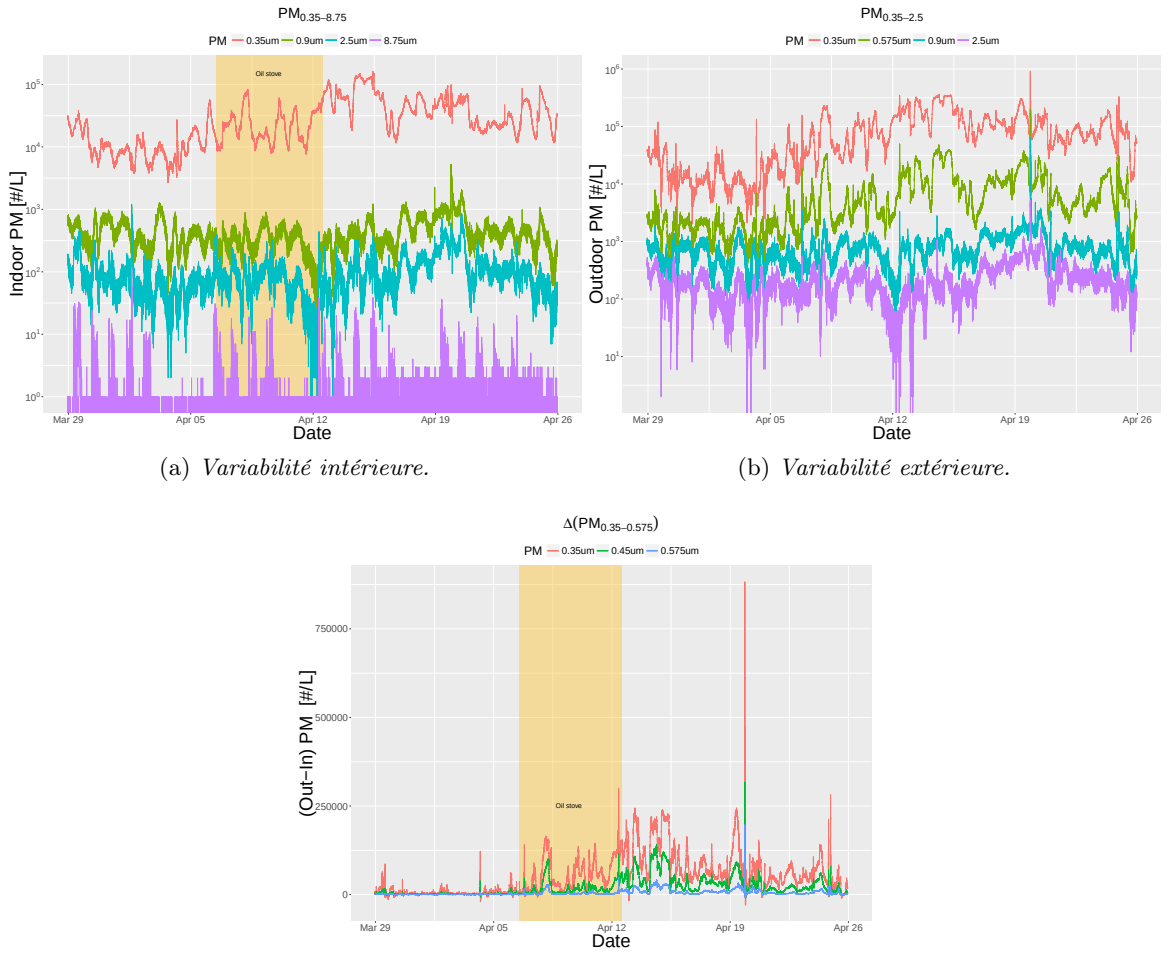
La Figure 2.5.10 montre les séries temporelles de la concentration (en $\#.\mathrm{L}^{-1}$) de quelques fractions de particules intérieures, extérieures et de la différence entre extérieur et intérieur. Pour les particules intérieures, il apparaît que l'utilisation du chauffage par poêle à pétrole ne montre pas d'effet significatif sur le niveau de concentration en particules.

La Figure 2.5.10c illustre la variabilité de la différence des concentrations des particules fines. Globalement, sur toute la période de mesure, l'écart extérieur-intérieur est positif pour les particules fines. Bien que les fenêtres soient fermées durant les phases expérimentales, les fluctuations des concentrations des particules fines intérieures sont très corrélées avec les concentrations extérieures de particules de mêmes taille. Ceci illustre l'importance des sources extérieures sur la concentration en particules fines dans l'air intérieur.

Le pic observé le 19-04 à 21 :54 dans l'air extérieur pour presque toutes les fractions n'a pas été reproduit -à une proportion près- dans les concentrations intérieures. La raison est probablement due au fait que cette fluctuation est très brève, elle dure moins de 10 min, ce qui a eu pour effet une dispersion rapide dans l'air extérieur. Une autre raison pourrait justifier ces observations : les ouvrants de maison expérimentale étaient fermés et les conditions météorologiques favorisaient la dispersion.

En revanche, il semblerait que certains pics extérieurs de concentrations fines particules influenceraient l'air intérieur avec un décalage temporel d'environ 15 minutes. Par exemple, le 04-04 à 5h51, les concentrations extérieures et intérieures de la fraction médiane 0.35 μm étaient respectivement de 33.5×10^3 #/L et 6.5×10^3#/L ; une augmentation d'ordre de 300% (134×10^3 #/L et 26×10^3 #/L) a été enregistrée pour les deux environnements après 5 min et 15 min, respectivement. Notons aussi qu'après ce pic, la vitesse de décroissance des PM extérieures est beaucoup plus élevée que la vitesse de décroissance des PM intérieures pour toutes les fractions : les conditions de dispersion extérieures sont plus favorables.

La distribution diurne de la concentration intérieure en particules de taille 0.35 μm montre une baisse de la valeur médiane entre 10h et 18h (*cf.* Figure 2.5.11a). Les concentrations maximales ont été observées principalement entre 11h et 12h. Quant aux particules intérieures de taille 2.5 μm, les niveaux de fluctuation sont plus élevées durant les heures de la journée.

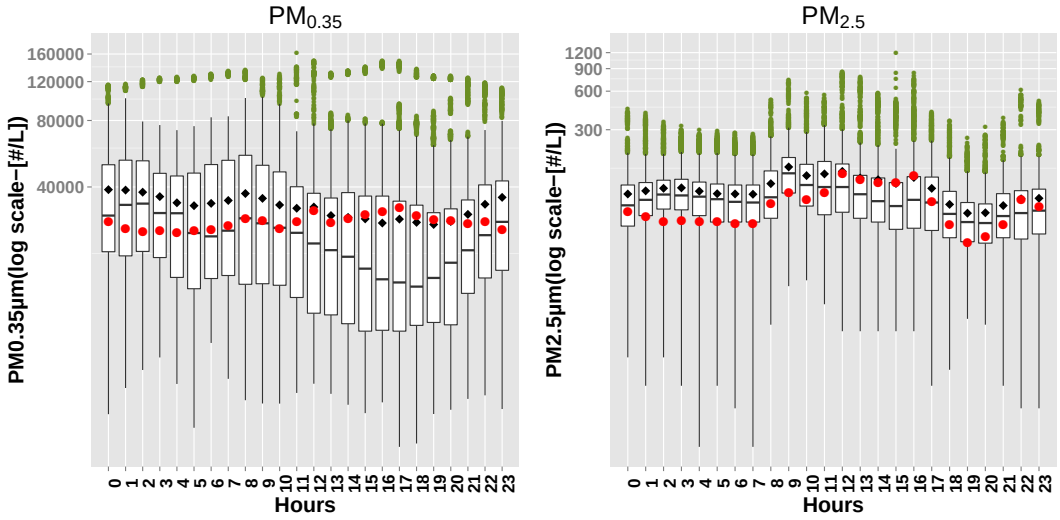


(a) *Variabilité intérieure.*

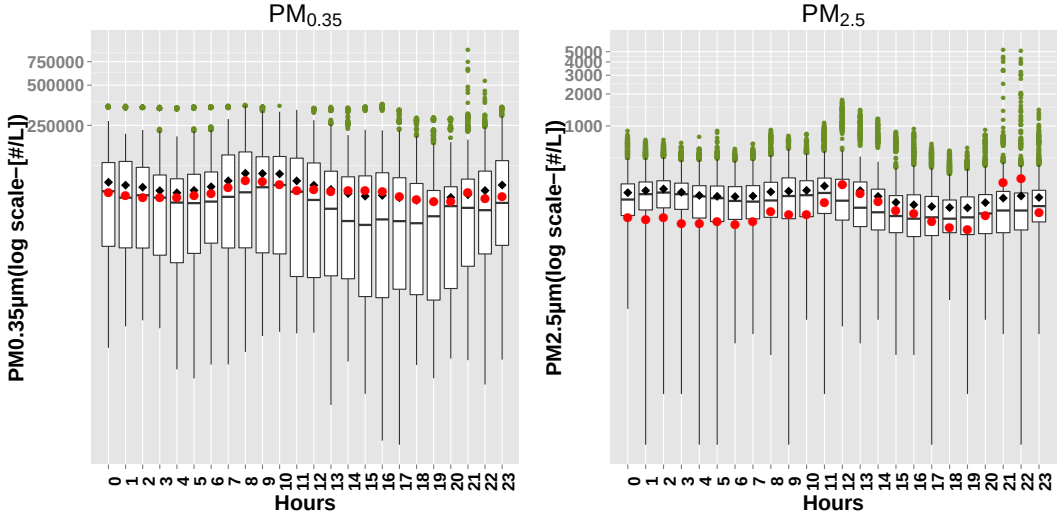
(b) *Variabilité extérieure.*

(c) *Variabilité de la différence entre extérieur et intérieur.*

FIGURE 2.5.10 – *Séries temporelles de quelques fractions de particules intérieures, extérieures et de la différence entre la concentration extérieure et intérieure en #/L mesurées en 2010 dans la maison expérimentale au pas de temps d'une minute. Les valeurs de concentrations dans la Figure 2.5.10a et dans la Figure 2.5.10b sont exprimées en Logarithme décimal* ($\log_{10}$). *La période ombrée (jaune) correspond à la semaine où l'utilisation de la poêle à pétrole a été en fonction pendant de courtes périodes (oil stove).*



(a) *Distribution diurne de la concentration en particules des fractions* $0.35\,\mu m$ *et* $2.5\,\mu m$ *extérieures. Les valeurs sont exprimées en échelle logarithmique par* $\#/L$.

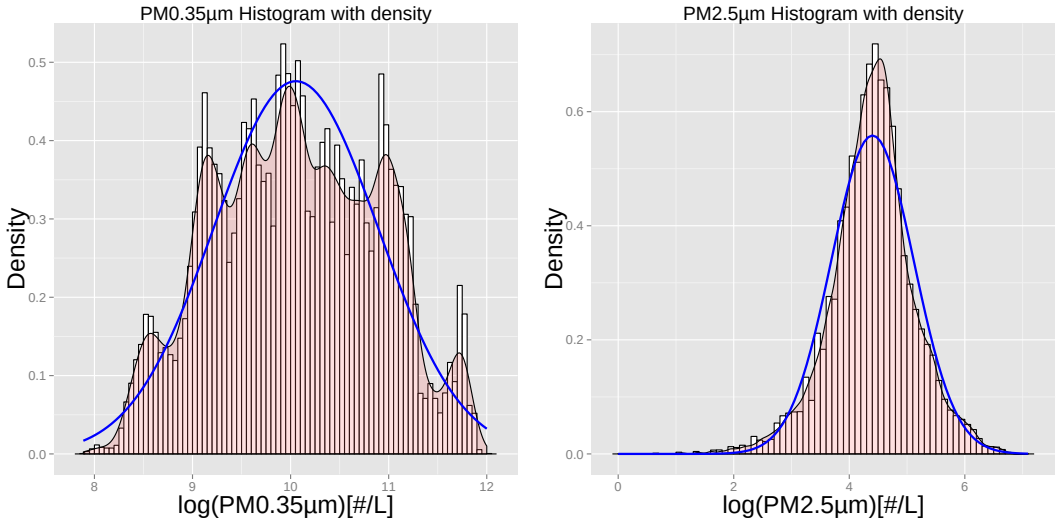


(b) *Distribution diurne de la concentration des fractions* $0.35\,\mu m$ *et* $2.5\,\mu m$ *extérieures. Les valeurs sont exprimées en échelle logarithmique par* $\#/L$.

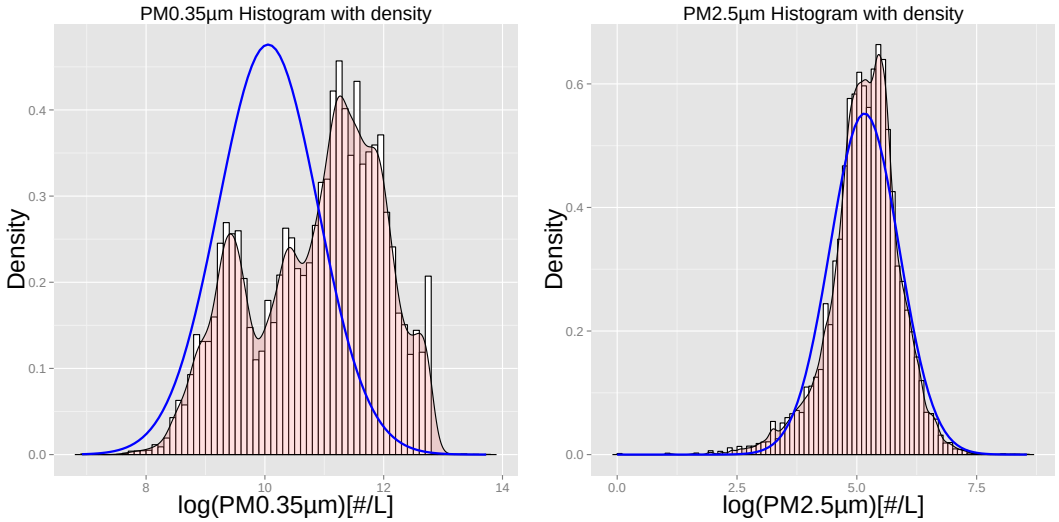
FIGURE 2.5.11 – *Profils diurnes des concentrations des particules de taille* $0.35\,\mu m$ *et* $2.5\,\mu m$ *observées simultanément en air intérieur dans la maison expérimentale et en air extérieure. La période de mesure couvre du 29-03-2010 au 26-04-2010 en pas de temps d'une minute. Le symbole losange noir représente la moyenne horaire, le point rouge représente l'écart-type horaire.*

Par contre, la variabilité diurne des variables extérieures ne révèle aucune forme particulière qui suggère l'existence d'une source majeure qui serait en mesure de l'expliquer (*cf.* Figure 2.5.11b). Deux raisons principales peuvent expliquer cette constatation : premièrement, les contributions relatives des sources extérieures sont comparables (aucune ne se démarque à une heure spécifique) et deuxièmement, l'étendue de la campagne de mesure et la taille de l'échantillon qui lisse les irrégularités. La Figure 2.5.11 montre

que les concentrations de particules fines sont beaucoup plus fluctuantes que les concentrations des particules de tailles moyennes.



(a) *Densité de probabilité des concentrations des particules à l'intérieur.*



(b) *Densité de probabilité des concentrations des particules à l'extérieur.*

FIGURE 2.5.12 – *Densité de probabilité de la concentration des particules à l'intérieur de la maison expérimentale et à l'extérieur, de tailles* $0.35\,\mu m$ *et* $2.5\,\mu m$*. Les mesures couvrent la période entre le 29-03-2010 et le 26-04-2010 au pas de temps d'une minute et les valeurs sont exprimées en logarithme.*

Les histogrammes des particules en suspension (intérieures et extérieures), l'estimation de leurs densités de probabilités ainsi que l'ajustement de la loi normale sont présentés dans la Figure 2.5.12. Pour les particules de tailles moyennes ($2.5\,\mu m$), la loi normale s'ajuste assez bien à la transformée logarithmique des données : les concentrations des PM moyennes suivent une loi de type log-normale. En revanche,

pour les particules fines ($0.35\,\mu m$), seules la distribution des concentrations intérieures semblent avoir une forme proche d'une log-normale, la densité de probabilité des niveaux de particules extérieures était étalée à gauche (la distribution décalée à droite de la médiane).

2.5.2.3 Les paramètres climatiques

Les paramètres climatiques mesurés à l'intérieur de la maison expérimentale et correspondant à la période qui s'étale du 01/03/2010 au 30/04/2010 avec un pas de temps d'une minute sont listés dans le tableau 2.3.2. .

Dans ce paragraphe, nous présentons uniquement les fluctuations de la température et de l'humidité spécifique intérieure. Les autres paramètres sont exposés dans l'Annexe C.2.1.

A partir de la mesure de l'humidité relative, de la température et de la pression, l'humidité spécifique est calculée en utilisant la formule de Rankine pour approximer la pression de vapeur saturante nécessaire au calcul.

$$H_{abs}\left(\frac{g}{kg}\ \text{air humide}\right) = \frac{Hr}{100} \times \frac{M_{\text{eau}}}{M_{\text{air}}} \times \exp(13.7 - \frac{5120}{273+T}) \times 1000 \tag{2.5.1}$$

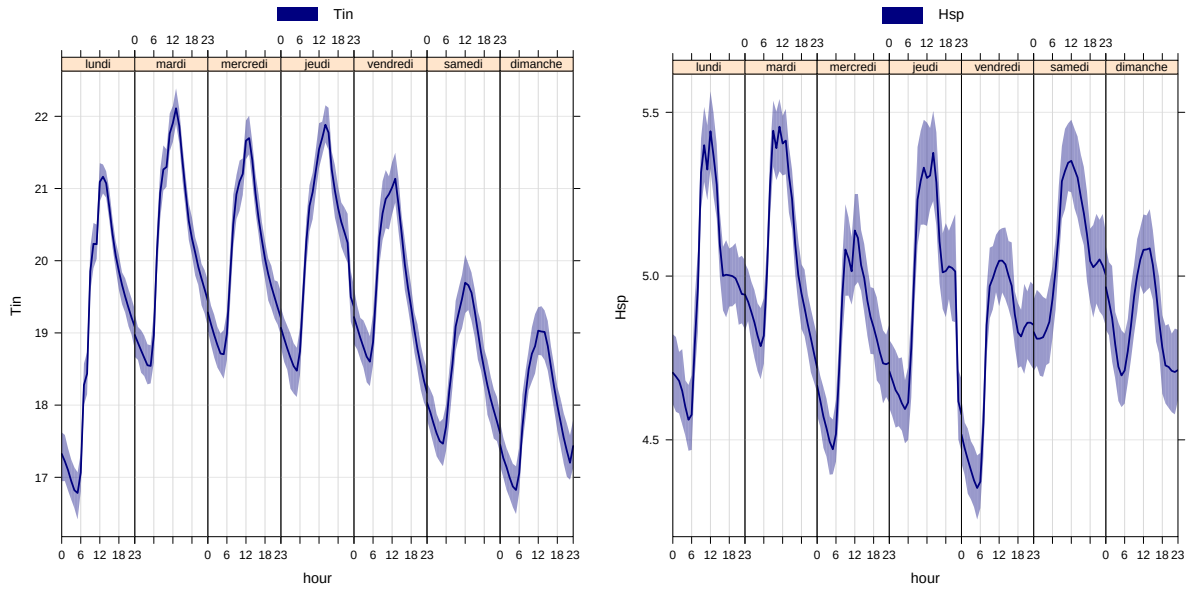
$$Hsp\left(\frac{g}{kg}\ \text{air sec}\right) = \frac{H_{abs}}{(1000 - H_{abs})} \times 1000. \tag{2.5.2}$$

La Figure 2.5.13 présente le profil de variation d'une semaine type et la distribution horaire des variables température et humidité spécifique intérieures.

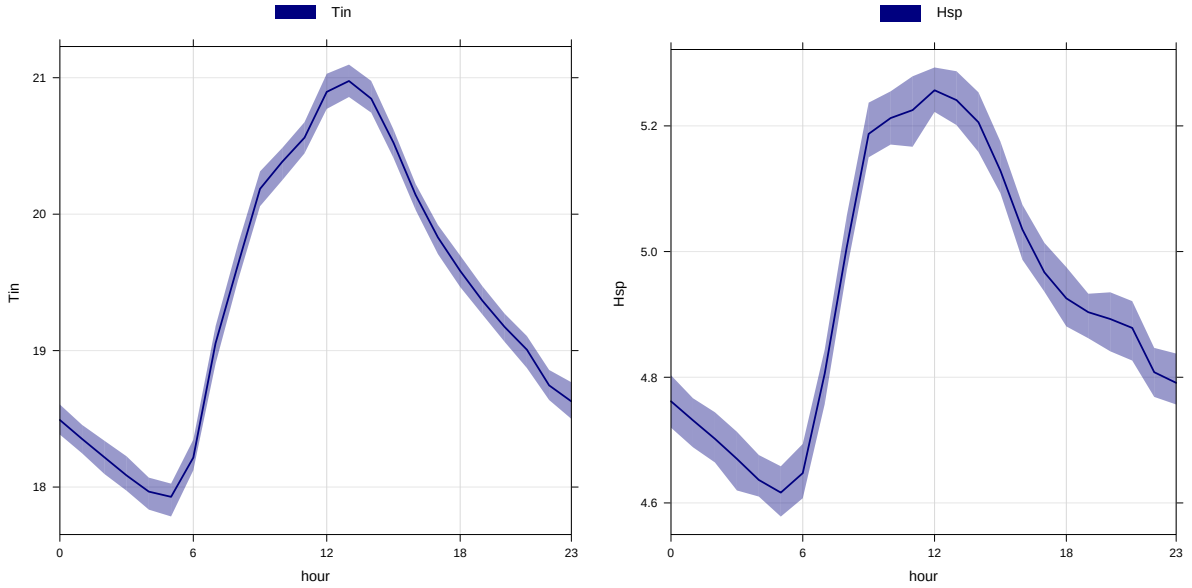
Durant toute la période de mesure, la température variait entre 10 °C et 27 °C et l'écart-type horaire variait entre 3.6 °C et 4 °C. Quant à la variabilité des fluctuations de l'humidité relative intérieure, elle était faible durant les heures de travail (*cf.* Figure C.2.2 en Annexe C.2.1).

Notons que durant le weekend, la température intérieure ne dépasse pas les 20 °C, on pourrait penser que les conditions extérieures influencent sensiblement le niveau de température intérieure par échange thermique. La distribution horaire des deux variables est quasi-sinusoïdale avec des valeurs maximales entre $11\,h$ et $14\,h$.

Afin de mettre en évidence l'existence d'une relation entre plusieurs variables de différents types, par exemple entre une variable scalaire avec une variable angulaire, le traitement statistique est différent et se base sur des techniques d'estimation qui leurs sont propres. L'une repose sur l'analyse de la statistique standard où l'espace $\mathbb{R}^n$, $n \in \mathbb{N}$ de l'échantillon est numérique et l'autre sur la statistique directionnelle où l'espace échantillon est une variété différentielle (Chikuse, 2012; Dryden & Kent, 2015). Dans le cadre de traitement des données d'un paramètre environnemental associé à la direction et à la vitesse du vent, il est nécessaire de construire un cadre qui permettrait de concilier les différents types de variables dans un même modèle. On se propose d'utiliser la représentation de la variabilité polaire de ces variables sur le plan $\mathbb{R}^2$. Elles représentent dès lors les caractéristiques des variables circulaires. La Figure 2.5.14 montre la variabilité de la température, de l'humidité et de la vitesse du vent en coordonnées polaires.



(a) *Semaine type des variables température intérieure (Tin) et humidité spécifique (Hsp) intérieure.*



(b) *Profil diurne des variables température intérieure (Tin) et humidité spécifique intérieure (Hsp).*

FIGURE 2.5.13 – *Profil de variabilité des paramètres d'ambiances (la température et l'humidité spécifique) moyennes mesurées dans l'air l'intérieur de la maison expérimentale.*

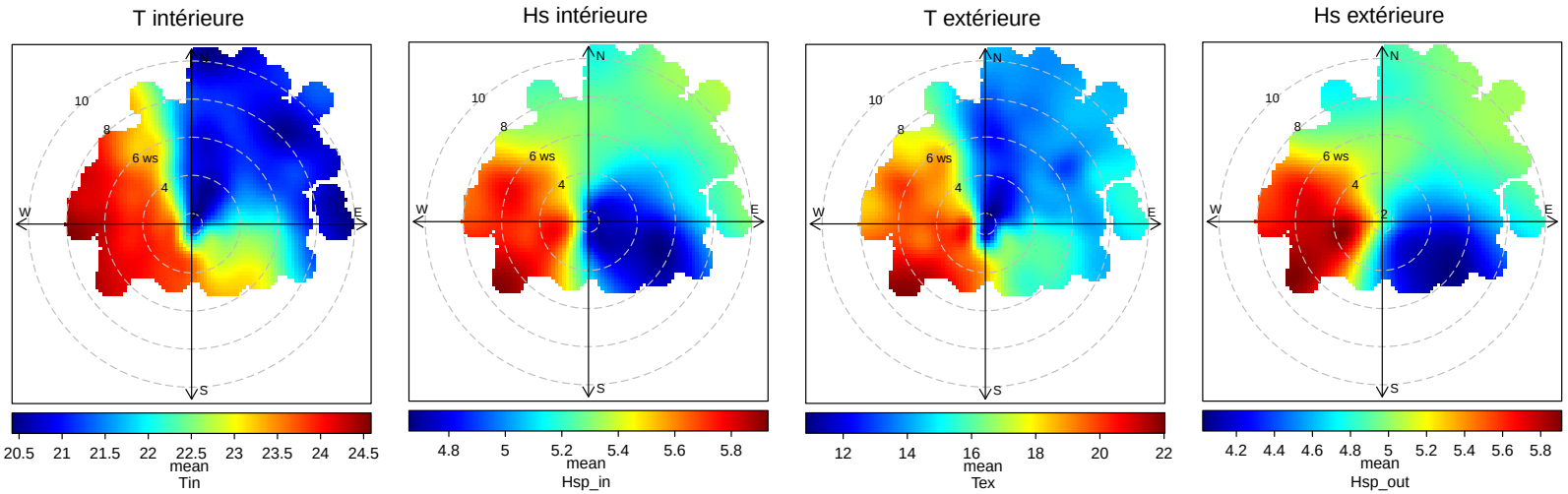


FIGURE 2.5.14 – *Distribution polaire de la température et de l'humidité relative (intérieures et extérieures) en fonction de la direction et de la vitesse du vent par la moyenne de chaque secteur.*

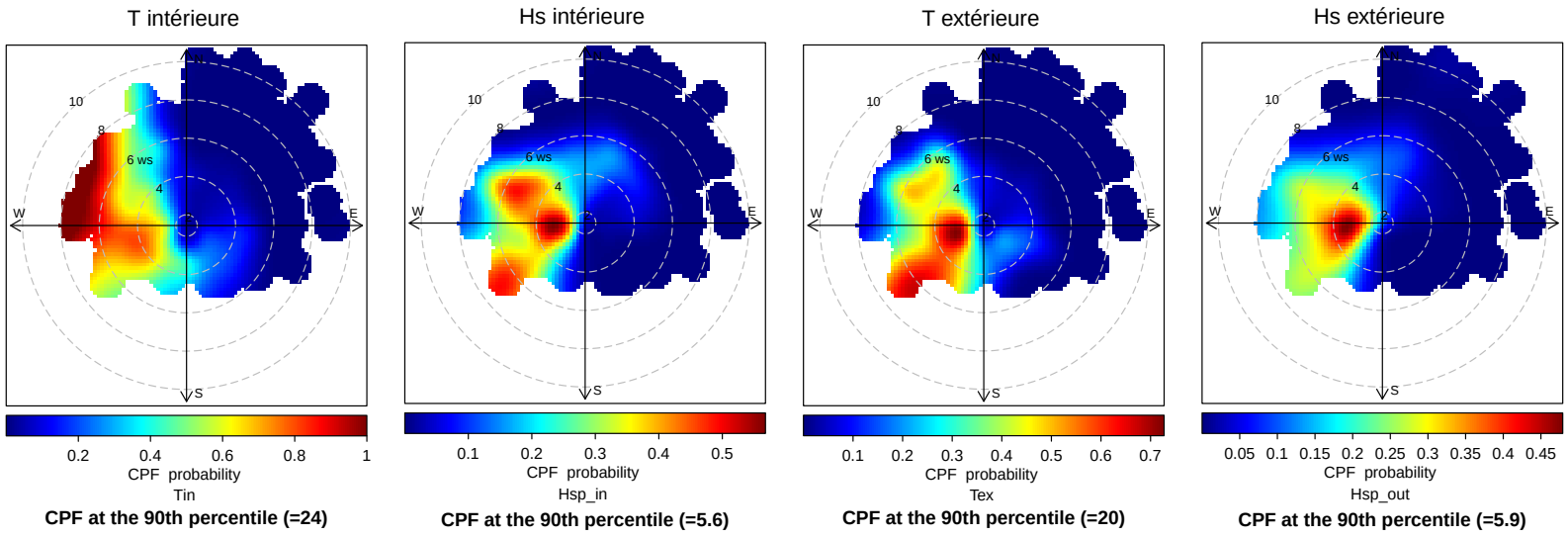


FIGURE 2.5.15 – *Distribution polaire de la température et de l'humidité relative (intérieures et extérieures) en fonction de la direction et de la vitesse du vent par la fonction de probabilités conditionnelles.*

Ces graphiques sont obtenus par un modèle d'interpolation (type Generalized Additive Model (GAM)) avec les techniques de régression non-paramétrique. On peut se référer à (Yu et al., 2004; Westmoreland et al., 2007) pour un exemple d'application sur les méthodes d'interpolation dans le cas des variables angulaires, au travaux de (Carslaw et al., 2006; Carslaw & Beevers, 2013) pour plus de détails sur la construction graphique et classification ou bien à (Batschelet et al., 1981; Fisher, 1995; Jammalamadaka & Sengupta, 2001; Mardia & Jupp, 2009; Pewsey et al., 2013) sur les fondements théoriques de la statistique circulaire.

La construction des graphes polaires bidimensionnels se fait comme suit :

- Données de la vitesse et de la direction du vent (ρ, θ) comme entrées ;
- Partitionnement de la direction et la vitesse du vent en secteurs. Généralement, le partitionnement en 10° pour θ et en 30 intervalles pour ρ permet une résolution suffisante des fluctuations de concentration. On calcule par la suite les indicateurs statistiques (moyenne, médiane, écart-type, et probabilités conditionnelles) d'un autre paramètre (température, concentration d'un polluant ...) pour chaque secteur ;
  Les composantes du vent sont alors $u = \bar{u} \cdot \sin\left(\frac{2\pi}{\theta}\right)$ et $v = \bar{u} \cdot \cos\left(\frac{2\pi}{\theta}\right)$ avec $\bar{u}$ la moyenne de ρ et θ est la direction du vent en degrés ;
- Application d'un modèle d'interpolation pour décrire la concentration C en fonction des paramètres du vent. Un cadre flexible comme l'interpolation par GAM (Wood, 2006) peut accomplir cette tâche. Le modèle GAM pour la concentration d'un polluant s'écrit comme une relation non-linéaire d'une certaine fonctionnelle S des covariables x_{ij} : $\sqrt{C_i} = \beta_0 + \sum_{j=1}^{n} S_j(x_{ij}) + \varepsilon_i$ avec C_i est la concentration du $j^{ème}$ polluant, β_0 est sa concentration moyenne, $S_j(x_{ij})$ est une fonction non-paramétrique des covariables x_{ij} et e_i est le $i^{ème}$ résidu. Dans cadre de notre étude, le modèle choisi est le suivant :

$$\sqrt{C_i} = \beta_0 + S(u, v) + \varepsilon_i, \tag{2.5.3}$$

  avec C fonction des paramètres du vent par une certaine fonction non-paramétrique S.

Il est possible d'obtenir le niveau de grille associé aux différentes variables environnementales par la fonction des probabilités conditionnelles (Ashbaugh et al., 1985). Cette fonction estime la probabilité que la valeur de la variable considérée dépasse un certain seuil pour le secteur du vent correspondant. Elle est définie comme :

$$\mathrm{CPF}_{\Delta\theta} = \frac{m_{\Delta\theta} \mid C_i \geq th}{n_{\Delta\theta}}, \tag{2.5.4}$$

avec $m_{\Delta\theta}$ le nombre de l'échantillon dans le secteur $\Delta\theta$ des concentrations C ayant une valeur supérieure ou égale à un seuil th, et $n_{\Delta\theta}$ est le nombre total des échantillons du même secteur $\Delta\theta$. Dans ce mémoire, la valeur seuil pour laquelle on définit les probabilités conditionnelles est posée au $90^{ème}$ centile.

L'équation 2.5.4 se généralise facilement pour les probabilités jointes. Ainsi, la concentration d'un paramètre environnemental définit par rapport au seuil s'obtient conditionnement à l'observation jointe des paramètres (ρ, θ). Pour une distribution bivariée, $\mathrm{CBPF}_{\Delta\theta,\Delta\rho}$ (Conditional Bivariate Probability Function) est donnée par :

$$\mathrm{CBPF}_{\Delta\theta,\Delta\rho} = \frac{m_{\Delta\theta,\Delta\rho} \mid C_i \geq th}{n_{\Delta\theta,\Delta\rho}}, \tag{2.5.5}$$

avec $m_{\Delta\theta,\Delta\rho}$ est nombre d'échantillons dans le secteur $\Delta\theta$ pour un intervalle $\Delta\rho$ de la vitesse du vent des concentrations C ayant une valeur supérieure ou égale à un seuil th, et $n_{\Delta\theta,\Delta\rho}$ le nombre total d'échantillons pour un intervalle de vitesse-direction $(\Delta\theta, \Delta\rho)$.
Dans la Figure 2.5.15, on représente d'abord la variabilité (en moyenne) de quatre paramètres climatiques en coordonnées polaires, ensuite on représente les distributions polaires de ces paramètres par la fonction de densité de probabilités conditionnelles. Les valeurs les plus élevées de la température ($> 19.5\,°C$ à l'extérieure et $> 23.5\,°C$ à l'intérieure) et de l'humidité spécifique correspondent aux secteurs Sud-Ouest et Nord-Ouest. On remarque que pour une vitesse $\rho \geq 8\,m \cdot s^{-1}$ sur le secteur Nord-Est, les températures extérieures varieraient entre 10 et 13°C. Quant aux niveaux de l'humidité spécifique, les concentrations basses ($< 4.4\,g \cdot kg^{-1}$ d'air sec) correspondent uniquement au secteur Sud-Est.

Dans la maison expérimentale et durant la phase de référence (aucune stratégie de chauffage n'a été mise en fonction), seules les sources intérieures influencées par les différents paramètres climatiques contribueraient de manière significative aux niveaux du HCHO. On représente dans la Figure 2.5.16 la distribution des fluctuations du HCHO ainsi que la variabilité par rapport aux probabilités conditionnelles des paramètres du vent.
Clairement, il est plus probable d'observer une concentration élevée en HCHO dans les secteurs Nord-Ouest et Sud-Ouest correspondant à $\rho \in [2, 6]\,m \cdot s^{-1}$ que dans les autres secteurs avec des $\rho > 6\,m \cdot s^{-1}$. En associant les observations de projection de la température et de l'humidité spécifique intérieures avec celles du HCHO, la relation entre l'ensemble de ces variables se précise sur les coordonnées polaires.
Avec une telle représentation, les paramètres climatiques pourraient permettre d'aider à estimer la contribution extérieure sur la concentration d'un polluant dans l'air intérieur. Notamment, l'observation de la température et/ou de l'humidité pris(es) conditionnellement à la direction et à la vitesse du vent fournit des informations complémentaires sur l'impact de la pollution extérieure. Ces informations traduisent généralement l'importance de l'origine géographique d'une source sur les concentrations de polluants. En effet, les conditions météorologiques favorisent l'accumulation ou la dispersion des polluants au voisinage des bâtis, donc de leur transfert potentiel vers l'air intérieur.

Pour les particules extérieures (voir Figure C.2.4 dans l'annexe C.2.1), les concentrations les plus élevées des particules de taille $0.35\,\mu m$ sont plutôt d'origine Nord-Est et ce avec des vitesses du vent allant de 2 à $10\,m \cdot s^{-1}$. Il semble que seules les sources associées à des vents de secteur Est avec une vitesse du vent supérieure à $8\,m \cdot s^{-1}$ contribueraient aux concentrations des particules ($0.35\,\mu m$) intérieures (Figure C.2.3 dans l'annexe C.2.1). Les niveaux les plus élevés pour les particules intérieures et extérieures de taille entre 0.9 et $2.5\,\mu m$ sont observés exclusivement dans le secteur Nord-Ouest avec des vitesses supérieures à $6.5\,m \cdot s^{-1}$. Les concentrations élevées pour les grosses particules ($> 10\,\mu m$) sont observées pour un ρ entre 4 et $6.5\,m \cdot s^{-1}$ dans le secteur Nord-Est.

2.5.3 Fluctuations dans l'espace de bureaux

2.5.3.1 Campagne 2012

Les données présentées dans ce paragraphe sont celles mesurées entre le 28 janvier et le 30 juin 2012 dans l'espace paysager. Contrairement aux deux environnements décrits précédemment où les paramètres étaient mesurés avec une résolution temporelle d'une minute, dans cet espace de bureaux, les mesures sont au pas de temps d'une heure et elles concernent les paramètres suivants :

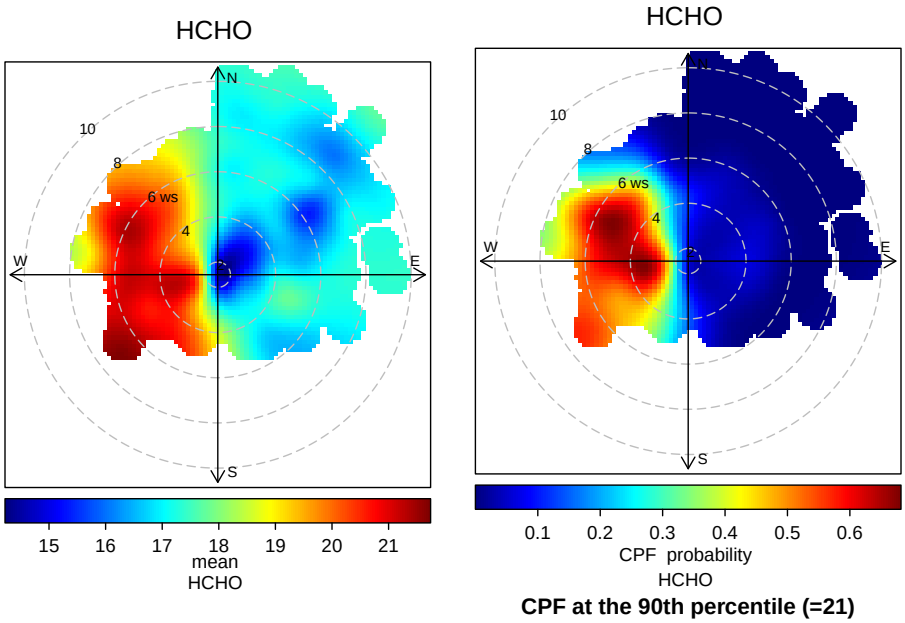


FIGURE 2.5.16 – *Variabilité polaire des concentrations du formaldéhyde dans la maison expérimentale associée aux variables circulaires par la fonction de probabilités conditionnelle (à droite) et à la moyenne (à gauche).*

TABLE 2.5.2 – *Statistiques de la concentration de CO_2 globale et en fonction de l'état d'occupation dans l'espace de bureaux sur la période allant du 28/01 jusqu'au 30/06/2012 au pas de temps horaire.*

État	*n*	*Mean*	*sd*	*Median*	*Min*	*Max*	*Range*	*Skew*	*Kurtosis*	*SE*
All	3391	420	98	382	306	888	582	1.7	2.3	1.7
1	3129	411	93	378	306	865	559	1.9	3.6	1.7
2	262	528	98	514	360	888	528	0.4	-0.2	6

- La concentration de cinq gammes de particules, dont la taille varie entre 0.35 et 8.75 μm et la concentration en CO_2 dans l'air intérieur ;
- La température, l'humidité relative et l'irradiance extérieure et celle intérieure ;
- La vitesse et la direction du vent (station météorologique propre à la maison expérimentale) ;
- L'état d'occupation et d'ouverture.

Le pas de temps d'une heure affecte la variable "ouverture", car plusieurs configurations peuvent être établies en fonction de la durée d'ouverture et du nombre de fenêtres et/ou de portes ouvertes. Pour cela, un codage spécifique a été effectué pour pouvoir utiliser l'information sur l'état des ouvertures : il s'agit de calculer le nombre de minutes pendant lesquelles au moins une fenêtre (respectivement une porte) est ouverte. Cette variable permet donc d'estimer la proportion horaire du temps de l'ouverture des fenêtres ouvertes (respectivement de la porte d'entrée principale). Quant à la variable occupation, elle est codée en binaire, par 1 pour désigner l'état d'inoccupation et par 2 pour occupation, si au moins un occupant était détecté dans le créneau horaire.

Les polluants mesurés

La concentration de CO_2

La série temporelle, le profil diurne ainsi que la densité de la concentration en CO_2 sont représentés dans la Figure 2.5.17. Les fluctuations dans l'espace paysager sont semblables à celles observées dans le bureau individuel, mais à des niveaux plus faibles (voir la sous-section 2.5.1.1). Pendant les six mois de mesure en 2012, les concentrations variaient entre 306 et 887 ppm avec une médiane de 381 ppm et un écart-type 98 ppm. Le profil moyen de la concentration reste stable au niveau de 360 ppm en dehors des heures d'occupation et atteint les 500 ppm durant ces heures. Quelques paramètres statistiques de la concentration de CO_2 en fonction de l'état d'occupation sont présentés dans le Tableau 2.5.2. Les deux distributions se chevauchent sur la quasi-totalité des valeurs observées, ceci montre qu'on peut observer de très fortes valeurs même si l'espace de bureaux est vacant, et inversement, on peut également observer de faibles valeurs alors que l'espace est occupé.

Pour un environnement intérieur de ce type, la relation entre le volume de la pièce, le taux de renouvellement de l'air et le nombre d'occupants est importante pour comprendre l'aire de chevauchement des deux distributions. En effet, les faibles valeurs dans le cas d'occupation peuvent être attribuées à une bonne aération de l'espace ou au rapport du nombre d'occupants sur le volume de la pièce.

En outre, les valeurs du 90^e (576 ppm) et du 95^e (641 ppm) centiles montrent que les concentrations restent faibles par rapport à la capacité d'occupation de l'espace.

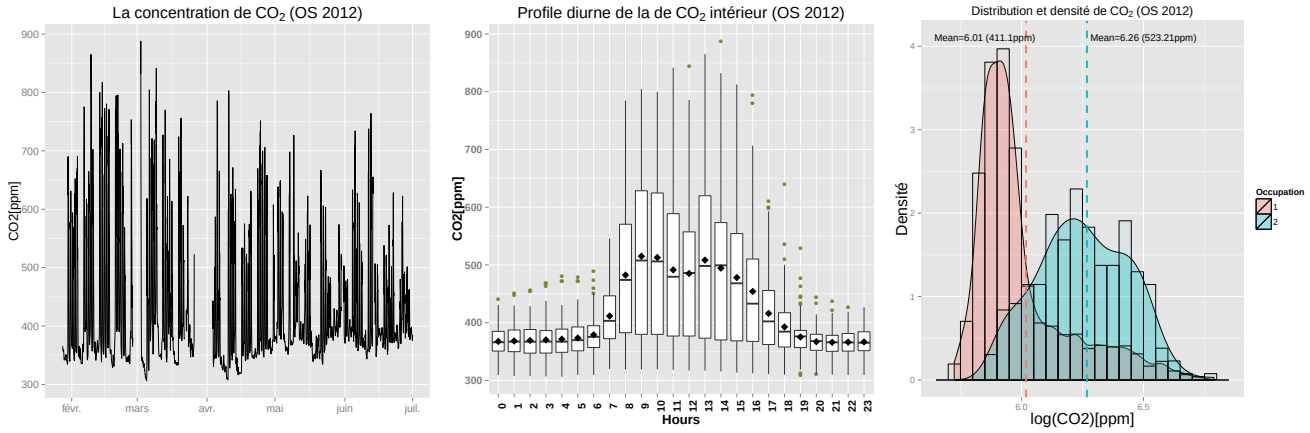


FIGURE 2.5.17 – *La série temporelle, le profil de la distribution horaire et la densité de la concentration de CO_2 dans l'espace paysager durant la campagne de 2012 (OS2012). Les mesures couvrent une période allant 28/01 jusqu'au 30/06/2012 au pas de temps horaire. La densité est représentée en fonction de l'état d'occupation et les concentrations ont été transformées en logarithme, les deux lignes verticales discontinues représentent les moyennes respectives de chaque groupe : occupation et inoccupation.*

Les concentrations en particules

Simultanément aux mesures de CO_2, cinq gammes de particules de taille allant de 0.35 à 8.75 μm ont été choisies pour analyser les fluctuations de la pollution particulaire durant cette campagne. Les séries temporelles au pas de temps horaire de la concentration en nombre de particules des différentes tailles (exemple de quatre gammes) sont présentées dans la Figure 2.5.18.

Pour les particules de tailles fines, les concentrations varient entre 0.7×10^3 #L^{-1} à 249×10^3 #L^{-1} pour 0.35 μm et de 0.02×10^3 #L^{-1} à 2×10^3 #L^{-1} pour 0.9 μm. Pour les particules de tailles 1.8 et 4.5 μm, les concentrations peuvent atteindre 570 et 215 #L^{-1} , respectivement. Le Tableau 2.5.3 récapitule la variabilité en terme de centiles des séries de particules en suspension dans l'air intérieur de l'espace paysager durant la campagne de 2012.

La variabilité temporelle des particules de taille médiane de 8.75 μm est dominée par une succession de sauts quasi-discrets allant de 0 à 31 #L^{-1} ; ce type de variabilité est presque similaire à celle observée pour le CO_2 (*cf.* la série temporelle dans la Figure 2.5.17).

TABLE 2.5.3 – Centiles des séries temporelles de la concentration en particules de l'air dans l'espace paysager durant la campagne de 2012 (OS2012). Les mesures couvrent une période allant du 28/01 jusqu'au 30/06/2012 au pas de temps horaire. La valeurs sont exprimées en $\#/L$.

Centiles	**5%**	**10%**	**20%**	**25%**	**50%**	**75%**	**80%**	**90%**	**95%**
0.35 μm	3918	4980	7168	8605	18168	38891	49013	72587	95418
0.9 μm	65	88	119	136	214	353	394	535	734
1.8 μm	20	28	40	44	75	122	136	176	218
4.5 μm	0.88	1.5	2.7	3.51	9.96	24	28	37.7	49
8.75 μm	0.018	0.037	0.074	0.09	0.36	2.42	3.1	4.4	5.66

Par ailleurs, les paramètres statistiques présentés sont influencés par l'état d'occupation et le taux de renouvellement de l'air. Par exemple, on constate une augmentation de la concentration médiane

de proportion de l'ordre de 14 % pour les tailles 0.35 et 0.9 μm, de 24 % pour les 1.8 μm, de 57 % pour le particules moyennes 4.5 μm, de 71.4 % pour la taille 8.75 μm. Ces observations indiquent que l'influence de l'occupation sur le niveau de fluctuation des particules en suspension se répartie de manière distributive en fonction de leur taille.

Hormis le fait que la concentration soit plus élevée pour les fines particules, leur niveau de variabilité présente des tendances et de fortes fluctuations sur de courtes périodes. Par exemple, entre février et avril, une tendance haussière (linéaire) et une tendance baissière en avril peuvent être dégagées de la trajectoire globale de la concentration des particules de taille 1.8 μm. Des mouvements quasi-périodiques intra-hebdomadaire sont observés pour les particules de tailles moyennes : 4.5 et 8.75 μm.

Notons encore, en fonction de l'occupation, que la distribution mensuelle indique qu'il n'y pas de profil exploitable pour les particules fines et la différence entre les fluctuations en fonction de l'occupation est très difficile à mettre en exergue. Au contraire, plus les particules sont grosses, plus l'impact de l'occupation sur la concentration est déterminant. Sur la Figure 2.5.19, on montre que la contribution de "l'occupation" aux valeurs de particules, notamment pour les gammes 4.5 et 8.75 μm est significative, et ceci principalement durant les mois de février et mars.

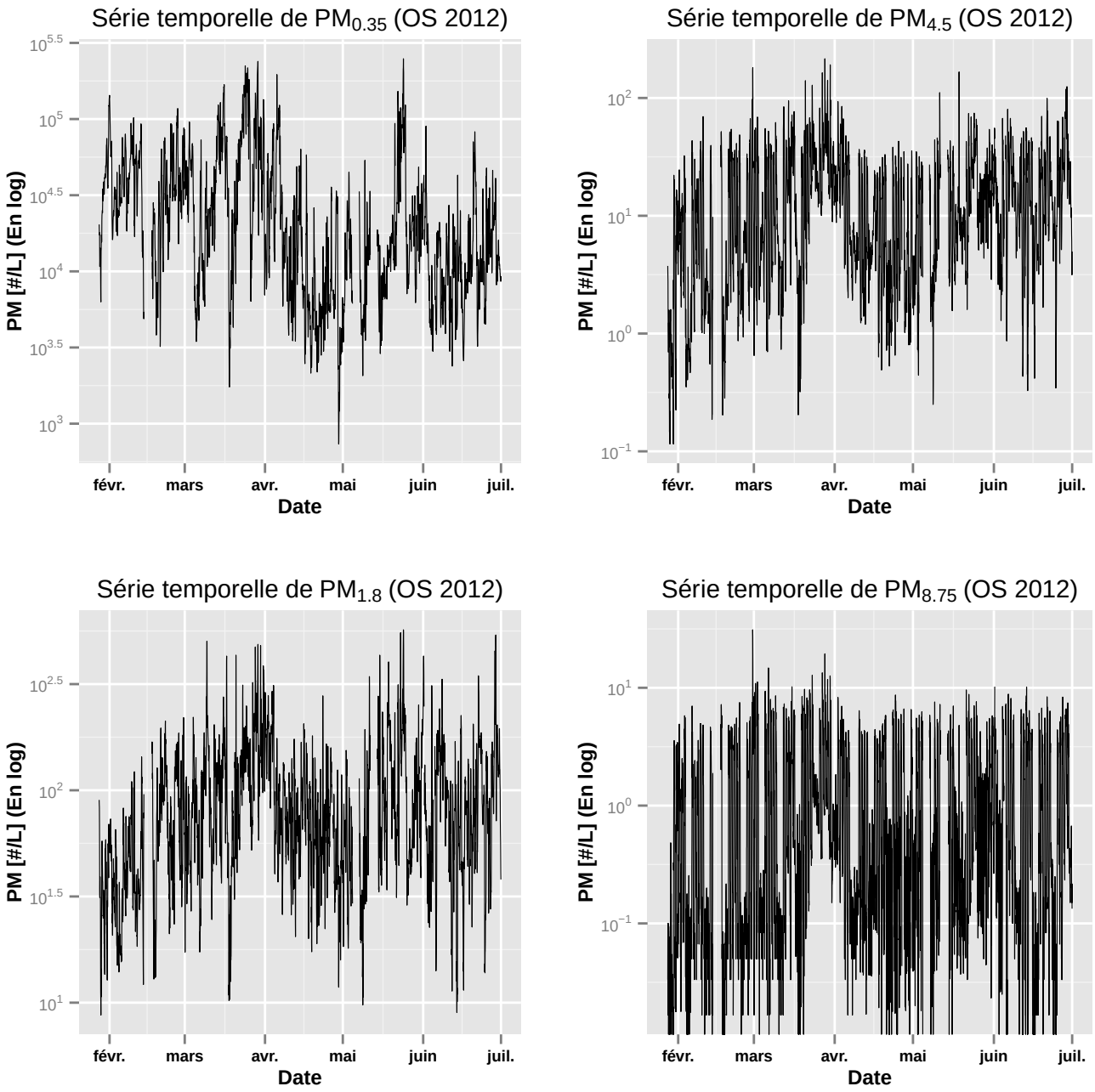
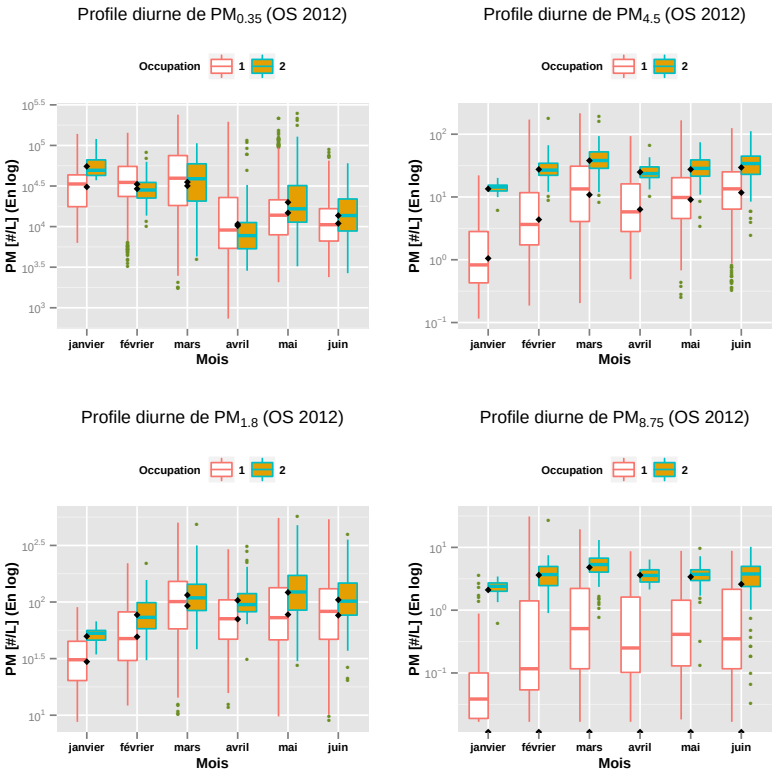
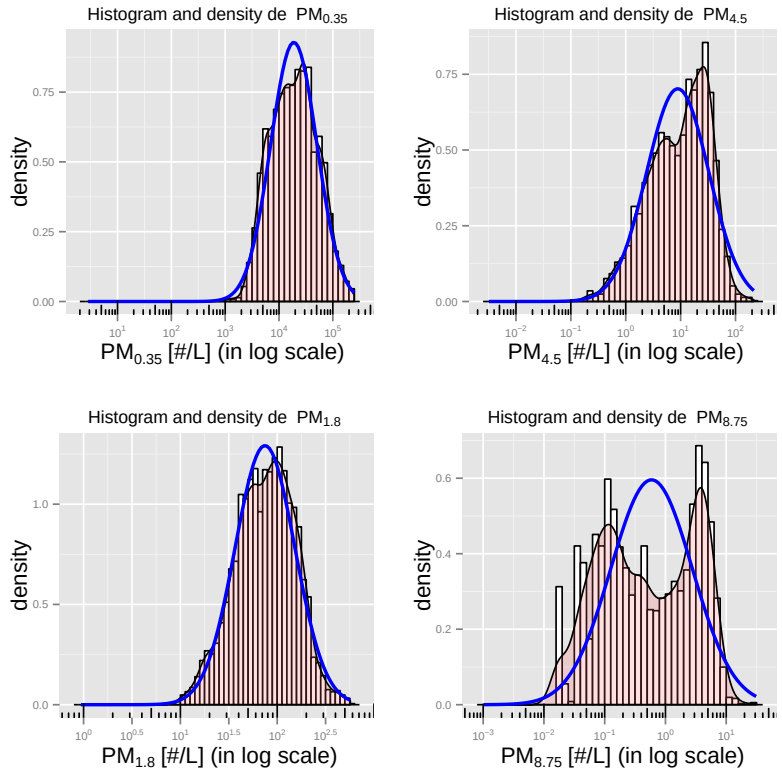


FIGURE 2.5.18 – *Les séries temporelles de la concentration horaire en particules dans l'espace de bureaux durant la campagne de 2012 (OS2012).*



(a) *La distribution mensuelle de la concentration de particules en fonction de l'occupation. L'état d'occupation est codée par 1 pour désigner l'inoccupation et 2 pour le cas d'occupation.*



(b) *Les densités de probabilités estimées par la méthode de Parzen-Rosenblatt (noyau Gaussien, l'aire sous la courbe) et par des Gaussiennes paramétriques (courbe bleu).*

FIGURE 2.5.19 – *Profils mensuels et densités de probabilités des concentrations de particules dans l'espace paysager durant la campagne de 2012. Les valeurs des concentrations sont exprimées en* $\log_{10}$ *par le nombre de particules par litre* $(\#/L)$*. Les mesures couvrent la période allant 28/01 jusqu'au 30/06/2012 au pas de temps horaire.*

TABLE 2.5.4 – Configurations les plus probables dans l'espace de bureaux.

	Occupation	**Porte principale**	**Fenêtre**	n_i	f_i (%)
Configuration 1	1	1	1	131	3.52
Configuration 2	1	1	0	29	0.8
Configuration 3	1	0	1	112	3
Configuration 4	1	0	0	41	1.1
Configuration 5	0	1	1	1828	50
Configuration 6	0	1	0	170	4.5
Configuration 7	0	0	1	716	19.4
Configuration 8	0	0	0	693	18.7

Les densités de probabilités montrent que les concentrations horaires des particules de tailles inférieures à 4.5μm peuvent être approximées par les densités log-normales. En revanche, pour les particules supérieures à 8.75μm, plusieurs modes se dessinent.

Influence des paramètres d'occupation et d'ouverture sur la concentration des polluants

En tout, l'espace paysager était occupé à 8.4 % du temps global sur toute la période de mesure (5 mois de mesure sans exclure les week-ends). Durant la période d'occupation, les occupants agissent sur les différentes composantes de l'environnement. Les actions les plus commodes sont l'ouverture des fenêtres et des portes. Plusieurs configurations peuvent être considérées dans un espace normalement occupé, elles sont présentées dans le Tableau 2.5.4. Pour l'occupation, le 1 indique que l'espace est occupé par au moins une personne, 0 sinon. La porte principale prend la valeur 1 lorsqu'elle est ouverte, et 0 sinon. Quant à l'ouverture des fenêtres, si au moins une fenêtre est ouverte, alors l'espace est considéré comme ayant les fenêtres ouvertes, sinon toutes les fenêtres sont considérées fermées. La $1^{ère}$ configuration (*cf.* Tableau 2.5.4), indique que l'espace est occupé, et que fenêtres et porte sont ouvertes. Pour la $7^{ème}$, l'espace est inoccupé, et tous les ouvrants sont fermés.

Le Tableau 2.5.4 fournit le nombre d'heures observé pour chaque configuration. La configuration la plus fréquente pendant 50% du temps, correspond à un espace de bureaux inoccupé, avec porte et fenêtres ouvertes. L'effet calendrier n'a pas été pris en compte dans ce Tableau.

Clairement, la combinaison de ces variables (ouverture et occupation) est assez délicate à traiter, car on peut envisager plusieurs autres configurations, et en fonction de celles-ci, on code l'ouverture des fenêtres pour chaque configuration. L'expression, "*au moins une fenêtre ouverte*" peut être contestée, le choix sur cette formulation est assujetti à la qualité des données recueillies pour toutes les fenêtres de l'espace paysager. En effet, en présence des données manquantes sur au moins une fenêtre, il est très difficile d'établir la relation entre le facteur d'ouverture, l'occupation et la concentration en polluant. Donc, l'agrégation d'enregistrements doit être le plus large possible et nuancée pour permettre l'interprétation des résultats.

2.5.3.2 Campagne 2013

La campagne de mesure de 2013 au niveau de l'espace paysager est assez complète et de bonne qualité. La mesure la plus exigeante est celle du formaldéhyde, qui nécessite du personnel qualifié.

L'enregistrement des mesures est complet à hauteur de 99 % pour le CO et l'ozone, de 88 % pour les oxydes d'azotes, de 86 % pour le CO_2 et enfin de 74 % pour le HCHO. Ces valeurs manquantes sont en partie liées à la récupération périodique des données, au temps de calibration mais également à des pannes techniques parfois. La concentration des particules n'a pas été renseignée sur cette période.

Pendant la campagne de 2013, l'état des ouvrants est renseigné au pas de temps d'une minute comme pour l'ensemble des autres mesures.

Variabilité temporelle du formaldéhyde

La mesure de HCHO a été réalisée entre le 27 avril et le 31 juillet en 2013 dans l'espace de bureaux au pas de temps d'une minute. La série temporelle ainsi que la densité de probabilité associées aux fluctuations du HCHO sont rapportées dans la Figure 2.5.20. Au total, 101642 valeurs ont été enregistrées sur 138240, soit environ 26 % de valeurs manquantes. Les interruptions d'enregistrement ont eu lieu sur des différents segments de temps réparties sur l'ensemble de la période de mesure, la panne technique étant la principale cause des valeurs manquantes.

La concentration de HCHO mesurée durant cette période variait entre 0.3 et 23 ppb, avec une moyenne de 9.5 ppb ; en outre 45 % des valeurs sont supérieures à la moyenne. Le niveau global de variation de ces concentrations est plus faible que celui observé dans la maison expérimentale. En revanche, les fluctuations observées dans l'espace paysager présentent plus de variations rapides que la concentration du HCHO dans la maison expérimentale. À titre indicatif, le coefficient de variation est estimé à 40 % pour HCHO dans l'espace paysager contre 23 % dans la maison expérimentale. Cette observation reflète le caractère plus aléatoire des fluctuations dans un espace normalement occupé.

La densité de probabilité estimée est plutôt polymodale, aplatie et légèrement étalée à droite. Sur l'ensemble de la série, 5 % des données fluctuent entre 16 ppb et 23 ppb et la moyenne est presque identique à la médiane (8.9 ppb). Par comparaison avec la distribution observée dans la maison expérimentale (*cf.* Figure 2.5.8), les coefficients d'aplatissement (-0.01) et de d'asymétrie (0.47) avoisinent le zéro.

Sur la Figure 2.5.21, le profil diurne montre qu'il n'existe pas de distribution horaire moyenne type des fluctuations. Le profil moyen reste stable au cours de la journée : la variabilité reste monotone et quasiment indépendante de l'heure de la mesure. Dans la maison expérimentale au contraire , le profil des fluctuations horaires dépend des heures et un profil journalier type a été mis en évidence.

Le profile diurne dans la Figure 2.5.21 montre que la moyenne horaire diminue légèrement au cours de la journée et remonte légèrement en soirée. Par ailleurs, une augmentation du niveau moyen de la concentration passant de 7.9 ppb à 10.6 ppb au cours des 4 mois de mesure. Une baisse de la concentration moyenne est généralement observée lors de l'occupation de l'espace.

Au regard de ces observations (voir aussi la Figure C.2.5 dans l'annexe C.2.2 pour une illustration complémentaire), on constate que le niveau moyen de la concentration en formaldéhyde dans un espace normalement occupé dépend au moins de deux paramètres simultanément : les conditions d'ambiances intérieures et extérieures (température et humidité) et l'occupation. En effet, la concentration dépend naturellement de la température qui augmente progressivement entre avril et juillet et durant la journée,

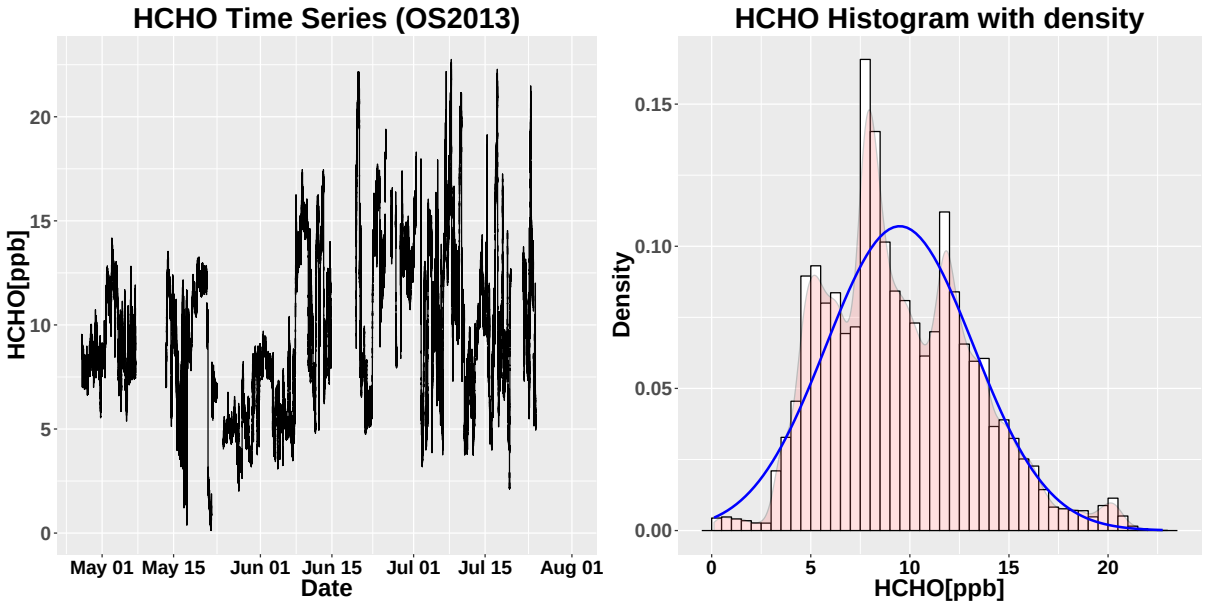


FIGURE 2.5.20 – *La série temporelle des concentrations de HCHO et la densité de probabilité associée sur toute la période de mesure (27/04 - 31/07/2013). Les mesures sont effectuées dans l'espace paysager au pas de temps d'une minute. La courbe en bleu correspond à une approximation de la distribution selon la densité de la loi Gaussienne.*

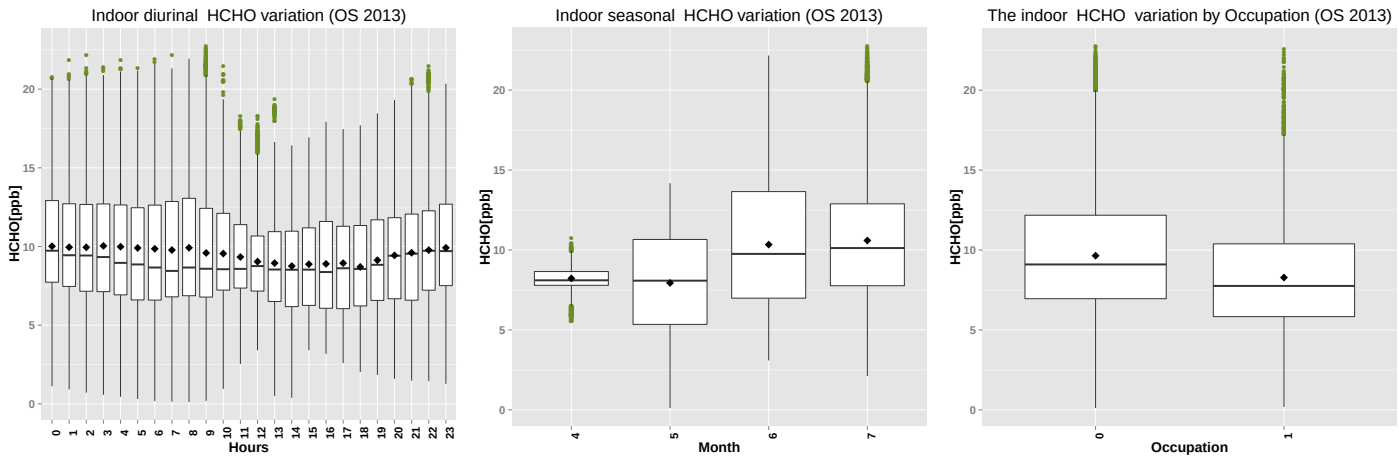


FIGURE 2.5.21 – *Distribution des concentrations de HCHO, à l'échelle de la journée, par mois et selon l'occupation (0 indique l'inoccupation et 1 l'occupation). Les mesures couvrent la période du 27/04/2013 au 31/07/2013 au pas de temps d'une minute.*

favorisant ainsi l'augmentation des émissions, mais d'un autre côté, une température plus élevée incite les occupants à aérer l'espace entraînant une action sur les ouvertures et favorisant donc la dilution du HCHO dans l'espace.

Il se trouve que ces deux paramètres interviennent pratiquement aux mêmes heures de la journée, ce qui a pour effet une dilution de la concentration par les mouvements d'air, expliquant ainsi la baisse des niveaux de HCHO.

Variabilité temporelle des autres polluants : l'ozone (O_3), dioxyde de carbone (CO_2), monoxyde de carbone (CO) et les oxydes d'azotes (NO_x)

Les mesures d'ozone et de dioxyde de carbone ont été renseignées sur la même période que le formaldéhyde (soit du 27/04/2013 au 31/07/2013) et au même pas de temps d'une minute. Les Figures 2.5.22a et 2.5.22b montrent les séries temporelles de ces deux paramètres, leurs distributions diurnes et leurs densités de probabilités.

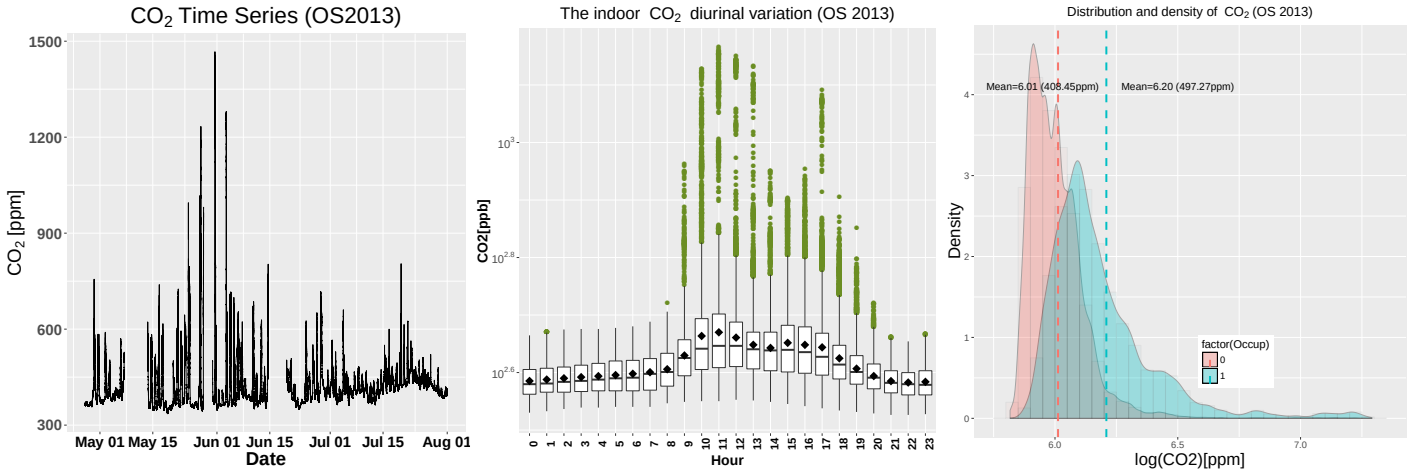
Hormis les valeurs extrêmes, la concentration de CO_2 a varié entre 400 et 744 ppm en atteignant une valeur maximale de 1450 ppm. La présence des valeurs extrêmes est attribuée à une sur-occupation ponctuelle de l'espace de bureaux.

L'aire de chevauchement des densités de probabilités par occupation est plus importante sur les mesures de 2013 par rapport aux mesures de 2012. Quant au profil diurne, les différences entre les différentes séries de concentration des trois environnements étudiés sont négligeables.

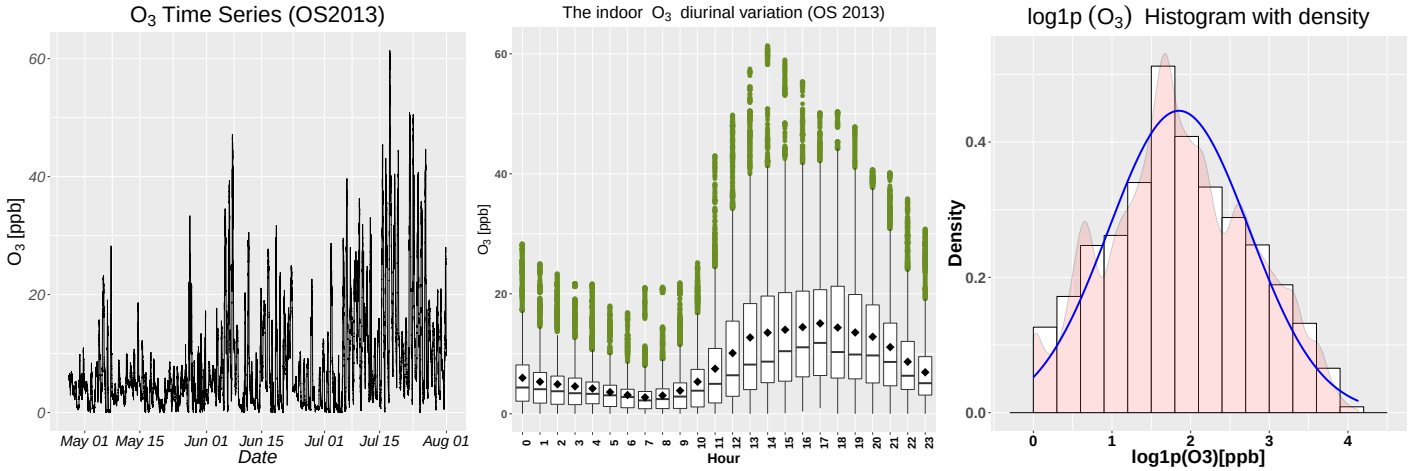
La concentration d'ozone peut atteindre les 62 ppb, mais sa moyenne est faible (8.4 ppb). Environ 70 % des valeurs sont inférieures à cette moyenne. La distribution horaire de la concentration de d'ozone dépeint un profil diurne-type au cours de la journée : les concentrations élevées sont observées entre 13 h et 15 h, elles diminuent lentement après 17 h et elles sont très faibles pendant la nuit. On remarque par ailleurs une augmentation graduelle des concentrations mensuelles entre avril et juillet, la concentration étant plus élevée en été. Les données semblent suivre une loi de distribution log-normale.

Les concentrations de monoxyde de carbone (CO) observées sont faibles (moyenne de 0.17 ppm), mais elles présentent beaucoup de variabilité. Pour les oxydes d'azote (NO_x), la concentration moyenne observée est de 10.3 ppb ; elle est inférieure à la concentration observée à l'extérieur (environ 17 ppb à la station du réseau AIRPARIF de Lognes sur la même période).

La Figure 2.5.23 reprend les principales informations concernant la variabilité de la concentration en CO et en NO_x dans l'espace paysager durant la campagne de 2013 (OS2013). Les profils moyens type mettent en évidence un pic matinal de NO_x (entre 7 h et 10 h), ainsi qu'une diminution des niveaux de CO durant les heures de la journée.



(a) *Série temporelle, distribution diurne et densité de probabilités des concentrations de CO_2 dans l'espace paysager durant la campagne de 2013. La densité est représentée en fonction de l'état d'occupation et les concentrations sont logarithmées, les deux lignes verticales discontinues représentent les moyennes respectives de chaque groupe : occupation (1) et inoccupation (0).*



(b) *Série temporelle, distribution diurne et densité de probabilité de la concentration d'ozone (O_3). Sur la figure de densité, les valeurs de concentrations de l'ozone sont translatées pas 1 ppb et logarithmées ($x \mapsto \log(x+1)$) ; la courbe en bleu correspond à l'ajustement de la densité par une loi Gaussienne de même moyenne et écart-type.*

FIGURE 2.5.22 – *Variabilité temporelle, profil diurne et densité de probabilités des données de l'ozone et de CO_2 mesurées des l'espace paysager du 27/04/2013 au 31/07/2013 au pas de temps d'une minute.*

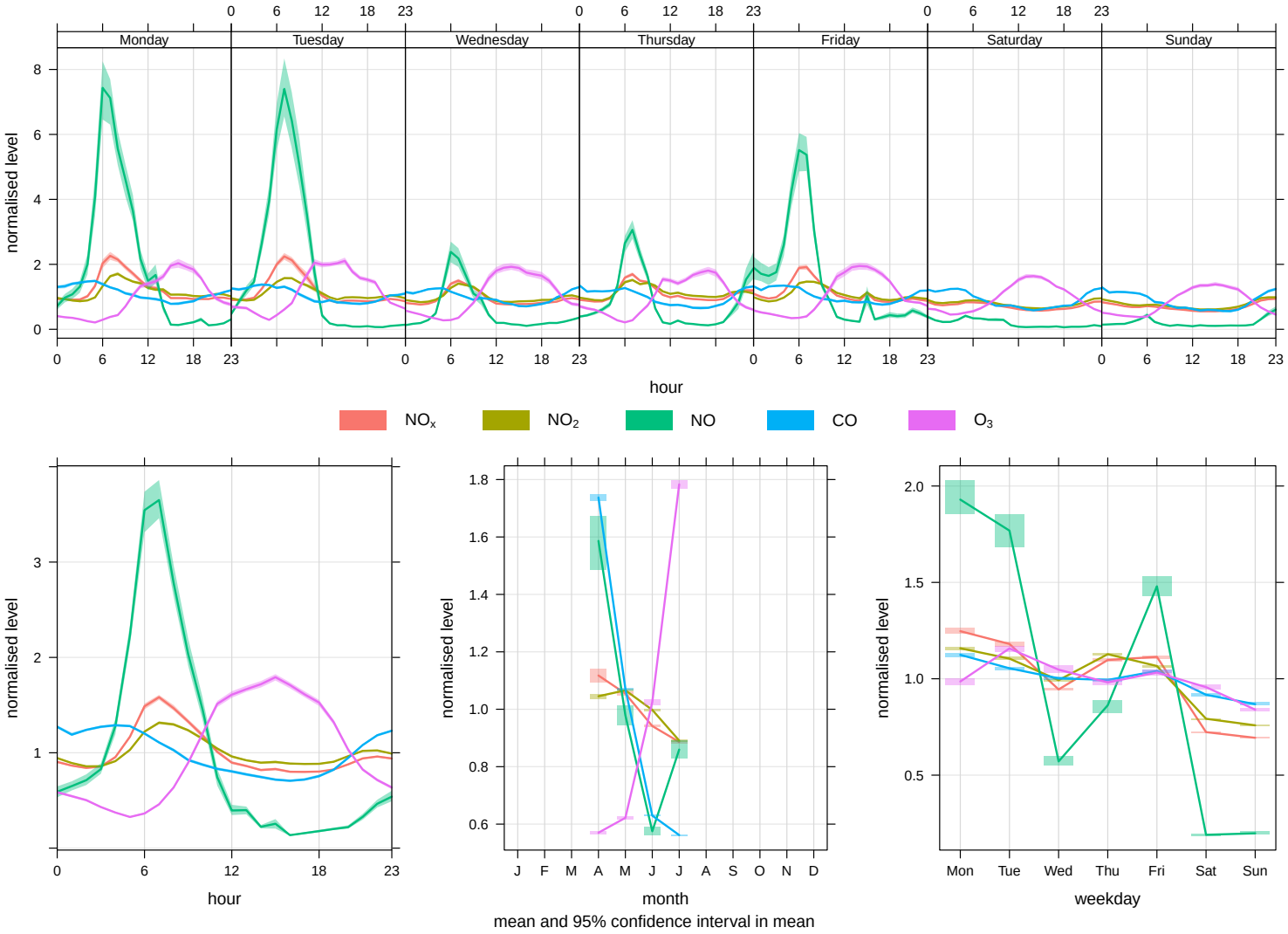


FIGURE 2.5.23 – *Fluctuations moyennes des concentrations en NO_x, NO, NO_2, CO et en O_3 dans l'espace paysager pendant la période du 27/04/2013 au 31/07/2013 avec pas de temps d'une minute : profile hebdomadaire type, distribution diurne, variation moyenne par mois et la moyenne de chaque jour. Les données ont été normalisées par $\tilde{x} = \frac{x}{\bar{x}}$.*

La principale caractéristique de la distribution hebdomadaire des NO_x est la baisse des concentrations le mercredi et l'absence du pic matinal le week-end. La plupart des fortes concentrations de NO_x sont dues à la variabilité du NO, qui est caractérisée par la présence des valeurs extrêmes qui se "greffent" de manière ponctuelle sur les concentrations moyennes. Ainsi, pendant la période de mesure, 97 % des valeurs de concentration de NO sont inférieures à 11 ppb et dans le reste des 3 % on peut atteindre un maximum de 60 ppb. Le pic observé à l'intérieur peut être attribué alors aux émissions provenant du trafic extérieur. Le trafic routier est la source principale des oxydes d'azote (NO_x) en île-de-France, avec une contribution de 50 % pour l'année de référence de 2000 (AirParif, 2008). Sa contribution est très visible sur ces graphiques. Notons que le bâtiment cible est à 600 m de la voie rapide D199 (600m), à 130 m d'une rue passante, à 4.5 km de l'autoroute A4 et à 4.7 km de l'autoroute A104.

Fluctuations des paramètres et leur influence sur la variabilité des polluants

L'état d'occupation et d'ouverture des fenêtres

Dans cette section, nous présentons d'abord les données relatives à l'état des ouvertures et d'occupation, ensuite nous combinons les différents jeux de données pour tenter d'évaluer en amont leur importance sur les niveaux de polluants. Il s'agit en particulier d'établir une taxonomie des variables par rapport à leurs fréquences d'occurrence, leur disponibilité, afin d'estimer leur potentiel explicatif sur les différents polluants.

La variable "*Occupation*" désigne l'état d'occupation de l'espace paysager, étant codée comme suit : elle prend la valeur 1 si au moins un des capteurs a détecté un mouvement, 0, sinon. Durant toute la période de mesure (du 27/04/2013 au 31/07/2013 au pas de temps d'une minute), cette variable est renseignée à 100 %, car on a éliminé les configurations dans lesquelles tous les capteurs sont en panne en même temps ou lors d'un dysfonctionnement lié à l'acquisition des données par la CSTBox.

De façon générale, l'espace paysager est occupé pendant 10.4 % du temps. Cette information ne nous renseigne pas réellement sur l'occupation de l'espace, mais uniquement sur le nombre total de détections de mouvement durant la période de mesure.

Si on somme le nombre de détections par heure, la variable "*Occupation*" ne dépasse pas les 30 min d'occupation/h, valeur maximale observée vers 10 h, soit 10 % du temps total lors de l'occupation. Le taux d'occupation horaire en journée est de l'ordre de 20 min d'occupation/h, et ce indépendamment des weekends et des jours fériés.

Le paramètre "*Occupation*" est très variable selon plusieurs facteurs (le nombre d'occupants, la couverture spatiale de l'environnement par les capteurs...), mais moins biaisée par son mode de quantification. En effet, si on étudie la distribution de l'occupation sur l'ensemble des heures de la journée (7-19 h) et indépendamment du jour, on constate que l'espace était occupé à 19 % du temps et à 26 % si on exclut les weekends. On pense que ces chiffres sont corrects, car en théorie on devrait trouver une proportion de l'ordre de 20-25 % pour 20 jours ouvrés dans le mois et une présence quotidienne de 8 heures. Mais étant donné que l'espace paysager est occupé par au moins 5 personnes mobiles dans leur travail, leur présence et par conséquent, l'occupation est très variable.

Durant les jours ouvrés et entre 7-19 h, l'occupation de l'espace de bureaux variait entre un minimum de 21 % pour le mercredi et un maximum de 30 % pour le vendredi, étant en moyenne de 25 % pour les autres jours. L'occupation par jour ouvré représente donc environ 5 % du temps durant toute la période de mesure.

On peut conclure que la variable "*Occupation*" est un bon indicateur de présence et nous renseigne sur la présence des occupants, mais elle sous-estime légèrement l'occupation réelle de l'environnement en l'absence du nombre d'occupants par minute. En effet, l'estimation de l'occupation peut être faussée par le fait que la portée des capteurs est limitée à une demi-sphère et ainsi l'espace paysager n'est pas couvert en totalité, par conséquent, certains mouvements peuvent ne pas être détectés, d'où une sous-estimation possible de l'occupation réelle.

L'information disponible sur l'état d'ouverture des fenêtres pour toutes les cinq fenêtres instrumentées était de 60%, soit trois fenêtres sur cinq. Deux capteurs avaient des défaillances techniques, mais leur emplacement était au niveau des bureaux individuels.

La Figure 2.5.24 reprend les informations liées aux deux paramètres étudiés : l'occupation et l'ouverture des fenêtres. Elle donne la distribution journalière et hebdomadaire de l'état d'ouverture des fenêtres associées à la variation de l'occupation de l'espace paysager. L'ouverture des fenêtres est renseignée par quatre modalités : elle vaut

- 0 : toutes les fenêtres renseignées étaient fermées ;
- 1/3 : une fenêtre est ouverte une fenêtre est ouverte sur trois fenêtres renseignées ;
- 2/3 : deux fenêtres sont ouvertes sur trois fenêtres renseignées ;
- 1 : toutes les fenêtres renseignées sont ouvertes.

Il arrive qu'on exprime les trois modalités d'ouverture par une seule modalité : au moins une fenêtre est ouverte, remplaçant ainsi "*une fenêtre sur trois est ouverte*", "*deux fenêtres sur trois sont ouvertes*" et *"toutes les fenêtres sont ouvertes*".

Toutes les fenêtres étaient fermées 52.5 % du temps et au moins une fenêtre ouverte 47.5 % du temps. Il est moins probable que les fenêtres soient ouvertes toutes en même temps : cette proportion est de l'ordre de 6 %, suivie par 22.12 % et 19.17 % pour une fenêtre et deux fenêtres ouvertes, respectivement.

Durant la présence, les occupants ont tendance à ouvrir deux fenêtres, avec une fréquence de 34.6 %, et rarement les trois au même temps (9.3 %). En outre, la probabilité de la configuration qu'aucune fenêtre ne soit ouverte lors de l'occupation est de l'ordre de 29 %.

D'après le profil diurne, les occupants ont tendance à laisser au moins une fenêtre ouverte en quittant l'espace de bureau, et ceci d'autant plus le jeudi et le vendredi. On observe aussi qu'il y a des situations où toutes les fenêtres étaient ouvertes le samedi, mais pas le dimanche, ce qui indique qu'il y a une présence durant ce jour qui aurait fermé les fenêtres (la ronde du gardien), même si de manière ponctuelle. On observe d'ailleurs que les détecteurs de présence indiquent qu'il y a de brèves occupations durant des heures très tard dans la nuit et le weekend.

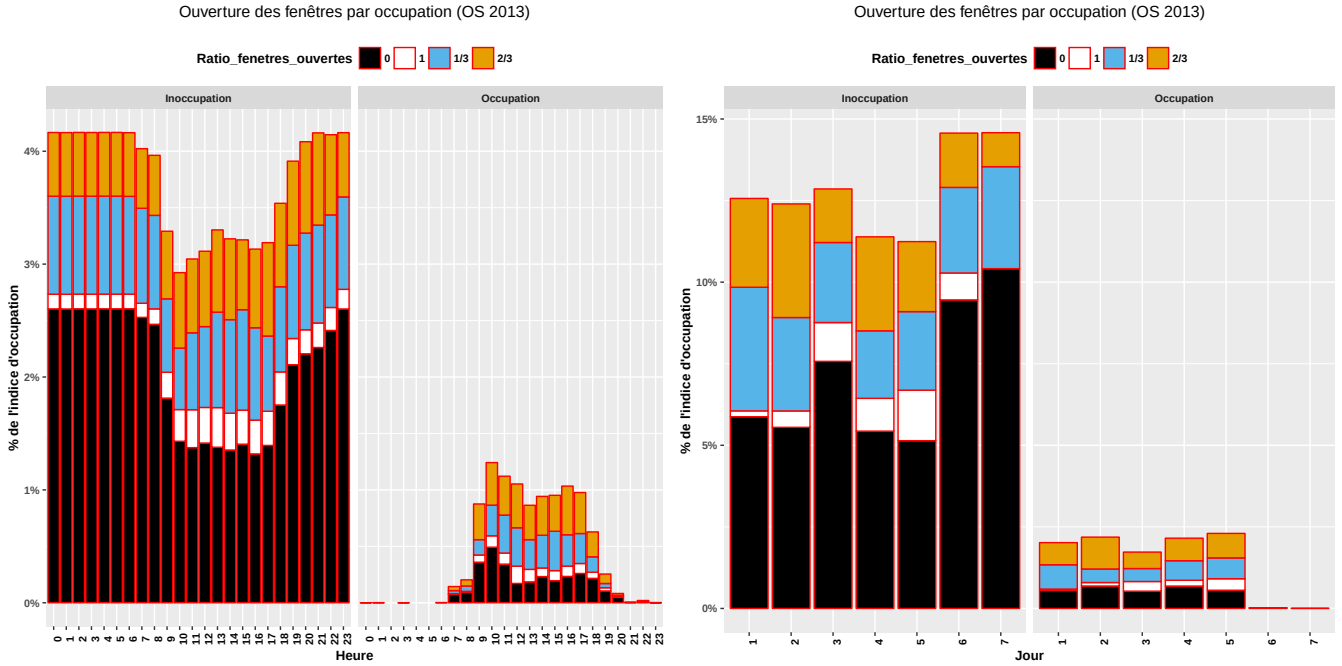


FIGURE 2.5.24 – *Variation journalière (à gauche) et hebdomadaire (à droite) de l'état des fenêtres en fonction de l'occupation durant la période de mesure : 27/04/2013 - 31/07/2013 au pas de temps d'une minute dans l'espace paysager. Uniquement trois fenêtres sur cinq instrumentées ont été renseignées. La variable ratio de fenêtres ouvertes prend la valeur 0 lorsque aucune fenêtre n'est ouverte,* 1/3 *si une fenêtre est ouverte sur trois fenêtres renseignées,* 2/3 *lorsque deux fenêtres sont ouvertes sur trois fenêtres renseignées et la valeur 1 pour toutes les fenêtres renseignées sont ouvertes.*

Relation entre les différents paramètres et la concentration en polluants

Les paramètres climatiques, d'occupation et d'ouverture des fenêtres peuvent influencer la concentrations polluants à l'intérieur. Ils forment un triplet de variables analysées simultanément par croisement des différents jeux de données. Pour ce faire, nous adoptons la même démarche que celle présentée dans la section 2.5.2.3 en s'inspirant de l'approche développée dans (Uria-Tellaetxe & Carslaw, 2014; Carslaw & Beevers, 2013; Carslaw et al., 2006).

Sur la Figure 2.5.25, la variabilité polaire [2] du formaldéhyde et celle de l'ozone sont présentées en fonction de quatre paramètres : la direction et la vitesse de vent, l'indice d'ouverture des fenêtres et l'état d'occupation de l'espace paysager. Notons que l'indice d'ouverture des fenêtres est renseigné comme suit : il vaut 0 si aucune fenêtre n'est ouverte, 0.5 si au plus deux fenêtres sur trois fenêtres renseignées sont ouvertes, il vaut 1 sinon. C'est la même variable présentée dans le paragraphe précédant, en remplaçant les deux modalités 1/3 et 2/3 par 1/2, et en gardant à l'esprit que le taux d'information disponible pour cette variable est de 60% pour les cinq fenêtres instrumentées (3/5).

Pour le HCHO, les niveaux les plus élevés sont observés dans le cas où aucune fenêtre n'est ouverte et aucune présence d'occupants n'est enregistrée. Vu que l'aération de l'espace se fait principalement *via* les fenêtres, la contribution de la variable vent dans cette configuration n'a d'importance que si on disposait des données de toutes les fenêtres instrumentées. Mais *a priori*, le nombre de fenêtres ouvertes n'a pas trop d'influence : l'état d'une ou de deux fenêtres suffirait pour expliquer l'impact de ces paramètres sur la variabilité de certains polluants.

Bien que les vents d'ouest ne soient pas les plus dominants durant la campagne de mesure, les concentrations élevées du HCHO ont été observées dans ce cas, mais uniquement lorsque les fenêtres sont fermées. Le secteur sud-ouest est le secteur dominant pour les valeurs médianes du HCHO autour de 16 ppb. Plus le ratio des fenêtres ouvertes est grand pour la configuration d'inoccupation, plus la concentration est faible. Par contre, lorsqu'il y a occupation et dans le cas où au plus deux fenêtres étaient ouvertes, les concentrations sont les plus faibles.

Pour la variabilité de l'ozone, les conséquences de l'ouverture des fenêtres sur la concentration intérieure de l'ozone sont beaucoup plus apparentes. En effet, il suffit qu'une seule fenêtre soit ouverte pour que la concentration médiane augmente, et ce indépendamment de l'occupation. L'impact de la pollution extérieure sur l'air intérieur pour cette variable est très important.

Les vents forts influencent les conditions au voisinage des bâtiments et leurs effets deviennent importants : la vitesse et la direction du vent impactent l'infiltration de l'air ainsi que les échanges de chaleur par convection à la surface de l'enveloppe des bâtiments, notamment les bâtiments peu isolés. Par exemple, si la vitesse du vent augmente et en fonction de sa direction, le ciel se couvrira ou se dégagera, influençant ainsi la photochimie, donc la composition de l'air. Ces mouvements d'air avec les variations de température favoriseront, entre autres les déperditions thermiques (s'il fait froid) et donc les besoins en énergie de chauffage des locaux augmentent : sollicitation des différentes composantes du bâtiment par l'occupant.

En outre, la température extérieure, quant à elle influence l'ambiance intérieure en termes de température ressentie et d'humidité. Ces ambiances sont généralement modulées par l'occupant afin d'optimiser son confort. La régulation des conditions intérieures se fait en intervenant sur la variable aération ou confinement de l'espace par l'ouverture ou par la fermeture des fenêtres.

Ces actions sur les différentes composantes du bâtiment donnent lieu à des variations considérables sur les niveaux de polluants.

2. Dans le sens où la variable est exprimée par rapport à une autre circulaire.

Les observations précédentes pour le HCHO ne peuvent pas être interprétées uniquement en se basant sur les paramètres extérieurs, mais nécessitent une analyse approfondie avec d'autres paramètres et relations avec d'autres polluants.

Pour remédier à cette difficulté, on utilise la même procédure mais avec d'autres variables de contrôle que la vitesse du vent. Par exemple, sur la Figure 2.5.26, on présente la variabilité médiane du HCHO en fonction de la variation de la température ou de l'humidité spécifique, et ceci par rapport à l'occupation et au ratio d'ouverture des fenêtres.

Clairement, les concentrations élevées du HCHO sont associées à des températures intérieures élevées et ceci d'autant plus pour la configuration de fenêtres fermées et l'espace vacant. Même observation pour l'humidité spécifique. En effet, les concentrations de HCHO dépassant les 20 ppb sont associées à une humidité spécifique (Hsp) intérieure supérieure à $14\,\mathrm{g.kg^{-1}}$ d'air sec. Notons que ces niveaux de HCHO correspondent exclusivement à une humidité relative supérieure à 60 %, dans le cas d'inoccupation et lorsqu'aucune fenêtre n'est ouverte (*cf.* Figure C.2.7 en annexe).

Cela concorde avec notre hypothèse sur les facteurs majeurs influençant la concentration intérieure du HCHO. Dans le cadre de cette étude, la direction du vent joue un rôle mineur : aucun secteur n'est privilégié lors des concentrations élevées du HCHO. En ce qui concerne l'influence de l'humidité spécifique, on observe que dans le cas où toutes les fenêtres sont ouvertes et une concentration de $\mathrm{Hsp} < 7\,\mathrm{g.kg^{-1}}$ d'air sec, aucune concentration du HCHO n'a été prélevée. Pour cette plage de variation de l'humidité et pour les autres configurations, la concentration du HCHO est très faible : < 5.5 ppb.

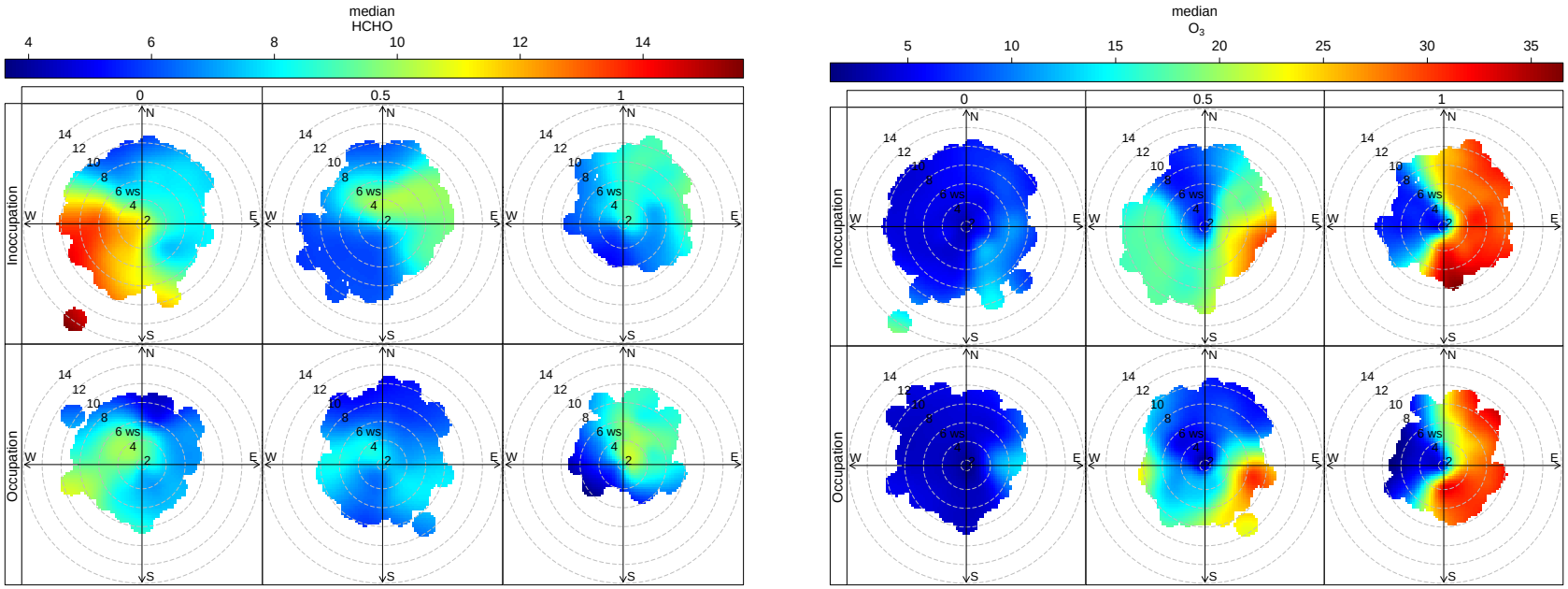


FIGURE 2.5.25 – *Influence des conditions extérieures par rapport à l'état d'occupation et en fonction de l'ouverture des fenêtres sur les niveaux des concentrations intérieures du formaldéhyde et de l'ozone. Les mesures couvrent la période du 27/04/2013 au 31/07/2013 toutes les minutes.*

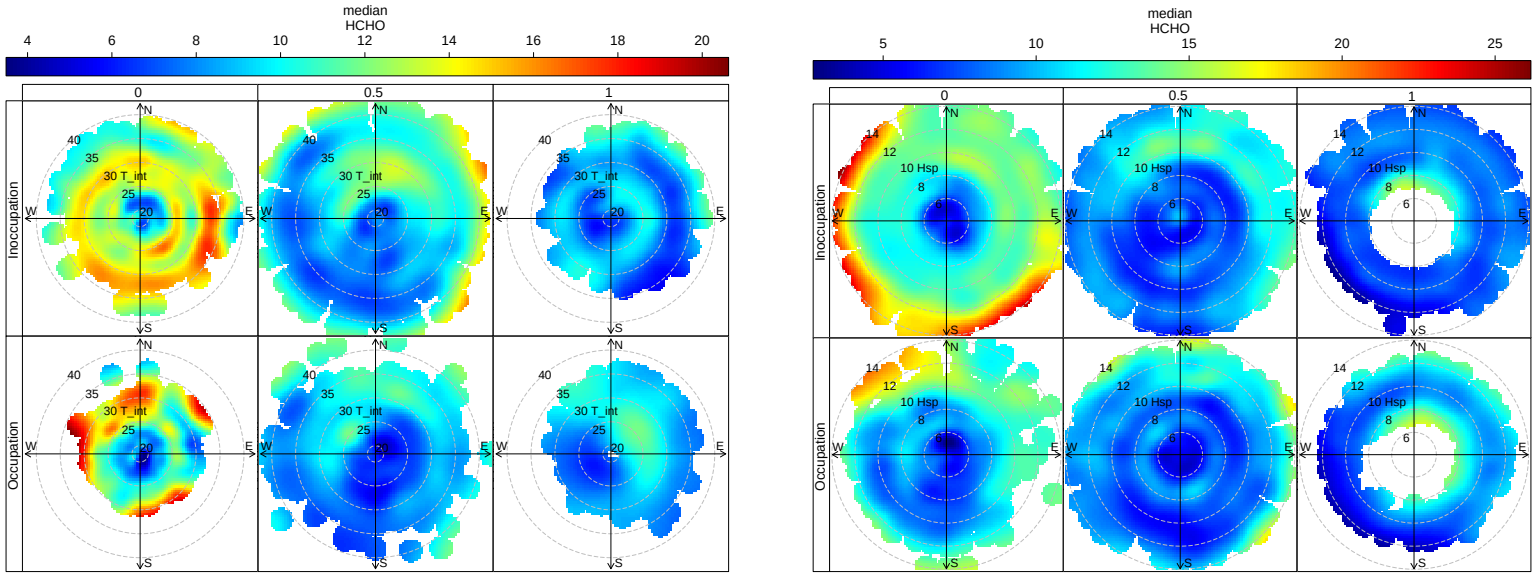


FIGURE 2.5.26 – *Concentration médiane du HCHO avec la variation de la température intérieure (à gauche) et de l'humidité spécifique intérieure (à droite) associée à l'état d'occupation et du ratio des fenêtres ouvertes. Les mesures couvrent la période du 27/04/2013 au 31/07/2013 au pas de temps d'une minute.*

Afin d'approfondir l'analyse de l'influence de l'humidité spécifique sur le HCHO, en cherche l'existence des structures temporelles de HCHO en relation avec les observations de H_{sp}, en appliquant une classification de une classification type *k-médoïdes.*

L'algorithme des $k-médoïdes$ peut être considéré comme une variante de l'algorithme des $k-means$. Cependant, au lieu de prendre la moyenne du cluster comme dans l'algorithme des $k-means$, l'algorithme $k-médoïdes$ considère la donnée qui minimise la somme des distances entre elle-même et le reste des données dans un même cluster comme le médoïde du cluster.

La Figure 2.5.27 montre les résultats de la classification en quatre classes de la variabilité du HCHO indépendamment de l'état des fenêtres et d'occupation, ainsi que la distribution des classes sur toute la période de mesures et leurs contributions horaires.

Hormis la première classe qui détecte la variabilité des concentrations de HCHO pour lesquelles la concentration de l'humidité spécifique est inférieure à $10\,\mathrm{g.kg^{-1}}$(air sec), les trois autres classes sont difficilement séparables.

Le récapitulatif de la variabilité horaire, journalière, hebdomadaire et mensuelle des fractions des classes obtenues par la classification *k-médoïdes* est donné dans la Figure 2.5.28. Clairement, le profil diurne de la première classe reprend globalement les concentrations faibles et sa contribution horaire par jour diminue au cours des mois (*cf.* Figure 2.5.27). Cette classe peut être associée au "bruit" de fond des concentrations de HCHO. On remarque en outre qu'elle est très présente en mois de mars et diminue en été, laissant ainsi la place à la quatrième classe qui traduit les fortes valeurs de H_{sp}.

La deuxième classe est associée au profil diurne de l'ensemble des fluctuations de la concentration de HCHO. D'ailleurs, si on compare le profil diurne de toute la série de HCHO (*cf.* voire la Figure C.2.5) avec celui de la deuxième classe, on remarque une similarité importante dans la manière dont les fluctuations diminuent entre 6 h et 18 h et augmentent après 18 h. La troisième classe dépeint un profil diurne presque stable au cour de la journée, les fluctuations importantes correspondent uniquement aux lundi, mardi et mercredi. Quant à la quatrième classe, elle correspond à une moyenne peu fluctuante dans le temps (entre 8 ppb et 12 ppb) sur l'ensemble des jours.

Bien que la classification a permis de rendre quelques types de variation, il est très difficile d'associer une classe à une source de variabilité. Seule l'intégration des autres paramètres relatifs à l'ouverture des fenêtres semble expliquer ces variations.

2.5.3.3 Campagne 2015 - variabilité du formaldéhyde

Dans cette section nous résumons les résultats de l'analyse statistique concernant la campagne de mesure de 2015 dans l'espace paysager. Les mesures effectuées couvrent la période du 01/01/2015 au 31/07/2015 toutes les minutes, soit environ 305220 valeurs pour chacun des paramètres mesurés. En raison de la forte proportion des valeurs manquantes sur le mois de juillet, nous traitons uniquement les données des six premiers mois de mesure.

Le jeu de données obtenu pour cette dernière campagne présente l'avantage d'avoir les mesures du formaldéhyde en air intérieur et en extérieur. Néanmoins, la résolution temporelle de celles-ci est élargie à 20 minutes pour permettre au module de multiplexage (AL V08 d'Aerolaser GmBh) de recueillir les

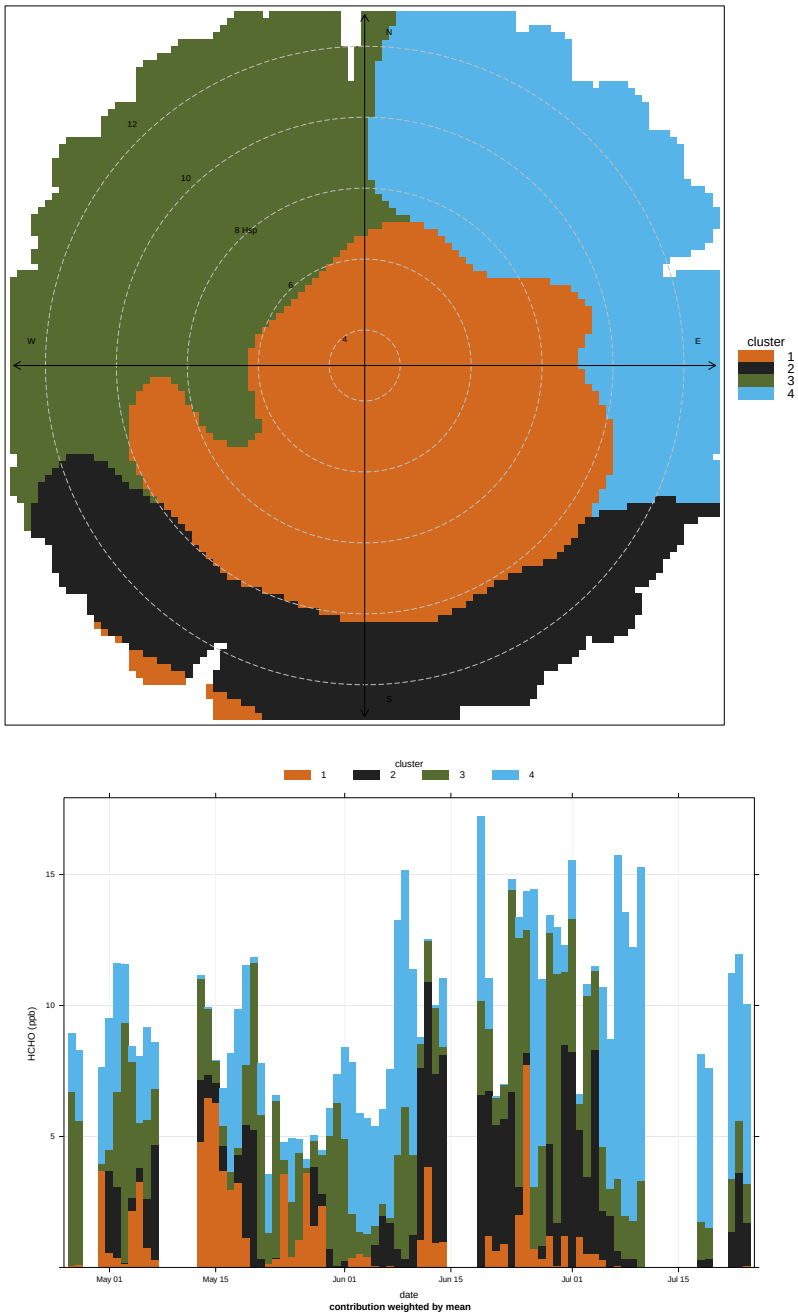


FIGURE 2.5.27 – *Classification k-médoïdes sur les concentrations du formaldéhyde associées à la variable direction du vent et aux concentrations de l'humidité spécifique intérieures. Les contributions temporelles horaires par jour de chaque classe sont données dans le graphe à droite. Le nombre d'éléments dans chaque classe est comme suit : classe* **1**=*11846, classe* **2**=*16053, classe* **3**=*23345 et classe* **4**=*25002.*

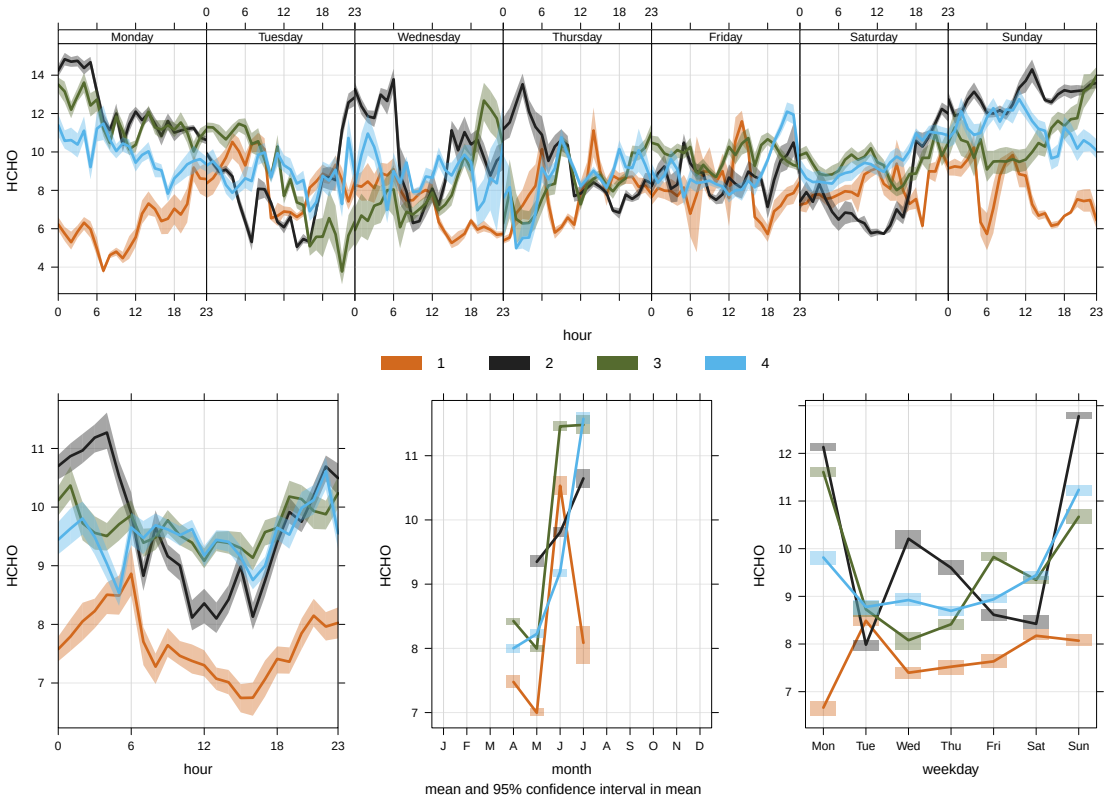


FIGURE 2.5.28 – *Variabilité des quatre classes (médoïdes) de la concentration du HCHO.*

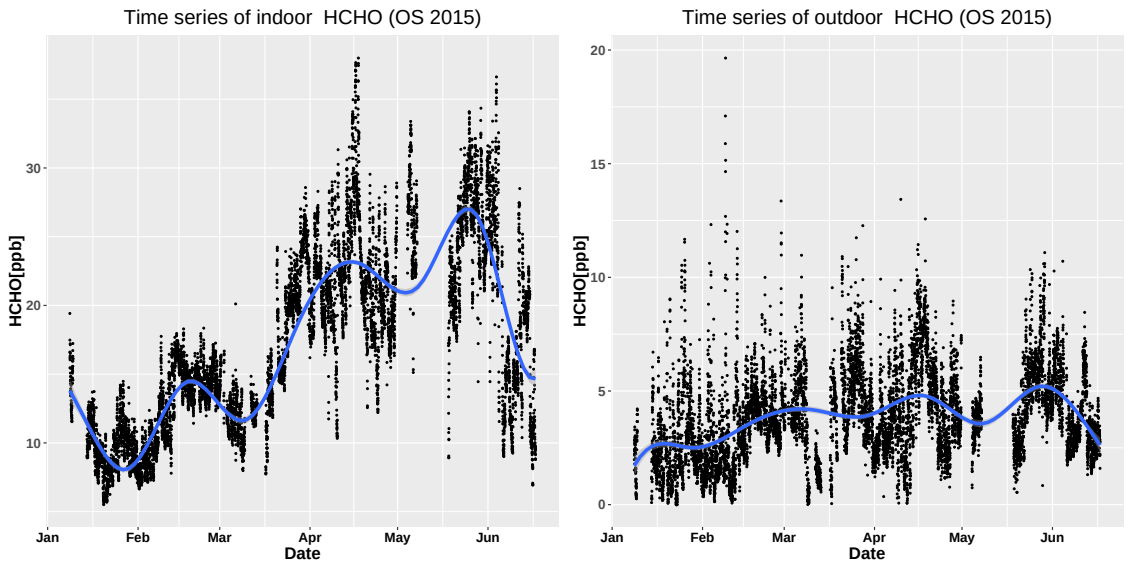


FIGURE 2.5.29 – *Séries temporelles de concentration de HCHO intérieur (à gauche) et extérieur (à droite) issues des mesures en mode séquentiel toutes les 20 minutes durant la période allant du 01/01/2015 au 30/06/2015. La courbe bleu correspond à l'estimation de la tendance non-linéaire par des fonctions splines cubiques.*

données de manière séquentielle. En effet, compte tenu du temps de réponse de l'instrument et des caractéristiques des lignes de prélèvement (en PTFE® longueur 8 m), le changement de voie est effectué toutes les 10 minutes. La dernière valeur de chaque période correspondant à une moyenne sur une minute est conservée. Ainsi, en un point de prélèvement donné, la concentration du formaldéhyde est renseignée toutes les 20 minutes.

Dans le cadre de la thèse seules les données du formaldéhyde sont présentées pour 2015.

Les séries temporelles des concentrations en formaldéhyde dans l'air intérieur et dans l'air extérieur durant la campagne de 2015 sont présentées sur la Figure 2.5.29. La concentration moyenne de HCHO est de 17 ppb en air intérieur et de 3.9 ppb en air extérieur, soit un ratio intérieur/extérieur égal à 4.5. Ce ratio est similaire aux résultats préliminaires de l'étude menée par Weisel et al. (2005) : 3 ppb à l'extérieur et 17 ppb à l'intérieur. En avril et mai, laconcentration en air intérieur augmente et est supérieure à la médiane générale. À l'extérieur, cette particularité n'est pas décelée.

Les niveaux du HCHO intérieur augmentent au cours des mois. Par exemple, durant les mois d'hiver, la concentration moyenne est en dessous de 13 ppb et augmente au mois de mars à 16 ppb jusqu'à atteindre 25 ppb en mois de mai. On observe aussi qu'au mois de juin, le niveau moyen baisse à 18 ppb alors que la variabilité augmente (écart-type = 6.3 ppb). Ceci peut être expliqué par l'aération de l'espace paysager qui intervient de manière significative au mois de juin.

Le maximum observé pour le HCHO intérieur est de 38 ppb et il est survenu à 14 h au mois d'avril.

Quant à la distribution mensuelle des concentrations de HCHO extérieur, elle ne montre aucun profil-type caractérisant les fluctuations. Néanmoins, on note que globalement les niveaux diminuent pendant les week-ends et restent stables durant les heures de la journée.

La Figure 2.5.30 montre la variabilité des concentrations de formaldéhyde intérieure et extérieure par rapport à leurs niveaux normalisés. Contrairement aux mesures effectuées en 2013, la distribution moyenne par heure dépeint un profil diurne pour le HCHO intérieur.

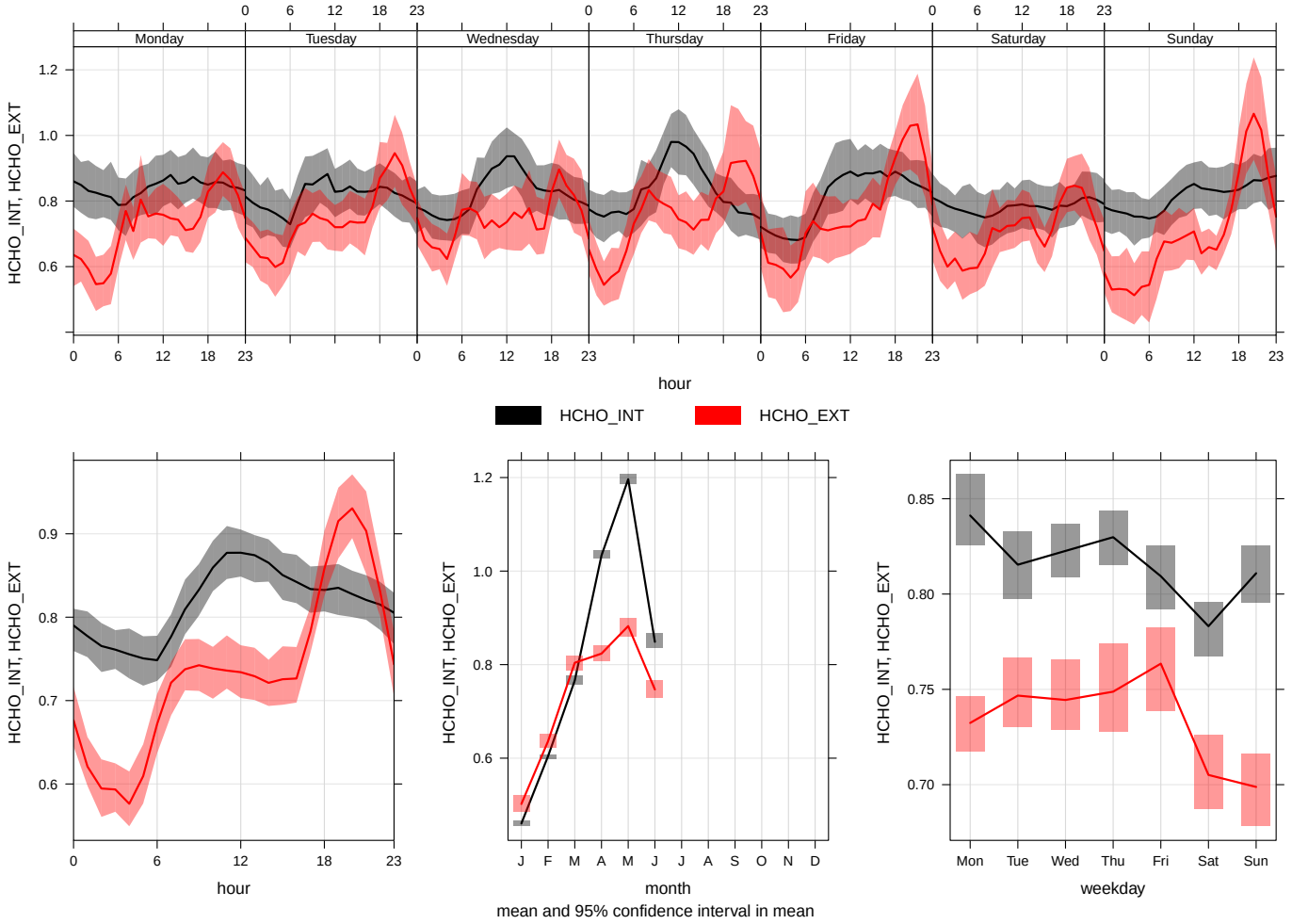


FIGURE 2.5.30 – *Récapitulatif de la variabilité temporelle des concentrations du formaldéhyde intérieures et extérieures. Le premier panel (haut) correspond au profil d'une semaine type sur l'ensemble de la période de mesure (01/01/2015 au 30/06/2015) toutes les 20 minutes. Le second donne respectivement le profil diurne, la variation par mois et la distribution des concentrations moyennes par type de jour (hebdomadaire). Les valeurs normalisées NC (Normalised Concentration), sont calculées par les formules suivantes : pour le HCHO intérieur,* $NC_{HCHO_{int}} = C_{HCHO_{int}} / \left(\bar{C}_{HCHO_{int}} + \bar{C}_{HCHO_{ext}}\right)$ *et pour le HCHO extérieur* $NC_{HCHO_{ext}} = 4 \times C_{HCHO_{ext}} / \left(\bar{C}_{HCHO_{int}} + \bar{C}_{HCHO_{ext}}\right)$*, où* $\bar{C}_x$ *est la moyenne de la variable* x*.*

2.6 Bilan et conclusion

Ce chapitre a pour but de restituer l'information statistique disponible dans toutes les bases de données sur les différents environnements étudiés. Ce survol nous permet d'appréhender -*a priori*- l'importance de certains facteurs par rapport aux autres. Pour ce faire, nous comparons les différents niveaux de fluctuations et les schémas de variation des concentrations afin de saisir les facteurs plausibles qui déterminent le type de variabilité. Cela nous a permis d'appréhender les différences qui pourraient subsister entre les différents environnements et faire une typologie sur les espaces étudiés.

Les premières observations montrent que la nature de la fluctuation dépend du type de polluant et de l'environnement. Ainsi, pour certains polluants, comme par exemple les particules fines, le schéma de variation ne varie pas d'un environnement à un autre. Mais pour les particules de tailles moyennes, un profil diurne, journalier et mensuel se dessine dans les environnements occupés. Ceci montre l'influence de la présence humaine sur les niveaux de certains polluants.

Quant au formaldéhyde, le schéma de variation en l'absence d'occupation est assez régulier et dépend des facteurs physiques, tels que les sources intérieures et les paramètres climatiques. La forme de la variation temporelle du HCHO est quasi-sinusoïdale dans la maison expérimentale. Or dans l'espace de bureaux durant la campagne de 2013, où l'occupation et l'activité des occupants sont importantes, la variabilité du HCHO est erratique et présente des changements abruptes *i.e* variation rapide par rapport à l'échelle de temps et la longueur de la série. En 2015, le profil moyen journalier de la concentration intérieure en formaldéhyde dans l'espace paysager montre un maximum en milieu d'après-midi et un minimum en fin de matinée. Ce profil est également plus marqué les jours de la semaine que le week-end. Cette différence entre les profils d'une campagne à une autre, pourrait être liée au comportement à l'ouverture des fenêtres qui diffère d'un jour à l'autre et d'une année sur l'autre.

Toutefois, les niveaux de variations dans la maison expérimentale sont en moyenne plus importants que ceux observés dans l'espace de bureaux en 2013 et similaires en 2015. Cette comparaison est à prendre à titre indicatif, car elle a été faite à partir des enregistrements effectués sur des périodes différentes et sur des séries de longueurs différentes.

L'étude menée dans la maison expérimentale est un cas assez intéressant dans la mesure où l'occupation et les usages domestiques sont quasi-absents. L'importance de ces facteurs peut être mise en évidence par leur absence.

En définitif, les résultats obtenus des différentes campagnes suggèrent l'existence de plusieurs facteurs qui se combinent pour expliquer les régularités ou/et les irrégularités des fluctuations de la concentration des polluants :

- le type d'environnement étudié : chaque situation réelle mettant en jeu plusieurs phénomènes caractérisés par une combinaison spécifique des différents paramètres d'ambiance (température, humidité et ventilation) ;
- les sollicitations internes au bâtiment : l'occupation et les activités des occupants sont des paramètres qui modifient en profondeur la structure de variabilité temporelle des polluants.

CHAPITRE 3

STRUCTURES DE VARIABILITÉ ET CARACTÉRISTIQUES DES FLUCTUATIONS

QUELLES *sont les structures intrinsèques à la variabilité temporelle des mesures issues de la QAI ? Comment se manifestent-elles sur les différents polluants et comment les exploiter ? Y a-t-il des structures communes pour certains groupes de variables ? Ces questions s'imposent préalablement à l'identification de différents "patterns" (caractéristiques) en vue d'obtenir des indications préliminaires pour la modélisation. L'existence d'une composante aléatoire dans les enregistrements d'un paramètre de l'environnement intérieur reflète non seulement la variabilité inhérente de ce paramètre, mais aussi la variabilité de l'ensemble des phénomènes qui le génèrent. Mesurer une telle quantité nécessite d'identifier certaines caractéristiques de variabilité. Pour ce faire, nous analysons quelques propriétés spectrales des séries. Le degré de stochasticité est une notion essentielle qui se définit comme étant la quantité de l'aléa qui peut se manifester aussi bien quantitativement que qualitativement. La décomposition d'une série temporelle peut révéler au moins deux propriétés, une caractéristique associée à la partie stochastique et l'autre, liée à la partie déterministe. En l'absence de toute étude traitant ce sujet pour la QAI, il nous a paru nécessaire de tester plusieurs méthodes (STL et SSA), afin de pouvoir déboucher une piste de travail, tant théorique que pratique, concernant la prévision par décomposition des séries.*

Sommaire

3.1 Introduction

Dans ce chapitre, nous introduisons d'abord quelques définitions et propriétés des séries temporelles, ainsi que leurs applications aux données de la QAI. Notre objectif est de donner un aperçu sur chacune des principales caractéristiques de l'analyse des séries temporelles et de les appliquer sur les données de la qualité de l'air intérieur.

Selon le type de données et la nature du polluant que l'on doit traiter, là ou les harmoniques sous-jacentes pourront être considérées comme des parties déterministes. En analyse des séries temporelles, la détection de cette caractéristique pourrait mettre en évidence des phénomènes de récurrence dans le temps, ce qui implique que la probabilité d'observer les mêmes caractéristiques dans un futur proche devient progressivement forte.

Ce chapitre est inspiré de la littérature suivante :

1. pour les fondements mathématiques de l'analyse des séries temporelles, les livres classiques de Box & Jenkins (1976) ; Brillinger (2001) ; Fuller (2009) ; Hannan (1966) et Brockwell & Davis (1991) qui sont incontestablement "les références" dans ce domaine. Notons aussi, les livres en langue française ayant une portée mathématique importante, qui sont : Azencott & Dacunha-Castelle (1984) et Gourieroux & Monfort (1995) ;
2. pour l'analyse spectrale des séries temporelles, on se réfère généralement à Koopmans (1995); Priestley (1982). Une revue de littérature sur l'analyse spectrale des données climatiques est exposée par Ghil et al. (2002) et Yiou et al. (1996) ;
3. pour l'analyse de dépendance à long terme, voir Doukhan et al. (2003); Rangarajan & Ding (2003); Taqqu (1988); Robinson (2003); Palma (2007), et pour la notion de l'auto-similarité, Embrechts (2009) ;
4. pour l'application des séries temporelles aux données environnementales, on cite le volume 12 de la monographie "Handbook of Statistics" par Rao & Patil (1994), dédiée à la statistique environnementale. Par ailleurs, le livre de Parnell (2013) expose les problèmes des systèmes complexes environnementaux du point de vue de l'analyse des séries temporelles, ou encore Chandler & Scott (2011) reprennent les différentes méthodes d'extraction de la tendance et leur application dans le domaine de l'environnement.

3.2 Préliminaires de l'analyse des séries temporelles pour la QAI

Nous introduisons succinctement, en guise de préambule, certains concepts de l'analyse des séries temporelles que nous adopterons tout au long du manuscrit. Il s'agit en particulier d'habiller la vision purement descriptive des chroniques de la QAI par des outils statistiques que nous considérerons nécessaires pour aborder la suite, notamment, la notion de prévision dans le cas des modèles paramétriques : l'espérance conditionnelle et le théorème de projection.

3.2.1 Trajectoire d'un processus aléatoire

3.2.1.1 Aspects théoriques

Depuis les axiomes de Kolmogorov (1956) en théorie des probabilités, il est commode de représenter l'intervention de l'aléa dans un phénomène par le choix d'un point ω (appelé épreuve) dans un ensemble adéquat Ω et de considérer les variables aléatoires X_t comme des applications [1] sur Ω à valeurs dans $\mathbb{R}$. L'ensemble Ω contient tous les résultats possibles du phénomène aléatoire étudié.

Pour quantifier l'aléa, il s'agit mathématiquement de construire une famille $\mathcal{A}$ des parties de Ω, formant une structure d'évènements, appelée $\sigma-algèbre$ [2]. La réalisation d'un évènement $A \in \Omega$ de $\mathcal{A}$ ($\sigma-algèbre$) est mesurée sur un espace $(\Omega, \mathcal{A})$ par une probabilité $\mathbb{P}(A) \in [0,1]$. Une famille de variables aléatoires $\{X_t, t \in \mathbb{Z}\}$ toutes définies dans l'espace de probabilité $(\Omega, \mathcal{A}, \mathbb{P})$ est un processus aléatoire (Brockwell & Davis, 1991) :

$$X = \{(X_t, t) \in \mathbb{R} \times \mathrm{T}\}. \tag{3.2.1}$$

On obtient avec cette construction l'objet principal de l'analyse des séries temporelles, c'est-à-dire la famille de variables aléatoires (v.a.) $\{X_t,\, t \in T\}$ accessibles à l'expérience ainsi que leurs propriétés statistiques. Dans l'analyse des séries temporelles, l'ensemble d'indices (ou le paramètre) T est un ensemble de points dans le temps, souvent $T = \{0, \pm 1, \pm 2, \ldots\}$, $\{0, 1, 2, \ldots\}$, $[0, \infty)$ ou $(-\infty, \infty)$. Lorsque $T \subset \mathbb{Z}$, on dit que le processus est à temps discret et, lorsque $T \subset \mathbb{R}$, que le processus est à temps continu. Dans la suite de cette thèse, nous nous intéresserons de façon prioritaire aux processus à temps discret.

Définition 3.2.1. (Trajectoire) On appelle trajectoire d'un processus aléatoire pour toute réalisation d'une épreuve $\omega \in \Omega$, les applications $t \mapsto X_t(\omega)$.

Notons qu'en fait un processus est une application $X : \Omega \times \mathrm{T} \to \mathbb{R}$ telle que :

- à chaque instant $t \in T$, l'application $\omega \longmapsto X(t, \omega) \in (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ est une variable aléatoire, où $\mathcal{B}(\mathbb{R})$ est la tribu borélienne de $\mathbb{R}$;
- pour chaque épreuve $\omega \in \Omega$, l'application $t \longmapsto X(t, \omega)$ est une fonction de $T \to \mathbb{R}$ (ou trajectoire).

1. Plus précisément, des fonctions mesurables de $(\Omega, \mathcal{A})$ dans $(\mathbb{R}, \mathcal{B})$.
2. La structure $\sigma-algèbre$ est une classe de parties de Ω (donc $\mathcal{A} \subset \mathcal{P}(\Omega)$), qui contient Ω ($\Omega \in \mathcal{A}$) et qui est stable par réunion dénombrable (si $(A_i)_{i \in I \subset \mathbb{N}}\, \mathcal{A}_i$ alors $\cup_{i \in I} A_i \in \mathcal{A}$) et par passage au complémentaire (pour $A \in \mathcal{A}$, alors $A^c \in \mathcal{A}$).

Bien que cette construction n'ait pas d'impact direct sur les calculs concrets, elle fournit cependant les ingrédients nécessaires pour définir certaines propriétés statistiques de la prévision, notamment l'importance des espaces des variables de carrés intégrables et l'espérance conditionnelle.

Pour une variable environnementale de l'air intérieur, on dispose en première instance des informations sur les plages de variations de ce paramètre, mais dont les valeurs peuvent varier sans restriction dans ces intervalles.

3.2.1.2 Aspects pratiques

Le phénomène de la qualité de l'air intérieur fournit généralement des suites numériques, et les valeurs prises par un paramètre donné sont des réalisations d'un ensemble de variables aléatoires et forment, avec l'indice temps ($t \in \mathrm{T}$) une "image de la trajectoire". Pour des applications empiriques, on ne dispose que d'un nombre fini de données, donc des réalisations d'un nombre fini de variables $(X(t_1), X(t_2), \ldots, X(t_n))$ pour une suite finie d'instants $\{t_1 < t_2 < \ldots < t_n\}$. Ici, X_{t_i} et $X(t_i)$ sont interchangeables du moment qu'on ne spécifie pas la nature discrète ou continue de T. Comme les séries temporelles traitées sont issues d'un phénomène réel, on travaille dans l'ensemble $\mathbb{Z}$. Dans la suite de ce manuscrit, nous utilisons donc la première écriture, $X(t)$ étant réservée généralement aux processus à temps continu. Pour alléger l'écriture, nous nous distinguerons pas le processus X_t de l'une de ses réalisations $X_t(\omega)$.

La suite d'observations possède au moins deux caractéristiques :

- *(a)* en dehors de l'évolution à peu près polynômiale d'une tendance, les observations présentent des aspects locaux très irréguliers, nécessitant un très grand nombre de degrés de liberté pour pouvoir modéliser la courbe associée aux observations ;
- (b) le processus "physique" est irréversible : impossible de reproduire la suite de mesures dans leurs conditions réelles.

Qu'elle soit exprimée par une concentration ou par un rapport de quantités, la série d'observations associée aux réalisations des variables aléatoires fournit des informations sur l'état du système qui l'a générée. Les informations portées par ces variables sont plus ou moins "bruitées", elles dépendent du type du paramètre étudié et les causes qui l'ont généré. En résumé, l'aspect irrégulier traduit la variabilité des observations, elle est d'autant plus importante que la résolution temporelle est fine.

Selon (a), la suite d'observations $x_1, x_2, \ldots, x_T$ est modélisée par un processus aléatoire X, ce qui revient à dire que $\{x_t,\ t = 1, 2, \ldots, T\}$ est une trajectoire "typique" du processus $\{X_t,\ t \in \mathbb{Z}\}$. En ce qui concerne (b), il est moins évident de traduire formellement les questions d'irréversibilités, néanmoins on peut dire que pour pouvoir déduire le modèle des données, il faut que nos observations $x_1, x_2, \ldots, x_T$ déterminent au moins les lois jointes du processus X_t pour $T \to +\infty$ (Azencott & Dacunha-Castelle, 1984).

3.2.2 La Fonction d'AutoCorrélation (ACF) et la stationnarité

Lorsqu'on travaille sur un nombre fini de variables aléatoires, nous cherchons souvent les dépendances ou des structures de corrélations entre les différents retards des observations, en d'autres termes leur niveau d'autocorrélation. Il existe, entre autres, deux outils permettant d'évaluer l'autocorrélation d'une série : l'autocorrélation simple et l'autocorrélation partielle.

Définition 3.2.2. *La Fonction d'AutoCoVariance ACV*
Soit $\{X_t, t \in T\}$ un processus aléatoire, tel que la $\mathbb{V}ar(X_t) < \infty$ pour tout $t \in T$; alors la fonction d'autocovariance $\gamma_x(\bullet, \bullet)$ de X_t est définie par

$$\gamma_x(r,s) = \mathbb{C}ov(X_s, X_r) = \mathbb{E}[(X_r - \mathbb{E}[X_r])(X_s - \mathbb{E}[X_s])], \text{ pour } r,s \in T. \tag{3.2.2}$$

Pour certaines conditions de la fonction d'autocovariance, on définit la stationnarité au sens large (stationnarité de second degré) :

Définition 3.2.3. *Stationnarité au sens large*
La série temporelle $\{X_t,\, t \in \mathbb{Z}\}$ d'indice $\mathbb{Z} = \{0, \pm 1, \pm 2, \ldots\}$, est dite stationnaire au sens large ou faiblement stationnaire, si

- (i) pour tout $t \in \mathbb{Z}$, X est un processus de second ordre, *i.e.* $\mathbb{E}\left[|X_t|^2\right] < +\infty$,
- (ii) pour tout $t \in \mathbb{Z}$, $\mathbb{E}(X_t) = m_x$ (indépendant de t),
- (iii) pour tout $r,s,t \in \mathbb{Z}$, $\gamma_x(r,s) = \gamma_x(r+t, s+t)$.

Le processus aléatoire est dit stationnaire au sens strict si la structure de probabilité est invariante par translation dans le temps. Sauf indication contraire, les processus considérés sont à valeurs dans $\mathbb{R}$ et le terme "stationnarité" renvoie à la stationnarité au sens large. Dans ce cas, $\gamma_x(r,s) = \gamma_x(r-s, 0)$ pour tout $r,s \in \mathbb{Z}$. Il est donc plus pratique de travailler avec un seul paramètre retard (ou avance) h : $\gamma_x(h) = \gamma_x(h,0) = \mathbb{C}ov(X_{t+h}, X_t)$, pour tout $t, h \in \mathbb{Z}$.

Les matrices de covariance de séquences de n valeurs consécutives du processus X_t sont positives, et elles possèdent de plus une structure de Toëplitz (caractérisée par le fait que $(\Gamma_n)_{i,j} = \gamma(i-j)$) :

$$\begin{aligned} \Gamma_n &= \mathbb{E}\left[[(X_t - m_x)\ldots(X_{t-n+1} - m_x)]^\top [(X_t - m_x)\ldots(X_{t-n+1} - m_x)]\right] \\ &= \begin{bmatrix} \gamma(0) & \gamma(1) & \ldots & \gamma(n-1) \\ \gamma(1) & \gamma(0) & \ldots & \gamma(n-2) \\ \vdots & \vdots & \ddots & \vdots \\ \gamma(n-1) & \gamma(n-2) & \ldots & \gamma(0) \end{bmatrix} \end{aligned} \tag{3.2.3}$$

La fonction d'autocorrélation (ACF, AutoCorrelation Function en Anglais) de X_t se définit de manière analogue par rapport à h comme suit :

$$\rho_x(h) = \mathbb{C}orr(X_{t+h}, X_t) = \frac{\gamma_x(h)}{\gamma_x(0)}, \text{ pour tout } t, h \in \mathbb{Z}. \tag{3.2.4}$$

Ces mesures quantifient l'influence "linéaire" du décalage temporel entre deux observations de la variable X_t. La fonction d'autocorrélation est symétrique ($\rho(h) = \rho(-h)$, pour $h \in \mathbb{Z}$) et à valeurs dans $[-1, 1]$. Cette dernière propriété s'obtient directement de l'inégalité de Cauchy-Schwarz appliquée à $\gamma_x(h)$:

$$|\gamma_x(h)| = |\mathbb{E}[(X_{t+h} - m_x)(X_t - m_x)]| \leq \sqrt{\mathbb{E}\left[(X_{t+h} - m_x)^2\right]\mathbb{E}\left[(X_t - m_x)^2\right]} = \gamma_x(0).$$

Afin d'avoir une première idée de la structure de dépendance temporelle du processus aléatoire, il est fréquent, à partir d'une réalisation de longueur n de la série, soit $X_1, \ldots, X_n$, de chercher à estimer la

fonction d'autocovariance du processus sous-jacent. Cette étape est préliminaire à toute construction d'un modèle approprié. On utilise généralement pour estimer $\gamma_x(h)$, l'autocovariance empirique définie, pour $0 \leq h < n$ par

$$\widehat{\gamma}(h) = \frac{1}{n}\sum_{j=1}^{n-h}\left(X_j - \overline{X}\right)\left(X_{j+h} - \overline{X}\right) = \widehat{\gamma}(-h), \tag{3.2.5}$$

où $\overline{X} = \frac{1}{n}\sum_{j=1}^{n} X_j$ désigne la moyenne empirique. Il est préférable d'utiliser un autre estimateur en remplaçant $1/n$ par $1/(n-h)$ pour que la matrice $\widehat{\gamma}(i-j)$ soit semi-définie positive (Brockwell & Davis (1991), page 221).

Introduisons maintenant deux exemples de processus aléatoires stationnaires : bruits blancs et le processus harmonique. Le premier est particulièrement intéressant pour définir la notion de la prédictibilité selon Goerg (2013).

Définition 3.2.4. *Bruit Blanc (BB)*
Une séquence ε_t est un bruit blanc de moyenne nulle et de variance σ_ε^2 si $\mathbb{E}(\varepsilon_t) = 0$, $\mathbb{E}\left(\varepsilon_t^2\right) = \gamma_\varepsilon(0) = \sigma_\varepsilon^2 < \infty$, $\gamma_\varepsilon(h) = 0$, $\forall h \in \mathbb{Z}$ quand $h \neq 0$; on le notera $\{\varepsilon_t\} \sim BB\left(0, \sigma_\varepsilon^2\right)$. On appelle bruit blanc fort, toute suite du second ordre de variables aléatoires $\{\varepsilon_t\}$, centrées, indépendantes et identiquement distribuées *(iid)* de variance finie $\left(\mathbb{E}\left(\varepsilon_t^2\right) = \sigma_\varepsilon^2 < +\infty\right)$. On le notera $\{\varepsilon_t\} \sim iid\left(0, \sigma_\varepsilon^2\right)$.

En général, il n'est pas nécessaire de faire l'hypothèse de bruit blanc fort lorsque l'on s'intéresse à des modèles de séries supposées stationnaires au second ordre. Dans le cas de bruit blanc faible, il est important à noter qu'aucune hypothèse d'indépendance n'est faite.

Définition 3.2.5. *Processus harmonique*
Soient N variables aléatoires (**v.a.**) $\{A_k\}_{1\leq k\leq N}$ et $\{\Phi_k\}_{1\leq k\leq N}$ telles que :

- *(i)* $\mathbb{C}ov(A_k, A_l) = \sigma_k^2 \delta(k-l)$, où $\delta(\bullet)$ est l'impulsion de Dirac,
- *(ii)* $\{\Phi_k\}_{1\leq k\leq N}$ une suite de **v.a.** indépendantes et identiquement distribuées $(i.i.d)$, de loi uniforme sur $[-\pi, \pi]$,
- *(iii)* $\{A_k\}_{1\leq k\leq N}$ sont indépendantes de $\{\Phi_k\}_{1\leq k\leq N}$.

Un processus harmonique est un processus défini comme suit :

$$X_t = \sum_{k=1}^{N} A_k \cos(\lambda_k t + \Phi_k), \tag{3.2.6}$$

où $\{\lambda_k\} \in [-\pi, \pi]$ représentent N pulsations.

On vérifie que $\mathbb{E}(X_t) = 0$ et que sa fonction d'autocovariance est donnée par

$$\gamma(h) = \mathbb{E}[X_{t+h}X_t] = \frac{1}{2}\sum_{k=1}^{N}\sigma_k^2 \cos(\lambda_k h), \tag{3.2.7}$$

où $\sigma_k = \mathbb{E}\left\{A_k^2\right\}$. Le processus harmonique est donc stationnaire au second ordre.

Les deux deux processus que nous venons de définir ont la particularité suivante : un processus bruit blanc ne peut pas être prédit de façon linéaire à partir de son passé, tandis que le processus harmonique est prédictible à partir de son passé. Cette constatation est une conséquence du célèbre théorème de décomposition de Wold (Voir Brockwell & Davis (1991), pages 187-189, pour plus de détails).

On s'intéresse plus précisément aux structures de décroissance de l'ACF. Les processus à mémoire courte se manifestent par une décroissance rapide (exponentielle) de l'ACF vers zéro. Au contraire, une série comporte une mémoire longue si l'ACF décroit très lentement (fonction puissance) vers zéro comme :

$$\rho(h) \sim h^{-v}, \text{ pour } h \to \infty. \tag{3.2.8}$$

Pour certaines valeurs de v dans la relation 3.2.8, on peut s'attendre à ce que $\sum_h \rho(h) \longrightarrow \infty$, d'où le qualificatif de "longue mémoire" pour la série chronologique. Les conséquences de cette caractéristique se traduisent par la répercussion d'un choc aléatoire sur un nombre très important de valeurs futures ; ce choc affectera donc le comportement à long terme de la série.

3.2.3 La Fonction d'AutoCorrélation Partielle (PACF)

Lorsqu'on cherche des corrélations entre les retards du processus, par exemple entre t et $t+h$, souvent on s'intéresse à l'influence exacte d'une observation passée sur la valeur courante du processus en ôtant toutes les observations intermédiaires. Cette mesure de corrélation est plus délicate à interpréter, mais elle est formalisée sous le nom d'autocorrélation partielle.

L'autocorrélation partielle $\alpha(h)$ au décalage temporel h peut être vue comme la corrélation entre X_1 et X_{h+1}, ajustée sur les variables intermédiaires $X_2, \ldots, X_h$. Sa construction récursive, appelée aussi " algorithme Durbin-Levinson" nécessite de construire une suite $\phi_{h,j}$ où l'indice $h \in \mathbb{N}^*$ est le décalage (lag) temporel et l'indice j varie entre 1 et h (voir la proposition 5.2.1 de (Brockwell & Davis, 1991), et les pages 97-99 pour plus de détails). Soit d'abord $\phi_{1,1} = \alpha(1) = \rho(1)$ et, pour tout $h > 1$, on construit

$$\phi_{h,h} = \left[\rho(h) - \sum_{k=1}^{h-1} \phi_{h-1,k}\rho(h-k)\right]\left[1 - \sum_{k=1}^{h-1} \phi_{h-1,k}\rho(k)\right]^{-1} \tag{3.2.9}$$

où, lorsque $j < h$, $\phi_{h,j} = \phi_{h-1,j} - \phi_{h,h}\phi_{h-1,j-1}$. La valeur de $\phi_{h,h}$ correspond au coefficient de corrélation linéaire entre les résidus issus de la régression X_h et celle de X_0 sur les observations intermédiaires, respectivement.

Pour tout $h \in \mathbb{N}^*$, la suite $\phi_{h,h}$ peut être décrite formellement par la formule suivante :

$$\phi_{h,h} = \alpha(h) = \mathbb{C}orr\left(X_h - P_{\overline{sp}\{1,X_1,\ldots,X_{h-1}\}}X_h, X_0 - P_{\overline{sp}\{1,X_1,\ldots,X_{h-1}\}}X_0\right), \tag{3.2.10}$$

où $P_{\overline{sp}\{1,X_1,\ldots,X_{h-1}\}}$ est la projection orthogonale de toute variable aléatoire de l'espace $\mathcal{L}^2(\Omega, \mathcal{A}, \mathbb{P})$ sur le sous-espace fermé engendré par $\{1, X_1, \ldots, X_{h-1}\}$.

Nous rappelons dans la section suivante la transformée de Fourier, qui représente la base de l'analyse spectrale. Il s'agit d'un autre outil pour représenter les séries temporelles, outil sans doute indispensable qui va permettre de décomposer les séries en fonction de leurs périodicités. Ainsi, les basses fréquences vont représenter les dynamiques de long terme et de faible variabilité tandis que les hautes fréquences vont décrire les dynamiques de court terme.

3.2.4 La densité et la mesure spectrale

D'un point de vue mathématique, une suite d'observations de plusieurs paramètres environnementaux de longueur n est représentée par des vecteurs dans un espace de dimension $\mathbb{R}^n$. Afin de mieux comprendre l'information contenue dans ces séries temporelles, il est commode de changer de repère ou de base dans l'espace $\mathbb{R}^n$. Il se trouve qu'il existe une base particulièrement adaptée aux séries temporelles stationnaires : la base de Fourier.

En effet, il est bien connu que les autocovariances d'un processus stationnaire $\{X_t, t \in \mathbb{Z}\}$ coïncident avec les coefficients de Fourier d'une mesure positive, appelée "mesure spectrale" (voir le théorème de Herglotz E.2 en annexe E). On suppose que cette mesure admet une densité par rapport à la mesure de Lebesgue sur le tore $\mathbb{T} = [-\pi, \pi]$.

Théorème 3.2.6. *(Densité spectrale) Soit le processus aléatoire $\{X_t, t \in \mathbb{Z}\}$ et sa fonction d'autocovariance γ_x de carré sommable* i.e. $\sum_{h\in\mathbb{Z}} \gamma_x(h)^2 < \infty$. *Il existe une densité spectrale $S \in \mathcal{L}^2([-\pi, \pi])$, dont la représentation sous forme de série de Fourier est donnée par*

$$S_x(\lambda) = \frac{1}{2\pi} \sum_{h\in\mathbb{Z}} \gamma_x(h)\, e^{-i\lambda h}, \quad \lambda \in \mathbb{T}. \tag{3.2.11}$$

On peut consulter l'ouvrage (Giraitis et al., 2012) page 11 pour une démonstration et les références suivantes pour plus de considérations mathématiques : (Priestley, 1982; Brockwell & Davis, 1991) ou (Brillinger, 2001).

L'intérêt pratique de la représentation de Fourier tient en très grande partie à l'existence d'une transformée de Fourier inverse :

$$\gamma_x(h) = \int_{\mathbb{T}} S_x(\lambda)\, e^{i\lambda h} d\lambda. \quad h = 0, \pm 1, \pm 2, \ldots, \tag{3.2.12}$$

D'après le théorème 3.2.11 et la définition du bruit blanc (*cf. Définition* 3.2.4), nous avons alors

$$S_\varepsilon(\lambda) = (2\pi)^{-1} \sigma_\varepsilon^2, \quad \lambda \in [-\pi, \pi]. \tag{3.2.13}$$

Pour le processus harmonique, défini dans 3.2.5, la suite des coefficients d'autocovariance n'est pas sommable au sens des conditions du théorème 3.2.6. Donc la mesure spectrale n'admet pas de densité : le processus harmonique admet une mesure spectrale mais pas une densité spectrale. En effet, soit la mesure de Dirac au point a, notée δ_a ; alors la mesure spectrale du processus harmonique [3] peut s'écrire :

$$\nu(d\lambda) = \frac{1}{4} \sum_{k=1}^{N} \sigma_k^2 \delta_{\lambda_k}(d\lambda) + \frac{1}{4} \sum_{k=1}^{N} \sigma_k^2 \delta_{-\lambda_k}(d\lambda). \tag{3.2.14}$$

La mesure spectrale d'un processus harmonique apparaît donc comme une somme de mesures de Dirac, dont les masses σ_k^2 sont localisées aux fréquences des différentes composantes harmoniques (Douc et al. (2014), page 12).

3. La définition et les propriétés d'un processus harmonique sont traités dans le cas réel ($\mathbb{R}$) ; pour les processus à valeurs complexes, on peut consulter les chapitres 4 et 5 dans Brockwell & Davis (1991) et le chapitre 1 (pages 11-12) du récent ouvrage de Douc et al. (2014).

3.3 Mesure de la prédictibilité au sens de Goerg

La description du degré de variabilité, en tant que structure inhérente aux fluctuations des concentrations, nous préoccupe dans cette thèse. Les observations associées à la qualité de l'air intérieur, comme nous l'avons vu précédemment dans les processus aléatoires, n'est que la réalisation d'un nombre fini de variables aléatoires. En statistique et en théorie de l'information, une variable aléatoire peut être appréhendée par la notion d'entropie de Shannon (1948), plus particulièrement par l'entropie différentielle de sa densité de probabilité (Michalowicz et al., 2013). Selon Cover & Thomas (2012), *"entropy is the minimum descriptive complexity of a random variable..."*, elle nous offre donc un autre outil en vue de mettre en évidence quelques structures de variabilités temporelles des paramètres de l'air intérieur.

Introduisons maintenant l'entropie différentielle au sens de Goerg (2013) et présentons ses arguments sur la mesure de prédictibilité. Pour ce faire, reprenons d'abord la formule 3.2.13 de la densité spectrale d'un processus bruit blanc normalisé :

$$f_{\varepsilon}(\lambda) = \frac{S_{\varepsilon}(\lambda)}{\sigma_{\varepsilon}^2} = \frac{1}{2\pi}, \text{ avec } \lambda \in [-\pi, \pi]. \tag{3.3.1}$$

Le même procédé peut être appliqué à un processus aléatoire stationnaire $\{X_t, t \in \mathbb{Z}\}$,

$$f_X(\lambda) = \frac{S_X(\lambda)}{2\pi} = \frac{1}{2\pi} \sum_{h=-\infty}^{+\infty} \rho_X(h) e^{ih\lambda} d\lambda. \tag{3.3.2}$$

Cette forme de densité possède -presque- les mêmes propriétés qu'une fonction de densité de probabilité d'une variable aléatoire : $f_X(\lambda) \geq 0$ et $\int_{-\pi}^{\pi} f_X(\lambda)\, d\lambda = 1$. Sur cette caractéristique et selon Goerg (2013), le niveau de prédictibilité d'un processus aléatoire stationnaire peut être estimé par la mesure de l'entropie différentielle de $f_X(\lambda)$

$$H(X_t) = -\int_{-\pi}^{\pi} f_X(\lambda) \log f_X(\lambda)\, d\lambda. \tag{3.3.3}$$

Au regard de cette considération et de ce que nous avons évoqué précédemment sur les processus $\{\varepsilon_t\}_{t\in\mathbb{Z}} \sim BB\left(0, \sigma_{\varepsilon}^2\right)$, alors

$$\begin{aligned} H(\varepsilon_t) &= -\int_{-\pi}^{\pi} \frac{1}{2\pi} \log\left(\frac{1}{2\pi}\right) d\lambda && (3.3.4)\\ &= \int_{-\pi}^{\pi} \frac{\log(2\pi)}{2\pi} d\lambda && (3.3.5)\\ &= \log(2\pi). && (3.3.6) \end{aligned}$$

Clairement, $H(\varepsilon_t)$ doit correspondre au niveau maximal "d'impredictibilité" d'un processus stationnaire, donc pour tout processus stationnaire X_t

$$H(X_t) \leq H(\varepsilon_t) = \log(2\pi). \tag{3.3.7}$$

Définition 3.3.1. *Prédictibilité au sens de* Goerg
Pour un processus faiblement stationnaire X_t, la fonction

$$\begin{aligned} \Omega_g : \quad X_t &\longmapsto [0, \infty) \\ \Omega_g(X_t) &= 1 - \frac{H(X_t)}{\log(2\pi)} \end{aligned}$$

définit la prédictibilité de X_t.

Cette mesure sert non seulement à quantifier la prédictibilité (au sens de Goerg) mais aussi à décomposer un ensemble de séries temporelles en composantes, appelées composantes "plus ou moins" prédictibles.

Remarque 3.3.2. Nous avons cherché une mesure qui se construit autour de deux processus : le premier étant imprédictible (comme le bruit blanc) et l'autre entièrement prédictible. Ce dernier processus existe, prenant par exemple le processus harmonique défini dans 3.2.5, s'il existe un rang n pour lequel la matrice de covariance Γ_n définie dans 3.2.3 est non inversible, le processus correspondant X_t est prédictible dans le sens où il existe une combinaison linéaire $a_1, \ldots, a_p$ avec $p \leq n-1$ telle que $X_t = \sum_{i=1}^{p} a_i X_{t-i}$, l'égalité ayant lieu presque sûrement ; voir Brockwell & Davis (1991) et Douc et al. (2014) pour la démonstration. Ce résultat est sans surprise compte tenu du fait que les trajectoires de ce processus sont des sommes de sinusoïdes de fréquences $\lambda_1, \lambda_2, \ldots, \lambda_N$ dont seules les amplitudes et les phases sont aléatoires. Néanmoins, la fonction d'autocovariance d'un processus harmonique, donnée dans 3.2.7 est non-absolument sommable, elle admet une mesure spectrale (voir le théorème d'Herglotz E.2 en annexe E) mais pas une densité spectrale. Le problème est donc que la densité spectrale n'existe pas, par conséquent $f_X(\lambda)$ et $\Omega_g(X_t)$ telles que définies dans 3.3.2 et dans 3.3.1 n'existent pas, non plus. On se contente pour le moment de la mesure proposée par Goerg.

3.4 L'analyse spectrale et prédictibilité des données de la QAI

3.4.1 Résultats sur l'analyse spectrale des séries temporelles issues des mesures de la QAI

La Figure 3.4.1 montre les densités spectrales et les fonctions d'autocorrélation de la concentration en CO_2, en HAP et en particules de taille 0.35μm dans un bureau individuel. Les propriétés descriptives de ces chroniques sont discutées dans les sections 2.5.1.1, C.1.2, 2.5.1.2, respectivement. Une première caractéristique spectrale partagée par de nombreux polluants se dégage : les spectres exhibent des pôles à la fréquence zéro, et ce, quelle que soit la résolution temporelle où l'étendu de la série.

Pour les enregistrements du CO_2 au pas de temps de 10 minutes, la période dominante est de 1 jour suivie par plusieurs pics au niveau des basses fréquences $((7\,\text{jour})^{-1}$, $(3.4\,\text{jour})^{-1}$, $(12\,\text{h})^{-1}$, $(8\,h)^{-1})$. Quant aux fluctuations des HAP_s, échantillonnées au pas de temps d'une minute, la fréquence dominante se situe au niveau de $(12\,\text{h})^{-1}$ et aucune fréquence ne prédomine le spectre au delà de cette fréquence. Par contre,

pour les particules de taille 0.35μm, il est moins évident de distinguer une fréquence fondamentale. Il apparaît que ce type de spectre n'est pas identique pour toutes les tailles, notamment pour les particules de tailles moyennes (voire l'annexe C.2.1), leurs spectres de puissance présentent des pics inférieurs à la fréquence d'un jour^{-1}, pareil que le CO_2.

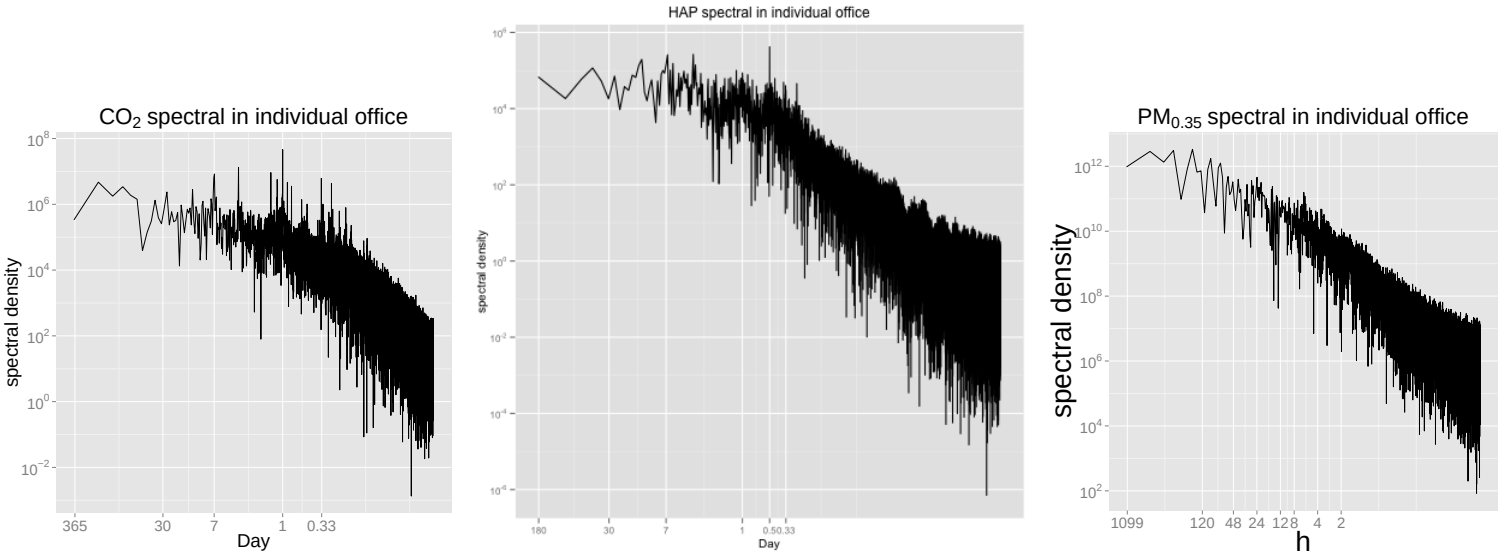
Visuellement, on remarque qu'au niveau de la vitesse de décroissance pour les spectres du CO_2 et des HAP_s, deux parties peuvent être analysées séparément : une décroissance moins rapide sur les basses fréquences (premier niveau de fluctuation) et plus rapide au niveau des hautes fréquences (deuxième niveau de fluctuation). En revanche, la vitesse de décroissance en bi-log du spectre des particules fines est quasi-linéaire et se décline sur un seul morceau. Pour ce dernier cas, la densité spectrale en échelle $\log - \log$ présente une forme de bruit de type $f^{-\alpha}$. Nous détaillerons cette caractéristique dans les prochaines sections.

La fonction d'autocorrélation peut renseigner sur quelques caractéristiques liées à la persistance du processus générateur. En tant qu'indicateur graphique, l'autocorrélogramme distingue, au moins qualitativement, trois types de mémoire : sans mémoire, à mémoire courte ou à mémoire longue. Pour les polluants étudiés, nous observons une décroissance de l'ACF assez lente pour le CO_2 et très lente pour les HAPs et les particules de taille 0.35 μm.

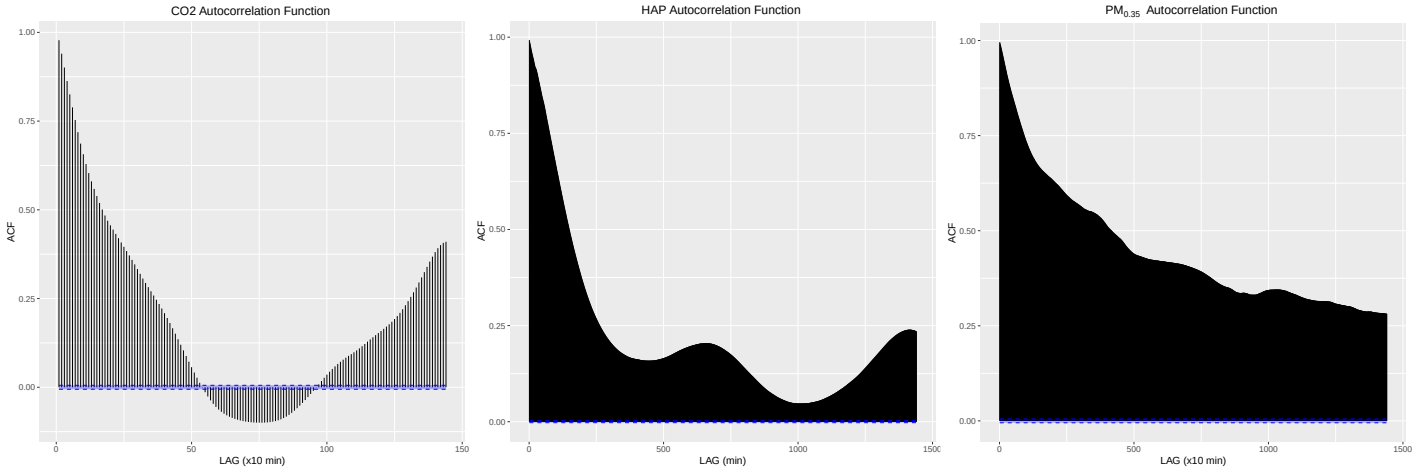
L'autocorrélation devient nulle au bout de 8.33 h, 30 h et 2.5 jours pour le CO_2, les HAP_s et les particules, respectivement (*cf.* Figure 3.4.1b). En d'autres termes, cela signifie qu'en présence de persistance pour les particules fines, la concentration à l'instant t peut avoir un impact sur la concentration 2.5 jours après, alors qu'en présence d'une mémoire "moins longue" comme pour le CO_2, l'impact est de l'ordre de $1/3$ jour.

Nous constatons par ailleurs que l'ACF des concentrations de CO_2 devient négative et reste à des niveaux faibles (< -0.12) et re-bascule sur des valeurs positives après un lag de 98 min. Quant aux particules fines et aux HAP_s, les autocorrélations persistent dans le positif pour des retards grands.

Les propriétés spectrales des autres gammes de particules dans le bureau individuel sont rapportées dans la Figure D.0.2 en annexe D. Globalement, les particules de tailles moyennes dépeignent les mêmes structures de variabilité spectrale que le CO_2 : deux pics de fréquences fondamentales ($(24\,\mathrm{h})^{-1}$ et $(8\,\mathrm{h})^{-1}$) et l'ACF alterne de signe toutes les 500 min sur un lag de 1500 min.



(a) *Densité spectrale de la concentration de CO_2, des HAP_s et de $PM_{0.35}$ dans le bureau individuel. L'abscisse est donnée en période (h) correspondante à la fréquence $(1/f)$.*



(b) *Autocorrélogramme des concentrations de CO_2, des HAP_s et de $PM_{0.35}$ dans le bureau individuel.*

FIGURE 3.4.1 – *Propriétés spectrales de différents polluants dans le bureau individuel.*

La Figure 3.4.2 montre les densités spectrales de fluctuations des quelques paramètres dans l'espace paysager (campagne 2012). Rappelons que pour les particules, 15 gammes de taille allant de $0.35\,\mu$m à $> 20\,\mu$m ont été mesurées à l'origine, mais seules 5 gammes agrégées ont été retenues pour cette étude, car très corrélées entre elles.

Contrairement aux fluctuations dans le bureau individuel d'une résolution d'une minute, les fréquences obtenues à partir des mesures dans l'espace paysager mettent bien en évidence une fréquence principale d'un $jour^{-1}$. Pour le CO_2, pratiquement les mêmes fréquences sont observées dans les deux environnements : $(8\,\mathrm{h})^{-1}$, $(12\,\mathrm{h})^{-1}$, $(24\,\mathrm{h})^{-1}$, $(168\,\mathrm{h})^{-1}$ suggérant que la variabilité du CO_2 est indépendante de l'environnement, elle dépend uniquement des fréquences de l'occupation.

Par ailleurs, le comportement de décroissance du spectre au niveau des hautes et basses fréquences est diffèrent entre les deux environnements. La densité spectrale des particules de taille $0.35\,\mu$m affiche néanmoins une période principale au niveau d'un jour avec une tendance à la baisse de manière linéaire. Cette fréquence peut être attribuée à un *artefact* lié à l'agrégation des autres gammes sur cette taille, mais le déclin du spectre est presque similaire entre les deux environnements. Pour les tailles 0.9 et $1.8\,\mu$m, on observe une fréquence d'un jour^{-1} qui se détache du périodogramme, constituant ainsi un argument en faveur de la cyclicité de la variabilité des particules.

À partir de la taille $1.8\,\mu$m, deux nouvelles fréquences importantes de $(8\,\mathrm{h})^{-1}$ et de $1\,\mathrm{semaine}^{-1}$ apparaissent. Plus les particules sont grosses, plus la densité spectrale exhibe des ruptures au niveau de ces deux fréquences, traduisant la présence d'au moins deux niveaux de variabilité. En effet, pour les tailles $4.5\,\mu$m et $8.75\,\mu$m, leurs densités présentent une allure quasi-plate sur les très basses fréquences $\left(> 1\,\mathrm{semaine}^{-1}\right)$ suivie par un autre niveau monotone sur une bande de $(168\,\mathrm{h})^{-1} - (8\,\mathrm{h})^{-1}$ et enfin une décroissance linéaire entre les fréquences $(8\,\mathrm{h})^{-1}$ et $1\,\mathrm{h}^{-1}$. Cette caractéristique, observée aussi pour le CO_2, reflète l'existence d'une certaine loi d'échelle déterminant les différentes structures de variabilité.

Par rapport aux périodogrammes de la température et de l'humidité spécifique intérieures, un pic fondamental à la fréquence de $1\,\mathrm{jour}^{-1}$ et une fréquence secondaire de $(12\,\mathrm{h})^{-1}$ dominent le spectre (Voir annexe D.0.1). Ces observations sont sans surprise du fait que les paramètres climatiques ont généralement une variation cyclique diurne.

Au regard de l'analyse des autocorrélations, on observe un comportement similaire entre le CO_2 et les particules de taille $8.75\,\mu$m : une oscillation sinusoïdale sur les 400 h de retard avec une valeur nulle au bout de 8 h d'intervalle. Pour la gamme $4.5\,\mu$m, le premier retard pour une ACF nulle survient après 35 h, tandis que l'ACF s'annule après 8.3 jours de retard pour la gamme $1.8\,\mu$m. Plus la taille des particules est fine, plus la structure des séries est persistante : l'information statistique se propage au cours du temps. Ainsi pour la gamme $0.35\,\mu$m, le premier retard pour lequel l'ACF devient nul est séparé de 15 jours du lag 1 (premier lag).

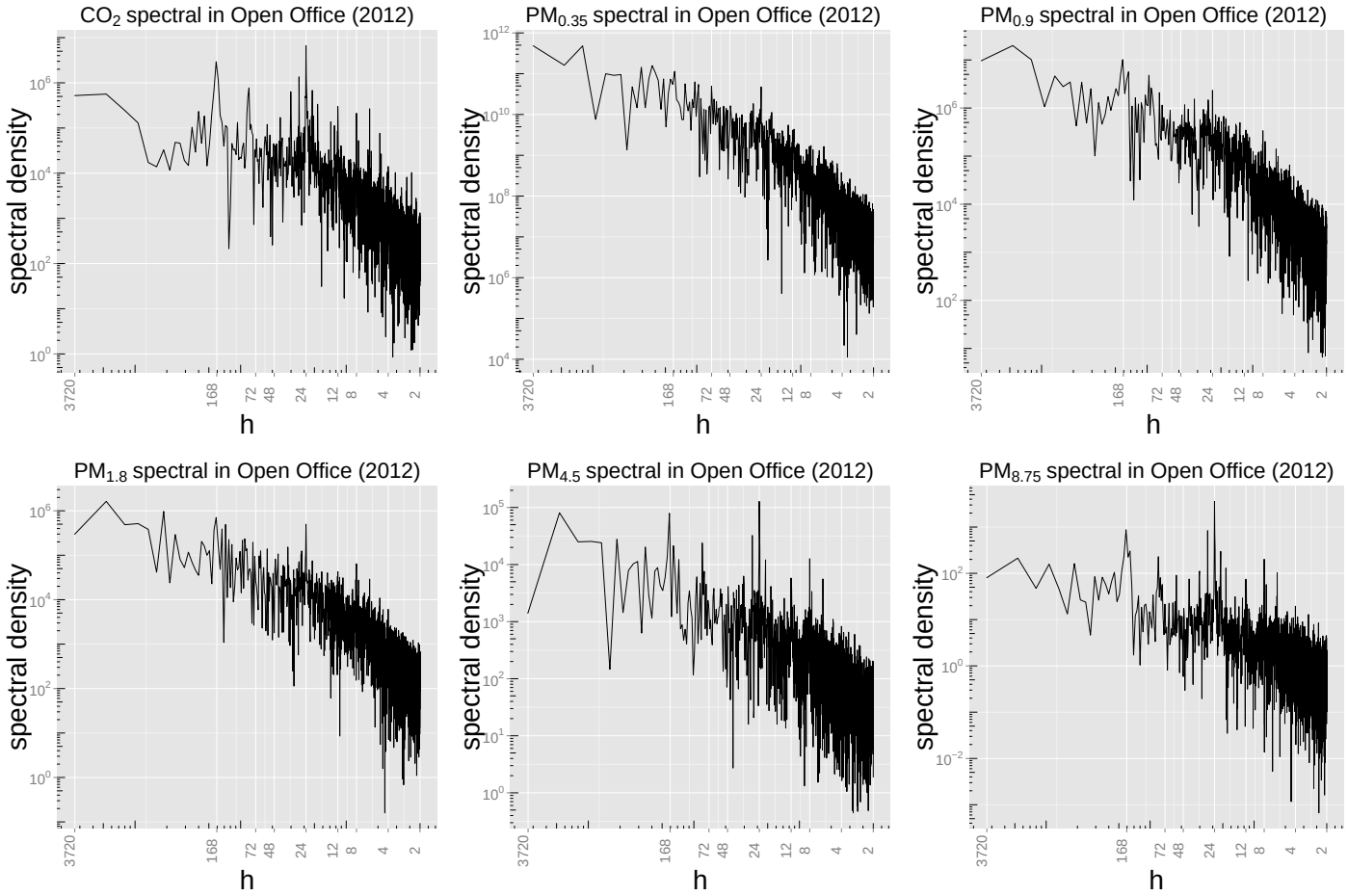


FIGURE 3.4.2 – *Densités spectrales des différents paramètres observées dans l'espace de bureaux durant la campagne de 2012 (Six mois de mesures au pas de temps horaire). L'abscisse est donnée en période (h) correspondante à la fréquence* $(1/f)$.

En revenant aux différents types de décroissances des spectres, la décroissance lente des ACF observées sur plusieurs paramètres peut refléter diverses propriétés statistiques : présence des composantes déterministes comme la tendance ou la saisonnalité, dépendance à long terme etc. Du point de vue de la QAI, on peut expliquer ce phénomène par trois mécanismes "plausibles" :

- agrégation de plusieurs micro-facteurs et variables en se combinant sur différentes échelles temporelles (essentiellement à mémoires courtes) ;
- changements abrupts dans la variabilité (non-stationnarité en moyenne) ou/et la présence des comportements explosifs (non-stationnarité en variance) ;
- présence de nonlinéairités dans le système.

Ces mécanismes peuvent être dépendants au point que certains indiquent la présence des autres. Par exemple, la présence d'une fluctuation accidentée est un bon marqueur de la présence d'un certain type de nonlinéarité. Aussi, la nonlinéairité peut se manifester par la présence des structures de non-stationnarité.

La prise en compte du type de mémoire est un aspect capital de la modélisation des séries temporelles. En effet, elle peut contribuer à une meilleure compréhension de la variabilité, et par conséquent à une meilleure prévision. Dans la section suivante, nous allons aborder les aspects théoriques ainsi que leurs applications aux fluctuations des mesures issues de la QAI.

Quoi qui l'en soit, tous ces traits traduisent la complexité du phénomène et la classification n'est pas une chose aisée.

Du point de vue de l'instrumentation de l'environnement intérieur, l'information sur la mesure du CO_2 peut être décrite uniquement sur quelques jours : l'influence des valeurs très lointaines n'est pas très importante. Au contraire, l'importance des retards sur la concentration immédiate des particules fines nécessite un historique long et un pas de temps fin.

3.4.2 Résultats sur la prédictibilité des séries temporelles des données de la QAI

3.4.2.1 Application directe de la mesure Ω_g sur les séries de la QAI

Au regard de la définition 3.3.1, Goerg (2013) propose une estimation de la prédictibilité Ω par l'analyse du spectre normalisé de la série temporelle :

$$\hat{f}_{j,X} = \frac{\widehat{S_X}(\lambda_j)}{\sum_{j=0}^{T-1} \widehat{S_X}(\lambda_j)}. \tag{3.4.1}$$

Pour un échantillon $\boldsymbol{x_T} = x_0, x_1, \ldots, x_T$ de X, la prédictibilité estimée est basée sur l'entropie discrète

$$\widehat{\Omega_g}(\boldsymbol{x_T}) = 1 + \sum_{j=0}^{T-1} \hat{f}_{j,X} \cdot \log\left(\hat{f}_{j,X}\right). \tag{3.4.2}$$

Ces relations traduisent le niveau de prédictibilité sur un intervalle de $[0, 1]$: $\widehat{\Omega_g}(\boldsymbol{x_T}) = 0 \Leftrightarrow \boldsymbol{x_T}$ issues d'un bruit blanc, $\widehat{\Omega_g}(\boldsymbol{x_T}) = 1 \Leftrightarrow$ l'échantillon $\boldsymbol{x_T}$ est un sinus parfait (Goerg, 2013).

Nous appliquons cette mesure aux données de la QAI, d'abord sur les données brutes, ensuite sur les séries différenciées. Le Tableau 3.4.1 donne l'estimation de $\hat{\Omega}_g$ sur les séries temporelles brutes de polluants, de la température et de l'humidité spécifique. Sur le Tableau 3.4.2, on prend la différence première des

concentrations et on estime la prédictibilité sur les données transformées. Cette transformation absorbe toutes les composantes déterministes, ce qui laisse présager des séries stationnaires.

Hormis la méthode basée sur le périodogramme pour l'estimation de la densité spectrale, la méthode Weighted Overlapping Segment Averaging (WOSA (Nuttall & Carter, 1982)) et la méthode Multitaper (Walden, 1989; Percival & Walden, 1993) donnent pratiquement les mêmes résultats pour les séries non-transformées.

En appliquant la méthode sur les données brutes, l'estimation de la prédictibilité au sens de Goerg donne une forte prédictibilité pour les particules fines ($0.35 - 0.45\,\mu$m) et inversement proportionnelle à la taille de ces particules. Ce résultat est complètement à l'inverse de nos suppositions, on s'attendait à ce que la prédictibilité soit plus grande pour les particules moyennes et le CO_2. Cette observation peut être attribuée au fait que nos séries ne sont pas stationnaires. En effet, les particules fines et les HAP_s présentent des fluctuations locales très irrégulières avec des sauts de variabilité sur une échelle temporelle fine. Pour les concentrations des particules de taille moyenne et le CO_2, la variabilité temporelle exhibe des schémas de variation type à l'échelle diurne et hebdomadaire (*cf.* section 2.5). Intuitivement, la présence de caractéristiques types, notamment l'effet saisonnier augmenterait le niveau de prédictibilité, mais les résultats montrent des estimations plus ou moins acceptables sur les données brutes.

Notons que pour le formaldéhyde, en supprimant le pic associé à l'activation d'une source (ici bâtonnets d'encens), la mesure de prédictibilité augmente. Cela montre que cette variation est aléatoire, réduit donc la prédictibilité.

Contrairement à la prédictibilité sur les données non-transformées, la mesure prédictive sur les séries différenciées concorde avec nos suppositions : plus les particules sont fines, plus la prédictibilité de leur vitesse de fluctuation diminue. Malgré la différentiation, les paramètres de température intérieure et extérieure révèlent une prédictibilité élevée, la vitesse de fluctuation conserve son niveau de prédictibilité. En tout, la différenciation des grandeurs climatiques présente les meilleurs résultats de prédictibilité avec une estimation d'environ 18 % pour la température intérieure et 26 % pour la température extérieure.

En dehors de ces considérations, on note que la prédictibilité des paramètres dans la maison expérimentale est élevée, suivie par la prédictibilité des fluctuations issues du bureau individuel et puis du bureau paysager. Cette observation va dans le sens de ce qu'on pourrait attendre de la mesure prédictive Ω. En fait, chaque série temporelle est formée de plusieurs parties qu'on pourrait appeler déterministes et stochastiques. Si les composantes déterministes dominent la variabilité des séries, alors on dira que la série est "fortement" prévisible. Si, au contraire, la variabilité est dominée par les composantes aléatoires, alors on dira que la série est imprédictible.

Du point de vue des sources de fluctuations, il est clair que la sollicitation des différentes composantes du bâtiment diffère d'un environnement à un autre. En effet, dans la maison expérimentale, seules les sources liées à l'environnement conditionné par les paramètres climatiques et indépendantes de l'occupation interviennent dans la détermination des niveaux de fluctuations, donc dans la structure interne de la série temporelle. Cela va se traduire en une contribution importante de la composante déterministe en l'absence de facteurs aléatoires, tels que l'occupation. En définitive, la prédictibilité augmente dans ce type d'environnement.

Au contraire pour les autres environnements normalement occupés (le bureau individuel et l'espace paysager), l'interaction entre l'occupant et les différents systèmes du bâtiment augmente. La variabilité de la concentration est donc soumise à des divers paramètres qui dépendent de plusieurs facteurs aléatoires : les caractéristiques stochastiques se manifestent d'avantage. Finalement, la mesure de prédictibilité diminue dans ce type d'environnement.

TABLE 3.4.1 – Estimation de la prédictibilité $\hat{\Omega}_g(\%)$ des différentes séries temporelles de la QAI par différentes méthodes d'estimation de la densité spectrale de puissance. L'estimation a été effectuée sur les variables non-transformées. Les environnements considérés sont : le bureau individuel lors de la campagne 2011 (BI2011), la maison expérimentale (campagne MARIA) et l'espace paysager lors la campagne 2012 (OS2012). Pour les particules, on présente deux valeurs par ligne (ce n'est pas un intervalle).

Environnement	Variable x_t	$\widehat{\Omega_g}(\%)$ par méthode spectrale		
		WOSA	**Périodogramme**	**Multitaper**
BI2011	$P_{0.35}$-$P_{0.45}$	48.12-46.75	52.22-50.35	48.79-47.07
	$P_{1.8}$-$P_{2.5}$	35.98-38.21	41.08-44.38	36.30-38.80
	$P_{3.5}$-$P_{4.5}$	36.22-33.49	42.23-39.33	36.90-34.24
	$P_{6.25}$-$P_{8.75}$	29.19-22.9	34.73-28.10	30.00-23.71
	$P_{12.5}$-$P_{17.5}$	20.17-8.96	25.64-13.46	21.06-9.62
	CO_2	30.56	37.73	31.81
	HAP_s	38.79	42.89	39.87
OS2012	$P_{0.35}$-$P_{0.9}$	33.36-27.11	37.81-30.78	34.04-26.72
	$P_{1.8}$-$P_{4.5}$	23.98-20.71	29.17-29.45	23.75-21.74
	$P_{8.75}$	22.70	34.09	24.76
	CO_2	28.75	41.04	30.81
	T_{int}-HS_{int}	39.51-51.23	49.39-56.70	40.71-52.96
	T_{ext}-HS_{ext}	49.27-47.29	56.85-52.85	50.80-48.23
MARIA	HCHO	52.64	66.69	55.27
		60	73	62.7
	$P_{0.35}int$-$P_{0.45}int$	63.19-60.49	65.56-63.81	63.91-61.41
	$P_{0.575}int$-$P_{0.725}int$	54.24-49.81	59.43-54.41	54.90-50.52
	$P_{0.9}int$-$P_{1.3}int$	49.17-49.78	54.22-55.35	50.22-50.97
	$P_{0.1.8}int$-$P_{2.5}int$	46.71-54.73	52.51-59.21	47.76-54.99
	$P_{3.5}int$-$P_{4.5}int$	45.27-37.71	49.97-42.71	45.60-38.30
	$P_{6.25}int$-$P_{8.75}int$	33.21-23.46	38.05-28.20	33.79-24.06
	$P_{12.5}int$-$P_{17.5}int$	21.18-7.63	24.05-10.7	21.12-7.84
	$P_{0.35}ext$-$P_{0.45}ext$	57.20-59.43	62.31-63.65	58.67-60.89
	$P_{0.575}ext$-$P_{0.725}ext$	49.43-29.23	56.97-35.56	52.11-31.14
	$P_{0.9}ext$-$P_{1.3}ext$	23.88-24.86	26.62-28.57	24.65-25.86
	$P_{0.1.8}ext$-$P_{2.5}ext$	27.01-43.56	32.57-51.83	28.52-45.78
	$\Delta P_{0.35}$-$\Delta P_{0.45}$	49.21-55.18	56.44-60.87	51.34-57.14
	$\Delta P_{0.575}$-$\Delta P_{0.725}$	44.03-23.32	52.98-28.22	47.42-24.77
	$\Delta P_{0.9}$-$\Delta P_{1.3}$	19.83-20.13	20.61-21.06	19.97-20.25
	$\Delta P_{0.1.8}$-$\Delta P_{2.5}$	20.49-30.17	22.32-36.43	20.69-31.45

Note : nous avons estimé deux fois la série de HCHO. D'abord avec le pic associé à l'activation ponctuelle d'une source suivie par une ouverture des fenêtres (cf. Figure 2.5.8), ensuite en lissant cette variation (cellules grisées). Les variables ΔP_i correspondent à la différence entre les concentrations extérieures et intérieures des particules de la même taille i.

TABLE 3.4.2 – Estimation de la prédictibilité $\widehat{\Omega_g}(\%)$ des différentes séries temporelles de la QAI par différentes méthodes d'estimation de la densité spectrale de puissance. L'estimation a été effectuée sur des séries transformées par une différenciation de premier ordre : $x_t \hookrightarrow \Delta x_t$. Les environnements considérés sont : le bureau individuel lors de la campagne 2011 (BI2011), la maison expérimentale (campagne MARIA) et l'espace paysager lors de la campagne 2012 (OS2012).

Environnement	Variable Δx_t	$\hat{\Omega}_g(\%)$ par méthode spectrale **WOSA**	**Périodogramme**	**Multitaper**
BI2011	$P_{0.35}$-$P_{0.45}$	1.17-1.35	3.78-3.80	1.24-1.34
	$P_{1.8}$-$P_{2.5}$	3.19-2.74	6.22-5.67	3.41-2.94
	$P_{3.5}$-$P_{4.5}$	2.25-2.46	5.29-5.46	2.43-2.66
	$P_{6.25}$-$P_{8.75}$	2.36-2.88	5.38-5.88	2.58-3.09
	$P_{12.5}$-$P_{17.5}$	2.72-3.22	5.74-6.27	2.92-3.45
	CO_2	2.84	6.56	3.23
	HAP_s	1.57	3.92	1.64
OS2012	$P_{0.35}$-$P_{0.9}$	3.37-2.67	6.63-6.50	3.21-2.92
	$P_{1.8}$-$P_{4.5}$	2.67-3.66	7.04-8.19	2.85-3.90
	$P_{8.75}$	3.99	9.36	4.58
	CO_2	8.86	17.80	10.20
	T_{int}-HS_{int}	16.81-6.62	21.76-13.08	17.06-7.38
	T_{ext}-HS_{ext}	24.05-4.01	31.49-8.58	24.93-4.46
MARIA	HCHO	1.55	4.75	1.75
		2.28	5.73	2.48
	$P_{0.35}int$-$P_{0.45}int$	2.21-2.32	5.39-5.65	2.40-2.53
	$P_{0.575}int$-$P_{0.725}int$	3.03-3.43	6.37-6.83	3.23-3.69
	$P_{0.9}int$-$P_{1.3}int$	3.62-3.81	6.96-7.16	3.89-4.07
	$P_{0.1.8}int$-$P_{2.5}int$	3.98-3.70	7.45-7.16	4.24-3.91
	$P_{3.5}int$-$P_{4.5}int$	3.15-3.10	6.42-6.49	3.32-3.24
	$P_{6.25}int$-$P_{8.75}int$	2.91-3.38	6.10-6.64	3.02-3.57
	$P_{12.5}int$-$P_{17.5}int$	3.60-4.64	6.66-7.63	3.75-4.77
	$P_{0.35}ext$-$P_{0.45}ext$	4.05-3.46	4.96-4.68	3.87-3.26
	$P_{0.575}ext$-$P_{0.725}ext$	4.23-4.75	4.87-5.17	4.07-4.62
	$P_{0.9}ext$-$P_{1.3}ext$	4.92-4.34	5.35-5.21	4.80-4.21
	$P_{0.1.8}ext$-$P_{2.5}ext$	3.38-2.69	5.18-5.42	3.33-2.87
	$\Delta P_{0.35}$-$\Delta P_{0.45}$	3.80-3.28	4.88-4.62	3.62-3.08
	$\Delta P_{0.575}$-$\Delta P_{0.725}$	4.06-4.50	4.79-5.02	3.90-4.37
	$\Delta P_{0.9}$-$\Delta P_{1.3}$	4.61-4.11	5.15-5.14	4.49-3.99
	$\Delta P_{0.1.8}$-$\Delta P_{2.5}$	3.10-2.76	5.29-5.63	3.14-2.94

Note : nous avons estimé deux fois la série du HCHO. D'abord avec le pic associé à l'activation ponctuelle d'une source suivie par une ouverture des fenêtres (cf. Figure 2.5.8), ensuite en lissant cette variation (cellules grisées). Les variables ΔP_i correspondent à la différence entre les concentrations extérieures et intérieures des particules de la même taille i.

3.4.2.2 Commentaires sur Ω_g et normalisation par rapport à une série sinusoïdale

Nous avons remarqué que l'estimation proposée par Goerg (2013) est très sensible à la taille de l'échantillon et à la résolution temporelle utilisée.

Pour différentes tailles de la même série, on observe une augmentation exponentielle de la prédictibilité. En effet, nous avons simulé plusieurs séries temporelles de type $y_t = \sin(2\pi x)$ de tailles allant de 10^2 à 10^7 points. Les résultats de prédictibilité $\hat{\Omega}_g(y_t)$ obtenus pour chaque série montrent une estimation croissante de la prédictibilité avec l'augmentation de la taille de l'échantillon.

Pour illustrer cette relation, la Figure 3.4.3 montre les résultats obtenus de la prédictibilité sur les différentes séries temporelles de sinus. Nous présentons un ajustement linéaire (courbe **rouge**) de la prédictibilité en fonction de la taille des séries, le coefficient de détermination de cette régression est de 0.79. Pour donner une idée sur la relation non-linéaire entre la mesure Ω et la taille T, une autre régression polynomiale de degré 3 a été effectuée (courbe **bleue**). Cette constatation laisse présager que le logarithme de la mesure Ω_g varie linéairement avec le logarithme de la taille de la série. Nous pensons donc qu'il y a une relation de la forme :

$$\Omega_g \propto T^{\xi}. \tag{3.4.3}$$

Donc l'estimation de la mesure prédictive doit tenir compte d'un biais lié à la longueur des séries. Pour cela, nous reprenons la même mesure et nous normalisons par rapport à une série sinusoïdale comme suit :

$$\widetilde{\Omega}^*(X_t) = \frac{\hat{\Omega}_g(X_t)}{\hat{\Omega}_{\text{sin}}}, \tag{3.4.4}$$

où $\hat{\Omega}_g(X_t)$ est la mesure de Goerg de la série X_t de taille T, $\hat{\Omega}_{\text{sin}}$ est l'estimation de la mesure de Goerg pour une série sinusoïdale de la même taille que X_t, enfin $\widetilde{\Omega}^*(X_t)$ est une mesure normalisée qui permet de comparer entre elles plusieurs séries de différentes tailles.

Le Tableau 3.4.3 donne l'estimation de la mesure de prédictibilité sur les séries de HCHO dans deux campagnes de mesures, en 2013 et en 2015. Nous donnons aussi l'estimation normalisée de la prédictibilité des mêmes séries par rapport au sinus correspondant, *i.e,* la prédictibilité d'une série sinusoïdale de même taille que la série de HCHO.

Nous remarquons que $\hat{\Omega}_g(X_t)$ donne presque les mêmes valeurs pour la série de 2013 que pour la série de 2015. Or, ce que nous avons pu observer lors de la description des données, la variabilité du HCHO en 2015 était beaucoup plus régulière que la variabilité en 2013. Les fluctuations de 2013 présentaient beaucoup de changements abrupts qui altèrent les variations régulières ; donc la présence des facteurs aléatoires était plus prononcée. Une mesure de prédictibilité doit détecter ces aspects liés à la manifestation des composantes déterministes et aléatoires. Même avec l'extraction des composantes déterministes, qui sont la saisonnalité et la tendance, les valeurs de $\hat{\Omega}_g(X_t)$ sont très proches, voire même une prédictibilité plus élevée pour la série de 2013 que celle de 2015.

En revanche, lorsqu'on normalise par rapport à la prédictibilité harmonique, comme dans la formule 3.4.4, la nouvelle mesure $\widetilde{\Omega}^*(X_t)$ corrobore le sens d'une mesure de prédictibilité (cellules grisées dans

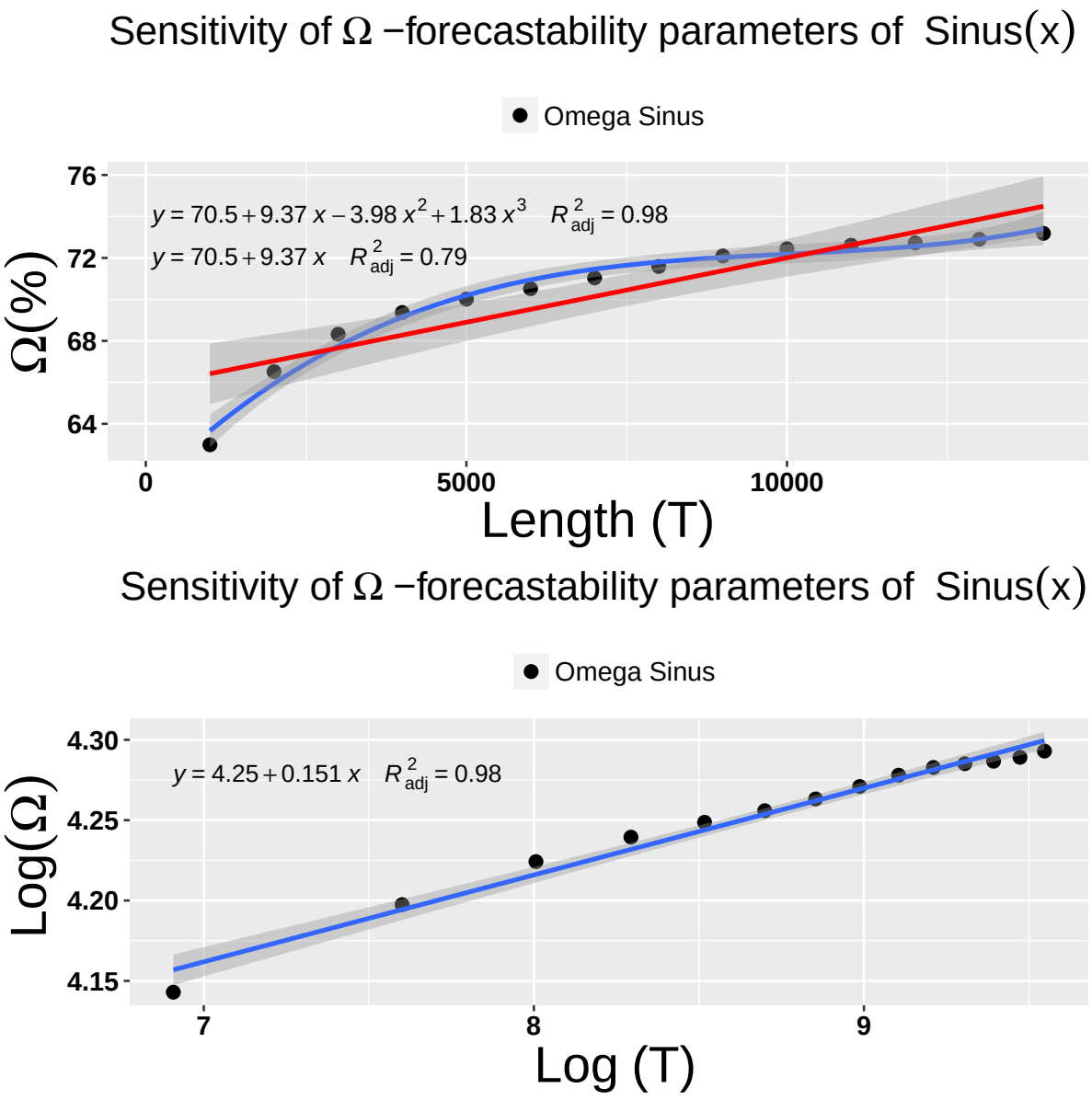


FIGURE 3.4.3 – *Sensibilité de la prédictibilité Ω_g des séries sinusoïdales par rapport à la taille.*

TABLE 3.4.3 – Estimation de la prédictibilité $\widehat{\Omega_g}(\%)$ des différentes séries de HCHO dans l'espace paysager durant les campagnes de 2013 et de 2015. Sont données aussi l'estimation de la première différentiation ΔHCHO et des différentes composantes issues de la décomposition **STL** : la saisonnalité, la tendance et le bruit. L'estimation de la prédictibilité normalisée est donnée par $\widetilde{\Omega}^*(X_t)$ et elle est rapportée à l'estimation de la mesure de prédictibilité de la série sinus de même taille, $\hat{\Omega}_{\text{sin}}$. T_1 est la taille de la série HCHO durant la campagne de 2013 (OS13) et le T_2 désigne la taille de la série du HCHO durant la campagne de 2015 (OS15).

	$\hat{\Omega}_g(X_t)$	Variable	**WOSA**	**Periodogramme**	**Multitaper**
OS2013		HCHO	58.11	63.81	58.66
		ΔHCHO	1.812	5.276	2.061
	STL	*Season*	54.28	70.4	58.4
		Trend	62.71	68.72	63.42
		Remainder	17.24	20.33	17.67
OS2015		HCHO	55.65	66.91	58.90
		ΔHCHO	2.63	6.975	2.93
	STL	*Season*	39.64	57.25	42.74
		Trend	67.23	77.58	70.4
		Remainder	17.88	21.78	16.8
	$\widetilde{\Omega}^*(X_t)$				
$\hat{\Omega}_{\text{sin}}$	$T_1 = 20000$		70.11	90.67	74.1
	$T_2 = 3251$		63.418	72.95	68.61
HCHO	OS13		82.89	70.37	79.17
	OS15		87.75	91.72	85.85

le Tableau 3.4.3). En effet, toutes les méthodes d'estimation (WOSA, Périodogramme et Multitaper) donnent une prédictibilité plus grande pour la série de 2015 qu'en 2013.

Cette observation traduit la nécessité d'utiliser les composantes déterministes, déjà présentes et qu'on peut extraire facilement à des fins de prévision. Tandis que pour la série de 2013, l'extraction de ces composantes semble plus délicate et, nécessite donc un traitement approprié pour explorer ses composantes latentes.

En ce qui concerne les méthodes spectrales d'estimation de la prédictibilité, nous remarquons que le périodogramme donne des résultats biaisés et très sensibles à la taille de l'échantillon. Cette constatation va dans le sens des remarques soulevées par Goerg (2013). En effet, on observe, lors de l'estimation de $\hat{\Omega}_g$ des sinus en fonction de la taille, une relation non-linéaire entre T et Ω_g après passage au logarithme. Donc, la relation $\Omega_g \propto T^{\xi}$ ne semble plus être vérifiée, nous ne la recommandons pas ici pour la mesure de prédictibilité.

3.5 Structure de dépendance : dimension fractale et l'exposant de Hurst

De nombreuses séries temporelles que nous avons pu observer précédemment présentent des périodogrammes avec un pic au pôle de la fréquence zéro. De manière équivalente, dans le domaine temporel,

la fonction d'autocorrélation diminue très lentement avec un taux hyperbolique. Cette caractéristique peut être étayée soit par la non-stationnarité des séries (présence des composantes déterministes) où par le phénomène de dépendance à long terme. Dans ce deuxième cas, des observations très espacées dans le temps afficheront tout de même une certaine dépendance : leurs autocorrélations décroissent à une vitesse hyperbolique.

Par ailleurs, la dimension fractale et l'exposant de Hurst offrent un cadre (parmi d'autres) qui permet de mettre en évidence ces aspects d'auto-dépendances des chroniques. Ces concepts, fournissent un moyen pour quantifier le degré d'irrégularité (ou rugosité) d'une série et d'extraire les paramètres indépendants de l'échelle.

3.5.1 Un peu de littérature

L'idée de représenter le monde réel et les phénomènes qui le constituent par des lois d'échelle n'est pas récente. Déjà, la première moitié du XX^e siècle fût riche tant au niveau conceptualisation que modélisation, partant de la thèse de Bachelier (1900), en passant par les travaux de Kolmogorov (1941), jusqu'aux travaux plus récents de Hurst (1951) et de Mandelbrot (1963; 1983). Tous venus des horizons et des champs disciplinaires très variés, ils reconnaissent l'existence d'une loi d'échelle sur les phénomènes étudiés, ce qui fait d'ailleurs dire à Barnsley (1988) que l'on peut parler de "fractals partout".

Les premiers travaux sur l'analyse des séries caractérisées par de fortes dépendances à long terme (Long-Range Dependence ou LRD) ont été développés dans le but de mettre en évidence la loi empirique de Hurst (1951) sur les séries hydrologiques. Les processus linéaires à mémoire courte de type ARMA ne pouvaient pas expliquer les observations sur les niveaux des crues du Nil. Ainsi, Hurst rejette dès lors ce type de modèle en montrant qu'il sous-estime la complexité des fluctuations hydrologiques. Suite à ces considérations, les travaux de Mandelbrot (1965; 1968; 1972; 1983) ont approfondi la notion de LRD par l'introduction des objets fractals. Largement inspirées par ces derniers, plusieurs méthodes d'estimation de la dimension fractale ont été proposées dans différents domaines de la science et de l'ingénierie. Depuis le début des années 90, de nombreux états de l'art ont été déjà publiés, regroupant les principales méthodes utilisées en vue d'estimer les caractéristiques d'irrégularité d'un objet, que ce soit mathématique ou physique. On peut consulter les ouvrages et les monographies suivantes : Klinkenberg & Goodchild (1992); Schepers et al. (1992); Theiler (1990); Cutler (1993) et Schmittbuhl et al. (1995). Récemment, on trouve plusieurs applications pour les séries de polluants ; par exemple Jayawardena (2014) (Chapitre 10) et Seuront (2009) présentent l'essentiel des méthodes utilisées pour la détection de structures de variabilité en mettant l'accent sur leurs applications aux systèmes environnementaux.

Varotsos & Kirk-Davidoff (2006) utilisent la DFA (Detrended Fluctuation Analysis) pour quantifier le niveau de persistance des concentrations totales d'ozone (TOC) et de températures de brillance (TRT). Les fluctuations de TOC présentent des dépendances à long terme et ce lien montre l'existence d'une dynamique conjointe entre la variabilité à petites et grandes échelles temporelles. Cette méthode est aussi utilisée sur les résidus après suppression des oscillations saisonnières des données TOC et montre que les fluctuations exhibent aussi une dépendance à long terme (Kiss et al., 2007). Pour montrer l'intermittence des fluctuations de certains polluants, Anh et al. (2000) utilisent un modèle basé sur une cascade multifractale sur les données de NO et SO_2. Néanmoins, ce modèle n'a pas été utilisé pour la prévision d'épisodes de pollution.

3.5.2 Définition de la mémoire longue et sa caractérisation

On reprend d'abord deux définitions correspondant aux comportements à mémoire longue des séries temporelles, l'une est basée sur la fonction d'autocorrélation (temporelle), et l'autre sur le comportement du spectre de puissance (fréquentielle).

Définition 3.5.1. *Mémoire longue I*
Un processus stationnaire X_t est un processus à mémoire longue s'il existe un nombre réel $\alpha \in]0,1[$, et une constante positive c, vérifiant :

$$\lim_{k\to\infty} \frac{\rho(k)}{c \cdot k^{-\alpha}} = 1 \tag{3.5.1}$$

Donc asymptotiquement, pour $k \to \infty$, $\rho(k) \propto c.k^{-\alpha}$.

Définition 3.5.2. *Mémoire longue II*
Un processus stationnaire X_t est un processus à mémoire longue s'il existe un nombre réel $\alpha \in]0,1[$, et une constante positive c_λ, tel que :

$$\lim_{\lambda\to 0} \frac{S_x(\lambda)}{|\lambda|^{-\alpha}} = 1. \tag{3.5.2}$$

Ainsi, on vérifie que $S_x(\lambda) \propto c_\lambda |\lambda|^{-\alpha}$ pour $\lambda \to 0$.

3.5.3 Classification des séries temporelles en fonction de la structure de dépendance

3.5.3.1 Relation entre le paramètre d du modèle $ARFIMA(0,d,0)$, l'exposant H de Hurst et la dimension fractale

Rappelons rapidement le cadre de la modélisation des modèles $ARFIMA$ et leurs propriétés statistiques (largement développées par Hosking (1981)). Soit $X_t \sim ARFIMA(0,d,0)$ ($i.e$ $(1-L)^d X_t = \varepsilon_t$ avec le paramètre de différentiation d fractionnaire) ; si $d \in \left]-\frac{1}{2},\frac{1}{2}\right[$, alors :

- Le processus X_t est stationnaire et inversible,
- La fonction d'autocovariance s'écrit :

$$\gamma_k = \frac{\Gamma(1-2d)\,\Gamma(k+d)}{\Gamma(d)\,\Gamma(1-d)\,\Gamma(k+1-d)}, \text{ avec } \gamma_k \sim \frac{\Gamma(1-2d)}{\Gamma(d)\,\Gamma(1-d)} k^{2d-1} \text{ pour } k \longrightarrow \infty. \tag{3.5.3}$$

où Γ est la fonction Eulérienne de seconde espèce :

$$\Gamma(x) = \begin{cases} \int_0^\infty t^{x-1}\exp^{-t} dt & \text{si } x > 0 \\ \infty & \text{si } x = 0 \\ x^{-1}\Gamma(x+1) & \text{si } x < 0 \end{cases} \tag{3.5.4}$$

- La densité spectrale est donnée par :

$$S(\lambda,d) = \left(2\sin\frac{\lambda}{2}\right)^{-2d} \text{ avec } 0 < \lambda \leq \pi \text{ et } \lim_{\lambda\longrightarrow 0} S(\lambda,d) = \lambda^{-2d}. \tag{3.5.5}$$

Avec ces propriétés, les processus ARFIMA donnent lieu à des autocorrélations qui diminuent à un taux hyperbolique, contrairement aux processus ARMA dont les autocorrélations décroissent avec un taux géométrique. En particulier, si l'on s'intéresse à la forme de la densité spectrale (relation 3.5.5), elle n'est pas limitée à une valeur finie au voisinage de la fréquence zéro. Cette caractéristique est similaire à celle observée dans les processus aléatoires qui font intervenir, non pas l'ordre d'intégration fractionnaire, mais l'exposant de Hurst. En effet, les travaux menés par Mandelbrot (1965; 1968) ont donné lieu à la première formulation des processus aléatoires fractionnaires de type "Browniens fractionnaires". Ce modèle ayant un paramètre $\mathbf{H}$ (exposant de Hurst) est une généralisation des processus Browniens standards (annexe E.1). Un processus Brownien fractionnaire d'exposant $\mathbf{H}$, noté $B_{\mathbf{H}}(t,\cdot)$ et est défini par :

$$B_{\mathbf{H}}(t,\cdot) = \frac{1}{\Gamma\left(\mathbf{H}+\frac{1}{2}\right)}\left\{\int_0^t (t-s)^{\mathbf{H}-\frac{1}{2}}\, dB(s,\cdot)\right\} \tag{3.5.6}$$

où $0 < \mathbf{H} < 1$ est l'exposant de Hurst (statistique R/S ci-après) et $B(s,\cdot)$ est le mouvement Brownien ordinaire de variance unitaire (voir l'annexe E.1, la formule E.1.1). La fonction d'autocovariance des incréments $\Delta B_{\mathbf{H}}(t) = B_{\mathbf{H}}(t,\cdot) - B_{\mathbf{H}}(t-1,\cdot)$ (bruit fractionnaire en temps discret) est donnée par :

$$\gamma_k = \frac{1}{2}\left[|k+1|^{2\mathbf{H}} - 2|k|^{2\mathbf{H}} + |k-1|^{2\mathbf{H}}\right] \tag{3.5.7}$$

et pour $k \to \infty$, on a la relation suivante :

$$\gamma_k \sim \mathbf{H}(2\mathbf{H}-1)\, k^{2\mathbf{H}-2}. \tag{3.5.8}$$

Le comportement asymptotique dans les relations 3.5.3 et 3.5.8 possède le même ordre de déclin hyperbolique : $2d-1 \backsimeq 2\mathbf{H}-2$. La relation entre l'exposant de Hurst et l'ordre d'intégration fractionnaire peut s'écrire :

$$\mathbf{H} = d + \frac{1}{2}. \tag{3.5.9}$$

Cette relation montre qu'une mauvaise estimation de l'un des paramètres ($\boldsymbol{H}$ où d) va se répercuter sur l'estimation de l'autre.

La densité spectrale $S(\lambda)$ définie dans 3.5.5 peut s'écrire pour $\lambda \longrightarrow 0$ comme

$$\begin{aligned} \log\left(S_X(\lambda)\right) &\sim c - 2d\log(\lambda) \\ &\sim c - (5-2D)\log(\lambda) \\ &\sim c - (2\mathbf{H}+1)\log(\lambda) \end{aligned} \tag{3.5.10}$$

où D est la dimension fractale du processus X_t. On peut déduire la relation entre $\mathbf{H}$ et D de la formule 3.5.10 :

$$\mathbf{H} = 2 - D \tag{3.5.11}$$

Pour une démonstration complète de la relation 3.5.11, on peut se référer à l'article de Orey (1970) et à l'ouvrage de Falconer (2004).

3.5.3.2 Classification des processus

Reprenant la forme de comportement du spectre au niveau des basses fréquences et en fonction de la valeur **H**, on peut rencontrer (voir l'exposé plus détaillé dans (Abraham-Frois et al., 1998)) les cas suivants :

- $\frac{1}{2} < \mathbf{H} < 1$: le processus est persistant et les autocorrélations sont positives et décroissent avec un taux hyperbolique vers 0 lorsque le retard (lag) est élevé. La densité spectrale est concentrée autour des basses fréquences (variabilité lente), elle tend vers l'infini lorsque la fréquence tend vers 0.
- $\mathbf{H} = \frac{1}{2}$: le processus est sans mémoire (ou mémoire très courte équivalant à un processus ARMA),
- $0 < \mathbf{H} < \frac{1}{2}$: le processus est anti-persistant, les autocorrélations alternent de signe et la densité spectrale est dominée par des composantes de hautes fréquences.

3.5.4 Estimation de la dimension fractale pour les séries temporelles

Comme évoqué dans la sous-section précédente, on peut d'une certaine manière, analyser les structures de dépendance des séries temporelles avec la relation établie entre la dimension fractale, l'exposant fractionnaire et l'exposant de Hurst. Cette sous-section est dédiée à l'estimation de la dimension fractale dans le cadre des séries temporelles.

3.5.4.1 Préliminaires mathématiques

La dimension fractale pour les séries temporelles (dimension [4] de Hausdorff, suivant Falconer (2004)) mesure le niveau de rugosité ou d'irrégularité de la trajectoire. Une série suffisamment lisse et différentiable dans l'espace $\mathbb{R}^d$ a donc une dimension topologique et fractale égale à d. En revanche, pour une série réelle (non-différentiable), la dimension fractale prend des valeurs entre la dimension topologique d et $d+1$.

Nous avons vu dans la section précédente, la relation entre l'exposant de Hurst et la dimension fractale. Par le biais de cette dernière, on pourrait classifier les séries par type de persistance et mémoire.

Nous suivons la pratique courante pour la définition de la dimension fractale d'un ensemble de points $X \subset \mathbb{R}^d$ pour qu'elle soit la dimension classique de Hausdorff (Falconer, 2004). Pour $\epsilon > 0$, un $\epsilon-$recouvrement de X est une collection $\{B_i, i = 1, 2, \ldots\}$ finie ou dénombrable de boules $B_i \subset \mathbb{R}^d$ ayant un diamètre $|B_i| \leq \epsilon$ qui couvre X. On peut se référer au livre de Tricot (1999) (Chapitre 2) pour plus de détails sur le recouvrement.

4. La définition précise, mais complexe : Valeur de D pour laquelle le volume de dimension D change de l'infini à zéro.

Définition 3.5.3. *Dimension de* HAUSDORFF

Soient $\delta, \epsilon > 0$ et la mesure de Hausdorff $\delta-$dimensionnelle de X définie par :

$$H^{\delta}(X) = \lim_{\epsilon \to 0} \inf \left\{ \sum_{i=1}^{\infty} |B_i|^{\delta}, \text{ avec } \{B_i : i = 1, 2, \ldots\} \text{ est un } \epsilon - \text{recouvrement de } X \right\};$$

il existe une seule valeur non-négative D telle que :

$$\begin{cases} H^{\delta}(X) = \infty & \text{si } \delta < D \\ H^{\delta}(X) = 0 & \text{si } \delta > D \end{cases}$$

Cette valeur est la dimension de HAUSDORFF de l'ensemble X.

Sous certaines conditions "faibles" de régularité, la dimension de Hausdorff coïncide avec la dimension de comptage des boîtes,

$$D_{BC} = \lim_{\epsilon \to 0} \frac{\log(N(\epsilon))}{\log(1/\epsilon)}, \tag{3.5.12}$$

avec $N(\epsilon)$ le plus petit nombre de cubes de largueur ϵ de $\mathbb{R}^d$ pour recouvrir X.

3.5.4.2 Principe général de construction

Pratiquement toutes les méthodes d'estimation suivent un schéma global assez similaire, mais sur le fond les méthodologies employées sont bien différentes (Gneiting et al., 2012). Un estimateur de la dimension fractale peut être construit comme suit :

1. Une certaine propriété numérique, soit $\mathcal{Q}$, d'une série temporelle X_t est calculée en fonction d'un paramètre d'échelle, noté généralement ϵ ;
2. Une loi de puissance asymptotique $\mathcal{Q}(\epsilon) \propto \epsilon^{\beta}$ (pour $\epsilon \to 0$) devient petite est calculée tel que :
   (a) l'exposant d'échelle β est une fonction linéaire avec la dimension fractale D ;
   (b) D est estimé par une régression de $\log \mathcal{Q}(\epsilon)$ sur $\log(\epsilon)$, pour les plus petites valeurs de ϵ.

Nous présentons, ci-après, quelques méthodes d'estimation de la dimension fractale en se basant sur le schéma global de construction.

3.5.4.3 Méthodes d'estimation

Nous présentons trois méthodes usuelles d'estimation de la dimension fractale pour une série temporelle : comptage des boîtes, l'estimateur de HALL-WOOD et l'estimation variationnelle. Néanmoins, il existe dans la littérature plusieurs procédures et approches heuristiques pour déterminer la dimension fractale.

Comptage des boîtes Sur la base de la formule 3.5.12, on peut construire une loi d'échelle par plusieurs recouvrements de la série temporelle. En effet, cette méthode consiste à appliquer successivement sur la trajectoire un quadrillage de plus en plus fin et compter à chaque itération le nombre de boîtes pleines, *i.e.* le nombre de boîtes contenant au moins une partie de la série. D'abord, la trajectoire est recouverte par une seule boîte, puis en quatre quadrants et on compte le nombre de cellules contenant la courbe. Ensuite, chaque quadrant est divisé en quatre sous-quadrants jusqu'à ce que la finesse du maillage ne permet plus de révéler des structures plus fines : la largeur de la boîte est égale à la résolution temporelle de la série. Enfin, on trace le graphe des points de coordonnées $(N(\epsilon), \epsilon)$ sur une échelle bi-logarithmique. La pente de la droite de régression linéaire correspond ainsi à la dimension fractale de la série.

Pour l'estimation de D_{BC}, une méthode algorithmique peut être utilisée. Pour une présentation simple, on pose $T = 2^K$ et le déroulement de l'algorithme peut être résumé comme suit : soient l'écart de la série, $R = \max_{0\leq j\leq T} X_{j/T} - \min_{0\leq j\leq T} X_{j/T}$ et l'échelle $\epsilon_k = 2^{k-K}$, $k = 0, 1, 2, \ldots K$. La plus grande échelle est donc $\epsilon_K = 1$: une seule boîte couvrant toute la trajectoire de hauteur R (la boîte englobante). À l'échelle ϵ_k, la boîte englobante est maillée par 4^{K-k} boîtes de largeur 2^{k-K} et de hauteur de $R \times 2^{k-K}$. $N(\epsilon_k)$ désigne le nombre de boîtes qui coïncident avec la trajectoire.

L'estimation (naïve) par le comptage des boîtes est obtenue en utilisant l'ajustement de la pente de la régression $\log N(\epsilon)$ sur $\log(\epsilon)$ par les Moindres Carrés Ordinaire (MCO) :

$$\hat{D}_{BC} = -\frac{\sum_{k=0}^{K} (w_k - \bar{w}) \log N(\epsilon_k)}{\sum_{k=0}^{K} (w_k - \bar{w})^2}, \tag{3.5.13}$$

avec $w_k = \log(\epsilon_k)$ et $\bar{w}$ est la moyenne de $w_0, w_1, \ldots, w_K$. Pour une estimation rapide, nous adoptons les recommandations de Liebovitch & Toth (1989), qui consistent à exclure toutes les échelles fines ϵ_k, tel que $N(\epsilon_k) > T/5$.

L'estimateur de Hall-Wood Hormis le fait que la méthode de comptage des boîtes doit écarter le comptage de la plus grande et de la petite échelle ($N(\epsilon_0) \geq T$, $N(\epsilon_K) = 1$), beaucoup de chercheurs ont souligné d'autres limitations relatives à l'estimation de D_{BC}. Leurs critiques concernent en particulier le caractère naïf d'introduire toutes les échelles dans la régression $\log(N(\epsilon))$ sur $\log(\epsilon)$. Plusieurs travaux ont cherché alors à modifier l'estimation par comptage des boîtes (Dubuc et al., 1989; Taylor & Taylor, 1991).

Face à cette faiblesse de $\hat{D}_{BC}$, Hall & Wood (1993) ont introduit une version modifiée de l'estimation par comptage des boîtes, les auteurs se basant sur la plus petite échelle observée au sein de la série. Pour soutenir leur proposition, soit $A(\epsilon)$ l'aire totale des boîtes à l'échelle ϵ qui coupe la trajectoire. Il y a $N(\epsilon)$ boîtes, on a alors la relation $A(\epsilon) \propto N(\epsilon)\epsilon^2$ qui permet de redéfinir la formule 3.5.12, comme suit :

$$D_{BC} = 2 - \lim_{\epsilon\to 0} \frac{\log(A(\epsilon))}{\log(\epsilon)}. \tag{3.5.14}$$

À l'échelle $\epsilon_\ell = \ell/T$, avec $\ell = 1, 2, \ldots$, l'estimation de $A(\ell/T)$ peut être approximée par :

$$\hat{A}\left(\ell/T\right)=\frac{\ell}{T}\sum_{i=1}^{\lfloor T/\ell\rfloor}\left|X_{i\ell/T}-X_{(i-1)\ell/T}\right|, \tag{3.5.15}$$

où $\lfloor n/T\rfloor$ est la partie entière de ℓ/T . L'estimateur de Hall-Wood (1993) est basé sur la régression par MCO du $\log\left(\hat{A}\left(\ell/T\right)\right)$ sur $\log\left(\ell/T\right)$:

$$\hat{D}_{HW}=2-\frac{\sum_{\ell=1}^{L}\left(w_\ell-\bar{w}\right)\log\left(\hat{A}\left(\ell/T\right)\right)}{\sum_{\ell=1}^{L}\left(w_\ell-\bar{w}\right)^2}, \tag{3.5.16}$$

où $w_\ell=\log\left(\ell/T\right)$, $L\geq 2$ et $\bar{w}$ est la moyenne de $w_1,w_2,\ldots,w_L$. Pour minimiser le biais de l'estimateur, Hall-Wood (1993) ont recommandé l'utilisation de $L=2$, $\hat{D}_{HW}$ devient

$$\hat{D}_{HW}=2-\frac{\log\left(\hat{A}\left(2/T\right)\right)-\log\left(\hat{A}\left(1/T\right)\right)}{\log 2}. \tag{3.5.17}$$

C'est cet estimateur qui a été implémenté dans (Gneiting et al., 2012) et utilisé dans nos applications.

Estimateur variationnel Cette méthode consiste à travailler sur les incrémentations stationnaires X_u-X_{u+t} du variogramme d'ordre p :

$$\gamma_p\left(t\right)=\frac{1}{2}\mathbb{E}\left|X_u-X_{u-t}\right|^p. \tag{3.5.18}$$

L'estimation classique par la méthode des moments d'ordre p de 3.5.18 en lag $t=\ell/T$ de la série X_t est :

$$\hat{V}_p\left(\ell/T\right)=\frac{1}{2\left(T-\ell\right)}\sum_{i=\ell}^{T}\left|X_{i/T}-X_{(i-\ell)/T}\right|^p. \tag{3.5.19}$$

L'estimateur de variation de l'ordre p pour la dimension fractale est défini par :

$$\hat{D}_{V;p}=2-\frac{1}{p}\frac{\sum_{\ell=1}^{L}\left(w_\ell-\bar{w}\right)\log\left(\hat{V}_p\left(\ell/T\right)\right)}{\sum_{\ell=1}^{L}\left(w_\ell-\bar{w}\right)^2}, \tag{3.5.20}$$

où $w_\ell=\log\left(\ell/T\right)$, $L\geq 2$ et $\bar{w}$ est la moyenne de $w_1,w_2,\ldots,w_L$.

Une question posée et discutée intensivement dans la littérature concernant cette méthode, est elle du choix de $p>0$. Dans cette thèse, on retient les recommandations de Gneiting et al. (2012), soit $p=1$ et $L=2$ et on renvoie le lecteur aux références citées par cette étude. Pour ce choix, il s'agit finalement d'estimer un "madogramme". Finalement l'estimation pour $L=2$ se fait comme suit :

$$\hat{D}_{V;p} = 2 - \frac{1}{p}\frac{\log\left(\hat{V}_p\left(2/T\right)\right) - \log\left(\hat{V}_p\left(1/T\right)\right)}{\log 2}. \tag{3.5.21}$$

On rappelle par ailleurs que l'estimation par variogramme est initialement développée dans le domaine de la statistique spatiale (Bruno & Raspa, 1989), ensuite dans la littérature de la statistique mathématique (Constantine & Hall, 1994; Chan & Wood, 2004).

3.5.5 L'exposant de Hurst

Plusieurs procédures d'estimation de l'exposant de Hurst ont été proposées dans la littérature, nous nous référons à l'exposé de Taqqu et al. (1995) et nous renvoyons à ces auteurs pour plus de détails. Néanmoins, ces méthodes sont généralement décrites dans le cadre des séries stationnaires.

3.5.5.1 Méthode basée sur la statistique R/S

L'analyse R/S, présentée par Hurst (1951), sur les séries hydrologiques, est la première méthode mise en place pour estimer le paramètre $\mathbf{H}$. Cette méthode a été développée dans divers travaux par Mandelbrot (1963; 1968; 1972; 1983; 1975; 1965). La statistique se définit comme l'étendue des sommes partielles des écarts d'une série temporelle à sa moyenne divisée par son écart-type (Mignon, 1998).

Formellement, la statistique $\mathbf{Q}_T$ consiste à calculer le rapport entre l'étendue R_T et l'écart-type $S_T = \left[\frac{1}{T}\sum_{i=1}^{T}\left(X_i - \bar{X}_T\right)^2\right]^{\frac{1}{2}}$ de la série X_t, $t = 1, \ldots, T$, *i.e.* $\left(\mathbf{Q}_T = \frac{R_T}{S_T}\right)$. L'étendue R_T est définie par :

$$R_T = \max_{i\leq k\leq T}\sum_{i=1}^{k}\left(X_i - \bar{X}_T\right) - \min_{i\leq k\leq T}\sum_{i=1}^{k}\left(X_i - \bar{X}_T\right). \tag{3.5.22}$$

Donc la statistique $\mathbf{Q}_T$ est donnée par le rapport suivant :

$$\mathbf{Q}_T = \frac{\max_{i\leq k\leq T}\sum_{i=1}^{k}\left(X_i - \bar{X}_T\right) - \min_{i\leq k\leq T}\sum_{i=1}^{k}\left(X_i - \bar{X}_T\right)}{\left[\frac{1}{T}\sum_{i=1}^{T}\left(X_i - \bar{X}_T\right)^2\right]^{\frac{1}{2}}} \tag{3.5.23}$$

Le premier terme de R_T est le maximum sur k des sommes partielles des k écarts des X_i par rapport à sa moyenne et le second est le minimum sur k de cette même séquence de sommes partielles. La statistique $\mathbf{Q}_T$ est toujours non-négative.

L'analyse des étendues normalisées peut détecter la présence de mémoire longue même dans une série temporelle fortement non gaussienne et peut, en outre ? déceler des cycles non périodiques (Mandelbrot & Wallis, 1969).

Le comportement asymptotique de la statistique est donné par la relation suivante :

$$\mathbf{Q}_T = \frac{R_T}{S_T} \sim cT^{\mathbf{H}}, \text{ pour } T \to \infty, \tag{3.5.24}$$

où la constante $\mathbf{H}$ est l'exposant de Hurst. Elle peut être approximée par une simple régression linéaire entre le $\log(\mathbf{Q}_T)$ en fonction $\log(T)$:

$$\log(\mathbf{Q}_T) \sim \log(c) + \mathbf{H}\log(T). \tag{3.5.25}$$

$$\mathbf{H} \sim \frac{\log(\mathbf{Q}_T)}{\log(T)} \tag{3.5.26}$$

Selon Lo (1991), la statistique R/S souffre de plusieurs inconvénients : elle manque de robustesse en présence de mémoire à court terme et sa distribution statistique est inconnue. C'est-à-dire, d'une part elle n'arrive pas à discerner, en "petit" échantillon, les structures de variabilité à court terme et les propriétés d'une mémoire à long terme ; d'autres part, elle ne constitue pas un test statistique. Afin d'apporter une solution à ces problèmes, Lo (1991) propose une statistique, appelée *R/S modifiée* qui est robuste par rapport à la dépendance à court terme et a dérivé sa distribution limite, voir (Mignon, 1998) plus de discussion. La *R/S modifiée,* notée $\tilde{\mathbf{Q}}_T$ s'écrit :

$$\tilde{\mathbf{Q}}_T = \frac{R_T}{\hat{\sigma}_T(q)} = \frac{1}{\hat{\sigma}_T(q)}\left[\max_{i\leq k\leq T}\sum_{i=1}^{k}\left(X_i - \bar{X}_T\right) - \min_{i\leq k\leq T}\sum_{i=1}^{k}\left(X_i - \bar{X}_T\right)\right], \tag{3.5.27}$$

où :

$$\hat{\sigma}_T(q) = \left\{\frac{1}{T}\sum_{i=1}^{T}\left(X_i - \bar{X}_T\right)^2 + \frac{2}{T}\sum_{i=1}^{q}\omega_i(q)\left[\sum_{j=i+1}^{T}\left(X_i - \bar{X}_T\right)\left(X_{j-i} - \bar{X}_T\right)\right]\right\}^{\frac{1}{2}} \tag{3.5.28}$$

et

$$\omega_i(q) = 1 - \frac{i}{q+1} \quad q < T. \tag{3.5.29}$$

On note que la statistique de Lo diffère de R/S de Hurst uniquement par son dénominateur. En effet, en présence d'autocorrélation, l'écart-type estimé ne représente plus seulement la somme des variances des termes individuels, mais inclut également les autocovariances pondérées en fonction des décalages q, les poids $\omega_i(q)$ ayant été suggérés par Newey & West (1987).

En utilisant des simulations Monte-Carlo, Andrews (1991) montre que lorsque q est trop grand par rapport à la taille T de la série, alors la distribution de l'estimateur peut être très différente de sa distribution asymptotique. A l'inverse, lorsque q est trop petit, alors l'estimateur est biaisé à cause des autocorrélations qui pourraient être très influentes. Pour optimiser le choix de q, Andrews (1991) a proposé la règle suivante :

$$q = \lfloor k_T \rfloor \text{ où } k_T = \left(\frac{2\hat{\rho}}{1-\hat{\rho}^2}\right)^{\frac{2}{3}}\left(\frac{3T}{2}\right)^{\frac{1}{3}}, \tag{3.5.30}$$

où $\lfloor k_T \rfloor$ est la partie entière de k_T et $\hat{\rho}$ désigne l'estimateur du coefficient d'autocorrélation d'ordre 1. Les poids $\omega_i(q)$ deviennent alors $\omega_i = 1 - \left|\frac{i}{k_T}\right|$.

Avec ces raffinements, il est possible de dresser la distribution limite, ainsi R/S-*modifiée* est donnée par la formule :

$$V = \frac{\tilde{\mathbf{Q}}_T}{\sqrt{T}}, \tag{3.5.31}$$

et converge vers un *pont Brownien* ; on peut se référer à Samorodnitsky (2007) pour une revue de la littérature et pour plus de détails mathématiques.

3.5.5.2 Méthode basée sur la variance agrégée

La procédure de cette méthode consiste à diviser la série temporelle de taille T en $[T/m]$ séquences de taille m (en prenant la partie entière). Sur chacun des m blocs on calcule la quantité suivante :

$$X_k^{(m)} = \frac{1}{m} \sum_{t=1+m(k-1)}^{km} X(t), \ k = 1, 2, \ldots, T/m. \tag{3.5.32}$$

Ensuite, la variance empirique "intra-bloc" des $X_k^{(m)}$ est estimée par :

$$\mathbb{V}\left(X_k^{(m)}\right) = \frac{1}{[T/m]} \sum_{k=1}^{[T/m]} \left[X^{(m)}(k)\right]^2 - \left[\frac{1}{[T/m]} \sum_{k=1}^{[T/m]} \left[X^{(m)}(k)\right]\right]^2. \tag{3.5.33}$$

Cette variance se comporte comme une fonction puissance (Mignon, 1998) de type :

$$\mathbb{V}\left(X_k^{(m)}\right) \sim c + m^{2\mathbf{H}-2}. \tag{3.5.34}$$

Enfin, le tracé en échelle log-log des variances et de m doit se comporter comme une application affine de coefficient directeur $2\mathbf{H} - 2$.

3.5.5.3 Méthode du log-périodogramme : estimateur Geweke et Porter-Hudak (GPH)

La méthode Geweke et Porter-Hudak (GPH) (1983) consiste à effectuer une régression log-périodogramme autour du logarithme des basses fréquences. La procédure nécessite donc de choisir un seuil de variabilité m à intégrer lors de la régression, c'est-à-dire, de choisir un nombre qui reproduit les fluctuations des basses fréquences. On retient en général $m = \sqrt{T}$, T étant le nombre d'observations de la série, mais on peut trouver dans la littérature des raffinements sur le choix optimal de m, par exemple Hurvich et al. (1998) rapportent $m \sim T^{0.8}$.

En fait, la méthode GPH est développée dans le cadre de l'estimation de l'ordre d'intégration fractionnaire d des processus ARFIMA, mais puisque la densité spectrale d'un processus à mémoire longue est proportionnelle à $|\lambda_j|^{1-2\mathbf{H}}$, où $\lambda_j = \frac{2\pi j}{T}, j = 1, \ldots, T/2$ sont les fréquences de Fourier, alors le périodogramme doit être proportionnel à $|\lambda_j|^{1-2\mathbf{H}}$ des basses fréquences. Par conséquent, une régression log-log du périodogramme en fonction de la fréquence fournit une droite de pente $1 - 2\mathbf{H}$. La suite de cette

description est inspirée des travaux de (Lardic & Mignon, 1999, 2002), on y trouve aussi un état de l'art sur les autres méthodes spectrales pour les questions de dépendance à long terme.

La méthode GPH est basée sur la forme de la densité spectrale :

$$S_X(\lambda) = \left|1 - e^{-i\lambda}\right|^{-2d} S_\eta(\lambda), \tag{3.5.35}$$

où $S_\eta(\lambda) = \frac{\sigma^2}{2\pi}\frac{|\Theta(\exp -i\lambda)|^2}{|\Theta(\exp -i\lambda)|^2}$ est la densité spectrale du processus $ARMA(p,q)$: $\eta_t = \nabla^d X_t$. Le périodogramme de la série X est défini par :

$$I_x(\lambda) = \frac{1}{2\pi T}\left|\sum_{j=1}^{T} X_j \exp^{ij\lambda}\right|^2. \tag{3.5.36}$$

Une régression du log-périodogramme autour des basses fréquences (λ_0) revient à travailler sur la formule suivante :

$$\log(S_x(\lambda)) = \log(S_\eta(\lambda_0)) - d\log\left|1 - e^{-i\lambda}\right|^2 + \log\left[\frac{S_\eta(\lambda)}{S_\eta(\lambda_0)}\right]. \tag{3.5.37}$$

En ajoutant $\log(I_X(\lambda_j))$ de part et d'autre de l'égalité et en remplaçant λ par les fréquences de Fourier λ_j, on obtient une forme plus exploitable :

$$\ln(I_x(\lambda_j)) = \ln(S_\eta(\lambda_0)) - d\ln\left|1 - e^{-i\lambda_j}\right|^2 + \ln\left[\frac{I_x(\lambda_j)}{S_x(\lambda_j)}\right] + \underbrace{\ln\left[\frac{S_\eta(\lambda_j)}{S_\eta(\lambda_0)}\right]}_{\approx 0 \text{ pour } \lambda_j \to 0}. \tag{3.5.38}$$

Soient $\boldsymbol{y_j} = \ln(I_x(\lambda_j))$; $a = \ln(S_\eta(\lambda_0))$; $b = -d$; $\boldsymbol{z_j} = \ln\left|1 - e^{-i\lambda_j}\right|^2$et $\boldsymbol{\xi_j} = \ln\left[\frac{I_x(\lambda_j)}{S_x(\lambda_j)}\right]$ pour $j = 1, 2, \ldots, m$ (où m est le paramètre discuté au début de cette section), alors on peut représenter la formule 3.5.38 par modèle linéaire simple

$$\boldsymbol{y_j} = a + b\boldsymbol{z_j} + \boldsymbol{\xi_j} \tag{3.5.39}$$

et l'estimateur $\hat{d}_{GPH}$ de d par les MCO est fourni par :

$$\hat{d}_{GPH} = -\frac{\sum_{j=1}^{m}(\boldsymbol{z_j} - \bar{\boldsymbol{z}}_{\boldsymbol{j}})(\boldsymbol{y_j} - \bar{\boldsymbol{y}})}{\sum_{j=1}^{m}(\boldsymbol{z_j} - \bar{\boldsymbol{z}})^2}. \tag{3.5.40}$$

Sous réserve de la condition de stationnarité ($-0.5 < d < 0.5$), Geweke et Porter-Hudak (1983) soutiennent que $\hat{d}_{GPH}$ converge asymptotiquement ($T \longrightarrow \infty$) vers la loi normale :

$$\hat{d}_{GPH} \sim \mathcal{N}\left(d, \frac{\pi^2}{6\sum_{j=1}^{m}(\boldsymbol{z_j} - \bar{\boldsymbol{z}})^2}\right). \tag{3.5.41}$$

3.5.5.4 L'analyse des fluctuations redressées (ou Detrended Fluctuations Analysis DFA)

Depuis les publications de Peng (1994; 1995), l'analyse des fluctuations redressées (Detrended Fluctuations Analysis : DFA) a connu un grand succès pour l'analyse des séries temporelles non-stationnaires, en particulier pour la détection de la loi d'échelle sur de nombreuses séries temporelles.

Cette approche consiste à transformer la série originale en série intégrée par le calcul de la somme cumulée et l'écart à la moyenne :

$$\widetilde{Y}(k) = \sum_{j=1}^{k} \left(X_T(j) - \bar{X}_T \right), \text{ avec } k \in \{1, \ldots, T\}. \tag{3.5.42}$$

La série intégrée est ensuite divisée en fenêtres indépendantes de longueur équivalente m : la partie entière de T/m, $[T/m]$. Dans chaque bloc, la droite des moindres carrées est estimée, représentant la tendance de cette fenêtre. Une fois la série intégrée est redressée en lui retranchant la tendance locale, l'analyse porte sur les résidus de la régression. La fonction DFA représente l'écart-type des résidus de cette régression pour toute la série :

$$F(m) = \left[\frac{1}{m\,[T/m]} \sum_{k=1}^{m[T/m]} \left(\widetilde{Y}(k) - \hat{Y}_m(k) \right)^2 \right]^{\frac{1}{2}}. \tag{3.5.43}$$

De la même manière que dans les méthodes précédentes, on représente le $\log(F(m))$ en fonction de $\log(m)$.

3.5.6 Application aux données de la QAI

Avant de commencer l'analyse des résultats, les données ayant une tendance suite à une dérive de capteur ont été corrigées par une régression linéaire sur le temps et réajustées à leur niveau moyen, en particulier, lorsque le suivi des mesures couvre une longue période, notamment pour la campagne dans le bureau individuel en 2011 et celle dans l'espace paysager en 2013. Pour donner une idée sur l'ordre de grandeur, le capteur de CO_2 utilisé a une dérive d'environ 20 ppm sur une année.

3.5.6.1 Comportement du spectre au voisinage de la fréquence zéro

Nous avons tracé sur la Figure 3.5.1 la représentation bi-logarithmique du spectre de puissance obtenu à partir de chaque série des concentrations des particules dans le bureau individuel. Les ajustements ont été effectués sur un niveau de variabilité $m = \sqrt{T}$ de fréquence $\lambda \sim (4.4\,\mathrm{h})^{-1}$. Ce spectre présente diverses caractéristiques notables. La présence d'au moins une pente spectrale donne une première confirmation du comportement scalant des fluctuations. Hormis les particules de taille $1.8\,\mu$m, les ruptures au niveau de la variabilité fréquentielle ne semblent pas être significatives : une approximation avec une seule droite suffirait pour expliquer la présence d'une loi de puissance de type $f^{-\alpha}$. Pour la taille de $1.8\,\mu$m, la rupture aurait lieu entre les fréquences $(6\ \mathrm{h})^{-1}$ et $(2\ \mathrm{h})^{-1}$, soit après la dernière fréquence significative. On remarque un comportement quasi-plat du spectre après la fréquence $(0.5\,\mathrm{h})^{-1}$ pour la taille de $1.8\,\mu$m, ce qui suggèrerait une fluctuation type bruit blanc au niveau de ces fréquences.

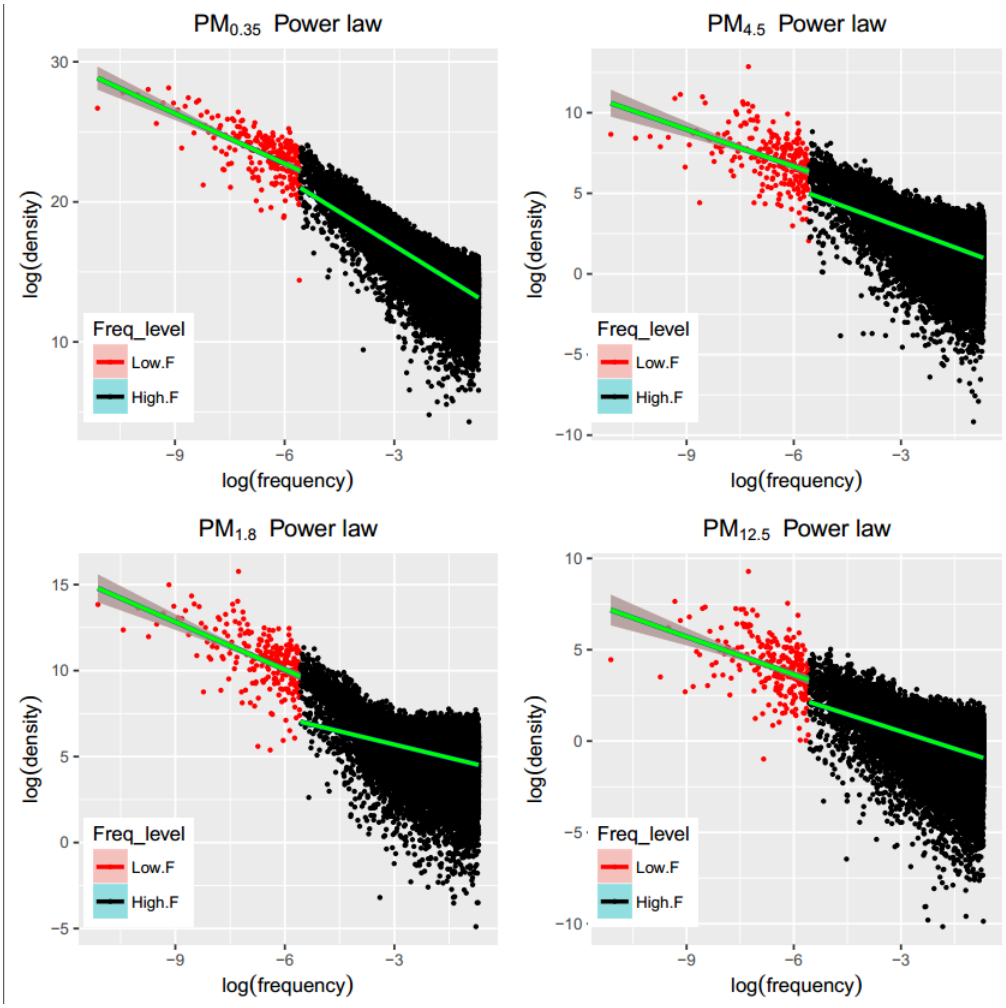


FIGURE 3.5.1 – *Comportement de la densité spectrale sur les données brutes de la concentration en particules par rapport au niveau de fluctuation de fréquence* $(4.4\,h)^{-1}$. *Les données sont issues des mesures effectuées toutes les minutes dans le bureau individuel en 2011 et sont exprimées en* $\# \cdot L^{-1}$.

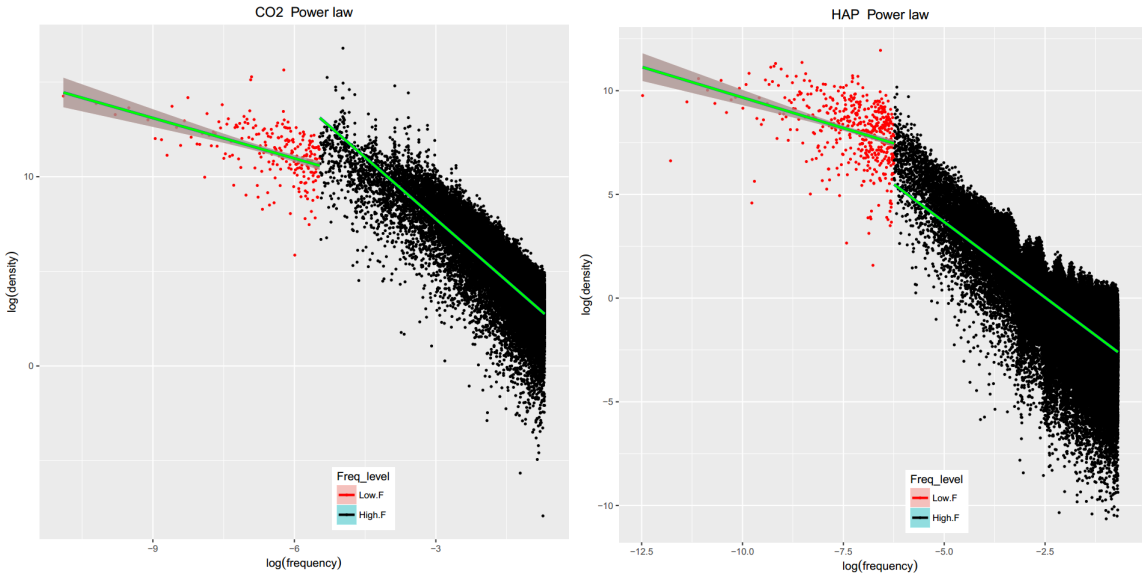


FIGURE 3.5.2 – *Comportement de la densité spectrale des données brutes de la concentration de CO_2 (à gauche) et des HAPs (à droite) par rapport au niveau des fréquences $(1.5\ jour)^{-1}$ et $(8.5\,h)^{-1}$, respectivement. Les données sont issues des mesures effectuées dans le bureau individuel en 2011, le pas de temps était de 10 minutes pour le CO_2 et d'une minute pour les HAP_s totaux.*

La même procédure a été appliquée pour les séries du CO_2 et des HAP_s. Sur la Figure 3.5.2, on présente les ajustements linéaires effectués sur les bi-log périodogrammes du CO_2 et des HAP_s à l'échelle de variabilité de $(1.5\ jour)^{-1}$ et $(8.5\,h)^{-1}$, respectivement. La pente de décroissance globale du CO_2 est plus élevée que celle des HAP_s. Néanmoins, le comportement spectral exhibe au moins deux régimes de variabilités fréquentielles. Sur les basses fréquences du spectre de puissance, la décroissance de la pente du spectre pour le CO_2 est beaucoup plus rapide que pour les HAP_s, et inversement pour les hautes fréquences.

La présence d'une fréquence principale et des pics secondaires pour le CO_2 augmente le niveau de la pente spectrale des hautes fréquences. De plus, il semblerait que la résolution temporelle joue un rôle important dans la détermination des différents régimes de variabilité spectrale.

Nous présentons sur la Figure 3.5.3 la variabilité fréquentielle des mesures de concentrations de CO_2 et des particules effectuées dans l'espace paysager durant la campagne de 2012. Rappelons que les données sont au pas de temps horaire. Ce que l'on peut dire, c'est que plus les tailles de particules sont fines, mieux on voit les niveaux de variabilité (ou plus). On remarque une coupure au niveau de la variabilité fréquentielle de $(2.6\ \text{jour})^{-1}$. La régression linéaire par morceaux sur le spectre bi-logarithmique des concentrations de particules de $1.8\,\mu$m ne semble pas très significative ; la décroissance de la densité spectrale se fait sur un seul morceau.

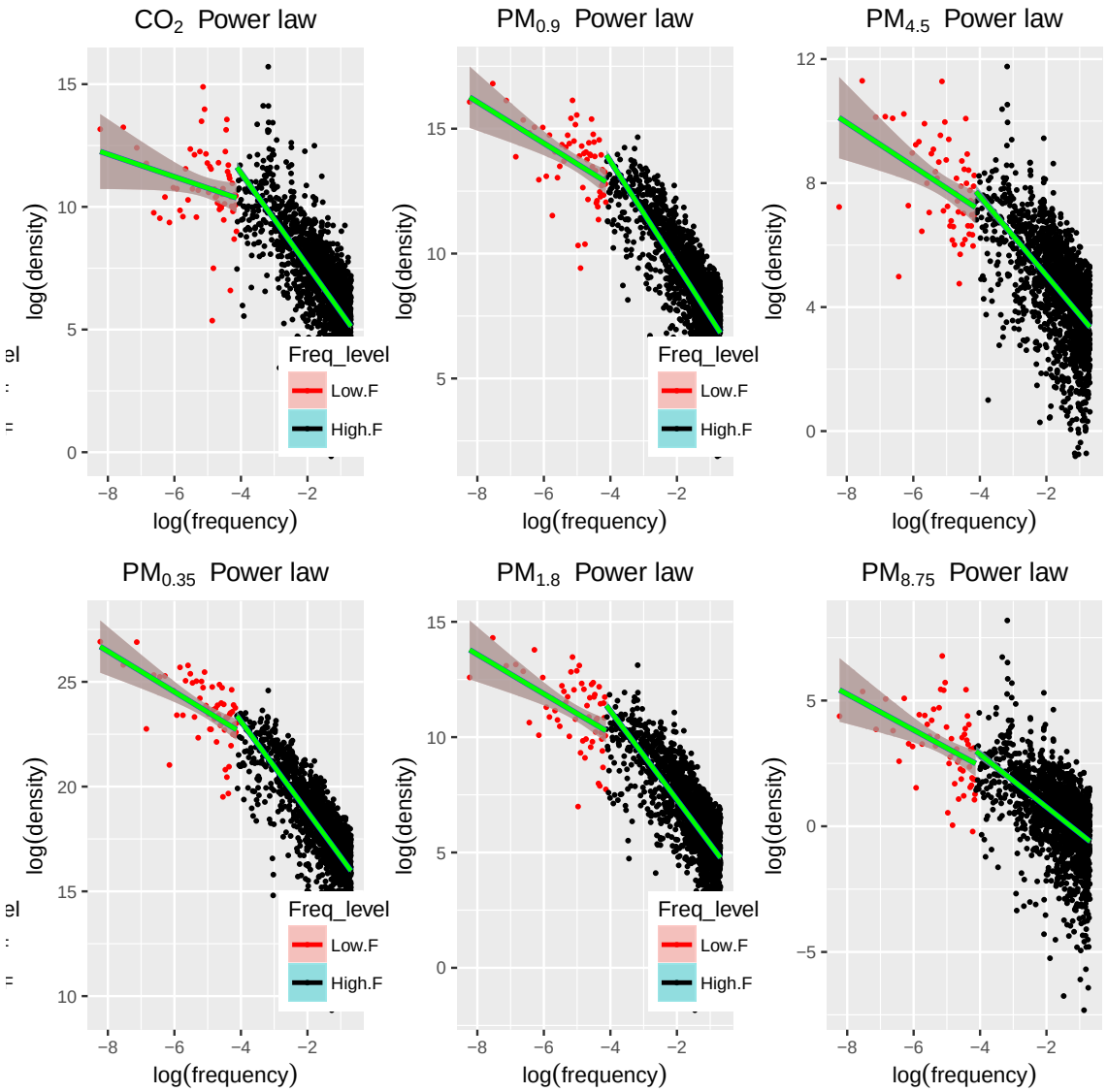


FIGURE 3.5.3 – *Comportement de la densité spectrale des données brutes de concentration en particules (différentes tailles) et du CO_2 en fonction de la de fluctuation de la fréquence* $(2.6\,jours)^{-1}$. *Les données correspondent aux mesures horaires effectuées dans l'espace de bureaux en 2012.*

Les pentes spectrales (droites ajustées aux densités) observées pour les différents paramètres des divers environnements sont reportées dans le Tableau 3.5.1. Tous les paramètres présentent un comportement scalant ; que ce soit sur les basses fréquences, sur les hautes ou sur tout le spectre, les coefficients de régression sont significativement différents de zéro (au sens de test statistique).

TABLE 3.5.1 – Estimation de la pente de régression bi-logarithmique des périodogrammes : BI (bureau individuel, campagne 2011) et OS12 (espace paysager, campagne 2012).

Pente $(-\alpha)$		λ_i		
Bureau	Paramètre	**Toutes**	**Hautes.freq**	**Basses.freq**
BI	$P_{0.35}\,\mu$m	1.635	1.608	1.19
	$P_{0.45}\,\mu$m	1.625	1.606	1.13
	$P_{1.8}\,\mu$m	0.628	0.513	0.92
	$P_{2.5}\,\mu$m	0.823	0.749	0.83
	$P_{3.5}\,\mu$m	0.944	0.902	0.78
	$P_{4.5}\,\mu$m	0.866	0.818	0.76
	$P_{6.25}\,\mu$m	0.852	0.81	0.75
	$P_{8.75}\,\mu$m	0.672	0.621	0.7
	$P_{12.5}\,\mu$m	0.675	0.629	0.69
	$P_{17.5}\,\mu$m	0.386	0.322	0.71
	HAP	1.477	1.456	0.58
	CO_2	1.99	0.711	2.17
OS12	$P_{0.35}\,\mu$m	1.925	0.968	2.15
	$P_{0.9}\,\mu$m	1.791	0.827	2.07
	$P_{1.8}\,\mu$m	1.661	0.855	1.92
	$P_{4.5}\,\mu$m	1.142	0.705	1.27
	$P_{8.75}\,\mu$m	0.946	0.714	1.04
	CO_2	1.562	0.461	1.89
	T	1.746	1.662	1.92
	HS	2.335	1.7289	2.51

Notes : la régression est effectuée sur toute la bande des fréquences $\lambda_j = \frac{2\pi j}{T}, j = 1, \ldots, T/2$, *où* T *est la taille de la série considérée (colonne* Toutes*). Sur les basses fréquences, la régression est effectuée sur l'ensemble des* m *premières fréquences :* $1, \ldots, m = \left\lfloor \sqrt{T} \right\rfloor$, *où* $\lfloor \bullet \rfloor$ *est la partie entière de* $\bullet$. *Sur les hautes fréquences, la régression est effectuée sur l'ensemble des dernières* $\frac{T}{2} - m$ *fréquences :* $m+1, \ldots, T/2$.

Enfin, nous passons à l'analyse du comportement spectral de la variabilité de formaldéhyde issu des mesures des trois campagnes : maison expérimentale (MARIA) et l'espace paysager (2013 et 2015). On présente sur la Figure 3.5.4 les périodogrammes des séries temporelles du HCHO ainsi que leurs régressions bi-logarithmiques.

Comme évoqué précédemment, la période principale dans la maison expérimentale et dans l'espace paysager durant la campagne de 2015 était d'un jour. En revanche, la densité spectrale de la série de 2013 n'exhibe pas de fréquence principale clairement séparable : le déclin du spectre est quasi-linéaire sur l'ensemble des fréquences, on ne remarque pas de rupture.

La régression bi-logarithmique sur le périodogramme de la série HCHO de la maison expérimentale présente une rupture au niveau de la période de 4 h. La variabilité fréquentielle des concentrations dans l'espace paysager montre un seul déclin. Néanmoins, en 2013, la résolution temporelle semble avoir un impact sur l'amplification du niveau de variabilité.

En plus de la régression linéaire, nous avons tenté une régression de type LOESS (voir la section ci-après) pour permettre plus de flexibilité sur la nature décroissante et irrégulière du spectre. Pour la série de 2015, on observe une forte variation fréquentielle sur la bande spectrale au niveau des fréquences entre $(1\,\text{jour})^{-1}$ et $(3.5\,\text{jour})^{-1}$ et une dynamique quasi-linéaire sur le reste du spectre. Ce comportement irrégulier de la série MARIA semble avoir lieu sur la bande $(4\,\text{h})^{-1}$ et $(1\,\text{jour})^{-1}$. Pour la série de 2013, la régression LOESS ne fournit pas plus d'informations par rapport à la régression linéaire par morceaux.

Ces observations, sur l'irrégularité au sein d'une bande spectrale, traduisent le niveau de fluctuation autour de la fréquence principale.

On note par ailleurs un artefact pour la série MARIA au niveau des hautes fréquences. Il est dû probablement à l'interpolation des données manquantes.

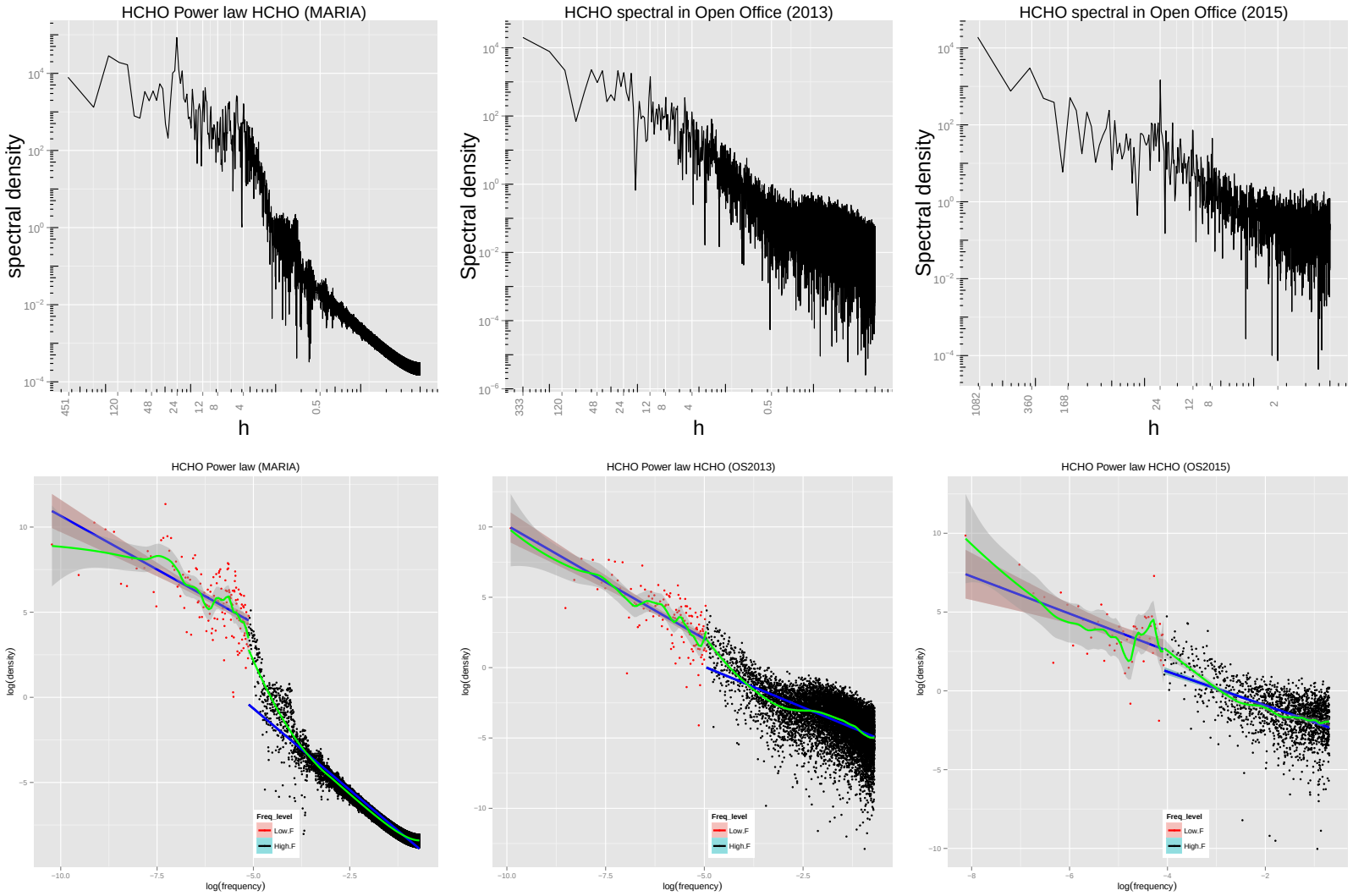


FIGURE 3.5.4 – *Densités spectrales de la concentration de HCHO observée dans deux environnements (MARIA et bureaux paysager) par rapport au niveau de fluctuation propre à chaque série. Les trois graphiques dans le panel du haut, sont présentés uniquement les densités spectrales en donnant en abscisse les périodes correspondantes au fréquences. Le panel des graphiques en bas donne les régressions bi-logarithmiques par morceaux linéaires (**en bleu**) ou par régressions localement linéaires (**en vert**).*

Quant à l'analyse quantitative sur le niveau de la pente spectrale, les résultats sont présentés dans le Tableau 3.5.2. Du point de vue test statistique ($t-$ Student), le coefficient de régression α est significatif.

Table 3.5.2 – Estimation de la pente de régression bi-logarithmique des périodogrammes des concentrations de HCHO dans les différents environnements étudiés et différentes campagnes. Le symbole $s.e$ représente l'erreur type (pour standard error en anglais) et l'intervalle de confiance de l'estimation du coefficient α est borné par sa valeur inférieure (Lower) et supérieure (Upper).

HCHO		α	$s.e$	Lower	Upper
MARIA	Toutes	-2.13	0.0056	-2.142	-2.119
	Basses.fréq	-1.25	0.1199	-1.495	-1.022
	Hautes.fréq	-1.91	0.0048	-1.921	-1.902
OS2013	Toutes	-1.3	0.013	-1.334	-1.281
	Basses.fréq	-1.61	0.132	-1.875	-1.351
	Hautes.fréq	-1.15	0.015	-1.187	-1.127
OS2015	Toutes	-1.23	0.032	-1.303	-1.175
	Basses.fréq	-1.17	0.239	-1.659	-0.698
	Hautes.fréq	-1.07	0.042	-1.154	-0.988

3.5.6.2 Mesure de dépendances

La mesure de dépendance au sein des séries temporelles traduit l'effet de persistance de l'information statistique au cours du temps. Elle permet donc de fournir une mesure de l'intensité de la dépendance à long terme. Nous nous proposons à présent d'appliquer les diverses méthodes d'estimation pour quantifier la persistance, en commençant par l'exposant de Hurst.

Estimation de l'exposant de Hurst des séries de la QAI

Les méthodes d'estimation de l'exposant de Hurst nécessitent généralement la condition de stationnarité (Mandelbrot, 1972). Afin d'éviter une "fausse" interprétation de la présence du phénomène de persistance, nous appliquons trois filtres sur la série brute (HCHO) pour rendre les séries stationnaires : différenciation d'ordre un (ΔHCHO), différentiation d'ordre fractionnaire d'ordre d estimé à partir de la méthode GPH et les résidus de la décomposition STL ($\epsilon - STL$). L'estimation de l'ordre fractionnaire des séries HCHO (OS2013 et OS2015) est donnée dans le Tableau 3.5.3.

Au vu des résultats figurant dans le Tableau 3.5.4, on constate d'abord que les deux méthodes utilisées conduisent à des résultats similaires sur la série différenciée de 2013 uniquement ; pour les autres séries, les valeurs sont très peu comparables. La statistique R/S de la série de HCHO (2013) est estimée à 0.72, alors que la méthode Higuchi donne une valeur de 0.968. Les autocorrélations sont toutes positives et décroissent à un taux hyperbolique. Cette observation peut s'expliquer par la présence des composantes déterministes au sein des séries. Ainsi malgré la taille de la série de 2013 (14 jours au pas de temps d'une minute), les basses fréquences sont très importantes. Quant à la série de 2015, la composante déterministe a été révélée par l'analyse spectrale : une fréquence principale se "détache" du spectre. On notera que la technique R/S tend à surestimer la valeur de l'exposant de Hurst sur une série fortement

TABLE 3.5.3 – Estimation de l'ordre fractionnaire $\hat{d}$ de la série HCHO par la méthode GPH.

$\hat{d}_{gph}$	$m = T^k$	$k = 0.2$	$k = 0.5$	$k = 0.8$
OS2013	HCHO	0.905	0.807	0.824
	ΔHCHO	-0.251	-0.194	-0.174
	$\epsilon - STL$	0.416	0.005	0.549
OS2015	HCHO	1.210	0.634	0.821
	ΔHCHO	-0.052	-0.367	-0.184
	$\epsilon - STL$	-1.824	-0.237	0.534

Notes : les valeurs en $k = 0.5$ pour la série brute de HCHO sont utilisée pour la différentiation fractionnaire $\Delta^d HCHO$. Le nombre de fréquences pris en considération est donné par la partie entière m.

saisonnière. D'ailleurs, la R/S fournit une valeur supérieure à 1 pour la série brute de 2015 (cellule grisée dans le Tableau 3.5.4), valeur difficilement interprétable.

L'analyse de la persistance sur la série brute n'est pas conseillée, car la dépendance à long terme doit être une caractéristique indépendante de la composante déterministe.

L'estimation de l'exposant de HURST par les deux statistiques, R/S et HIGUCHI sur les séries différenciées donne une valeur inférieure à 0.5 ; la variabilité se manifeste d'avantage sur les hautes fréquences. Ce résultat met en évidence le caractère de courte mémoire lorsque le filtre de différentiation absorbe toute fluctuation de basse fréquence ; donc la variabilité de haute fréquence recouvre la série. Avec ce filtre, les autocorrélations alternent de signe, donc le processus est anti-persistant : des phases de hausse ont tendance à être suivies par des phases de baisse.

Lorsque la série de 2013 est différenciée par un ordre fractionnaire de $d = 0.807$, la valeur de son exposant de HURST varie entre 0.44 par la méthode R/S et 0.59 par la méthode HIGUCHI. Ainsi, pour la même série différenciée de même ordre, la première technique conclut sur un phénomène d'anti-persistance alors que l'autre nous conduit à opter en faveur de la persistance (linge en gras dans le Tableau 3.5.4). Ce résultat nous montre la difficulté à conclure quant à la structure de dépendance des séries. Par contre par un $d = 0.634$ sur la série de 2015, les résultats tendent à conclure sur le phénomène de persistance des séries de HCHO.

Donc, entre une différenciation entière et une autre fractionnaire, les traitement statistique sous-jacent est très différent.

Par ailleurs, la statistique R/S du reste de la décomposition STL ($\epsilon - STL$) de la série de 2013 donne une valeur très proche de 0.5, ce qui laisse présager un comportement dont la dépendance à différents retards temporels quasi-nul. Au contraire, l'estimation par la statistique HIGUCHI tend à conclure sur la persistance de la série $\epsilon - STL$ ($H = 0.77 > 0.5$). Quant à la série $\epsilon - STL$ de 2015, l'estimation de l'exposant de HURST conduit à la présence d'un phénomène à mémoire longue, quelque soit la méthode d'estimation.

Globalement, nous retiendrons que les résidus de la décomposition STL semblent présenter une structure de dépendance à long terme, alors que les séries différenciées de premier ordre sont anti-persistantes.

Notons que cette thèse n'étudie pas la relation entre les estimations, la taille de l'échantillon et la résolution temporelle. Comment se propage l'erreur d'estimation due à une méthode par rapport à ces paramètres ?

TABLE 3.5.4 – Estimation de l'exposant de HURST des concentrations de HCHO. La variable HCHO est la série temporelle des donnée brutes ; $\Delta HCHO$ est la première différenciation de la série HCHO et la variable $\epsilon - STL$ représente la série des résidus de la décomposition STL.

Méthodes		$R/S-$modifiée			HIGUCHI		
Environnement	Variable	*Estimate*	*Std.Err*	*t-value*	*Estimate*	*Std.Err*	*t-value*
OS2013	HCHO	0.722	0.123	5.858	0.968	0.021	45.274
	ΔHCHO	0.412	0.029	13.974	0.399	0.030	13.504
	Δ^dHCHO	**0.446**	**0.026**	**17.402**	**0.594**	**0.037**	**15.887**
	$\epsilon - STL$	0.517	0.066	7.883	0.770	0.033	23.675
OS2015	HCHO	1.028	0.032	32.246	0.963	0.020	48.564
	ΔHCHO	0.407	0.016	24.707	0.298	0.016	18.804
	Δ^dHCHO	0.617	0.026	23.331	0.543	0.014	38.257
	$\epsilon - STL$	0.750	0.025	30.243	0.617	0.021	29.283

Notes : En fonction de la méthode d'estimation, les valeurs de $\hat{H}$ peuvent s'écarter d'une estimation à une autre. Le code source (en C et R) des fonctions pour l'estimation peut être consulté dans la page de web MURAD S. TAQQU(1995). *La méthode proposée par* HIGUCHI (1988) *est similaire à la procédure des valeurs absolues de la série agrégée.*

Enfin, le sujet étant très vaste, mais nécessite d'être approfondi ; un modèle de prévision doit prendre en compte toutes ces caractéristiques.

Estimation de la persistance par la dimension fractale des séries de la QAI L'existence d'un comportement scalant de la densité spectrale de la plupart des séries étudiées jusqu'à maintenant nous conduit à s'interroger sur l'existence d'une mesure invariante à l'échelle du temps d'observation. L'estimation de la dimension fractale permet, entre autres, de donner un indice caractérisant cette structure de variabilité.

La mesure fractale permet de quantifier les notions d'auto-similarité : propriétés statistiques comparables quelque soit l'échelle temporelle d'observation [5]. Cette dimension, qui caractérise le degré d'irrégularité d'une série temporelle, mesure la façon dont une série occupe l'espace. De manière très schématique, la relation essentielle caractérisant les séries fractales impose que la variance du signal soit une fonction puissance de la longueur de l'intervalle temporel sur lequel cette variance a été calculée :

$$\mathbb{V}(X_t) \propto \Delta t^{2\boldsymbol{H}}. \tag{3.5.44}$$

On retrouve dans cette équation l'exposant $\boldsymbol{H}$ de HURST. En dehors des considérations pratiques liées à l'environnement intérieur, le point de vue théorique sur la relation d'invariance d'échelle stipule que les séries sont des produits inhérents des systèmes complexes dynamiques, opérant aux frontières du chaos (Marks-Tarlow, 1999).

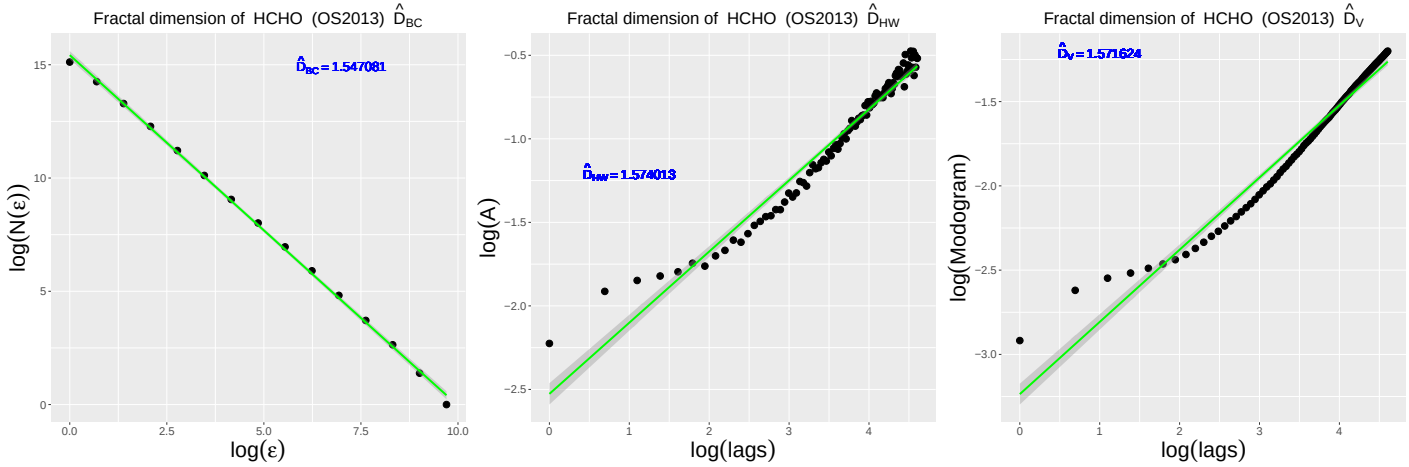
Nous avons appliqué les différentes méthodes d'estimation de la dimension fractale sur les différentes séries temporelles de la QAI.

5. À la question, "Quelle est la dimension d'un objet quelconque ?", B MANDELBROT, faisant référence aux côtes de la Grande Bretagne, répond que cela dépend de la distance à laquelle vous vous trouvez. Au fur et à mesure que nous nous approchons d'un objet, le travail de mesure devient de plus en plus complexe.

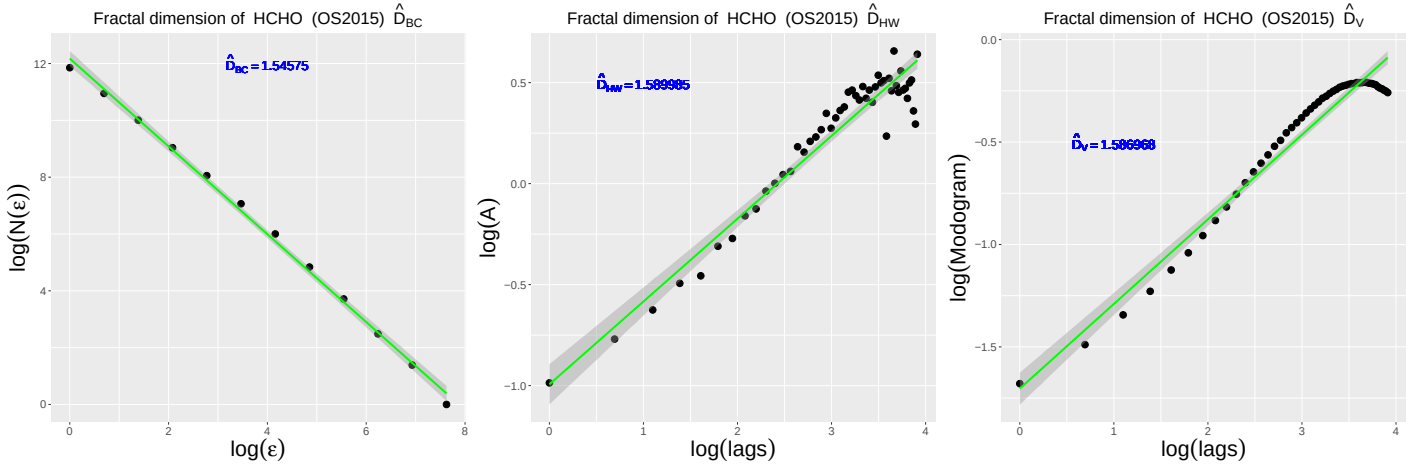
Nous présentons sur les Figures 3.5.5 et 3.5.6 l'estimation de la dimension fractale sur différentes séries temporelles de la QAI. Au vu des résultats globaux, la méthode Hall-Wood ($\hat{D}_{HW}$) et l'estimation variationnel ($\hat{D}_v$) sont très comparables, la dimension fractale de la série du formaldéhyde est estimée à une valeur avoisinant 1.54. Quant à la méthode de remplissage des boîtes ($\hat{D}_{BC}$), elle donne une valeur de 1.54. L'échelle de temps était d'une minute en 2013 et de 20 minutes en 2015.

Pour le CO_2, la dimension fractale estimée était de 1.48 pour les méthodes $\hat{D}_{HW}$ et $\hat{D}_v$ et de 1.54 avec la méthode de boîtes ($\hat{D}_{BC}$). Pour les fluctuations des HAPs, la dimension fractale estimée est similaire à l'estimation fractale de la série du formaldéhyde.

Ce résultat ne permet cependant pas de conclure sur la nature fractale des série étudiées, mais elle offre une mesure définissant la complexité de la variabilité temporelle des polluants.

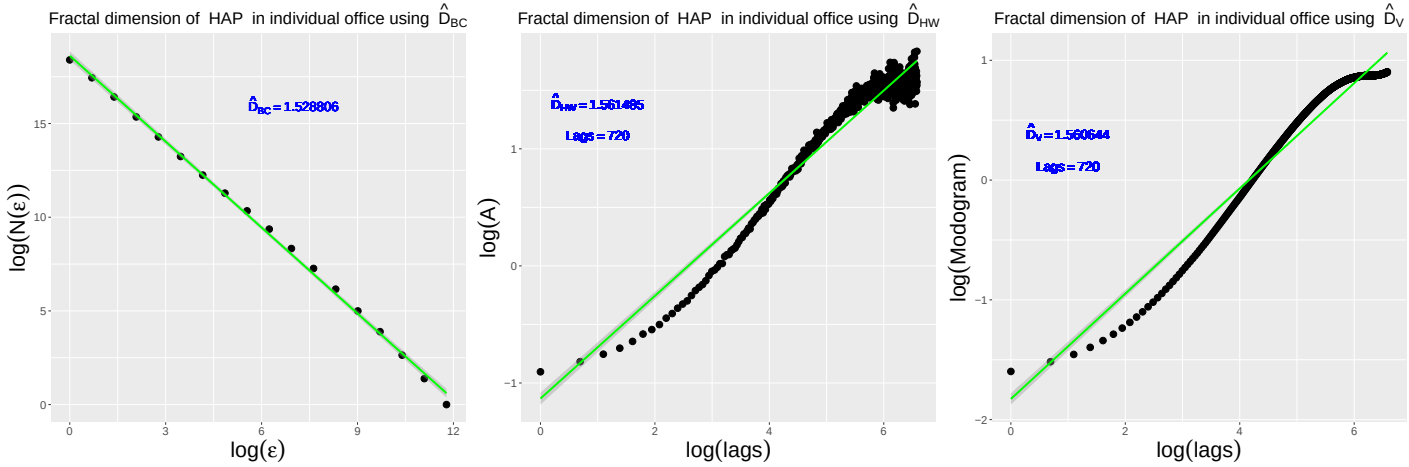


(a) *Estimation de la dimension fractale $\hat{D}$ pour la série de HCHO en 2013 avec $\hat{D}_{BC}$ (méthodes des boîtes), $\hat{D}_{HW}$ (*Hall-Wood*) et $\hat{D}_{V}$ (estimation variationnelle).*

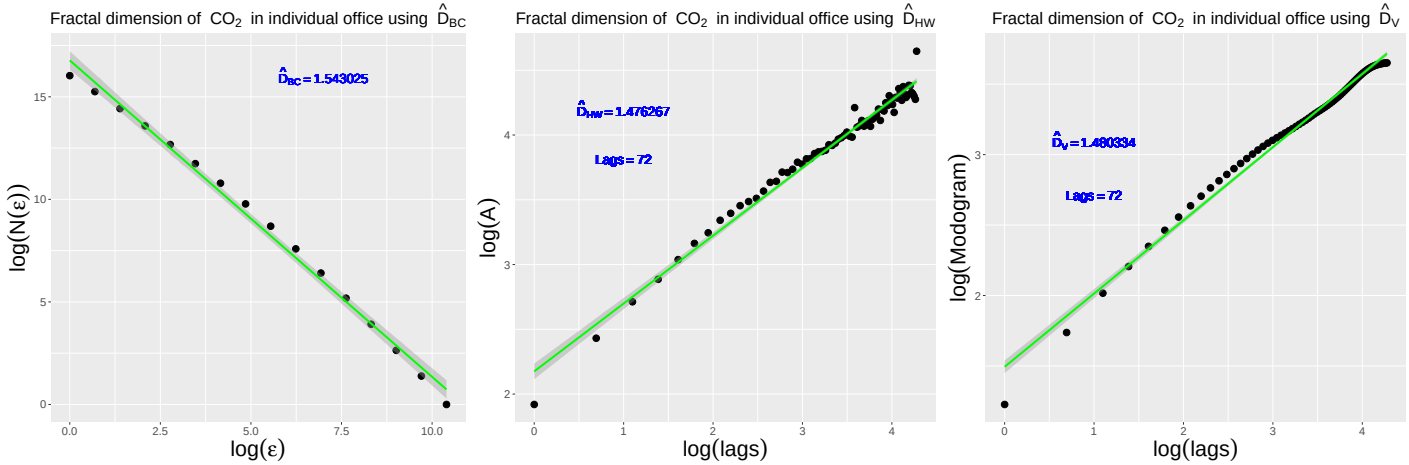


(b) *Estimation de la dimension fractale $\hat{D}$ pour la série de HCHO en 2015 avec $\hat{D}_{BC}$ (méthodes des boîtes), $\hat{D}_{HW}$ (*Hall-Wood*) et $\hat{D}_{V}$ (estimation variationnelle).*

Figure 3.5.5 – *Estimation de dimension fractale D des fluctuations des HCHO dans l'espace paysager (campagne de 2013 et 2015), par trois méthodes.*



(a) *Estimation de la dimension fractale $\hat{D}$ pour les HAP_s avec $\hat{D}_{BC}$ (méthodes des boîtes), $\hat{D}_{HW}$ (Hall-Wood) et $\hat{D}_V$ (estimation variationnelle)*



(b) *Estimation de la dimension fractale $\hat{D}$ pour les CO_2 avec $\hat{D}_{BC}$ (méthodes des boîtes), $\hat{D}_{HW}$ (Hall-Wood) et $\hat{D}_V$ (estimation variationnelle)*

Figure 3.5.6 – *Estimation de dimension fractale D des fluctuations des HAP_s et du CO_2 dans le bureau individuel (campagne de 2011), par trois méthodes.*

3.6 Décomposition des séries temporelles

3.6.1 Introduction

L'analyse des séries temporelles peut être appréhendée de différentes manières. L'une des façons est la décomposition en tendance, cycles (et/ou saisonnalité) et en composantes aléatoires. Cette décomposition repose sur l'idée selon laquelle la structure inhérente de la série temporelle est un composite de plusieurs éléments déterministes et stochastiques qu'il faut étudier séparément. Par ailleurs, l'étude de certaines propriétés, comme la dépendance à longue portée (la mémoire longue) requiert l'étude seulement de la partie aléatoire. En effet, la présence d'une tendance ou d'une composante périodique dégrade l'analyse de cette propriété.

Traditionnellement retenue, une tendance correspond à l'évolution sur le long terme, alors que les cycles correspondent à la dynamique sur le court terme (Doz et al., 1995). Cette caractérisation est néanmoins difficile à mettre en évidence sans prendre en compte le type de données utilisées (résolution temporelle et l'étendue de la série).

Toutefois, la diversité des méthodes proposées dans la littérature montre qu'il n'existe pas une construction théorique unifiée pour une telle décomposition. L'absence de consensus est due au fait qu'il est difficile de définir le poids de chaque composante, la forme de la tendance (linéaire, quadratique, exponentielle etc.), le nombre de fréquences pour la composante saisonnière et la nature de la partie aléatoire (homo/hétéro-scédasticité).

De plus, les hypothèses faites sur la superposition des composantes (additive et/ou multiplicatives) sont souvent déterminées par des connaissances *a priori* sur le phénomène étudié. En revanche, l'absence de ces informations sur les chroniques de la QAI requiert une analyse en utilisant des outils statistiques "neutres".

Dans cette section, deux méthodes de décomposition sont présentées et qui seront appliquées dans la partie prévision.

3.6.2 Problème de décomposition

Souvent, l'analyse des séries temporelles propose des méthodes exploratoires afin de comprendre la dynamique des variables. Une façon d'y procéder est d'extraire certaines composantes "latentes" de la série initiale. La dynamique temporelle est alors attribuée à un mélange d'une composante déterministe (la tendance et/ou la saisonnalité) avec une composante aléatoire (idéalement stationnaire) qui évoluent à des rythmes différents et de façon indépendante.

Dans le cadre de la modélisation à grande échelle pour des durées très longues (plusieurs dizaines d'années), on y ajoute souvent la composante cyclique qui correspond à un phénomène répétitif de période inconnue ou variable dans le temps. Dans le cas de notre étude, cette composante est difficile à détecter et ceci en raison l'étendue de nos séries (qui s'arrête à quelques mois).

Les composantes souvent explorées sont :

Tendance (T) : correspond à la dynamique de long terme, déterminant ainsi le sens de variation de la série ;

Saisonnalité (S) : correspond à l'évolution régulière autour de la tendance, possédant fréquemment une forme quasi-sinusoïdale ;

Bruit (e) : capture toute la partie aléatoire.

Les trois composantes peuvent être combinées : par une structure (i) additive : $X_t = T + S + e$; (ii) multiplicative $X_t = T \cdot S \cdot e$; ou (iii) par une structure mixte, comme par exemple $X_t = (T + S) \cdot e$. Il est évident que le modèle multiplicatif peut se ramener très vite en modèle additif par passage au logarithme.

3.6.3 Méthode basée sur une régression non-paramétrique

De façon générale, la régression non-paramétrique permet de décrire la relation entre une variable aléatoire Y et une variable explicative X, sans supposer la forme de la relation entre les deux variables. Nous avons présenté la méthode basée sur les fonctions splines, elle-même rentre dans le cadre de la régression non-paramétrique.

Le plus souvent, si une tendance apparaît clairement sur le graphique de la série temporelle, la décomposition commence par l'extraction de celle-ci. Il existe différents procédés, selon la nature des données à traiter, permettant d'analyser séparément les composantes latentes des séries temporelles.

3.6.3.1 Considérations théoriques

La méthode "Seasonal Trend Loess" (STL) proposée par Cleveland (1979) est une méthode de régression non-paramétrique par une pondération localement lisse (locally weighted scatterplot smoother (lowess) [6]). Pour une application aux concentrations de CO_2 extérieures, Cleveland et al. (1990) montrent l'effet de la tendance au cours des dernières années.

Considérons un échantillon aléatoire (x_i, y_i), $i = 1, \ldots, n$, du couple (X, Y). Alors le modèle de régression non paramétrique est le suivant :

$$y_i = \mathcal{G}(x_i) + \epsilon_i, \quad i = 1, \ldots, n, \tag{3.6.1}$$

où $\mathcal{G}(x)$ est la fonction de régression que l'on cherche à estimer, supposée lisse. Sous la condition que Y soit intégrable ($i.e.\, \mathbb{E}[|Y|] < \infty$), la fonction de régression $\mathcal{G}(x)$ est définie comme l'espérance conditionnelle de Y sachant $X = x$, soit

$$\mathcal{G}(x) = \mathbb{E}[Y \mid X = x]. \tag{3.6.2}$$

L'estimation de cette fonction est appelée fonction de lissage. Sa flexibilité est généralement contrôlée par des paramètres de lissage qui dépendent de la méthode utilisée. De façon générale, le contrôle de ces paramètres tient surtout à la dualité *biais-variance* associée à la minimisation de l'erreur quadratique moyenne MSE :

$$MSE = biais^2 + variance.$$

6. Loess et Lowess sont équivalentes.

La fonction de lissage par polynômes locaux consiste à approcher, par un polynôme de degré p, la fonction de régression $\mathcal{G}$ du modèle général 3.6.1 autour d'un point x. Si de plus, $\mathcal{G}$ est de classe $\mathscr{C}^p$ (fonctions réelles p continûment différentiables) au voisinage de x, alors par le développement de Taylor, on obtient :

$$\mathcal{G}\left(x'\right) \approx \sum_{j=0}^{p} \frac{\mathcal{G}^{(j)}\left(x\right)}{j!} \left(x' - x\right)^j = \sum_{j=0}^{p} \beta_j \left(x' - x\right)^j . \tag{3.6.3}$$

L'estimation des coefficients $\beta = (\beta_0, \ldots, \beta_p)^\top$ du polynôme local au point x se fait par minimisation du critère des moindres carrés pondérés :

$$\min_{\beta} \sum_{i=1}^{n} K\left(\frac{x_i - x}{\omega}\right) \left\{ y_i - \sum_{j=0}^{p} \beta_j \left(x_i - x\right)^j \right\}^2 , \tag{3.6.4}$$

où K est une fonction noyau (K pour kernel) et ω représente la fenêtre de lissage. Lorsque $p = 2$, la régression est dite localement quadratique et lorsque $p = 1$, elle est appelée régression localement linéaire. Cette dernière consiste à utiliser la fonction tricubique comme fonction noyau :

$$K\left(x\right) = \begin{cases} \left(1 - |x|^3\right)^3 & \text{pour } |x| < 1 \\ 0 & \text{pour } |x| \geq 1 \end{cases}$$

Ce noyau est proposé par Cleveland (1979) ; il a été dès lors, le plus utilisé et cette méthode de régression est appelée LOESS (ou LOWESS pour *LOcally WEighted Scatter-plot Smoother*).

La procédure STL est conçue principalement pour des séries à basses fréquences, donc pour des données journalières, voire même mensuelles. Ceci constitue un inconvénient majeur pour l'analyse de nos données, qui sont enregistrées avec un pas de temps fin (1, 10, et 20 minutes jusqu'à 1 heure). Néanmoins, on peut appliquer cette méthode pour des séries qui exhibent un comportement saisonnier.

La procédure d'extraction des composantes latentes consiste à :

- établir le profil de variabilité sur la période principale. La composante saisonnière S_t reproduit exactement le comportement de cette variabilité. La période principale est généralement prédéfinie par l'utilisateur en utilisant l'analyse spectrale.
- désaisonnaliser la série : $X_t^{(s)} = X_t - S_t$, avec $X_t^{(s)}$ est la série désaisonnalisée
- extraire la tendance $T(t)$ à partir de la série désaisonnalisée par la méthode LOESS : $X_t^{(ds)} = X_t^{(s)} - T(t)$;
- calculer le reste obtenu par la relation $e_t = X_t - S_t - T(t)$.

Dans ce type de décomposition, plusieurs paramètres peuvent influencer la nature de chaque composante. En premier lieu l'extraction de la composante saisonnière requiert la définition d'une seule fréquence principale ; or la série peut présenter plusieurs harmoniques. Deuxièmement, l'extraction de la tendance, elle aussi nécessite de définir le "span" ou la fenêtre dans la régression LOESS. Enfin, le reste ou le bruit dépend des deux autres composantes.

3.6.3.2 Application aux données de la QAI

Nous appliquons la méthode de décomposition STL aux différents jeux de données de l'air intérieur. On présente sur la Figure 3.6.1 une décomposition STL de la concentration en HCHO dans la maison expérimentale (MARIA). Pour cette série, le pic correspondant à l'activation de la source bâtonnets d'encens a été supprimé et remplacé par la moyenne de la série, car celui-ci altère considérablement la décomposition.

On se propose maintenant de donner une indication sur la contribution de chaque composante dans les fluctuations de la série originale. Elle se calcule en termes d'écart-type relatif par rapport à l'écart-type de la série originale et aux écarts-types des autres composantes. Soit $\widehat{\sigma_p}$ l'écart-type estimé de la série temporelle du polluant p ; après décomposition de la série, la somme des écarts-types des trois composantes est généralement supérieure à $\widehat{\sigma_p}$. La contribution relative (Ct_r) de chaque composante au niveau total de variabilité peut être estimée comme suit :

$$\widehat{Ct}_r = \frac{\widehat{\sigma_C} - \frac{\widehat{\sigma_{diff}}}{3}}{\widehat{\sigma_p}}, \tag{3.6.5}$$

où $\widehat{\sigma_C}$ est l'écart-type de chaque composante : S représente la saisonnalité, T la tendance T et e le reste, ($\widehat{\sigma_S}$, $\widehat{\sigma_T}$, $\widehat{\sigma_e}$), et $\widehat{\sigma_{diff}} = (\widehat{\sigma_S} + \widehat{\sigma_T} + \widehat{\sigma_e}) - \widehat{\sigma_p}$.

Le Tableau 3.6.1 donne les contributions relatives de chaque composante de la décomposition STL de quelques séries temporelles de polluants en air intérieur. On peut toujours se référer à la section 2.5 pour la description détaillée des séries.

Bien que la série du HCHO soit de pas de temps d'une minute, la décomposition semble relativement facile et une prévision par un modèle basé sur cette décomposition est envisageable car la composante saisonnière (ici diurne) est très importante. D'ailleurs, le calcul de la contribution relative (formule 3.6.5) de la composante diurne est majoritaire, elle est de l'ordre de 60 %, contre 35 % pour la tendance et 3.72 % pour le reste de la décomposition. Autrement dit, les principales composantes qui déterminent la variabilité du HCHO dans la maison expérimentale sont déterministes. Ceci laisse présager que l'occupation est la principale source de stochasticité dans les environnements intérieurs. Cette hypothèse est désormais vérifiée avec une telle décomposition.

La contribution de la tendance polynomiale reste majoritaire pour les particules de taille entre 0.35 – 2.5 μm et diminue avec l'accroissement de leur taille. Quant à la contribution relative de la saisonnalité, elle augmente avec la taille des particules, et elle est associée à une contribution importante des résidus de la décomposition.

Il s'avère que la méthode STL est très sensible au caractère discret de la variabilité des grosses particules ; l'estimation de chaque composante nécessite le contrôle de la taille de la fenêtre nécessaire pour l'extraction de la tendance, sinon elle est biaisée. Cette méthode est alors inadaptée pour la prévision des fluctuations des séries des grosses particules. Nous allons voir, par contre, son utilité pour la prévision des séries ayant des composantes très marquées par la saisonnalité.

Des exemples de décomposition par la méthode STL sont représentés dans les Figures 3.6.1 et 3.6.2.

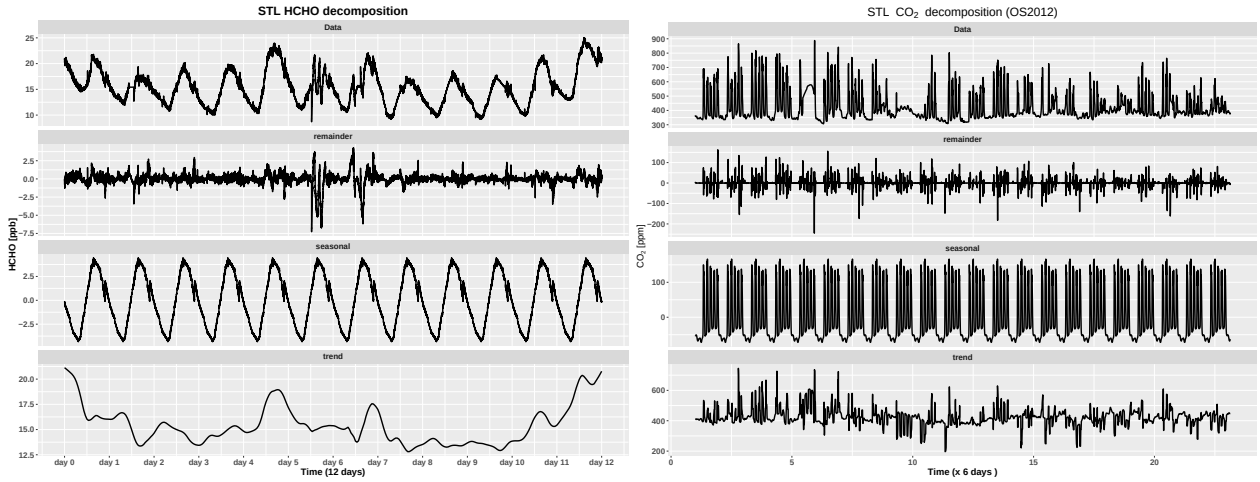


FIGURE 3.6.1 – *Décomposition par la méthode STL de la variabilité du HCHO dans la maison expérimentale (MARIA) mesurée toutes les minutes et du CO_2 dans l'espace paysager (campagne 2012) au pas de temps horaire. Les périodes fondamentales de la décomposition saisonnière sont d'un jour pour le formaldéhyde et de 7 jours pour le CO_2.*

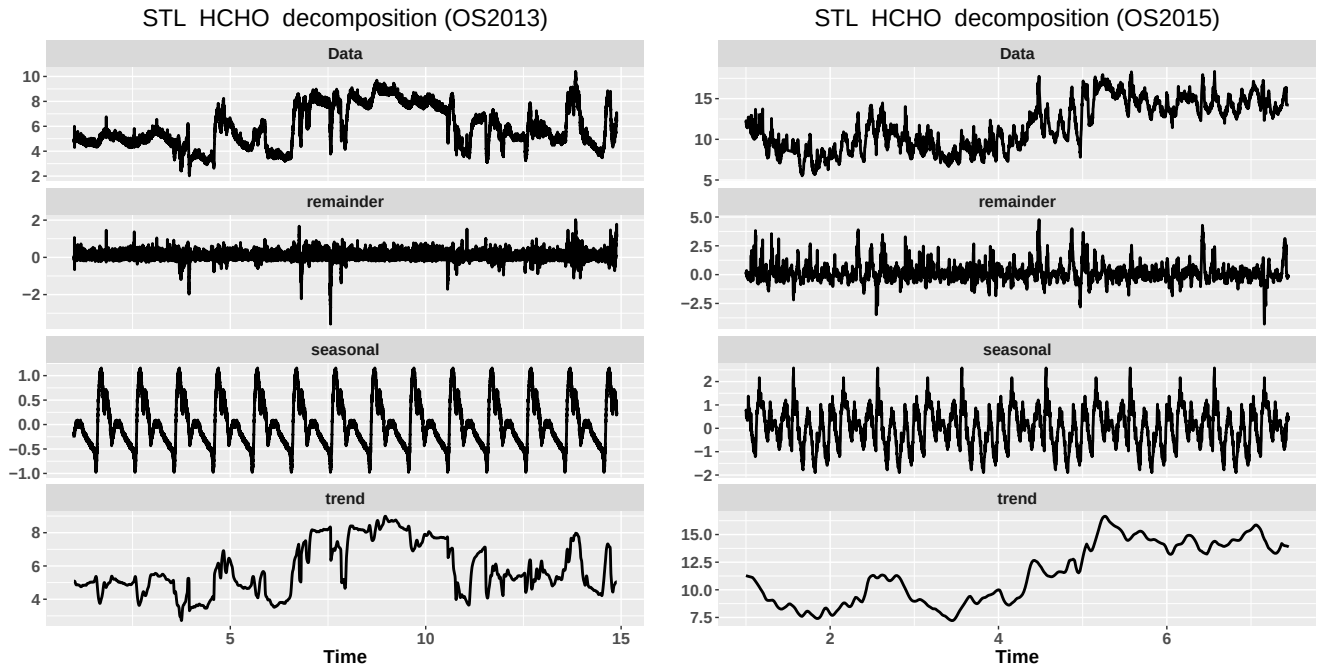


FIGURE 3.6.2 – *Décomposition par la méthode STL de la variabilité du HCHO dans l'espace paysage durant deux campagnes : en 2013 (pas de temps d'une minute) et en 2015 (pas de temps toutes les 20 minutes).*

TABLE 3.6.1 – Récapitulatif de la contribution en variabilité de chaque composante dans les séries de polluants de l'air et facteurs climatiques. La colonne *t.span* (trend span) correspond à la taille de la fenêtre retard (délai) pour l'extraction de la tendance.

Environnement	Variable	*t.span* (h)	Contribution (Ct_r %) Saisonnalité	Tendance	Reste
MARIA	HCHO	8	60.34	35.94	3.72
	$P_{0.35}$-$P_{0.45}$	8	7.36-6.21	79.26-75.60	13.38-18.19
	$P_{0.575}$-$P_{0.725}$	8	5.48-9.35	64.01-44.92	30.51-45.72
	$P_{0.9}$-$P_{1.3}$	8	8.26-7.52	46.56-54.62	45.18-37.86
	$P_{1.8}$-$P_{2.5}$	8	7.52-6.78	54.62-64.47	37.86-28.75
	$P_{3.5}$-$P_{4.5}$	8	7.27-14.46	42.48-20.12	50.25-65.41
	$P_{6.25}$-$P_{8.75}$	8	20.20-20.26	6.12-3.27	73.67-76.47
	$P_{12.5}$-$P_{17.5}$	1.2	19.46-23.38	27.57-18.32	52.96-58.28
OS2012	CO_2	6	57.25	39.19	3.56
	$P_{0.35}$-$P_{0.9}$	4	9.16-12.96	87.12-80.37	3.72-6.67
	$P_{1.8}$-$P_{4.5}$	4	17.05-36.61	75.65-58.05	7.29-5.34
	$P_{8.75}$	5	46.69	42.24	11.07
	T_{int}-HS_{int}	8-9	34.43-11.04	62.01-88.37	3.56-0.59
	T_{ext}-HS_{ext}	12-6	19.23-9.37	80.46-89.21	0.31-1.42
OS2013	HCHO	1.2	16.11	80.76	3.12
OS2015		24	11.68	76.52	11.78

3.6.4 L'analyse spectrale à décomposition singulière (SSA)

3.6.4.1 Introduction

La décomposition-reconstruction par l'analyse spectrale à décomposition singulière (Singular Spectrum Analysis, SSA) est une méthode non-paramétrique de l'analyse des séries temporelles. Elle est au carrefour de plusieurs disciplines, l'analyse classique des séries temporelles, systèmes dynamiques et la statistique/géométrie multivariée. Elle tire ses racines de deux importants résultats en statistique : la transformation de Kosambi-Karhunen–Loève et le théorème de plongement de Mañé-Takens (1981). L'adjectif "spectrale" est relatif à la décomposition en valeurs singulières des matrices, celle-ci est issue du théorème spectral très connu en algèbre linéaire ; donc, elle n'a pas de rapport direct avec l'analyse fréquentielle.

Ce sont les travaux de (Broomhead & King, 1986b,a) qui ont donné naissance à la méthode SSA. Depuis, un flux d'articles dans différents domaines de développement méthodologique et d'application l'on utilisée (Vautard et al., 1992; Vautard & Ghil, 1989; Groth & Ghil, 2011; Viljoen & Nel, 2010). Plusieurs monographies et livres sont dédiés à cette méthode (Elsner & Tsonis, 1996; Golyandina et al., 2001).

3.6.4.2 Algorithme et méthodologie

Une version condensée de la méthodologie peut être décrite de la manière suivante : soient la série temporelle $X_t = (x_1, x_2, \ldots, x_N)$, L un entier définissant la longueur de la fenêtre (nombre de lignes) et $K = N - L + 1$. On construit $L-$vecteurs retards et on constitue avec les $K-$vecteurs délais une matrice trajectoire $\boldsymbol{X}$. La décomposition de la matrice $\boldsymbol{X}\boldsymbol{X}^\top$ consiste en la recherche de L valeurs et vecteurs propres ; généralement cette procédure se fait par la Décomposition en Valeurs Singulières (SVD). Une combinaison particulière d'un certain nombre r de ces vecteurs propres détermine un sous-espace vectoriel $\mathcal{L}_r$ dans $\mathbb{R}^L$ $(r < L)$. Une projection des données L-dimensionnelles $[X_1 : \ldots : X_K]$ est alors effectuée sur le sous-espace $\mathcal{L}_r$; ensuite en faisant la moyenne des diagonales, le résultat "converge" à une certaine matrice de Hankel $\widetilde{\boldsymbol{X}}$. La série temporelle $(\widetilde{x}_1, \widetilde{x}_2, \ldots, \widetilde{x}_N)$ est une bijection avec $\widetilde{\boldsymbol{X}}$ et fournit ainsi une approximation de la série originale X_t avec un certain degré de fluctuation. On peut néanmoins reconstruire la totalité de la série en prenant tous les vecteurs.

Phase I : La décomposition

Étape 1 : plongement. Le célèbre théorème de Takens assure une reconstruction non ambigüe de l'espace des phases à partir de d'une série temporelle par la méthode du vecteur retard (ou décalage). De manière équivalente, Broomhead & King (1986b,a) ont proposé l'analyse des structures qualitatives des séries temporelles à partir du plongement :

$$X_i = (x_1, \ldots, x_{i+L-1})^\top, \quad i = 1, 2, \ldots, K. \tag{3.6.6}$$

La matrice trajectoire de la série X_N est donnée par la forme suivante :

$$\boldsymbol{X} = [X_1 : \ldots : X_K] = (x_{ij})_{i,j=1}^{L,K} = \begin{pmatrix} x_1 & x_2 & \ldots & x_{K-1} & x_K \\ x_2 & x_3 & \ldots & x_K & x_{K+1} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_L & x_{L+1} & \ldots & x_{N-1} & x_N \end{pmatrix}. \tag{3.6.7}$$

Cette matrice de dimension $L \times K$ possède deux propriétés principales :

1. Les vecteurs colonnes ou lignes de $\boldsymbol{X}$ sont des sous-séries temporelles décalées ;
2. Tout élément de $\boldsymbol{X}$ est égal à son anti-diagonal, $\boldsymbol{X}$ est une matrice de Hankel : $x_{ij} = x_{i+j-1}$.

Étape 2 : décomposition. Nous considérons maintenant l'espace vectoriel $\mathbb{R}^L$ de dimension $L \in \mathbb{N}^*$, et $\mathcal{E} = \{E_i\}_{i=1}^L = (\overrightarrow{e_1}, \overrightarrow{e_2}, \ldots, \overrightarrow{e_L})$ une base orthonormale [7] de $\mathbb{R}^L$.

La décomposition de la matrice trajectoire se fait de la manière suivante :

$$\boldsymbol{X} = \sum_{i=1}^{L} \boldsymbol{X}_i = \sum_{i=1}^{L} E_i Q_i^\top, \tag{3.6.8}$$

avec $Q_i = \boldsymbol{X}^\top E_i$, et $\|Q_i\|^2 = \|\boldsymbol{X_i}\|^2 = \lambda_i$. Pour une telle décomposition, deux choix de la base $\mathcal{E}$ peuvent être considérés :

7. $\mathcal{B}$ est orthonormée ssi : $\|\overrightarrow{e_1}\| = \|\overrightarrow{e_2}\| = \cdots = \|\overrightarrow{e_L}\| = 1$ et pour tout $i \neq j$, $\overrightarrow{e_i} \cdot \overrightarrow{e_j} = 0$ $(\overrightarrow{e_i} \perp \overrightarrow{e_j})$

1. *Basique* : $\{E_i\}_{i=1}^{L}$, les vecteurs propres de $\boldsymbol{X}\boldsymbol{X}^\top$; alors ce cas correspond à la décomposition en valeurs singulières (SVD) de $\boldsymbol{X}$, telles que $\boldsymbol{X} = \sum_i \sqrt{\lambda_i} U_i V_i^\top$, $U_i = E_i$ et $Q_i = \sqrt{\lambda_i} V_i$. Les réels $\lambda_1, \lambda_2, \ldots, \lambda_L$ sont les valeurs propres de $\boldsymbol{X}\boldsymbol{X}^\top$ ($\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_L \geq 0$) et Q_i. Dans la littérature, le triplet $\left(\sqrt{\lambda_i}, U_i, V_i\right)$ est appelé *"eigentriple"*.
2. *Toeplitz* : $\{E_i\}_{i=1}^{L}$ sont les vecteurs propres de la matrice G d'élément :

$$g_{ij} = \frac{1}{N - |i-j|} \sum_{m=1}^{N-|i-j|} x_m x_{m+|i-j|}, \ 1 \leq i, j \leq L. \tag{3.6.9}$$

Cette deuxième décomposition n'est approprié que dans le cas d'une série stationnaire et centrée (Golyandina, 2010). Donc, nous privilégions dès lors l'utilisation de la décomposition SVD dans les applications.

Phase II : La reconstitution

Étape 3 : groupage des triplets singuliers "*eigentriple grouping*" Une fois la décomposition faite, la procédure de groupage consiste en la séparation d'un ensemble d'indices $\{1, \ldots, d\}$ en m sous-ensemble disjoints $I_1, \ldots I_m$, où $d = \max\{j \ : \ \lambda_j \neq 0\}$. Soit $\boldsymbol{X}_I = \sum_{i \in I} \boldsymbol{X}_i$. La relation 3.6.8 donne lieu à

$$\boldsymbol{X} = \boldsymbol{X}_{I_1}, \boldsymbol{X}_{I_2}, \ldots, \boldsymbol{X}_{I_m}. \tag{3.6.10}$$

La procédure du choix d'ensembles I est appelée "eigentriple grouping". Si $m = d$ et $I_j = \{j\}$, $j = 1, \ldots, d$, alors le groupage correspondant est appelé "élémentaire" .

Étape 4 : moyennisation diagonale. À cette étape, nous transformons chaque matrice $\boldsymbol{X}_{I_j}$ de l'étape précédente (formule 3.6.10) en séries temporelles de la même taille N que la série initiale. La procédure est la suivante : soit $\boldsymbol{Y}$ une $L \times K$-matrice des éléments y_{ij}, avec $1 \leq i \leq L$ et $1 \leq j \leq K$. On pose $L^* = \min(L, K)$, $K^* = \max(L, K)$ et $N = L + K - 1$. On peut distinguer deux cas sur les éléments de $\boldsymbol{Y}$ en fonction des retards K et de L,

$$y_{ij}^* = \begin{cases} y_{ij} & si \ L < K, \\ y_{ji} & sinon. \end{cases}$$

Dans le cas le plus simple, $L \leq K$, la moyennisation consiste à calculer simplement la moyenne des éléments antidiagonaux :

$$\widetilde{y}_s = \frac{1}{\#(A_s)} \sum_{(l,k) \in A_s} y_{lk}, \tag{3.6.11}$$

où $A_s = \{(l,k) \ : \ l + k = s + 1, 1 \leq l \leq L, \ 1 \leq k \leq K\}$ et $\#(A_s)$ désigne le nombre d'éléments dans A_s.

De manière plus générale, la "moyennisation" consiste en la transformation de $\boldsymbol{Y}$ en une série $y_1, \ldots, y_N$ en utilisant les formules suivantes :

$$y_k = \begin{cases} \frac{1}{k}\sum_{m=1}^{k} y^*_{m,k-m+1} & \text{pour } 1 \le k \le L^*, \\ \frac{1}{L^*}\sum_{m=1}^{L^*} y^*_{m,k-m+1} & \text{pour } L^* \le k \le K^*, \\ \frac{1}{N-k+1}\sum_{m=1}^{N-k+1} y^*_{m,k-m+1} & \text{pour } K^* \le k \le N. \end{cases} \tag{3.6.12}$$

Ceci correspond à moyenner les éléments de la matrice sur les antidiagonales $i + j = k + 1$. Par exemple, pour $k = 1$, la première valeur y_1 de la série $y_1, \ldots, y_N$ est $y_{1,1}$; pour la deuxième valeur $y_2 = \frac{1}{2}(y_{1,2} + y_{2,1})$ et ainsi de suite jusqu'à la dernière valeur y_N.

Le processus (3.6.12) (ou dans le cas simple 3.6.11) appliqué aux matrices $\boldsymbol{X}_{I_k}$ reconstruit la série initiale X_N au travers des séries reconstruites associées à un certain degré de fluctuation : $\widetilde{X} = \left(\widetilde{x}_1^{(k)}, \widetilde{x}_2^{(k)}, \ldots, \widetilde{x}_N^{(k)}\right)$. En conséquence, la série initiale est décomposée en m sous-composantes latentes additives :

$$X_n = \sum_{k=1}^{m} \widetilde{x}_n^{(k)}, \ n = 1, \ldots, N. \tag{3.6.13}$$

3.6.4.3 Choix des paramètres et séparabilité

La question sur le choix des paramètres surgit du fait que la décomposition admet plusieurs formes. La notion de séparabilité pourrait permettre de répondre à cette question. La séparabilité d'un processus additif de deux séries temporelles $X_N^{(d)}$et $X_N^{(p)}$, $d \neq p$, signifie la possibilité d'extraire une variable de la somme des deux variables. SSA peut, approximativement, séparer un bruit additif d'un signal. Les composantes latentes de la série peuvent être identifiées en se basant sur le principe suivant :

la forme des vecteurs propres duplique la structure inhérente de la série qui a produit ces vecteurs propres. Le plongement de la série conserve les propriétés topologiques.

Par conséquent, les graphes des vecteurs propres peuvent aider dans le processus d'identification. Par exemple, le nuage de points des vecteurs propres d'une série sinusoïdale produit un T-vertex régulier ; car une série sinusoïdale mono-fréquentielle génère, exactement ou approximativement, deux sinusoïdales avec la même fréquence mais déphasées de $\pi/2$.

Par ailleurs, la matrice de corrélation pondérée, appelée $\boldsymbol{w}-$matrice de corrélation, entre les différentes composantes des séries reconstituées donne des informations très utiles pour la séparation : une bonne séparation fournit une très faible corrélation dans la $\boldsymbol{w}-$matrice. Donc en analysant cette dernière, on peut trouver les groupes des composantes fortement corrélées et construire une composante "intégratrice".

Sur la Figure 3.6.3, nous représentons ce point de vue par la décomposition SSA d'une série sinusoïdale mono-fréquentielle de taille 10^4 avec un $L = 200$. Il est évident qu'il suffit uniquement de deux vecteurs pour récupérer la totalité de l'information. Pour une série à trois fréquences, comme par exemple $f(t) = 0.2\sin(5\omega t) + 0.3\sin(9\omega t) + 0.1\sin(15\omega t)$, où $\omega = 2\pi f_0$ avec f_0 la fréquence principale, la décomposition de la matrice de délais avec les combinaisons des 7 premiers vecteurs propres est représentée dans la Figure 3.6.4. Clairement, avec une représentation à six vecteurs, la décomposition SSA sépare le signal du bruit. La contribution de la $7^{ème}$ composante est nulle, alors le nuage de points avec une autre composante significative, donne lieu à une contraction de cette composante sur elle même : aucun remplissage homogène de l'espace. La figure de la matrice de corrélation le montre plus clairement.

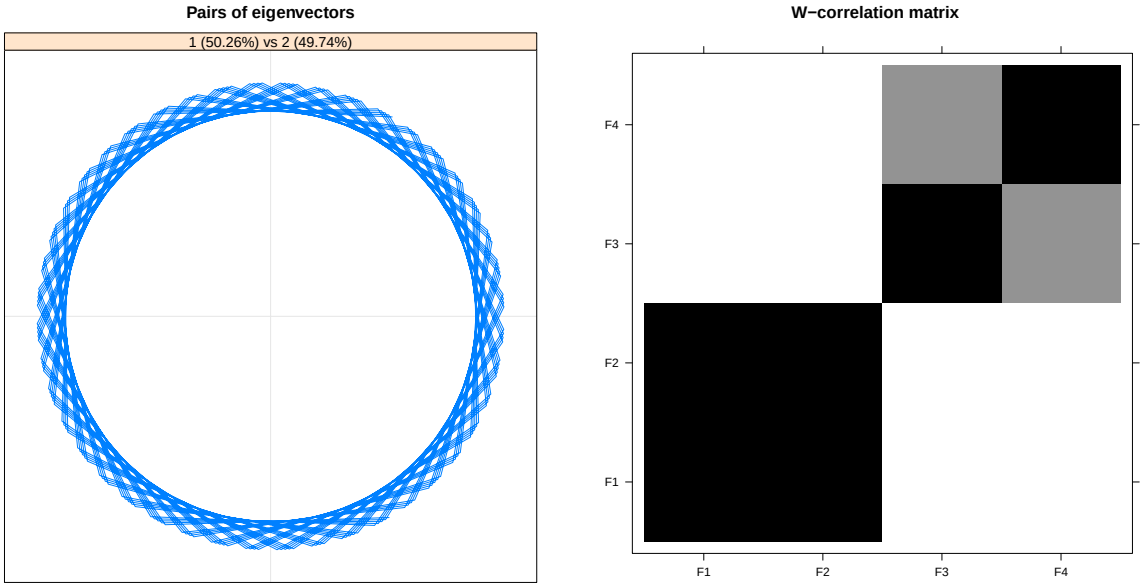


FIGURE 3.6.3 – *La procédure SSA sur une série temporelle sinus d'une seule fréquence principale. La taille de la série originale est de 10 000 et la reconstitution est faite sur* $L = 200$. *Les deux premiers vecteurs propres donnent un T.vertex avec une représentation de 100% (à gauche), la matrice* $w-$*correlation donne la corrélation entre les 4 différentes composantes séries reconstruites ; la séparabilité est observée à partir de deux composantes.*

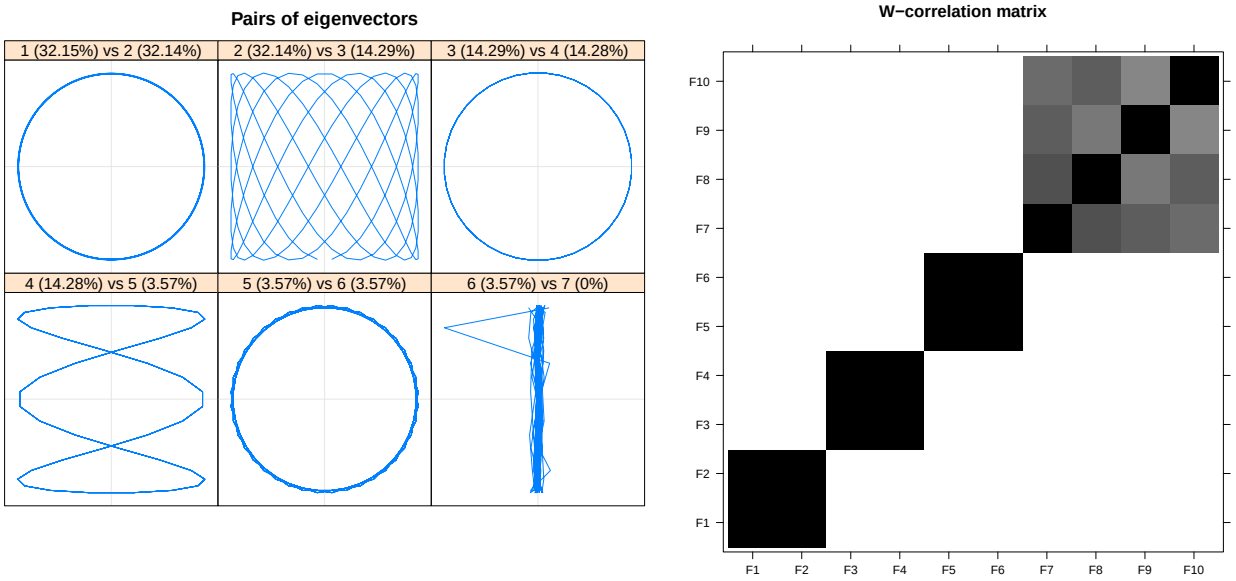


FIGURE 3.6.4 – *La procédure SSA sur une série temporelle sinus avec trois fréquences. La taille de la série originale est de 9901 et la reconstitution est faite sur* $L = 200$. *Les T.vertex sont localisés sur les combinaisons de vecteurs propres 1-2, 3-4 et 5-6. Le reconstitution avec 6 composantes représente 100% de la série originale (à gauche). La matrice* $w-$*correlation donne la corrélation entre les 10 différentes composantes séries reconstruites ; la séparabilité est observée à partir de six composantes.*

En situation réelle, la présence du bruit complique le choix des longueurs K et L, donc diminue la qualité de séparation. Cette question est très discutée par Golyandina et al. (2001) (section 1.6) et dans (Golyandina, 2010).

Globalement, si la structure de variabilité est assez régulière, par exemple une forte composante saisonnière, $L \sim N/2$ on peut extraire la composante la plus importante. Si, par contre, la trajectoire de la série exhibe des caractéristiques très irrégulières, il est recommandé d'utiliser la procédure séquentielle de SSA, c'est-à-dire, d'extraire d'abord la tendance avec un L petit, puis récupérer les résidus de cette décomposition et extraire en une deuxième étape la composante saisonnière en posant $L \sim N/2$.

3.6.4.4 Applications aux fluctuations de formaldéhyde

La procédure de la décomposition SSA a été appliquée aux séries temporelles du HCHO des deux campagnes de mesures dans l'espace paysager : 2013 et 2015. Les conditions de décomposition (formule 3.6.7) et de reconstitution SSA (formule 3.6.10) sont reportées dans le Tableau 3.6.2. La fenêtre de plongement a été fixée à 2880 pour la série de 2013, beaucoup moins de ce que préconisent Golyandina et al. (2001), car avec les données de hautes fréquences, l'utilisation de $L = N/2$ lisse énormément les fluctuations, on arrive à peine à reconstruire la variabilité diurne. En revanche, nous avons utilisé environ un tiers des données pour le plongement de la série de 2015 : $L \sim N/2$. En ce qui concerne le groupage des sous-matrices, donné dans la formule 3.6.10, celui-ci dépend de la nature des fluctuations. Pour les représentations, le choix est généralement fait par rapport à ce qu'on veut montrer, ou plus précisément ce qui est suffisant de montrer, car il est très difficile de représenter, par exemple 50 séries temporelles et chercher l'interprétation de chacune.

Nous représentons sur la Figure 3.6.5 la décomposition de la série des concentrations de HCHO durant la période de 2013 et sur la Figure 3.6.6, celle de 2015. Les figures donnent, de plus, les 12 premiers vecteurs propres présentés par paire. La structure de variabilité de la série est dupliquée par la forme des vecteurs propres issus de la décomposition SVD.

Contrairement à la série de 2015, il est très difficile de trouver une composante harmonique clairement identifiable sur la série de 2013. Nous remarquons aussi que la troisième composante issue de la décomposition de la série de 2013 présente une variance non constante au cours du temps.

Table 3.6.2 – Condition d'extraction des composantes latentes et de reconstitution des séries HCHO dans l'espace paysager par la méthode SSA.

	HCHO	
Campagne	*2013*	*2015*
N	19 000	2963
L	2880	1008
Tendance	(1,2)	(1,2,3)
Saison 1	(3,4)	(3,4,5)
Saison 2	(5,6)	(6,7,8)
Saison 3	(7,8)	(9,10)

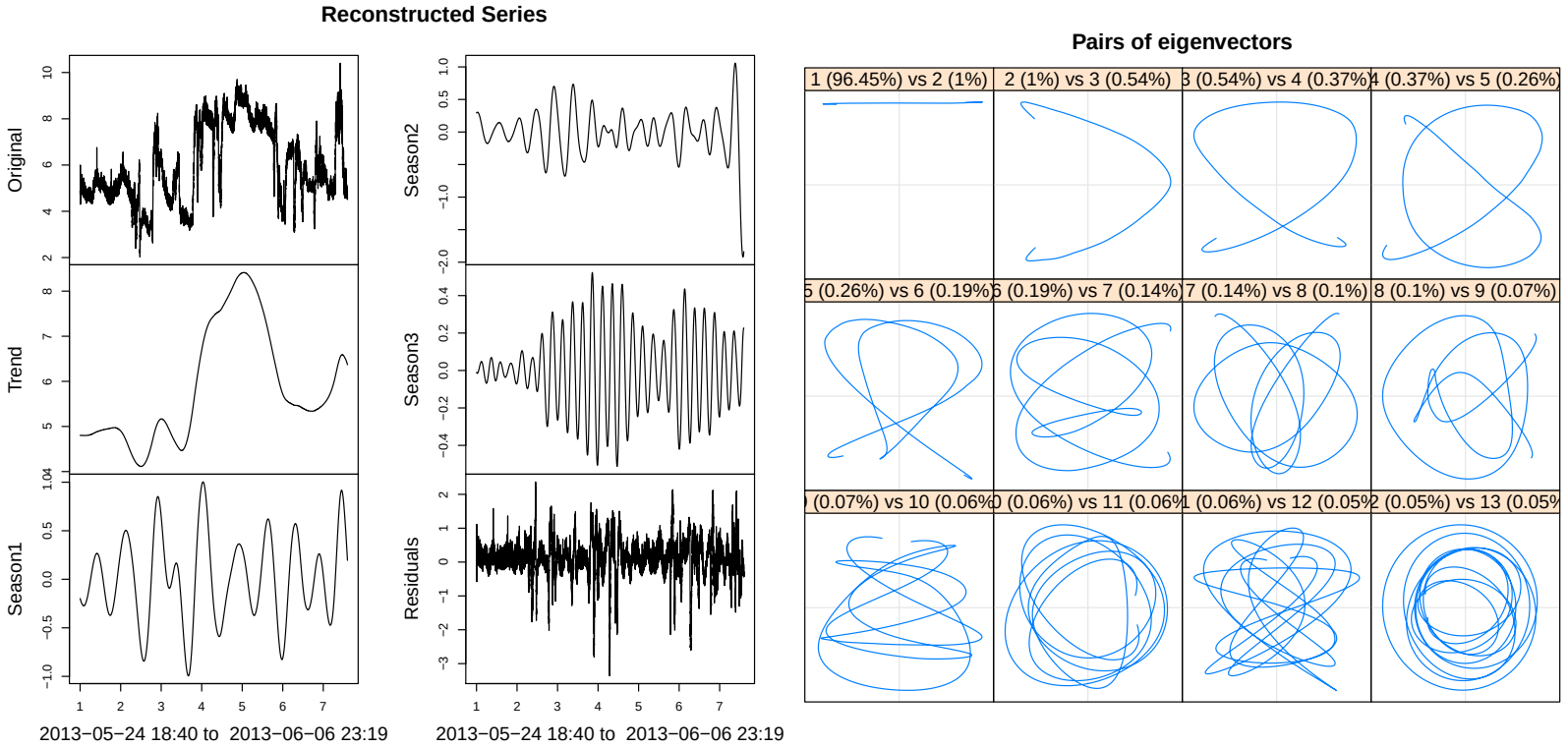


FIGURE 3.6.5 – *La reconstitution par SSA de la série temporelle de HCHO durant la campagne de 2013 (à gauche) dans l'espace paysager. La projection des 12 premiers vecteurs propres deux à deux (à droite). La taille de la série initiale est de* $N = 19\,000$ *minutes (série d'apprentissage utilisée plus tard pour la prévision) et la fenêtre de plongement est* $L = 2880$ *minutes (2 jours).*

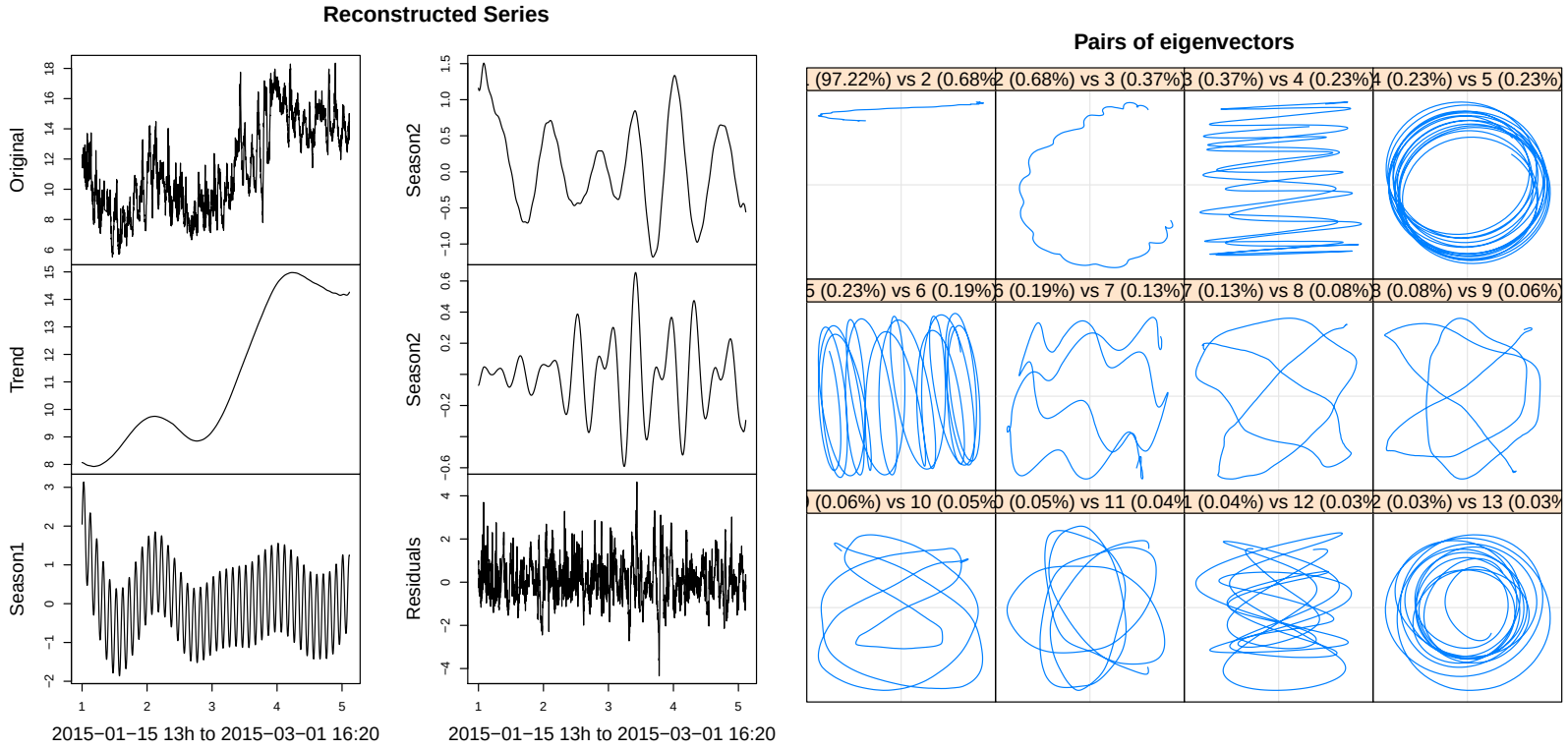


FIGURE 3.6.6 – *La reconstitution par SSA de la série temporelle de HCHO durant la campagne de 2015 (à gauche), le pas de temps est de 20 minutes. La projection des 12 premiers vecteurs propres deux à deux (à droite). La taille de la série initiale est de* $N = 2963 \times 20$ *minutes (série d'apprentissage utilisée plus tard pour la prévision) et la fenêtre de plongement est* $L = 1008 \times 20$ *minutes.*

3.7 Conclusion, discussion et perspectives

Nous avons étudié quelques propriétés spectrales ainsi que leurs relations avec l'effet de persistance des séries. En particulier, notre recherche s'est orientée vers les questions de fractalité(s) des séries et l'estimation des quantités relatives à la dépendance à long terme. L'interprétation de ces quantités dans le domaine de la QAI est encore sujette à discussion et mérite d'être approfondie. Les méthodes d'estimation, toutes empruntées des différentes disciplines peuvent s'écarter entre elles et nécessitent donc une étude à part. D'ailleurs, rien qu'avec la définition de la dépendance à long terme, et en fonction de ce qu'on en entend dire par "dépendance" on peut faire dériver plusieurs critères et quantités qui estiment la mémoire longue.

Étant donné le pas temps d'une minute utilisé dans la plupart des campagnes, l'utilisation de la fenêtre m donnée par Hurvich et al. (1998) ($m \sim T^{0.8}$) nous semble très élevée. Par exemple, elle représente environ 11 % de l'ensemble des fréquences pour les mesures de particules dans le bureau individuel ; l'estimation de la structure de dépendance dans ces conditions est très influencée par les hautes fréquences.

Par ailleurs, la persistance des séries et les changements de régime sont des phénomènes qui tendent aisément à être confondus (Diebold & Inoue, 2001; Kuswanto & Sibbertsen, 2008). Si on appréhende de modéliser nos séries par le changement de régimes, ce qui est d'ailleurs légitime pour certains polluants, il se peut qu'en réalité, le processus qui a généré les données soit à mémoire longue, ou les deux. Cette "confusion" sur les structures des séries expliquerait probablement le fait que dans quelques études, elles présentent des résultats acceptables que ce soit par la mémoire longue ou par le changement de régime. Une piste très peu exploitée dans la littérature pourrait permettre de mettre en évidence simultanément les deux phénomènes, à savoir des modèles de type Markov-Switching Multifractal (MSM).

Une autre question s'impose à nous lors de la caractérisation des séries temporelles ; elle concerne les relations entre les processus de type mémoire longue et les séries dont le processus est généré par une dynamique du chaos. Selon Peters (1994; 1996), cette relation peut se matérialiser, d'une part, par le fait que ces processus partagent les mêmes caractéristiques d'échelle : fractales, exposant de Hurst et le degré d'intégration fractionnaire. D'autre part, l'exposant de Lyapunov permet de détecter la présence de mémoire longue, dans le sens où ce dernier peut être relié à l'analyse R/S de Hurst. Sur le premier argument de Peters (1994; 1996), il nous apparaît très difficile de justifier "empiriquement" ces relations par le simple fait que les estimations données par différentes méthodes peuvent, "parfois" conduire à des conclusions contradictoires. Malheureusement, la relation $D = 2 - H$ livre une dimension d'un processus stochastique, il est donc très difficile de la relier à la dimension fractale de l'attracteur chaotique.

Aussi, dans la monographie de Guégan (2003), l'auteur soutient que les systèmes chaotiques déterministes peuvent présenter une mémoire longue ce qui permet alors de les utiliser en prévision, donc elle remet en cause l'idée communément admise que les systèmes chaotiques ne sont pas prédictibles.

Dans un environnement maîtrisé de la QAI, l'information portée par la dynamique des concentrations arrive régulièrement et de manière prédictible. Au contraire, dans un environnement occupé, l'occupant comme principale source de stochasticité, introduit une déformation irrégulière dans la dynamique. Ces déformations se répercutent à des échelles de temps très fin, avec alternances, de périodes plus ou moins fortes dilatations, qui ne répondent pas seulement à une échelle de temps calendaire de type heure, jours, mois, mais aussi à la régularité de phénomènes statistiques : l'échelle de temps est dans ce cas, un objet fractal.

Durant mon travail, j'ai pu remarquer que très peu d'études, voir aucune, n'a été rapportée dans l'analyse de ces propriétés pour la QAI. En outre, la littérature existante pour les applications à l'environnement,

bien que constituée d'un corpus déjà assez vaste, présentait encore des zones d'ombre, tant au niveau mathématique qu'application.

Enfin, pour pouvoir comparer, la plupart des études effectuées dans l'air extérieur ne correspondent pas à la nature des données que nous traitons, non seulement parce qu'elles n'explorent pas des échelles plus fines, mais la structure de variabilité diffère : les séries de l'environnement intérieur sont plus "accidentées". Alors, en entendant cette assertion, on peut dire qu'elles sont plus difficiles à prévoir, c'est ce que nous avons cherché à quantifier au début de ce "long" chapitre. L'estimation de la mesure de prédictibilité proposée par Goerg (2013), bien qu'intuitivement très séduisante et statistiquement solide, nous a permis de soulever quelques critiques liées à son estimation. Les deux points critiques les plus importants, à nos yeux sont :

- la mesure ($\Omega - \textit{prédictibilité}$) est très sensible à la taille de la série ; nous en avons fourni un exemple pour les séries sinusoïdales. La comparaison de deux séries de tailles différentes, éventuellement de différentes résolutions temporelles est très difficiles car les deux séries ne possèdent pas les mêmes fréquences ; alors que la définition est basée sur l'entropie différentielle de la densité spectrale.

- il n'y a pas de borne supérieure ; au sens de ce qu'on pourrait appeler "prédictibilité" d'une série temporelle devient très difficile cerner.

Face à ces problèmes, la comparaison nous semble très délicate ; voire dépourvue de sens. Alors pour pallier à ce problème, nous avons proposé de normaliser la série par rapport à la Ω_{sinus} correspondant à la taille de la série à comparer.

En définitif, la caractérisation des séries temporelles de concentrations de polluants de l'air intérieur nécessite une étude beaucoup plus approfondie. Ce que nous avons présenté n'est finalement qu'un regard très furtif sur le domaine l'**environnemétrie intérieure.**

Deuxième partie

SOURCES DE VARIABILITÉ ET MODÈLES DE PRÉVISION POUR LA QUALITÉ DE L'AIR INTÉRIEUR

Courte introduction de la partie

> "*La prévision empirique, c'est l'attente des cas similaires et ne requiert pas une connaissance rationnelle des causes et des effets, mais seulement le souvenir des faits observés et de la manière dont ils ont coutume de se succéder : ce sont des expériences répétées qui font naître l'habileté*". Emmanuel KANT, didactique anthropologique, page 60.

> *Remember that all models are wrong ; the practical question is how wrong do they have to be to not be useful*[8]. GEORGE, E. P. BOX et DRAPER, N. R (page 74, Box & Draper, 1987).

Parmi les plus importants succès scientifiques, et sans doute les plus recherchés est la prévision ou de façon générale la prédiction (Leconte, 2013). Ainsi, Thalès prédit une éclipse, Newton le retour de la comète de Halley. Quelques années plus tard, la découverte de Neptune par le calcul à partir de la trajectoire et des caractéristiques d'Uranus, Einstein prédit la courbure des rayons lumineux par le soleil et le prévisionniste James Stagg convainquit D. Eisenhower de changer le jour du débarquement pour le matin du 6 juin 1944. Récemment, Higgs prédit l'existence d'une nouvelle particule, et a reçu à cet effet le prix Nobel de physique en 2013. Ces exploits historiques prouvent que l'activité de prévision occupe une place centrale dans dans tous les domaines, notamment dans un contexte où la complexité est élevée.

Dans leur monographie *"Forecast verification : a practitioner's guide in atmospheric science"*, Jolliffe & Stephenson (2011) considèrent que le processus de prévision revient à considérer que disposer d'une information sur le futur est préférable à l'ignorance. Cette assertion suggère, comme la définit Clements & Hendry (2002), que tout jugement porté sur le futur est une prévision [9]. Mais la prévision doit répondre à un besoin scientifique ou pratique. Elle est *utile* lorsqu'on peut optimiser le processus de prise de décision.

Ceci est d'autant vrai si ce que l'on cherche à prévoir concerne non seulement l'ensemble des états possibles dans le futur, mais aussi leurs valeurs aux instants donnés précis dans le futur. Cette définition stipule qu'un modèle de prévision doit s'appuyer sur un ensemble d'informations disponible jusqu'à l'instant t, c'est ce que nous avons appelé par la filtration $\mathcal{F}_t^X$ sur un espace de probabilité $(\Omega, \mathcal{F})$, représentant ainsi l'information véhiculée par le processus X jusqu'à la date t.

La prévision est difficile quel que soit le domaine d'application ; prévoir la variabilité des chroniques issues des mesures de la QAI ne déroge pas de cette règle. En effet, la qualité de l'air intérieur est la résultante de nombreux phénomènes physico-chimiques complexes ; elle est en outre conditionnée par de différentes composantes qui influencent les ambiances intérieures, notamment les occupants et leurs activités. La singularité des ces ambiances tient au fait que l'occupant agit sur les différentes composantes du bâtiment de manière aléatoire, mais réagit presque de la même manière aux conditions de l'ambiance intérieure pour optimiser son confort. Comment tenir compte de ces variables dans un modèle de prévision ? Dans quelles conditions sont-elles utiles ? Ces questions ont été abordées partiellement dans les chapitres 2,

8. *"Rappelez-vous que les modèles sont faux, la question pratique est de savoir jusqu'à quel point ils mauvais pour qu'ils ne soient pas utiles".*

9. Emmanuel Kant dans son livre **Didactique Anthropologique** donne presque la même réflexion : "*C'est la faculté de se représenter quelque chose comme futur [page 57]... En effet, tout désir comporte une prévision, douteuse ou certaine, de ce que ces forces permettent : on ne tourne les yeux vers le passé (souvenir) que pour rendre possible la prévision du futur car, en général si nous regardons autour de nous, c'est du point de vue du présent, pour nous décider ou pour nous préparer à quelque chose [page 60]*".

3 et dans la partie II de ce manuscrit. Dans cette partie, nous verrons en quoi ces variables servent en prévision.

L'approche que nous adoptons est complètement statistique : elle s'affranchit donc des différents processus phusico-chimiques de formation, de transformation et de transport de ces polluants ; seuls les effets sont pris en compte. Cette approche nécessite une exploitation des composantes inhérentes aux fluctuations de la série ainsi que l'exploitation des différents facteurs. Dans la partie I, nous avons identifié quelques unes : la non-stationnarité, la non-linéarité, les sauts abrupts, la dépendance à long terme, qui sont les plus importants.

Pour explorer le champ de recherche des modèles de prévision, nous nous appuyons d'abord sur la typologie générale des modèles de séries temporelles proposées par Gouriéroux et Montfort : (*i*) les modèles auto-projectifs, (*ii*) les modèles explicatifs et (*iii*) les modèles de décomposition (ajustement) (Gourieroux & Monfort, 1995). Ensuite, nous déclinons cette typologie par deux classes de modèles : les modèles paramétriques et les modèles non-paramétriques.

Cette distinction repose sur le rôle des variables dans la modélisation et la forme du modèle qui décrit ces variables. Pour les modèles auto-projectifs, les variables peuvent en effet être descriptives dans le sens où seule la variable endogène retardée explique l'intégralité de l'information contenue dans la chronique. C'est-à-dire, seul le temps constitue une variable exogène non aléatoire. Dans ce cas, plusieurs modèles paramétriques (TAR, GARCH, bilinéaire etc.) ou non paramétriques (réseaux de neurones artificiel (ANN), chaînes de Markov cachées (HMM), dynamique du chaos, etc.) peuvent être utilisés.

D'autres variables, dites explicatives peuvent s'introduire dans les deux types de modèles afin de prendre en compte leurs influences sur la variable endogène. Dans ce cas, on est face à la modélisation multivariée des séries temporelles, c'est l'extension directe des modèles univariés.

L'approche de décomposition, quant à elle, consiste en l'exploitation des composantes latentes du signal intégrateur et fournit la prévision de chaque composante. Dans ce type de modélisation, rarement les modèles paramétriques sont utilisés. Une approche plus complète est possible : l'hybridation de plusieurs modèles de différentes manières.

Une description théorique exhaustive des modèles de prévision n'est pas envisageable dans ce manuscrit, puisque les conditions pour lesquelles ils sont développés sont variables selon le domaine d'application. Cependant, une revue de littérature actuelle sur les différents modèles statistiques appliqués à l'environnement, complétée par des exemples est proposée dans le premier chapitre de cette partie. Il s'agit pour nous, de discuter le positionnement de nos choix dans le cadre plus général des approches de modélisation possibles. Et ceci, en raison d'absence "complète" de modèles de prévision pour la qualité de l'air intérieur.

Nous nous souscrivons dès lors, par une approche partant du simple aux plus complexe, respectant ainsi le principe de parcimonie du rasoir d'Ockham : ne pas multiplier les hypothèses au-delà du nécessaire, en d'autres termes, **privilégier l'hypothèse la plus simple tant que cela reste compatible avec les observations** [10] **.**

Donc, nous avons d'abord tenté une modélisation classique de type linéaire associée une méthode de décomposition en les appliquant aux différentes séries de mesures (Chapitres 7-8). En fonction des exigences liées à la nature des différents polluants et à leurs structures de variabilité, nous avons proposé une approche qui consiste à utiliser la décomposition en bandes spectrales couplée avec des modèles non-linéaires de type TAR ou dynamique du chaos.

10. Une formule plus au moins similaire partagée par Aristote en 350 avant notre ère qui, déclara : "*C'est en vain que l'on fait avec plusieurs ce que l'on peut faire avec un petit nombre*".

CHAPITRE 4

SÉPARATION ET CONTRIBUTIONS DES SOURCES DE VARIABILITÉ DE LA QUALITÉ DE L'AIR INTÉRIEUR

"Les sciences n'essaient pas d'expliquer ; c'est tout juste si elles tentent d'interpréter ; elles font essentiellement des modèles. Par modèle, on entend une construction mathématique qui, à l'aide de certaines interprétations verbales, décrit les phénomènes observés. La justification d'une telle construction mathématique réside uniquement et précisément dans le fait qu'elle est censée fonctionner". JOHN VON NEUMANN.

DANS *un environnement intérieur réel, les informations sur les sources d'émission de polluants sont transmises à travers leurs mélange de type MIMO (pour Multiple Input Multiple Output) et sont reçus par des capteurs. Les signaux constitués à partir de ce mélange servent de point de départ pour la reconstitution des séries initiales. Depuis deux décades, ce problème a suscité un grand intérêt et a engendré de nombreuses contributions par les techniques de séparation des sources.*

L'objectif de ce chapitre est de montrer en quoi certains outils statistiques de séparation des sources les plus récents peuvent apporter des éclairages nouveaux sur la compréhension des fluctuations des sources de pollution dans un environnement intérieur réel.

Sommaire

4.1 Introduction

Les méthodes développées pour résoudre le problème de séparation des sources à partir d'un signal intégrateur ont fait couler beaucoup d'encre. Il est intéressant de noter que le problème de séparation des sources à partir d'un mélange n'est pas un problème propre à l'environnement, on le retrouve aussi dans d'autres domaines. L'exemple classique donné en traitement du signal est celui de la reconnaissance des voix de plusieurs personnes qui parlent en même temps, à partir des enregistrements de cocktails de sons réalisés par plusieurs microphones. On peut employer un formalisme mathématique équivalent dans les deux domaines. Bien que le problème de séparation de sources soit formalisé mathématiquement de la même manière que ça soit en traitement du signal ou en environnement, les hypothèses assumées ou les contraintes imposées ont donné naissance à plusieurs méthodes de résolution. Dans le domaine de la qualité de l'air, le but est de mettre en évidence les "signatures" des différentes sources, permettant ainsi leur identification. En effet, en faisant l'hypothèse que les éléments émis par une source doivent se retrouver dans l'environnement de celle-ci groupés (statistiquement corrélés), l'utilisation des méthodes de classification et de reconnaissance de formes demeure très appropriée par leur capacité à mettre en évidence les groupes correspondant aux plus fortes corrélations (interprétés ensuite en termes de "signatures" des sources) (Ionescu, 2010).

La Séparation Aveugle des Sources (SAS ou BSS pour *Blind Source Separation*) est un terme générique regroupant les problèmes qui consistent à restaurer un ensemble de séries temporelles sources non observées à partir d'observations sur les mélanges de ces sources. L'adjectif *"aveugle"* dans l'expression de SAS renvoie au fait qu'on ne dispose pas (ou très peu) d'informations au sujet du mélange de sources. En statistique, il s'agit de méthodes d'apprentissage non-supervisées, c'est-à-dire sans connaissance préalable sur les mélanges. Le terme "aveugle", est imposé dans la littérature de télécommunications, est maintenant universellement utilisé. Or, d'un point de vue méthodologique, le fait de remonter aux sources et à l'estimation de leurs contributions au mélange est au contraire "*extralucide*" (voir la préface de Comon & Jutten (2007)).

Les premiers travaux en séparation aveugle des sources datent des années 80 et traitaient à l'origine le traitement des signaux physiologiques, plus exactement dans décodage du mouvement des vertébrés (Roll, 1981). Dans Comon & Jutten (2010) [Chap-1] les auteurs décrivent le problème biologique qui a initié les travaux sur la séparation de sources. Celui-ci consistait à étudier les réponses musculaires

émises à l'issue de différentes sortes d'excitations. Dès lors, la résolution des problèmes de séparation des sources s'est répandu dans d'autres champs disciplinaires et à suscité l'intérêt de la communauté scientifique. Et ce afin de répondre, en utilisant les algorithmes de SAS, aux questionnements dans les applications très variées, traitant des signaux de natures différentes.

Le point de vue traité dans ce chapitre est donc celui des méthodes inverses : caractérisation de la variabilité temporelle des sources à partir des observations réelles. Pour ce faire, on fait appel aux méthodes appliquées de manière générale en traitement du signal pour la séparation des sources à la problématique de la qualité de l'air intérieur, notamment l'Analyse en Composantes Indépendantes (ACI) (Comon & Jutten, 2007, 2010; Hyvärinen et al., 2004) et les méthodes de factorisation en matrices non-négatives (NMF) (Lee & Seung, 1999; Cichocki & Amari, 2002; Cichocki et al., 2009), factorisation matricielle positive (PMF) (Paatero et al., 1991; Paatero & Tapper, 1994), ainsi que l'Analyse en composantes principales (Wold et al., 1987; Saporta, 1990; Jolliffe, 2002).

4.2 Séparation des sources de pollution : survol de la littérature

Le problème de séparation des sources dans l'air intérieur n'a été abordé que récemment. En effet, un nombre important d'études se focalisent sur la recherche d'une source particulière extérieur, en cherchant à analyser la variabilité du rapport des concentrations intérieur-extérieur. Certaines études s'arrêtent à ce niveau, d'autres cherchent plus de détails, en séparant les sources (et leurs contributions) au niveau de l'environnement intérieur. Plus généralement, il s'agit des ambiances de type résidentiel, écoles ou bien des environnements mobiles, comme les intérieurs des voitures ou des bus.

Il existe un large spectre de méthodes qui abordent les questions d'identification et de contributions des sources de pollution. La Figure 4.2.1 illustre ce point en mettant en perspective le niveau d'information requis pour la résolution des problèmes d'identification des sources. Clairement, le modèle de bilan massique (CMB, pour Chemical Mass Balance) requiert une "parfaite" connaissance *a priori* du type de sources influençant le site de mesure (et de leur profil chimique) et cherche plutôt leurs contributions. Par contre, les méthodes dites "statistiques" comme la factorisation matricielle positive (PMF pour *Positive Matrix Factorization*) ou l'ACP (pour l'Analyse en Composantes Principales) s'appuient sur la détermination de divers paramètres physico-chimiques permettant l'identification de sources *a posteriori,* à partir de leurs "signatures".

Plusieurs études sur l'identification et la contribution des sources de pollution particulaire extérieure ont été rapportées dans la littérature. Chao & Cheng (2002) ont utilisé la CMB afin de mettre en évidence l'importance des sources intérieures liées à l'activité des occupants. La contribution de la source cuisson serait responsable de 61.9% des concentrations de $PM_{2.5}$. Pour une revue de littérature sur l'identification des sources par CMB, on peut consulter (Chow & Watson, 2002).

En ce qui concerne les applications des méthodes factorielles, l'ACP serait la méthode la plus utilisée de tous les articles qu'on a pu consulter. Beaucoup d'études l'ont utilisée pour la caractérisation des sources des COV.

Guo (2011) a étudié les sources de COV dans des maisons à Hong Kong. Il a appliqué l'ACP avec VARIMAX, ensuite il a déterminé les contributions par la technique ACPA (PCA/ACPS). Cette technique diffère de l'ACP classique par la contrainte de positivité imposée aux profils et aux contributions. Le

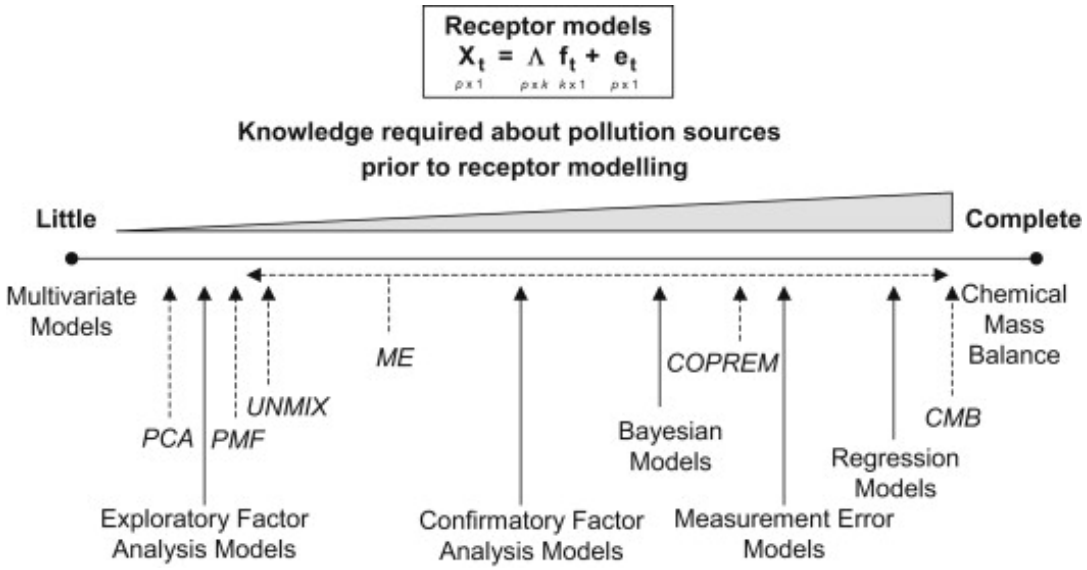


FIGURE 4.2.1 – *Niveau d'information requis par les modèles récepteur pour l'estimation des sources de pollution, d'après Viana et al. (2008).*

choix de leur méthode a été justifié par le fait que la technique ACPA demande un minimum d'entrées concernant les caractéristiques des sources, mais fournit des informations, à la fois sur les profils et les contributions. Ce modèle récepteur, comme tous les autres d'ailleurs, peut échouer dans la séparation des sources si elles sont fortement corrélées entre elles (colinéaires). Les auteurs ont analysé des échantillons pour 15 espèces de COV et formaldéhyde. Ils ont conclu que les principales sources étaient les produits d'entretien pour les maisons, le bois peint, désodorisant de chambre, les matériaux de construction et les boules de naphtaline.

En ce qui concerne les concentrations des particules, dans (Martuzevicius et al., 2008), les auteurs ont étudié les sources de $PM_{2.5}$ dans des résidences situées à côté (30-300 m) d'importantes voies expresses de circulation, en regardant plus spécifiquement l'influence du trafic, ainsi que la relation entre les niveaux extérieur et intérieur de particules. Afin de déterminer la quantité de particules provenant du trafic à l'intérieur des résidences, les auteurs ont appliqué un modèle multilinéaire de factorisation matricielle positive, en particulier, trilinéaire, appelé PARAFAC. La base de données a été construite en faisant des analyses chimiques des échantillons (EC, OC, Si, S, Mn, Fe, Zn, Br, Pb). La conclusion des auteurs a été que les sources intérieures contribuaient plus au niveau total de $PM_{2.5}$ que l'air extérieur, même dans des conditions de proximité au trafic routier. Le modèle PARAFAC a été utilisé également par Yakovleva et al. (1999), Hopke et al. (2003), et Larson et al. (2004).

Un modèle étendu de la PMF a été testé par Zhao et al. (2007) afin d'étudier l'exposition des enfants asthmatiques dans des écoles. Les auteurs ont cherché les sources communes de $PM_{2.5}$ pour trois types d'environnements : personnel, intérieur (de l'école) et extérieur (de l'école) à Denver. Ils ont trouvé quatre sources extérieures et trois sources intérieures pour les trois types d'environnement. La cuisson s'est avéré être la plus importante source intérieure. L'influence du tabac dans les maisons concernées était significative en ce qui concerne l'exposition des personnes aux particules. L'influence du trafic important à l'extérieur de l'école en est ressortie. Les échantillons ont été collectés pendant deux périodes d'hiver sur des filtres en Teflon qui ont été pesés et analysés par XRF pour déterminer les concentrations élémentaire à partir de Na jusqu'au Pb. Une caractéristique importante de cette étude est le fait qu'un agenda détaillé des activités des personnes a été relevé et que les concentrations de cotinine ont été mesurées. Les auteurs

ont utilisé un modèle étendu par rapport à la PMF, avec des matrices 4-dimensionnelles d'élément représentant la concentration d'une espèce dans un échantillon d'un certain type (sur une personne, à l'intérieur, à l'extérieur) collecté sur un sujet donné à une certaine date. Le journal d'activités des personnes a constitué une information précieuse pour l'identification des sources obtenues par la PMF.

Une autre étude menée par Molnár et al. (2014) utilise la PMF pour l'identification des sources de $PM_{2.5}$, la contribution majoritaire étant les sources extérieures avec 69% de l'ensemble des sources étudiées. L'activité des occupants participe à augmenter l'exposition aux $PM_{2.5}$ de 21% (2.2 $\mu g/m^3$). Notons que la contribution des sources extérieures dépend principalement la localisation géographique du site d'exposition (urbain, rural ...) et de la variation saisonnière.

4.3 Position du problème de séparation des sources pour la QAI

4.3.1 Cadre général

Le modèle de mélange décrivant les transformations liant les sources aux observations peut avoir plusieurs configurations

- linéaire ou non linéaire ;
- convolutif ou instantané ;
- variant ou invariant dans le temps.

Le cadre de modélisation le plus utilisé dans la littérature est celui des modèles linéaires et instantanés. Pour ce dernier, on suppose qu'à chaque instant t, les m observations $\{x_{i,t}\}_{i=1,\dots,m}$ sont des mélanges linéaires instantanés des p sources $\{s_{j,t}\}_{j=1,\dots,p}$:

$$x_{i,t} = \sum_{j=1}^{p} a_{ij} s_{j,t} + e_{i,t}, \tag{4.3.1}$$

où $a_{ij} \in \mathbb{R}$ pour $i = 1, \dots, m$ et $j = 1, \dots, p$ sont les coefficients de mélange. Ayant T observations pour $t = 1, \dots, T$ et en utilisant une notation matricielle, l'expression 4.3.1 s'écrit

$$\boldsymbol{X} = \boldsymbol{A}\boldsymbol{S} + \boldsymbol{E}, \tag{4.3.2}$$

où $\boldsymbol{X} \in \mathbb{R}^{m \times T}$ est la matrice des observations, $\boldsymbol{A} \in \mathbb{R}^{m \times p}$ est la matrice de mélange, $\boldsymbol{S} \in \mathbb{R}^{p \times T}$ est la matrice des sources et $\boldsymbol{E} \in \mathbb{R}^{m \times T}$ est la matrice des perturbations aléatoires additives. La matrice $\boldsymbol{E}$ prend en compte uniquement les *incertitudes liées à la modélisation*. Des méthodes de factorisation, notamment la factorisation en matrices positives, nécessite en entrée, non seulement les observation $\boldsymbol{X}$ mais aussi une *matrice d'incertitudes de mesure* ; à ne pas confondre avec la matrice inconnue $\boldsymbol{E}$.

Posé comme tel, deux grandes familles tentent de résoudre ce problème selon des hypothèses qui leurs sont propres. Ces hypothèses peuvent être formulées soit d'un point de vue statistiques ou soit d'un point de vue "algébrique". Le problème de séparation des sources se présente donc comme un problème d'optimisation sous contrainte dont l'objectif final est :

1. l'identification des sources du mélange et de leurs contributions ;
2. la reconstruction des sources.

4.3.2 Quelle problématique pour les séries temporelles de QAI ?

L'application des méthodes d'identification-contribution des sources en sciences environnementales apparaît souvent dans une perspective de reconnaissance des sources à partir des **signatures chimiques** (association des éléments chimiques propre à la source). Souvent, ces connaissances existent et sont requises pour l'interprétation des résultats. Parfois, elles ne sont pas uniques, car on retrouve des signature "partielles" qui peuvent correspondre à plusieurs sources. C'est la bonne connaissance du site de l'étude qui peut aider à leur identification. La méthode la plus utilisée en environnement est la factorisation matricielle positive (PMF) et elle a été conçue par Paatero (1994) spécialement pour la séparation des sources en environnement. Cette méthode part du principe que la concentration d'une espèce chimique j dans un échantillon i est égale au produit de la contribution d'une source k dans cet échantillon et la concentration de l'espèce chimique dans cette même source k. Formellement, le problème se ramène mathématiquement à la formulation suivante :

$$\boldsymbol{X} = \boldsymbol{GF} + \boldsymbol{E}. \tag{4.3.3}$$

La matrice $\boldsymbol{X} \in \mathbb{R}^{n\times m}$ représente des données enregistrées pour m espèces (les colonnes de $\boldsymbol{X}$) pour les n échantillons (les lignes de $\boldsymbol{X}$). Les éléments de la matrice $\boldsymbol{G}$ représentent les contributions et $\boldsymbol{F}$ la matrice des profils.

En revanche, dans l'analyse des séries temporelles, ce problème d'identification-contribution des sources se présente avec une différence notable : les connaissances relatives à la variabilité temporelle des sources, donc aux **signatures temporelles**, demeurent inexistantes. La matrice $\boldsymbol{X}$ dans 4.3.3 représente les observations enregistrées par m capteurs aux T moments différents : $\boldsymbol{X} \in \mathbb{R}^{m\times T}$. Une colonne de $\boldsymbol{X}$ représente les enregistrements des capteurs à un moment donné. Ces enregistrements proviennent de n sources. Les intensités des émissions des sources à un moment donné t se trouvent dans la colonne t d'une matrice $\boldsymbol{Y} \in \mathbb{R}^{n\times T}$.

Par conséquent, l'application de ces techniques sur les séries chronologiques de polluants intérieurs nécessitent de poser quelques hypothèses supplémentaires sur la nature des sources et leurs mélanges : linéaire ou non linéaire, convolutif ou instantané et variant ou invariant dans le temps.

Nous gardons à l'esprit ces considérations afin de nuancer, lors de l'interprétation des résultats, la nature de la variabilité des sources de pollution. On peut obtenir un groupement de sources, au lieu d'une source spécifique. Plus précisément, l'interprétation qu'on doit accorder à un profil temporel d'une série issue d'une factorisation quelconque doit être compris comme la représentation des fluctuations temporelles d'un "*cluster*" de sources. Cette distinction tient au fait que paramètres climatiques, l'occupation et les activités des occupants (notamment l'ouverture) agissent simultanément sur la concentration des polluants et l'intensité de leurs sources.

Prenant par exemple à la Figure 1.3.3 qui représente les processus affectant les concentrations de particules. L'importance des phénomènes de dépôt (puits) et la remise en suspension des particules dépend non-seulement de leur taille, mais aussi de l'importance des facteurs d'usage liée au bâtiment et les paramètres climatiques.

Les profils temporels estimés par une méthode de factorisation mettent en évidence plus l'interaction de facteurs-sources que seulement les sources, prises individuellement.

4.4 L'Analyse en Composantes Indépendantes (ACI)

4.4.1 Cadre général

Cette classe des méthode pour la résolution des problèmes BSS fait apparaître une hypothèse très importante : l'indépendance statistique des signaux sources. La définition rigoureuse, donnée dans (Jutten & Herault, 1991; Comon, 1994), part du principe général des modèles statistiques à variables latentes. Le terme "modèle statistique latent" suggère que la composante n'est pas directement observable en plus d'autres hypothèses liées à l'exogénéité du bruit additif et la nature probabiliste des sources. La formule 4.3.1 illustre le fait que les sources et le leurs mélange correspondent aux composantes latentes du modèle statistique.

Pour simplifier, considérons l'écriture matricielle (4.3.2) du problème BSS sans la partie résiduelle et sous forme vectorielle :

$$\boldsymbol{x} = \boldsymbol{A}\boldsymbol{s}, \tag{4.4.1}$$

où $\boldsymbol{x}$ est un vecteur aléatoire dont les éléments $x_1, \ldots, x_n$ sont des mélanges, $\boldsymbol{s}$ le vecteur aléatoire d'élément $s_1, \ldots, s_n$ et $\boldsymbol{A}$ la matrice de mélange d'élément a_{ij}. Le point de départ de l'ACI est la simple hypothèse statistique d'indépendance des s_i qui permet de choisir une factorisation de $\boldsymbol{x}$ (à un facteur près). Une autre hypothèse s'ajoute à l'indépendance, elle concerne la non-gaussianité des composantes indépendantes.

4.4.2 Hypothèses

Pour les problèmes de séparation aveugle des sources, dans le cadre du mélange instantané, nous supposons que :

H1. Les signaux sources sont indépendants, identiquement distribués et mutuellement indépendants.

H2. Le nombre de capteurs est supérieur ou égal au nombre de sources.

H3. La matrice de mélange $\boldsymbol{A}$ est de rang complet.

H4. La matrice résiduelle est indépendante des sources et sa structure est additive.

H5. Au plus un seul signal source a une distribution gaussienne.

4.4.3 La séparation

Le principe générale de la séparation aveugle des sources dans le cadre de l'ACI consiste à trouver une transformation $\boldsymbol{W}$ qui permet d'obtenir, à partir des observations, des signaux sources mutuellement indépendants :

$$\hat{\boldsymbol{s}} = \boldsymbol{W}^\top \boldsymbol{x}. \tag{4.4.2}$$

Le problème se réduit alors en la détermination d'une transformation linéaire, avec $\boldsymbol{W} \in \mathbb{R}^{p \times n}$ de telle manière que les composantes de $\hat{\boldsymbol{s}} \in \mathbb{R}^n$ soient aussi indépendantes que possible par maximisation d'une fonction mesurant l'indépendance statistique (Hyvarinen, 1999; Hyvärinen et al., 2004). Cette matrice permet de récupérer les signaux sources à un facteur d'échelle et une permutation près.

La matrice $\boldsymbol{W}$ se dit matrice de séparation, si et seulement si elle satisfait (en l'absence de bruit), l'égalité suivante :

$$\boldsymbol{W}^{\top}\boldsymbol{x} = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{s}, \tag{4.4.3}$$

ou encore :

$$\boldsymbol{W}^{\top}\boldsymbol{A} = \boldsymbol{P}\boldsymbol{\Lambda}, \tag{4.4.4}$$

où $\boldsymbol{P}$ est une matrice de permutation et $\boldsymbol{\Lambda}$ est est une matrice diagonale non singulière.

4.4.4 Indépendance statistique et non-Gaussianité

Avant de présenter l'ACI, rappelons le concept d'indépendance statistique. Des variables aléatoires $\{x_j\}_{j=1}^{p}$ sont dites statistiquement mutuellement indépendantes si et seulement si :

$$\mathscr{P}(x_1, \ldots, x_p) = \prod_{j=1}^{p} \mathscr{P}(x_j). \tag{4.4.5}$$

Par conséquent si deux variables aléatoires x_1 et x_2 sont statistiquement indépendantes, alors pour toute fonction f et g, on a

$$\mathbb{E}[f(x_1)\, g(x_2)] = \mathbb{E}[f(x_1)]\, \mathbb{E}[g(x_2)]. \tag{4.4.6}$$

4.4.4.1 L'information mutuelle

L'indépendance peut être quantifiée par la quantité d'information mutuelle entre les vecteurs aléatoires. Soient $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$, p vecteurs aléatoires de densité conjointe $\mathscr{P}_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_p}$ et de densités marginales $\mathscr{P}_{\boldsymbol{x}_1}, \ldots, \mathscr{P}_{\boldsymbol{x}_p}$; l'information mutuelle entre ces vecteurs est définie comme la divergence de Kullback-Leibler (entropie relative) entre les densité de probabilité $\prod_{k=1}^{p} \mathscr{P}_{\boldsymbol{x}_k}$ et $\mathscr{P}_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_p}$:

$$\begin{aligned} \mathcal{I}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p) &= -\mathbb{E}\left[\log \frac{\mathscr{P}_{\boldsymbol{x}_1}(\boldsymbol{x}_1),\ldots,\mathscr{P}_{\boldsymbol{x}_p}(\boldsymbol{x}_p)}{\mathscr{P}_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_p}(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_p)}\right], \\ &= \sum_{i=1}^{p} H(\boldsymbol{x}_k) - H(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p) \end{aligned} \tag{4.4.7}$$

où $H(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)$ et $H(\boldsymbol{x}_1), \ldots, H(\boldsymbol{x}_p)$ sont l'entropie conjointe et les entropies différentielles marginales de $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$:

$$H(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p) = -\mathbb{E}\left[\log \mathscr{P}_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_p}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)\right] \tag{4.4.8}$$

et $H(\boldsymbol{x}_k)$ est définie de façon analogue avec $\mathscr{P}_{\boldsymbol{x}_k}$ à la place de $\mathscr{P}_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_p}$. Considérons que les variables $\{x_j\}_{j=1}^{p}$ sont les estimées des sources obtenues par l'application d'une matrice, notée $\boldsymbol{B}$, *i.e.* $\boldsymbol{x} = \boldsymbol{B}\boldsymbol{s}$. Pour chercher l'information mutuelle de ces variables, on aura besoin du lemme suivant :

Lemme 4.4.1. *Soit $\boldsymbol{y}$ un vecteur aléatoire et $\boldsymbol{x} = g(\boldsymbol{y})$ où g est une transformation inversible dérivable de Jacobien (matrice des dérivées) g'. Alors :*

$$H(\boldsymbol{x}) = H(\boldsymbol{y}) + \mathbb{E}\left[\left|\log\left(g'(\boldsymbol{x})\right)\right|\right]. \tag{4.4.9}$$

Par conséquent, l'information mutuelle des variables aléatoires $\{x_j\}_{j=1}^{p}$ sera exprimée par

$$\mathcal{I}(\boldsymbol{x}) = \sum_{j=1}^{p} H(x_j) - H(\boldsymbol{y}) - \log\det \boldsymbol{B}. \tag{4.4.10}$$

4.4.4.2 Kurtosis

Les statistiques d'ordre supérieur sont des outils pratiques pour évaluer l'indépendance statistique des variables aléatoires non-gaussiennes ; le kurtosis, comme le quatrième cumulant, est une mesure classique. Les v.a. gaussiennes ont le kurtosis, qui mesure entre autres niveau d'aplatissement, égal à zéro. L'ACI cherche des transformations des données $\boldsymbol{X}$ telles que les données $\boldsymbol{Y}$ obtenues conduisent à un kurtosis maximum positif ou minimum négatif. Le kurtosis est défini comme

$$\text{kurt}(x) = \mathbb{E}\left[x^4\right] - 3\left(\mathbb{E}\left[x^2\right]\right)^2. \tag{4.4.11}$$

La première version de l'algorithme FastICA, proposée par Hyvärinen & Oja (1997), utilise la valeur absolue du kurtosis comme une mesure de non-gaussianité, mais, cette dernière étant insuffisamment robuste, a été vite remplacée par la néguentropie, présentée ci-après.

4.4.4.3 Néguentropie

La néguentropie est une mesure très importante de la non-gaussianité ; son principe est basé sur les résultats de la théorie de l'information par la mesure de l'entropie différentielle (Cover & Thomas, 2012). L'entropie différentielle, que nous avons vue dans le cadre de la mesure de prédictibilité dans le chapitre 3 constitue donc un outil essentiel dans le traitement. Un résultat fondamental en théorie de l'information stipule que la variable aléatoire Gaussienne possède la plus grande entropie parmi toutes les autres variables de variance égale.

Si l'on considère un vecteur aléatoire $\boldsymbol{x} = \{x_1, \ldots, x_p\}$ de densité de probabilité Gaussienne de covariance $\boldsymbol{\Sigma}$, alors

$$H(\boldsymbol{x}) = \frac{1}{2}\left(p\left(1 + \log 2\pi\right) + \log\det \boldsymbol{\Sigma}\right). \tag{4.4.12}$$

La néguentropie est définie comme étant une mesure de l'éloignement entre la distribution d'une variable aléatoire et la densité gaussienne. Pour un vecteur aléatoire Gaussien $\boldsymbol{y}_{\text{Gauss}}$ et un autre $\boldsymbol{y}$ quelconque de même covariance, la néguentropie est définie par

$$\mathcal{J}(\boldsymbol{y}) = H\left(\boldsymbol{y}_{\text{Gauss}}\right) - H(\boldsymbol{y}). \tag{4.4.13}$$

La néguentropie, telle que définie ci-dessus, est toujours non-négative et nulle si et seulement si $\boldsymbol{y}$ est Gaussienne. D'autres propriétés intéressantes, comme l'invariance par translation, peuvent être consultées dans (Comon, 1994; Hyvarinen, 1999). L'estimation de cette mesure est très documentée dans la littérature, par exemple, en utilisant les moments d'ordres supérieurs, Jones & Sibson (1987) donnent l'approximation suivante :

$$\mathcal{J}(y) \approx \frac{1}{12}\mathbb{E}\left[y^3\right]^2 + \frac{1}{48}\left[\text{kurt}\,(y)\right]^2. \tag{4.4.14}$$

Afin d'obtenir des estimateurs plus robustes, Hyvärinen & Oja (1998) utilisent des fonctions non-quadratiques G ; la néguentropie est alors approchée par :

$$\mathcal{J}(y) \propto \left\{\mathbb{E}\left[G(y)\right] - \mathbb{E}\left[G(\nu)\right]\right\}^2, \tag{4.4.15}$$

où G est choisie entre ces deux fonctions :

$$G_1(u)) = \frac{1}{\alpha_1}\log\cosh(\alpha_1 u), \qquad G_2(u)) = -\exp\left(-\frac{u^2}{2}\right) \tag{4.4.16}$$

avec $1 \leq \alpha_1 \leq 2$ est une constante.

4.4.5 L'algorithme FastICA (Hyvarinen, 1999)

Comme dans beaucoup d'algorithmes, les étapes de pré-traitement des données, en l'occurrence le "centrage" et la décorrélation sont nécessaire pour respecter quelque hypothèses liées à la position du problème.

On suppose dans la suite que $\mathbb{E}\left[\left(\boldsymbol{w}^\top \boldsymbol{z}\right)^2\right] = \|\boldsymbol{w}\|^2 = 1$; le gradient de l'estimation de la néguentropie dans (4.4.15) par rapport à $\boldsymbol{w}$ de la fonction

$$J(\boldsymbol{w}) = \mathbb{E}\left[G\left(\boldsymbol{w}^\top \boldsymbol{z}\right)\right] \tag{4.4.17}$$

est

$$\frac{\partial J(\boldsymbol{w})}{\partial \boldsymbol{w}} = \mathbb{E}\left[\boldsymbol{z}G\left(\boldsymbol{w}^\top \boldsymbol{z}\right)\right]. \tag{4.4.18}$$

Le procédé d'optimisation, par exemple la recherche du maximum de $J(\boldsymbol{w})$, peut s'effectuer par les multiplicateurs de Lagrange. Dans ce cas, on arrive à résoudre le système :

$$\begin{cases} \mathbb{E}\left[\boldsymbol{z}G\left(\boldsymbol{w}^\top \boldsymbol{z}\right)\right] + \gamma \boldsymbol{w} & = 0 \\ \|\boldsymbol{w}\| & = 1 \end{cases}. \tag{4.4.19}$$

Il en résulte que $\gamma = -\mathbb{E}\left[\boldsymbol{w}^\top \boldsymbol{z}G\left(\boldsymbol{w}^\top \boldsymbol{z}\right)\right]$.

Sous l'hypothèse $\mathbb{E}\left[\boldsymbol{z}\boldsymbol{z}^\top G'\left(\boldsymbol{w}^\top \boldsymbol{z}\right)\right] \approx \mathbb{E}\left[\boldsymbol{z}\boldsymbol{z}^\top\right]\mathbb{E}\left[G'\left(w^\top z\right)\right] = \mathbb{E}\left[G'\left(w^\top z\right)\right]$, la résolution de (4.4.19) par la méthode itérative de Newton conduit à l'algorithme (Hyvärinen et al. (2004), page 189) :

$$\begin{cases} \boldsymbol{w} \leftarrow \boldsymbol{w} - \frac{\mathbb{E}\left[\boldsymbol{z}G'\left(\boldsymbol{w}^\top \boldsymbol{z}\right)\right]+\gamma \boldsymbol{w}}{\mathbb{E}\left[G'\left(w^\top z\right)\right]+\gamma} \\ \boldsymbol{w} \leftarrow \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|} \end{cases} . \tag{4.4.20}$$

En multipliant la première formule dans (4.4.20) par $\left(\mathbb{E}\left[G'\left(w^\top z\right)\right]+\gamma\right)$ et quelques simplifications, l'algorithme devient

$$\begin{cases} \boldsymbol{w} \leftarrow \left\{\mathbb{E}\left[\boldsymbol{z}G\left(\boldsymbol{w}^\top \boldsymbol{z}\right)\right]-\mathbb{E}\left[G'\left(\boldsymbol{w}^\top \boldsymbol{z}\right)\right]\right\} \boldsymbol{w} \\ \boldsymbol{w} \leftarrow \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|} \end{cases} \tag{4.4.21}$$

L'algorithme FastICA a les propriétés, selon Hyvarinen (1999), suivantes :
- La convergence est cubique (ou au moins du second degré)
- L'algorithme trouve les composantes directement indépendantes de (presque) toute distribution non-gaussienne en utilisant n'importe quelle non-linéarité G.
- La performance de la méthode peut être optimisée en choisissant une non-linéarité G convenable.

4.5 Factorisation Matricielle Positive (PMF)

La résolution du problème de SAS par l'ACI est basée sur l'hypothèse d'indépendance des sources au sens statistique de terme : les sources sont modélisées comme des processus stochastiques indépendants les uns des autres (Comon & Jutten, 2007). En revanche, la contrainte statistique ne résout pas les problèmes "d'interprétabilité" des sources à contributions négatives. C'est pour cette raison que d'autre méthodes factorielles ont été proposées dans la littérature pour palier à ce problème, en imposant dans le processus de décomposition la contrainte de non-négativité.

Durant les années 90, la méthode de Factorisation Matricielle Positive (PMF), développée par Paatero (1991; 1994; 1997a; 1997b), a constitué un tournant très important pour les problème séparation des sources de pollution. La PMF est devenue une technique très connue dans le domaine de l'environnement, car son pouvoir d'interprétation dans le cadre des modèles récepteurs est très grand (Hopke, 1991, 2010). En outre, cette méthode n'exige que très peu d'informations *a priori*.

On souhaite par cette méthode, et comme son nom l'indique, avoir une factorisation d'une matrice positive en un produit de matrices positives. Nous verrons dans la section suivante une différence notable avec la méthode NMF (Non-Negative Matrix Factorisation), qui impose la même contrainte de non-négativité des sources, mais qui résout le problème autrement.

Cette contrainte "de signe", exigée pour la résolution des problèmes BSS facilite l'interprétation des résultats de décomposition. En effet, le principe de conservation de la masse stipule que la concentration d'un composé j mesuré au temps t, représente une la combinaison linéaire des sources.

Le problème de base est donné dans la formule 4.3.3, où le bilan de masse peut s'exprimer sous une forme linéaire élémentaire de type :

$$x_{tj} = \sum_{k=1}^{p} g_{tk} f_{kj} + e_{tj} = c_{tj} + e_{tj}, \tag{4.5.1}$$

où x_{tj} est la concentration du polluant j à l'instant t, les éléments g_{tk} représentent les contributions de la source k à l'instant t et f_{kj} représentes les profiles des émissions des p sources. Les éléments de la matrice des résidus $\boldsymbol{E}$ (*i.e.* e_{tj}) correspondent à la partie inexpliquée par le modèle. Formellement, p sources de pollution émettent m polluants, la quantité émise se retrouve ensuite dans l'environnement des sources où sont effectuées des mesures sur une durée de temps T, pour les m polluants.

La résolution du 4.3.3 (ou de 4.5.1) permettant d'estimer $\boldsymbol{F}$ et $\boldsymbol{G}$ consiste à résoudre un problème d'optimisation sous contrainte avec une fonction objectif Q :

$$\underset{\boldsymbol{G},\boldsymbol{F}\geq 0}{\arg\min} Q\left(\boldsymbol{X},\boldsymbol{G},\boldsymbol{F}\right). \tag{4.5.2}$$

Ce problème d'optimisation consiste à minimiser la quantité de la forme :

$$Q\left(\boldsymbol{E}\right)=\sum_{t=1}^{T}\sum_{j=1}^{m}\left(\frac{e_{tj}}{\varsigma_{tj}}\right)^2=\sum_{t=1}^{T}\sum_{j=1}^{m}\left(\frac{x_{tj}-\sum_{k=1}^{p}g_{tk}f_{kj}}{\varsigma_{tj}}\right)^2, \tag{4.5.3}$$

où ς_{tj} représente l'incertitude de mesure pour chaque élément x_{tj} de la matrice d'observation $\boldsymbol{X}$ à chaque pas de temps t. Encore une fois, une hypothèse supplémentaire sur les incertitudes doit être fixée dans le cadre des séries temporelles. Elle concerne principalement la question de savoir si les erreurs faites lors de la mesures, donc des incertitudes associées, sont indépendantes du temps ou pas. Notre hypothèse dans ce cadre est la suivante, les séries d'incertitudes sont constantes dans le temps. Cette hypothèse facilite largement le traitement mathématique de la résolution du problème PMF. À notre connaissance, la question de stationnarité de la série d'incertitude n'a pas fait l'objet de grande discussion dans la communauté scientifique, car elle relève de la question du processus stochastique qui génère les incertitudes.

De nombreuses variantes algorithmiques ou sur la forme des pénalisations ont été publiées pour la résolution de problème de la PMF. Pour la factorisation de cette dernière, trois familles d'algorithmes sont souvent citées :

1. Alternate Least Square (ALS) algorithm (Paatero & Tapper, 1994; Paatero, 1997a) ;
2. Two-way PMF (Paatero, 1997a, 2000) ;
3. Multilinear Engine (ME) (Paatero, 1997a).

La résolution du problème de la PMF avec le premier algorithme consiste à utiliser la méthode des moindres carrés alternés "Alternating Least Square" de (Paatero & Tapper, 1994; Paatero, 1997a). Dans cette procédure, on fixe une matrice et on estime la deuxième en minimisant la fonction objectif Q (avec une pénalisation quadratique) sous contrainte de non-négativité. Au pas suivant, le rôle des matrices $\boldsymbol{G}$ et $\boldsymbol{F}$ est interchangé ; ainsi la matrice estimée à l'étape antérieure sera fixée dans ce qui suit, et l'autre sera estimée avec la formule 4.5.3.

En faisant varier les deux matrices $\boldsymbol{G}$ et $\boldsymbol{F}$ à chaque pas, d'où la dénomination PMF2 (Paatero, 1997a,b), l'algorithme two-way PMF permet, à l'aide d'une fonction de pénalisation, de réduire encore plus le degré de liberté rotative de la solution. Autrement dit, les fonctions de pénalisation permettent de respecter la contrainte de non-négativité en évitant la boucle de fixation et d'estimation des matrices $\boldsymbol{G}$ et $\boldsymbol{F}$. La fonction objectif Q à minimiser devient $\mathring{Q}$:

$$\mathring{Q}\left(\boldsymbol{E},\boldsymbol{G},\boldsymbol{F}\right)=Q\left(\boldsymbol{E}\right)+P\left(\boldsymbol{G}\right)+P\left(\boldsymbol{F}\right)+R\left(\boldsymbol{G}\right)+R\left(\boldsymbol{F}\right), \tag{4.5.4}$$

avec $P(\boldsymbol{G})$ et $P(\boldsymbol{F})$ deux fonctions de pénalisation pour forcer le respect de la contrainte de non-négativité, les termes $R(\boldsymbol{G})$ et $R(\boldsymbol{F})$ de régularisation ont pour rôle de supprimer les "singularités" (causées par les rotations et le changement d'échelle) du modèle. Ces fonctions sont définies par :

$$\begin{aligned} P(\boldsymbol{G}) &= -\alpha \sum_{t=1}^{T} \sum_{k=1}^{p} \log(g_{tk}), \\ P(\boldsymbol{F}) &= -\beta \sum_{k=1}^{p} \sum_{j=1}^{m} \log(f_{kj}), \\ R(\boldsymbol{G}) &= -\gamma \sum_{t=1}^{T} \sum_{k=1}^{p} (g_{tk})^2, \\ R(\boldsymbol{F}) &= -\delta \sum_{k=1}^{p} \sum_{j=1}^{m} (f_{kj})^2. \end{aligned} \tag{4.5.5}$$

Les coefficients α et β contrôlent l'intensité des pénalités et γ et δ celles de régularités. Les valeurs de ces coefficients sont très petits pendant les itérations pour que leurs valeurs finales soient négligeables, mais pas nulles. En pratique, les fonctions logarithmiques sont rapprochées par un développement en série de Taylor jusqu'à des termes quadratiques (Khlaifi, 2007). Les valeurs négatives sont passagères, car elles sont immédiatement réajustées par le développement en série ; seulement des valeurs non-négatives sont stockées dans les matrices $\boldsymbol{G}$ et $\boldsymbol{F}$.

La dernière méthode, qui est "Multilinear Engine" pour la PMF, fut développé dans (Paatero, 1999) pour le traitement des problèmes bilinéaires. On trouve néanmoins, un développement beaucoup plus complet pour les problèmes de séparation des sources dans l'excellente monographie de Cichocki et al. (2009) et aussi dans (Cichocki & Amari, 2002).

4.6 Factorisation en Matrices Non-Négatives (NMF)

Le principe de la factorisation d'une matrice en plusieurs matrices est un problème très connu en algèbre linéaire, le traitement des matrices non-négatives ont un place particulière (voir l'ouvrage de Berman & Plemmons (1979)). Le principe mathématique a été emprunté et largement appliqué en statistique multidimensionnelle, l'analyse en composantes principales qui utilise la décomposition en valeurs singulières (SVD) en un exemple indéniable. L'ACP et d'autres méthodes factorielles voisines, construisent des facteurs orthogonaux deux à deux. La PMF de Paatero & Tapper (1994), par la suite la NMF de Lee & Seung (1999), sont construites sur une décomposition sans contrainte d'orthogonalité mais avec celle de non-négativité des matrices des facteurs. Cette nouvelle contrainte est posée afin d'en simplifier l'interprétation et sur la base d'une motivation "neuronale" : les neurones ne fonctionnent que de façon additive, pas soustractive.

C'est l'article de Lee & Seung, paru dans la revue *Nature* en 1999 (Lee & Seung, 1999) qui a créé l'enthousiasme autour de la NMF, qui prend alors son nom définitif. La méthode a connu dès lors des modifications au niveau algorithmique donnant lieu a plusieurs variantes ; ces variantes diffèrent souvent par le type de la fonction objectif, le problème de l'initialisation, de la pénalisation (L^1, L^2..) et le choix des algorithmes.

Compte tenu du fait que ces méthodes n'ont été utilisés que récemment en sciences atmosphérique Plouvin et al. (2014); Kfoury et al. (2014); Chreiky et al. (2015); Kfoury et al. (2016) et que pratiquement aucune étude en environnement intérieur, l'analyse exhaustive des variantes NMF n'est pas envisageable dans cette thèse. La description présente de la méthode NMF ne se veut pas exhaustive ; elle est axée uniquement sur quelques aspects nécessaires pour les applications empiriques. Pour lecture plus complète, on se réfère aux deux ouvrages de Cichocki et al. (2009) et (Cichocki & Amari, 2002).

La plupart des implémentations ont été réalisées à l'aide du package (en R (2015)) éponyme de Renaud Gaujoux et Cathal Seoighe (2010).

4.6.1 Le modèle

Soit une matrice $\boldsymbol{X} \in \mathbb{R}_+^{m\times T}$ non-négative des données enregistrées par m capteurs aux T moments différents. Ces enregistrements proviennent de n sources. Les intensités des émissions des sources à un moment donné t se trouvent dans la colonne t d'une matrice $\boldsymbol{Y} \in \mathbb{R}_+^{n\times T}$. La factorisation non-négative de la matrice $\boldsymbol{X}$ consiste en la recherche de deux matrices $\mathbb{R}_+^{m\times n} \ni \boldsymbol{A} = [\boldsymbol{a_1}, \boldsymbol{a_2}, \ldots, \boldsymbol{a_n}]$ et $\mathbb{R}_+^{n\times T} \ni \boldsymbol{Y} = \boldsymbol{B}^\top = [\boldsymbol{b_1}, \boldsymbol{b_2}, \ldots, \boldsymbol{b_n}]^\top$ ne contenant que des valeurs positives ou nulles et dont le produit approche $\boldsymbol{X}$:

$$\boldsymbol{X} \quad \approx \quad \boldsymbol{AY} \tag{4.6.1}$$

ou de manière plus générale [1] en introduisant la matrice résiduelle $\boldsymbol{E} \in \mathbb{R}^{m\times T}$:

$$\boldsymbol{X} = \boldsymbol{AY} + \boldsymbol{E} = \boldsymbol{AB}^\top + \boldsymbol{E}. \tag{4.6.2}$$

Le problème NMF est aussi représenté sous sa forme des modèles bilinéaires :

$$\boldsymbol{X} = \sum_{j=1}^{n} \boldsymbol{a}_j \circ \boldsymbol{b}_j + \boldsymbol{E} = \sum_{j=1}^{n} \boldsymbol{a}_j \boldsymbol{b}_j^\top + \boldsymbol{E}, \tag{4.6.3}$$

où le symbole $\circ$ désigne le produit extérieur des vecteurs (élément par élément). Le choix du rang de factorisation $r \ll \min\{m, T\}$ assure une réduction drastique de dimension et donc des représentations parcimonieuses. Par rapport aux notations précédentes, le nombre des sources p devient n.

La différence essentielle entre la PMF et la NMF provient du fait que la deuxième n'opère pas de pondération par rapport aux incertitudes de mesure, mais la factorisation contient toujours des matrices positives. La factorisation NMF du modèle 4.6.2 est résolue par la recherche d'un optimum local du problème d'optimisation :

$$\underset{\boldsymbol{A},\boldsymbol{Y}\geq 0}{\arg\min} \left[\mathcal{D}\left(\boldsymbol{AY}, \boldsymbol{X}\right) + R\left(\boldsymbol{A}, \boldsymbol{Y}\right)\right], \tag{4.6.4}$$

où

— $\mathcal{D}$ est une fonction perte mesurant la qualité d'approximation. Les plus utilisées sont celles basées soit sur un critère des moindres carrés, comme la norme de Frobenius des matrices,

$$\mathcal{D} \ : \ \mathcal{D}_F\left(\boldsymbol{A}, \boldsymbol{B}\right) = \mathrm{TR}\left[\left(\boldsymbol{A} - \boldsymbol{B}\right)\left(\boldsymbol{A} - \boldsymbol{B}\right)^\top\right], \tag{4.6.5}$$

ou la divergence de Kullback-Leibler :

$$\mathcal{D} \ : \ \mathcal{D}_{KL}\left(\boldsymbol{A} \parallel \boldsymbol{B}\right) = \sum_{i,j} a_{ij} \log\left(\frac{a_{ij}}{b_{ij}}\right) - a_{ij} + b_{ij}. \tag{4.6.6}$$

1. Puisque nous travaillons sur les vecteurs colonnes des matrices, il est plus pratique d'utiliser $\boldsymbol{Y} = \boldsymbol{B}^\top$ plutôt que $\boldsymbol{Y}$. Le symbole $\top$ renvoie à la transposée d'une matrice alors que T représente la taille des séries.

— R une fonction de pénalisation optionnelle de régularisation pour forcer les propriétés recherchées des matrices $\boldsymbol{A}$ et $\boldsymbol{Y}$ (Cichocki et al., 2009).

Notons que non seulement les solutions sont locales car la fonction objectif du problème d'optimisation n'est pas convexe en $\boldsymbol{A}$ et $\boldsymbol{Y}$, mais en plus les solutions ne sont pas uniques. Toute matrice $\boldsymbol{K}^{r\times r}$ non-négative et inversible fournit des solutions équivalentes en termes d'ajustement :

$$\boldsymbol{X} = \boldsymbol{A}\boldsymbol{K}\boldsymbol{K}^{-1}\boldsymbol{Y}. \tag{4.6.7}$$

4.6.2 Types de divergence utilisables

4.6.2.1 Les divergences de Csiszár

Ce type de divergence s'obtient à partir d'une fonction réelle, convexe, différentiable $f : (0,\infty) \to \mathbb{R}$ avec $f(1) = 0$ et $f'(1) = 0$. Pour deux vecteurs $\boldsymbol{u} = (u_1, u_2, \ldots, u_n)^\top$ et $\boldsymbol{v} = (v_1, v_2, \ldots, v_n)^\top$, la divergence est définie par :

$$\mathcal{D}_f(\boldsymbol{u},\boldsymbol{v}) = \sum_{i=1}^{n} v_i f\left(\frac{u_i}{v_i}\right) \tag{4.6.8}$$

La condition $f'(1) = 0$ est nécessaire pour que les conditions de divergence pour $\mathcal{D}_f$ soient remplies pour n'importe quels vecteurs $\boldsymbol{u},\boldsymbol{v}$ positifs. Dans le cas $f'(1) \neq 0$, il est nécessaire que $\sum_{i=1}^{n} u_i = \sum_{i=1}^{n} v_i = 1$ pour que $\mathcal{D}_f$ réponde aux demandes de divergence.

Des exemples de divergences à partir de la fonction de Csiszàr sont donnés dans le Tableau 4.6.1.

Table 4.6.1 – Exemples de divergences à partir de la fonction de Csiszár.

Fonction de Csiszár	**Formule de la divergence** $\mathcal{D}_f(\boldsymbol{u},\boldsymbol{v})$	**Nom de la divergence**
$f(t) = \lvert t-1\rvert$	$\sum_i \lvert u_i - v_i\rvert$	Variation totale
$f(t) = (t-1)^2$	$\sum_i \frac{(u_i-v_i)^2}{v_i}$	Pearson χ^2
$f(t) = \frac{(t-1)^2}{t}$	$\sum_i \frac{(u_i-v_i)^2}{u_i}$	Neyman
$t\ln(t) - t + 1$	$\sum_i u_i \ln\left(\frac{u_i}{v_i}\right) - u_i + v_i$	I ou Kullback-Leibler
$(t-1)\ln(t)$	$\sum_i (u_i - v_i)\ln\left(\frac{u_i}{v_i}\right)$	J
$\frac{t^\alpha - 1 - \alpha(t-1)}{\alpha(\alpha-1)}$ $\alpha \neq \{0,1\}$	$\frac{1}{\alpha(\alpha-1)} \sum \left(v_i \left(\left(\frac{u_i}{v_i}\right)^\alpha - 1\right) - \alpha(u_i - v_i)\right)$	α

4.6.2.2 Divergences de Bregman

Ces divergences s'obtiennent à partir d'une fonction strictement convexe $\phi : \mathbb{R}^n \to \mathbb{R}$ de classe $\mathcal{C}^1$, et pour des vecteurs $\boldsymbol{u},\boldsymbol{v} \in \mathbb{R}^n$, la divergence est définie par

$$\mathcal{D}_\phi(\boldsymbol{u},\boldsymbol{v}) = \phi(\boldsymbol{u}) - \phi(\boldsymbol{v}) - (\boldsymbol{u}-\boldsymbol{v})^\top \nabla\phi(\boldsymbol{v}) \tag{4.6.9}$$

Si nous avons une fonction strictement convexe $f : \mathbb{R} \to \mathbb{R}$, alors nous pouvons définir $\phi : \mathbb{R}^n \to \mathbb{R}$, $\phi(t_1, t_2, , \ldots, t_n) = \sum_{i=1}^{n} f(t_i)$ et la formule de la divergence de BREGMAN (formule 4.6.9) devient

$$\mathcal{D}_\phi(\boldsymbol{u}, \boldsymbol{v}) = \sum_i f(u_i) - f(v_i) - (u_i - v_i) f'(v_i) \tag{4.6.10}$$

Des exemples de divergences à partir de la fonction de BREGMAN sont donnés dans le Tableau 4.6.2.

TABLE 4.6.2 – Exemples de divergences de BREGMAN

Fonction convexe	**Formule de la divergence**	**Nom de la divergence**
$\frac{1}{2}\boldsymbol{t}^\top = \frac{1}{2}\sum t_i^2$	$\frac{1}{2}\sum_i (u_i - v_i)^2$	Le carré de la distance EUCLIDIENNE et pour $\boldsymbol{u}$ et $\boldsymbol{v}$ des matrices le carré de la distance de FROBENIUS.
$\frac{1}{2}\boldsymbol{t}^\top W \boldsymbol{t}$, W est semi définie positive	$\frac{1}{2}(\boldsymbol{u} - \boldsymbol{v})^\top W (\boldsymbol{u} - \boldsymbol{v})$	Le carré de la distance de MAHALANOBIS
$\sum t_i \ln(t_i)$	$\sum_i u_i \ln\left(\frac{u_i}{v_i}\right) - u_i + v_i$	Divergence de type KULLBACK-LEIBLER

4.6.2.3 La divergence de type β,($\beta \neq -\{1, 0\}$)

Cette divergence s'obtient avec :

$$\mathcal{D}_\beta(\boldsymbol{u}, \boldsymbol{v}) = \sum_i \left(u_i \frac{u_i^\beta - v_i^\beta}{\beta} - \frac{u_i^{\beta+1} - v_i^{\beta+1}}{\beta + 1} \right), \tag{4.6.11}$$

avec β un nombre réel ($\beta \neq -1$ et $\beta \neq 0$). Il est important de noter $\beta = 1$, on obtient la distance Euclidienne standard. Pour $\beta \to 0$ on obtient la divergence de KULLBACK-LEIBLER et pour $\beta \to -1$, on obtient distance appelée "distance ITAKURA-SAITO".

Parfois il est nécessaire d'imposer des restrictions supplémentaires sur les inconnues $\boldsymbol{A}$ et $\boldsymbol{Y}$ qui sont de la forme $\Phi(\boldsymbol{A})$ et $\Psi(\boldsymbol{Y})$ minimum. Alors à la place de (4.6.4), on obtient une autre formulation du problème NMF :

déterminer $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, $\boldsymbol{Y} \in \mathbb{R}^{n \times T}$ telles que

$$\mathcal{D}(\boldsymbol{AY}, \boldsymbol{X}) + C_A \Phi(\boldsymbol{A}) + C_Y \Psi(\boldsymbol{Y}) \tag{4.6.12}$$

soit minimum, avec C_A et C_Y des constantes positives. Les fonctions $\Phi(\boldsymbol{A})$ et $\Psi(\boldsymbol{Y})$ sont connues comme fonctions de pénalisation.

Une technique standard pour la minimisation d'une fonction $f(u_1, u_2, \ldots, u_n)$ est la suivante :

- *(i)* On choisit un point initial u
- *(ii)* On rajoute à u la grandeur

$$\delta u = -\eta \nabla f(u_1, u_2, \ldots, u_n) = -\sum_i \eta_i \frac{\partial f}{\partial u_i} \tag{4.6.13}$$

(déplacement dans la direction de descente maximale $-\nabla f$ multipliée par les paramètres positifs η appelés taux d'apprentissage)

— *(iii)* On arrête les itérations si une condition d'arrêt est vérifiée : soit δu est très petit ou bien le nombre d'itérations est largement suffisant pour espérer trouver un minimum local.

— Sinon, on refait l'étape *(ii)* avec un nouveau u.

4.6.3 Algorithmes multiplicatifs pour la NMF linéaire instantané

4.6.3.1 Algorithmes multiplicatifs pour la distance Euclidienne

Soit $\mathcal{D}$ la distance Euclidienne :

$$\begin{aligned} \mathcal{D}_2\left(\boldsymbol{AY}, \boldsymbol{X}\right) &= \tfrac{1}{2}\sum_{i,t}\left(\left[\boldsymbol{AY}\right]_{i,t} - x_{i,t}\right)^2 \\ &= \sum_{i,t}\left(\sum_{j=i}^{n} a_{i,j}y_{j,t} - x_{i,t}\right)^2 . \end{aligned} \tag{4.6.14}$$

Alors la fonction (4.6.12) à minimiser devient

$$F\left(\boldsymbol{A}, \boldsymbol{Y}\right) = \sum_{i,t}\left(\sum_{j=1}^{n} a_{i,j}y_{j,t} - x_{i,t}\right)^2 + C_A\Phi\left(\boldsymbol{A}\right) + C_Y\Psi\left(\boldsymbol{Y}\right). \tag{4.6.15}$$

Nous avons

$$\frac{\partial F\left(\boldsymbol{A}, \boldsymbol{Y}\right)}{\partial a_{i,j}} = \left[-\boldsymbol{XY}^\top + \boldsymbol{AYY}^\top\right]_{i,j} + C_A\frac{\Phi\left(\boldsymbol{A}\right)}{\partial a_{i,j}} \tag{4.6.16}$$

$$\frac{\partial F\left(\boldsymbol{A}, \boldsymbol{Y}\right)}{\partial y_{j,t}} = \left[-\boldsymbol{A}^\top\boldsymbol{X} + \boldsymbol{A}^\top\boldsymbol{AY}\right]_{j,t} + C_Y\frac{\Psi\left(\boldsymbol{Y}\right)}{\partial y_{j,t}} \tag{4.6.17}$$

Si les taux d'apprentissage sont choisis selon Lee et Seung (voir le livre de Cichocki et al. (2009), page 136) :

$$\eta_{i,j} = \frac{a_{i,j}}{\left[\boldsymbol{AYY}^\top\right]_{i,j}}, \tag{4.6.18}$$

$$\eta_{i,t} = \frac{y_{i,t}}{\left[\boldsymbol{A}^\top\boldsymbol{AY}\right]_{i,j}}, \tag{4.6.19}$$

la règle de mise à jour multiplicative (découlant du principe (4.6.13)) devient

$$a_{i,j} \leftarrow a_{i,j} - n_{i,j}\frac{\partial F(\boldsymbol{A},\boldsymbol{Y})}{\partial a_{i,j}} \approx a_{i,j}\frac{\left[\left[\boldsymbol{X}\boldsymbol{Y}^\top\right]_{i,j} - C_A\frac{\Phi(\boldsymbol{A})}{\partial a_{i,j}}\right]_+}{\left[\left[\boldsymbol{A}\boldsymbol{Y}\boldsymbol{Y}^\top\right]_{i,j}\right]_+}, \tag{4.6.20}$$

$$y_{j,t} \leftarrow y_{j,t} - n_{j,t}\frac{\partial F(\boldsymbol{A},\boldsymbol{Y})}{\partial y_{i,t}} \approx y_{j,t}\frac{\left[\left[\boldsymbol{A}^\top\boldsymbol{X}\right]_{j,t} - C_Y\frac{\Psi(\boldsymbol{Y})}{\partial y_{j,t}}\right]_+}{\left[\left[\boldsymbol{A}^\top\boldsymbol{A}\boldsymbol{Y}\right]_{j,t}\right]_+}. \tag{4.6.21}$$

On voit qu'en l'absence de Φ et Ψ, il n'est pas nécessaire de prendre les parties positives dans le processus de mise à jour, car les grandeurs qui interviennent sont non négatives. Afin d'éviter le cas $a_{i,j} = 0$ à une itération donnée (ce qui implique $a_{i,j} = 0$ dans les itérations suivantes), on prend pour la partie positive $[\bullet]_+ = \max\{\bullet, \varepsilon\}$ avec $\varepsilon > 0$ petit, par exemple $\varepsilon = 10^{-9}$. Pour éviter que le dénominateur devienne excessivement petit, on utilise aussi $[\bullet]_+$.

Cichocki et al. (2009) recommandent les mises à jour suivantes

$$a_{i,j} \leftarrow a_{i,j}\left\{\frac{\left[\left[\boldsymbol{X}\boldsymbol{Y}^\top\right]_{i,j} - C_A\frac{\Phi(\boldsymbol{A})}{\partial a_{i,j}}\right]}{\left[\left[\boldsymbol{A}\boldsymbol{Y}\boldsymbol{Y}^\top\right]_{i,j}\right]}\right\}_+^\omega \tag{4.6.22}$$

$$y_{j,t} \leftarrow y_{j,t}\left\{\frac{\left[\left[\boldsymbol{A}^\top\boldsymbol{X}\right]_{j,t} - C_Y\frac{\Psi(\boldsymbol{Y})}{\partial y_{j,t}}\right]}{\left[\left[\boldsymbol{A}^\top\boldsymbol{A}\boldsymbol{Y}\right]_{j,t}\right]}\right\}_+^\omega \tag{4.6.23}$$

avec $\omega \in [0.5, 2]$.

4.6.3.2 Algorithmes multiplicatifs pour la divergence Kullback-Leibler

Pour la divergence de Kullback-Leibler $\mathcal{D}_{KL}$

$$\mathcal{D}_{KL}(\boldsymbol{X}, \boldsymbol{A}\boldsymbol{Y}) = \sum_{i,t} x_{i,t}\log\left(\frac{x_{i,t}}{[\boldsymbol{A}\boldsymbol{Y}]_{i,t}}\right) - x_{i,t} + [\boldsymbol{A}\boldsymbol{Y}]_{i,t}, \tag{4.6.24}$$

les formules de mise a jour de $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, $\boldsymbol{Y} \in \mathbb{R}^{n\times T}$ s'obtiennent de manière analogue avec la fonction à minimiser qui est :

$$F(\boldsymbol{A},\boldsymbol{Y}) = \mathcal{D}_{KL}(\boldsymbol{X}, \boldsymbol{A}\boldsymbol{Y}) + C_A\Phi(\boldsymbol{A}) + C_Y\Psi(\boldsymbol{Y}). \tag{4.6.25}$$

Les formules de mise à jour sont :

$$a_{i,j} \leftarrow a_{i,j} - n_{i,j}\frac{\partial F(\boldsymbol{A},\boldsymbol{Y})}{\partial a_{i,j}} \approx a_{i,j}\frac{\left[\sum_t y_{j,t}\left(\frac{x_{i,t}}{[\boldsymbol{AY}]_{i,t}}\right) - C_A\frac{\Phi(\boldsymbol{A})}{\partial a_{i,j}}\right]_+}{\sum_t y_{j,t}} \tag{4.6.26}$$

$$y_{j,t} \leftarrow y_{j,t} - n_{j,t}\frac{\partial F(\boldsymbol{A},\boldsymbol{Y})}{\partial y_{i,t}} \approx y_{j,t}\frac{\left[\sum_i a_{i,j}\left(\frac{x_{i,t}}{[\boldsymbol{AY}]_{i,t}}\right) - C_Y\frac{\Psi(\boldsymbol{Y})}{\partial y_{j,t}}\right]_+}{\sum_i a_{i,j}} \tag{4.6.27}$$

avec des des taux d'apprentissage $\eta_{i,j} = \frac{a_{i,j}}{\sum_t y_{j,t}}$ et $\eta_{j,t} = \frac{y_{i,t}}{\sum_t a_{i,j}}$.

Pour améliorer la convergence, on peut utiliser (voir le livre de Cichocki et al. (2009) , page 140) :

$$a_{i,j} \leftarrow \left\{ a_{i,j}\left(\frac{\left[\sum_t y_{j,t}\left(\frac{x_{i,t}}{[\boldsymbol{AY}]_{i,t}}\right) - C_A\frac{\Phi(\boldsymbol{A})}{\partial a_{i,j}}\right]_+}{\sum_t y_{j,t}}\right)^{\omega}\right\}^{1+\alpha_A} \tag{4.6.28}$$

$$y_{j,t} \leftarrow \left\{ y_{j,t}\left(\frac{\left[\sum_i a_{i,j}\left(\frac{x_{i,t}}{[\boldsymbol{AY}]_{i,t}}\right) - C_Y\frac{\Psi(\boldsymbol{Y})}{\partial y_{j,t}}\right]_+}{\sum_i a_{i,j}}\right)^{\omega}\right\}^{1+\alpha_Y} \tag{4.6.29}$$

avec $\omega \in [0,2]$ et $\alpha_A, \alpha_Y \in [0.001, 0.005]$. La présence de ω produit une amélioration de la convergence et la présence de α_A, α_Y force la décroissance des composantes de $\boldsymbol{A}$ et de $\boldsymbol{Y}$ pour obtenir des résultats avec moins de composantes non nulles. On peut aussi utiliser $[\bullet]_+$ au dénominateur pour éviter qu'il devienne excessivement petit.

4.6.3.3 Algorithmes multiplicatifs pour la divergence α

La divergence Alpha est définie par

$$\mathcal{D}^{(\alpha)}(\boldsymbol{X},\boldsymbol{AY}) = \frac{1}{\alpha(\alpha-1)}\sum_{i,t}\left(x_{i,t}^{\alpha}[\boldsymbol{AY}]_{i,t}^{1-\alpha} - \alpha x_{i,t} + (1-\alpha)[\boldsymbol{AY}]_{i,t}\right), \tag{4.6.30}$$

et les formules de mise à jour sont données par

$$y_{j,t} \leftarrow y_{j,t}\left(\frac{\sum_{i=1}^{m} a_{i,j}\left(\frac{x_{i,t}}{[\boldsymbol{AY}]_{i,t}}\right)^{\alpha}}{\sum_{i=1}^{m} a_{i,j}}\right)^{\frac{1}{\alpha}} \tag{4.6.31}$$

$$\alpha_{i,j} \leftarrow \alpha_{i,j}\left(\frac{\sum_{t=1}^{T}\left(\frac{x_{i,t}}{[\boldsymbol{AY}]_{i,t}}\right)^{\alpha} y_{j,t}}{\sum_{t=1}^{T} y_{j,t}}\right)^{\frac{1}{\alpha}} \tag{4.6.32}$$

4.6.3.4 Algorithmes multiplicatifs pour la divergence β

Pour minimiser la fonction objectif à base d'une divergence β

$$F(\boldsymbol{A},\boldsymbol{Y}) = \underbrace{\sum_{i,t}\left(x_{i,t}\frac{x_{i,t}^{\beta}-[\boldsymbol{AY}]_{i,t}^{\beta}}{\beta}+\frac{[\boldsymbol{AY}]_{i,t}^{\beta+1}-x_{i,t}^{\beta+1}}{\beta+1}\right)}_{\mathcal{D}^{(\beta)}(\boldsymbol{X},\boldsymbol{AY})} + C_A\|\boldsymbol{A}\|_1 + C_Y\|\boldsymbol{Y}\|_1, \tag{4.6.33}$$

les formules de mise à jour sont données par

$$y_{j,t} \leftarrow y_{j,t}\frac{\left[\sum_{i=1}^{m} a_{i,j}\left(\frac{x_{i,t}}{[\boldsymbol{AY}]_{i,t}^{1-\beta}}\right)-C_Y\right]_+}{\sum_{i=1}^{m} a_{i,j}\,[\boldsymbol{AY}]_{i,t}^{\beta}} \tag{4.6.34}$$

$$\alpha_{i,j} \leftarrow \alpha_{i,j}\frac{\left[\sum_{t=1}^{T}\left(\frac{x_{i,t}}{[\boldsymbol{AY}]_{i,t}^{1-\beta}}\right)y_{j,t}-C_A\right]_+}{\sum_{t=1}^{T}[\boldsymbol{AY}]_{i,t}^{\beta}\,y_{j,t}} \tag{4.6.35}$$

4.6.4 Algorithmes multiplicatifs convolutifs pour la NMF

Dans cette situation le mélange est convolutif (retardé), ce qui peut être formalisé comme :

$$\boldsymbol{X} = \sum_{p=0}^{p-1}\boldsymbol{A}_p\overset{p\rightarrow}{\boldsymbol{Y}}+\boldsymbol{E}, \tag{4.6.36}$$

où $\overset{p\rightarrow}{\boldsymbol{Y}}$ est la matrice des estimations des sources retardée de p unités :

$$\overset{p\rightarrow}{\boldsymbol{Y}} = \boldsymbol{Y}_{t-p}. \tag{4.6.37}$$

Pour le modèle estimé

$$\widetilde{\boldsymbol{X}} = \sum_{p=0}^{p-1}\boldsymbol{A}_p\overset{p\rightarrow}{\boldsymbol{Y}}, \tag{4.6.38}$$

où $\boldsymbol{A}_p = [\boldsymbol{a}_{i,j,p+1}]_{i=1,\ldots,m\,;\,j=1,\ldots,n}$, $p=0,1,\ldots,p-1$, la divergence bêta (par exemple) de $\boldsymbol{E}=\boldsymbol{X}-\widetilde{\boldsymbol{X}}$ est :

$$\mathcal{D}^{(\beta)}\left(\boldsymbol{X},\widetilde{\boldsymbol{X}}\right) = \sum_{i,t}\left(x_{i,t}\frac{x_{i,t}^{\beta}-\widetilde{x}_{i,t}^{\beta}}{\beta}+\frac{\widetilde{x}_{i,t}^{\beta+1}-x_{i,t}^{\beta+1}}{\beta+1}\right) \tag{4.6.39}$$

et les formules de mise à jour sont données par

$$\alpha_{i,j,p+1} \leftarrow \alpha_{i,j,p+1} \frac{\sum_{t=1}^{T} \left(x_{i,t} \widetilde{x}_{i,t}^{\beta-1} \right) y_{j,t-p}}{\sum_{t=1}^{T} \widetilde{x}_{i,t}^{\beta} y_{j,t-p}} \tag{4.6.40}$$

$$y_{j,t} \leftarrow y_{j,t} \frac{\sum_{i=1}^{m} \sum_{p=1}^{p-1} a_{i,j,p+1} x_{i,t+p} \widetilde{x}_{i,t+p}^{\beta-1}}{\sum_{i=1}^{m} \sum_{p=1}^{p-1} a_{i,j,p+1} \widetilde{x}_{i,t+p}^{\beta}} \tag{4.6.41}$$

4.7 Applications aux données de la QAI

4.7.1 Sur les données des concentrations de particules dans le bureau individuel

Nous avons commencé par la caractérisation des sources de particules dont la concentration a été mesurée dans le bureau individuel, car cet environnement est moins complexe que l'espace paysager. Plus précisément, il s'agit de caractériser la structure des profils temporels des sources de particules. Pour cela, nous avons d'abord entrepris une analyse statistique des concentrations en nombre de particules dans cet environnement de bureau, étape préalable au développement de modèles permettant de quantifier la contribution des sources. Ensuite, nous avons fait appel aux méthodes appliquées en statistique exploratoire des données et en traitement du signal pour la séparation des sources : ACP, ACI, PMF et NMF. Les résultats préliminaires ont été présentés dans (Ouaret et al., 2014c) et étendus dans (Ouaret et al., 2014a). Dans cette dernière communication, nous avons mené une étude comparative entre la PMF, l'ACI et la NMF.

Les données prises en compte dans cette étude concernent 45 jours consécutifs, du 19 février 2011 à minuit jusqu'au 4 avril 2011 à 23h50. Le pas de temps est de 10 minutes. Les résultats de la statistique descriptive montrent qu'en termes de variabilité, plus les particules sont fines, plus leur écart-type est grand et plus elles sont corrélées entre elles. En effet, les coefficients de corrélation de Pearson pour les particules entre 0.3 μm et 0.80 μm varient entre 0.75 et 0.98. Les coefficients de corrélation entre les particules $>$ 20 μm et $<$ 4.5 μm ne dépassent pas 0.35. Les coefficients d'asymétrie des paramètres mesurés sont positifs, leurs queues de distribution sont étalées vers la droite ; ceci laisse présager des distributions log-normales.

Outre la mesure d'aplatissement et d'asymétrie, les distributions de probabilités ont été estimées par une méthode non paramétrique (à noyau gaussien). Une distribution log-normale a été également ajustée aux données. De façon générale, on remarque que pour les particules fines, les distributions en nombre sont caractérisées par une densité de probabilité de type log-normale ; ce n'est pas le cas pour les grosses particules, en particulier du fait que leur nombre est très souvent voisin de zéro. On peut se référer au chapitre 2 pour une analyse statistique plus complète.

En utilisant l'analyse de la fonction d'autocorrélation et l'estimation de la dépendance temporelle, nous étudions les propriétés de la structure temporelle des sources de particules.

La Figure 4.7.1 présente la méthodologie générale suivie en vue de caractériser la typologie des émissions des sources. Dans le but de comprendre la structure de ces fluctuations, on s'intéresse, d'une part aux propriétés intrinsèques des séries (la forme de la fonction d'autocorrélation), et d'autre part aux spécificités de la variabilité, en faisant une classification pour trouver des journées-types. L'analyse des profils diurnes des séries temporelles dans un bureau individuel met en évidence la présence des niveaux de concentration élevés durant les heures de travail, notamment, pour les grosses particules et le CO_2,

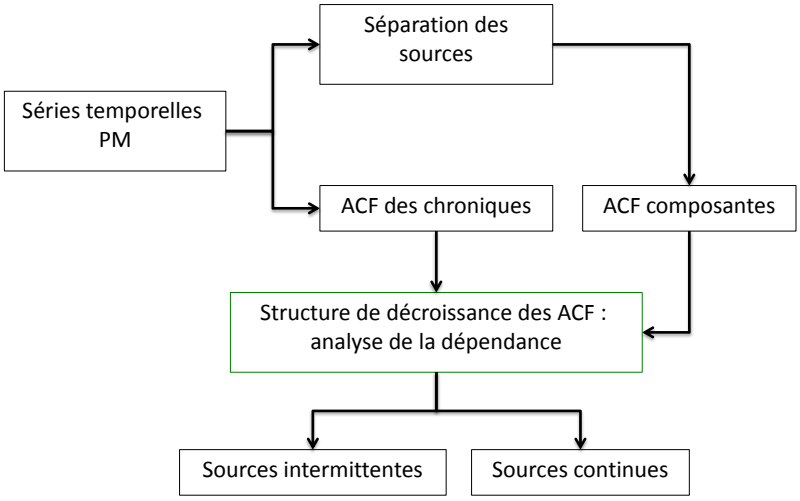


FIGURE 4.7.1 – *Démarche suivie pour la caractérisation de la structure des profiles des sources.*

tandis que pour les particules comprises entre 0.30 et 0.5 μm de diamètre, les profils diurnes exhibent une faible variabilité.

Les caractéristiques de la variabilité des phénomènes qui ont généré les séries temporelles ne sont pas forcément connues, la seule information dont nous disposons est l'historique des séries mesurées. Justement, la fonction d'autocorrélation (ACF) nous permet d'explorer en partie les structures statistiques des séries.

La Figure 4.7.2 présente les courbes de la fonction d'autocorrélation (ACF) (les corrélogrammes) calculées à partir des séries temporelles de la concentration en nombre de particules et du niveau du CO2. On remarque que les valeurs de l'ACF de la série des concentrations de CO_2 sont significativement non nulles jusqu'à ce que le corrélogramme soit tronqué à partir de 8h (480 minutes). Le même type de comportement est observé pour les fractions de taille supérieure à 1.6 μm. La plupart des fonctions d'autocorrélation se comportent comme un mélange de fonctions exponentielles/sinusoïdales amorties, en particulier pour les particules de taille moyenne et le CO_2. En effet, l'ACF de ces fractions traduit l'aspect de saisonnalité de ces concentrations.

En revanche, les corrélogrammes correspondants aux tailles de particules fines présentent des autocorrélations relativement élevées avec une décroissance très lente, le premier passage à zéro se produisant après 2 jours seulement. Les fortes valeurs de l'ACF expriment ici, à la fois la force et l'étalement de la persistance. En fait, cette persistance matérialisée par une décroissance lente des autocorrélations (corrélation à long terme) représente un mécanisme complexe associé aux sources de particules fines. Ces sources présentent des fluctuations multi-fréquentielles, c'est-à-dire selon différentes échelles temporelles.

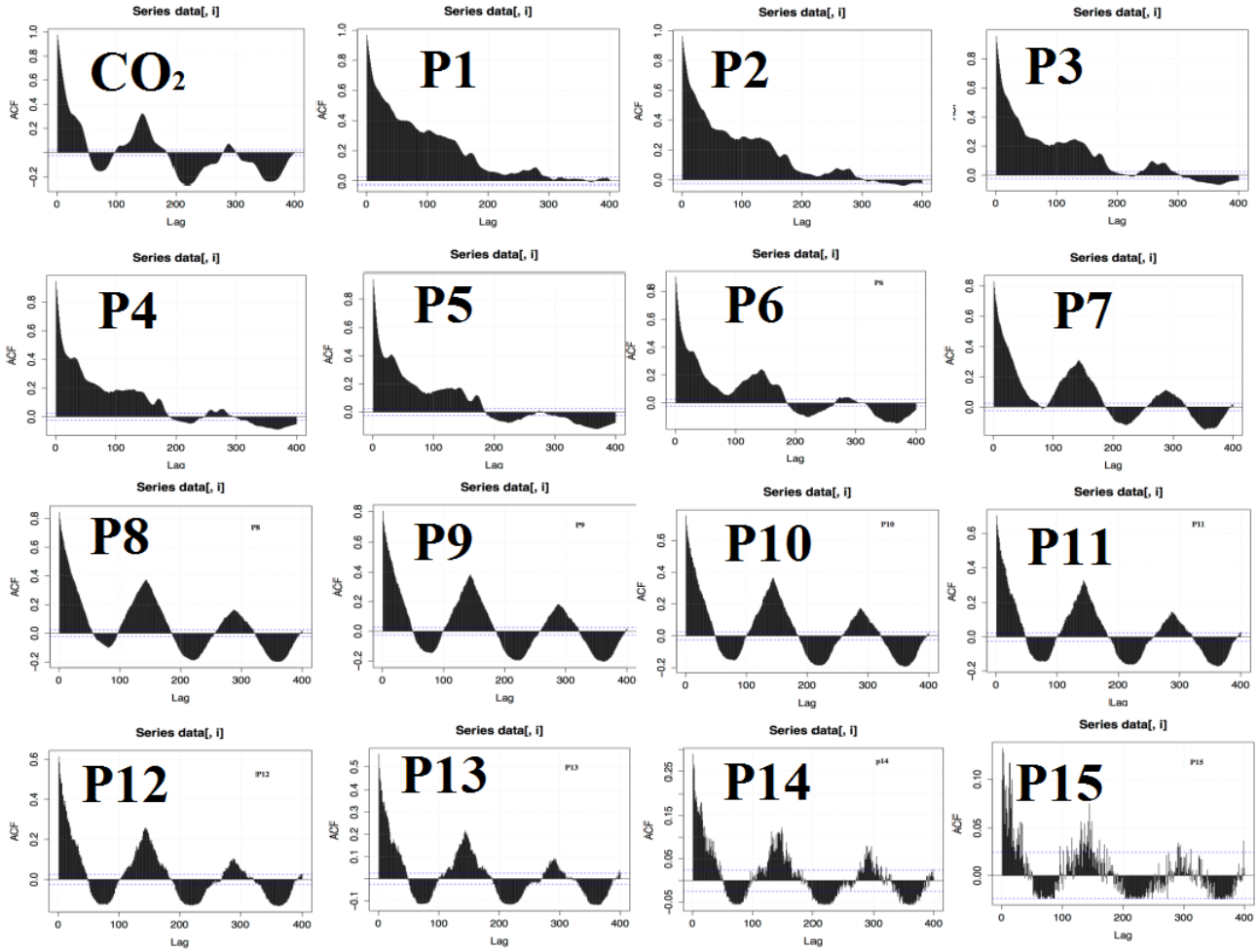


FIGURE 4.7.2 – *Fonctions d'autocorrélation des séries temporelles particulaires des 15 fractions granulométriques (de P1 au P15) et du CO_2 ; le retard (lag) est en ×10 minutes. Les deux lignes horizontales pointillées représentent l'intervalle de confiance à 95% pour les autocorrélations.*

Nous avons adopté les notations suivantes pour désigner les différentes fractions médianes de particules :

- P1 pour $0.30-0.40$ μm ; P2 pour $0.40-0.50$ μm ; P3 pour $0.50-0.65$ μm ; P4 pour $0.65-0.80$ μm ; P5 pour $0.80-1.0$ μm ; P6 pour $1.0-1.6$ μm ;
- P7 pour $1.6-2.0$ μm ; P8 pour $2.0-3.0$ μm ; P9 pour $3.0-4.0$ μm ; P10 pour $4.0-5.0$ μm ; P11 pour $5.0-7.5$ μm ; P12 pour $7.5-10$ μm ; P13 pour $10.0-15.0$ μm ; P14 pour $15.0-20.0$ μm ; pour >20 μm.

L'analyse en composantes principales a été appliquée aux séries temporelles des polluants (toutes les fractions de particules), chaque série étant considérée comme une variable. Les variables ont été standardisées (centrées-réduites) pour neutraliser l'effet d'échelle provoqué essentiellement par le nombre des particules fines. Chaque variable reçoit ainsi la même importance, quelle que soit sa moyenne et son écart-type.

Les six premières composantes principales expliquent 89% de la variance totale, avec 55% de variabilité expliquée par la première composante principale (Figure 4.7.3a). Le corrélogramme associé à la première composante principale possède une forme sinusoïdale et alterne de signe (Figure 4.7.3b).

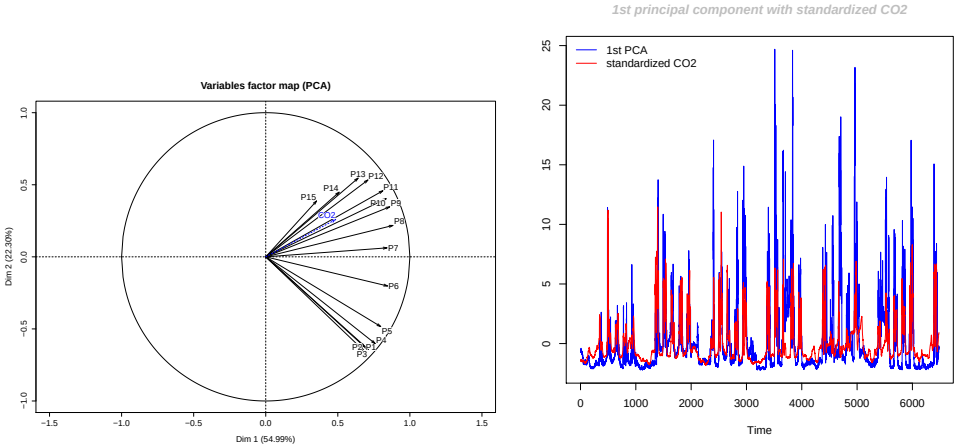
On retrouve la même forme de corrélogramme que celle associée à la concentration de CO_2 et des particules de tailles moyennes (1.6 μm à 7.5 μm). Cette structure semblable d'autocorrélation laisse présager l'existence d'une source commune, vraisemblablement liée à l'occupation, dont le principal indicateur est le CO_2. Les trois autres composantes traduisent la variabilité associée aux particules fines comme on peut en juger avec la forme des fonctions d'autocorrélations associées. Elles peuvent être associées à trois sources distinctes de variation de la concentration des particules fines. Les profils temporels des composantes indiquent le schéma de variation de ces sources dans le temps.

La trajectoire des concentrations du CO_2 et de la première composante principale mettent en évidence le même type de variation (Figure 4.7.3a). Elles sont caractérisées par des phases d'alternances (expansion et contraction) des fluctuations. Ceci montre l'importance de l'occupation et l'activité des occupants au regard de la variabilité de la pollution particulaire.

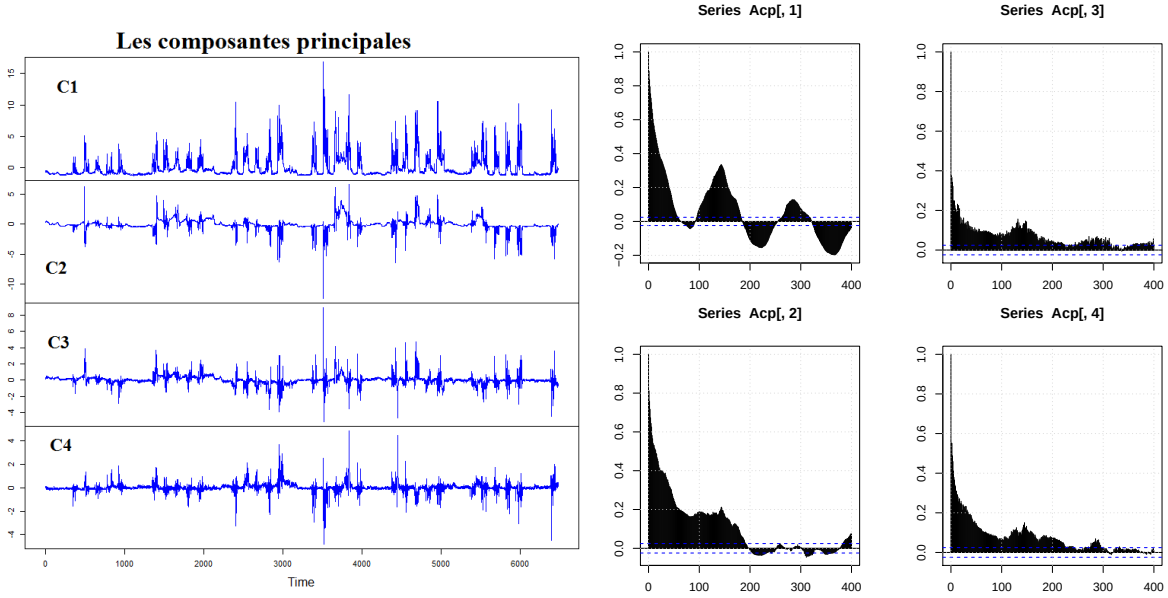
On peut interpréter la première composante de l'ACP appliquée aux particules comme le facteur associé à l'occupation dans le bureau individuel. Elle explique 55% de la variance totale des concentrations des particules. En projetant le CO_2 (en tant que variable passive) sur le cercle de corrélations, on peut remarquer qu'il est très corrélé aux variables P9, P10 et P11, ce qui représente des particules de taille moyenne.

Les variables les plus corrélées à la première composante principale sont P6 ($1.0-1.6$ μm) et P7 ($1.6-2.0$ μm). Selon la deuxième composante de l'ACP qui explique 22% de la variance totale, on a deux classes de particules : un groupe dans le premier quadrant du cercle de corrélations (les particules P7,..., P15) correspondant aux grosses particules et les particules P1,...,P6 dans le quatrième quadrant. En effet, ces deux classes exhibent des variabilités différentes, comme on a pu le remarquer sur l'ACF de ces variables.

L'analyse en composantes indépendantes (ACI) requiert de fixer en entrée le nombre de composantes (ou sources de variation). Plusieurs simulations ont été faites afin d'évaluer le critère RMSE (racine carrée de l'erreur quadratique moyenne) en fonction du nombre de composante indépendantes. Cette



(a) *Cercle de corrélations définie par les deux première composantes de l'ACP appliquée aux concentrations de particules de P1 au P15 (à gauche) ; la représentation de la 1ère composante principale et de la concentration de CO_2 standardisée (à droite).*



(b) *Profils temporels des quatre premières composantes principales (à gauche) et les corrélogrammes associés à ces composantes (à droite).*

FIGURE 4.7.3 – *ACP pour les concentrations des particules dans le bureau individuel en 2011.*

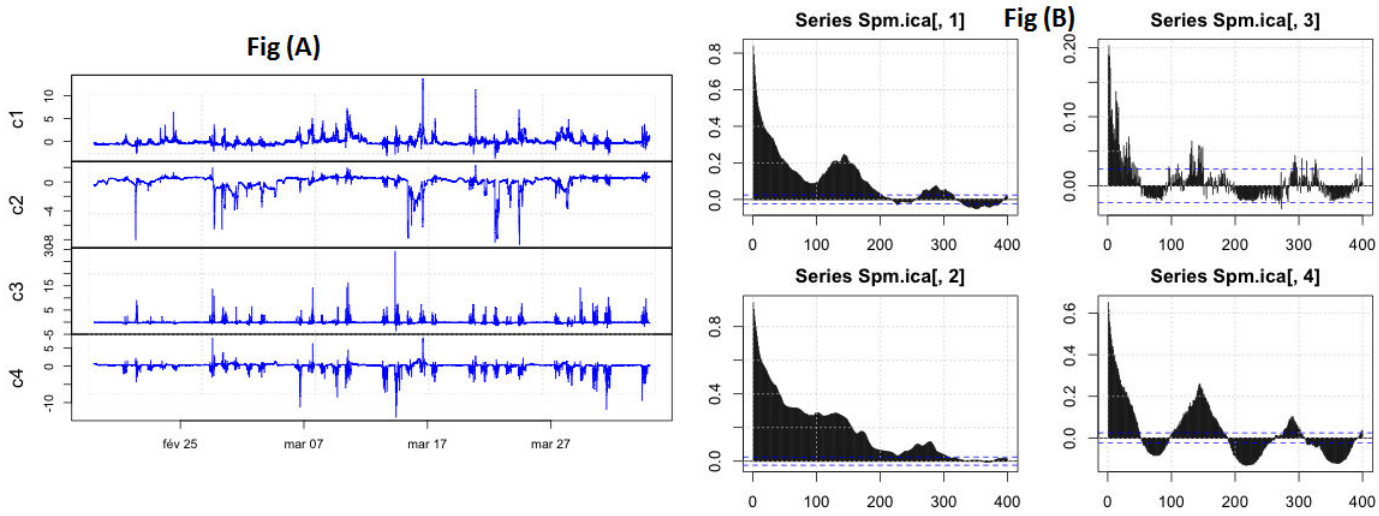


FIGURE 4.7.4 – *Fig (A) : Profils temporels des quatre premières composantes indépendantes. Fig (B) : corrélogrammes associés aux composantes indépendantes.*

erreur représente l'écart entre les séries temporelles initiales et les variables reconstituées à partir des composantes indépendantes.

Le choix du nombre de composantes retenues est basé sur la recherche du premier minimum local du RMSE. On remarque qu'en se basant sur ce critère, le nombre de composantes à retenir pour chaque fraction varie en fonction de la taille médiane des particules. Pour les séries 0.72 µm, 0.9 µm, 6.25 µm et 8.75 µm, seules 2 composantes suffisent pour atteindre un minimum local. Sinon pour la plupart des fractions, 4 composantes sont nécessaires, et par conséquent, nous retiendrons 4 composantes indépendantes pour l'ACI.

La Figure 4.7.4 montre les fluctuations temporelles des quatre composantes indépendantes obtenues sur les 15 fractions, ainsi que les graphiques de la fonction d'autocorrélation de toutes les composantes indépendantes (c1,. . .,c4).

La deuxième composante indépendante obtenue est inversée en signe, mais ceci n'est qu'un artefact, car les solutions sont déterminées à un facteur près, donc la multiplication par le facteur (-1) est envisageable. Les autocorrélations de cette composante sont significatives pour des retards importants (2 jours), montrant ainsi l'influence des observations très éloignées sur les plus récentes, d'où l'idée d'une mémoire de la série. La lente décroissance hyperbolique de la fonction d'autocorrélation indique que la variabilité des concentrations des particules fines admet les caractéristiques des séries à mémoire longue. De façon générale, une série présentant des corrélations à long terme est définie comme une série présentant une fonction d'autocorrélation de la forme : $\zeta(\tau) \propto \tau^{\lambda}$ avec $0 < \lambda < 1$.

En d'autres termes, la fonction d'autocorrélation suit une loi de puissance. Ce type de comportement est lié aux émissions des sources chroniques. Tout comme la première composante principale, la structure de décroissance du corrélogramme de la 4ème composante indépendante indique que celle-ci présente des fluctuations du niveau de particules similaires aux fluctuations du niveau de CO_2. Avec quatre composantes, l'ACI permet de mettre en évidence l'existence d'une source responsable des fluctuations des grosses particules (la 3ème composante indépendante), tandis qu'il faut retenir six dimensions pour l'ACP (95% de variance expliquée) pour pouvoir faire apparaître ce type de variation au niveau des autocorrélations. Ceci étant dû au fait que la contribution des grosses particules en termes de variance par rapport à la variance totale de toutes les particules est faible.

Les valeurs de l'autocorrélation de la 3ème composante indépendante se présentent comme des pics discrets avec de très faibles corrélations. Cette discontinuité est souvent liée à des phénomènes de faible probabilité d'occurrence. Les fonctions d'autocorrélation des composantes indépendantes 1 et 2 semblent proches de celles observées pour les composantes principales 2, 3 et 4. Pourtant, les profils temporels de ces composantes ne se ressemblent pas.

La contribution des sources estimée par l'ACI pour chaque composante indépendante sur chaque fraction de particules est présentée à la Figure 4.7.5. Les quatre composantes indépendantes contribuent de manière distributive en fonction de la taille des particules. Pour les fractions les plus fines de particules (0.35 µm, 0.45 µm, 0.57 µm), la contribution de la composante 2 avoisine les 80% de l'ensemble des contributions des autres composantes. Elle représente en plus, la source la moins contributive pour les grosses particules (>20 µm).

La composante 3 représente sur cette gamme de particules plus de 95% de l'ensemble des autres contributions de sources. Notons par ailleurs la particularité de discontinuité de la fonction d'autocorrélation pour la composante 3 : elle caractérise les émissions dues aux phénomènes aléatoires, tandis que pour la composante 2, les autocorrélations de type fonction de puissance caractérise les émissions des sources déterministes.

Pour la fraction de taille inférieure à 0.9 µm, les composantes indépendantes 2 et 1 partagent à part égale plus de 93% de l'ensemble des contributions. La composante 4 caractérisée par une décroissance sinusoïdale des autocorrélations correspond à la source majoritaire pour la gamme de taille 2-10 µm. Elle est de plus associée aux fluctuations de concentration du CO_2, elle est par conséquent reliée à la présence d'occupants dans la pièce. Il s'agit vraisemblablement de particules remises en suspension ou générées par les activités de l'occupant dans le bureau.

Les contributions des composantes principales sont très différentes de celles déterminées à partir des composantes indépendantes. Au regard de l'association dans chaque cas d'une des composantes avec la variation du niveau de CO_2, la contribution de la composante indépendante 4 semble mieux correspondre avec ce qui a été observé par ailleurs : à savoir que les particules les plus fines ne sont pas corrélées avec le CO_2. Au contraire, la première composante principale, très corrélée avec le CO_2 fournit une contribution pratiquement identique quelle que soit la taille des particules. Dès lors, les composantes indépendantes semblent plus facilement interprétables en termes de sources de fluctuations de particules que les composantes principales.

4.7.2 Comparaison des méthodes séparation des sources

Suites aux travaux initiés dans (Ouaret et al., 2014c) (section précédente), dans (Ouaret et al., 2014a), nous avons mené une étude comparative entre les différentes méthodes de séparation aveugle des sources. Enfin dans (Ouaret et al., 2016), nous comparons les résultats obtenus avec la NMF dans deux campagnes de mesures différentes : l'espace paysager en 2012 et en 2015.

Quatre méthodes de séparation des sources, en l'occurrence l'analyse par composantes indépendantes (ACI), la factorisation positive (Positive Matrix Factorization), la factorisation en matrices non-négatives (NMF) et l'analyse par composantes principales (ACP) ont été appliquées aux données mesurées dans le bureau individuel en 2011 (Ouaret et al., 2014a).

Les conditions générales des simulations par rapport à l'algorithme et la métrique utilisés sont présentées dans le Tableau 4.7.1.

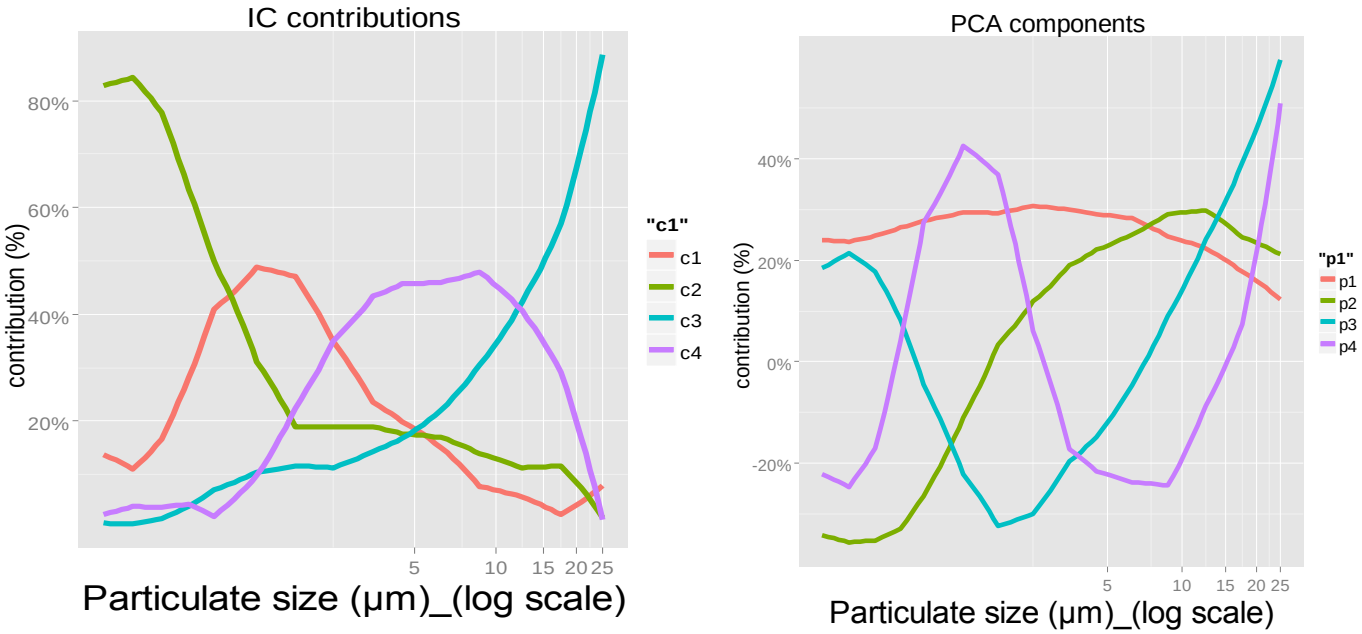


FIGURE 4.7.5 – *Contribution relative des composantes indépendantes (gauche) et vecteurs propres de l'ACP selon le diamètre des particules (droite).*

TABLE 4.7.1 – Conditions des simulations des méthodes de séparation des sources.

Méthode	**Algorithme**	**Métrique**	**Itérations**
ICA	Fast-ICA	Néguentropie	200
NMF	Brunet	Kullback-Leibler	20
PMF	Multilinear Engine	Euclidienne	20

L'un des paramètres critiques de ces méthodes est l'estimation du nombres de sources/facteurs en jeu. Dans la plupart des études, il est imposé considérant que le processus global est connu et seules les contributions respectives sont recherchées. Ce n'est pas le cas, lorsque ces méthodes sont utilisées en aveugle. Ce nombre de sources peut toutefois être estimé en comparant les erreurs résiduelles de plusieurs simulation (run) (*cf.* Figure 4.7.6). Par exemple, dans cette étude le nombre optimal de facteurs à prendre en compte est de 4 dans le bureau individuel et de 3 dans l'espace de bureaux, correspondant au point d'inflexion (Ouaret et al., 2014a). Par ailleurs, l'analyse en composantes indépendantes (ACI) requiert de fixer en entrée le nombre de composantes (ou sources de variations). Plusieurs simulations ont été faites afin d'évaluer le critère RMSE (racine carrée de l'erreur quadratique moyenne) en fonction du nombre de composante indépendantes (*cf.* Figure 4.7.6). Cette erreur représente l'écart entre les séries temporelles initiales et les variables reconstituées à partir des composantes indépendantes. Le choix du nombre de composantes retenues est basé sur la recherche du premier minimum local du RMSE. On remarque qu'en se basant sur ce critère, le nombre de composantes à retenir pour chaque fraction varie en fonction de la taille médiane des particules. Pour les séries 0.72 µm, 0.9 µm, 6.25 µm et 8.75 µm, 2 composantes suffisent pour atteindre le minimum local. Sinon pour la plupart des fractions, 4 composantes sont nécessaires, et par conséquent, nous retiendrons 4 composantes indépendantes pour l'ACI. En résumé, le nombre de composantes-facteurs dépend non seulement de la taille médiane des particules, mais aussi des environnements étudiés et du pas d'échantillonnage.

Les profils temporels normalisés des 4 facteurs extraits par la PMF et par l'ACI pour le bureau individuel sont représentés sur la Figure 4.7.7. Chaque facteur est associé à la gamme de taille de particules à laquelle il contribue le plus, à hauteur : de plus de 10 % pour la PMF et à hauteur de plus de 30 % pour l'ACI. Les facteurs le plus à droite F1 pour la PMF et C3 pour l'ACI présentent des fluctuations temporelles qui rappellent celles du CO_2. Les composantes de l'ACI apparaissent plus bruitées avec des valeurs négatives. Pourtant, les profils des composantes de l'ACI présentent des allures similaires aux facteurs de la PMF. Ainsi la composante C2 évoque les facteurs F2 et F4 par ailleurs très ressemblants. Les composantes C1 et C4 rappellent respectivement les facteurs F1 et F4.

Les contributions de chaque facteur aux différentes tailles de particules selon les différentes méthodes sont représentées sur la Figure 4.7.8. La numérotation des composantes varie non seulement selon la méthode utilisée, mais aussi en fonction des "runs", donc l'ordre des composantes n'a pas une signification particulière comme dans le cas de l'ACP.

Les composantes obtenues par ACI et NMF présentent des profils des contributions relativement similaires. La composante C2 de l'ACI et H3 de la NMF pourraient correspondre en termes de profils de contributions à l'association des facteurs F2 et F4 de la PMF. A l'inverse, le facteur 1 de la PMF semble

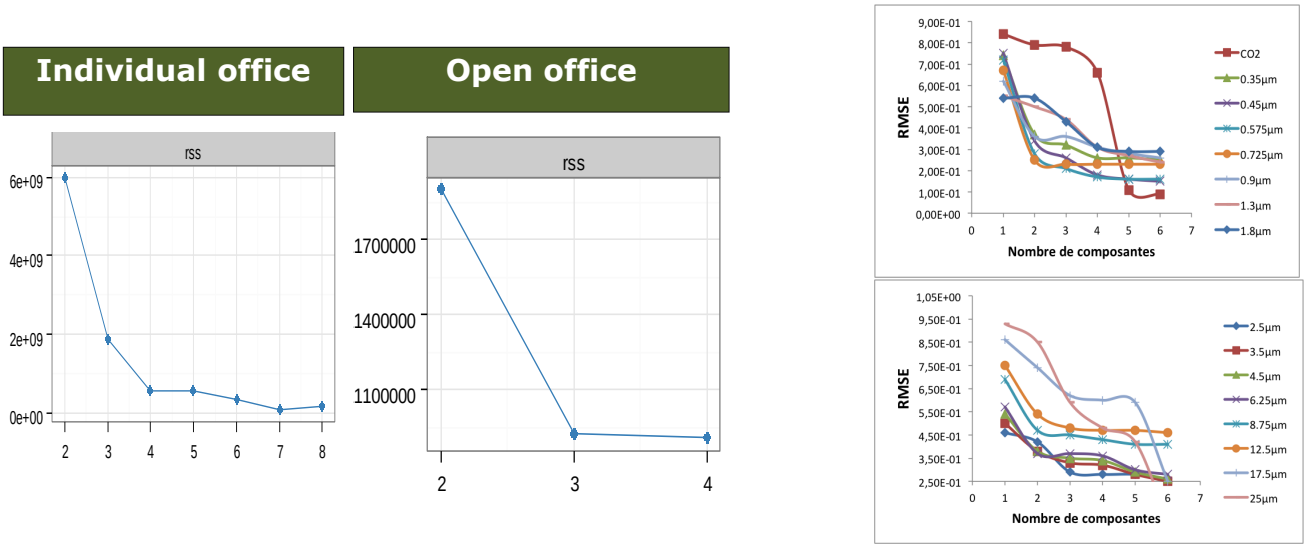


FIGURE 4.7.6 – *Évolution de l'erreur résiduelle obtenue après plusieurs simulations (rss : residual sum of squares) obtenue avec la NMF pour les particules dans le bureau individuel (première Figure à gauche) et dans l'espace paysager (deuxième figure à gauche). À droite en premier panel, évolution du RMSE en fonction du nombre de composantes indépendantes des fractions de tailles médianes 0.35-1.8 µm et du CO_2. À droite en deuxième panel : fractions de tailles médianes 2.5-25 µm dans le bureau individuel 2011 (toutes les 10 minutes).*

plutôt combiner les composantes C3 et C4 de l'ACI ou H2 et H4 de la NMF. Les trois méthodes fournissent des résultats assez proches en termes d'interprétation possible. La PMF se différencie des deux autres par le fait que ses facteurs affichent un contraste marqué dans les contributions des différentes tailles de particules, alors que pour l'ACI et la NMF, les contributions sont rarement nulles.

Les composantes extraites par l'ACI représentent des sources (ou groupes de sources) "indépendantes" au sens statistique. Quelque soit le critère statistique utilisé (indépendance) ou algébrique (non-négativité des matrices), les trois méthodes de séparation des sources fournissent des facteurs robustes interprétables selon le profil temporel et par rapport à la contribution moyenne des particules.

4.7.3 Séparation et contributions des sources : campagne de 2015

4.7.3.1 Séparation des sources des concentrations de particules

Les méthodes de séparation des sources avec la NMF et la PMF ont été appliquées sur les 6 mois de mesures des concentrations de particules durant la campagne 2015.

Étant donné le "volume" considérable des données utilisées ($\approx 300000 \times 15$ canaux), la factorisation de la matrice par la méthode PMF avec l'algorithme ME nécessite une gestion efficace de la mémoire de l'ordinateur. Le logiciel fourni par EPA (voir Norris et al. (2015)) ne fonctionne plus à partir de quelques centaines de milliers de lignes. Nous avons donc utilisé $\approx 100000 \times 15$ valeurs après la suppression des lignes ayants des valeurs manquantes.

La Figure 4.7.9 donne les contributions relatives aux différentes tailles de particules selon les deux méthodes. Clairement, les profils de contributions sont très similaires à ceux obtenus durant la campagne

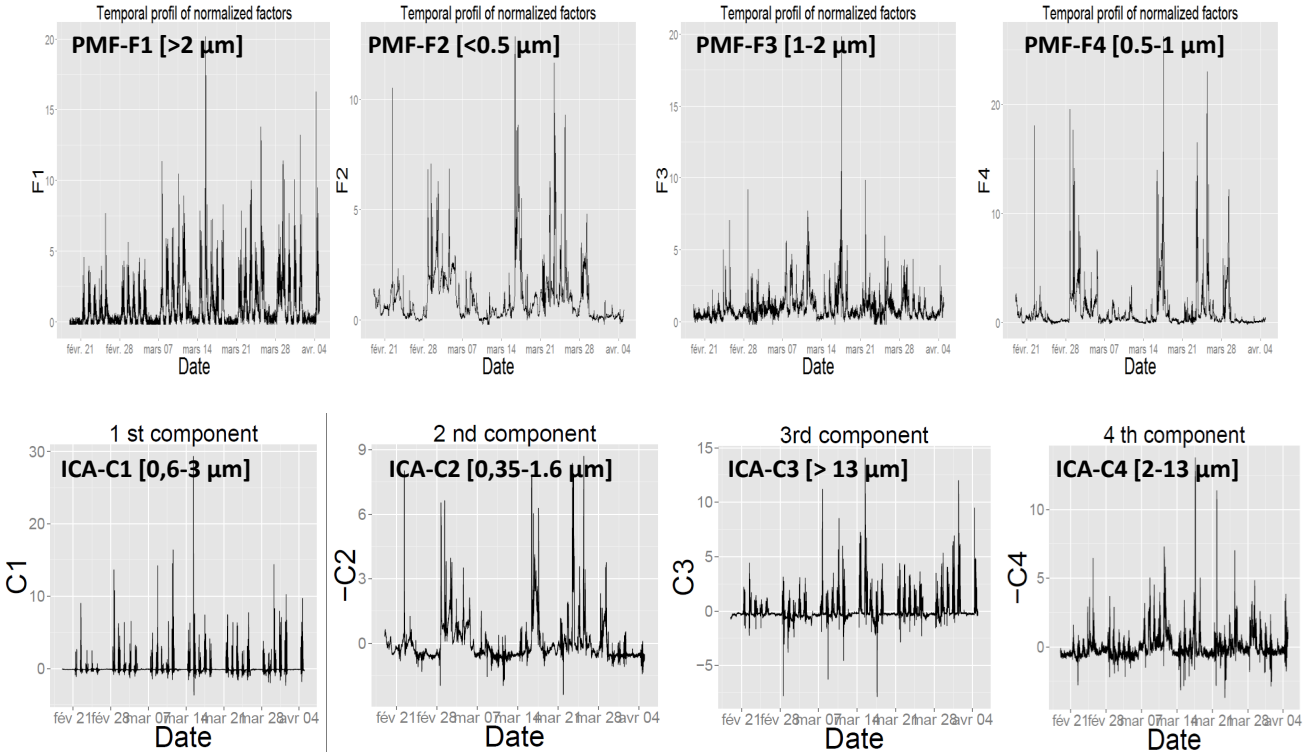


FIGURE 4.7.7 – *Profils temporels des facteurs-composantes extraits par la PMF ou l'ACI de la variation des concentrations de particules dans le bureau individuel de la campagne 2011.*

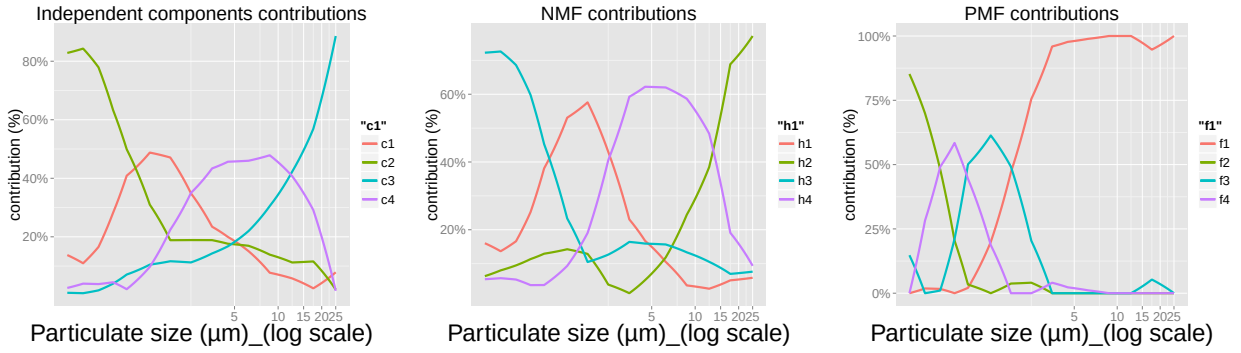


FIGURE 4.7.8 – *Contribution moyenne des facteurs aux différentes tailles de particules selon la méthode de séparation. Les mesures de concentrations de particules sont issues de la campagne 2011 dans le bureau individuel. Le pas de temps était d'une minute.*

2011 dans le bureau individuel. En effet, les composantes obtenues avec la factorisation NMF contribuent de manière distributive en fonction de la taille des particules. Alors que la méthode PMF met en évidence deux groupes bien distincts : les facteurs 1, 2 et 4 forment un cluster majoritaire des particules de taille inférieure à 0.755 µm, tandis que le troisième facteur contribue seul sur une large fraction de particules, de 0.9 à 20 µm.

En ce qui concerne les résultats de la NMF, la contribution du troisième facteur sur les fractions des particules fines (0.35 µm, 0.45 µm, 0,575 µm) avoisine les 90%. On retrouve cette proportion pour la contribution de la composante 4 sur les grosses particules (>12.5 µm). Les contributions majoritaires des deux dernières composantes (2 et 3) se trouvent au niveau du mode des fractions 1.3 et 3.5 µm.

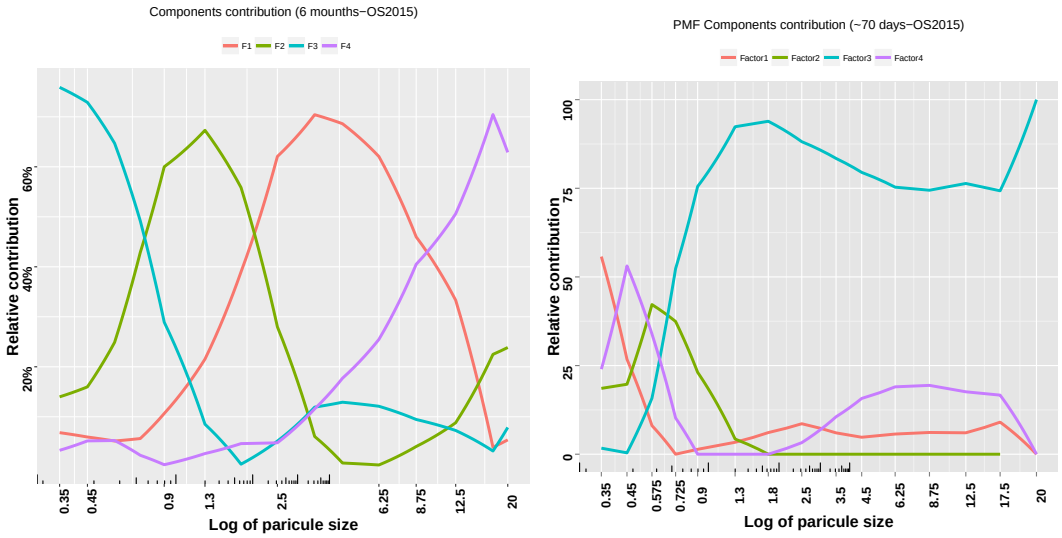


FIGURE 4.7.9 – *Contributions relatives des sources/facteurs de particules obtenues avec les méthodes NMF (à gauche) et PMF (à droite) dans l'espace paysager campagne 2015, pas de temps d'une minute.*

La déconvolution sur toute la période de mesure fournit une image instantanée de la contribution moyenne. Nous effectuons une analyse de NMF pour chaque mois de mesures. Donc, pour les fluctuations des concentrations de particules durant la campagne de 2015, six bases de données ont été extraites selon les mois. Cette analyse aborde la question de savoir si les contributions des sources/facteurs est invariante dans le temps ou pas. On s'intéresse à la variabilité mensuelle des facteurs extraits par NMF. Sur la Figure 4.7.10, nous présentons les profils de contributions par NMF pour chaque mois. Pratiquement, aucune différence significative ne permet de mettre en évidence un effet "mois" sur les contributions relatives des composantes extraites par la NMF. En effet, les contributions moyennes observées sur toute la période représente les contributions effectives de chaque mois :

- La première composante contribue à environ 80% aux fluctuations des fractions de tailles 0.3-0.75 µm ;
- La deuxième et la troisième partagent leurs impact (60% chacune) sur les gammes de tailles moyennes (0.9-4.5 µm) ;
- La quatrième composante est associée aux fluctuations des grosse particules, avec une contribution de 80%.

Cette analyse montre que finalement, sur une échelle mensuelle, l'impact des contributions à la variabilité des particules est minime. Nous regardons maintenant les profils des fluctuations sur les autres échelles temporelles afin de déceler les principales caractéristiques.

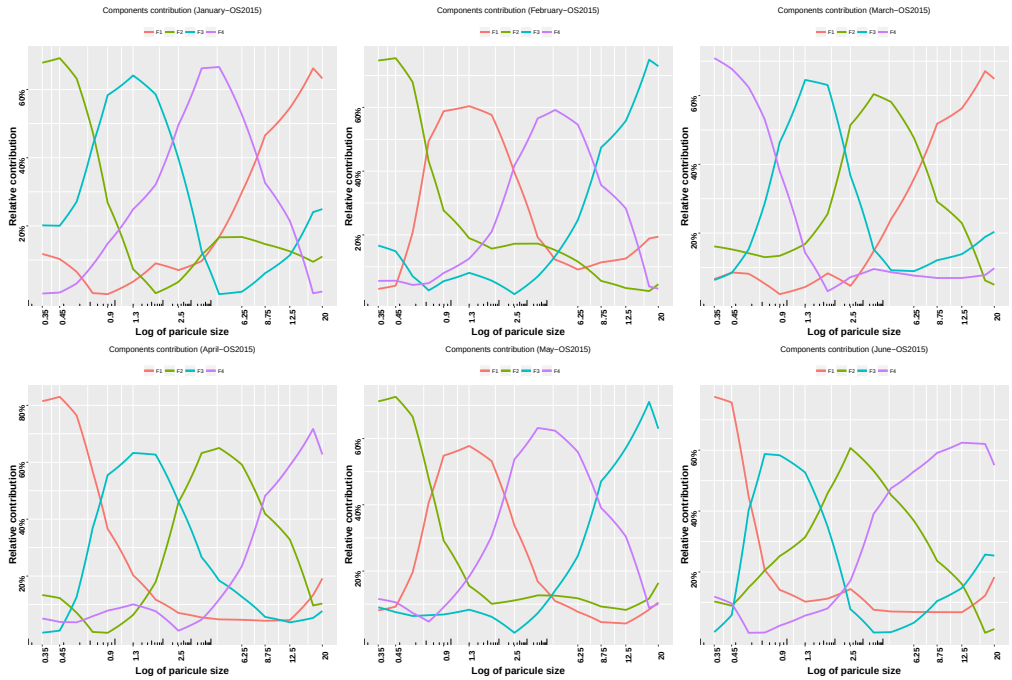


FIGURE 4.7.10 – *Contributions mensuelle des sources de fluctuations obtenues par la NMF sur les données des concentrations de particules mesurées pendant la campagne 2015 dans l'espace de bureaux.*

En vue d'examiner en détail l'origine de ces fluctuations, nous combinons les résultats de la factorisation avec les données d'occupation et d'ouverture. On présente sur la Figure 4.7.11, la variabilité polaire des quatre composantes NMF selon différents paramètres : la direction et la vitesse du vent, l'occupation et l'ouverture des fenêtres. On utilise la base complète de 6 mois.

La variabilité polaire de F1 illustre trois faits assez marquants :

1. les valeurs élevées sont observées uniquement lorsqu'au moins une fenêtre était ouverte, ceci suggère l'impact des sources extérieures sur les niveaux intérieurs ;
2. les paramètres du vent jouent un rôle prépondérant en l'absence d'occupants avec l'ouverture des fenêtres ;
3. l'occupation est très importante et réduit l'effet de la direction du vent.

En ce qui concerne la variabilité polaire de F2, les valeurs élevées sont indépendantes de tous les paramètres : elles se répandent sur pratiquement toutes les directions et pour tous les paramètres mis en jeu. Quant à la variabilité polaire de F3, il est très difficile d'identifier l'origine des niveaux élevés : une "bouffée" ponctuelle est observée durant l'inoccupation et lorsqu'au moins une fenêtre est ouverte. Enfin, il s'avère que la composante F4 dépend essentiellement de la direction du vent.

La Figure 4.7.12 montre l'évolution des fluctuations temporelles des composantes obtenues avec la méthode NMF. En ce qui concerne le profil-type, la première composante se distingue avec une courbe typique très marquée. La quatrième composante F4 semble corrélée avec la première en ayant une forme en cloche dans le profile journalier.

La variabilité de F2 et F3 semble suivre le même schéma de variation : une diminution durant les heures de journée et des valeurs élevées durant les heures de nuit. Pour ces facteurs, les variations sont indépendante du type de jour, alors que pour F1 et F4, aucune fluctuation significative n'apparaît durant le week-end.

Le profil de la première composante nous rappelle celui de variabilité du CO_2 observé durant toutes les campagnes dans les environnements occupés (hormis la maison expérimentale). En effet, une variabilité quasi-plate durant les week-ends et durant les périodes d'inoccupation et très fortes en période d'occupation.

4.7.3.2 Séparation des sources de concentration de formaldéhyde dans l'espace paysager : application de la NMF

Dans cette section, nous présentons les résultats obtenus par l'application de la méthode NMF sur les concentrations de HCHO intérieures et extérieures durant la campagne de 2015 dans l'espace de bureaux. Rappelons que durant la campagne 2013, l'instrument de mesure, l'analyseur AL4021 (Aerolaser GmBh), a été utilisé de manière continue en enregistrant une moyenne toutes les minutes mais pour un seul point de mesure. Pour les donnés de 2015, les mesures ont été effectuées de manière séquentielle toutes les 20 minutes par un module de multiplexage : une mesure à l'intérieur, 20 minutes après une mesure à l'extérieur, ensuite on revient à l'intérieur etc.

Afin de synchroniser les données sur une même base temporelle, les mesures extérieures ont été décalées de 20 minutes en arrière. Ce retard n'a pas été traité donc uniquement dans le cadre des mélanges instantanés. Évidemment, cette hypothèse est forte compte tenus de la fenêtre de délai, mais c'est une solution "acceptable" pour permettre la factorisation à partir des deux récepteurs (l'un à l'intérieur, et l'autre à l'extérieur).

Une factorisations par NMF ont été effectuées pour deux variables de HCHO (mesures intérieures et extérieures). Le nombre de facteurs a été fixé à 4 et les résultats de la décomposition sont présentés sur la Figure 4.7.13a. Les séries temporelles représentées sont obtenues en un facteur près.

Sur la Figure 4.7.13b, on présente les contributions des composantes obtenues de la factorisation simulation. Ce que l'on peut retenir comme résultat important, c'est l'existence d'une source spécifique aux concentrations intérieures de HCHO. Ainsi, la composante F3 dans les deux situations n'apparaît pas dans les profils de contributions pour HCHO extérieur (**HCHO_EXT**).

Ce résultat va dans le sens de tous les travaux de recherche mettant en exergue la particularité des environnements intérieures. Il rejoint donc les conclusions de la campagne logements effectuée par l'OQAI (Kirchner et al., 2007a,b; Mandin et al., 2009) : les concentrations de HCHO ont été observées dans tous les bâtiments et à des niveaux beaucoup plus supérieurs qu'à l'extérieur.

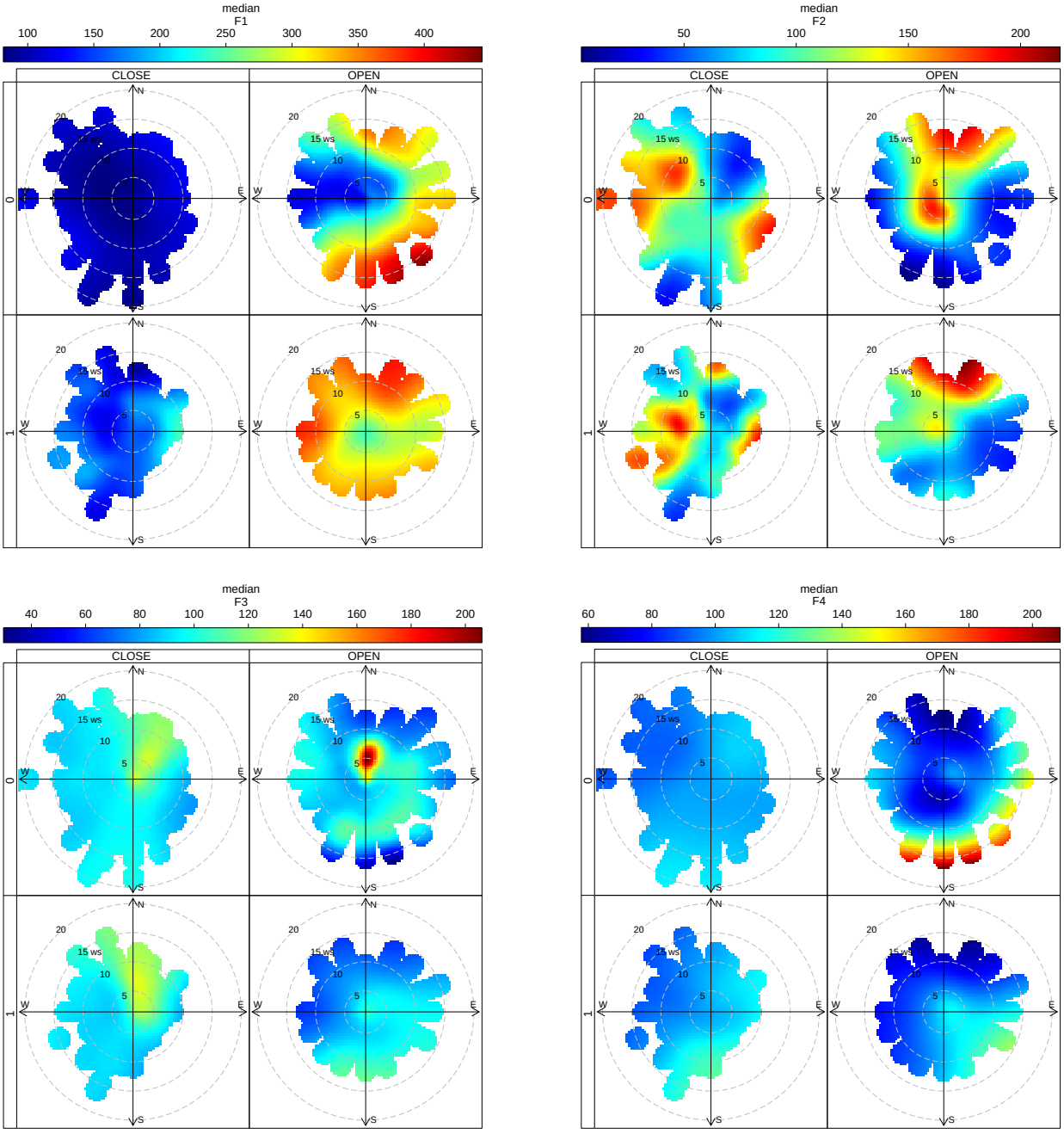


FIGURE 4.7.11 – *Variabilité polaire des sources/facteurs de fluctuations obtenus par la NMF selon l'occupation (0 inoccupation et 1 occupation) et l'ouverture des fenêtres (OPEN/CLOSE) par rapport aux paramètres du vents pour la base de données de particules, campagne 2015 avec un pas de temps d'une minute.*

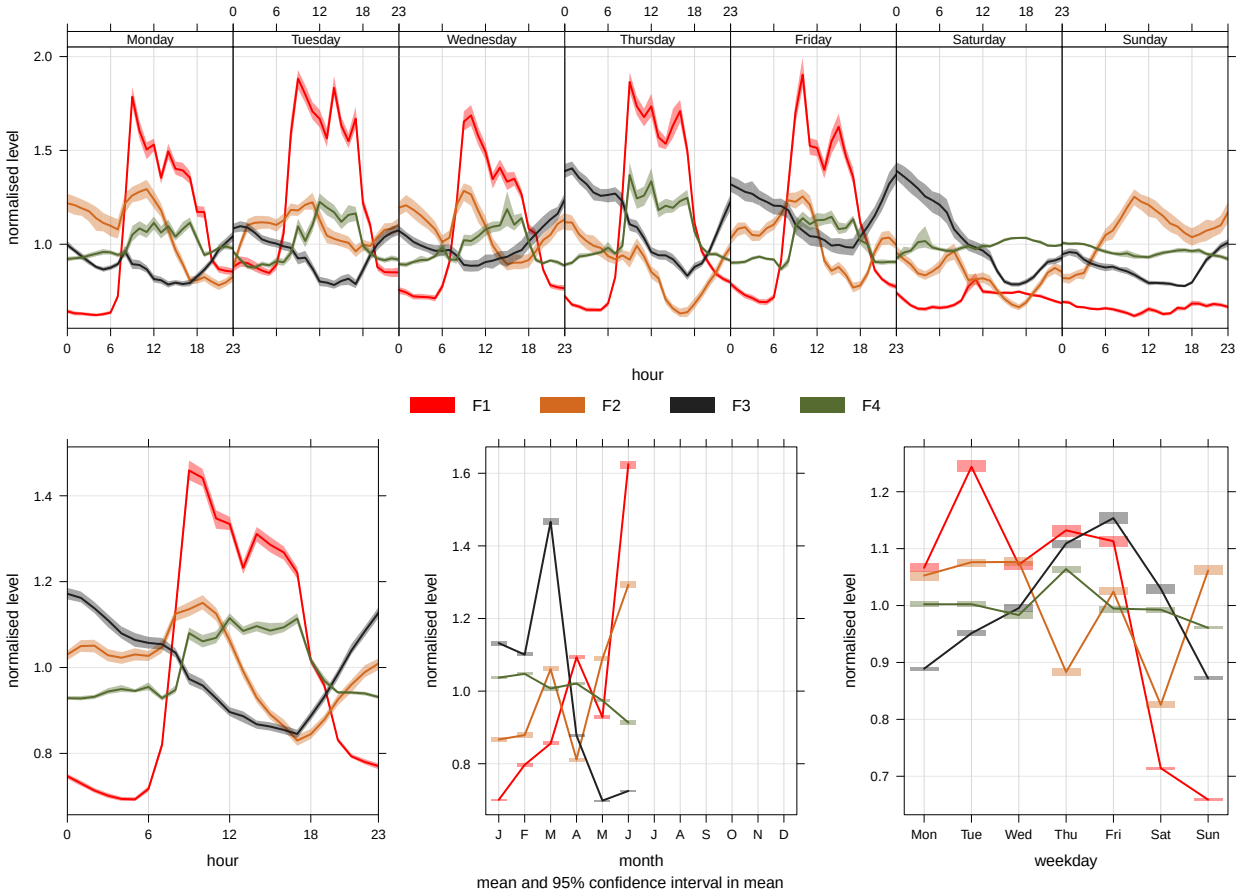
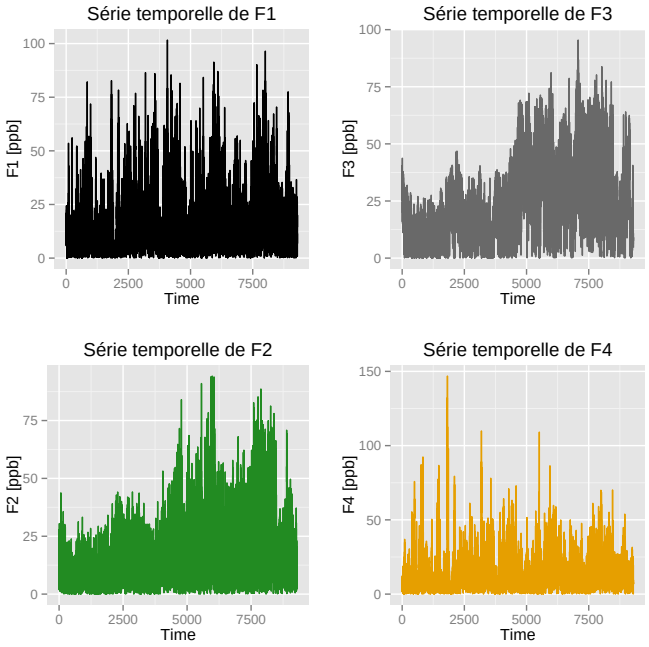
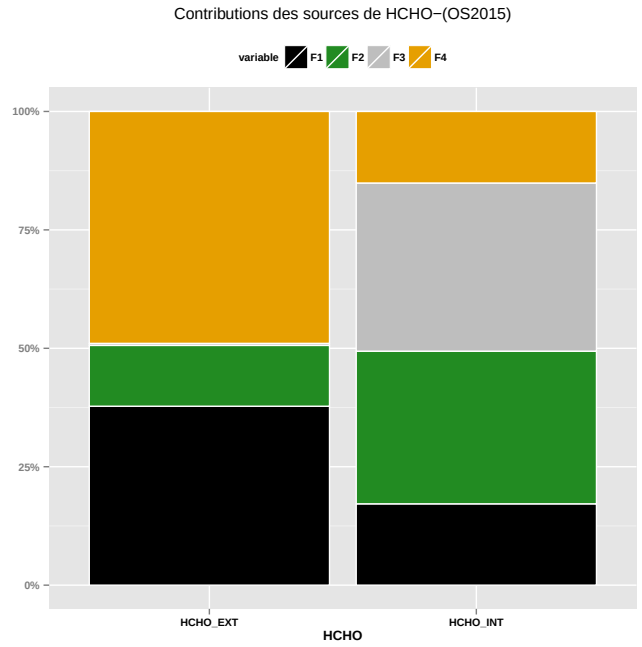


FIGURE 4.7.12 – *Évolution du profil moyen des quatre composantes obtenues avec la méthode NMF appliquée aux mesures des concentrations de particules de la campagne de 2015 (6 mois de mesures dans l'espace paysager). Les valeurs ont été normalisées par rapport à leurs moyennes respectives.*



(a) *Profiles de la variabilité temporelle des facteurs obtenus par la NMF sur les concentrations intérieures et extérieures de HCHO.*



(b) *Contributions des sources de HCHO par la méthode NMF appliquée aux mesures intérieures et extérieures de la campagne 2015.*

FIGURE 4.7.13 – *Profiles de la variabilité temporelle contributions des sources de HCHO par la méthode NMF appliquée aux mesures intérieures et extérieures issues de la campagne 2015. Les valeurs en ordonnée sont données à un facteur près : αF_i, $i = 1, \ldots, 4$.*

4.8 Discussion, conclusion et perspectives

Arrivée au terme de ce volet de la thèse, nous pouvons en faire plusieurs commentaires et en particulier, examiner dans quelle mesure nous pouvons désormais répondre à quelques questions sur les méthodes BSS appliquées aux données de la QAI.

Dans la première partie des résultats obtenus, nous avons tenté une approche consistant à caractériser les groupements des sources/facteurs par l'analyse des structures de la fonction d'autocorrélation. Le point de vue de cette intuition part du principe que la structure de la fonction d'autocorrélation de la source duplique, même de façon linéaire, les structures de dépendance au sein de la série observée. Grâce à cette proposition, deux typologies d'émission ont été mises en évidence : *(i)* la décroissance des autocorrélations de type fonction de puissance qui caractérise les émissions des sources déterministes (chroniques ou diffuses) et *(ii)* la décroissance de l'ACF de type exponentielle caractérise une variabilité aléatoire des sources. Cette dernière a été associée aux fractions de particules moyennes ayant un profil qui ressemble aux fluctuations de CO_2. Ceci traduit encore une fois le fait que l'occupation est le facteur le plus aléatoire, modifiant ainsi les structures régulières des séries. Ce résultat va dans le sens des observations évoquées tout au long de la première partie.

Ces premiers travaux montrent la similitude des différentes méthodes pour séparer à l'aveugle les sources contributives au niveau de particules observé dans un environnement intérieur. L'ACP classique n'offre pas le même confort d'interprétation que les autres méthodes.

Pour autant, séparer n'est pas identifier. Sans informations extérieures, les facteurs extraits restent difficilement identifiables autrement que par l'observation de fluctuations caractéristiques d'un processus particulier. Après leur extraction, une seconde phase d'exploitation doit donc être menée pour rechercher les associations avec d'autres phénomènes observés.

Ces travaux ont pris appui sur des mesures de particules en nombre relativement corrélées entre elles dont l'ACI, la NMF et la PMF ont su extraire les composantes indépendantes ou caractéristiques. Il est donc nécessaire de disposer de plusieurs enregistrements d'un même polluant pour que l'ACI soit utilisable. Néanmoins, nous nous posons la question de son utilisation sur un jeu de données hétérogènes, mais pour lequel nous suspectons des processus ou sources communes permettant d'expliquer les différentes variations observées. Ainsi, des variables exogènes comme la température, l'humidité, l'occupation et l'ouverture des fenêtres seraient autant de paramètres qui permettraient de se prononcer sur l'interprétation des composantes extraites. Pour l'instant, ces données sont utilisées *a posteriori* de l'analyse. L'analyse de la variabilité polaire offre un moyen très utile pour répondre à ces attentes, son utilisation couplée avec l'analyse NMF améliore l'interprétation des résultats.

Enfin, la séparation et la contribution des sources pour les mesures de formaldéhyde par NMF met en évidence une particularité très importante : pour un nombre de facteurs supérieur à 3, il ressort une composante sytématiquement inhérente aux concentrations intérieures. Ceci montre en définitif l'importance des sources intérieures dans la détermination des niveaux de HCHO.

CHAPITRE 5

PRÉVISION DES PARAMÈTRES ENVIRONNEMENTAUX : ÉTAT DE L'ART DES MODÈLES STATISTIQUES

La variabilité des mesures de la qualité de l'air présente des caractéristiques aléatoires, donc trop bruitée pour accéder à une description analytique débouchant sur une modélisation déterministe. Un ensemble d'approches a été élaboré dans la littérature afin de décrire au mieux le comportement des processus aléatoires à partir d'une série d'observations. L'une des plus importantes approches est la modélisation inverse à des fins de prévision. Ce chapitre a pour objectif de dresser un panorama général des méthodes d'analyse et de prévision des séries temporelles en sciences de l'environnement.

Sommaire

5.1 Introduction

Les tentatives d'expliquer les actions qui régissent l'atmosphère et la prévision du temps sont toutes de très anciennes activités qui ont accompagné l'Homme durant son évolution civilisationnelle. Mais c'est avec l'apparition des premiers instruments de mesure au $XVII^e$ siècle, la définition des grandeurs physiques fondamentales et des lois qui les régissent que la prévision météorologique s'est transformée en activité scientifique. L'une des premières tentatives d'utilisation des probabilités pour la prévision remonte aux communications de Cleveland Abbe en 1869. Mais c'est à partir des travaux pionniers de Lewis Fry Richardson (1922), qui entreprirent une première utilisation pratique de la mécanique des fluides dans le but de la prévision météorologique que la météorologie se transformait dès lors en une science à part entière (Treut, 2009).

Pour caractériser le phénomène de la pollution de l'air intérieur, on pourrait dire que les équations, généralement définies par un système d'équations différentielles sont connues et peuvent être obtenues avec des situations idéalisées. Cependant, ces conditions sont loin de la réalité physique, en particulier lorsqu'il est difficile d'avoir les mesures de certains paramètres. Dans la modélisation de la QAI par des modèles physico-chimiques, l'hypothèse d'homogénéité de l'air est souvent mise au devant de toutes les autres. Même si nous acceptons que les équations du modèle puissent être formulées dans des conditions idéales, il faudrait des mesures de toutes les variables d'état dans un volume très petit (point). En revanche, les mesures sont habituellement recueillies uniquement à des emplacements discrets. Farmer & Sidorowich (1987) ont fait valoir que les équations formulées pour des cas continus, comme le modèle photochimique, ne peuvent pas fournir des sorties fiables fondées sur des données discrètes.

Un besoin de développer une méthode basée sur une autre approche est donc fondamental. Le point de vue de la modélisation inverse s'avère prometteur pour appréhender ces problèmes. Nous considérons ici la modélisation inverse comme la procédure qui consiste à inférer sur les causes en partant des effets, ou à projeter la dynamique des observations sur un horizon futur (prévision).

Le problème de prévision se résume à une question fondamentale, a savoir comment à partir des observations $x_1, \ldots, x_T$, on peut avoir des valeurs futures $x_{T+1}, x_{T+2}, \ldots$ jugées "utiles" compte tenu du phénomène étudié. Pour des observations univariées, il s'agit de modéliser la dynamique à travers laquelle les valeurs passées influencent la valeur présente. Dans le cadre des processus stochastiques, les valeurs observées seront supposées être des réalisations des variables aléatoires $X_1, \ldots, X_T$ dont il faudra spécifier la dynamique. La prévision à la date $T+1$ consiste, d'un point de vue mathématique, à projeter X_{T+1} sur un ensemble de fonctionnelles de $X_1, \ldots, X_T$ (projection linéaire, espérance conditionnelle etc.), et à estimer l'intervalle de précision associé à cette projection.

La littérature en matière de prévision de la QAI est dans sa phase embryonnaire. Les besoins scientifiques de combler ce vide sont fondamentales. En revanche, la littérature de spécialité est bien fournie en ce qui concerne la prévision des niveaux de pollution extérieure. Il est nécessaire pour nous d'abord de définir le cadre mathématique dans lequel la prévision statistique et les modélisations des processus aléatoires sont fondées, ensuite, de répondre en quoi cette théorie apporte des réponses pour les cas concrets de la QAI.

Dans la section 5.4, un bilan des connaissances sur la modélisation statistique de la pollution atmosphérique est présenté et sera consacré essentiellement aux modèles de prévision des concentrations des polluants dans l'air extérieur.

5.2 La prévision linéaire des processus stationnaires

Au début de cette thèse, nous nous sommes positionnés d'abord dans le cadre "simple" de l'analyse des séries temporelles ; ensuite, nous avons intégré progressivement les différentes composantes liées aux structures inhérentes des séries de la QAI, correspondant à des situations plus complexes. Nous abordons donc la question de prévision dans ses aspects les plus classiques.

Avant de présenter le déroulement des procédures de prévision, plusieurs concepts de la statistique mathématique sont nécessaires pour la suite, en particulier les espaces de carrés intégrables L^2 et le théorème de projection. Cette thèse a d'abord un aspect pratique fondamental, qui est celui de la prévision de la QAI ; elle présente donc les approches de prévision sous cet angle. Néanmoins, il existe une pléiade de beaux résultats concernant la connexion entre la théorie de mesure, les processus stochastiques et la prévision (Feller, 1950; Neveu, 1965; Fuller, 2009).

Cette section est inspirée principalement de la littérature suivante :

- pour les fondements théoriques de la prévision, on se réfère aux livres d'Azencott & Dacunha-Castelle (1984) et de Brockwell & Davis (1991).
- pour la présentation des processus stationnaires, on se réfère aux monographies suivantes : Douc et al. (2014); Fuller (2009); Gourieroux & Monfort (1995).

On commence notre présentation par quelques définitions sur les fondements théoriques de la prévision.

Définition 5.2.1. (Espace $L^2(\Omega, \mathcal{A}, \mathbb{P})$)
Soit un espace de probabilité $(\Omega, \mathcal{A}, \mathbb{P})$ et une collection de toutes les variables aléatoires X définies sur Ω ; l'espace $L^2(\Omega, \mathcal{A}, \mathbb{P})$ (noté L^2) est dit espace des variables de carré intégrable si l'espérance satisfait à la condition

$$\mathbb{E}\left[X^2\right] = \int_{\Omega} X(\omega)^2 \, \mathbb{P}(d\omega) < \infty. \tag{5.2.1}$$

L'espace $L^2(\Omega, \mathcal{A}, \mathbb{P})$ est un espace de Hilbert, noté $\mathscr{H}$ (voir la démonstration dans (Brockwell & Davis, 1991), pages 46-48).

L'ensemble des variables aléatoires de carré intégrable $L^2(\Omega, \mathcal{A}, \mathbb{P})$, *i.e.* $\mathbb{E}\left[X^2\right] < \infty$, est un espace vectoriel normé sur $\mathbb{R}$, la norme étant $\|X\| = \sqrt{\mathbb{E}[X^2]}$. La norme utilisée est associée à un produit scalaire, ce qui permet de définir la notion d'orthogonalité dans L^2. Pour deux variables dans cet espace, X, Y, il est possible de calculer l'espérance de leur produit $\mathbb{E}[XY]$; l'application $\langle X, Y\rangle \to \mathbb{E}[XY]$ définit un produit scalaire sur l'espace L^2 ; elles sont dites orthogonales au sens de L^2, si $\mathbb{E}[XY] = 0$. Ces deux éléments, avec la notion de convergence permettent de définir la projection orthogonale.

Théorème 5.2.2. ***(de projection)*** *Si $\mathscr{M}$ est un sous-espace fermé d'un espace de Hilbert $\mathscr{H}$ et $x \in \mathscr{H}$, alors*

i) il existe un élément unique $\hat{x} \in \mathscr{M}$, tel que

$$\|x - \hat{x}\| = \inf_{y \in \mathscr{M}} \|x - y\| \tag{5.2.2}$$

et

ii) $\hat{x} \in \mathscr{M}$ et $\|x - \hat{x}\| = \inf_{y \in \mathscr{M}} \|x - y\|$ si et seulement si $\hat{x} \in \mathscr{M}$ et $(x - \hat{x}) \perp \mathscr{M}$.

L'élément $\hat{x}$ dans 5.2.2 est la projection de x sur $\mathscr{M}$. La démonstration de ce théorème est donnée dans (Brockwell & Davis, 1991). On note par $P_{\mathscr{M}}$ l'application qui associé à chaque x sa projection sur $\mathscr{M}$, et par $\mathscr{M}^{\perp}$ le complément orthogonal d'un sous-ensemble $\mathscr{M}$ de $\mathscr{H}$: tous les éléments de $\mathscr{H}$ qui sont orthogonaux à chaque élément de $\mathscr{M}$ (*i.e.* $x \in \mathscr{M}^{\perp}$ *ssi* $\langle x, y\rangle = 0$ *pour tout* $y \in \mathscr{M}$*).*

Définition 5.2.3. (Équation de prédition)
Soient l'espace de Hilbert $\mathscr{H}$, un sous-espace fermé $\mathscr{M}$ et un élément $x \in \mathscr{H}$. Le théorème 5.2.2 montre que l'élément de $\mathscr{M}$ le plus proche de x est l'unique élément $\hat{x} \in \mathscr{M}$ tel que

$$\langle x - \hat{x}, y\rangle = 0 \text{ pour tout } y \in \mathscr{M}. \tag{5.2.3}$$

Dans ce qui suit, nous considérons le problème de prévision des valeurs de $\{X_t,\, t \geq n,\}$ d'un processus stationnaire des termes $\{X_1, \ldots, X_n\}$. Le principe de prévision réside dans l'utilisation des observations jusqu'au moment n pour prévoir les valeurs futures de la série. Soit $\widehat{X}_{n+1}$, $n \geq 0$ la prévision à l'étape 1 définie par

$$\widehat{X}_{n+1} = \begin{cases} 0 & si \;\; n = 0 \\ P_{\mathscr{H}_n} X_{n+1} & si \;\; n \geq 1 \end{cases}$$

où $\mathscr{H}_n = \overline{\text{span}}\,\{X_1, \ldots, X_n\}$ est le sous-espace vectoriel fermé engendré par les combinaisons linéaires des $(X_i)_{i \leq n}$. Puisque $\widehat{X}_{n+1} \in \mathscr{H}_n$, alors on peut écrire

$$\widehat{X}_{n+1} = \sum_{j=1}^{n} \phi_{n,j} X_{n+1-j},\;\; n \geq 1, \tag{5.2.4}$$

où les coefficients d'autocorrélations partielles (chapitre 3) $\phi_{n,1}, \ldots, \phi_{n,n}$ satisfont à l'équation de "prédiction" 5.2.3 :

$$\left\langle \sum_{j=1}^{n} \phi_{n,j} X_{n+1-j}, X_{n+1-i} \right\rangle = \langle X_{n+1}, X_{n+1-i}\rangle\,, \;\; i = 1, \ldots, n,$$

avec $\langle X_{n+1},\, X_{n+1-i}\rangle = \mathbb{E}\left[X_{n+1} X_{n+1-i}\right]$ (le produit scalaire est défini dans $L^2(\Omega, \mathcal{A}, \mathbb{P})$).

Par linéarité du produit scalaire, ces équations peuvent être réécrites sous la forme

$$\sum_{j=1}^{n} \phi_{n,j} \gamma(j-i) = \gamma(i)\,, \;\; i = 1, \ldots, n, \tag{5.2.5}$$

ou de manière plus compacte, par

$$\mathbf{\Gamma_n \Phi_n} = \boldsymbol{\gamma_n} \tag{5.2.6}$$

où $\Phi_n = (\phi_{n,1}, \ldots, \phi_{n,n})^{\top}$, $\boldsymbol{\gamma_n} = (\gamma(1), \ldots, \gamma(n))^{\top}$ et $\mathbf{\Gamma_n} = [\gamma(j-i)]_{j,i=1}^{n}$. L'exposent $\top$ désigne le transposé. Le théorème de projection garantit que l'équation 5.2.6 a au moins une solution puisque $\widehat{X}_{n+1}$ doit être exprimée sous la forme de 5.2.4 pour $\boldsymbol{\Phi_n} \in \mathbb{R}^n$. Pour l'unicité de $\widehat{X}_{n+1}$, $\mathbf{\Gamma_n}$ doit être non-singulière. Dans ce cas, les solutions sont

$$\mathbf{\Phi_n} = \mathbf{\Gamma_n^{-1}}\boldsymbol{\gamma_n}.$$

Les conditions $\gamma(0) > 0$ et $\lim_{h\to\infty} \gamma(h) \to 0$ sont suffisantes pour que la matrice d'autocovariance $\mathbf{\Gamma_n} = [\gamma(j-i)]_{j,i=1}^{n}$ de $(X_1, \ldots, X_n)^\top$ soit non-singulière pour tout n (voir la démonstration dans Brockwell & Davis (1991), chapitre 5).

Il est alors possible de définir l'erreur quadratique moyenne de prévision (MSE)

$$MSE_n = \mathbb{E}\left[\left(X_{n+1} - \widehat{X}_{n+1}\right)^2\right], \ n \geq 1. \tag{5.2.7}$$

Pour $\phi_{1,1} = \gamma(1)/\gamma(0)$ et $MSE_0 = \gamma(0)$, l'algorithme de récurrence de Durbin-Levinson permet de calculer les coefficients $\phi_{n,j}$ et les MSE_n (les erreurs quadratiques moyennes) par les formules suivantes :

$$\phi_{n,n} = \tfrac{1}{MSE_n}\left[\gamma(n) - \sum_{j=1}^{n-1}\phi_{n-1,j}\gamma(n-j)\right] \tag{5.2.8}$$

$$MSE_n = MSE_{n-1}\left[1 - \phi_{n,n}^2\right]. \tag{5.2.9}$$

Dans la pratique, on utilise plus un autre algorithme que celui de Durbin-Levinson, c'est l'algorithme récursif sur les innovations. Ce dernier est plus général car la non-stationnarité du processus centré $\{X_t, t \in \mathbb{Z}\}$ n'est pas requise.

Proposition 5.2.4. ***(Algorithme des innovations)***

Soit un processus centré $\{X_t\}$ de fonction d'autocovariance $\kappa(i,j) = \mathbb{E}[X_i X_j]$, où la matrice $[\kappa(i,j)]_{i,j=1}^{n}$ est non-singulière pour tout $n \geq 1$. Alors les prévisions $\widehat{X}_{n+1}$, $n \geq 0$ à un un pas en avant (one-step predictors) et les erreurs quadratiques moyennes MSE_n, $n \geq 1$ sont exprimées par

$$\widehat{X}_{n+1} = \begin{cases} 0 & si \ \ n = 0, \\ \sum_{j=1}^{n}\theta_{nj}\left(X_{n+1-j} - \widehat{X}_{n+1-j}\right) & si \ \ n \geq 1, \end{cases} \tag{5.2.10}$$

et

$$\begin{cases} MSE_0 = \kappa(1,1), \\ \theta_{n,n-k} = \tfrac{1}{MSE_k}\left[\kappa(n+1,k+1) - \sum_{j=0}^{k-1}\theta_{k,k-1}\theta_{n,n-j}MSE_j\right], \ k = 0, 1, \ldots, k-1, \\ MSE_n = \kappa(n+1,n+1) - \sum_{j=0}^{n-1}\theta_{n,n-j}^2 MSE_j. \end{cases} \tag{5.2.11}$$

Alors que l'algorithme de Durbin-Levinson donne les coefficients de $X_1, \ldots, X_n$ dans la représentation 5.2.4, la proposition 5.2.4 fournit les coefficients des "innovations" $\left[X_j - \widehat{X}_j\right]_{j=1}^{n}$ sous leur forme orthogonale $\widehat{X}_{n+1} = \sum_{j=1}^{n}\theta_{nj}\left(X_{n+1-j} - \widehat{X}_{n+1-j}\right)$. Cette forme s'avère très utile dans la pratique, surtout pour les processus $ARMA(p,q)$.

5.3 Modèles linéaires des séries temporelles

5.3.1 Considérations théoriques

George Box et Gwilym Jenkins ont contribué, dans les années 70, à populariser la théorie des séries temporelles dans leur célèbre ouvrage *"Time Series Analysis : Forecasting and Control"* (Box & Jenkins, 1970).

Avant de présenter la démarche proposée dans (Box & Jenkins, 1970), nous donnerons quelques définitions et résultats sur les processus $ARMA$ (AutoRegressive–Moving-Average). Ce modèle est le résultat de la combinaison de deux processus aléatoires : les processus autorégressif (AR) et les processus moyenne mobile (MA).

Définition 5.3.1. Les processus autorégressifs moyennes mobiles ($ARMA$)
Un processus stationnaire X admet une représentation $ARMA(p,q)$ s'il satisfait :

$$X_t + \varphi_1 X_{t-1} + \varphi_2 X_{t-1} + \cdots + \varphi_p X_{t-p} = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} \quad (5.3.1)$$

$$\Phi(L) X_t = \Theta(L) \varepsilon_t \quad (5.3.2)$$

où

- *i)* l'opérateur retard L tel que $LX_t = X_{t-1}$ est linéaire et inversible,
- *ii)* $\varphi_p \neq 0$ et $\theta_q \neq 0$,
- *iii)* les polynômes Φ et Θ ont leurs racines de modules strictement supérieurs à 1,
- *iv)* Φ et Θ n'ont pas de racines communes,
- *v)* $\varepsilon = (\varepsilon_t,\, t \in \mathbb{Z})$ est un bruit blanc, de variance $\sigma^2 \neq 0$.

Un processus $ARMA(p,q)$ est stationnaire si toutes les racines de l'équation $\Phi(L) X_t = 0$ sont à l'extérieur du disque unité. De même pour que le processus soit inversible, il faut que toutes les racines de l'équation $\Theta(L) \varepsilon_t = 0$ soient à l'extérieur du disque unité.

Propriétés : *Si X est un processus stationnaire de représentation $ARMA(p,q)$:*

$$\Phi(L) X_t = \Theta(L) \varepsilon_t$$

- *i) X admet la représentation $MA(\infty)$:*

$$X_t = \frac{\Theta(L)}{\Phi(L)} \varepsilon_t = \Psi(L) \varepsilon_t = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j},\ \psi_0 = 1.$$

- *ii) X admet la représentation $AR(\infty)$:*

$$\varepsilon_t = \frac{\Phi(L)}{\Theta(L)} X_t = \sum_{j=0}^{\infty} \pi_j X_{t-j},\ \pi_0 = 1.$$

La modélisation univariée et la démarche générale de prévision par la procédure de Box-Jenkins sont fondées sur les processus $ARIMA\,(p,d,q)$. La procédure proposée est en quatre étapes :

- **Identification** : on utilise généralement l'estimation de la fonction d'autocorrélation (ACF) et de la fonction d'autocorrélation partielle (PACF). En revanche, il existe d'autres méthodes d'identification développées à partir des critères d'entropie. Citons, entre autre le critère AIC d'Akaike (1974) (Akaike's Information Criterion), le critère BIC de Schwarz et al. (1978) (Bayesian Information Criterion) et le critère CAT de Parzen (1975) (Criterion Autoregressive Transfer Function) ;
- **Estimation** : différentes procédures itératives peuvent être appliquées pour l'estimation des paramètres du processus $ARMA$. Les plus utilisées pour l'optimisation du critère des moindres carrés sont les méthodes de type Gauss-Newton ou le "compromis de Marquardt" ;
- **Tests de diagnostic** : il s'agit d'examiner si le modèle estimé est ou non compatible avec les hypothèses sous-jacentes aux modèles. Il en existe plusieurs : des tests graphiques de l'autocorrélation des résidus et des tests basés sur la statistique de χ^2 tel que le test de Ljung-Box ou celui de Box-Pièrce ;
- **Prévisions** : consiste à respecter le principe d'extrapolation en utilisant les formules de la prévision optimale pour les processus ARMA (voir ci-après).

Dans le cadre de l'analyse des structures de variabilité de la concentration en polluants dans l'air intérieur (*cf.* Chapitre 2 et 3), l'hypothèse de stationnarité n'est pas respectée. Donc, la prévision par un modèle de type ARMA sur ces séries conduit dans la plupart des cas à un échec. En revanche, si l'on considère une série de transformation des données, par exemple, désaisonnalisation du logarithme de la série puis différentiation, l'hypothèse de stationnarité devient plus vraisemblable.

Dans nos applications, il s'agit d'une procédure de décomposition de la série combinée avec une modélisation classique d'un modèle linéaire ou non linéaire. En effet, rappelons qu'à l'origine dans la modélisation des phénomènes aléatoires non-stationnaires, il a été suggéré d'écrire $Y_n = f(n) + X_n$ où X_n est stationnaire et $f\,(n)$ est une fonction déterministe à préciser [1]. Pour nous, l'extraction d'une saisonnalité, par différentes méthodes est un bon moyen d'estimer $f\,(n)$. En outre, la prévision de cette composante consiste à projeter par répétition de la dynamique de quelques fréquences principales, sur les horizons futurs. Dans ce cas, il est naturel de supposer que l'erreur de prévision sur la composante fréquentielle est nulle. Le résultat de la désaisonnalisation (la série désaisonnalisée) est ensuite modélisé par, une approche type ARMA, lissage exponentiel,....

Notons que les différents types de transformations dépendent énormément de la nature du polluant, de la présence des variations abruptes, ainsi que de la résolution temporelle (pas de temps).

On peut donc considérer que le résultat de ces transformations peut satisfaire une représentation ARMA. Si l'on s'intéresse uniquement à la différentiation première, ou plus généralement les différences d'ordre d et en adoptant le point de vue pratique de Box & Jenkins (1976), alors on a :

$$\triangle^d X_t = (1-L)^d\, X_t, \tag{5.3.3}$$

et on aboutit à la forme générale des processus autorégressifs moyennes mobiles intégrés ($ARIMA$) de type

$$\Phi\,(L)\, X_t = \Theta\,(L)\, \varepsilon_t, \tag{5.3.4}$$

1. Souvent prise comme somme d'un polynôme en n et de combinaisons linéaires de $\sin(n\lambda_j)$ et $\cos(n\lambda_j)$, $j = 1, 2, \ldots, k$

avec $\Phi(L) = \varphi(L)(1-L)^d$. Comme le font remarquer Gourieroux & Monfort (1995), la relation 5.3.4 doit satisfaire à une condition sur le mécanisme de démarrage : il faut que les conditions initiales soient non corrélées avec les valeurs futures du bruit. En effet, on ne peut pas supposer que la relation 5.3.4 est vraie pour tout indice t et en déduire X_t par inversion de $\Phi(L)$, car la série en ε_t obtenue est divergente (Gourieroux & Monfort, 1995).

Définition 5.3.2. Les processus autorégressifs moyennes mobiles intégrés ($ARIMA$)
Un processus $X = (X_t,\, t \geq 0)$ centré est un processus $ARIMA(p,d,q)$ s'il satisfait une équation de type :

$$\varphi(L)(1-L)^d X_t \quad = \quad \Theta(L)\varepsilon_t \quad t \geq 0 \tag{5.3.5}$$

où :

$$\begin{aligned} \varphi(L) &= 1 + \varphi_1 L + \cdots + \varphi_p L^p, \quad \varphi_p \neq 0, \\ \Theta(L) &= 1 + \theta_1 L + \cdots + \theta_p L^p, \quad \theta \neq 0, \end{aligned}$$

sont des polynômes dont les racines sont de module supérieur à 1, avec des conditions initiales $Z_{-1} = \left(X_{-1}, \ldots, X_{-(p+d)}, \varepsilon_{-1}, \ldots, \varepsilon_{-q}\right)^\top$ non corrélées avec $\varepsilon_0, \varepsilon_1, \ldots, \varepsilon_t, \ldots$ et le processus $\varepsilon = (\varepsilon_t,\, t \geq -q)$ est un bruit blanc.

5.3.2 Prévision dans les modèles ARIMA

Dans cette section, nous présentons le calcul récursif pour obtenir les prévisions optimales dans le cadre d'un modèle linéaire ARIMA, leurs intervalles de prévisions, ainsi les conséquences d'un prétraitement de type transformation logarithmique sur la formule de prévision.

Considérons un processus $ARMA(p,q)$ et reprenons la forme moyenne mobile infinie :

$$X_t = \sum_{j \geq 0} \psi_j \varepsilon_{t-j},\ \psi_0 = 1, \tag{5.3.6}$$

et notons par $\widehat{X_t}(h)$ la prévision de X_{t+h}, $h > 0$ faite au moment t pour un horizon de prévision h, et les $\widehat{e_t}(h) = X_{t+h} - \widehat{X_t}(h)$ seront considérées comme les erreurs de la prévision pour le moment $t+h$. Par définition, la prévision linéaire est donnée par l'espérance conditionnelle :

$$\widehat{X_t}(h) = \mathbb{E}\left[X_{t+h} \mid \mathcal{F}_t\right] \tag{5.3.7}$$

où $\mathcal{F}_t$ est l'ensemble d'informations disponibles au moment t, soit $\mathcal{F}_t = (X_1, X_2, \ldots, X_t, \varepsilon_1, \varepsilon_2, \ldots, \varepsilon_1)$. L'expression générale de la prévision d'un processus $ARMA$ s'écrit :

$$\widehat{X_t}(h) = \sum_{j \geq 0} \psi_{h+j} \varepsilon_{t-j}, \tag{5.3.8}$$

et l'erreur de prévision est :

$$\widehat{e_t}(h) = X_{t+h} - \widehat{X_t}(h) = \sum_{j=0}^{h-1} \psi_j \varepsilon_{t+h-j}, \text{ avec } \psi_0 = 1. \tag{5.3.9}$$

Sous l'hypothèse de l'indépendance des ε_t, la variance des erreurs de prévision est donnée par l'expression :

$$\mathbb{E}\left[\widehat{e_t}(h)^2\right] = \sigma_\varepsilon^2 \sum_{j=0}^{h-1} \psi_j^2. \tag{5.3.10}$$

Si de plus, les ε_t sont normaux, alors l'erreur de prévision est Gaussienne

$$\widehat{e_t}(h) \sim \mathcal{N}\left(0, \sigma_\varepsilon^2 \sum_{j=0}^{h-1} \psi_j^2\right), \tag{5.3.11}$$

et l'intervalle de prévision au seuil α, s'écrit

$$\widehat{X_t}(h) \pm u_{1-\frac{\alpha}{2}} \widehat{\sigma}_\varepsilon \sqrt{\sum_{j=0}^{h-1} \widehat{\psi}_j^2} \tag{5.3.12}$$

où $u_{1-\frac{\alpha}{2}}$ est le quantile d'ordre $\left(1-\frac{\alpha}{2}\right)$ de la loi normale centrée réduite, $\widehat{\sigma}_\varepsilon$ et $\widehat{\psi}_j$ sont des estimateurs consistants de σ_ε et ψ_j, respectivement.

Nous avons vu dans les chapitres précédents que la plupart des séries de la QAI possèdent des distributions de probabilité de type log-normales. Une forte variabilité des processus naturels, se manifestant par un étalement à droite de la distribution, est à l'origine de la non-normalité (Parkin & Robinson, 1992). En outre, les mesures de la qualité de l'air ne peuvent pas être négatives, par conséquent la plupart des données se concentrent au niveau des 25 centiles, ce qui les rend plus fréquentes à gauche. La non-stationnarité peut aussi être à l'origine de la positivité du coefficient d'asymétrie, en particulier la non-stationnarité en variance. On doit donc tout d'abord transformer les séries d'origine afin de neutraliser une partie de la dynamique de la variance, soit $Y_t = \ln(X_t)$. Les prévisions seront exprimées en logarithme, mais si on veut les prévisions de la série d'origine, il ne suffit pas de calculer les exponentielles. En effet, soit $Y_t = \ln(X_t)$ un processus $ARMA$, la prévision naïve consiste à dire que $\widehat{X_t^{naïve}}(h) = \exp\left(\widehat{Y_t}(h)\right)$. En revanche, en vertu de l'inégalité de Jensen, on a

$$\begin{aligned} \widehat{X_t}(h) &= \mathbb{E}\left[\exp\left(\ln\left(X_{t+h}\right) \mid X_t, X_{t-1}, \ldots\right)\right] \\ &> \exp\left(\mathbb{E}\left[\ln\left(X_{t+h}\right) \mid X_t, X_{t-1}, \ldots\right]\right) = \widehat{X_t^{naïve}}(h) \end{aligned} \tag{5.3.13}$$

ce qui impose de tenir compte de la variance de l'erreur de prévision de la série Y_t :

$$\widehat{X_t}(h) = \exp\left(\widehat{Y_t}(h) + \frac{\sigma_\varepsilon^2}{2} \sum_{j=0}^{h-1} \psi_j^2\right). \tag{5.3.14}$$

5.4 Bibliographie sur les applications des modèles statistiques pour la prévision des concentrations des polluants dans l'air

Cette section propose un état de l'art sur l'utilisation des modèles de séries temporelles pour la prévision des concentrations des polluants dans l'air.

5.4.1 Application des modèles linéaires

Après le succès de la méthode Box-Jenkins dans la communauté scientifique, ces modèles ont été très vite appliqués pour la prévision de la qualité de l'air, notamment pour l'air extérieur. Durant les années 80, on trouve par exemple des applications de cette méthode pour la prévision des niveaux de l'ozone (Simpson & Layton, 1983; Robeson & Steyn, 1990), de dioxyde de soufre (Finz et al., 1980) et des oxydes d'azote (Sawaragi et al., 1979; Inoue et al., 1986). Depuis, la méthodologie de Box-Jenkins a été largement utilisée pour la prévision d'autres paramètres. En effet, durant les années 90, les travaux de recherche se sont penchés sur la prévision des niveaux d'ozone et des NO_x. Durant les années 2000, les études concernant la prévision des concentrations des particules $PM_{2.5}$ et PM_{10} ont été plus documentées et elles suscitent toujours l'intérêt des chercheurs.

Dans l'étude menée par Kumar et al. (2004), un modèle $ARIMA\,(1,0,1)$ a été estimé et appliqué pour la prévision des concentrations maximales journlières d'ozone. La performance de ce modèle pour un critère de MAPE est de 13.14% pour une prévision d'un jour à l'avance.

Souvent, dans la littérature, l'analyse des niveaux de polluants atmosphériques est présentée en intégrant les informations sur les conditions climatiques. Jian et al. (2012) ont étudié la contribution des facteurs météorologiques sur la variabilité des niveaux de particules utrafines (UFP). Les auteurs ont utilisé un modèle SARIMA (Seasonal ARIMA) pour modéliser la trajectoire des accroissements des UFP en appliquant le filtre de différentiation. Pour l'étape de validation des modèles, plusieurs tests statistiques ont été appliqués afin de vérifier la structure des résidus issus de la régression, notamment, la statistique de Ljung-Box. L'adéquation du modèle est mesurée par le coefficient de détermination R^2. Les résultats de cette étude montrent que la température, l'humidité relative, la vitesse du vent et la pression sont des prédicteurs significatifs pour la prévision des concentrations UFP et PM_1. En revanche, les précipitations et la direction du vent n'ont pas d'impact significatif sur le niveau de variabilité des polluants étudiés. Shi & Harrison (1997) rapportent que la vitesse du vent est un facteur très important influençant les prévisions des concentrations des NO_x et NO par un modèle $AR\,(1)$.

Lee et al. (2012) ont étudié un modèle saisonnier des séries des concentrations de l'indice de pollution de l'air (air pollution index **API**) ; cet indice est basé sur cinq polluants (PM_{10}, SO_2, NO_2, CO et O_3). Pour l'identification des paramètres des modèles ARIMA candidats, les fonctions d'autocorrélation et d'autocorrélation partielle ont été utilisées. Pour l'évaluation des performances du modèle en prévision, plusieurs critères ont été utilisés, MAE, MAPE, MSE et RMSE. En se basant sur le critère MAPE pour les prévisions, les résultats sont inférieurs à 20 % pour les différents modèles utilisés. L'indice de la pollution de l'air a été largement analysé à des fins de prévision en utilisant la méthodologie classique de Box-Jenkins. Par exemple, Wang & Lu (2006) ont utilisé un modèle ARMA pour la prévision de la moyenne journalière de l'**API**, la série étudiée semble stationnaire et aucun test statistique n'a été rapporté dans cette étude. La prévision obtenue par deux modèles $ARMA$, ($ARMA\,(1,1)$ pour l'été et $ARMA\,(2,2)$ pour les autres saisons) donne de bonnes prévisions. El Raey et al. (2006) rapportent les mêmes ordres de paramètres de ARMA (*i.e.* de 0 à 2) dans une étude sur la prévision de l'**API** en Égypte.

TABLE 5.4.1 – Analyse des séries temporelles linéaires pour la prévision des concentrations des polluants dans l'air extérieur : synthèse bibliographique (exemples).

Polluant	Transformation	Modèle	Identification	Validation ε_t	Auteurs
O_3	$\triangle X_t$	$ARIMA\,(1,1,1)$	ACF, PACF	ACF	Slini et al. (2002)
PM	$\triangle X_t$	$SARIMA$	-		Sami et al. (2012)
O_3	-	$ARIMA\,(1,1)$	ACF, PACF	ACF et χ^2	Kumar et al. (2004)
O_3, NO NO_2, CO	$\triangle X_t$ $\log X_t$, $\triangle \log X_t$ $\triangle\sqrt{X_t}$ $\sqrt{X_t}$	$ARIMA\,(p,1,q)$ $p=0-3$, $q=1-3$	ACF, PACF, AIC, BIC FPE HIC	ACF	Kumar & Jain (2010)
UFP PM_{10}	$\triangle X_t$	$ARIMA(0,1,1)$ $ARIMA(0,1,0)$		Ljung Box	Jian et al. (2012)

Note : les critères HIC (Hannon–Quinn Information Criterion) et FPE (Final Prediction Error) sont largement utilisés dans la littérature, mais les résultats entre les différents critères peuvent s'avérer contradictoires et la décision sur le choix du modèle est problématique. Dans l'étude de Kumar & Jain (2010), les différents types de transformations ont été utilisés pour tous les polluants.

La Tableau 5.4.1 présente quelques travaux concernant l'analyse des séries temporelles linéaires pour la prévision des concentrations des polluants en air extérieur. Le pas de temps dans la plupart des séries de mesures analysées dans la littérature sont de l'ordre de 1 h à 1 jour. Ce type de données influence le type de modèle à utiliser. En effet, souvent les ordres des paramètres des modèles linéaires de type ARIMA ne dépassent pas 3.

On remarque aussi que dans l'étape d'identification et de validation, très peu d'études utilisent les tests habituels pour l'analyse de la stationnarité et de la structure des résidus. Souvent, l'analyse des résidus est effectuée avec les fonctions ACF et PACF ; par exemple dans l'étude menée par Durdu (2010), le modèle ARIMA retenu met en évidence la fidélité des hypothèses classiques de normalité et d'homoscédasticité.

L'analyse spectrale est souvent utilisée comme étape de prétraitement des données, les périodicités des séries sont aussi mises en évidence par l'analyse des profils diurnes ou mensuels. Avec ces deux techniques, Vingarzan & Taylor (2003) estiment que le maximum de l'ozone journalier possède une autocorrélation dans les retards 1 et 2.

Enfin, pour une revue de littérature sur l'application des séries temporelles aux problèmes de la pollution atmosphérique, on peut consulter (Milionis & Davies, 1994; Salcedo et al., 1999).

5.4.2 Application des modèles non-linéaires

Le phénomène de la pollution est complexe et de nature non-linéaire (Raga & Le Moyne, 1996). Plusieurs auteurs font l'éloge de ces modèles et rapportent leurs avantages pour la prévision (Chen et al., 1998; Weng et al., 2008).

La prise en compte de la "non-linéarité" a modifié profondément les approches de prévision, tant au niveau théorique que pratique. Cette modification a fait irruption dans tous les domaines des sciences et l'analyse de la qualité de l'air dans l'environnement n'échappe pas à ces modifications.

Deux classes de modèles paramétriques sont largement développées dans la littérature statistique des séries temporelles : modèles non-linéaires en moyenne et modèles non-linéaires en variance. Rarement les modèles non-linéaires en variance, comme par exemple la famille GARCH, sont appliqués dans le domaine de l'environnement (Kumar & De Ridder, 2010; Tol, 1996). Ceci est probablement dû aux faits que l'interprétation de ces modèles dans le domaine de l'environnement reste à préciser, ou bien simplement, ce type de modèles ne sont pas adaptés aux problématiques liées à l'environnement.

Par ailleurs, les modèles non-paramétriques développés pour la qualité de l'air sont largement discutés dans la littérature. Par exemple, Chen et al. (1997) utilisent plusieurs modèles paramétriques de type ARIMA et non-paramétriques de type régression splines multivariée (Multivariate Adaptive Regression Splines MARS). Les auteurs ne recommandent pas forcément l'utilisation de l'un ou de l'autre, mais en fonction du type de données, préconisent l'amélioration de chaque modèle pour chaque série.

La suite de cette section reprend quelques travaux de la prévision des concentrations de polluants et discute à la fin le positionnement de nos choix concernant leur application à la QAI.

5.4.2.1 Modèles Autorégressifs non-linéaires

La classe des modèles autoregressifs non-linéaires est une extension directe des modèles ARMA, notamment les modèles autorégressifs à seuil (TAR).

Kim & Kumar (2005) ont étudié deux modèles non-linéaires pour la prévision des concentrations de l'ozone. Les deux modèles utilisés sont Functional ARX (FARX) et les modèles TARX (à seuil avec des entrées exogènes). Les variables exogènes sont la température, la vitesse et la direction du vent. Pour les mêmes performances en prévision, FARX nécessite moins de prédicteurs que le modèle TARX. Néanmoins, les modèles à variables exogènes, tels que présentés dans cette étude, nécessitent la connaissance des valeurs futures des variables exogènes pour la prévision sur plusieurs horizons (multi-step forecast). Pisoni et al. (2009) proposent l'utilisation des modèles NARX (réseaux de neurones autorégressifs à entrées exogènes) pour l'estimation des pics de pollution en ozone. Cette étude est plus orientée sur l'analyse des performances du modèle NARX et l'influence de la pondération dans les réseaux de neurones autorégressifs.

La détection et la modélisation des changements abruptes de la variabilité des séries temporelles environnementales sont rarement mises en évidence dans la littérature. C'est pour cela que Fassò & Negri (2002a) ont proposé le modèle SFI-SETARX-ARCH (Seasonal Fractionally Integrated Self-Exciting AutoRegressive processes avec des variables eXogenous et des erreurs de type AutoRegressive Conditionally Heteroscedastic) pour la prévision des concentrations de l'ozone avec la méthode de Monte-Carlo (MC). Globalement, les prévisions fournies par ce modèle sont très bonnes pour un horizon de 20 h (R^2 varie entre 0.9 à 0.6), mais demeurent acceptables jusqu'à 50 h (R^2 varie entre 0.4 à 0.5).

Fassò & Negri (2002b) ont proposé le modèle SFI–ARX–ARCH pour la caractérisation des séries temporelles des concentrations de l'ozone à haute fréquence. Ce modèle a permis d'identifier les structures de variabilité rarement évoquées dans les autres travaux, en l'occurrence : l'héteroscédacticité des erreurs, la mémoire longue, la saisonnalité et la non-linéarité.

En résumé, les modèles autorégressifs non-linéaires de type NARX, NARIMA, SETAR,.. ont été très peu utilisés pour la prévision des niveaux de polluants jusqu'à présent.

5.4.2.2 Réseaux de neurones artificiels

Les applications des réseaux de neurones artificiels (ANN) sur les données de la qualité de l'air extérieur ont fait coulé beaucoup d'encre. La plupart des chercheurs dans ce domaine s'accordent sur la supériorité des ANN en prévision des concentrations des polluants par rapport aux modèles linéaires. Plusieurs états de l'art ont été réalisés à ce sujet. Par exemple, Gardner & Dorling (1998) ont présenté les principales propriétés des réseaux de type perceptron multicouche (MLP) ainsi que les différentes applications de ces modèles en environnement. Néanmoins, la plupart des applications présentées ne concernent pas directement la prévision, mais plutôt les prédictions. Récemment, l'étude de Pasero & Mesin (2010) a mis l'accent sur les différents types de ANN (MLP, SVM, ...) appliqués à l'environnement ainsi que la prévision des concentrations des polluants, en particulier de l'ozone.

Dans (Krasnopolsky & Chevallier, 2003a,b), les auteurs analysent l'utilisation des réseaux de neurones comme approche des méthodes inverses en sciences environnementales. Récemment Hanrahan (2011) ont publié une monographie traitant le développement des réseaux de neurones ainsi que leur application dans les sciences de la biologie et l'environnement. Les questions relatives à la prévision dans ces différents états de l'art sont très sommaires.

Souvent, les modèles ANN ont été appliqués comme support d'un autre modèle des séries temporelles. Díaz-Robles et al. (2008) ont proposé une combinaison d'un modèle ARIMA et les ANN pour la prévision des niveaux de PM. La méthodologie développée consiste en un premier temps à la modélisation par ARMAX (paramètres climatiques comme variables exogènes), ensuite un modèle neuronal par ANN a été appliqué sur les résidus de sortie. L'indice de succès obtenu pour la prévision de la valeur maximale sur les 24 heures suivantes a été de 87 %.

Brunelli et al. (2007) ont conçu un réseau neural récursif d'Elman pour la prévision à 48 heures de la concentration maximale journalière de SO_2, O_3,PM_{10}, NO_2, et de CO dans la ville de Palerme (Italie), utilisant comme prédicteurs météorologiques : la vitesse et la direction du vent, la pression atmosphérique et la température ambiante, moyennées sur les 12 heures précédentes. Le coefficient de corrélation obtenu entre les valeurs prédites et celles enregistrées varie entre 0.72 et 0.97, pour les différents polluants testés, montrant une bonne performance du modèle proposé.

Ce succès est probablement dû au fait que les réseaux de neurones sont très adaptés aux données environnementales, données dans lesquelles la non-linéarité est la principale caractéristique (Raga & Le Moyne, 1996) et par définition, un neurone artificiel réalise dans un réseau, une fonction non-linéaire, paramétrée par ses variables (Dreyfus et al., 2011).

En dépit de leur grand potentiel en modélisation, les ANN échouent dans les tests de prévision pour quelques types de séries temporelles. Leur statut d'approximateurs universels (Hornik et al., 1989, 1990) s'applique uniquement dans la phase d'apprentissage. En effet, un modèle linéaire de type ARIMA peut s'avérer plus performant en prévision que les ANN (Crone et al., 2005). Dans une étude de comparaison entre un modèle linéaire autorégressif, un modèle STAR (Smooth Transition AutoRegressive) et un modèle de réseaux de neurones menée par Teräsvirta et al. (2005), les ANN présentent quelques problèmes dans les performances en prévision ; sans aucune contrainte, la prévision peut diverger.

5.4.2.3 Systèmes dynamiques et chaos

Les modèles sous-jacents s'appuient sur l'idée selon laquelle, un système complexe comme l'environnement peut être appréhendé de manière simple en analysant les structures de récurrences (similitudes)

dans des séries temporelles. La prévision est alors possible en reproduisant les comportements récurrents des séries. Le chapitre 7 sera dédié à la description de ces modèles.

L'approche des systèmes dynamiques a été peu utilisée pour la prévision de la qualité de l'air. On peut citer l'étude de Chen et al. (1998) pour la prévision des concentrations de l'ozone. En effet, les auteurs ont mis en évidence la supériorité des modèles du chaos sur un modèle autorégressif, notamment pour la prévision à court terme. Koçak et al. (2000) ont développé un modèle de séries temporelles non-linéaires de systèmes dynamiques. La prévision a été effectuée avec une approximation polynomiale des trajectoires de l'espace des phases reconstruit et la dimension de plongement a été ré-estimée de manière séquentielle par le coefficient de corrélation entre les valeurs prédites et les valeurs observées des concentrations de l'ozone.

Dans (Chelani, 2005), l'étude propose d'utiliser les sorties de la reconstitution de l'espace d'état, après estimation de la dimension de plongement, et de l'intégrer dans un réseau de neurones pour la prévision de PM_{10}. La même procédure a été utilisée dans une autre étude pour la prévision de dioxyde d'azote (Chelani et al., 2005). Un autre exemple d'application pour la prévision des concentrations de NO_2 et de SO_2 est donné par Khokhlov et al. (2008) ; les performances en prévision sont exprimées par le coefficient de corrélation r et il est supérieur à 0.9.

Récemment, Kříž (2014) a utilisé l'approche de reconstitution de l'espace des phases : d'abord pour la caractérisation de la variabilité des concentrations de NO_2 : le retard $\tau = 5$ et la dimension de plongement $m = 7$; ensuite, la matrice de délai obtenue a servi comme entrée dans un modèle de réseaux de neurones de type RBF pour la prévision. Globalement, les résultats de prévision montrent une amélioration prédictive du modèle proposé par rapport à la moyenne de la variabilité.

D'autres travaux se concentrent principalement sur la caractérisation de la structure de variabilité des systèmes dynamiques pour le domaine de pollution. On trouve une littérature de spécialité très riche pour l'estimation de la dimension de corrélation des séries temporelles des variables environnementales (Islam et al., 1993). Souvent appelée dimension fractale de l'attracteur [2], la dimension de corrélation distingue "clairement" la partie déterministe de la partie purement aléatoire d'une série temporelle (Grassberger & Procaccia, 1983). Cependant, l'estimation de ce paramètre est largement discutée dans la littérature de spécialité. Ainsi, Ruelle (1990) estime qu'on doit être vigilent pour les estimations de la dimension qui ne sont pas bonnes : inférieures à $2 \log_2(T)$, où T est le nombre d'observations dans la série. Ce point met en évidence l'importance de la longueur de la série pour une estimation fiable de la dimension de corrélation. Pour les applications aux données des concentrations des polluants, Lee & Lin (2008) mettent en évidence la présence des caractéristiques liées à la dynamique du chaos par l'estimation de la dimension de corrélation ($\sim$ 3.42-4.71) dans de nombreuses séries de polluants.

En résumé, la plupart des applications de l'analyse des séries temporelles de polluants par cette approche consistent en la recherche de structure de variabilité. En revanche, par rapport à l'ensemble des méthodes inverses déployées dans ce domaine de recherche, très peu d'études sont dédiées aux problématiques de prévision.

5.4.3 Modèles de décomposition avec hybridation

Les méthodes par décomposition regroupent un très large spectre de procédures et de possibilités. L'une des premières méthodes proposées est la décomposition en tendance, saisonnalité et bruit. Par exemple,

2. La dimension de corrélation fait partie de la famille d'un grand nombre de méthodes pour l'estimation de la dimension fractale.

Cleveland (1979) a proposée la méthode STL pour les concentrations du CO_2 extérieures. Compte tenu de la forte saisonnalité et une tendance linéaire très importante sur ces données, la méthode STL a permis de démontrer l'importance des émissions en CO_2 dans le monde (Cleveland, 1979; Cleveland et al., 1990).

La détection des composantes saisonnières est la plus recherchée dans la littérature, car elle permet non seulement de comprendre la nature cyclique du phénomène étudié, mais aussi c'est une composante facile à prévoir. Pour la prévision un jour à l'avance des concentrations d'ozone, Kumar & De Ridder (2010) ont proposé d'utiliser la composante de Fourier associée à la fréquence saisonnière avec une modélisation ARIMA. Ensuite, les résidus de la modélisation ont été utilisés comme entrés dans un modèle non-linéaire en variance de type GARCH, ceci pour optimiser l'estimation de l'intervalle de prévision. Les résultats de cette méthode mettent en évidence la stabilité de la structure dans le modèle ARMA (1,2) sur les trois sites étudiés, mais des structures différentes pour le modèle GARCH. Il apparaît que l'introduction du modèle GARCH dans le cas étudié n'apporte que très peu d'informations sur l'intervalle de prévision. Dans une autre étude publiée récemment, le même auteur propose une approche qui consiste en l'utilisation de la méthode SSA avec des procédures de type ARIMA pour la prévision de l'ozone (Kumar, 2015). Les résultats du modèle SSA-ARIMA sont comparés avec ceux obtenus avec FFT-ARIMA, le RMSE et le MAE montrent que la qualité de prévision avec le modèle SSA-ARIMA est meilleure que par la seconde approche.

5.4.4 Comparaison entre plusieurs modèles

En raison de l'existence d'une corrélation entre les paramètres climatiques et les concentrations des polluants, plusieurs modèles de comparaison sont utilisés, notamment les modèles de type boîte noire. Ces derniers couvrent un important nombre de techniques d'analyse des données : les arbres de décisions (CART, Classification and Regression Tree Analysis), les modèles de régression, techniques de partitionnement (Clustering, Classification Non-supervisée) et les réseaux de neurones. Bruno et al. (2004) proposent le modèle de classification par arbre de décision pour la prévision de la moyenne journalière des dépassements des concentrations en ozone. Le modèle est capable de prévoir deux jours en avance les dépassements. Néanmoins, l'intégration des paramètres climatiques prévus est nécessaire pour ce type de modélisation.

Le Tableau 5.4.2 donne quelques travaux relatifs à l'application des modèles hybrides pour la prévision des concentrations de polluants. Clairement, la plupart des chercheurs s'accordent sur le fait que la combinaison de plusieurs modèles peut produire des meilleurs prévisions.

Plusieurs études ont été consacrées à la comparaison des différents modèles de prévision sur les mêmes jeux de données de la qualité de l'air. Les réseaux de neurones sont les plus utilisés comme référence. Par exemple, dans l'étude de Kukkonen et al. (2003), une comparaison entre cinq modèles neuronaux, un modèle linéaire statistique et un modèle déterministe (non spécifié) a été effectuée pour prévoir les concentrations de NO_2 et de PM_{10}. Les résultats obtenus ont mis en évidence que les ANN sont légèrement plus performants que le modèle déterministe ou celui linéaire statistique.

Malgré toutes ces comparaisons, un consensus sur un modèle fiable n'est pas facile à mettre en évidence, et ceci d'autant plus que plusieurs modèles proposés n'ont pas fait l'objet d'une comparaison.

TABLE 5.4.2 – Exemples de modèles utilisés dans la littérature pour la prévision des concentrations des polluants atmosphériques.

Modèle statistique	Polluant	Performances	Référence
ANN Persistance	PM_{10}	ANN > Persistance	Hooyberghs et al. (2005)
MLR, CART ACP+MLP	PM_{10}	RMSE : ANN=7.13 ; ACP+MLP = 8.14 CART=33.55 ; MLR =11.24.	Slini et al. (2006)
ANN+AG MLR	PM_{10}	R^2 : ANN+AG= 0.80- 0.89 GLM= 0.29-0.35	Grivas & Chaloulakou (2006)
MLR+PCA ANN+PCA	O_3	RMSE= 28.13 RMSE = 21.78	Sousa et al. (2007)
RQ	O_3	RMSE = 16.86 analyse valeurs extrêmes	Sousa et al. (2009)

Notes : Dans l'étude de Hooyberghs et al. (2005), le modèle de persistance consiste à dire que la pollution des PM au moment t est identique à celle au moment t+1. MLR représente les modèles de régression linéaire multiple ; AG représente les algorithmes génétiques et RQ représente les modèles de régression quantile.

5.4.5 Modélisation et prévision des paramètres climatiques dans l'environnement intérieur

Contrairement à la QAI, d'énormes efforts ont été accomplis dans le domaine de l'énergie des bâtiments pour la prédiction des paramètres climatique des environnement clos. Ces prédictions vont dans le sens de l'efficacité énergétique : processus d'optimisation du bilan thermique des bâtiments. Néanmoins, les modèles présentés dans ce domaine parlent souvent de la prédiction à une étape, très peu sont utilisés pour une extrapolation sur des horizons plus lointains. En effet, on trouve dans les travaux de Kusiak et al. (2010), une modélisation de deux paramètres d'ambiance intérieure et du CO_2 (appelés de cet article "IAQ") par un réseau de neurones multicouche et les cartes de contrôles.

Les modèles les plus utilisés dans ce domaine sont les modèles ARIMAX et NARX ; plusieurs revues de littérature sont proposées dans Kumar et al. (2013); Paudel et al. (2014). Une particularité des données utilisées dans ce domaine et qu'on ne voit pas souvent pour la QAI est la longueur des séries et le pas de temps. En effet, dans l'air intérieur, rarement le pas de temps utilisé est inférieur à 30 min. Par exemple Frausto & Pieters (2004) ont utilisé un an de données pour la prédiction de la température ; ces données ont été modélisées avec les paramètres d'ambiances extérieurs *via* le modèle NARX. La modélisation par ANN montre qu'il est possible de prédire ces paramètres.

Mustafaraj et al. (2011) ont présenté à leur tour une comparaison entre un modèle linéaire classique de type ARX et un modèle non-linéaire de type réseaux de neurones autorégressif à entrées exogènes. Les comparaisons ont été effectuées à deux niveaux d'analyse : la validation et le test. En ce qui concerne les prévisions de la température et de l'humidité relative, les indicateurs MAE correspondants aux différents horizons (de 30 min à 3 h) de prévision montrent que le modèle non-linéaire est plus performant.

5.5 Discussion et conclusions

La première partie de ce chapitre est consacrée aux fondements théoriques de la prévision linéaire, indépendamment de la nature de la variable étudiée. Ensuite un état de l'art a été entrepris sur les applications des modèles de prévision des séries temporelles pour la qualité de l'air extérieur.

Dans les sciences environnementales de l'air extérieur, les données présentent souvent des régularités très marquées liées aux différents processus, notamment pour le NO_2 et l'O_3. Aussi, la résolution temporelle utilisée, qui est généralement exprimée en moyenne horaire ou journalière, requiert une modélisation qui ne nécessite pas de tenir compte les fluctuations de hautes fréquences. Au contraire, les mesures actuelles de la QAI sont recueillies généralement avec un pas de temps très fin, entre une et dix minutes. En fonction de la nature du polluant, la variabilité est généralement plus fluctuante, donc des composantes aléatoires se "greffent" au sein de la série en modifiant les structures de régularité. Elles nécessitent donc une modélisation plus complexe.

A l'exception de quelques travaux, la majorité des articles présentent la prévision des concentrations de polluants dans l'air par les modèles de réseaux de neurones. Plusieurs études ont été consacrées à la comparaison de différents modèles ; la plupart présentent des confrontations entre les performances des ANN et les modèles linéaires, dont on sait qu'ils sont assez mal adaptés aux données environnementales et constituent donc de bien modestes étalons. Notons que beaucoup d'articles utilisent les valeurs futures des régresseurs pour fournir les prévisions de la variable cible. Cette situation est très problématique d'un point de vue pratique, car l'échéance de prévision doit être prise en compte même pour les régresseurs.

Malheureusement, la qualité de prévision présentée dans la plupart des travaux ne tient pas compte du "niveau de prédictibilité" des séries utilisées. En d'autres termes, la prévision des concentrations de certains polluants est plus "facile" que d'autres. Par exemple, une variabilité qui présente un profil type s'avère plus facile à estimer la reproductibilité des régularités. Par contre, une forte variabilité nécessite une exploitation plus approfondie dans la recherche des périodicités. En outre, la prévision sur certaines périodes est plus facile que pendant d'autres périodes (Tol, 1996), car la complexité d'un phénomène est conditionnée par la dynamique de ses différentes composantes.

En définitive, ce que l'on retient est qu'il est très difficile de conclure quant à la classe de modèle la plus adaptée à nos données. La problématique de prévision de la QAI nécessite un développement de méthodes spécifiques aux fluctuations qu'elle génère. Donc, la position de cette thèse est double :

- Utiliser les structures inhérentes à la variabilité des concentrations comme informations *a priori* pour la prévision ;
- Identifier les modèles de prévision susceptibles de reproduire ces caractéristiques.

Le chapitre suivant est dédié à la prévision par les méthodes de lissage et de décomposition. Justement, ces méthodes tiennent compte de certaines caractéristiques de fluctuation : reproduction en moyenne de la fréquence principale, la prévision devient donc maniable.

CHAPITRE 6

MODÈLES PAR DÉCOMPOSITION : PRÉVISION DES PARAMÈTRES DE LA QAI

L'ANALYSE *des séries temporelles repose souvent sur l'exploitation des différentes structures de variabilité. Un modèle de prévision "doit" tenir compte des informations fournies par ces structures. L'approche développée dans ce chapitre consiste en l'exploitation des différentes composantes obtenues par décomposition des séries temporelles initiales dans le but de prévision.*

Sommaire

6.1 Introduction

Dans tout système marqué par une dynamique aléatoire, la prévision statistique est indispensable à la prise de décision. En environnement intérieur, ces décisions vont au niveau des règlementations étatiques. Les modèles de prévision pour la QAI sont les éléments essentiels, qui lorsqu'ils sont disponibles, conditionnent la gestion d'une éventuelle forte exposition aux différents polluants. En fonction des valeurs des concentrations prédites, seront engagées des actions pour réduire ces dernières à des niveaux jugés "normaux". Ces procédures préventives permettront de garantir une bonne qualité de l'air dans les ambiances intérieures. Ce chapitre propose la prévision de quelques séries temporelles de la QAI par décomposition en composantes latentes. Les modèles classiques de type ARIMA et de lissage sont appliqués sur les résultats de la décomposition, donc une modélisation à double support.

L'analyse des différentes structures de variabilité temporelle de polluants intérieurs a fait apparaître diverses régularités, les plus importantes étant celles liées aux aspects de la variabilité diurne. Il s'agit pour nous de montrer l'intérêt pratique de ces méthodes pour la prévision de la QAI et les conditions de leur succès. En se basant sur les résultats de la décomposition présentés dans le chapitre 3, l'approche appliquée dans cette partie consiste à utiliser les différentes composantes déterministes extraites par différentes méthodes et modéliser le reste de cette décomposition par des modèles statistiques de prévision.

6.2 Le lissage exponentiel

6.2.1 Préliminaires

La famille des modèles de prévision par lissage exponentiel est fondée sur une idée simple : les observations affectent d'autant moins la prévision qu'elles sont éloignées du moment auquel on fait la prévision. Dans ces modèles, l'influence des observations autour de la valeur présente est grande ; on suppose en outre que cette influence est décroissante quand on remonte dans le passé. Soit T nombres réels constituant la série $X_t = (X_1, \ldots, X_T)$ et on souhaite prévoir les valeurs futures X_{T+1}, $X_{T+2} \ldots, X_{T+h}$.

Définition 6.2.1. Lissage exponentiel simple
La prévision de la série X_T à l'horizon $h \in \mathbb{N}^*$, notée $\widehat{X_T}(h)$ par la méthode du lissage exponentiel simple est définie par :

$$\widehat{X_T}(h) = (1-\beta)\sum_{j=0}^{T-1} \beta^j X_{T-j}, \tag{6.2.1}$$

où $\beta\,(0 < \beta < 1)$ une constante de lissage. Cette prévision ne dépend de h qu'à travers β. Notons que si β est choisi indépendamment de h, $\widehat{X_T}(h)$ ne dépend pas de h.

Par définition, on voit que plus la constante de lissage β est proche de 1, plus la prévision est rigide, c'est-à-dire peu sensible aux fluctuations récentes. Au contraire, plus la constante β est proche de 0, plus la prévision est influencée par les observations récentes, c'est-à-dire la prévision est souple. La notation $\hat{X}_T(h)$ dans 6.2.1 est souvent ramenée à $\widehat{X}_T$ lorsque β est indépendant de h.

Au-delà des considérations qualitatives (voire subjectives) basées sur l'appréhension de "rigidité" ou de "souplesse", le choix du β peut être estimé par minimisation de la somme des carrés des erreurs de prévision aux différents moments $1, \ldots, T - h_0$, pour un horizon h_0 donné :

$$\widehat{\beta} = \arg\min_{\beta} \left\{ \sum_{t=1}^{T-h_0} \left[X_{t+h_0} - \widehat{X}_t(h_0) \right]^2 \right\}. \tag{6.2.2}$$

Ainsi pour $h_0 = 1$, le problème devient :

$$\widehat{\beta} = \arg\min_{\beta} \left\{ \sum_{t=1}^{T-1} \left[X_{t+1} - (1-\beta) \sum_{j=0}^{T-1} \beta^j X_{t-j} \right]^2 \right\}. \tag{6.2.3}$$

Il est difficile d'établir un lien entre l'erreur de prévision et le paramètre β. L'influence du choix de ce dernier reste faible dans des plages assez grandes. Pour soutenir ce propos et en s'inspirant des résultats fournis par Cox (1961), Gourieroux & Monfort (1995) donnent un exemple pour un processus autorégressif d'ordre un $\left(X_T \sim AR(1), \text{ avec } \mathbb{E}(X_t X_{t+h}) = \rho^{|h|}\right)$ de corrélation ρ. L'erreur de prévision à l'horizon h avec la méthode du lissage exponentiel simple est

$$\Delta(\rho, \beta, h) = \frac{2}{1+\beta} + \frac{2(1-\beta)\left(\beta\rho - \rho^h - \beta\rho^h\right)}{(1+\beta)(1-\beta\rho)}. \tag{6.2.4}$$

Pour $h = 1$, le problème se ramène à la discussion des coefficients ρ en fonction de β :

— si $\frac{1}{3} < \rho < 1$, la fonction $\Delta(\rho, \beta, 1)$ admet un minimum en $\frac{1-\rho}{2\rho}$;

— si $-1 < \rho < \frac{1}{3}$ la fonction $\Delta(\rho, \beta, 1)$ est décroissante sur l'intervalle $[0, 1]$.

En résumé, si ρ est négatif, la méthode donne de très mauvais résultats. Si au contraire, ρ est positif, alors les meilleures valeurs de β sont de l'ordre de 0.7 ou 0.8 ; d'ailleurs ce sont les valeurs les plus utilisées (Gourieroux & Monfort, 1995; Brown, 2004).

Bien que ces méthodes aient été introduites à la fin des années 50 (Holt, 1957; Winters, 1960; Muth, 1960) [1], le cadre de modélisation stochastique, l'analyse de sélection du modèle ainsi que l'estimation de l'intervalle de prévision n'ont été développés que récemment (Ord et al., 1997; Hyndman et al., 2002, 2008). On peut trouver dans (Harvey, 1984; Gardner, 1985, 2006) un état de l'art sur la méthode de prévision par le lissage exponentiel.

6.2.2 Taxonomie des méthodes de prévision par lissage exponentiel

L'analyse des méthodes de lissage exponentiel commence généralement par l'analyse de la tendance. Cette dernière est composée par deux termes : le niveau ℓ et la croissance b. La combinaison de ces sous-composantes au sein d'une série temporelle peut être construite de diverses manières. Soit T_h la prévision de la tendance sur un horizon h et soit ϕ $(0 < \phi < 1)$ un paramètre d'amortissement qui contrôle le niveau de croissance. Il est possible de construire au moins cinq modèles pour la tendance :

1. Les travaux de Holt sont très largement cités, mais n'ont été publiés qu'en 2004 (voir Holt (2004))

TABLE 6.2.1 – Types de modèles pour la méthode de lissage exponentiel

Tendance	Saisonnalité		
	Aucun	Additif	Multiplicatif
N	N,N	N,A	N,M
A	A,N	A,A	A,M
$\mathbf{A}_d$	A_d, N	A_d, A	A_d, M
M	M, N	M, A	M, M
M_d	M_d, N	M_d, A	M_d, M

$$
\begin{aligned}
\text{Aucun } \mathbf{N} &: T_h = \ell \\
\text{Additif } \mathbf{A} &: T_h = \ell + b \\
\text{Additif Amorti } \mathbf{A_d} &: T_h = \ell + \sum_{j=1}^{h} \phi^j b \\
\text{Multiplicatif } \mathbf{M} &: T_h = \ell b^h \\
\text{Multiplicatif Amorti } \mathbf{M_d} &: T_h = \ell b^{\sum_{j=1}^{h} \phi^j}
\end{aligned}
$$

Le choix du type de la tendance est fait généralement par l'analyse "graphique" de la série. L'introduction d'un terme d'amortissement suggère qu'il existe une relation entre l'impact des dernières observations sur l'horizon de prévision.

Dans un deuxième temps, la composante saisonnière est introduite, soit de manière additive, soit multiplicative. Ensuite, la composante aléatoire ou "l'erreur" est incluse de différentes manières, tout comme la saisonnalité. La nature de la composante aléatoire est souvent ignorée dans ce type de modélisation, mais l'introduction des modèles d'espace-états (state-space models) dans (Hyndman et al., 2002, 2008) suggère la possibilité de calculer les intervalles de prévision pour chaque modèle.

La classification des méthodes de lissage exponentiel peut être résumée dans le Tableau 6.2.1. Par exemple la cellule N, N décrit la méthode du lissage exponentiel simple, la cellule A, N décrit la méthode traditionnelle de Holt (1957) et la cellule A, A décrit le modèle additif de HOLT-WINTERS.

Hyndman et al. (2008) proposent d'étendre cette classification des modèles de lissage par l'intégration du terme des chocs ε_t. Ces termes interviennent soit de manière additive ou de manière multiplicative et on aboutit à la forme des modèles espace-états (state-space models). Trente modèles espace-états peuvent être dérivés du tableau 6.2.1. Ces modèles peuvent fournir non seulement la prévision moyenne comme le lissage exponentiel, mais aussi les intervalles de prévision autour des valeurs prédites. Dans cette thèse, nous utilisons l'approche proposée dans (Hyndman et al., 2002, 2008) pour le calcul des intervalles de prévision.

6.2.2.1 Prévision récursive de la méthode de lissage exponentiel simple

Notons que dans la littérature, par exemple dans (Hyndman et al., 2008), la constante de lissage est souvent notée par $\alpha = (1 - \beta)$. Nous gardons cette dernière notation.

On se place au moment $t-1$; la prévision de la valeur suivante de X_t est $\widehat{X_t}$ et lorsque l'on observe la valeur réelle X_t, alors l'erreur de prévision $\widehat{X_t} - X_t$ est disponible. La prévision suivante $\widehat{X}_{t+1}$ s'interprète alors comme la prévision au moment précédent corrigée d'un terme proportionnel à la dernière erreur de prévision :

$$\widehat{X}_{t+1} = \widehat{X_t} + \alpha \left(X_t - \widehat{X}_t \right), \tag{6.2.5}$$

$$\widehat{X}_{t+1} = \alpha X_t + (1-\alpha)\, \widehat{X}_t. \tag{6.2.6}$$

6.2.2.2 Méthode de Holt-Winters et représentation espace-état

Cette description est réduite au modèle saisonnier additif, c'est-à-dire nous envisageons uniquement le cas d'une série présentant deux composantes additives : une tendance et une saisonnalité de période m. La méthode Holt-Winters prédit X_t connaissant la série jusqu'à $t-1$ par :

$$\widehat{X_t} = \ell_{t-1} + b_{t-1} + S_{t-m}, \tag{6.2.7}$$

où ℓ_t et b_{t-1} sont respectivement le niveau et la pente au moment $t-1$, et S_{t-m} désigne la composante saisonnière au moment $t-m$. L'erreur de prévision s'écrit :

$$\hat{e}_t = X_t - \widehat{X_t}. \tag{6.2.8}$$

Quand l'observation X_t devient disponible, les composantes sont mises à jour en commençant par le niveau :

$$\ell_t = \alpha \left(X_t - S_{t-m} \right) + (1-\alpha) \left(\ell_{t-1} + b_{t-1} \right). \tag{6.2.9}$$

Notre connaissance de la saisonnalité pour le moement t remonte en effet à $t-m$. Ensuite, on met à jour la pente

$$b_t = \beta^* \left(\ell_t - \ell_{t-1} \right) + (1-\beta^*)\, b_{t-1}, \tag{6.2.10}$$

et enfin la saisonnalité :

$$S_t = \gamma \left(X_t - \ell_{t-1} - b_{t-1} \right) + (1-\gamma)\, S_{t-m}. \tag{6.2.11}$$

Les équations de base de la méthode de Holt-Winters, aussi appelée *lissage exponentiel triple* pour le cas saisonnier additif sont donc données comme suit :

$$\text{Niveau} : \ell_t = \alpha \left(X_t - S_{t-m} \right) + (1-\alpha) \left(\ell_{t-1} + b_{t-1} \right) \tag{6.2.12}$$

$$\text{Pente} : b_t = \beta^* \left(\ell_t - \ell_{t-1} \right) + (1-\beta^*)\, b_{t-1} \tag{6.2.13}$$

$$\text{Saison} : S_t = \gamma \left(X_t - \ell_{t-1} - b_{t-1} \right) + (1-\gamma)\, S_{t-m} \tag{6.2.14}$$

$$\text{Prévision} : \widehat{X}_{t+h} = \ell_t + b_t h + S_{t-m-h_m^+} \tag{6.2.15}$$

où m est la longueur de la fréquence saisonnière et $h_m^+ = [(h-1) \mod m]+1$. Les paramètres α, β^* et γ sont entre 0 et 1. On garde la notation β^* au lieu de β, car cette dernière est souvent réservée aux modèles espace-états, décrits ci-après.

En faisant intervenir l'erreur de prévision $\hat{e}_t$ dans les différentes équations, on peut écrire un modèle probabiliste parallèle au modèle de HOLT-WINTERS standard (Hyndman et al., 2002).

$$\begin{aligned} \widehat{X}_t &= \ell_{t-1} + b_{t-1} + S_{t-m} + \varepsilon_t \\ \ell_t &= \ell_{t-1} + b_{t-1} + \alpha\varepsilon_t \\ b_t &= b_{t-1} + \beta\varepsilon_t \\ S_t &= S_{t-m} + \gamma^*\varepsilon_t, \end{aligned} \tag{6.2.16}$$

où $\beta = \alpha\beta^*$, $\gamma^* = \frac{1}{1-\alpha}\gamma$ et ε_t est une innovation. La prévision par modèle d'espace-état est identique au modèle classique de HOLT-WINTERS, seul le calcul des intervalles de prévision diffère (Hyndman et al., 2008).

Considérons le cas d'une série temporelle des concentrations de polluant en air intérieur au pas de temps horaire. Après l'identification de la saisonnalité, faite par d'autres analyses préalables, la représentation espace-état peut être décrite comme suit en reprenant le système 6.2.16. L'état $\boldsymbol{y}_t$ est défini par

$$\boldsymbol{y}_t = [\ell_t \; b_t \; S_t \; \ldots \; S_{t-m+1}]^\top, \tag{6.2.17}$$

et en introduisant trois matrices $\boldsymbol{w}, \boldsymbol{F}$ et $\boldsymbol{g}$:

$$\boldsymbol{w} = \begin{bmatrix} 1 \\ 1 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}, \quad \boldsymbol{F} = \begin{bmatrix} 1 & 1 & 0 & 0 & \ldots & 0 & 0 \\ 0 & 1 & 0 & 0 & \ldots & 0 & 0 \\ 0 & 0 & 0 & 0 & \ldots & 0 & 1 \\ 0 & 0 & 1 & 0 & \ldots & 0 & 0 \\ 0 & 0 & 0 & 1 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \ldots & 1 & 0 \end{bmatrix} \quad \text{et } \boldsymbol{g} = \begin{bmatrix} \alpha \\ \beta \\ \gamma \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

Le modèle général d'*innovation de l'espace-d'état* est

$$X_t = \boldsymbol{w}^\top \boldsymbol{y}_{t-1} + \varepsilon_t, \tag{6.2.18}$$

$$\boldsymbol{y}_t = \boldsymbol{F}\boldsymbol{y}_{t-1} + \boldsymbol{g}\varepsilon_t, \tag{6.2.19}$$

où X_t sont les observations au moment t, $\boldsymbol{y}_t$ est le vecteur d'état et $\{\varepsilon_t\}$ est un processus BB (bruit blanc) Gaussien de variance σ_ε^2. Du point de vue modélisation, cette représentation permet d'identifier un large spectre de modèles plus complexes en les regroupant sur des parties plus "maniables", réduisant ainsi le risque d'erreur dans la spécification des modèles (Hyndman et al. (2008), page 34). Le terme $\boldsymbol{w}^\top \boldsymbol{y}_{t-1}$ dans l'équation de mesure 6.2.18 décrit l'effet de la mémoire sur la valeur présente de X_t, et ε_t décrit la part aléatoire, elle est donc l'innovation du processus : la seule source de "stochasticité" pour la série observée, $\{X_t\}$. L'équation 6.2.19 est appelée équation de transition et sa matrice de transition est $\boldsymbol{F}$. Le terme $\boldsymbol{F}\boldsymbol{y}_{t-1}$ traduit l'effet mémoire sur l'état actuel $\boldsymbol{y}_t$ et $\boldsymbol{g}\varepsilon_t$ représente l'imprédictibilité sur le vecteur d'état $\boldsymbol{y}_t$. Quant au vecteur $\boldsymbol{g}$, il traduit l'effet de persistance sur le vecteur d'état en contrôlant

la transmission d'erreur *via* les variables latentes. L'équation de transition est donc un mécanisme créant des dépendances inter-temporelles entre les différentes composantes d'une série temporelle.

On note par ailleurs que les modèles issus de 6.2.18 et 6.2.19, peuvent être étendus aux modèles encore plus généraux, en permettant aux matrices $\boldsymbol{w}, \boldsymbol{F}$ et $\boldsymbol{g}$ de varier dans le temps.

Les problèmes d'initialisation et d'estimation de ces modèles sont largement discutés dans la littérature de prévision, voir par exemple (Hyndman et al., 2002). En revanche, la plupart des travaux décrivent les solutions par des schémas heuristiques basés sur des séries temporelles de résolution très large : de journalière à annuelle. Or, le pas de temps utilisé dans nos applications est d'environ 1000 fois plus réduit (1440 minutes par jour). Donc, la sélection de la fréquence principale pour l'extraction de la composante saisonnière est très peu documentée dans le cas des données de hautes fréquences.

6.3 Prévision par décomposition SSA (Singular Spectrum Analysis) : aspects théoriques

Comment est-il possible de réaliser une analyse multivariée à partir d'une seule série temporelle et fournir les prévisions de cette dernière ? La technique SSA propose une solution. Rappelons que cette question été soulevée durant la fin des années 80 (Fraedrich, 1986; Broomhead & King, 1986b,a; Vautard & Ghil, 1989; Vautard et al., 1992) et reprise après les années 2000 (Golyandina et al., 2001; Golyandina & Zhigljavsky, 2013).

Cette sous-section est complémentaire à celle développée dans 3.6.4 (chapitre 3). Donc, dans la description suivante, on garde les mêmes notations. La discussion sur le choix de L (fenêtre de la matrice retard) et la notion de séparabilité des séries sont deux éléments très importantes pour la prévision.

Soient :

- *i)* I l'ensemble des triplets-propres choisi ;
- *ii)* $E_i \in \mathbb{R}^L,\ i \in I$, les vecteurs propres correspondants et $\underline{E_i}$ leurs premières $L-1$ coordonnées ;
- *iii)* π_i la dernière coordonnée de E_i ;
- *v)* $\nu^2 = \sum_{i \in I} \pi_i^2$;
- *vi)* $\widetilde{X}_N$ les séries reconstruites par I ;

et $R = (a_{L-1}, , a_1)^\top$ comme

$$R = \frac{1}{1-\nu} \sum_{i \in I} \pi_i \underline{E_i}. \tag{6.3.1}$$

La prévision par récurrence (Golyandina & Zhigljavsky, 2013) de $Y_{N+M} = (y_1, \ldots, y_{N+M})$ peut être décrite par

$$y_i = \begin{cases} \widetilde{x_i} & \text{pour } i = 1, \ldots, N, \\ \sum_{j=1}^{L-1} a_j y_{i-j} & \text{pour } i = N+1, \ldots, N+m. \end{cases} \tag{6.3.2}$$

Une telle prévision dépend fortement des paramètres de reconstitution des séries, notamment la longueur de la fenêtre L de la matrice retard dans 3.6.7. Basé sur une étude de simulation, le choix de L ne doit pas dépasser $N/2$; Golyandina (2010) recommande $L = N/3$.

TABLE 6.4.1 – Récapitulatif des modèles de lissage et de décomposition appliqués à la variabilité des différents polluants.

Environnement	**BI2011**	**MARIA**	**OS2012**		**OS2013**	**OS2015**			
Polluant	CO_2	HCHO	CO_2	$PM_{[0.35-8.75]}$	HCHO	HCHO	$PM_{0.35}$	$PM_{2.5}$	CO_2
HW	✔	✔	✘	✘	✔	✔	✔	✔	✔
STL+ARIMA	✘	✔	✔	✔	✔	✔	✘	✘	✘
SSA	✘	✔	✘	✘	✔	✔	✔	✔	✘
N	52560	17281	3720		20000	3251	∼118 j		∼138 j
N.app	3 mois	11521	3624		19000	2963	60 j		60 j
N.test	1008	5760	96		1000	288	2 j		7 j
Pas de temps	10 min	1 min	1 h		1 min	20 min	1 min		

Note : Le symbole ✔ *désigne que le modèle est appliqué et comparé avec d'autres méthodes et le symbole* ✘ *désigne le cas contraire.*

6.4 Données utilisées et procédures de prévision par les modèles de décomposition

Cette sous-section est dédiée à la présentation des données utilisées pour la prévision des concentrations de polluants dans les différents environnements par les méthodes de lissage et de décomposition. Les détails sur les données de mesure sont présentés dans le chapitre 2 et les caractéristiques des fluctuations dans le chapitre 3.

Nous présentons dans le Tableau 6.4.1 le récapitulatif des méthodes utilisées pour chaque polluant ainsi que les détails sur les parties d'apprentissage et test utilisées.

6.4.1 Procédure pour l'application de la méthode de Holt-Winters

La question à laquelle on voudrait apporter un éclairage est : y-a-t-il un modèle "standard" et simple qui soit adapté à un grand nombre de situations pour la prévision de la QAI ? Ici, on se propose de donner les conditions de succès d'un modèle de type HOLT-WINTERS pour différents polluants. Les paramètres obtenus sont $\alpha = 0.00033$, $\beta^* = 0.003$ et $\gamma = 0.095$, nous expliquons ci-après la procédure générale.

La méthode de HOLT-WINTERS (HW) est utilisée d'abord pour la prévision des concentrations de CO_2 dans le bureau individuel. On commence par le polluant "le plus simple" du point de vue variation temporelle, dans l'environnement le moins complexe : le bureau individuel (la maison expérimentale ne présentait pas d'intérêt, car celle-ci était inoccupée).

L'objectif est de déterminer les plages de variation des paramètres α, β^* et γ de manière à ce que la prévision par le modèle 6.2.15 soit acceptable dans cet environnement. On utilise ensuite ces paramètres comme initialisation pour les autres séries : d'autres environnements et d'autres polluants.

La procédure consiste à vérifier si avec une bonne prévision des concentrations de CO_2 dans le bureau individuel, servant de base initiale pour des modèles optimisés pour la suite, on arrive à prévoir les concentrations du même polluant dans l'espace paysager. En gardant les paramètres optimisés pour le

CO_2 dans l'espace paysager, on teste si on aboutit à une prévision acceptable pour d'autres polluants, notamment les PM et le HCHO.

Concrètement, les étapes que nous avions adoptées sont présentées ci-après :

Etape I *Dans le bureau individuel*

- Sur une année de mesure des concentrations de CO_2 avec un pas de temps de 10 minutes (*cf.* Tableau 6.4.1), on découpe la série en trois parties : l'ensemble d'apprentissage, l'ensemble de validation et l'ensemble de test. Soit par exemple un découpage de trois mois pour l'apprentissage, une semaine pour la validation et le reste de l'année pour le test. La taille de la série de validation doit être identique avec celle de la série test. De préférence, la série de validation (donc, de test) doit correspondre à la période principale détectée par l'analyse spectrale, qui est pour le CO_2, une semaine.

- Sur l'ensemble de validation, nous effectuons les prévisions par la méthode de Holt-Winters en faisant varier les paramètres du modèle HW jusqu'à ce que la prévision soit bonne : un RMSE < 70 ppm sur une semaine de prévision ;

- Pour tester le modèle, on peut procéder de différentes manières :
  - On tire, de manière aléatoire, une semaine test sur l'ensemble de test, qu'on compare avec la semaine prédite dans l'étape de validation. Dans ce cas, il faut conditionner le tirage de la série test, c'est-à-dire tirer une semaine qui commence avec le même jour que le jour de la prévision dans l'étape de validation (dans notre cas, c'est un lundi) ;
  - Avec une série d'apprentissage plus longue (incluant la série de validation de l'étape précédente), on procède à une prévision de la semaine test (Figure 6.4.1). Les résultats que nous présentons dans la phase de test sont ceux de ce découpage.

FIGURE 6.4.1 – *Procédure de test avec le modèle HW pour les concentrations de CO_2*

- À ce stade, les prévisions que nous avons obtenues sont satisfaisantes (RMSE < 70 ppm sur une semaine de prévision) ; nous récupérons alors les valeurs des paramètres de ce modèle et on passe à l'étape II. Les valeurs obtenues pour la prévision du CO_2 dans le bureau individuel sont $\alpha = 0.00199$, $\beta^* = 0.0065$ et $\gamma = 0.075$.

Etape II *Dans l'espace paysager*

- Une fois la plage de variation des paramètres déterminée, on procède à la prévision des concentrations de CO_2 dans l'espace paysager. Les valeurs des paramètres obtenues ne sont pas loin de celles obtenues dans la première étape : $\alpha = 0.00033$, $\beta^* = 0.003$ et $\gamma = 0.095$;
- Ensuite, nous appliquons le modèle HW avec les mêmes paramètres obtenus pour la prévision du CO_2 sur le modèle de prévision des concentrations de HCHO et des particules.

On rappelle que cette procédure est motivée par le fait que la variabilité de CO_2 est "facile" à prévoir et générée par un environnement moins complexe (Bureau individuel). Les données de CO_2 ont servi donc à tester les modèles sur plusieurs jeux de données des autres polluants.

6.4.2 Décomposition par STL (Seasonal Trend Decomposition using Loess) associée à un modèle de prévision

La prévision associée à la décomposition STL consiste à modéliser les composantes latentes issues de cette dernière. Plus précisément, on modélise la série désaisonnalisée soit par un modèle ARIMA ou par un modèle de type lissage exponentiel. La prévision de la composante saisonnière réside dans la projection de la série comprenant la fréquence principale sur l'horizon de prévision. Cette composante est déterministe, donc l'erreur de prévision est supposée nulle.

L'intervalle de prévision est calculé à partir de la série désaisonnalisée. Pour le modèle ARIMA, le calcul est effectué en utilisant la formule 5.3.12 pour la série X_t et 5.3.14 pour la série $Y_t = \log(X_t)$.

6.5 Application de la méthode de lissage exponentiel sur les concentrations des polluants

Bien qu'elle soit ancienne par rapport aux développements récents de l'analyse des séries temporelles, la méthode de lissage exponentiel s'avère assez performante pour la prévision des séries de certains polluants, notamment, pour les données dont les fluctuations sont dominées par les composantes déterministes.

De ce point de vue, nous pensons que la détection et l'intégration de la fréquence principale dans les modèles de ce type vont affecter la prévision des concentrations des polluants. La fréquence principale, va traduire le rôle de la composante déterministe "saisonnière", elle dépend donc du type de polluant à traiter.

La Tableau 6.5.1 résume l'utilisation de la composante saisonnière pour les séries de polluants appliquée aux différents modèles de prévision. En se référant aux résultats des structures de variabilité temporelles, traitées dans le chapitre 3, il est difficile pour certains polluants, notamment le HCHO et les particules fines, d'utiliser une seule période principale. Les densités spectrales des séries de ces polluants exhibent un pôle à la fréquence zéro. Le "choc" des perturbations aléatoires se manifeste tout au long du spectre.

Nous traitons ce problème dans le chapitre suivant en se proposant de donner une méthode d'analyse par bandes spectrales.

6.5.1 La variabilité du CO_2

En se référant aux résultats préliminaires présentés dans le chapitre 2, la variabilité du CO_2 dépend fortement de la densité de l'occupation et de l'aération. Rappelons que les périodes principales obtenues par l'analyse spectrale sont d'une semaine et d'un jour (*cf.* chapitre 3).

On présente dans la Figure 6.5.1a les prévisions de la variabilité du CO_2 et les performances du modèle de Holt-Winters dans l'étape de validation. Après plusieurs simulations, on arrive à une combinaison des paramètres qui fournit des prévisions acceptables : $\alpha = 0.00199$, $\beta^* = 0.0065$ et $\gamma = 0.075$ ($RMSE < 70\,ppm$). Malgré une légère baisse de la tendance de la prévision dans l'étape de validation, les résultats restent acceptables et reproduisent globalement la variabilité journalière d'occupation et de l'inoccupation pendant le week-end.

TABLE 6.5.1 – Périodes principales jouant le rôle de la composante saisonnière appliquée aux différents modèles de prévision.

Polluant	Environnement	Pas de temps	Période principale
CO_2	BI2011	10 min	7 jours
	OS2015	1 min	
HCHO	MARIA	1 min	1 jour
	OS2013	1 min	4 jour*
	OS2015	20 min	7 jour
$PM_{0.35-0.9}$	OS2012	1 h	7 jours*
$PM_{1.8-8.75}$			7 jours
$PM_{0.35}$	OS2015	1 min	7 jours*
$PM_{2.5}$			7 jour

Notes : (*) désigne que la série temporelle ne présente pas réellement de période principale : la densité spectrale exhibe un pôle à fréquence zéro pour les fines particules et le HCHO.

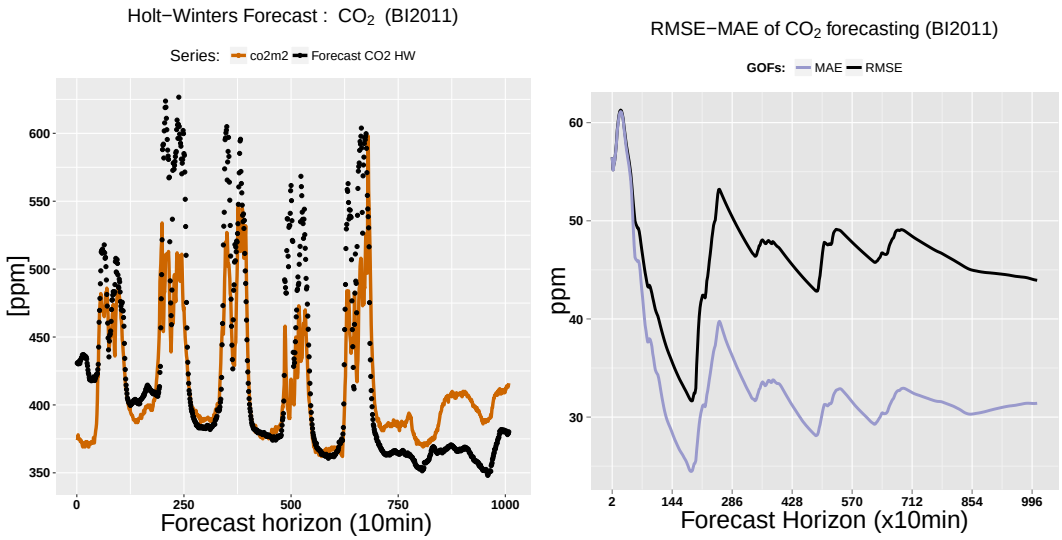
Sur la Figure 6.5.1b, on présente les prévisions sur la partie test ainsi que les performances associées au modèle HW. Clairement, du point de vue RMSE et MAE, les performances globales sont très bonnes. En effet, bien qu'on projette la trajectoire sur 1008 données, le RMSE global est au dessous de 70 ppm et le critère MAE est toujours inférieur à 75 ppm quelque soit l'horizon de prévision. Rappelons que la limite de détection du capteur de CO_2 est de 50 ppm.

Les paramètres obtenus par le modèle HW doivent être interprétés par rapport à la résolution temporelle utilisée, qui est de 10 minutes. En effet, le pas de temps des séries influence sur le niveau de variabilité ainsi que sur les composantes latentes de tendance et de saisonnalité. Un pas de temps large a pour effet de lisser la série, donc fait apparaître plus les composantes déterministes. Au contraire, un pas de temps fin augmente le niveau de variabilité.

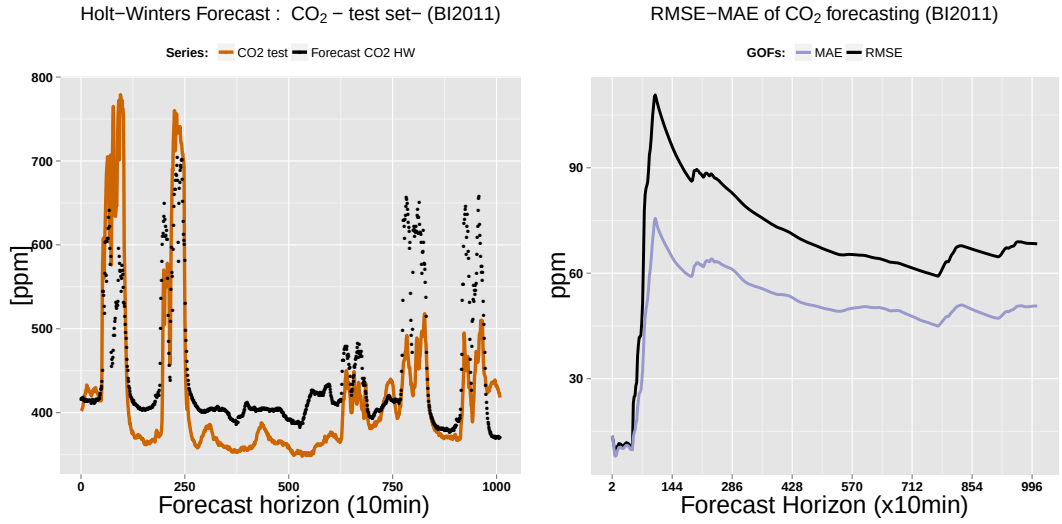
Les paramètres obtenus pour la série du CO_2 montrent que le paramètre de lissage saisonnier γ est le plus important de tous les autres ; la tendance (paramètre α) et la pente de croissance (paramètre β^*) sont quasi nulles.

Nous reprenons les paramètres du modèle HW obtenus pour le CO_2 dans le bureau individuel et nous les utilisons comme valeurs initiales pour la prévision du même polluant, mais dans l'espace paysager. Rappelons ici que pour cette deuxième série, le pas de temps est d'une minute, alors que pour la première, le pas de temps était de 10 minutes.

La Figure 6.5.2 donne les prévisions sur une semaine (10080 points) des concentrations de CO_2 dans l'espace paysager durant la campagne de 2015. Hormis un début médiocre de prévision, d'un écart d'environ 90 ppm, la qualité de prévision fournie par HW avec les paramètres optimisés est bonne. En effet, sur l'ensemble de la série test, les indicateurs MAE et le RMSE avoisinent les 45 ppm et 55 ppm, respectivement.



(a) *Prévision sur l'ensemble de validation des concentrations du CO_2 après optimisation des paramètres du modèle HW (en noir) et comparaison avec les mesures (en marron) et performances du modèle.*



(b) *Prévision sur l'ensemble de test des concentrations de CO_2 en utilisant les paramètres obtenus de l'étape de validation.*

FIGURE 6.5.1 – *Prévision d'une semaine des concentrations du CO_2 (n=52560) des mesures effectuées dans le bureau individuel avec la méthode de* HOLT-WINTERS. *La partie apprentissage a été effectuée sur environ 3 mois d'observations (apprentissage et validation) et la partie test, sur une semaine. Les paramètres du modèle sont : $\alpha = 0.00199$, $\beta^* = 0.0065$ et $\gamma = 0.075$ et les performances du modèle en termes de RMSE et MAE sont données dans le graphique à droite.*

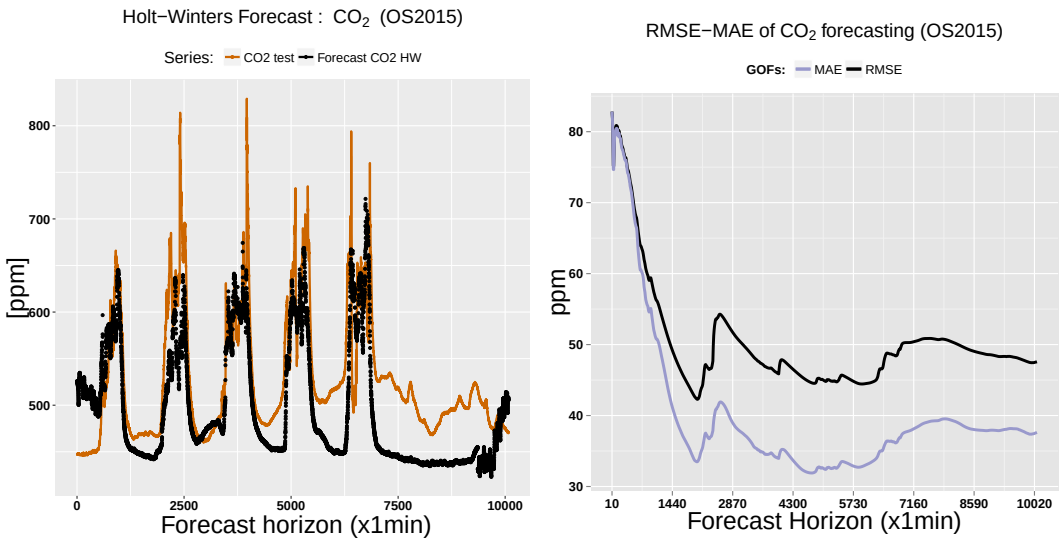


FIGURE 6.5.2 – *Prévision d'une semaine des concentrations du CO_2 (n=200 000) des mesures effectuées dans le bureau paysager durant la campagne de 2015, avec la méthode de* HOLT-WINTERS. *La partie apprentissage a été effectuée sur environ 2 mois d'observations et la partie test, sur une semaine. Les paramètres du modèle sont : $\alpha = 0.00033$, $\beta^* = 0.003$ et $\gamma = 0.095$ et les performances du modèle en termes de RMSE et MAE sont données dans le graphique à droite.*

6.5.2 Prévision de la variabilité des particules

En se basant sur les résultats de la prévision du CO_2 fournies par le modèle HW, nous testons le pouvoir prédictif de ce dernier sur la prévision des concentrations des particules dans l'espace paysager durant la campagne de 2015. Nous présentons uniquement les résultats relatifs à la prévision des particules fines ($0.35\,\mu$m) et moyennes ($2.5\,\mu$m). La variabilité des grosses particules est dominée par des sauts "quasi-discrets", leur prévision nécessite alors un développement des modèles à valeurs entières de type modèle de comptage des séries temporelles ; cette thèse ne traitera pas cette problématique.

Sur la Figure 6.5.3, on présente les résultats de la prévision des concentrations des particules de tailles $0.35\,\mu$m et $2.5\,\mu$m par le modèle de HOLT-WINTERS. Clairement, la qualité de la prévision par le modèle appliqué à la série de particules moyennes est meilleure que celle appliquée à la série des particules fines. Pour les particules fines, le modèle HW surestime largement les concentrations, d'environ $40\,\#\text{cm}^{-1}$ en début de prévision et d'environ $20\,\#.\text{cm}^{-1}$ sur un horizon de deux jours de prévision. Néanmoins, on constate une reproduction de la variation générale, c'est-à-dire deux pics au niveau d'un jour et d'un jour et demi de prévision avec un retard d'environ 3 h ; donc le modèle semble déphasé.

En ce qui concerne la prévision des particules de taille moyenne, le modèle HW arrive à reproduire la variation globale sur 3000 minutes (50 h) de prévision. En visualisant le caractère très "bruité" de la variabilité des $PM_{2.5}$, il paraît naturel d'obtenir un indice mesurant la qualité des performances très faible. En effet, on présente dans 6.5.4 le coefficient de détermination (R^2) et l'indice d'agrément modifié (md) entre la série test et la série prédite. Pour les particules de taille $2.5\,\mu$m, le R^2 est toujours inférieur à 0.3 alors que le md peut atteindre 0.5 sur l'ensemble de la série test. En revanche, la qualité de prévision

exprimée par R^2 pour les particules fines est très médiocre. Cela s'explique par une forte variabilité de ces dernières.

Par ailleurs, lorsqu'on calcule les indices de performance à l'échelle de quelques minutes (d'une minute à dix minutes), l'évolution de ces derniers n'est pas graduelle en fonction de l'horizon de prévision. On constate que l'erreur globale donnée par RMSE ou le MAE à la fin de l'échéance de prévision n'est pas fortement inférieure à l'erreur moyenne obtenue en début de la prévision.

Cette caractéristique est observée surtout lorsqu'au début de la prévision, le modèle échoue dans la détermination des premières valeurs, mais récupère, au cours du temps, l'évolution moyenne de la variabilité. De plus, on remarque que le R^2 est beaucoup plus affecté par cette caractéristique que l'indice *md*.

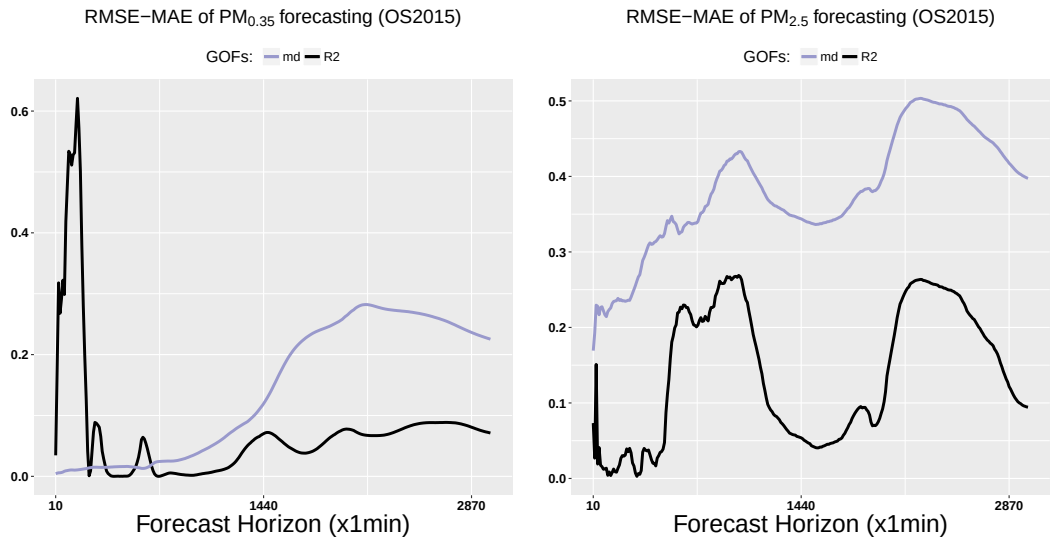


FIGURE 6.5.4 – *Le coefficient de détermination R^2 et l'indice d'agrément modifié md de la prévision des concentrations de particules (0.35 µm et 2.5 µm) par la méthode HW.*

6.5.3 Prévision de la variabilité du HCHO

La particularité la plus importante qu'on a pu remarquer pour les différentes séries de HCHO est la diversité des structures de variabilité temporelles :

- Un niveau de concentration élevé dans la maison expérimentale avec une variation quasi-sinusoïdale ;
  - Absence d'occupation ;
  - Les composantes déterministes sont les plus déterminantes dans les fluctuations avec quelques inégalités.
- Faible niveau avec de très forts sauts abrupts et changement de régimes pour la série de 2013 ;
  - Densité d'occupation et actions sur les fenêtres très importantes.
  - Fortes irrégularités autour des composantes déterministes, les perturbations des chocs aléatoires sont très importantes ;

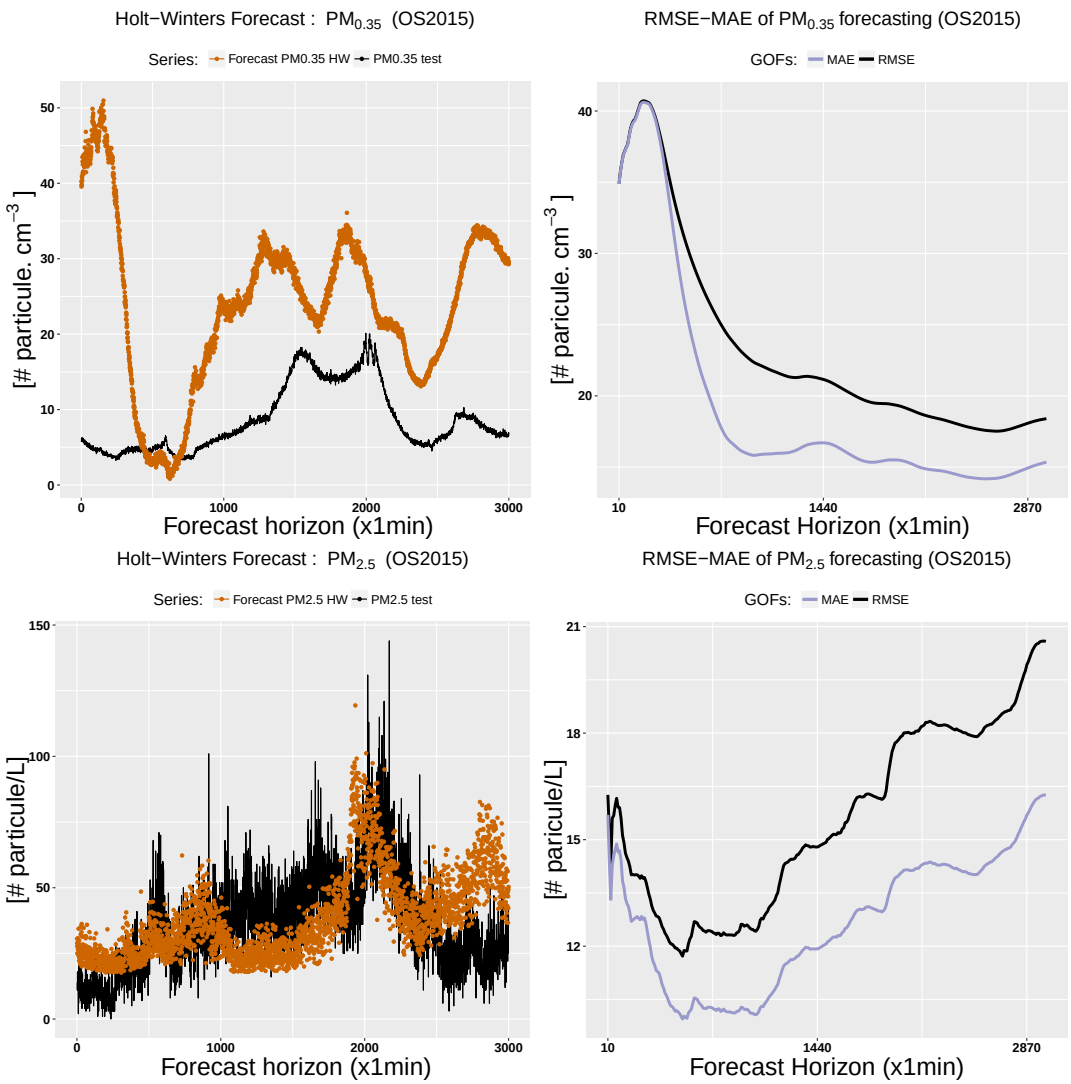


FIGURE 6.5.3 – *Prévision des concentrations des particules de tailles 0.35 µm et 2.5 µm dans l'espace de bureaux durant la campagne de 2015 par le modèle* HOLT-WINTERS *en utilisant les paramètres :* $\alpha = 0.00033$, $\beta^* = 0.003$ et $\gamma = 0.0095$. *Les performances du modèle en termes de RMSE et MAE sont données dans les graphiques à droite.*

— Durant la campagne de 2015, les niveaux moyens de concentration avec un profil diurne et hebdomadaire sont très marqués autour d'une tendance mensuelle ;
— Densité d'occupation normale, mais très peu d'actions sur les fenêtres (hormis le mois de juin) ;
— Les composantes déterministes dominent les fluctuations.

Au regard de ces aspects, il est très clair qu'un modèle de prévision doit tenir compte de ces paramètres. La difficulté réside non seulement dans la détermination de l'importance de chaque composante, mais aussi de savoir à quel moment une perturbation aléatoire est plus importante qu'une variation saisonnière.

À présent, on s'affranchit de certaines questions qui relèvent de "raffinement des modèles" et on les traitera dans le chapitre suivant. L'exposé ci-dessous part du principe qu'il est possible d'obtenir des prévisions satisfaisantes à partir d'un modèle de type Holt-Winters optimisé pour le CO_2 dans le bureau individuel.

Les paramètres α, β^* et γ du modèle HW appliqués aux séries du formaldéhyde des différents environnements sont les mêmes que ceux appliqués aux séries des PM dans l'espace paysager durant la campagne de 2015.

Sur la Figure 6.5.5, on présente les résultats de la prévision de HCHO obtenus avec le modèle HW en fixant les paramètres optimisés : $\alpha = 0.00033$, $\beta^* = 0.003$ et $\gamma = 0.0095$. La Figure 6.5.5a donne les séries temporelles de test comparées aux séries temporelles de prévision. La Figure 6.5.5b illustre la qualité de prévision mesurée par deux indices : RMSE et MAE.

En imposant une fréquence principale d'un jour^{-1} dans le modèle HW, les résultats de la prévision sont sans surprise pour la série du HCHO dans la maison expérimentale. Néanmoins, le paramétrage du modèle ne détecte pas la tendance à la hausse après le début du troisième jour de prévision. Les indices de performances de la prévision, exprimés par le MAE et le RMSE, donnent un écart inférieur à 1.5 ppb sur les premiers 30000 minutes.

Le même paramétrage du modèle HW est appliqué aux mesures du HCHO durant la campagne de 2013. Les prévisions obtenues par ce dernier sur un horizon de 1000 minutes s'écartent légèrement des valeurs observées : l'écart moyen est d'environ 1.73 ppb sur l'ensemble de la série test. On remarque néanmoins que la variation abrupte, survenue après l'horizon de 10 h, a été détectée avec la période principale de 4 jours imposé au modèle. Cette performance tient principalement à la caractéristique de la fréquence principale. Pour une autre série test, on aurait pu voir une prévision médiocre, sauf si on réajuste la fréquence saisonnière sur une autre valeur.

Comme évoqué en début de cette section, la variabilité du HCHO durant la campagne de 2015 est caractérisée par une oscillation diurne et hebdomadaire importante. En initialisant le modèle HW avec cette dernière période, la prévision de toutes les 20 min montre une très bonne approximation des valeurs observées sur un horizon de deux jours de prévision. Le paramétrage du modèle est très stable, reproduisant ainsi la variation diurne et le faible changement abrupt, mais à partir de cette période, *i.e.* 2 jours, le modèle diverge de la valeur moyenne de la série test et surestime les fluctuations. L'écart moyen entre la série prédite et la série test est de 1.18 ppb, très faible par rapport au niveau global des fluctuations.

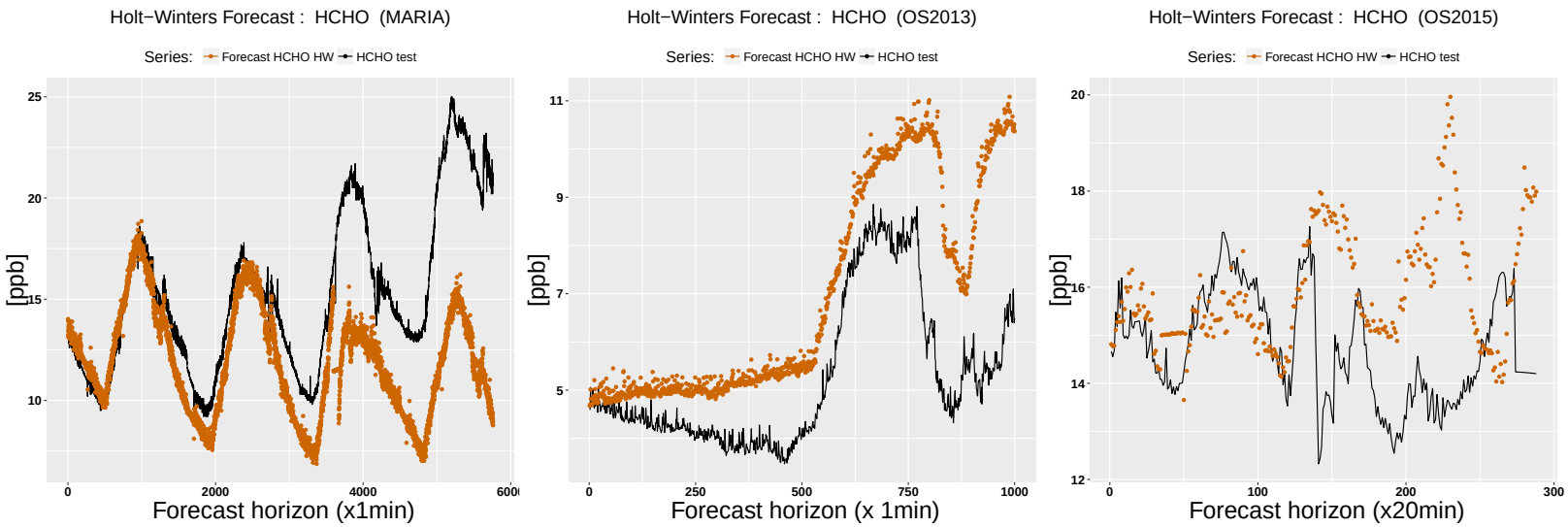
En ce qui concerne les mesures de performance, exprimées en RMSE et MAE, elles révèlent une augmentation graduelle de l'erreur de prévision. Au contraire, la courbe des indices symétriques pour les particules fines est inversée pour les prévisions de 2015. Pour le formaldéhyde dans l'espace de bureaux, les performances du même modèle montrent une différence entre les deux périodes de mesure (2013 et 2015). En fait, deux hypothèses peuvent expliquer cette différence :

— *i)* la prévision des fluctuations de HCHO en 2013 est plus difficile : plusieurs facteurs perturbent la trajectoire de la série de mesure. Le modèle ne peut pas mettre en évidence tous ces facteurs ;

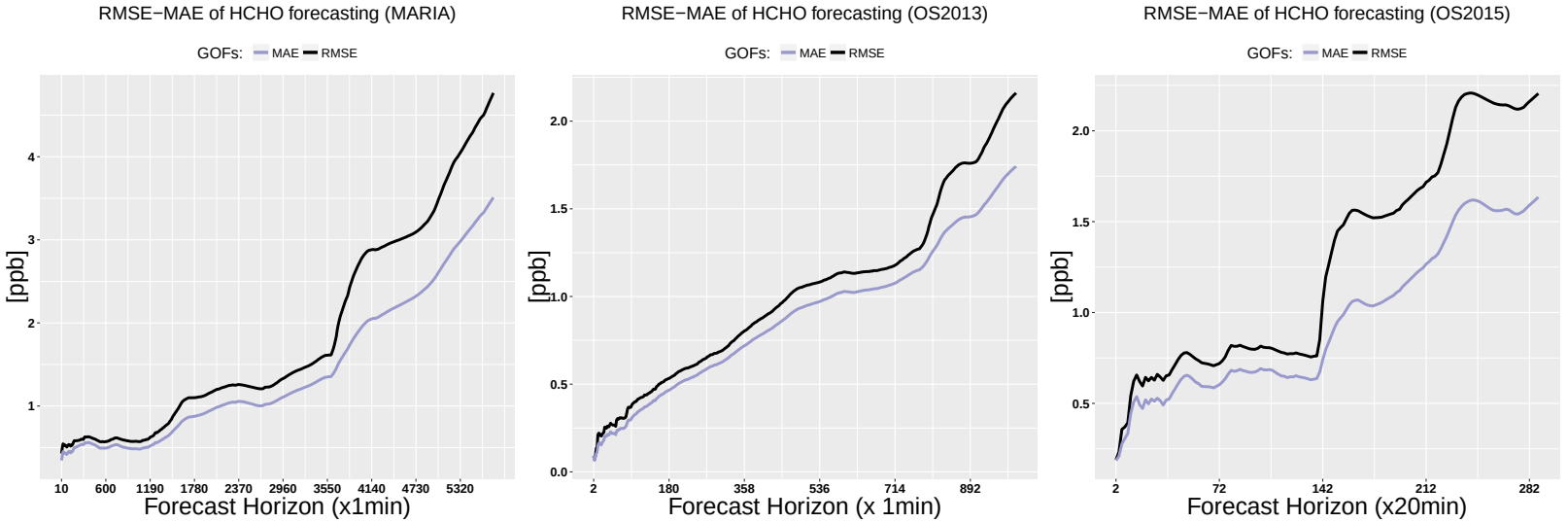
— *ii)* le pas de temps utilisé en 2015, qui est de 20 minutes, réduit le niveau de variabilité, donc peu de perturbations aléatoires, tandis qu'en 2013, le nombre de pas en prévision est très grand, ce qui a pour effet une transmission de l'erreur à chaque pas de prévision.

Malgré l'importance du point *ii),* il nous semble peu vraisemblable pour le modèle HW. La transmission d'erreur se manifeste d'avantage dans l'intervalle de prévision que sur le niveau de la série (moyenne).

Au vu de *i),* on pense que la nature inhérente à la série originale générée dans des conditions particulières est en cause de cette différence. Pour s'en convaincre, la visualisation de la série issue des mesures dans la maison expérimentale et sa prévision montrent qu'il existe très peu de facteurs aléatoires qui éloignent la série de sa trajectoire moyenne.



(a) *Prévision des concentrations de HCHO dans les différents espaces intérieurs par le modèle HW.*



(b) *Qualité de prévision exprimée par le MAE et le RMSE du modèle HW appliqué aux différentes séries de HCHO.*

FIGURE 6.5.5 – *Prévision des concentrations du HCHO avec la méthode de* HOLT-WINTERS *en utilisant les paramètres :* $\alpha = 0.00033$, $\beta^* = 0.003$ et $\gamma = 0.0095$. *dans les environnements suivants : maison expérimentale (MARIA), espace paysager en 2013 (OS2013) et en 2015 (OS2015)*

6.6 Applications des modèles de décomposition sur les données de concentrations des polluants de l'air intérieur

La prévision par une décomposition part de l'idée selon laquelle, une série temporelle peut être mieux prédite avec le traitement séparé de ses composantes latentes.

Plusieurs approches seront envisagées :

- une décomposition classique de type STL+ARIMA qui consiste en :
  - prévision de la composante saisonnière en projetant la composante saisonnière sur l'horizon futur.
  - prévoir le reste de la désaisonnalisation par un modèle ARIMA.
- une décomposition et prévision par SSA.

Cette section a pour objectif de mettre en évidence les caractéristiques de ces modèles ainsi que leur pouvoir prédictif.

6.6.1 Prévision par décomposition STL+ARIMA

La méthode STL+ARIMA a été appliquée aux différentes séries de HCHO ; d'abord, sur les mesures effectuées dans la maison expérimentale, ensuite sur les mesures effectuées dans l'espace paysager durant les campagnes de 2013 et de 2015.

De plus, la même procédure a été testée sur les séries de mesures des concentrations horaires en nombres de particules dans l'espace paysager durant la campagne de 2012. On se réfère aux deux tableaux 6.4.1 et 6.5.1 pour les informations relatives au découpage de la série initiale en ensemble d'apprentissage et de test, ainsi qu'à l'extraction de la composante saisonnière par la fréquence principale utilisée.

6.6.1.1 Résultats de prévision

Sur la Figure 6.6.1, on présente les séries de prévision, l'intervalle de prévision associé, ainsi que les performances des modèles. En ce qui concerne les ordres des modèles obtenus, toutes les séries désaisonnalisées sont intégrées à l'ordre 1 : les séries sont stationnaires avec le premier filtre de différenciation. La non-stationnarité de la série désaisonnalisée peut être expliquée par le fait que la tendance polynomiale modifie, au cours du temps, les propriétés statistiques de deuxième ordre (espérance et les autocovariances). Pour l'obtention des paramètres des modèles ARIMA, il apparaît que la résolution temporelle et la taille de la série y jouent un rôle important. En effet, pour les séries au pas de temps d'une minute, l'ordre de AR varie entre 2 à 5, obtenu par la minimisation du critère AIC ; alors que pour la série au pas de temps de vingt minutes, l'ordre d'autorégression est nul. Quant à la partie moyenne mobile, les coefficients interviennent sur toutes les séries.

Tout comme les autres modèles présentés précédemment, la prévision de la série du HCHO dans la maison expérimentale est la plus précise par rapport aux séries prédites de l'espace paysager. En effet, l'aspect sinusoïdal est prédit avec moins d'erreur, et ceci malgré le nombre de minutes prédites, mais la tendance associée à la série test est mal interceptée par la prévision (*cf.* Figure 6.6.1a).

Bien que la prévision ait été faite que sur 1000 minutes, la prévision de la série durant la campagne de 2013 reste acceptable sur un horizon de 10 h de prévision. Néanmoins, le changement abrupt a été sous-estimé par le modèle, contrairement au modèle de lissage, qui surestime cette variation. Pour remédier à cette mauvaise spécification, nous développerons dans le chapitre suivant une méthode de prévision plus appropriée.

Le modèle de prévision pour la série de 2015 montre qu'elle est stable en moyenne et récupère la variation diurne associée la variabilité globale. Ce trait montre en outre, l'importance de la spécification de la composante régulière dans le processus de prévision. Nous avons soulevé le problème de la fréquence qui traduirait la saisonnalité sur les données hautes fréquences.

L'intervalle de prévision graduel (55% à 99%) est obtenu uniquement sur le modèle ARIMA. Clairement, il dépend principalement du nombre de points de la série de prévision ainsi que de l'erreur de prévision associée. Notons aussi que cet intervalle est symétrique, donc sur certains horizons de prévision, on pourrait s'attendre qu'il nous livre des valeurs négatives sur la borne inférieure. C'est le cas observé au bout d'un jour et demi de prévision dans la maison expérimentale. Bien que l'étendu temporel de l'échéance de prévision soit le même avec la série de 2015 (h =4 jours), la borne inférieure associée à la prévision de cette dernière est toujours positive. À ce niveau, on voit clairement que la résolution temporelle amplifie l'erreur de prévision et donne, par conséquent, des valeurs "aberrantes" physiquement.

Pour palier à ce problème, il faut faire un choix, qui peut, par ailleurs être décisif :

— accepter une prévision à long terme, mais en gardant à l'esprit l'importance de l'incertitude liée à cet horizon ;

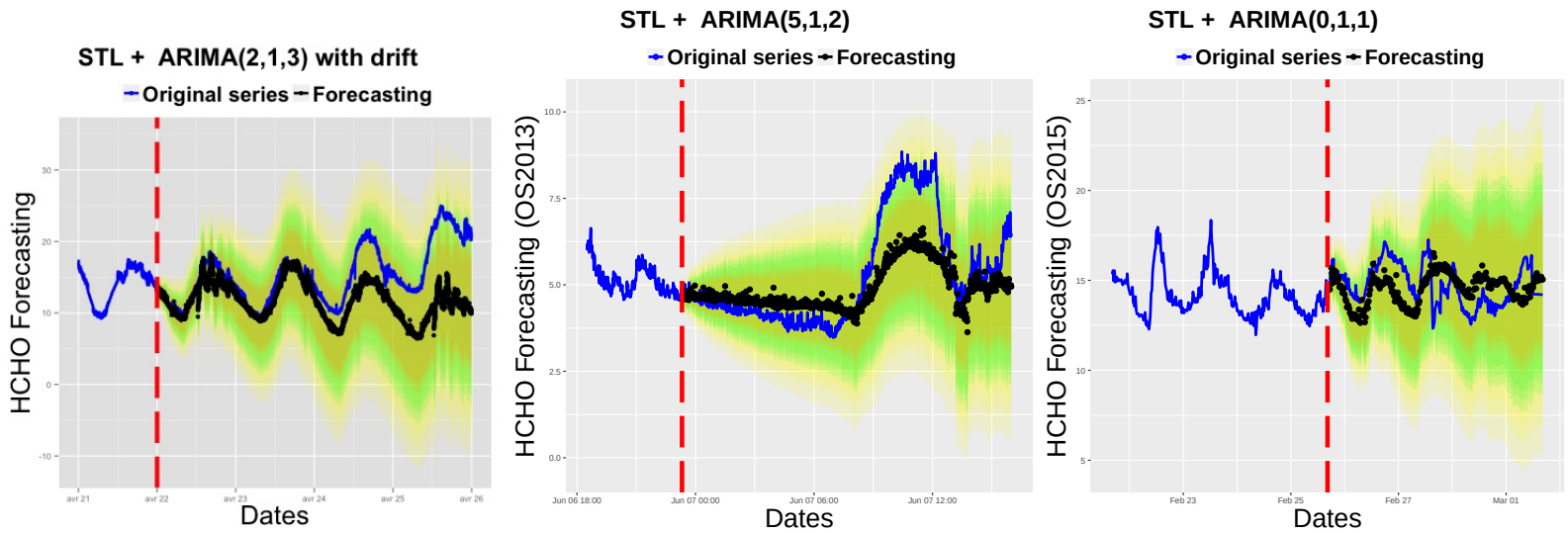
— tenir compte uniquement des prévisions à très court terme, mais que le modèle garantit "*l'efficience calculatoire*[2]" .

On donne dans la Figure 6.6.1b les performances prédictives du modèle STL+ARIMA sur les différentes séries utilisées. On distingue trois comportements des critères RMSE et MAE en fonction de l'horizon de prévision. Les zones ombrées R_1, R_2 et R_3 désignent les zones caractéristiques de l'évolution des indices. L'évolution globale de ces indices est graduelle et ils sont très corrélés entre eux : plus on s'éloigne de la première prévision, plus l'erreur augmente.

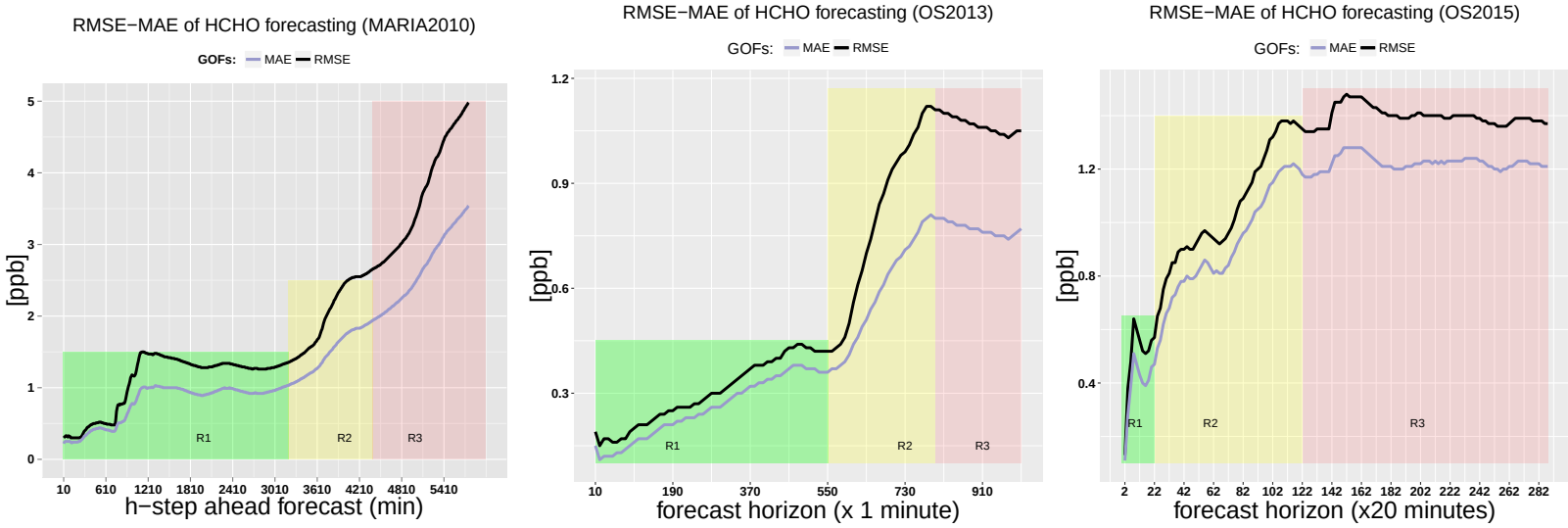
Avec ce modèle, on obtient un RMSE inférieur à 1 ppb après 12 h de prévision pour les séries au pas de temps d'une minute, et inférieur à 1.42 ppb les 4 jours de prévision pour la série de pas de temps de 20 minutes. Ceci montre, encore une fois, l'importance de la résolution temporelle utilisée.

On note que les deux indices, RMSE et MAE, donnent presque les mêmes valeurs sur le très court terme (8 h) et s'écartent pour un h grand. D'ailleurs, cette observation se précise d'avantage au moment de la variation abrupte de la série de 2013.

2. Je défini ici l'efficience calculatoire par le temps de réponse du système de prévision avant l'échéance de l'horizon de prévision.



(a) *Prévision des concentrations de HCHO dans la maison expérimentale (à gauche) et dans l'espace paysager (au centre et à droite) par un modèle de décomposition de type STL combiné avec un modèle ARIMA. La ligne en pointillée (rouge) sépare la partie estimation de la partie prévision : à partir de cette date, on donne la série test* (**en bleu**) *et sa prévision* (**en noir**).



(b) *Performances en termes de RMSE et MAE du modèle STL+ARIMA en prévision des concentrations du HCHO dans la maison expérimentale (à gauche) et dans l'espace paysager (à droite).*

FIGURE 6.6.1 – *Prévision des concentrations de HCHO et performances du modèle STL+ARIMA. La partie ombrée des prévisions est associée à l'intervalle de prévision obtenu par la modélisation ARIMA.*

La technique de prévision par décomposition STL a été aussi appliquée aux différentes séries temporelles de particules. Ces séries sont issues des mesures de concentration en nombre dans l'espace paysager durant la campane de 2012. On tente de prévoir quatre jours sur des séries de concentrations horaires de trois mois.

Comme évoqué dans le chapitre (2) de présentation des données, les séries temporelles en nombre de particules exhibent généralement une forme de densité de probabilité de type log-normal. L'examen graphique de ces séries montre que la variance varie au cours du temps : des moments de fortes agitations et des moments de variabilité calme, toutes sur une tendance polynomiale. Afin de réduire l'effet de ces caractéristiques sur la non-stationnarité, un prétraitement par transformation logarithmique nous semble nécessaire.

Nous présentons donc les prévisions des séries temporelles transformées, mais les indices de performances ont été calculés à partir des données initiales. Rappelons la formule de retour (donnée dans 5.3.14) en valeurs d'origine :

$$\widehat{X_t}(h) = \exp\left(\widehat{Y_t}(h) + \frac{\sigma_\varepsilon^2}{2}\sum_{j=0}^{h-1}\psi_j^2\right), \tag{6.6.1}$$

où $\widehat{Y_t}(h)$ sont les valeurs prédites à l'horizon h sur l'échelle logarithmique.

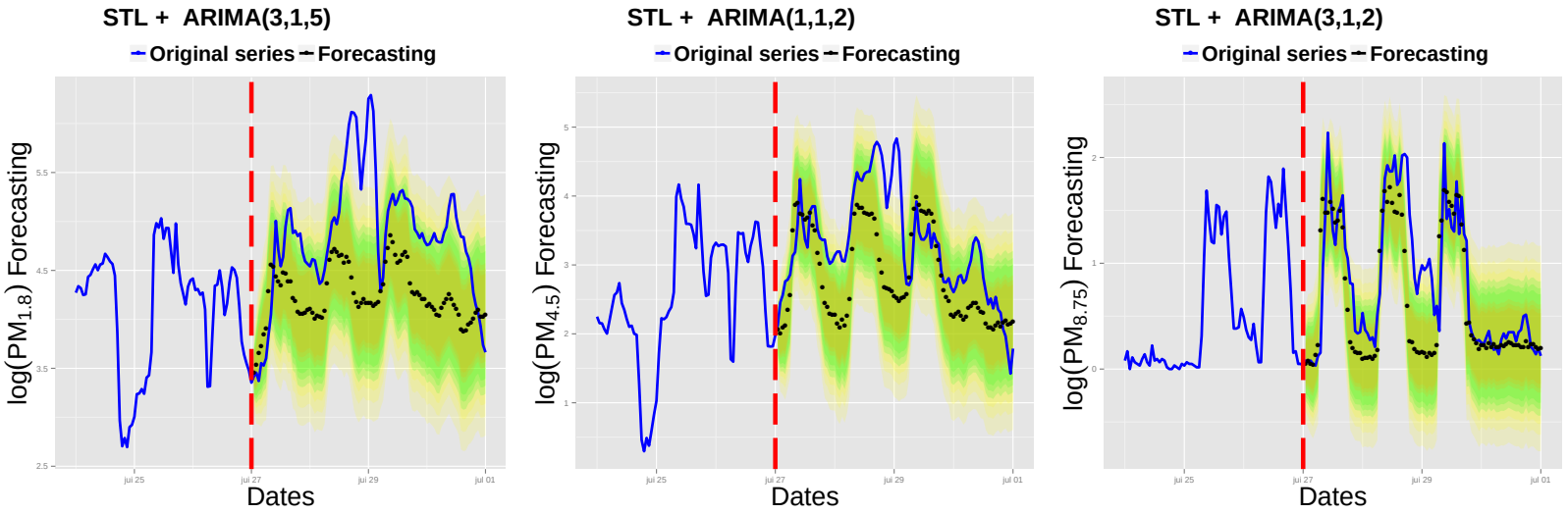
La Figure 6.6.2 donne les prévisions par le modèle STL+ARIMA des concentrations horaires en nombre de particules sur trois gammes : 1.8, 4.5 et 8.75 μm. L'horizon de prévision était fixé à 4 jours.

on remarque que plus les particules sont grosses, plus le modèle de prévision reconnaît la régularité saisonnière de type diurne.

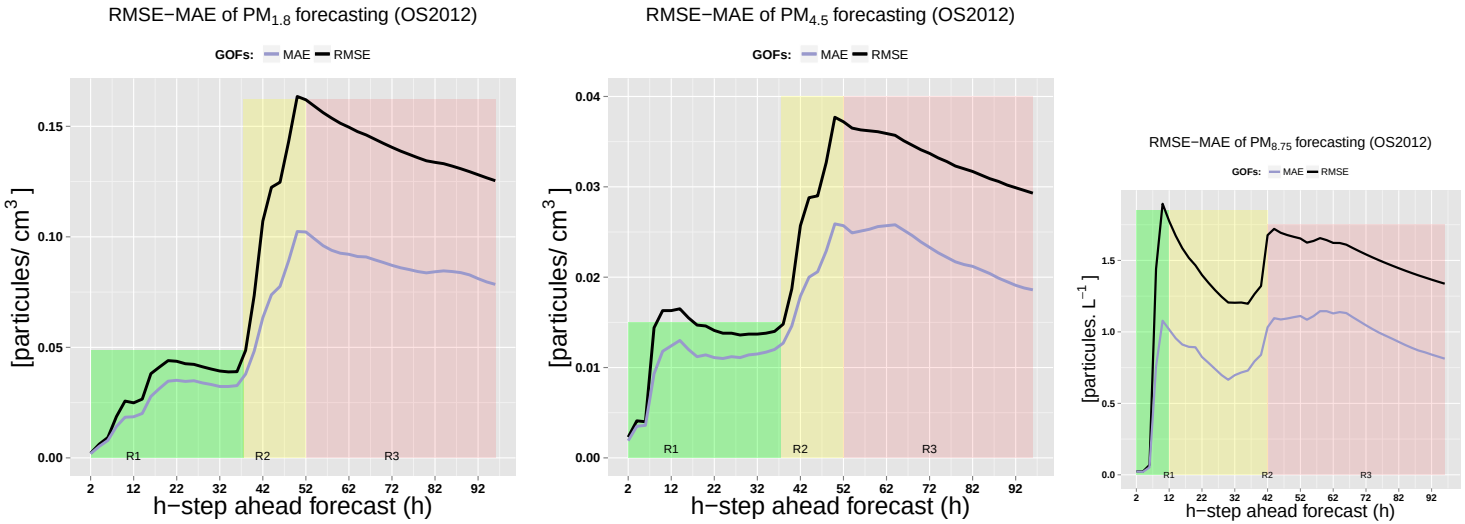
Par contre, en dépit d'une bonne approximation des concentrations de particules de taille 1.8 μm sur le très court terme ($<$ 10 h), le modèle perd très vite en précision quand l'horizon de prévision augmente. Notons néanmoins que l'oscillation moyenne est plus ou moins détectée, certes en gommant complètement les fortes variations abruptes. Pour les particules de tailles inférieures à 1.8μm, la série test chevauche la borne supérieure de l'intervalle de prévision, elle n'est donc pas contenue dans l'intervalle.

La prévision des concentrations de particules de taille 8.75 μm est juste sur les quatre jours. Le modèle est capable de reconnaître les épisodes de forte variabilité et les épisodes de variabilité faible. Cette prévision peut être expliquée par la présence d'une régularité, essentiellement diurne et hebdomadaire, que la décomposition STL à pu extraire.

Pour le calcul des indices de la qualité de prévision, les valeurs prédites en logarithme ont été ensuite exprimées en valeurs d'origine par la formule 6.6.1, ensuite comparées avec la série de test. Seules les premières 22 h de prévision peuvent être considérées acceptables pour les particules inférieures à 4.5 μm. Au contraire, avec une erreur moyenne absolue inférieure à une particule/L, la prévision des PM de 8.75μm sur les quatre jours de prévision est correcte.



(a) *Prévision des concentrations des particules de taille* 1.8, 4.5, *et* 8.75 μm *dans l'espace paysager par un modèle de décomposition de type STL combiné avec un modèle ARIMA. Trois jours avant la première prévision sont présentées. La ligne en pointillées rouge sépare la prévision de l'apprentissage.*



(b) *Performances en terme de RMSE et MAE du modèle STL+ARIMA en prévision des concentrations de particule de tailles* 1.8, 4.5, *et* 8.75 μm *l'espace paysager durant la campagne de 2012.*

FIGURE 6.6.2 – *Prévision des concentrations horaires en nombre des particules par le modèle STL+ARIMA dans l'espace paysager durant la campagne de 2012. .*

6.6.1.2 Discussion sur la spécification de ARMA et de l'hétéroscédasticité conditionnelle autorégressive des résidus

la spécification des ARMA Malgré toutes les tentatives effectuées dans la spécification ARIMA des séries désaisonnalisées et filtrées par une différence, on constate souvent quelques éléments qui rendent difficile le choix du modèle pour ce type de données :

- *i)* En adoptant la procédure "quasi-automatique" proposée dans (Hyndman & Khandakar, 2008; Hyndman & Athanasopoulos, 2014), certains coefficients d'ordre supérieur ne sont pas significativement différents de zéro ($t - student$ inférieur à 1.96), voir le Tableau 6.6.2. Cette procédure "automatique" souffre d'une mauvaise spécification très claire qui tend à rejeter "systématiquement" tous les modèles. D'ailleurs, cette observation va dans le sens des commentaires soulevés récemment par Shumway & Stoffer (2013).
  Au contraire, si on réduit les ordres du modèle, c'est-à-dire qu'on accepte uniquement les coefficients statistiquement significatifs, les résidus de la régression présentent une particularité très erratique au niveau de la variabilité de la variance (volatilité), ce qui rend plus problématique la procédure de choix.
- *ii)* Les résidus ainsi extraits, présentent des structures de variabilité qui "trahissent" l'hypothèse de bruit blanc Gaussien et souvent, de nature hétéroscédastique.

Pour les données de type hautes fréquences, il est clair que le modèle ARMA intervient uniquement dans la stabilisation moyenne des prévisions ainsi que dans le calcul des intervalles de prévision. En effet, la prévision par ARMA converge rapidement quand l'horizon de prévision augmente. C'est précisément ce que nous voulions éviter étant donné le pas de temps des séries.

Quoi qu'il en soit, la modélisation de telles séries est un problème complexe ; la difficulté n'est pas seulement due à la diversité des environnements étudiés, du type de polluant ou à l'importance de la fréquence d'observation. Elle tient surtout, d'après ce qu'on a pu observer, à l'existence des patterns statistiques inhérents à QAI. Par exemple, les structures de la mémoire et "l'explosion" du spectre au pôle de la fréquence zéro nécessitent des traitements plus adéquats. Par exemple, on présente sur la Figure 6.6.3 les séries résiduelles de la modélisation STL+ARIMA, leurs fonctions d'autocorrélation ainsi que leurs densités de probabilités.

Les statistiques du test de normalité des résidus et les $p - value$ sont présentées dans le Tableau 6.6.1. Les hypothèses alternatives pour chaque test sont :

- Omnibus, H_1 : *la distribution n'est pas normale par son aplatissement ou par son asymétrie.*
- Skewness, H_1 : *la distribution n'est pas normale par son asymétrie.*
- Kurtosis, H_1 : *la distribution n'est pas normale par son aplatissement.*

Pour ces tests, la $p-value$ est nulle, on rejette alors l'hypothèse de normalité des résidus de la régression. Cette constatation est probablement due à un excès d'aplatissement plutôt qu'à une asymétrie (*cf.* les densité de probabilités dans Figure 6.6.3).

Faits stylisés

Plusieurs traits, qu'on appellera "*faits stylisés*"[3], se dégagent de l'examen graphique :

1. *Regroupement des extrêmes.* On remarque des sous-périodes de forte agitation résiduelle, suivies par un groupement de périodes dont la variabilité est très faible. La volatilité[4] temporelle, qui

3. Ce terme est emprunté de la mathématique financière.
4. On trouve le terme "volatile" pour les espèces chimiques (COV), mais là il s'agit d'une caractéristique purement statistique liée à la variabilité de la variance conditionnelle.

TABLE 6.6.1 – Test de normalité des résidus de la modélisation des séries désaisonnalisées du polluant HCHO

$\widehat{\varepsilon_t}$ de STL+ARIMA	Statistique	Test	Stat. de test	p-value
HCHO 2013	χ^2	Omnibus	3520.53654	0.00
	Z3	Skewness	34.46281	0.00
	Z4	Kurtosis	48.2996	0.00
HCHO 2015	χ^2	Omnibus	429.4881	0.00
	Z3	Skewness	10.85423	0.00
	Z4	Kurtosis	17.65428	0.00
HCHO MARIA	χ^2	Omnibus	1566.405348	0.00
	Z3	Skewness	4.203565	0.00
	Z4	Kurtosis	39.353975	0.00

se traduit par ces aspects de fluctuation, met en évidence la non-constance de la *variance conditionnelle* : la probabilité d'observer de fortes valeurs de $\widehat{\varepsilon_{t-1}}^2$ semble augmenter la probabilité d'observer de fortes valeurs pour $\widehat{\varepsilon_t}^2$ (l'hétéroscédasticité conditionnelle).

2. *Queues de distribution épaisses.* Les distributions de probabilités montent de très forts pics en zéro : distribution leptokurtiques. Les tests classiques de normalité (ci-après) tendent à rejeter nettement l'hypothèse d'une distribution normale.
3. *Autocorrélations faibles.* Les autocorrélations observées sont généralement très faibles, surtout pour les séries au pas de temps de 20 minutes. Ce fait peut nous renvoyer à l'hypothèse de bruit blanc des résidus, mais on constate qu'il existe des autocorrélations ponctuelles significativement différentes de zéro sur des retards faibles pour la série dans la maison expérimentale (MARIA) et lointains pour les séries de l'espace paysager.

Lorsque la différentiation première est appliquée sur les séries brutes, le filtre absorbe complètement toutes les composantes déterministes, produisant ainsi les faits stylisées listés ci-dessus. Ils sont en outre, souvent plus prononcés, et ceci d'autant plus avec la résolution des observations fines.

La non-linéarité en variance : l'hétéroscédastité conditionnelle

Compte tenu des différentes structures de variabilité et faits stylisés observés dans de nombreuses séries temporelles de polluants de type hautes fréquences, on soupçone l'existence d'une structure hétéroscédastique des erreurs. Avant de rappeler à cet effet le test ARCH basé sur le multiplicateur de Lagrange, nous présentons brièvement les modèles ARCH.

Introduit par Engle (1982) [5] lors d'une étude sur la variance de l'inflation en Grande Bretagne, les modèles ARCH ont connu très vite un large succès. Les modèles ARCH (q) sont basés sur la paramétrisation quadratique de la variance conditionnelle. Le σ_t^2 apparaît comme fonction linéaire du processus du carré des innovations ε_t.

5. Les faits stylisés observés dans notre cas ressemblent étrangement à ceux observés dans d'autres domaines. Alors, encore une fois, nous empruntons le vocabulaire de "l'économétrie financière" à *l'environnemtrie* intérieur. Rappelons que l'auteur des modèles ARCH, ROBERT FRY ENGLE a été (conjointement avec CLIVE WILLIAM JOHN GRANGER) récompensé du prix nobel en 2003.

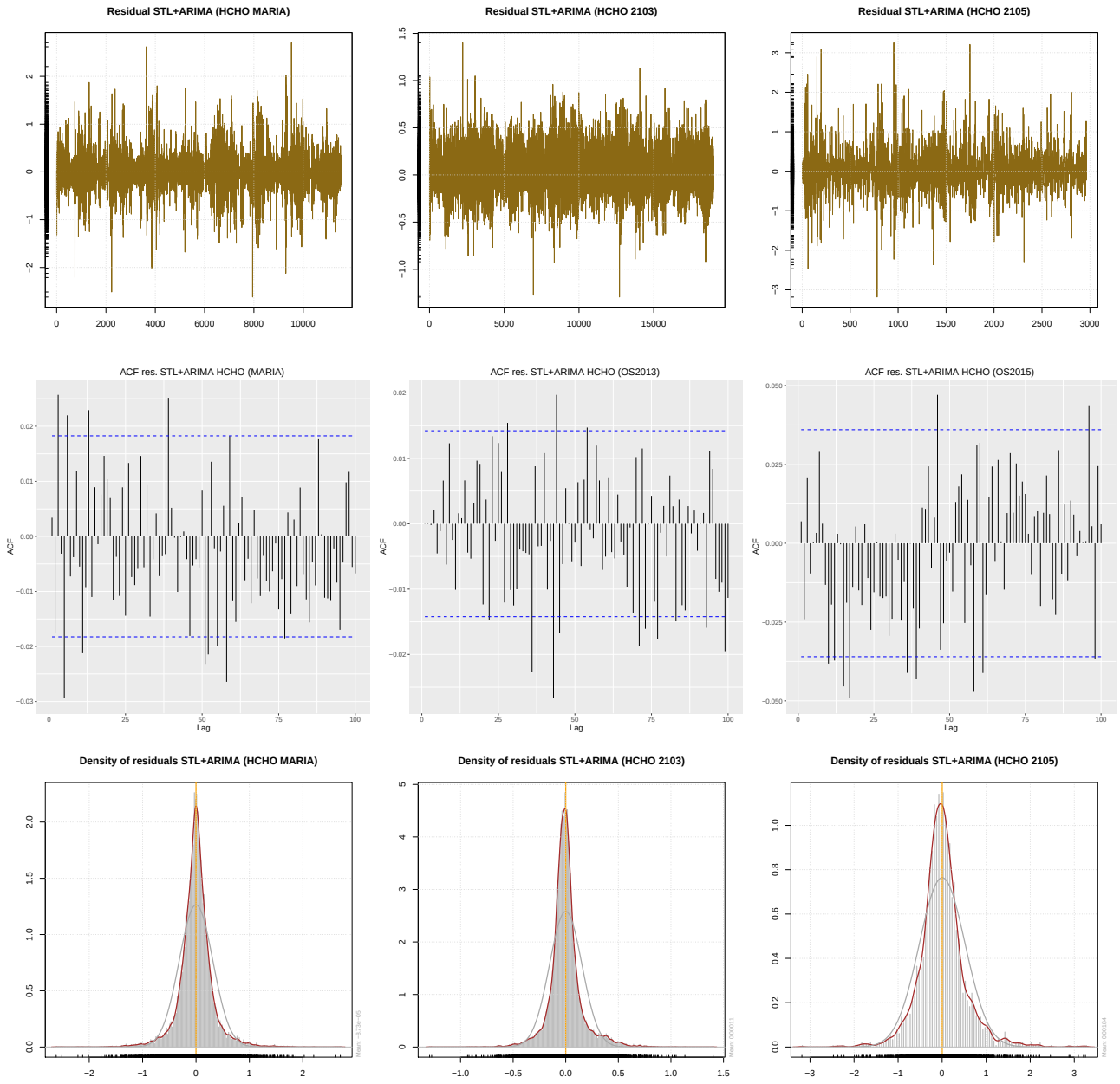


FIGURE 6.6.3 – *Diagnostic sur des résidus de la modélisation STL+ARIMA pour les séries de HCHO. Le premier panel (en ligne) donne les séries temporelles des résidus, le panel du milieu représente les ACF des résidus au 100 premiers retards, le dernier donne leurs densités de probabilités ainsi l'ajustement des Gaussiennes associées.*

Définition 6.6.1. *Les processus AutoRegressive Conditional Heteroscedasticity) ARCH (q)*
Pour les réels $\alpha_0 > 0$ et $\alpha_i > 0$, $\forall i$, un processus ARCH (q) est donné par

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-i}^2, \tag{6.6.2}$$

$$= \alpha_0 + \alpha(L)\,\varepsilon_t^2. \tag{6.6.3}$$

Le modèle défini dans 6.6.2 permet de prendre en compte les clusters (regroupements) de hautes et basses volatilités, mais dont le signe reste imprévisible. Ainsi, dans cette configuration le carré des perturbations suit un processus autorégressif d'ordre q. Ce premier modèle à variance conditionnellement hétéroscédastique a été rapidement généralisé par Bollerslev (1986) en établissant le modèle $GARCH\,(p,q)$ (Generalized AutoRegressive Conditional Heteroskedasticity). Cette extension consiste en l'introduction de valeurs retardées de la variance dans son équation et est, de ce fait, similaire à l'extension des modèles AR aux $ARMA$. Pour une description plus complète, on peut consulter les monographies de Francq & Zakoian (2011) et de Bollerslev et al.(2010) (Chapitre 8 : *Glossary to ARCH (GARCH)*) ainsi que leurs références.

Définition 6.6.2. *Les processus GARCH (p, q)*
On dit que (ε_t) est un processus $GARCH(p,q)$ si ses deux premiers moments conditionnels existent et vérifient

- *i)* $\mathbb{E}(\varepsilon_t \mid \varepsilon_u,\ u < t) = 0, \qquad t \in \mathbb{Z}$;
- ii) *Il existe des constantes* α_0, α_i, $i = 1, \ldots, q$ et β_j, $j = 1, \ldots, p$ *telles que*

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-i}^2 + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^2, \quad t \in \mathbb{Z}. \tag{6.6.4}$$

Pour le test ARCH, supposons que l'équation de la moyenne soit décrite par un processus $ARMA$ et les perturbations par le modèle $ARCH$:

$$\begin{cases} \Phi(L) = \Theta(L)\,\varepsilon_t, \\ \sigma_t^2 = \alpha_0 + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-i}^2. \end{cases} \tag{6.6.5}$$

On teste la significativité de la deuxième régression dans 6.6.5, dont on veut tester l'hétéroscédaticité. L'hypothèse nulle testée est celle d'homoscédasticité : $\alpha_1 = \alpha_2 = \cdots = \alpha_q = 0$ contre l'hypothèse alternative d'hétéroscédaticité : au moins un coefficient $\alpha_i\,(i = 1, \cdots, q)$ est différent de zéro. Si l'hypothèse nulle est acceptée, alors la variance conditionnelle est constante : $\sigma_t^2 = \alpha_0$. Au contraire, si l'hypothèse nulle est rejetée, les résidus suivent un processus $ARCH\,(q)$.

Pour nous, le test est effectué sur les résidus $\widehat{\varepsilon_t}$ de la première équation de 6.6.5. Sa mise en œuvre peut s'effectuer en plusieurs étapes :

TABLE 6.6.2 – Estimation du modèle STL+ARIMA à erreurs ARCH.

	Coefficients	HCHO 2015	HCHO 2013	HCHO maria
ARMA	$\hat{\varphi}_1$	-	**1.21**	**1.12**
	(s.e)		(0.0321)	(0.0716)
	$\hat{\varphi}_2$	-	-0.42	-0.52
	(s.e)		(0.0172)	(0.0798)
	$\hat{\varphi}_3$	-	0.054	–
	(s.e)		(0.0133)	
	$\hat{\varphi}_4$	-	-0.022	-
	(s.e)		(0.0118)	
	$\hat{\varphi}_5$	-	-0.003	-
	(s.e)		(0.0107)	
	$\hat{\theta}_1$	-0.43	**-1.61**	**-1.64**
	(s.e)	(0.0167)	(0.0313)	(0.0749)
	$\hat{\theta}_2$	-	0.69	1.01
	(s.e)		(0.022)	(0.1264)
	$\hat{\theta}_3$	-	-	-0.1939
	(s.e)			(0.0522)
	Drift	-	-	-0.0006
				(0.0013)
	AIC	4560	-17044.3	6129.85
	AICc	4560	-17044.3	6129.86
	BIC	4572	-16981.5	6181.31
	RMSE	0.52	0.154	0.315
LM	χ^2	207.37	719.48	1322
	df	10	12	12
ARCH TEST	*p-value*	0.00	0.00	0.00

Notes : les cellules ombrées en gris représentent les ordres dans les modèles dont les coefficients sont non-significatifs (à 5%). Pour la série de 2013 il est beaucoup plus difficile de trouver un compromis entre la significativité des coefficients tout en respectant les hypothèses liées à la structure des résidus. Les ***coefficients en gras*** *donnent une idée sur le caractère erratique de la série : problème de stationnarité.*

1. calculer le carré de la série des résidus : $\widehat{\varepsilon_t}^2$.
2. régresser $\widehat{\varepsilon_t}^2$ sur une constante et ses q valeurs passées, dont seuls les retards significatifs sont conservés (on se sert des PACF des carrés des résidus, *cf* : Figure G.1.1 de l'annexe G).
3. Calculer la statistique TR^2, où T est la taille de la série et le R^2 est le coefficient de détermination.

Sous l'hypothèse nulle d'homoscédasticité, la statistique TR^2 suit une loi de $\chi^2(q)$. Comme tout test basé sur la distance du χ^2 on rejette l'hypothèse nulle pour une distance élevée et la $p-value$ donne la probabilité de dépasser la distance observée si l'hypothèse nulle est vérifiée (*cf.* Tableau 6.6.2).

Dans le tableau 6.6.3, on donne l'estimation de la variance conditionnelle par le $GARCH(1,1)$. Souvent dans la littérature, on utilise au plus un ordre supérieur à 2 dans $GARCH(p,q)$. Donc le modèle final peut s'écrire comme STL+ARIMA+GARCH, on lit : décomposition par STL avec un modèle ARIMA de la série désaisonnalisée à erreur GARCH.

TABLE 6.6.3 – Estimation de variance conditionnelle par $GARCH\,(1,1)$

$GARCH(1,1)$		Estimate	Std. Error	t value	*prob*
HCHO2015	$\widetilde{\alpha}_0$	0.020137	0.002931	6.87	00.00
	$\widetilde{\alpha}_1$	0.157968	0.015665	10.084	00.00
	$\widetilde{\beta}_1$	0.77832	0.019018	40.925	00.00
HCHO2013	$\widetilde{\alpha}_0$	1.48E-04	1.73E-05	8.58	00.00
	$\widetilde{\alpha}_1$	1.80E-02	1.10E-03	16.32	00.00
	$\widetilde{\beta}_1$	9.76E-01	1.61E-03	606.55	00.00
HCHOMaria	$\widetilde{\alpha}_0$	0.0046479	0.0004242	10.957	00.00
	$\widetilde{\alpha}_1$	0.1242331	0.0094328	13.17	00.00
	$\widetilde{\beta}_1$	0.8302401	0.0120528	68.884	00.00

6.6.2 Prévision par décomposition Singular Spectrum Analysis (SSA)

Cette section présente les résultats des prévisions effectuées sur les séries du HCHO dans l'espace paysager par la procédure SSA.

Sur la Figure 6.6.4, sont présentées la série test avec une partie de la série d'apprentissage, la série prédite ainsi que les performances du modèle. Les prévisions sont obtenues avec la méthode présentée dans 6.3. Pour obtenir les intervalles de prévision, trente simulations ont été effectuées. On prend les centiles comme bornes inférieures et supérieures.

Pour la série de 2013, bien qu'on arrive à reproduire la variation globale, le modèle sur-estime les valeurs au début de la prévision (après 6 h de prévision) et la variation abrupte est mal interceptée. La prévision sur cette série ne dérive pas du minimum et du maximum observés. Les critères RMSE et le MAPE restent inférieurs à 1 ppb sur les 6 premières heures de prévision et se stabilisent à ce niveau sur l'ensemble de 16 h.

Pour la série de 2015, la prévision à très court terme reste précise, mais diverge rapidement après 1.5 jour. Ainsi, le modèle arrive à prévoir environ 33 h avec une erreur inférieure à 0.5 ppb et les erreurs restent en deçà de 1 ppb sur les deux premiers jours.

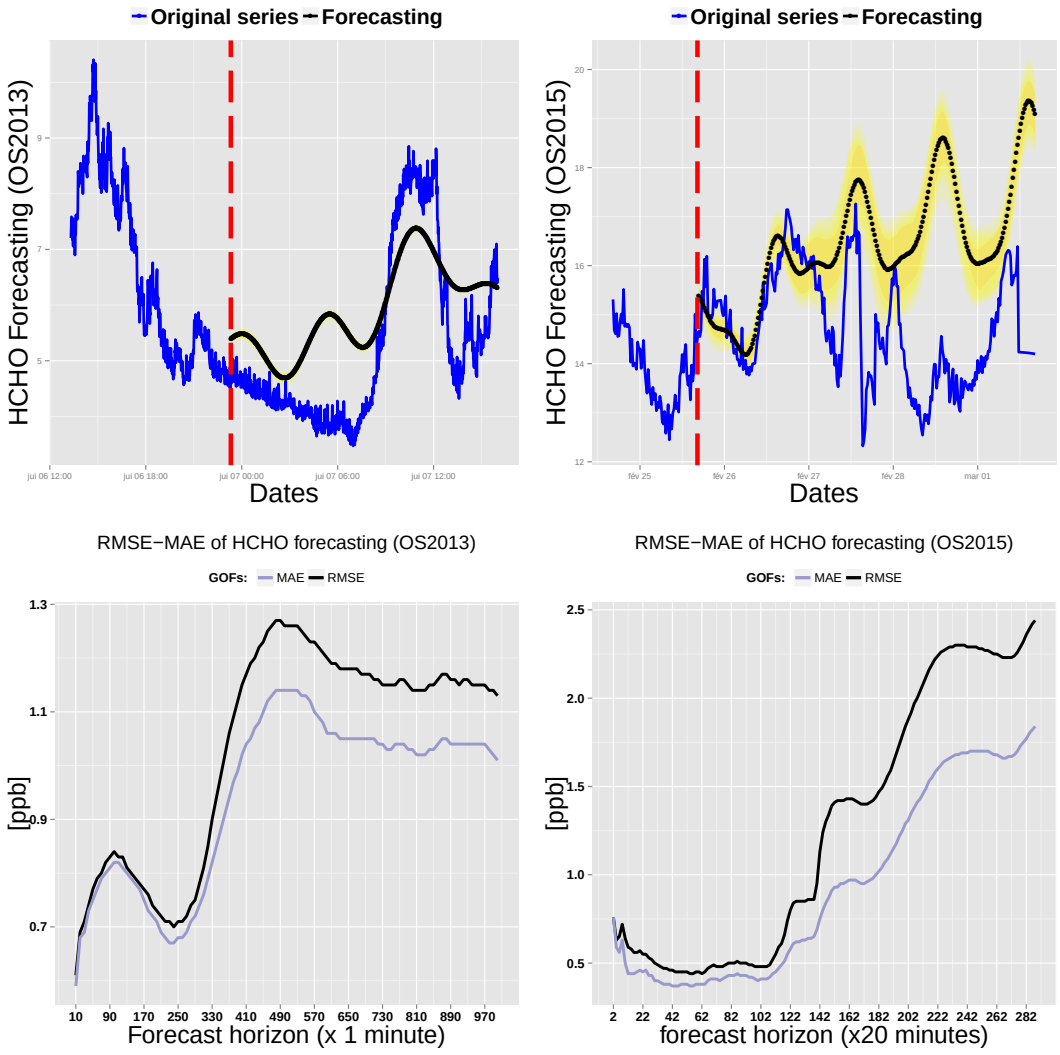


FIGURE 6.6.4 – *Prévision des concentrations de HCHO sur un horizon de quatre jours dans l'espace paysager durant la campagne de 2015 par la méthode SSA.*

6.7 Discussion, conclusions et perspectives

Au début de cette thèse et dans les conditions d'une littérature "quasi-inexistante" sur les modèles de prévision dans le domaine la QAI, nous avions une double intention derrière nos objectifs tracés :

- éclairer les principaux défis de la prévision de la qualité de l'air intérieur ;
- examiner les différentes approches possibles pour fournir des prévisions *bonnes* et *utiles* aux gestionnaires-décideurs.

Pour ces raisons, notre travail s'est appuyé sur les résultats des modèles de prévision obtenus dans différentes disciplines, tant au niveau théorique que pratique. Par exemple, l'analyse classique des séries temporelles par les modèles ARIMA-GARCH et le modèles non-paramétriques de prévision de type SSA.

Dans la modélisation STL+ARIMA à variance des erreurs hétéroscédastiques, les prévisions délivrées par un modèle type ARMA convergent rapidement vers la moyenne parce que l'on perd très vite de l'information sur les perturbations aléatoires quand l'horizon de prévision augmente. C'est précisément ce que nous voulions éviter étant donnée la résolution temporelle utilisée dans la plupart de nos séries. La qualité de prévision est donc déterminée par la qualité d'extraction des composantes déterministes, notamment la saisonnalité. Les erreurs de la régression ARIMA ont été modélisées par $GARCH\,(1,1)$ et le modèle général est donné par l'expression : STL+ARIMA+GARCH.

Finalement, c'est l'extraction et la prévision des composantes latentes qui rendent la prévision globale bonne, ou pas. Donc c'est la manière de décomposer la série qui détermine en amont le niveau de prédictibilité de ces dernières. Cette décomposition semble plus contraignante sur la séries du HCHO en 2013 et sur les concentrations des particules fines. Cela est dû à la nature de la variabilité de ces polluants.

Très peu utilisée dans la littérature, la méthode SSA non-paramétrique a été appliquée pour la prévision du HCHO dans l'espace paysager. Elle s'avère performante dans certains cas bien précis, notamment pour le HCHO en 2015. Le choix de la longueur de la fenêtre pour la matrice de délais est très important dans la récupération de la régularité au sein de la série. Encore une fois, c'est la période principale qui est recommandée dans l'analyse de la matrice de délais. Pour les séries ayant un spectre au pôle de la fréquence et le "bruit" qui se manifeste sur l'ensemble de la densité spectrale, le choix de L est très délicat. En fonction de la longueur des séries, nous proposons un L qui soit multiple de la période principale.

Les modèles de décomposition appliqués ici pour la QAI ne reposent pas forcément ni sur une construction théorique unique ni sur la théorie physique de l'environnement intérieur. L'extraction des composantes déterministes *via* une méthode statistique consiste à identifier l'équilibre de long terme (tendance) de la variabilité et la fluctuation oscillatoire clairement reconnaissable. Cette distinction ne suffit pas pour l'extraction des différentes composantes, elle nécessite donc des hypothèses supplémentaires.

Ce chapitre ne se propose que de poser quelques "jalons" ; les éléments de réponses pour certaines séries de la QAI ainsi ébauchés demanderont des études supplémentaires pour être précisées. Le chapitre suivant se propose de donner une autre décomposition avec une autre classe de modèle de prévision ; la non-linéairité et le chaos s'installent.

CHAPITRE 7

DÉCOMPOSITION EN BANDES SPECTRALES, MODÈLES NON-LINÉAIRES ET PRÉVISIONS

"All linear systems are the same. Each nonlinear system is nonlinear in its own way". Gilmore & Lefranc (2002) paraphrasant Tolstoy dans Anna Karénine : *"Les familles heureuses se ressemblent toutes ; les familles malheureuses sont malheureuses chacune à leur façon."*

Les *prévisions fournies par les différentes approches appliquées jusqu'à présent souffrent de l'incapacité à détecter les changements de forte amplitude. Dans ce chapitre, on propose une méthode pour remédier à ce problème. La procédure consiste à décomposer en bandes spectrales la série initiale en obtenant plusieurs composantes, ensuite à modéliser chaque composante par des modèles non-linéaires de séries temporelles, comme les modèles à changement de régime ou les modèles basés sur la théorie des systèmes dynamiques (théorie du chaos).*

Sommaire

7.1 Introduction

Nous avons vu dans le chapitre précédent que la variabilité des concentrations de formaldéhyde mesurées en 2013 (espace paysager) exhibe des phénomènes de rupture (des sauts abrupts) ; de manière plus précise, on peut l'appeler variabilité à changement de régime. De plus, de nombreuses séries temporelles de la QAI qu'on a pu observer présentent une densité spectrale croissante lorsque la fréquence tend vers zéro. La présence de ces caractéristiques nécessite donc une modélisation adaptée qui puisse prendre en compte les variations de forte amplitude.

Un aspect "trivial" de la non-linéarité réside justement dans l'abandon de l'hypothèse de stabilité de la série et la présence des comportements à mémoire longue. Dans le cadre des fluctuations des concentrations de polluants de l'air intérieur, on a pu observer des sauts abrupts, ce qui représente une particularité de la non-linéarité du signal. Par exemple, dans le cas d'un environnement normalement occupé, les valeurs de plusieurs paramètres sont modifiées au cours du temps de façon imprévisible. Souvent, l'amplitude de ces variations est élevée par rapport au niveau global de la trajectoire. D'un point de vue statistique, on peut associer à la variation abrupte la caractéristique d'un processus non-stationnaire en moyenne. Cette particularité des séries peut être identifiée par un basculement marqué d'un régime de variabilité à un autre pour une certaine période, à l'image d'un comportement dû à la succession de l'occupation-inoccupation ou à l'état des ouvrants par une suite de séquences de fermé-ouvert.

Le changement de régime s'opère généralement sur le niveau moyen de la série. Donc, dans le cadre des modèles non-linéaires paramétriques, seule sera traitée ici la modélisation de l'espérance conditionnelle. En particulier, nous n'aborderons pas ici les modèles non-linéaires en variance conditionnelle, comme les modèles GARCH et leurs dérivés ; nous les avons employés uniquement dans le cadre de l'analyse des résidus des modèles STL-ARMA.

Les modèles à changement de régime décrivent un ensemble d'états dans lesquels chaque portion de variabilité se révèle explicative du phénomène environnemental étudié. Si la succession de ces phases de variabilité se manifeste de façon régulière, celle-ci peut être attribuée à une forme de saisonnalité. En effet, la succession des états occupations-inoccupations dans un environnement réel serait la cause directe et indirecte d'au moins deux types de variation : un comportement quasi-saisonnier des concentrations du CO_2 et d'un comportement très aléatoire des concentrations du formaldéhyde. Pour ce dernier, la dynamique des concentrations de formaldéhyde est liée (d'une certaine mesure) aux variations de l'état des ouvrants qui sont déterminés par la succession des phases occupations-inoccupations. Loin d'une variation stable, on accepte d'écarter la linéarité des processus générateurs et donc leur unicité.

Ce chapitre est consacré à quelques outils permettant de quantifier les phénomènes de non-linéarité des séries temporelles, en se servant de deux facettes complémentaires : les modèles à changement de régime (paramétriques) et les modèles des systèmes dynamiques (non-paramétriques). Nous proposons l'utilisation de la décomposition en bandes spectrales et la modélisation de chaque bande par l'un de ces modèles.

7.2 Décomposition en Bandes Spectrales (SBD)

Cette section expose le principe de base de la décomposition en bandes spectrales **SBD** (pour Spectral Band Decomposition) des séries temporelles. Bien qu'elle soit simple dans sa construction et efficace dans les applications, cette méthode n'a pas fait l'objet d'une grande préoccupation dans la littérature des séries temporelles. On trouve néanmoins d'autres méthodes plus complexes (telle que la décomposition en ondelettes, qui peut être une alternative), mais qui ne seront pas présentées dans ce manuscrit.

Soient les observations $(x_t)_{0\leq t\leq T-1}$ du processus $(X_t)_{t\in\mathbb{Z}}$ ayant un spectre de fréquences W. La décomposition en bandes spectrales consiste à choisir un ensemble de m intervalles $B_1, B_2, \ldots, B_m$ contenant un certain nombre de fréquences, intervalles définis par les fréquences de coupure normalisées $0 = f_0 < f_1 < \ldots < f_m < \frac{1}{2}$, tels que

$$B_k = \{f \in W \mid f_{k-1} \leq f < f_k\}\,. \tag{7.2.1}$$

La $k^{ème}$ composante FFT, qu'on note (FFT_k) est une série temporelle de la même taille que la série d'observations et elle est obtenue en prenant la partie réelle de la transformée inverse, IFFT, correspondant uniquement à la bande de fréquences B_k. Plus précisément, soit $(y_j)_{0\leq j\leq T-1}$ la FFT de $(x_t)_{0\leq t\leq T-1}$:

$$y_j = \sum_{t=0}^{T-1} x_t e^{2\pi i \frac{jt}{T}}, \tag{7.2.2}$$

où $i^2 = -1$. On définit, sur les bandes de fréquences, les quantités y_t^k, comme suit :

$$y_t^k = \begin{cases} y_t & \text{si } n_{k-1} \leq t \leq n_k \text{ ou } T - n_k < t \leq T - n_{k-1} \\ 0 & \text{sinon ,} \end{cases} \tag{7.2.3}$$

où $n_k = \lfloor T \times f_k \rfloor$ est la position de l'indice pour la fréquence normalisée $f_k \in \left[0, \frac{1}{2}\right[$. Notons que $y_t = \bar{y}_{T-t}$, où $\bar{y}_{T-t}$ est le conjugué complexe de y_t, alors

$$x_t^k = \frac{1}{T} \sum_{t=0}^{T-1} y_t^k e^{-2\pi i \frac{kt}{T}} \tag{7.2.4}$$

est réel. La $k^{ème}$ série résiduelle $\left(e_t^k\right)_{0\leq t\leq T-1}$ issue du processus SBD est définie par la différence entre la série originale et la somme des k composantes FFT reconstruites :

$$e_t^k = x_t - \sum_{j=1}^{k} x_t^j \;\; k = 1, \ldots, m. \tag{7.2.5}$$

Au delà de la simplicité de la méthode SBD, au moins trois autres avantages sont en la faveur pour l'appliquer lors de la décomposition :

- *i)* aucune perte d'information sur la taille des séries reconstruites, donc nous récupérons toujours des séries temporelles de taille T reproduisant un certain niveau de fluctuation ;
- *ii)* pour une série fortement saisonnière, il suffit de prendre une bande de largeur très fine autour de la fréquence principale pour reconstruire relativement bien l'oscillation périodique ;
- *iii)* la méthode peut être perçue comme une procédure de décomposition en tendance, saisonnalité et bruit. Ainsi, les basses fréquences correspondent à la tendance, la bande autour de la fréquence principale détecte la saisonnalité et les hautes fréquences mettent en évidence un mélange entre un bruit aléatoire et l'hétéroscédaticité des fluctuations.

Nous avons étudié quelques propriétés de la variabilité des séries temporelles de la QAI et nous avons montré quelques aspects de leur complexité. La plupart des séries partagent la propriété suivante : la densité spectrale est concentrée au niveau du pôle à la fréquence zéro. Cette caractéristique rend le point (*iii*) plus difficile à mettre en œuvre, car une initialisation arbitraire est nécessaire pour développer un procédure pour le choix des bandes. Plus précisément, au moins deux paramètres des trois listés ci-après, déterminent la qualité de la décomposition :

- *i)* le nombre de bandes à utiliser ;
- *ii)* le nombre de fréquences à inclure dans chaque bande ;
- *iii)* l'étendue de la fenêtre spectrale ou la largeur de la bande.

Les points *i)* et *ii)* sont similaires, puisque choisir le nombre de fréquences d'une bande spectrale, c'est exactement la même chose que de prendre une largeur particulière de cette bande. Par ailleurs, selon les caractéristiques de la densité spectrale, l'étendue de chaque bande peut être différent d'une fenêtre à une autre. Dans nos applications, nous utilisons souvent une largeur de bande identique.

Par exemple, la Figure 7.2.1 montre la densité spectrale des concentrations de HCHO ainsi que les bandes spectrales définissant le domaine de décomposition de la série temporelle.

7.3 Modèles autorégressifs à changement de régime

Bien que le concept de non-linéarité (ou l'indice de non-linéarité) soit difficile à définir (D'Agostino, 1986; Guégan, 1994, 2003; Ghosh, 1996; Barahona & Poon, 1996), la nécessité de prendre en compte les comportements d'asymétrie, les sauts abrupts et la persistance sur les structures de la dynamique des chroniques de la QAI est imparable. La modélisation des concentrations de polluants par un modèle linéaire par segments où le changement de régime est régi par un seuil, permet de reproduire de tels phénomènes. La voie qui s'est alors révélée plus fructueuse est celle des modèles à changement de régime qui ont l'avantage de permettre aussi une interprétation physique. Cette classe de modèles a été introduite initialement pour mettre en lumière la structure dynamique des cycles d'une série temporelle, suite à des chocs de taille et de signe différents (Tong, 1977; Tong & Lim, 1980; Tong, 1993). Leurs formes paramétriques et leurs propriétés statistiques permettent d'avoir une dynamique différente suivant les régimes ou les états du système dans lesquels il se trouve ; on peut se reporter à (Tong, 1983, 1993) pour une présentation de nombreuses applications.

Suivant les mécanismes de transition stochastique entre les différents états, les modèles à seuil peuvent être classés selon l'observabilité de la variable de transition, en deux catégories : les modèles à changement de régime à seuil de type TAR (*Threshold Auto-Regression*) où la variable de transition est observée, et les modèles à changement de régime de type Markovien, où la variable de transition est cachée.

On se place dans le cadre paramétrique des modèles autorégressifs non-linéaires : trouver un modèle pour les observations revient à déterminer les paramètres d'une fonction dont la forme est connue *a priori* et

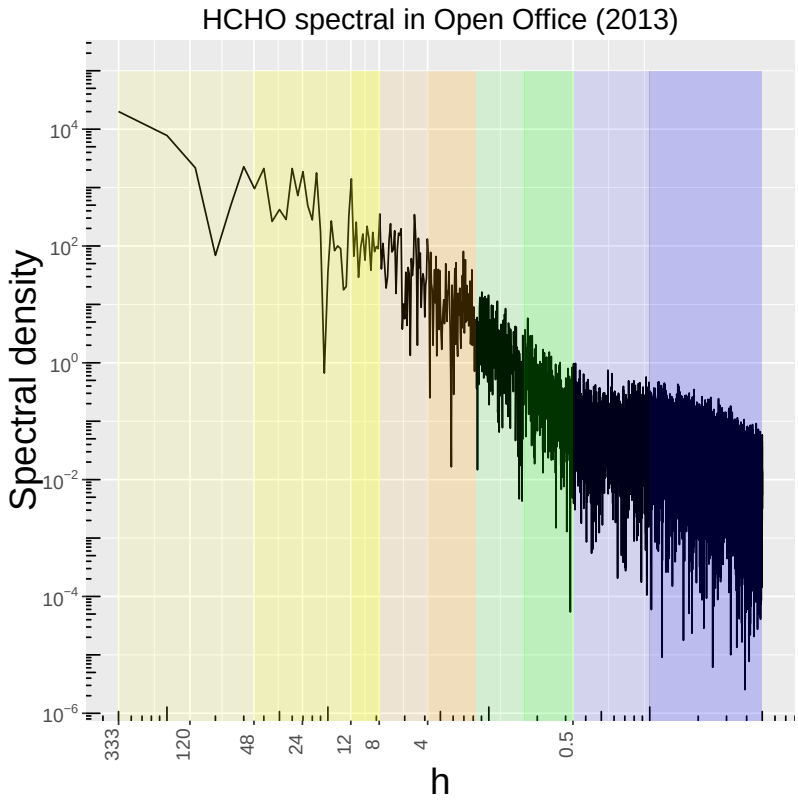


FIGURE 7.2.1 – *Exemple de découpage en bandes spectrales sur un périodogramme pour les fluctuations des concentrations de HCHO en 2013 dans l'espace paysager. La bande en bleu clair ressemble à un spectre presque horizontal, reproduisant ainsi la caractéristique du bruit blanc.*

qui vérifie

$$X_t = f\left(X_{t-1}, X_{t-2}, \ldots\right) + \varepsilon_t \tag{7.3.1}$$

où f est une fonction non-linéaire sur le passé de X_t et ε_t est un bruit blanc.

Une façon d'introduire la non-linéarité est de considérer que la relation 7.3.1 est linéaire par morceaux. Cette classe de modèles a été initiée à l'origine par (Tong, 1978; Tong & Lim, 1980) et étendue par les travaux de (Teräsvirta, 1994; Dijk et al., 2002).

Dans cette section, nous présentons deux principaux modèles non-linéaires en moyenne utilisés à des fins de prévision : modèles à transition brutale et modèles à transition lisse. Il existe néanmoins plusieurs autres modèles qui décrivent les transitions entre les régimes, la classe des modèles les plus utilisés étant les modèles à changement de régime Markovien. Cette thèse n'abordera cette approche qu'au **niveau de l'estimation du modèle**.

La Figure 7.3.1 illustre une vision globale (non-exhaustive) des modèles à changement de régime selon l'observabilité de la variable de transition.

7.3.1 Modèles à seuil à transition brutale

7.3.1.1 Présentation générale des modèles

Les modèles autorégressifs à seuil à transition brutale TAR (Threshold Auto-Regressive) supposent une transition entre les régimes par une fonction de transition indicatrice signalée par une variable de transition. La variable de transition peut être une date connue ; par exemple on sait que la variabilité du CO_2 dans un environnement de type bureau est déterminée par la présence des occupants, celle-ci dépendant du temps calendrier. On peut aussi envisager, dans ce cas, que l'évènement responsable du changement de régime soit signalé par une variable exogène ou par une variable endogène retardée.

Définition 7.3.1. *Processus à changement de régime à transition brutale*
Pour un processus aléatoire $\{X_t\}$, la spécification générale d'un modèle TAR à un nombre quelconque de régimes ℓ est notée par $TAR\left(\ell; p^{(1)}, p^{(2)}, \ldots, p^{(j)}\right)$ et s'écrit :

$$X_t = \sum_{j=1}^{\ell} \left[\phi_0^{(j)} + \sum_{i=1}^{p^{(j)}} \phi_i^{(j)} X_{t-i} + \varepsilon_t^{(j)} \right] \mathbb{I}\left(J_{t-d} \in A_j\right). \tag{7.3.2}$$

Pour chaque régime j, on considère un processus autorégressif AR d'ordre $p^{(j)}$ (entier positif) avec les différents paramètres $\phi_k^{(j)}$, où les perturbations aléatoires $\varepsilon_t^{(j)}$ sont supposées iid de variance σ_j^2 : $\left\{\varepsilon_t^{(j)}\right\} \sim iid\left(0, \sigma_j^2\right)$, pour $j = 1, \ldots, \ell$. Les segments $A_j \in]c_{j-1}, c_j]$ forment une partition de l'espace $\mathbb{R}$ par rapport aux valeurs seuil c_j, tels que $-\infty = c_0 < c_1 < \ldots < c_{\ell-1} < c_\ell = \infty$. Enfin, d est un entier positif appelé paramètre de délai (ou de retard) et $\mathbb{I}(\bullet)$ est une indicatrice prenant la valeur 1 si $J_{t-d} \in A_j$ et 0 sinon.

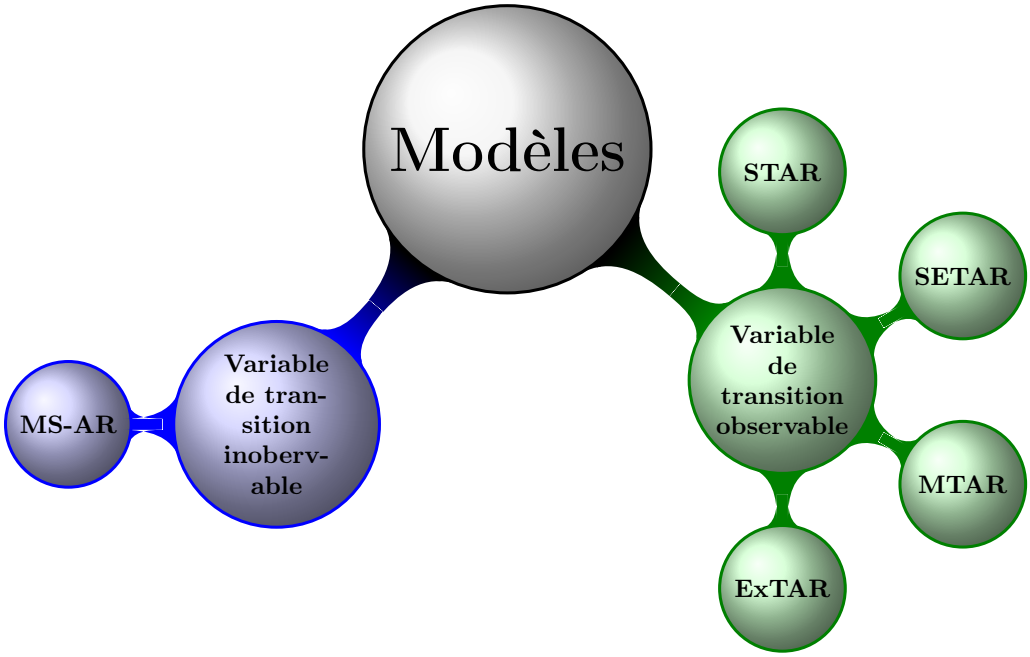


FIGURE 7.3.1 – *Récapitulatif des modèles à changement de régime selon l'observabilité de la variable de transition. Lorsque la transition est induite par une variable inobservable générée par un processus de Markov, alors le passage d'un régime à un autre est signalé par une matrice de probabilités* (**MS-AR**-*Markov switching model). Au contraire, si la variable est observée, on distingue deux classes de modèles, selon la forme de la fonction de transition : (i) brutale, aboutissant à au moins trois types de modèles, qui sont les modèles à variable exogène (***ExTAR***), à variable endogène retardée (***SETAR***) ou à variable de différence (***MTAR***) ou bien (ii) par une fonction de transition lisse donnant les modèles* ***STAR.***

De manière équivalente aux modèles linéaires autorégressifs (AR), on peut obtenir une forme plus générale des modèles TAR en complétant la partie des perturbations $\varepsilon_t^{(j)}$ par une forme moyenne mobile générale (TMA). Le modèle est appelé TARMA (*Threshold AutoRegressive Moving Average*) et se définit comme suit :

$$X_t = \begin{cases} \phi_0^{(1)} + \sum_{i=1}^{p^{(1)}} \phi_i^{(1)} X_{t-i} + \varepsilon_t^{(1)} + \sum_{j=1}^{q^{(1)}} \theta_j^{(1)} \varepsilon_{t-j}^{(1)} & \text{si } J_{t-d} \leq c_1, \\ \phi_0^{(2)} + \sum_{i=1}^{p^{(2)}} \phi_i^{(2)} X_{t-i} + \varepsilon_t^{(2)} + \sum_{j=1}^{q^{(2)}} \theta_j^{(2)} \varepsilon_{t-j}^{(2)} & \text{si } c_1 < J_{t-d} \leq c_2, \\ \vdots & \vdots \\ \phi_0^{(\ell)} + \sum_{i=1}^{p^{(\ell)}} \phi_i^{(\ell)} X_{t-i} + \varepsilon_t^{(\ell)} + \sum_{j=1}^{q^{(\ell)}} \theta_j^{(\ell)} \varepsilon_{t-j}^{(\ell)} & \text{si } c_{\ell-1} < J_{t-d}. \end{cases} \tag{7.3.3}$$

Ce modèle (le système 7.3.3) est une forme de généralisation des modèles ARMA en modèles non-linéaires, mais linéaires par morceaux, car ils permettent des changements dans les coefficients ARMA au fil du temps. Ces changements sont déterminés en comparant les valeurs précédentes à des valeurs de seuils fixes (Douc et al., 2014). Chaque modèle ARMA se réfère à un régime. Dans cette définition, les valeurs de $\left(p^{(j)}, q^{(j)}\right)$ des ordres ARMA peuvent varier dans chaque régime, mais souvent, dans les applications, elles sont égales.

La variable J_{t-d} est appelée variable de transition, représentant ainsi le paramètre qui signale le changement entre les régimes. Dans les modèles 7.3.2 et 7.3.3, la variable de transition a pris la forme générale, le choix de cette dernière peut être guidé par la théorie de la QAI. De manière générale, on peut distinguer au moins trois formes déterminant la variable de transition :

1. la variable endogène retardée $J_{t-d} = X_{t-d}$: on obtient les modèles **SETAR** (*Self-Exciting Threshold AutoRegressive*) (Tong, 1983) ou les modèles **STAR** (*Smooth Transition AutoRegressive*) (Teräsvirta, 1994) ;
2. la variable exogène retardée $J_{t-d} = Y_{t-d}$: on aboutit aux modèles **ExTAR** (*Exogenous Self-Exciting Threshold AutoRegressive) ;*
3. la variation retardée, endogène ou exogène $\Delta J_{t-d} = J_t - J_{t-d}$: on obtient la classe des modèles **MTAR** (*Momentum Threshold AutoRegression*).

Souvent, dans les applications, la variable de transition utilisée est endogène retardée, car elle est facile à interpréter et nécessite des hypothèses faibles sur l'impact réel sur la variable observée à l'instant t. En revanche, pour les variables de transition exogènes retardées, le traitement statistique nécessite non seulement de prouver la significativité de son influence, mais aussi l'impact des ses valeurs retardées sur le processus de variables endogènes X_t. Pour le denier cas (ΔJ_{t-d}), largement défendu par Enders & Granger (1998); Caner & Hansen (2001) dans le cadre des modèles économétriques, les auteurs stipulent que si le processus générant X est globalement stationnaire et possède une racine très proche de l'unité, il est préférable d'utiliser la série en différence.

Du point de vue de la QAI, la variable de transition par différence première d'un décalage d peut être justifiée par l'argument suivant : la variabilité des concentrations d'un polluant peut dépendre de l'amplitude des variations d'un paramètre exogène, plutôt que de son niveau.

Par exemple, la variabilité de la concentration en formaldéhyde dans l'air intérieur est influencée plus par le basculement de la variable fenêtre d'un état à un autre, que par le nombre de fenêtres ouvertes ou fermées. Les processus aléatoires qui peuvent prendre en compte cette assertion sont appelés MTAR (*Momentum Threshold AutoRegression*), voir par exemple [1] l'article de Bohl & Siklos (2004).

1. Comme évoqué dans le chapitre précédant, on veut développer un cadre cohérent de l'analyse des séries temporelles pour la QAI en gardant un œil sur le développement théorique des autres domaines de recherche, car ces derniers ont une portée intuitive de la modélisation très intéressante.

La stationnarité et l'inversibilité des modèles linéaires de type ARMA sont bien connues (voir la section 5.3.1) dans la littérature. Au contraire, pour les modèles non-linéaires à changement de régime, ces propriétés statistiques sont moins spécifiées (Douc et al., 2014).

7.3.1.2 Estimation des modèles SETAR

L'estimation des processus à changement de régime requiert l'estimation de tous ses paramètres :

- les coefficients du modèle $\boldsymbol{\phi} = \left(\phi_i^{(j)}\right)_{i=0,\dots,p^{(j)}}^{j=1,\dots,\ell}$ (dans le cas de SETARMA, on estime aussi le vecteur $\boldsymbol{\theta}$ des coefficients de $\varepsilon_t^{(j)}$) ;
- le nombre de régimes ℓ ;
- le délai d de la variable de transition J_{t-d} signalant le changement de régime ;
- les valeurs seuil de transition $c^{(j)}$.

Il est évident que l'estimation simultanée de tous ces paramètres est une tâche très difficile et très coûteuse en temps de calcul. Compte tenu de la forme des modèles à changement de régime, les variables explicatives dépendent des seuils, donc les méthodes usuelles d'estimation du type moindres carrés ordinaires (MCO) ne sont alors pas applicables dans cette situation. En outre, la fonction de vraisemblance, sur ce modèle n'est pas dérivable en fonction de ces paramètres, alors la méthode du maximum de vraisemblance (MV) n'est pas applicable non plus. La méthode proposée est l'estimation alternée ou séquentielle récursive des MCO.

On se restreint dans cette présentation à la procédure générale d'estimation de Tong & Lim (1980) pour un SETAR $(2; p_1, p_2)$. Pour plus de détails sur le traitement statistique, il est possible de se référer à Tong & Lim (1980); Tong (1993); Hansen (1997).

Afin d'alléger les notations et simplifier la représentation, nous adoptons pour l'ordre d'auto-régression du premier régime le symbole p_1 et pour le deuxième régime, le symbole p_2. De plus, nous considérons uniquement une seule perturbation ε_t se propageant de manière symétrique sur les deux régimes : la transmission de chocs est partagée par les deux régimes.

De ce fait, on estime un seul seuil c du modèle :

$$X_t = \begin{cases} \phi_0^{(1)} + \sum_{i=1}^{p_1} \phi_i^{(1)} X_{t-i} + \varepsilon_t & \text{si } X_{t-d} \leq c \\ \phi_0^{(2)} + \sum_{i=1}^{p_2} \phi_i^{(2)} X_{t-i} + \varepsilon_t & \text{si } X_{t-d} > c \end{cases} \tag{7.3.4}$$

La méthode de Tong et Lim (1980) consiste à appliquer séquentiellement plusieurs étapes :

1. Fixer un ensemble de valeurs pour le paramètre de délai d de la variable de transition X_{t-d} et le seuil c. Le choix de ces ensembles peut être arbitraire, mais on peut, pour des raisons de simplicité, poser c comme étant la moyenne de série.
2. Estimer par MCO chaque régime en choisissant les valeurs de p_1 et p_2 qui minimisent le critère AIC (pour *Akaike Information Criterion*) et calculer le AIC global (somme des deux critères) ;
3. Garder d fixe et faire varier la valeur de c. Pour l'estimation de c, répéter l'étape 1 et 2 pour les différentes valeurs de c et retenir comme valeur optimale du seuil celle qui minimise le critère AIC global ;
4. Estimer d en répétant les étapes précédentes pour chaque d et retenir comme valeur optimale celle qui minimise le critère AIC.

Outre le critère AIC, défini ci-après, on peut en utiliser d'autres, comme le critère BIC (pour *Bayesian Information Criterion*) dans ce processus d'estimation (Colletaz & Hurlin, 2007) :

$$\begin{aligned} AIC\left(p^{(1)},\ldots,p^{(\ell)}\right) &= \sum_{j=1}^{\ell}\left[T_j \log\hat{\sigma}_j^2 + 2\left(p^{(j)}+2\right)\right] \\ BIC\left(p^{(1)},\ldots,p^{(\ell)}\right) &= \sum_{j=1}^{\ell}\left[T_j \log\hat{\sigma}_j^2 + \left(p^{(j)}+1\right)\log T_j\right] \end{aligned}$$

où T_j, $j = 1,\ldots,\ell$ est le nombre d'observations contenu dans le $j^{ème}$ régime, $\hat{\sigma}_j^2$ est la variance des résidus dans le $j^{ème}$ régime et ℓ est le nombre de régimes fixé *a priori*.

Rappelons que ces procédures ont été testés, lors de leurs développements, dans un cadre de séries temporelles à courte durée et au pas de temps large. L'application de cette technique sans aucun contrôle *a priori* sur nos données est extrêmement coûteuse en temps de calcul. Nous avons supposé, dans la plupart de nos applications, que l'ordre des processus autorégressifs de chaque régime est identique et déterminé à partir de l'estimation d'un modèle linéaire. Cette hypothèse ainsi que le nombre de régimes, ne reposent pas forcément sur une justification théorique, mais elle est souvent retenue pour des raisons pratiques.

7.3.2 Modèles à seuils à transition lisse : STAR

A cause du caractère discret de la fonction indicatrice dans les modèles préconisés ci-dessus, les transitions entre les régimes sont brutales et le passage d'un régime à l'autre se fait en une période. Le problème de ces modèles se situe au niveau de la prévision des régimes. Plus précisément, soit un modèle SETAR à deux régimes de variabilité ; si la valeur de X_{t-d} est supérieure à la valeur du seuil c, mais l'observation X_{t+1-d} est inférieure à c, alors la prévision en $t+1$ sera tirée d'une équation différente de celle utilisée en t.

Il est possible dès lors de considérer cette transition par une fonction lisse et on introduit alors les modèles STAR (*Smooth Threshold Auto-Regressive*) (Chan & Tong, 1986; Teräsvirta, 1994). Dans ces modèles, la transition d'un régime vers un autre se fait en introduisant un lissage au cours du temps qui permet d'atténuer les changements abrupts entre les régimes d'un modèle SETAR. Ces modèles ont été introduits par Chan & Tong (1986); Luukkonen et al. (1988) et popularisés par Granger et al. (1993) et Teräsvirta (1994). Pour une revue de littérature voir Dijk et al. (2002); Chow & Zhang (2013); Kock et al. (2011).

Nous traitons ici uniquement le cas le plus simple, celui d'une seule fonction de transition lisse, ayant le rôle de pondérer les régimes AR. On suppose, en outre, que la propagation de la variance des perturbations est identique sur les segments AR : un seul processus ε_t.

Définition 7.3.2. *Processus à changement de régime de transition lisse* (Teräsvirta, 1994)
Le processus $(X_t)_{t\in\mathbb{Z}}$ suit un modèle $STAR(2;p)$ s'il admet l'écriture suivante :

$$\begin{aligned} X_t &= \left(\phi_0^{(1)} + \sum_{i=1}^{p}\phi_i^{(1)}X_{t-i}\right)\times\left[1-G\left(J_t;\mathscr{D}\right)\right] + \\ &\quad \left(\phi_0^{(2)} + \sum_{i=1}^{p}\phi_i^{(2)}X_{t-i}\right)\times G\left(J_t;\mathscr{D}\right) + \varepsilon_t, \end{aligned} \tag{7.3.5}$$

où $G(\bullet)\in[0,1]$ est une fonction continue des paramètres $\boldsymbol{\theta}$ et J_t est la variable de transition.

Le modèle STAR à deux régimes d'ordre p (l'ordre d'autorégression des deux régimes est identique) remplace la fonction indicatrice $\mathbb{I}$ dans 7.3.2 par une fonction $G(\bullet)$ continue et bornée entre 0 et 1.

De la même manière que pour les modèles à changement de régime à transition brutale, la variable de transition peut être endogène retardée, exogène ou une variable de différenciation. Généralement, l'ensemble des paramètres $\mathscr{D} = (\gamma, c)$ est représenté par le seuil de transition c et un paramètre de lissage γ, donc la fonction de transition est de la forme $G(J_t; \gamma, c)$. Deux fonctions de transition sont fréquemment utilisées conduisant à deux variantes du modèle **STAR** : la fonction de transition logistique (LSTAR) ou exponentielle (ESTAR) (*cf.* Tableau 7.3.1).

Table 7.3.1 – Variantes du modèle STAR

Modèle	$G(J_t; \gamma, c)$
Logistic Smooth Auto-Regressive (LSTAR)	$[1 + \exp\{-\gamma(J_t - c)\}]^{-1}$
Exponential Smooth Threshold Auto-Regression (ESTAR)	$1 - \exp\left\{-\gamma(J_t - c)^2\right\}$

La fonction $G(J_t; \gamma, c)$ dans le modèle LSTAR passe instantanément de 0 à 1 dès que la quantité $(J_t - c)$ change de signe et ce, pour un paramètre de lissage suffisamment grand, approximant ainsi la fonction indicatrice $\mathbb{I}_{J_t}$ dans le modèle SETAR. En revanche, si $\gamma \to 0$ alors la fonction de transition avoisine 0.5 ($G(J_t; \gamma, c) \rightsquigarrow 1/2$) et le modèle LSTAR se réduit à un modèle AR simple.

La configuration quadratique de la fonction $G(J_t; \gamma, c)$ dans le modèle ESTAR permet de neutraliser l'effet signe de la quantité $(J_t - c)$; par conséquent, plus le paramètre de lissage est important ($\gamma \to +\infty$), plus la probabilité de passage entre les régimes est faible, *i. e.* le modèle ESTAR reste dans le même régime plus longtemps. Le modèle ESTAR quitte un régime pour rejoindre un autre dans deux cas : soit $\gamma \to 0$, soit $J_t \to c$.

7.3.3 Modèles à variable de transition cachée : Markov Switching AutoRegression (MS-AR)

7.3.3.1 Présentation générale

Les transitions entre les régimes peuvent être gouvernées par un processus aléatoire inobservé. Le modèle le plus connu pour refléter cette caractéristique de transition probabiliste est celui défini dans le cadre Markovien. Introduits dans la littérature économétrique par Goldfeld & Quandt (1973); Hamilton (1989, 1990); Krolzig (2013), les modèles à variables cachées ont été ensuite largement développés dans le traitement automatique [2] de la parole (Douc et al., 2001, 2004).

On trouve très peu d'études sur l'application de cette classe de modèles pour les données environnementales ; on peut se référer aux travaux de Ailliot (2004) pour des applications sur la variable vent ou à

2. Je note ici que les modèles utilisés dans la littérature du traitement du signal sont principalement les HMM (Hidden Markov Model ou Chaînes de Markov Cachées), voir les travaux de Baum & Petrie (1966); Baum et al. (1967, 1970); Cappé et al. (2009). Les modèles autorégressifs à changement de régime Markovien englobent les HMM car ils permettent une relation (même linéaire) entre les observations décalées dans le temps. Or dans les modèles HMM, la loi conditionnelle de l'observation à l'instant t sachant les observations passées et les états cachés, dépend uniquement de l'état caché à l'instant t.

la thèse de Page (2007) pour la simulation de l'activité des occupants dans un environnement intérieur. Cette dernière propose un modèle de chaîne de Markov pour décrire l'activité des occupants dans un environnement intérieur.

Dans le cadre des processus à changement de régime Markovien, le mécanisme de transition repose sur une variable d'état cachée, qu'on note S_t, qui est supposée suivre une chaîne de Markov. Pour simplifier la représentation, nous supposons qu'il existe seulement deux régimes : $S_t = \{1, 2\}$.

Le processus générant l'état inobservable est défini par une matrice des probabilités de transition $\mathscr{P}$ d'éléments suivants

$$\begin{cases} \mathscr{P}_{ij} = \mathbb{P}\left(S_{t+1} = j \mid S_t = i, S_{t-1} = k, \cdots\right) = \mathbb{P}\left(S_{t+1} = j \mid S_t = i\right), \\ \mathscr{P}_{ij} + \mathscr{P}_{ii} = 1, \\ \forall t = 1, 2, \ldots, T,\ i, j, k \in \{1, 2\}. \end{cases} . \tag{7.3.6}$$

Pour chaque période de temps, il existe donc une certaine probabilité d'appartenir à un régime donné. Selon le processus 7.3.6, la probabilité que le système soit dans l'état j à $t+1$ dépend uniquement de l'état dans lequel il se trouvait à l'instant t. Autrement dit, dans l'évolution au cours du temps, l'état du processus à l'instant $t+1$ ne dépend que de celui à l'instant précédent, mais pas de ses états antérieurs (Foata & Fuchs, 2002). Ce type de processus est appelé processus à temps discret sans mémoire, comme les modèles à mémoire courte de type ARMA. Chaque élément $\mathscr{P}_{ij}$ traduit la probabilité de passer d'un régime $S_t = i$ à un régime $S_{t+1} = j$, ou de rester dans le même régime.

Le modèle à deux régimes d'ordre p_1 et p_2, $MS\left(2\right) - AR\left(p_1, p_2\right)$ peut être donné par la définition 7.3.3.

Définition 7.3.3. Le processus $\left(X_t\right)_{t\in\mathbb{Z}}$ suit un modèle $MS\left(2\right) - AR\left(p_1, p_2\right)$ s'il admet l'écriture suivante :

$$X_t = \begin{cases} \phi_0^{(1)} + \sum_{i=1}^{p_1} \phi_i^{(1)} X_{t-i} + \varepsilon_t^{(1)} & \text{si } S_t = 1 \\ \phi_0^{(2)} + \sum_{i=1}^{p_2} \phi_i^{(2)} X_{t-i} + \varepsilon_t^{(2)} & \text{si } S_t = 2 \end{cases} \tag{7.3.7}$$

Nous avons vu l'algorithme EM (*Expectation Maximization*) lors du traitement des données manquantes. Cet algorithme permet d'estimer les paramètres d'un modèle statistique lorsque ce dernier dépend des variables latentes non-observables. Cet algorithme, dû à l'article fondateur de Dempster et al. (1977), est appliqué pour l'estimation des processus $MS - AR$. Pour un exposé plus détaillé, on se réfère à l'article de Hamilton (1990) ainsi qu'à son ouvrage (Hamilton (1994), chapitre 22) et à la thèse de Ailliot (2004) ; les propriétés statistiques des estimateurs sont discutées dans (Douc et al., 2004).

Sans rentrer dans les détails techniques de l'inférence statistique, la contribution à la log −vraisemblance conditionnelle de X_t (*i.e.* $\log f\left(X_t \mid \Omega_{t-1}, S_t = j; \Theta\right)$) se calcule à partir de la densité de X_t conditionnellement à l'ensemble d'informations Ω_{t-1} et au vecteur de paramètres Θ uniquement :

$$f\left(X_t \mid \Omega_{t-1}; \Theta\right) = \sum_{j=1}^{2} f\left(X_t \mid S_t = j, \Omega_T; \Theta\right) . \mathbb{P}\left(S_t = j \mid \Omega_{t-1}; \Theta\right). \tag{7.3.8}$$

Le problème d'estimation du modèle revient, au premier abord, à estimer à tout instant la probabilité d'occurrence de chaque état, étant données leurs réalisations passées et le vecteur des paramètres Θ,

c'est-à-dire $\mathbb{P}(S_t = j \mid \Omega_{t-1}; \Theta)$. L'estimation des probabilités associées aux régimes inobservés conditionnellement à l'ensemble de l'information disponible peut fournir une indication sur l'historique de ces régimes, alors trois types de probabilités conditionnelles, selon le contenu informationnel de Ω, peuvent être considérées :

1. la probabilité associée au filtrage (*filtred regime probabilities*) $\mathbb{P}(S_t = j \mid \Omega_t)$: elle donne la probabilité que l'état $S_t = j$ se réalise étant donné l'ensemble d'informations jusqu'au moment t, Ω_t ;
2. les probabilités prévues (*predicted regime probabilities*) $\mathbb{P}(S_t = j \mid \Omega_{t-1})$: elle donne la probabilité que l'état $S_t = j$ se réalise étant donné l'ensemble d'informations jusqu'au moment $t-1$, Ω_{t-1} ;
3. les probabilités lissées (*smoothing regime probabilities*) : $\mathbb{P}(S_t = j \mid \Omega_T)$: elle donne la probabilité que l'état $S_t = j$ se réalise étant donné l'ensemble d'informations jusqu'au moment t ainsi que celle de l'ensemble de l'échantillon jusqu'à l'instant T, Ω_T.

L'estimation par maximum de vraisemblance (MV) des probabilités de transition est donnée par :

$$\widehat{\mathscr{P}_{ij}} = \frac{\sum_{t=2}^{T} \mathbb{P}\left(S_t = j, S_{t-1} = i \mid \Omega_T; \hat{\Theta}\right)}{\sum_{t=2}^{T} \mathbb{P}\left(S_{t-1} = i \mid \Omega_T; \hat{\Theta}\right)} \qquad i, j \in \{1, 2\}, \tag{7.3.9}$$

où $\hat{\Theta}$ est l'estimation par MV des paramètres du modèle et Ω_T désigne l'ensemble d'informations de l'échantillon jusqu'à l'instant T.

7.3.3.2 Résultats de l'estimation d'un modèle à changement de régime Markovien sur la série des concentrations de HCHO

Les résultats préliminaires exposés dans le deuxième chapitre nous ont permis de mettre en évidence, à l'aide de la variabilité polaire, quelques relations liant la concentration de HCHO, les paramètres climatiques et l'ouverture des fenêtres (voir la sous-section 2.5.3.2).

Pour compléter ces considérations, nous avons adopté une procédure séquentielle d'un modèle à changement de régime Markovien pour élucider quantitativement les relations entre ces différents paramètres : un $MS(2) - AR(2)$ a été appliqué dans plusieurs configurations que nous allons expliquer ci-après.

Qautre hypothèses ont été posées :

— *(i)* la chaîne de Markov est à deux états $S_t = \{1, 2\}$, donc on a deux régimes de variabilité ;
— *(ii)* l'ordre d'autorégression de la variable cible (HCHO) est identique pour les deux régimes, et il est fixé à 2 ;
— *(iii)* la matrice de transition est fixe ;
— *(iv)* les variables explicatives agissent de manière instantanée sur la variable à expliquer.

Ces hypothèses ont été formulées par rapport à la seule variable de la série des concentrations de formaldéhyde en 2013 mesurées dans l'espace paysager. On peut certainement lever certaines hypothèses pour d'autres configurations, notamment lorsque le pas de temps utilisé est supérieur às 20 min.

Pour la première hypothèse, nous avons d'abord estimé un modèle à trois régimes, mais il s'est avéré que le régime intermédiaire n'est pas très important : les probabilités de séjour sur le régime 2 sont très faibles, donc elles sont très élevées pour quitter le régime 2 pour aller vers les régimes 3 et 1. Nous avons donc, fixé le nombre de régimes à deux pour les différentes simulations.

En ce qui concerne l'hypothèse *(ii)*, il est évident qu'elle est réductrice, étant donné le pas d'échantillonnage d'une minute. En revanche, en augmentant les p, l'estimation nécessite un traitement numérique très lourd du fait de la présence de hautes fréquences.

La troisième hypothèse tient du fait qu'il est très difficile d'estimer les probabilités non-stationnaires de passage d'un régime à l'autre. D'ailleurs, très peu de travaux ont été menés en essayant de lever la contrainte de la constance des probabilités de transition.

La dernière hypothèse peut être levée dans la modélisation vectorielle des séries temporelles, mais dans le cas de cette étude, aucun retard dans l'analyse de l'impact des covariables n'a été pris en compte.

Avant de présenter les résultats de l'estimation et de prévision à l'aide des modèles à changement de régime (SETAR, TAR, MS-AR), on propose une remarque générale sur les régimes de variabilité de formaldéhyde, qui peut être extrapolée pour la variabilité de certains polluants :

La vitesse de chute d'un régime de fluctuation (niveau élevé) vers un régime de fluctuation à un niveau bas est plus grande que dans le cas contraire. Le processus de disparition du HCHO (effet puits) est plus rapide que les processus de son émission. En termes statistiques, les chocs aléatoires sont asymétriques.

Pour illustrer ce propos, la Figure 7.3.2 montre un exemple d'occurrence des régimes et leur vitesse de variabilité par rapport à la variable d'ouverture des fenêtres. Malheureusement, le mode de renseignement de cette dernière n'était pas sans faute durant la campagne de 2013, car seulement 3 capteurs d'ouverture des fenêtres sur 5 capteurs étaient fonctionnels. Par conséquent, certaines phases d'élimination de HCHO n'ont pas été mises en évidence par l'ouverture des fenêtres, les deux derniers jours et demis de cet exemple illustrent ce propos. En effet, bien que l'état des fenêtres ait été signalé comme fermé, des changements abrupts ont été enregistrés, notamment le 2013-06-12 à 9h10 et le 2013-06-13 à 8h45. Probablement, un occupant a ouvert au moins une fenêtre, mais qui n'a pas été renseignée par les capteurs d'ouverture.

En tenant compte de ces observations, plusieurs questions surviennent dès lors qu'on s'intéresse aux modèles de prévision. Admettant que la variable sur l'ouverture de fenêtres a été intégrée dans un modèle de prévision à chargement de régime, alors :

1. comment l'erreur faite sur la validité d'une variable exogène (ici l'ouverture des fenêtres) se transmet sur la prévision de celle-ci ?
2. l'échec en prévision est-il dû à une mauvaise spécification du modèle ou à une altération du modèle par la variable exogène ? Ou aux deux simultanément ? Dans ce cas, comment peut-on distinguer une cause de l'autre ?

Sans formulation d'un modèle physique de l'environnement intérieur, ces questions s'avèrent encore plus difficiles, bien au delà de la validité des données.

La démarche adoptée dans la suite pour esquisser quelques éléments concernant l'importance des covariables (variables explicatives) sur la variabilité du HCHO est la suivante :

— Estimer le modèle $MS(2) - AR(2)$ sur la série des concentrations de HCHO **(Modèle 1)** ;

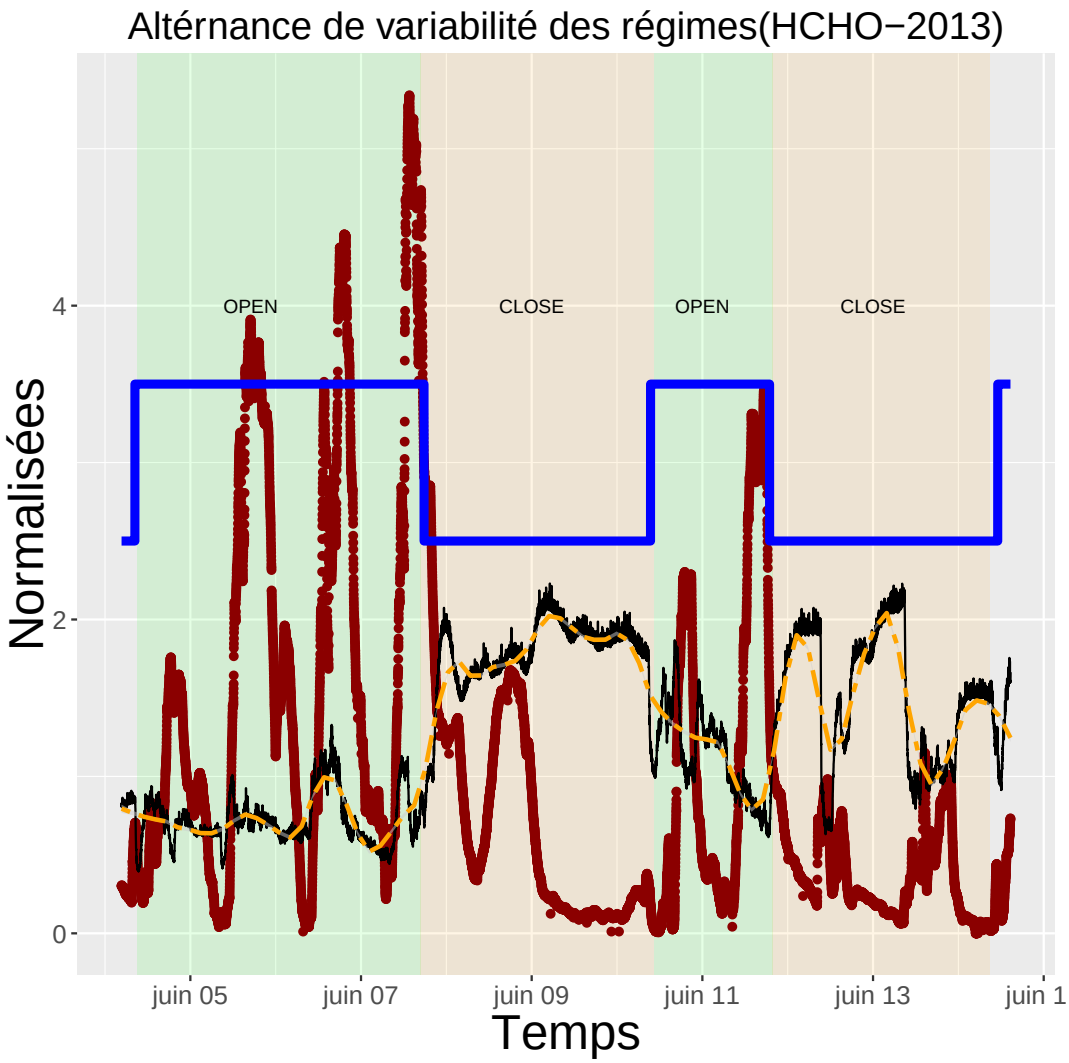


FIGURE 7.3.2 – *Altérnances des régimes et l'importance de la variabilité due à un effet "puits". Les concentrations normalisées sont obtenues en divisant la concentration réelle par la concentration moyenne de toute la série. Sur les fluctuations du HCHO (courbe trait plein en noir), on a ajusté une courbe par la méthode de régression LOESS (courbe tiretées en orange). Les fluctuations de l'ozone intérieur sont représentées avec des points (en rouge foncé). Les zones ombrées reflètent les périodes dues à l'importance des processus d'émissions (réctangles oranges) et les processus dus à l'effet puits (réctangles verts) associés à l'occurrence de la variable fenêtre : au moins une fenêtre est ouverte (OPEN) ou toutes les fenêtres sont fermées (CLOSE), la courbe bleue. Le taux d'informations disponible sur la variable d'ouverture est de 60%.*

— Estimer le modèle $MS(2)-AR(2)$ en ajoutant une variable explicative : la température intérieure (**Modèle 2**) ;
— Estimer le modèle $MS(2)-AR(2)$ avec deux variables explicatives : la température intérieure et la concentration d'ozone à l'intérieur (**Modèle 3**) ;
— Estimer le modèle $MS(2)-AR(2)$ avec trois variables explicatives : la température intérieure, les concentrations d'ozone intérieures et l'état des fenêtres (**Modèle 4**) ;
— Estimer le modèle $MS(2)-AR(2)$ en utilisant uniquement l'état des fenêtres comme covariable (**Modèle 5**) ;
— Estimer le modèle $MS(2)-AR(2)$ avec deux variables explicatives : la concentration d'ozone à l'intérieur et l'état des fenêtres (**Modèle 6**) ;
— Estimer le modèle $MS(2)-AR(2)$ sur une seule composante obtenue à partir de la série de HCHO par SBD en se limitant uniquement à la bande définie de la fréquence de coupure (**Modèle 7**).

Formellement, cela revient à estimer un modèle de la forme générale suivante :

$$\text{HCHO}_t = \begin{cases} \sum_{j=1}^{n} \alpha_j^{(1)} \text{Paramètre}\,(j) + \sum_{i=1}^{2} \phi_i^{(1)} \text{HCHO}_{t-i} + \varepsilon_t^{(1)} & \text{si } S_t = 1 \\ \sum_{j=1}^{n} \alpha_j^{(2)} \text{Paramètre}\,(j) + \sum_{i=1}^{2} \phi_i^{(2)} \text{HCHO}_{t-i} + \varepsilon_t^{(2)} & \text{si } S_t = 2 \end{cases}, \tag{7.3.10}$$

où $\alpha_j^{(1)}$, $\alpha_j^{(2)}$ sont les coefficients du paramètre $j = \{\text{température, ozone, fenêtre}\}$ correspondant au régime 1 et au régime 2, respectivement ; n représente le nombre de covariables utilisées. Par exemple, dans le premier modèle, aucune covariable n'est utilisée, dans le $4^{ème}$ modèle, on a utilisé les trois paramètres suivant : la température intérieure, la concentration intérieure d'ozone et l'état des fenêtres.

L'impact des covariables sur la variabilité du HCHO peut être analysé par deux informations complémentaires :

1. La significativité [3] des variables par des tests classiques de type $t-student$;
2. L'analyse de la matrice des probabilités de transition.

Dans le Tableau 7.3.2, on présente l'estimation des modèles présentés ci-dessus. Une remarque générale peut être dégagée de ce tableau : tous les coefficients des ordres autorégressifs de tous les modèles sont significativement différents de zéro, mais pour les coefficients des variables explicatives, la significativité dépend du régime dans lequel elles sont définies. On peut traduire cette constatation par le fait que la covariable intervient uniquement pour augmenter ou diminuer le niveau d'émission des concentrations de HCHO.

Par exemple, dans le modèle **2**, nous avons traité l'intervention d'une seule variable explicative (la température intérieure) ; le coefficient au niveau du premier régime est négatif (significatif au seuil de 5%), tandis qu'il est positif dans le second régime. La significativité du coefficient dans le premier régime de la température disparaît lorsqu'on rajoute l'ozone (*cf.* Modèle **3**, Tableau 7.3.2).

Il s'avère que l'ozone pourrait expliquer la variabilité de HCHO. En effet, dans les Modèles **3** et **4**, les coefficients de la variable de l'ozone sont négatifs dans les deux régimes de variabilité du HCHO. Autrement dit, lorsque l'ozone augmente, les concentrations de formaldéhyde diminuent, et inversement, lorsque la concentration de l'ozone diminue, celle du HCHO augmente. Cela peut être expliqué par l'assertion suivante : l'ozone est principalement d'origine extérieure, sa concentration à l'intérieur est plus faible ;

3. Je reste néanmoins un peu dubitatif sur ces tests lorsque les séries temporelles sont de grande dimension, ou lorsque le pas de temps est très fin, mais surtout lorsque la décision faite *via* ce test est à contresens du phénomène. Récemment, l'ASA (American Statistical Association) a publié un rapport sur les pratiques abusives de la $p-value$ dans la validation des résultats (Wasserstein & Lazar, 2016), la conclusion de l'article étant : "*No single index should substitute for scientific reasoning*".

lorsqu'on ouvre les fenêtres, l'ozone extérieur rentre à l'intérieur et en même temps la concentration du formaldéhyde chute par dilution de l'air *via* les ouvertures. Donc le taux de renouvellement de l'air par l'ouverture des fenêtres joue un double rôle sur la QAI : il accentue les concentrations de certains polluants (d'origine extérieure majoritairement) et diminue d'autres (d'origine intérieure principalement).

Si on continue ce raisonnement, c'est la variable d'ouverture qui est à l'origine de la variation brusque, l'ozone n'est là que pour la confirmer en tant qu'indicateur. D'ailleurs, sur la Figure 7.3.2, on peut voir cette relation. En fait, le traitement statistique d'une variable discrète (ouverture) avec une variable continue (HCHO) est plus difficile à mettre en évidence leurs relations que lorsqu'elles sont de même nature (discrète ou continue).

Pour s'en convaincre, le modèle **5** où seule la variable état des fenêtres est traitée comme variable explicative, son impact sur les chutes de concentrations en HCHO dans l'air est clair dans le deuxième régime (faible variabilité). Même dans le cas où on a analysé les covariables ozone et état des fenêtres (Modèle **6**), le coefficient de l'ozone reste négatif sur les deux régimes, tandis que celui de la fenêtre est de signe négatif pour le régime 2 et positif pour le premier régime.

Ce qui est intéressant dans les modèles à changement de régime Makovien c'est qu'ils fournissent, en plus de l'estimation des paramètres, une nature probabiliste de la transition entre les régimes. Cette dernière est riche en informations, car elle permet d'indiquer le niveau de stochasticité des variables latentes et éventuellement leur origine.

En effet, le processus latent S_t (à deux états) affectant la variable d'intérêt tend à être plus aléatoire lorsque la probabilité de passer d'un régime à un autre, et donc de rester dans le même régime, est de 1/2.

En observant les probabilités obtenues dans chaque modèle, il apparaît que les transitions se font à une probabilité allant de 20 à 30%. Sans variables explicatives, le modèle estimé tend à rester dans le régime 2 avec une probabilité de 80%, la transition du régime 2 vers le régime 1 est moins fréquente ($\simeq 20\%$) que dans le cas contraire. La probabilité de rester dans le régime 1 est de 70%, et celle de passage vers le régime 2 est de 30%.

On remarque que lorsqu'on évalue les modèles avec des covariables (la température et l'ozone intérieur dans les modèles **2** et **3**), la probabilité de transition du régime 1 vers le second régime diminue, mais revient à son niveau (30%) quand on intègre la variable fenêtre.

Lorsque seul l'état des fenêtres est appliqué comme variable explicative (modèle **5**), les coefficients de chaque régime pour cette dernière sont relativement significatifs et les coefficients autorégressifs de la variable HCHO demeurent inchangés que lorsqu'on n'utilise aucune variable explicative (modèle 1). Le signe négatif du coefficient de la variable "état des fenêtres" dans le modèle **5** au niveau du régime 2 traduit l'importance de cette dernière sur les concentrations de faibles variabilités.

Quant au modèle **7**, nous avons utilisé une série temporelle des concentrations de HCHO issue de la décomposition SBD ; une seule bande a été utilisée : la fréquence de coupure qui est de $f_c = (10\,h)^{-1}$. Il s'avère que le processus latent est presque absorbant dans la mesure où les probabilités de transition sont quasi nulles.

TABLE 7.3.2 – Estimation de $MS\,(2) - AR(2)$ pour la série temporelle des concentrations de formaldéhyde. 30 000 valeurs de HCHO (toutes les minutes) ont été utilisées de la série issue de la campagne de 2013 dans l'espace paysager.

	Modèle 1	Modèle 2	Modèle 3	Modèle 4	Modèle 5	Modèle 6	Modèle 7
$\widehat{\phi}_1^{(1)}$	0.74***	1.037***	1.038***	0.739***	0.74***	0.74***	1.998***
	(0.015)	(0.0103)	(0.0106)	(0.0122)	(0.0114)	(0.011)	0
$\widehat{\phi}_2^{(1)}$	0.25***	-0.038***	-0.038***	0.254***	0.25***	0.25***	-0.998***
	(0.011)	(0.0103)	(0.0106)	(0.0123)	(0.0115)	(0.011)	0
$\widehat{\phi}_1^{(2)}$	1.035***	0.7454***	0.74***	1.037***	1.0347***	1.034***	2***
	(0.0113)	(0.0121)	(0.012)	(0.0107)	(0.01)	(0.0095)	0
$\widehat{\phi}_2^{(2)}$	-0.036**	0.2516***	0.254***	-0.038***	-0.035***	-0.034***	-1***
	(0.0113)	(0.0121)	(0.0122)	(0.0107)	(0.01)	(0.0095)	0
$\widehat{\alpha}_{(T_i)}^{(1)}$	-	(-0.0002)*	0	0.0038***	-	-	-
	-	(0.0001)	(0.0001)	(0.0004)	-	-	-
$\widehat{\alpha}_{(T_i)}^{(2)}$	-	0.0019***	0.0035***	0	-	-	-
	-	(0.0003)	(0.0004)	(0.0001)	-	-	-
$\widehat{\alpha}_{(\mathrm{O}_3)}^{(1)}$	-	-	-0.0003**	-0.0017***	-	-0.0002	-
	-	-	(0.0001)	(0.0004)	-	(0.0003)	-
$\widehat{\alpha}_{(\mathrm{O}_3)}^{(2)}$	-	-	-0.0019***	-0.0003**	-	-0.0003**	-
	-	-	(0.0004)	(0.0001)	-	(0.0001)	-
$\widehat{\alpha}_{(Win)}^{(1)}$	-	-	-	-0.0143*	0.009*	0.0112*	-
	-	-	-	(0.007)	(0.0045)	(0.0053)	-
$\widehat{\alpha}_{(Win)}^{(2)}$	-	-	-	-0.003	-0.0045***	-0.0031*	-
	-	-	-	(0.0019)	(0.0013)	(0.0015)	-
$\widehat{\mathscr{P}}_{1,1}$	**0.70**	**0.79**	**0.79**	**0.703**	**0.70**	**0.69**	**0.995**
$\widehat{\mathscr{P}}_{1,2}$	**0.29**	**0.2**	**0.2**	**0.29**	**0.29**	**0.3**	**0.005**
$\widehat{\mathscr{P}}_{2,1}$	**0.20**	**0.29**	**0.29**	**0.203**	**0.2**	**0.20**	**0.004**
$\widehat{\mathscr{P}}_{2,2}$	**0.79**	**0.704**	**0.7**	**0.796**	**0.79**	**0.79**	**0.996**

*Notes. Le code de la significativité des coefficients est donné par la $p-value$. Tant que celle-ci est très petite (inférieure 0.05), nous rejetons l'hypothèse nulle de nullité des coefficients, Pr(>|t|) : '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '. Les modèles sont estimés sans constante, elle est souvent non significative. Les valeurs entre parenthèses correspondent à l'erreur type de chaque modèle. Le modèle 7 est estimé sur une série issue de la SBD pour les concentrations du HCHO ; une seule bande a été appliquée avec la fréquence de coupure $f_c = (10\,h)^{-1}$.*

7.3.4 Prévision des modèles non-linéaires paramétriques

7.3.4.1 Problématique et solution naïve

Nous avons étudié dans le chapitre 5 le concept de prévision dans le cadre des modèles linéaires. La relation récursive entre les différents horizons de prévision s'obtient grâce aux propriétés algébriques de l'espérance mathématique : l'opérateur $\mathbb{E}(\bullet)$ est une application linéaire. Autrement dit, l'espérance mathématique transforme la combinaison linéaire des variables aléatoires dans le processus en espérance mathématique de ces variables. Nous avons, par ailleurs fait le point sur les conséquences d'une transformation non-linéaire (de type logarithmique) sur la formule de prévision. En vertu de l'inégalité de Jensen, la fonction (convexe) de l'espérance d'une variable aléatoire X est inférieure à l'espérance de la fonction :

$$f(\mathbb{E}(X)) \leq \mathbb{E}[f(X)]. \tag{7.3.11}$$

Cette transformation indique un certain nombre de difficultés techniques liées à la construction d'une prévision dans le cadre des modèles non-linéaires. En fait, il s'agit particulièrement de certains modèles à seuil que nous avons présentés au début de ce chapitre.

Face à cette problématique, il existe trois manières pour construire des prévisions : prévisions ponctuelles, intervalles de prévision ou densité de prévision. La première forme consiste à résumer la distribution conditionnelle des observations futures par une seule valeur, la deuxième par un intervalle et la dernière par une densité de probabilité des prévisions. Dans le cas des modèles linéaires, l'intervalle de prévision construit autour de la valeur prédite est symétrique et continue. En revanche, dans les modèles non-linéaires, les intervalles de prévision peuvent non-seulement être **asymétrique**s, mais aussi **discontinus** (Hyndman, 1995, 1996).

Soit un modèle non-linéaire autorégressif d'ordre p :

$$X_t = f(X_{t-1}, X_{t-2}, \ldots, X_{t-p}; \Theta) + \varepsilon_t, \tag{7.3.12}$$

où ε_t est $i.i.d\left(0, \sigma_\varepsilon^2\right)$ et Θ est l'espace des paramètres du modèle paramétrique sous-jacent. Pour un modèle à seuil c à deux régimes de même ordre d'auto-régression p avec un délai d de la variable endogène décalée X_{t-d} *i.e.* SETAR$(2; p, p; d)$, l'espace des paramètres est $\Theta = \left(p, \phi_0^{(1)}, \phi_1^{(1)}, \ldots, \phi_p^{(1)}, \phi_0^{(2)}, \phi_1^{(2)}, \ldots, \phi_p^{(2)}, d, c\right)$.

Compte tenu de l'ensemble d'informations disponibles au moment t, noté Ω_t, la prévision optimale $\widehat{X}_{t+h}$ (skelton) s'obtient par l'espérance conditionnelle

$$\widehat{X}_{t+h} = \mathbb{E}(f(X_{t+h-1}, X_{t+h-2}, ..., X_{t+h-p}, \Theta) + \varepsilon_{t+h}|\Omega_t) \tag{7.3.13}$$

et l'erreur de prévision associée à $\widehat{X}_{t+h}$,

$$\widehat{\varepsilon}_{t+h} = X_{t+h} - \widehat{X}_{t+h}. \tag{7.3.14}$$

Quel que soit le modèle nonlinéaire envisagé, la prévision à l'horizon d'une période ($h = 1$) ne pose pas de problème particulier, car $\mathbb{E}(\varepsilon_{t+1}|\Omega_t) = 0$. En revanche, pour un ordre de prévision $h > 1$, il est très

difficile d'établir (voire impossible) la forme analytique permettant d'obtenir les prévisions ponctuelles. Pour $h = 2$, la prévision optimale peut être écrite comme suite

$$\begin{aligned} \widehat{X}_{t+2} &= \mathbb{E}\left(f\left(X_{t+1},\ldots,X_{t+2-p};\Theta\right)+\varepsilon_{t+2}\mid\Omega_t\right) \\ &= \mathbb{E}\left(f\left(f\left(X_t,\ldots,X_{t+1-p};\Theta\right)+\varepsilon_{t+1},\,X_t,..,X_{t+2-p};\Theta\right)+\varepsilon_{t+2}\mid\Omega_t\right). \end{aligned} \tag{7.3.15}$$

Étudions la difficulté d'estimer cette quantité avec un modèle non-linéaire d'ordre 1 de type $X_t = f\left(X_{t-1};\Theta\right)+\varepsilon_t$, plus simple. Si on souhaite prévoir le niveau X_{t+2}, la prévision optimale s'écrit alors sous la forme :

$$\begin{aligned} \widehat{X}_{t+2} &= \mathbb{E}\left[X_{t+2}\mid\Omega_t\right] \\ &= \mathbb{E}\left[f\left(X_{t+1};\Theta\right)+\varepsilon_{t+2}\mid\Omega_t\right] \\ &= \mathbb{E}\left[f\left(f\left(X_t;\Theta\right)+\varepsilon_{t+1};\Theta\right)\mid\Omega_t\right] \\ &= \mathbb{E}\left[f\left(\widehat{X}_{t+1}+\varepsilon_{t+1};\Theta\right)\mid\Omega_t\right] \end{aligned}. \tag{7.3.16}$$

Tout le problème vient du fait que, en règle générale, l'opérateur linéaire d'espérance ne peut pas être interchangé avec la fonction non linéaire $f(\bullet)$, *i.e.* : $\mathbb{E}\left[f(\bullet)\right] \neq f\left(\mathbb{E}\left[\bullet\right]\right)$(voir (Franses & Van Dijk, 2000)).

Dès lors, si l'on souhaite prévoir la valeur X_{t+2} conditionnellement à Ω_t, une solution "naïve" serait de négliger le terme d'erreur de prévision ε_{t+1} dans la formule 7.3.16. Le terme "négliger" employé ici consiste à dire qu'il est possible d'interchanger les opérateurs $f(\bullet)$ et $\mathbb{E}\left[\bullet\right]$. Par conséquent, la prévision optimale naïve de X_{t+2}, notée $\widehat{X}^N_{t+2}$ est obtenue par :

$$\widehat{X}^N_{t+2} = \mathbb{E}\left[f\left(\widehat{X}_{t+1}+\underbrace{\varepsilon_{t+1}}_{=0};\Theta\right)\mid\Omega_t\right] = f\left(\widehat{X}_{t+1};\Theta\right), \tag{7.3.17}$$

avec $\widehat{X}_{t+1} = f\left(X_t;\Theta\right)$. Pour un horizon h quelconque, la valeur de la prévision naïve est définie par une relation récursive de la forme :

$$\widehat{X}^N_{t+h} = f\left(\widehat{X}^N_{t+h-1};\Theta\right), \qquad h \geq 1. \tag{7.3.18}$$

La méthode naïve aboutit à des prévisions biaisées.

7.3.4.2 Solution par simulation numérique

La question de prévision dans le cadre général des modèles non-linéaires consiste à évaluer des quantités très complexes, qui nécessitent généralement un développement numérique afin d'approximer l'espérance conditionnelle. Reprenons l'exemple précédant d'un processus autorégressif non-linéaire d'ordre 1 et la forme de prévision dans la formule 7.3.16. La prévision repose dans l'évaluation de la quantité :

$$\widehat{X}^c_{t+2} = \int_{-\infty}^{\infty} f\left(\widehat{X}_{t+1}+\varepsilon_{t+1};\Theta\right) g\left(\varepsilon\right) d\varepsilon, \tag{7.3.19}$$

où $g\left(\varepsilon\right)$ désigne la fonction de densité du choc ε_{t+1} (Colletaz & Hurlin, 2007). Dans Brown & Mariano (1989), cette expression se réfère en "*closed-form forecast*", d'où l'exposent c.

Une autre manière d'exprimer cette intégrale consiste à dériver la prévision optimale en fonction de la distribution conditionnelle du "vrai" processus X_{t+1} par rapport à l'ensemble d'informations disponibles au moment t.

De 7.3.19, on peut aboutir à une autre expression de type

$$\begin{array}{rcl} \widehat{X}^{c}_{t+2} = & \int_{-\infty}^{\infty} f\left(X_{t+1};\Theta\right) h\left(X_{t+1} \mid \Omega_t\right) dX_{t+1} \\ = & \int_{-\infty}^{\infty} \mathbb{E}\left[X_{t+2} \mid X_{t+1}\right] h\left(X_{t+1} \mid \Omega_t\right) dX_{t+1}, \end{array} \tag{7.3.20}$$

où $h\left(X_{t+1} \mid \Omega_t\right)$ est la distribution conditionnelle de X_{t+1} par rapport à Ω_t. Selon Franses & Van Dijk (2000), cette distribution correspond à la distribution du résidu ε_{t+1} avec une moyenne égale à $f\left(X_t;\Theta\right)$. La généralisation de l'expression 7.3.20 à un horizon h est donnée par :

$$\mathbb{E}\left[X_{t+h} \mid \Omega_t\right] = \int_{-\infty}^{\infty} f\left(X_{t+h} \mid X_{t+h-1}\right) h\left(X_{t+h-1} \mid \Omega_t\right) dX_{t+h-1}. \tag{7.3.21}$$

Naturellement, ce développement ne résout en rien le problème de l'évaluation de cette intégrale. L'évaluation se fait donc par des méthodes numériques fondées sur des simulations.

Deux méthodes connues dans la littérature peuvent répondre à ce besoin : Monte-Carlo (MC) et Bootstrap (Boot). Les simulations Monte-Carlo nécessitent une hypothèse sur la distribution des résidus ε_{t+1} ; il s'agit de faire k tirages indépendants dans cette distribution. La prévision ponctuelle à l'étape 2 par MC, notée $\widehat{X}^{MC}_{t+2}$, s'obtient par la quantité moyenne :

$$\begin{array}{rcl} \widehat{X}^{MC}_{t+2} & = & \frac{1}{k}\sum_{i=1}^{k} \widehat{X}^{(i)}_{t+2} \\ & = & \frac{1}{k}\sum_{i=1}^{k} f\left(\widehat{X}_{t+1};\Theta\right) + \varepsilon^{(i)}_{t+1} \end{array}. \tag{7.3.22}$$

La méthode Bootstrap consiste à ré-échantillonner les résidus historiques dans l'étape de l'estimation $(\hat{\varepsilon}_t)$, en tirant des séquences $\left\{\hat{\varepsilon}^{(1)}_{t+1},\ldots,\hat{\varepsilon}^{(i)}_{t+1},\ldots,\hat{\varepsilon}^{(k)}_{t+1}\right\}$ dans la séquence $\{\hat{\varepsilon}_1,\hat{\varepsilon}_2,\ldots,\hat{\varepsilon}_t\}$.

La prévision ponctuelle à l'étape 2 par Bootstrap, notée $\widehat{X}^{Boot}_{t+2}$, s'obtient par la quantité moyenne :

$$\begin{array}{rcl} \widehat{X}^{Boot}_{t+2} & = & \frac{1}{k}\sum_{i=1}^{k} \widehat{X}^{(i)}_{t+2} \\ & = & \frac{1}{k}\sum_{i=1}^{k} f\left(\widehat{X}_{t+1};\Theta\right) + \hat{\varepsilon}^{(i)}_{t+1} \end{array}. \tag{7.3.23}$$

7.4 Modèles issus de la théorie des systèmes dynamiques

Plaçons-nous maintenant dans la perspective des modèles non-linéaires motivés par la théorie des systèmes dynamiques. Les fluctuations de la QAI ont des dynamiques complexes : c'est-à-dire qu'ils sont rarement strictement périodiques, mais fluctuent de façon irrégulière au cours du temps. Ce type de comportement est à la frontière (probablement floue) d'un système déterministe et des processus stochastiques. De ce fait, il semble donc invraisemblable que toutes les structures de variabilités qu'on a pu observer tout au long de cette thèse soient uniquement le produit de l'aléa ; l'avènement des systèmes dynamiques non linéaires du chaos pour les séries temporelles peuvent apporter des explications notables à ces dynamiques.

Le point de vue de ces approches en séries temporelles consiste à dire que l'évolution dynamique d'un phénomène est définie dans un certain espace de phase qu'on doit "reconstruire" à partir des mesures. Quand l'irrégularité dans la série temporelle est forte, ces systèmes non-linéaires peuvent exhiber des structures chaotiques déterministes (Kantz & Schreiber, 2004). Dans ce cas, l'influence des perturbations aléatoires est supposée minime ou bien supprimée par une méthode quelconque. Plusieurs travaux théoriques ont discuté sur la reconstitution en présence de bruit Casdagli et al. (1991); Guégan (2003) ; la séparation du bruit du signal est la méthode la plus courante. L'élimination du bruit est très conseillée (Kantz & Schreiber (2004), chapitre 4, page 58, Casdagli (1992)) pour fournir de bonnes prévisions avec la méthode dite de chaos. On trouve par exemple dans (Guégan, 2008; Cao, 2002) l'utilisation des ondelettes comme outils de pré-traitement et de déconvolution du bruit ; pour nous il s'agit de la décomposition en bandes spectrales (SBD).

Les techniques de détection et d'analyse dans ce type de systèmes sont regroupées sous le nom de "traitement de signaux non-linéaires". Il s'agit principalement de reconstruire, à partir d'une série temporelle, une trajectoire "topologiquement" équivalente aux trajectoires de l'espace des phases. Le problème qui se pose est celui de la fiabilité de la reconstitution dans l'espace effectif. La méthode de délais basée sur le théorème de plongement (Takens, 1981; Mañé, 1981; Sauer et al., 1991) donne les conditions pour y parvenir.

Nous avons étudié la méthode SSA qui consiste dans sa première étape à plonger la série temporelle par la méthode des délais, qui repose et se justifie à partir du théorème de (Takens, 1981; Mañé, 1981). Comme évoqué par Manneville (2004) (page 143), la représentation de la reconstitution peut être améliorée par rapport aux données au sens des moindres carrés par la décomposition en valeurs singulières, telle que proposée par Broomhead & King (1986b); Vautard et al. (1992). Cette assertion suggère une très forte relation entre la méthode SSA et les outils de détection de chaos.

7.4.1 Éléments de la théorie des systèmes dynamiques

Nous avons discuté la notion de prédictibilité dans un cadre plutôt stochastique, car la mesure fait intervenir l'entropie différentielle d'un processus de bruit aléatoire pur. Par ailleurs, Kantz & Schreiber (2004) (page 5) suggéraient un point de vue similaire, mais très nuancé : "*The predictibility of the signal can be taken as a signature of the deterministic nature of the system*" . Par conséquent, selon la part de stochasticité ou déterministe de la série, on peut définir une distance (éventuellement une divergence) qui caractérise la prédictibilité.

Cette thèse part du principe qu'il est possible de fournir des prévisions en se basant aussi sur les structures déterministes (éventuellement non-linéaires et complexes) ; la théorie des systèmes dynamiques donne un cadre exploratoire riche pour répondre à ce point.

7.4.1.1 Systèmes dynamiques

Dans cette section, une brève introduction des bases théoriques de l'analyse des systèmes dynamiques est présentée, ainsi que les attracteurs des séries temporelles associés à ces systèmes. Le concept de l'espace d'état est introduit dans le cadre de l'étude de l'environnement intérieur, et quelques exemples typiques d'attracteurs sont présentés par la suite.

7.4.1.2 L'espace d'état et systèmes dynamiques

Aspects théoriques L'étude d'un système complexe fait entrer en ligne de compte l'ensemble des positions de la trajectoire, des vitesses de variation que le système est *a priori* susceptible d'adopter. Cet ensemble est appelé espace d'état. Celui-ci possède souvent une structure de variété (voir l'annexe E.3 à la page 372 pour une introduction) ; la structure et la dimension de cette variété sont importantes pour construire une théorie solide.

On suppose pouvoir décrire l'espace d'état, ici l'environnement intérieur, par un ensemble de d variables d'état ; tel que chaque état de ce système corresponde à un point $S \in \mathscr{M}$, avec $\mathscr{M}$ une variété de dimension d. Il est généralement souhaitable que cette variété soit suffisamment lisse et compacte (différentiable). La variété $\mathscr{M}$ est appelée le *vrai espace d'état* et d la dimension effective de l'espace des phases [4].

La dimension du système peut évidemment être grande, mais le nombre "*effectif*" de degrés de liberté pour le système dynamique peut être très faible. Dans certains cas, la variété $\mathscr{M}$ peut être l'espace euclidien $\mathbb{R}^d$, mais la présence de certaines variables temporelles, comme les variables circulaires (direction du vent) ou l'absence d'information sur la nature de la variété sur laquelle le système évolue, $\mathscr{M}$ pourrait avoir une topologie toroïdale, donc non-Euclidienne. La topologie non-euclidienne peut être plongée seulement dans les espaces euclidiens de dimensions suffisamment grandes (voir la présentation dans l'annexe E.3.2).

Plaçons-nous maintenant dans la perspective des séries temporelles d'un système dynamique, où la dimension d et la topologie de $\mathscr{M}$ sont inconnues ; cette distinction est essentielle, puisqu'on ne dispose pas d'une alternative pour travailler sur les espaces euclidiens simples.

Tout système dont les états varient dans le temps est *un système dynamique* ; pour un tel système, un état est une fonction $S(t)$ du temps t. Si de plus, cette fonction est continue, c'est ce qu'on suppose plus généralement, l'évolution de l'état sur l'espace des phases forme une trajectoire paramétrée par t.

Dans un *système dynamique déterministe* [5], l'état initial $S(0)$ est le seul qui détermine les états futurs $S(t), t > 0$. Dans le cas d'un système dynamique stochastique, une application unique entre les états (futur, présent et passé) peut ne pas exister (Honerkamp, 1993). Il est donc possible qu'il n'y ait aucune "corrélation" entre les états ; d'ailleurs un bruit aléatoire pur comme un bruit blanc, l'est par définition.

Formellement et sans entrer dans les détails techniques (nous nous référons à Mawhin & Rouche (1973) pour plus de considérations mathématiques), l'unique fonction f^t reliant l'état initial $S(0)$ à l'instant t, $(S(t))$ est donnée par :

$$S(t) = f^t(S(0)). \tag{7.4.1}$$

4. Clairement ceci relève d'une description simplifiée, il existe des situations, notamment en hydrodynamique, où un nombre fini de variables n'est pas suffisant pour une description complète de l'espace d'état. Cette thèse n'abordera pas ces considérations.

5. Pour le cas continu, le système est noté $S(t)$; voir le point fait sur les notations dans le Chapitre 3.

De manière alternative, on peut représenter le système dynamique déterministe par un ensemble d'équations différentielles [6] :

$$\overset{\bullet}{S}(t) = \frac{d}{dt}S(t) = \boldsymbol{F}(S(t)), \quad t \in \mathbb{R}, \tag{7.4.2}$$

où $\boldsymbol{F}$ est le champ de vecteurs (lisse) dans $\mathbb{R}^d$. Sous certaines conditions (problème de Cauchy d'existence et d'unicité), l'ensemble des équations différentielles $\boldsymbol{F}$ a une solution unique pour toute valeur initiale $S(0)$. Si $\boldsymbol{F}$ ne dépend pas du temps, on note par $f^t(S)$ l'unique solution de 7.4.2 qui passe par S en $t = 0$ et l'application (à temps t fixé) $S \to f^t(S)$ s'appelle le *flot* du système dynamique 7.4.2.

Attracteurs En fonction de la structure de $\boldsymbol{F}$ (ou de manière équivalente de f^t), il existe plusieurs possibilités du comportement de $S(t)$ pour $t \to \infty$. Nous ne considérons pas la possibilité de divergence du système *i.e* : $S(t) \to \infty$, (puisqu'elle est inutile), mais seulement les cas bornés.

Si le système est dissipatif [7] (*i.e.* l'énergie n'est pas conservée), tous les volumes dans l'espace des phases vont se contracter sous l'action du flot ; pour un temps suffisamment long, le système va évoluer sur un ensemble très réduit d'états $\mathcal{A}$ appelé attracteur.

Pratiquement, un attracteur délimite une zone dans laquelle il est "très probable" que le système évoluera à long terme ; il existe quatre différents types d'attracteurs (*cf.* la Figure 7.4.1) :

1. Points fixes : tout état initial peut converger (au moins pour un certain bassin d'attraction) vers un état final de repos en point fixe. Il caractérise simplement un système atteignant un état stationnaire (*cf.* Figure (A) 7.4.1). Une série temporelle reproduisant le point fixe serait par exemple $X(t) = X(0)$.
2. Cycles limites : plutôt de converger en un point, le système évolue sur un ensemble d'états qui sont visités périodiquement. Il caractérise un système atteignant un état répétitif : la trajectoire de phase se referme sur elle-même. L'évolution temporelle est alors cyclique, le système présentant des oscillations permanentes (*cf.* Figure (B) 7.4.1).
3. Tores limites : plusieurs fréquences interviennent dans la trajectoire périodique dans le système à travers les états du cycle limite. Le système présente au moins deux périodes simultanées dont le rapport est irrationnel. La trajectoire de phase ne se referme pas sur elle-même, mais s'enroule sur une variété de dimension 2 (*cf.* Figure (C) 7.4.1), par exemple un tore. Une écriture paramétrique d'un tore $\boldsymbol{T}^2$ dans $\mathbb{R}^3$ est la suivante :

$$\begin{aligned} x &= a_1 \sin(2\pi f_1 t + \alpha_1) + a_2 \sin(2\pi f_2 t + \alpha_2) \\ y &= b_1 \sin(2\pi f_1 t + \beta_1) + b_2 \sin(2\pi f_2 t + \beta_2) \\ z &= c_1 \sin(2\pi f_1 t + \gamma_1) + c_2 \sin(2\pi f_2 t + \gamma_2) \end{aligned}$$

où $a_i, b_i, c_i, \alpha_i, \beta_i, \gamma_i$ pour $i = \{1, 2\}$ sont des constantes réelles et f_1 et f_2 sont les deux fréquences de base du mouvement (Vialar, 2005).

6. Dans son introduction, Lorenz (1969) dit " *The physical laws which govern the behavior of the earth's atmospher may be formulated as a system of differential equations. The problem of weather forecasting may be identified with the problem of discovering [...] a particular solution of these equations, whose initial conditions correspond to the present state of the atmosphere*".

7. Globalement un tel système se caractérise par le fait qu'un élément de volume de l'espace de phase voit en moyenne son volume diminuer lorsque t augmente (il s'agit d'une conséquence du théorème de Liouville) : cela se traduit par l'existence d'attracteurs.

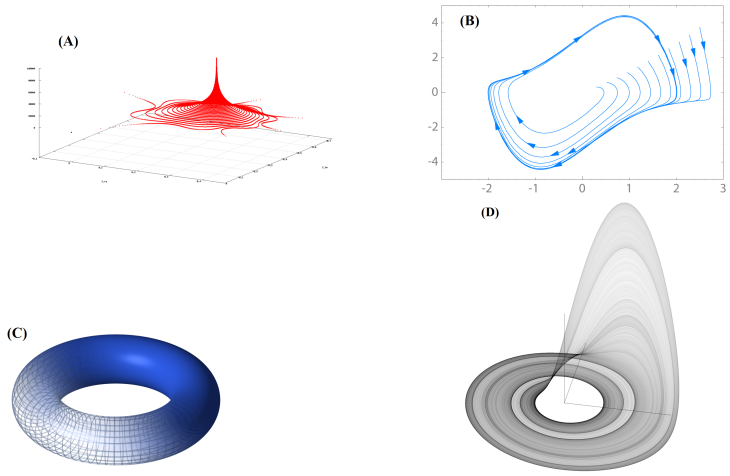


FIGURE 7.4.1 – *Différents types d'attracteurs pour les systèmes différentielles déterministes.*

4. Attracteurs étranges [8] : lorsque le système dissipatif produit des rythmes qui ne sont pas strictement périodiques. L'attracteur étrange désigne une figure dans l'espace représentant le comportement d'un système dynamique. La contraction des trajectoires autour de l'attracteur est liée au caractère chaotique du système réel (*cf.* système Rossler Figure (D) 7.4.1).

La qualité de l'air intérieur comme système dynamique

Ce paragraphe découle directement de la discussion dans la section 1.6 à la fin du premier chapitre de la première partie. Elle peut donc être perçue comme complémentaire.

De très nombreux travaux ont été élaborés pour modéliser le comportement dynamique des grandeurs physiques du bâtiment d'un point de vue thermo-aéraulique, mais très peu pour les questions de la QAI. En effet, on distingue généralement quatre types [9] de modèles pour le bâtiment :

- (*i*) modèles monozone : sont les plus simples car ils considèrent le bâtiment comme un seul volume communiquant avec l'extérieur, donc incapables d'estimer des grandeurs locales sur des échelles fines ;
- (*ii*) modèles multizones (nodaux) : les modèles sont construits à partir des équations de conservation de la masse, de l'énergie et de la quantité de mouvement sur des zones qui correspondent à plusieurs pièces ;
- (*iii*) modèles zonaux : avec une discrétisation plus fine que les modèles multizones, les modèles sont appliqués à des volumes très fins dans l'espace global du bâtiment ;
- (*iv*) codes de champs ou CFD (pour Computational Fluid Dynamics) : ces modèles cherchent à résoudre les équations physiques dans les volumes du domaine. Généralement, le domaine est découpé en un très grand nombre de mailles.

Tous ces modèles tentent de prendre en compte l'ensemble des différents processus physico-chimiques : le transport, la remise en suspension et les puits dans les environnements intérieurs ainsi que les réactions chimiques. Les mécanismes de formation et de transformation sont intrinsèquement liés aux différentes sources, aux paramètres climatiques et à l'occupation.

8. MANDELBROT (1983) juge que la dénomination de "étrange" est un choix inadapté qui devrait être plutôt "attracteur fractal".

9. voir les travaux de César (2014); Damian (2003) pour une présentation plus détaillée sur ces modèles.

La représentation mathématique de toutes ces transformation des modèles physico-chimiques aboutit généralement à une analyse des systèmes d'équations différentielles ordinaires (EDO) ou partielles (EDP). Ces représentations linéaires ou non-linéaires décrivent le système dynamique de la QAI.

Notre approche s'affranchit de ces mécanismes et de ces raffinements en modélisation ainsi que des connaissances en termes sources et paramètres influents. Ces paramètres peuvent intervenir dans nos modèles, mais en tant que variables exogènes. Les modèles présentés ici n'étudient pas le comportement du système dynamique dans son espace effectif, mais dans un pseudo-espace reconstruit à partir des observations par la méthode des délais.

Pour ce faire, on se base principalement sur les travaux de Kantz & Schreiber (2004) et les développements d'outils logiciels associés Hegger et al. (1999), et on fait appel aux différentes monographies traitant les séries temporelles non-linéaires par l'approche des systèmes dynamiques (chaotiques) comme :

— (Tong, 1993) pour une approche plus théorique ;
— Vialar (2005); Abraham-Frois (1998) et (Guégan, 2003) pour les applications en mathématique financière (économétrie). Ce dernier fournit aussi beaucoup de réflexions théoriques sur la relation entre la prévision et les systèmes chaotiques ;
— Galka (2000) pour les applications aux signaux physiologiques ; l'approche intuitive développée dans ce livre rend la présentation simple (même si au fond les concepts sous-jacents sont très complexes) ;
— Manneville (2004) pour applications aux différents phénomènes hydrodynamiques (stabilités et turbulences) ;
— comme le titre du livre l'indique : "*Applied nonlinear time series analysis*", (Small, 2005) présente les applications de l'analyse des séries temporelles par l'approche des systèmes dynamiques dans différents domaines des sciences.

7.4.1.3 Séries temporelles et systèmes dynamiques

Le problème de reconstitution

L'observation directe de tous les états de l'environnement intérieur comme états d'un système dynamique $S(t)$ n'est pas directement accessible. En effet, d'un point de vue pratique, il est très difficile de mesurer simultanément toutes les d variables du système ; l'accès aux informations liées aux variations du $S(t)$ demeurent partielles. L'attention va se focaliser sur les propriétés évolutives du processus mesuré, tant pour sa prévision que dans sa relation avec son passé.

Soit un système dynamique à temps discret [10] $S_{t+1} = f(S_t)$ et une série temporelle X_t. La reconstitution de la dynamique consiste en la détermination de la relation empirique entre les états S_t et la seule connaissance des X_t, $t = 0, 1, ...$

Nous utilisons, par conséquent, les informations fournies par un ensemble de séries temporelles $X_t = h(S_t)$ par le biais de l'application d'une *fonction de mesure* $h : \mathbb{R}^d \to \mathbb{R}^{d'}$ sur l'état effectif S, avec $d' < d$. Typiquement, la fonction h est inconnue car la nature de l'espace $\mathbb{R}^d$ et les états S sont inconnus.

Cette fonction fournit un moyen d'inférer sur l'évolution du système à partir d'un ensemble de variables très réduit, voir une seule variable. Une première mesure $X_0 = h(S_0)$ n'est évidement pas suffisante pour déterminer l'état S_0 car il faut plus de coordonnées pour le caractériser. L'existence de la fonction f amplifie l'accès aux informations contenues dans le système S_t par des mesures successives (Manneville, 2004). Ainsi, une deuxième mesure ajoute une information : $X_1 = h(S_1) = h(f(S_0))$, une troisième encore un peu plus car $X_2 = h(S_2) = h(f(f(S_0)))$ *etc.*

10. L'écriture $S(t)$ ou S_t désigne la quantité $S(t) \equiv S_t$

Les relations suivantes lient les états effectifs et les observations :

$$\begin{array}{rcl} X_0 = & h(S_0) & \\ X_1 = & h(S_1) & = h(f(S_0)) \\ X_2 = & h(S_2) & = h(f(f(S_0))) \\ \vdots & \vdots & \vdots \\ X_{d-1} = & h(S_{d-1}) & = h(f \circ \cdots \circ f(S_0)) \end{array} . \tag{7.4.3}$$

Une suite suffisamment longue de d mesures successives $\{X_0, X_1, \cdots, X_{d-1}\}$ devrait permettre de spécifier l'état initial S_0 ; la spécification successive par récurrence des états se fait en gardant la même longueur de la mémoire associée aux observations X_k. Autrement dit :

$$\begin{array}{ccc} \{X_0, X_1, \cdots, X_{d-1}\} & \xrightarrow{specifie} & S_0 \\ \{X_1, X_2, \cdots, X_d\} & \xrightarrow{specifie} & S_1 \\ \vdots & & \vdots \end{array} .$$

Enfin, la série de vecteurs :

$$\boldsymbol{V}_k = [X_k; \ldots; X_{k+d-1}] \tag{7.4.4}$$

donnerait accès à la trajectoire du système dans un espace de reconstruction $\mathbb{R}^d$, équivalente à celle dans son propre espace des phases.

Théorème de plongement de Takens

L'espace d'état effectif du système est relativement complexe et mal connu ; même en dimension faible. Il apparaît donc justifié de consacrer toute une partie l'analyse des séries temporelles dans une perspective des systèmes dynamiques, par la reconstitution de l'espace d'état.

Les techniques pour obtenir des informations sur le système dynamique à partir des séries temporelles mesurées dérivent toutes de la méthode des délais formalisée par Takens (1981). Le problème se situe au niveau de la qualité de la reconstitution dans l'espace effectif en correspondance avec la partie de l'espace des phases. Pour que cette reconstitution soit "acceptable", elle doit être injective : $\boldsymbol{V}_k = \boldsymbol{V}_{k'} \Longrightarrow S_k = S_{k'}$. La dimension définie dans 7.4.3, doit être plus grande ou égale à la dimension de l'espace dans lequel il faut, mathématiquement parlant, *plonger* le système pour le représenter de façon fiable (voir l'annexe mathématique E). Cette dimension est appelée dimension de plongement, notée d_E.

Le théorème de Takens (1981) stipule que les vecteurs

$$\boldsymbol{V}_k = [X_k; \ldots; X_{k+d_E-1}] \tag{7.4.5}$$

définis grâce à une fonction différentiable sur l'espace des phases réalise une reconstruction fiable de la dynamique pour une dimension de plongement d_E de données suffisante, soit $d_E = 2d' + 1$ où d' est la dimension topologique de la variété qui supporte la dynamique effective. En fait, la reconstruction proposée par Takens est plus générale que 7.4.5 puisqu'elle est construite avec le vecteur de délais suivant :

$$\boldsymbol{X}_t = \left(X_t, X_{t-\tau}, X_{t-2\tau}, \ldots, X_{t-(d_E-2)\tau}, X_{t-(d_E-1)\tau}\right)^\top. \tag{7.4.6}$$

Le théorème de Takens (1981) garantit que si la série est suffisamment longue et que d_E est suffisamment large vis-à-vis la dimension de la variété sur laquelle l'attracteur se trouve, alors l'image de dimension d_E de l'attracteur fournit une représentation topologique correcte de la dynamique. Ainsi, les orbites périodiques sur d'un attracteur correspondent aux orbites périodiques dans l'espace des phases reconstruit, et les orbites chaotiques du système d'origine paraîtront chaotiques dans cet espace (Vialar, 2005).

7.4.1.4 Choix des paramètres de plongement (τ, d_E)

Le délai (time delay) τ

L'application directe du théorème de plongement sur des données réelles est sujette à une mauvaise spécification. En effet, le théorème repose sur deux hypothèses difficiles à réaliser dans un cas réel, notamment l'absence de bruit dans les observations et la contrainte d'une structure très lisse de la variété.

En principe, il n'y a pas de justification théorique sur le choix du délai ; le théorème de plongement (Takens, 1981; Mañé, 1981; Sauer et al., 1991) est "*muet*" sur ce point (Abarbanel, 2012). La reconstitution peut se faire alors avec n'importe quel délai d'une série temporelle, excepté près pour certains délais dans les signaux périodiques ayant des fréquences bien prononcées (Sauer et al., 1991).

En revanche, d'un point de vue pratique, le choix du délai ainsi que la dimension de plongement sont de grande importance, car l'information contenue dans la représentation du vecteur de délai est très sensible au choix de ces deux paramètres. Ainsi, le choix d'un délai trop petit entraînerait une corrélation forte entre les composantes du vecteur délai, *i.e.* si le délai τ est relativement petit par rapport à la taille (T) de la série (X_t) et au pas d'échantillonnage τ_s, les coordonnées X_t et $X_{t-\tau}$ auront une dépendance forte. A l'inverse, le choix d'un délai trop grand pourrait gommer l'effet du bassin d'attraction et déstructurer la forme géométrique de la reconstitution. Et ce, en raison du caractère intrinsèque de l'instabilité des systèmes chaotiques. En présence d'un bruit aléatoire, la recherche et le choix des paramètres se compliquent avec sa variance. En effet, le choix d'un délai trop petit entraînerait une confusion entre la structure d'autocorrélation et la dynamique engendrée par les chocs aléatoires du bruit.

En absence d'un théorème sur un critère ou sur des conditions d'optimalité permettant de choisir le délai pour une meilleure approximation du système dynamique, plusieurs méthodes d'estimation ont été proposées dans la littérature. Abarbanel (2012) suggère un ensemble "*prescriptions*" afin de dépeindre au mieux[11] les propriétés des paramètres (τ, d_E) qui caractérisent les systèmes chaotiques.

Deux méthodes souvent citées dans la littérature pour estimer le délai τ : la fonction d'autocorrélation et l'information mutuelle entre les différents retards.

La dimension de plongement m

Une méthode efficace pour choisir la dimension de plongement consiste, après la fixation de τ, à considérer la reconstruction de la dynamique avec une dimension d'essai d, $[X_k; \ldots; X_{k+d-1}]$ et la reconstitution de la dimension $d+1$ obtenue en ajoutant une composante X_{k+d}, puis de déterminer le nombre de faux voisins. Le nombre de faux voisins correspond au nombre de pairs de points qui sont

11. Plus précisément, à la page 25 dans Abarbanel (2012), on peut lire *« ..To put it another way, we have no hope of finding the very best or optimal time delay without giving some additional rationale for that optimality. Any such rationale will, of course, determine some T, but such optimality is tautological ».*

voisins en dimension d mais qui ne le sont plus en dimension $d+1$. On augmente d jusqu'à ce que la fraction de faux voisins devient négligeable. La dimension de plongement *optimale* serait celle pour laquelle le nombre de faux voisins a diminué pour la première fois de façon significative (Manneville, 2004).

7.4.2 Reconstitution des séries temporelles de la QAI : l'impact de la filtration par bandes spectrales

En complément de l'analyse des structures de variabilité de séries temporelles des polluants dans l'air intérieur (*cf.* Chapitre 3), la dynamique du chaos offre des outils d'analyse très intéressants pour mettre en évidence certains aspects de prédictibilité (dynamique déterministe). La reconstitution du portrait de phase par la théorie du plongement donne une première impression de la nature de ces structures. En revanche, en présence du bruit, cette tâche s'avère très difficile à réaliser, car la présence d'une moindre fluctuation aléatoire dénature les structures régulières de variabilité censées être faciles à prédire.

La littérature sur les moyens de filtrage des séries temporelles pour la dynamique du chaos est très abondante ; la plupart évoquent les outils comme la transformée de Hilbert (Manneville, 2004; Gilmore & Lefranc, 2002) ou les ondelettes (Guégan, 2008). Nous proposons d'explorer les structures de variabilité par la décomposition de type SBD.

La Figure 7.4.2 présente le portrait de phase des fluctuations de CO_2 des mesures issues de la campagne de 2011 dans le bureau individuel. Nous présentons trois types de portait : *(i)* sur les données brutes, *(ii)* sur la série reconstruite par une bande spectrale définie par une fréquence de coupure au niveau $(5.83\,\mathrm{h})^{-1}$ et *(iii)* sur la composante saisonnière extraite par la méthode STL. Clairement, le portrait de phase sur les données brutes donne l'impression d'observer un point stationnaire, l'attraction vers point fixe ; le lissage par SBD met davantage en évidence cet aspect. Quant au portait de phase constitué par la composante déterministe saisonnière, le plongement de la série forme un attracteur de type torique. Cet ensemble est contracté sur lui-même pour des niveaux de faible variabilité et relaxé pour de fortes variabilité en formant des rondelles tout au long de l'attracteur. La série temporelle qui forme ce type d'attracteur a un profil diurne en forme de "**M**" ; il suffit, par conséquent de prendre la dimension de plongement égale à la période principale (en supposant que $\tau = 1$) pour visualiser un ensemble attractif fermé. La composante saisonnière dégage un profil diurne en forme de "**M**" généré par la succession des états suivants :

1. inoccupation, 19 h-7 h ;
2. occupation, 7 h-12.30 h ;
3. pause de midi (inoccupation), 12.30 h-13.30 h ;
4. occupation, 13.3 h-19 h.

Rappelons que la composante saisonnière détermine la part la plus importante dans la variabilité de quelques polluants, en particulier pour les concentrations du CO_2 et les particules de taille moyenne. Par conséquent, la manifestation de cet ensemble attractif au sein de l'attracteur général est très importante.

Sur la Figure 7.4.3 et la Figure 7.4.4, on présente la reconstitution des séries temporelles de HCHO issues des campagnes 2013 et 2015, respectivement. Sur chacune, trois reconstitutions ont été effectuées : *(i)* sur la série brute, *(ii)* sur la série reconstruite par une bande spectrale définie par une fréquence de coupure et *(iii)* sur la composante saisonnière extraite par la méthode STL.

Bien que la reconstitution de la partie saisonnière de la série HCHO de 2013 soit formée par un tore très régulier, l'impact de cette composante sur les fluctuations globales est minime, alors que la contribution

de celle de 2015 est élevée. Cette dernière forme elle aussi une tore contracté est allongé selon la partie de l'espace des phases.

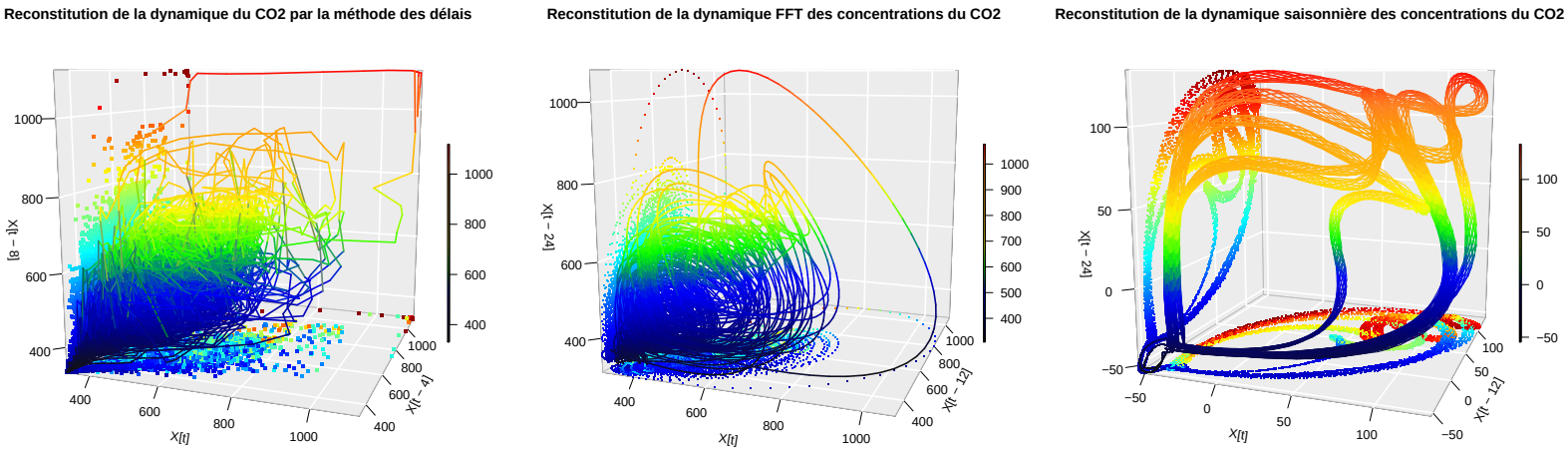


FIGURE 7.4.2 – *Reconstitution de l'espace des phases par la méthode des délais des différentes structures de la dynamique des concentrations du CO_2 toutes les 10 minutes dans le bureau individuel (campagne 2011, le pas de temps est de 10 minutes). Les paramètres de plongement sont : $m = 144$ et $\tau = 1$ pour tous les types de reconstitutions. Pour les données brutes (graphique à gauche), la reconstitution est formée par trois vecteurs : $X_t, X_{t-4},$ et X_{t-8}. Sur les données filtrées par une bande spectrale (de FFT) définie par une fréquence de coupure de $f_c = (5.83\,h)^{-1}$, la reconstitution est constituée par les vecteurs X_t, X_{t-12}, et X_{t-24} (graphique au centre). La composante saisonnière a été extraite avec la méthode STL, lissée par une régression Loess et la reconstitution est composée par X_t, X_{t-12}, et X_{t-24}.*

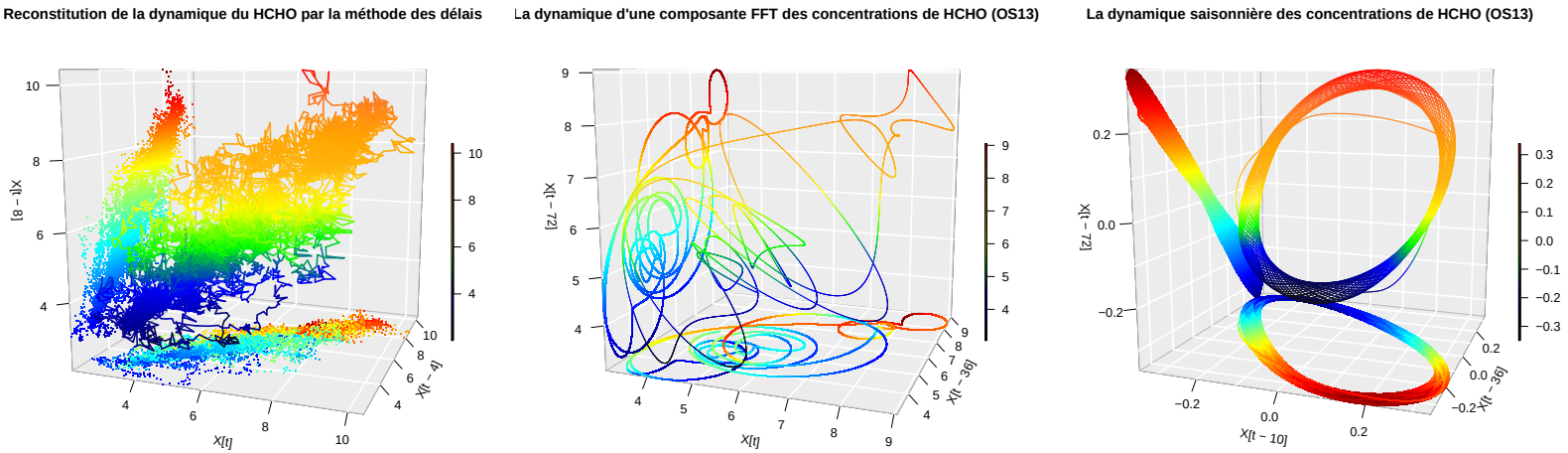


FIGURE 7.4.3 – *Reconstitution de l'espace des phases par la méthode des délais des différentes structures de la dynamique des concentrations de HCHO dans l'espace paysager pendant la campagne 2013 (pas de temps 1 minute). Les paramètres de plongement sont :* $m = 72$ *et* $\tau = 10$. *La reconstitution pour les données brutes (graphique à gauche) est formée par trois vecteurs :* X_t, X_{t-4}, *et* X_{t-8}. *Sur les données filtrées par une bande spectrale (de FFT) définie par une fréquence de coupure de* $f_c = (16.6\,h)^{-1}$, *la reconstitution est effectuée à partir de* X_t, X_{t-36}, *et* X_{t-72} *(graphique au centre). La composante saisonnière a été extraite avec la méthode STL, lissée par une régression Loess et sa reconstitution est basée sur* X_{t-10}, X_{t-36}, *et* X_{t-72}, *voir le graphique à droite. Pour les valeurs des paramètres de la méthode STL, voir la section 3.6.3.*

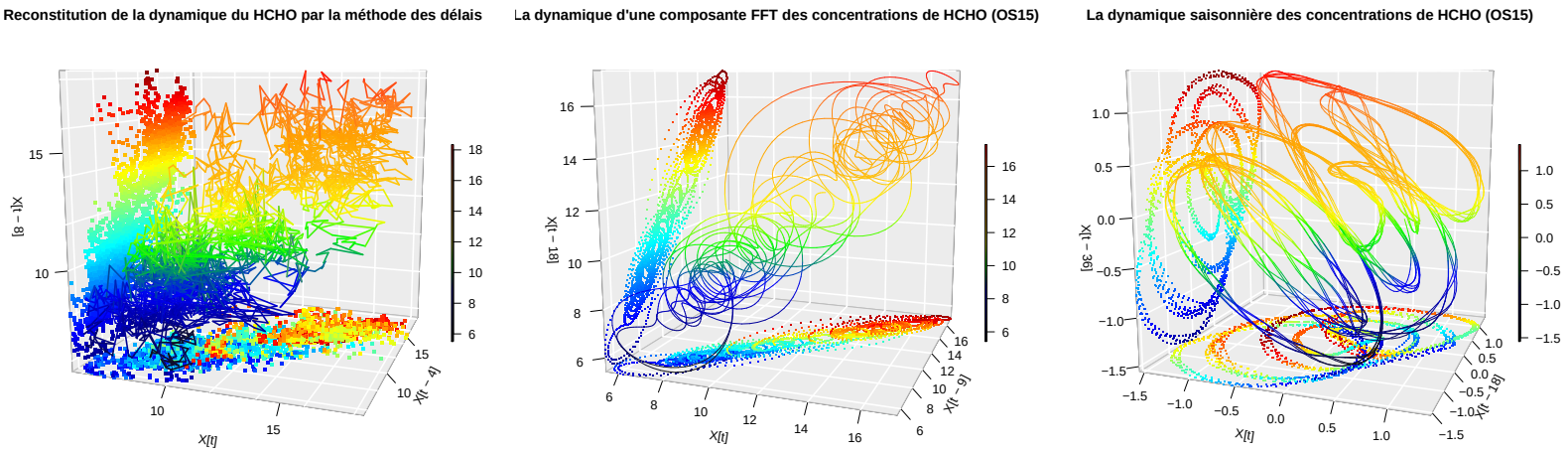


FIGURE 7.4.4 – *Reconstitution de l'espace des phases par la méthode des délais des différentes structures de la dynamique des concentrations de HCHO dans l'espace paysager pendant la campagne 2015 (pas de temps 20 minutes). Les paramètres de plongement sont :* $m = 72$ *et* $\tau = 10$ *pour tous les types de reconstitutions. La reconstitution pour les données brutes (graphique à gauche) est basée sur trois vecteurs :* $X_t, X_{t-4},$ *et* X_{t-8}*. Sur les données filtrées par une bande spectrale (de FFT) définie par une fréquence de coupure de* $f_c = (10.8\,h)^{-1}$*, la reconstitution est effectuée à partir* $X_t, X_{t-9},$ *et* X_{t-18} *(graphique au centre). La composante saisonnière a été extraite avec la méthode STL, lissée par une régression Loess, la reconstitution basée sur* $X_t, X_{t-18},$ *et* X_{t-36} *est donnée dans le graphique à droite. (Pour les valeurs des paramètres de la méthode STL, voir la section 3.6.3).*

7.4.2.1 Être, ou ne pas être chaotique, où se cachent les attracteurs étranges ?

D'une certaine manière, cette question pourrait se traduire ainsi : si le chaos existe vraiment dans les fluctuations de la QAI, où se trouve-il et comment le retrouve-t-on ? Notre réponse à cette interrogation est la suivante : il existe et on le retrouve dans les structures profondes de variabilité des séries temporelles. C'est en explorant les caractéristiques des fluctuations à des échelles fines, voire très fines, qu'on arrive à trouver des structures, qui en apparence, semblent très régulières mais qui cachent une grande complexité.

Évidement, nous ne prétendons pas explorer en détail ces structures [12], mais nous posons quelques jalons de réflexion qui nous permettront de *justifier* la nécessite de prendre en compte ces structures dans la prévision.

Nous avons évoqué précédemment l'importance accordée dans la littérature aux méthodes d'extraction de l'aléa au sein des séries temporelles. On se propose d'examiner la méthode de décomposition en bandes spectrales pour analyser les aspects liés aux attracteurs étranges.

En éliminant le "bruit aléatoire" par la FFT définie par une seule fréquence de coupure, on observe que les portraits de phase des fluctuations se structurent davantage sur les différents vecteurs délais. Cette structure "étrange" apparaît très nettement pour des composantes hautement déterministes comme la saisonnalité.

Observons maintenant le plongement des séries temporelles et leur portrait de phase associés à une certaine bande spectrale définie en amont. Les séries de polluants considérées dans cette étude sont les concentrations de formaldéhyde et les particules.

Dans les Figures 7.4.5 à 7.4.7, nous présentons le portrait de phase des séries temporelles reconstruites par FFT et IFFT correspondant à des bandes spectrales des concentrations de formaldéhyde et des concentrations en nombre de particules.

Pour les fluctuations de formaldéhyde de la campagne de 2013, la recherche des structures "*chaotiques*" s'avère plus difficile, la partie aléatoire semble se manifester sur des échelles temporelles très fines. La Figure 7.4.5 montre le plongement de quatre séries temporelles associées à quatre bandes de fréquences distinctes. Dans cette figure, on souhaitait mettre en évidence l'influence de la largeur de la fenêtre associée à une bande spectrale et la qualité de la reconstitution de l'attracteur. Pour une série temporelle de HCHO (OS13) associée à une bande spectrale $[f_1, f_2[$ autour de la fréquence principale $(12\,\mathrm{h})^{-1}$, pour une dimension de plongement $d_E = m = 144$, la forme géométrique du portrait de phase reconstruit présente une régularité assez marquée (*cf.* Figure 7.4.5). En augmentant la largeur de la fenêtre fréquentielle $[f_2, f_3[$, on observe une perturbation de la structure géométrique de l'attracteur mais gardant l'aspect régulier compact (la Figure pour la bande B_2). Plus on s'approche des hautes fréquences, plus il est nécessaire de rétrécir la fenêtre de la bande spectrale pour que la forme de l'attracteur reste régulier ; les exemples donnés pour les bandes B_3 et B_4 soutiennent cette observation.

12. Plus précisément, il s'agit de l'analyse topologique du chaos, voir Gilmore & Lefranc (2002) et Letellier & Gilmore (2013).

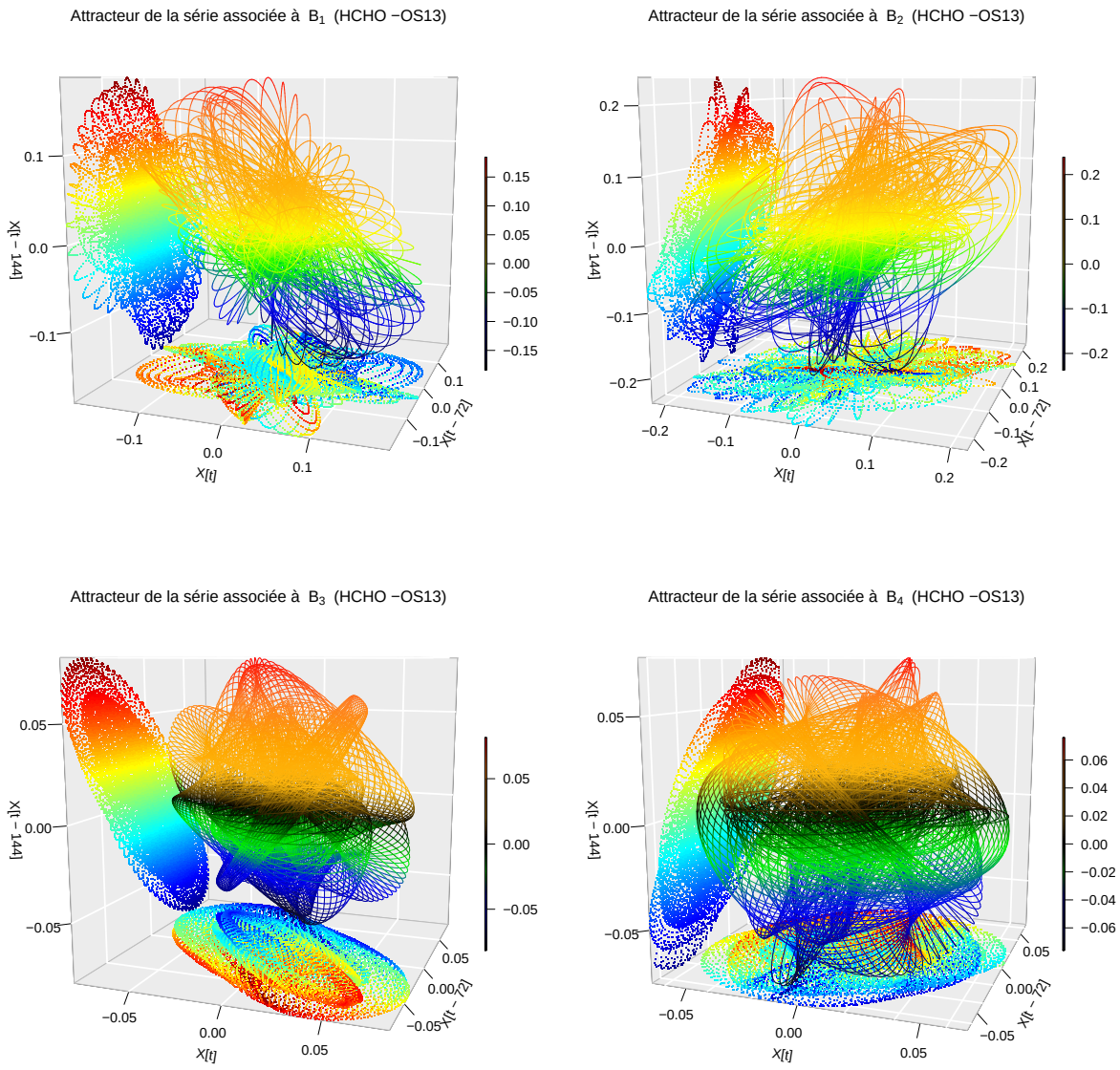
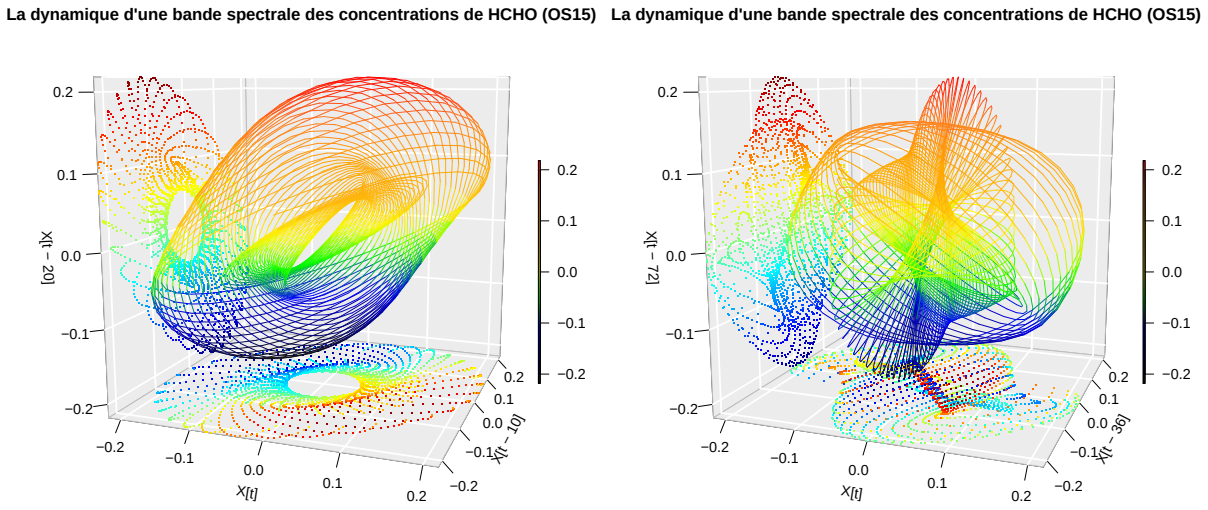
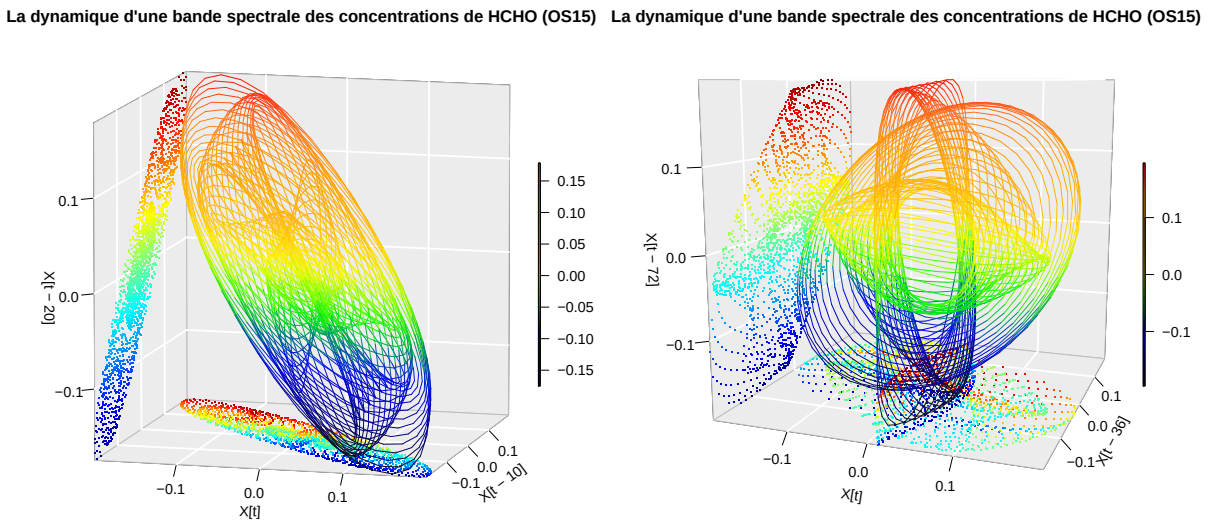


FIGURE 7.4.5 – *Attracteurs de quatre bandes spectrales (FFT) par la méthode des délais de la série temporelle de HCHO issue de la campagne de 2013 dans l'espace de bureaux (mesures toutes les minutes). La bande* $B_1 = [f_1, f_2[= \left[(100\,min)^{-1}, (91\ min)^{-1}\right[$, *la deuxième bande est* $B_2 = [f_2, f_3[= \left[(91\,min)^{-1}, (77\ min)^{-1}\right[$, *la troisième est* $B_3 = [f_3, f_4[= \left[(77\,min)^{-1}, (74\ min)^{-1}\right[$ *et la quatrième est* $B_4 = [f_4, f_5[= \left[(74\,min)^{-1}, (70\ min)^{-1}\right[$. *La reconstitution est effectuée à partir de trois vecteurs :* $X_t, X_{t-72},$ *et* X_{t-144}. *Les paramètres de plongements sont* $m = 144$ *et* $\tau = 12$ *pour les quatre bandes spectrales.*



(a) *Attracteur étrange en forme "harmonographique" d'une série temporelle associée à une bande spectrale B_1 d'un filtrage FFT pour la dynamique de HCHO. Sur le graphique à gauche, la reconstitution est basée sur les vecteurs X_t, X_{t-10} et X_{t-20} ; sur le graphique à droite, la reconstitution est effectuée avec les vecteurs X_t, X_{t-36} et X_{t-72} (tout l'espace de plongement).*



(b) *Attracteur étrange en forme "harmonographique" d'une série temporelle associée à une bande spectrale B_2 d'un filtrage FFT pour la dynamique de HCHO. Sur le graphique à gauche, la reconstitution est basée sur les vecteurs X_t, X_{t-10} et X_{t-20} ; sur le graphique à droite, la reconstitution est effectuée à partir des vecteurs X_t, X_{t-36} et X_{t-72} (tout l'espace de plongement).*

FIGURE 7.4.6 – *Attracteurs de deux bandes spectrales (FFT) par la méthode des délais de la série temporelle du HCHO issue de la campagne de 2015 (pas d'échantillonnage 20 minutes). La première bande $B_1 = [f_1, f_2[$ correspond à une plage de variabilité fréquentielle de $f_1 = (619\ min)^{-1}$ et $f_2 = (650\ min)^{-1}$ et la deuxième bande est $B_2 = [f_2, f_3[= \left[(650\ min)^{-1}, (591\ min)^{-1}\right[$ (graphique à droite). Les paramètres de plongement sont $m = 72$ et $\tau = 15$ pour les deux bandes spectrales.*

7.5 Prévision non linéaire par les systèmes dynamiques

La prévision basée sur la théorie du chaos requiert un travail une série temporelle suffisamment longue et sur sa structure ; par conséquent, le choix des paramètres de plongement (τ, d_E) est capital pour la prévision.

L'idée principale de la procédure de prévision consiste à dire que si la série temporelle est suffisamment longue, il existe probablement des états représentés dans le passé qui sont similaires aux états présents par rapport à une certaine métrique. Ceci impliquerait que d'autres états vont probablement se répéter avec plus ou moins de variations (Hegger et al., 1999). Pour illustrer ce propos, soit le système dynamique à temps discret :

$$S_{t+1} = f\left(S_t\right). \tag{7.5.1}$$

Cette écriture précise le fait que l'état futur est complètement spécifié par l'état récent au moment t, ce qui implique la prévision sans *ambigüité*. Or, toute incertitude ou erreur dans la spécification de l'état présent va se répercuter au cours du temps, et de manière exponentielle pour les systèmes chaotiques. Mais, même pour ce cas (du chaos), l'incertitude est amplifiée uniquement à un taux fini (Kantz & Schreiber, 2004) ; nous espérons par conséquent fournir des prévisions raisonnables à court terme. Le problème majeur dans la formule 7.5.1 vient principalement du fait que la connaissance de f est irréaliste pour les données réelles. Par une hypothèse minimale de continuité, on peut construire un schéma de prévision très simple. L'origine de cette idée remonte à Lorenz (1969) (méthode analogue de Lorenz), stipulant que la trajectoire dans l'espace d'état pour le plus proche voisin d'un point donné va suivre ce voisin en prévision. Plus précisément, on suppose que $t_0 < T$ et $t_0 + 1 \leq T$; pour prédire l'état futur S_{T+1} ayant observé l'état présent S_T, on cherche une liste dans tout le passé de S_t $(t < T)$ le plus proche de S_T. Si l'état au temps t_0 est similaire au présent, la continuité de f garantit que S_{t_0+1} devrait être très proche de S_{T+1}.

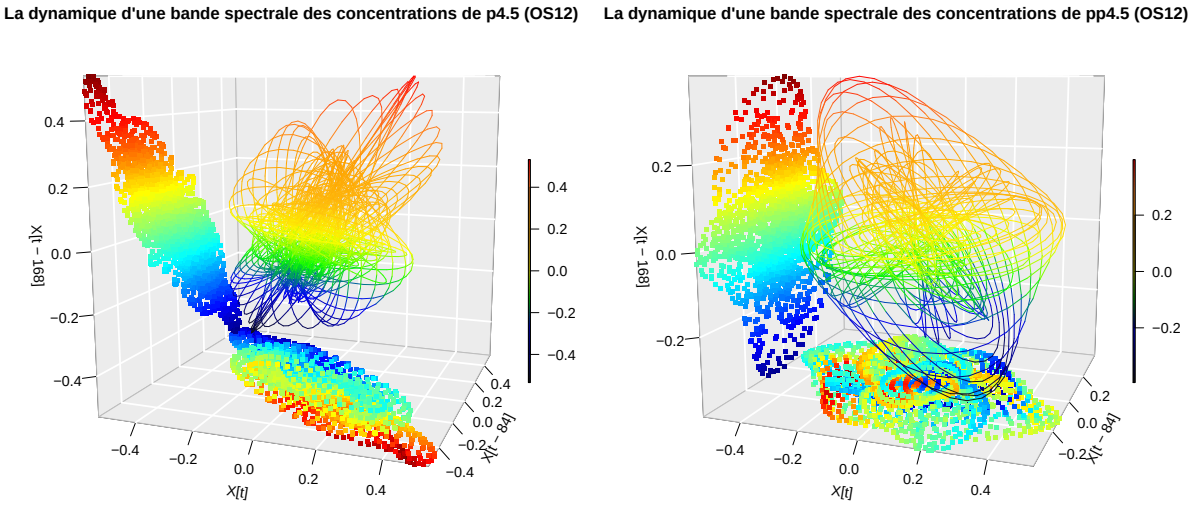
En revanche, l'observation directe de certains de ces états est très difficile, même dans le cas d'expériences physiques bien maîtrisées d'un système déterministe. Les observations faites dans un environnement réel sont uniquement les effets de ces états effectifs du système. La thèse s'inscrit dans le cadre des méthodes inverses, donc les causes (sources) sont inférées par l'observation des effets (concentrations en polluants). En termes mathématiques, les sorties du système dépendent fonctionnellement des entrées :

$$X_t = h\left(S_t\right), \qquad t = 1, \ldots, T. \tag{7.5.2}$$

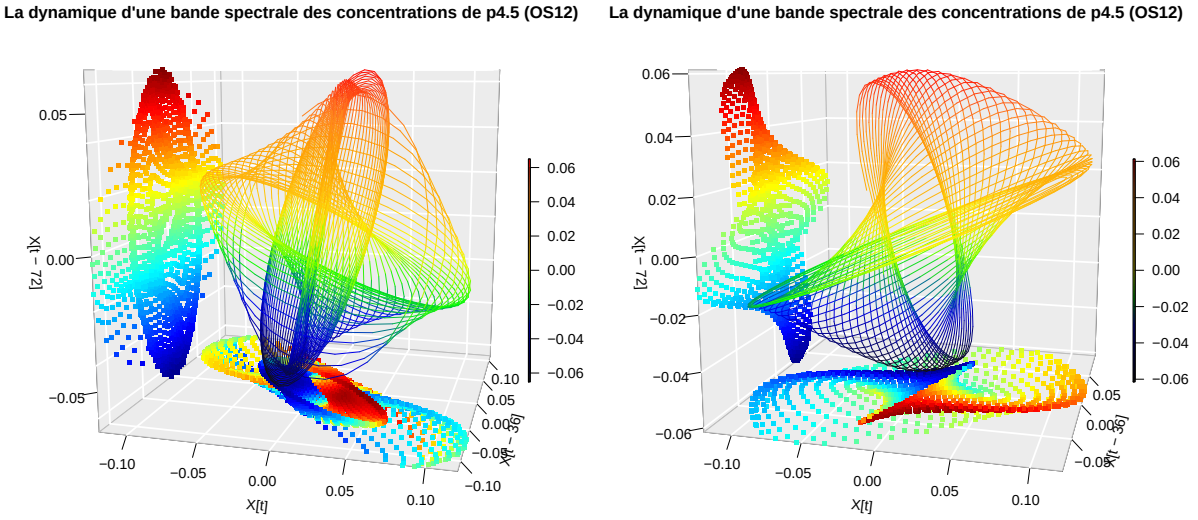
Le plus souvent, h comme f sont inconnues . Le théorème de Takens (1981) de la reconstitution de l'espace d'état par la méthode de délais offre une alternative très utile pour contourner ces difficultés en prévision [13]. En effet, la représentation du vecteur de délais

$$\boldsymbol{X}_t = \left(X_t, X_{t-\tau}, X_{t-2\tau}, \ldots, X_{t-(d_E-2)\tau}, X_{t-(d_E-1)\tau}\right)^\top, \tag{7.5.3}$$

13. Il faut noter que le théorème fait abstraction du bruit et la taille finie des observations.



(a) *Attracteurs étranges en forme "harmonographique" d'une série temporelle associée à une bande spectrale B_1 d'un filtrage FFT pour la dynamique des particules inférieures à 4.5 µm. Les paramètres de plongement sont $m = 72$ et $\tau = 1$ pour le graphique de droite et $m = 72$ et $\tau = 10$ pour le graphique de gauche.*



(b) *Attracteurs étranges en forme "harmonographique" d'une série temporelle associée à une bande spectrale B_2 d'un filtrage FFT pour la dynamique des particules inférieures à 4.5 µm. Les paramètres de plongement sont $m = 72$ et $\tau = 10$ pour le graphique de droite et $m = 72$ et $\tau = 15$ pour le graphique de gauche.*

FIGURE 7.4.7 – *Attracteurs étranges en forme "harmonographique" d'une série temporelle associée à une bande spectrale B_1 d'un filtrage FFT pour la dynamique des particules de taille inférieures à 4.5 µm à l'intérieur de l'espace paysager, campagne 2012, pas de temps horaire.*

donne lieu à un accès très "rapide" aux informations véhiculées par le système.

En vue d'obtenir les prévisions, plusieurs algorithmes et procédures ont été proposés dans la littérature par Farmer & Sidorowich (1987); Casdagli (1989, 1992); Abarbanel et al. (1993); Abarbanel (2012).

7.5.1 Les méthodes locales

Les méthodes locales sont basées sur la recherche des évènements similaires dans le passé. Formellement, il s'agit de trouver les points $\boldsymbol{X}_k$ telles que $|\boldsymbol{X}_T - \boldsymbol{X}_k| \leq \epsilon$ *i.e.* $\boldsymbol{X}_k \in \mathcal{U}_\epsilon(\boldsymbol{X}_T)$, où $\mathcal{U}_\epsilon(\boldsymbol{X}_T)$ est le voisinage du $\boldsymbol{X}_T$; ce voisinage forme une boule fermée. Dans la pratique, on suppose uniquement un nombre correspondant aux plus proches voisins et on ignore le rayon ϵ.

7.5.1.1 Moyenne au voisinage local (Locally constant predictor)

Cette méthode est similaire à la "méthode analogue de Lorenz" dans son concept, mais différente dans les calculs. On choisit un certain nombre k des voisins proche de $\boldsymbol{X}_T$, on calcule leur moyenne et cette valeur sera la prévision de $\boldsymbol{X}_T$.

Cette méthode est aussi appelée approximation locale d'ordre zéro (**LZO**), car elle ne fait intervenir aucun ordre polynomial d'une régression particulière.

On suppose dans ce qui suit que le choix de τ et de d_E est fait. Pour une prévision à un horizon $T+h$, nous choisissons un paramètre ϵ de même ordre que la résolution temporelle des observations avec lequel on définit un voisinage $\mathcal{U}_\epsilon(\boldsymbol{X}_T)$ de rayon ϵ autour du point $\boldsymbol{X}_T$. Pour tout point $\boldsymbol{X}_t \in \mathcal{U}_\epsilon(\boldsymbol{X}_T)$, donc pour tous les points ayant une distance au plus de ϵ par rapport à $\boldsymbol{X}_T$, on prend comme "prévision" la moyenne des prévisions individuelles de chaque X_{t+h} :

$$\widehat{X}_{T+h} = \frac{1}{\#\left[\mathcal{U}_\epsilon(\boldsymbol{X}_T)\right]} \sum_{\boldsymbol{X}_t \in \mathcal{U}_\epsilon(\boldsymbol{X}_T)} X_{t+h}, \tag{7.5.4}$$

où $\#[\bullet]$ désigne le nombre d'éléments du voisinage $\mathcal{U}_\epsilon(\boldsymbol{X}_T)$. Dans le cas où aucun point voisin n'est trouvé, on peut augmenter le rayon de la boule qui balaye l'attracteur jusqu'à ce que cette dernière contienne des points voisins.

7.5.1.2 Pondération linéaire au voisinage local

Dans cette méthode, la prévision est obtenue en résolvant le problème de régression linéaire en prenant des évènements similaires dans le passé, *i.e.* les voisinages dans l'espace de phase (reconstruit). Cette méthode est appelée ajustement linéaire local (**LFO**) ; elle nécessite un paramétrage particulier pour éviter que le système soit singulier.

Pour prédire $\boldsymbol{X}_{T+1}$, la procédure est donné pas la formule suivante

$$\widehat{X}_{T+1} = \boldsymbol{A}\boldsymbol{X}_T + \boldsymbol{b}, \qquad \boldsymbol{A} \in \mathbb{R}^{d_E \times d_E}, \boldsymbol{b} \in \mathbb{R}^{d_E}, \tag{7.5.5}$$

ou, de manière équivalente,

$$\widehat{X}_{T+1} = a_1 X_T + a_2 X_{T-\tau} + \cdots + a_{d_E} X_{T-d_E(\tau-1)} + b, \tag{7.5.6}$$

où $\boldsymbol{A}$ et $\boldsymbol{b}$ sont déterminés par la méthode des moindres carrés de

$$\arg\min_{a,b} \sum_k \left(a_1 x_k + a_2 x_{k-\tau} + \cdots + a_d x_{k-\tau(d-1)} + b - x_{(k+1)}\right)^2. \tag{7.5.7}$$

Cette somme s'effectue sur le nombre de point supérieur à d_E+1 points $\boldsymbol{X}_k = \left(x_k, x_{k-\tau}, \cdots, x_{k-\tau(d-1)}\right)$.

7.5.2 Les méthodes globales

Ces méthodes consistent à paramétrer la fonction f dans l'équation 7.5.1 et estimer ses paramètres. Les modèles les plus utilisés sont de type :

— fonctions pôlynomiales ;
— fonctions radiales (RBF) ;
— réseaux de neurones.

Dans cette thèse, nous utilisons uniquement les méthodes locales.

7.6 Applications aux concentrations de polluants de la QAI

Nous présentons dans cette section la méthodologie de prévision des concentrations des polluants par la méthode de la dynamique du chaos. La méthode suivie est basée sur l'optimisation des paramètres de plongement de la méthode des délais sur un ensemble de validation, ensuite les paramètres obtenus sont appliqués sur un ensemble de test. Comme dans le chapitre précédent, nous ne discuterons pas le choix des ensembles (validation et test) ; le découpage est fait selon des considérations pratiques : environ 75% pour l'apprentissage, 15% pour la validation et 10% pour le test.

On peut se référer au Tableau 6.4.1 pour les détails relatifs aux données d'apprentissage et de test. En raison de la présence de plusieurs paramètres qui interviennent dans cette approche, on se propose de donner une stratégie globale de prévision par la méthode de la dynamique du chaos. Cette méthodologie est plus ou moins applicable pour une certaine classe des séries temporelles de polluants où aucun prétraitement n'est appliqué ; nous y reviendrons ci-après.

On commence par fixer les notations qui serviront pour la suite. Ces notations sont en accord avec celles posées par Kantz & Schreiber (2004) dans le package `TISEAN(3.0.1)` [14] de Hegger et al. (1999) :

— la dimension de plongement : m ;
— le délai : τ ;
— le nombre minimum de voisins : k ;
— le rayon de la boule formée par les voisins proches : ϵ ;
— le facteur d'accroissement $r \geq 1$ de la boule si aucun voisin n'est trouvé (ce paramètre est noté f dans `TISEAN(3.0.1)` [15]).

14. **TI**me **SE**ries **AN**alysis : un ensemble de fonctions écrites sous C et Fortran pour l'analyse des séries temporelles non-linéaires. Ce package développe plus les modèles non-paramétriques du chaos. Il a été inclus dans les bibliothèques de R (2015) jusqu'à la version 2.15 de R uniquement.

15. C'est le seul changement effectué, car tout au long de la section f renvoie à la relation entre les états du système dynamique.

L'optimisation de ces paramètres fait intervenir plusieurs questions relatives à leur stabilité en prévision, notamment la relation entre les structures de variabilité et les paramètres de cette méthode. Par exemple, lorsque les conditions de prévision sont favorables pour un polluant donné, le sont-elles pour la variabilité d'un autre polluant ? Quel rôle joue la résolution temporelle utilisée pour les différents polluants ? Y a-t-il une règle générale pour les applications empiriques ?

Ces questions sont très liées entres elles et nécessitent une discussion à partir de plusieurs points de vue. Nous adoptons pour cela une stratégie de prévision qui se décline sur deux niveaux (notés (**I**) et (**II**) dans la description qui suit).

(I) Le compromis entre le nombre k des voisins et le facteur d'accroissement r

Dans cette étape, nous déterminons les conditions générales pour lesquelles les paramètres k et r forment un compromis pour un certain couple de paramètres m et τ fixés. Ce compromis, obtenu pour les concentrations de CO_2, servira de base pour tous les autres polluants. On commence par le CO_2, car c'est le polluant le plus "facile" à modéliser.

1. On fixe le nombre k de plus proches voisins, on fait varier le paramètre r et on calcule les prévisions ainsi que les performances ;
2. On fixe le paramètre r obtenu dans (1), on fait varier k et on calcule les prévisions ainsi que les performances.

Le meilleur compromis consiste à prendre les paramètres qui minimisent le RMSE (pour *Root-Mean-Square Error*) dans l'étape de validation.

Dans cette étape, on a constaté que plus le nombre k est grand, plus les prévisions sont bruitées. En outre, si $r > 1.1$ avec un k petit ($k < 5$), les prévisions fournies par la dynamique du chaos sont très mauvaises. En se basant sur la méthode **LZO**, le résultat (compromis final) est le suivant : $k \in \{1, 2, 3\}$ et $r \in [1.01, 1.05]$. Nous présenterons ci-après les résultats obtenus pour les prévisions des concentrations de CO_2.

Sauf mention contraire, les résultats de prévision pour les polluants autres que le CO_2, on prend $k^* = 3$ et $r^* = 1.05$.

(II) Compromis entre la dimension de plongement m et le délai τ

Cette étape consiste en :

1. l'optimisation des paramètres de plongement (la dimension de plongement m et le délai τ) sur un ensemble de validation ;
   (a) fixer le délai τ ;
      i. varier la dimension de plongement m ;
      ii. calculer le RMSE pour chaque horizon de prévision sur l'ensemble de validation et obtenir la meilleure dimension de plongement m^*.
   (b) pour m^* fixé :
      i. varier le délai τ ;
      ii. calculer le RMSE pour chaque horizon de prévision sur l'ensemble de validation et obtenir le meilleur délai τ^*.
2. Utiliser le couple (m^*, τ^*) pour la prévision (en aveugle) sur l'ensemble de test.

À présent, nous discuterons les résultats de prévision obtenus pour différentes conditions de simulation. Les prévisions du CO_2 par la méthode **LZO** donnent les plages de variation des paramètres à utiliser pour la prévision des autres polluants.

Table 7.6.1 – Conditions de simulation pour l'optimisation des paramètres de voisinage pour la dimension de plongement $m = 50$ et le délai $\tau = 6$.

		k	r
bloc 1	Sim1	2	1.01
	Sim2	2	1.05
	Sim3	2	1.09
	Sim4	2	1.2
	Sim5	2	1.5
	Sim6	2	1.8
bloc 2	Sim7	3	1.05
	Sim8	5	1.05
	Sim9	10	1.05
	Sim10	15	1.05
	Sim11	20	1.05
	Sim12	25	1.05

Notes : les cellules en vert clair fournissent la meilleure combinaison entre k et r. Pour optimiser le temps de calcul (∼1h/simulation), nous avons choisi de prendre dans le bloc 2 $r = 1.05$ (cellule grisée) au lieu de prendre la valeur optimale de 1.01.

7.6.1 Prévision des concentrations du CO_2

Nous avons appliqué la méthode **LZO** aux fluctuations du CO_2 visant un double objectif :

— Déterminer la relation entre la prévision et la variation des paramètres de voisinage *i.e.* le facteur d'accroissement r et le nombre de voisins proches.

— Optimiser les paramètres de plongement (m et τ) et tester de la prévision sur un ensemble "neutre" de test.

Nous avons suivi la procédure présentée précédemment et le Tableau 7.6.1 fournit les conditions générales de simulation pour la prévision des concentrations de CO_2 sur un ensemble de validation d'une semaine (1008 points). En effet, sur 52560 observations, les 20000 premières ont été utilisées en apprentissage (constitution de la matrice des délais) et une semaine pour la validation. Ensuite, la semaine de validation a été incorporée dans les 20000 observations d'apprentissage ($n = 21008$) pour prédire une semaine de test.

Les mesures sont au pas de temps de 10 minutes et elle proviennent de bureau individuel. Le bloc 1 du Tableau 7.6.1 concerne l'optimisation du facteur d'accroissement r en fixant le nombre de voisins à 2 et le bloc 2 donne les conditions d'optimisation du nombre de voisins k avec un facteur d'accroissement $r = 1.05$. Bien que la prévision optimale en validation soit donnée par un accroissement d'un facteur $r = 1.01$, l'optimisation du nombre des plus proches voisins est obtenue pour un facteur d'accroissement de 1.05, et ceci pour des raisons de temps de calcul.

La performance des résultats de la prévision des différentes simulations (*cf.* Tableau 7.6.1) est présentée graphiquement dans la Figure 7.6.1. L'évolution du RMSE en fonction de l'horizon de prévision est présentée pour les deux étapes de la procédure de simulation : optimisation du nombre des plus proches voisins k, et du facteur d'accroissement r.

Globalement, ni r ni k ne détermine le niveau de prévision à très court terme : toutes les combinaisons de ces paramètres sont équivalentes. En effet, le RMSE est quasiment identique pour toutes les simulations pour un horizon de prévision inférieur à 32 h. Au delà de cette échéance, le choix des paramètres de

voisinage devient de plus en plus important. Pour un nombre minimal des plus proches voisins ($k = 2$), le critère RMSE correspondant à $r = 1.01$ est très faible (RMSE < 74 ppm), même pour un horizon d'une semaine de prévision (*cf.* Figure 7.6.1a). En revanche, la complexité du calcul pour de tels paramètres s'avère plus grande. Alors, on choisit des valeurs plus faibles pour r et k.

En tenant compte de cette contrainte, nous avons choisi, pour l'optimisation de k, un $r = 1.05$. Sur la Figure 7.6.1b, nous présentons les résultats de l'évolution du critère RMSE des simulations pour $r = 1.05$. Clairement, la simulation 7 correspondant à un nombre $k = 3$ donne les meilleures prévisions. Ainsi, on peut conclure, quant à la relation entre les paramètres de voisinage et la qualité de prévision : de façon générale, il faut prendre des valeurs faibles pour k et de r afin d'espérer atteindre des prévisions acceptables.

En outre, un aspect très important, mais qui n'a pas été assez discuté dans cette thèse est celui de la relation entre les paramètres de voisinage et la résolution temporelle (pas d'échantillonnage) des observations. En fait, nous avons remarqué que plus le temps échantillonnage est fin, plus k doit être fin lui aussi, car le pas de temps influe sur le niveau de variabilité de la série, donc sur la variance. Compte tenu de cette remarque, une valeur élevée de k va générer des prévisions, par les méthodes locales, dont la trajectoire sera très bruitée.

Nous avons, par ailleurs, recherché le compromis entre les paramètres de plongement et les performances en prévision. En faisant varier m et τ séquentiellement, avec k et r fixes, il apparaît que le choix "optimal" correspond à une dimension de plongement égale à une semaine ($m = 1008$) et le retard d'une valeur unitaire ($\tau = 1$). En effet, environ 40 simulations ont été réalisées afin de permettre d'identifier les plages de variation ou les conditions dans lesquelles l'erreur de prévision en termes de RMSE est minimale.

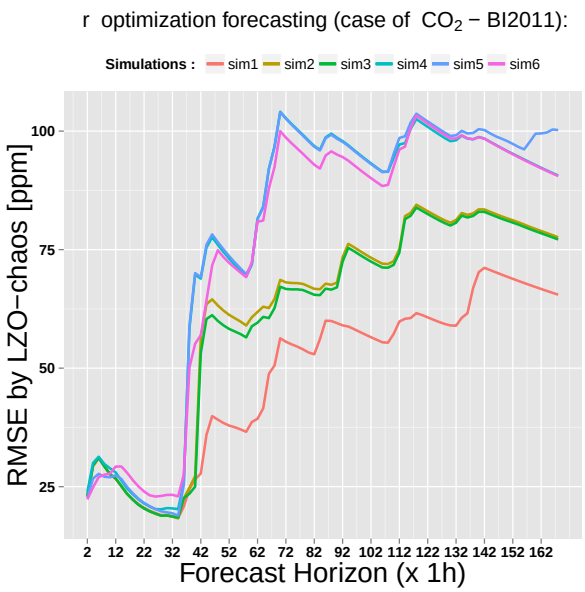
La condition nécessaire (mais pas suffisante) pour une prévision acceptable est la suivante : le produit des deux paramètres de plongement doit avoir une valeur approximative correspondant à la période principale de la série.

Une question immédiate survient avec ce type d'assertion, car il existe plusieurs combinaisons pour le produit de deux entiers ayant un résultat identique ; quelle est alors la bonne combinaison ? Pour l'instant, on a une solution à un facteur près. Il faut donc trouver encore une contrainte. Pour cela, on regarde l'influence de la résolution temporelle de la série.

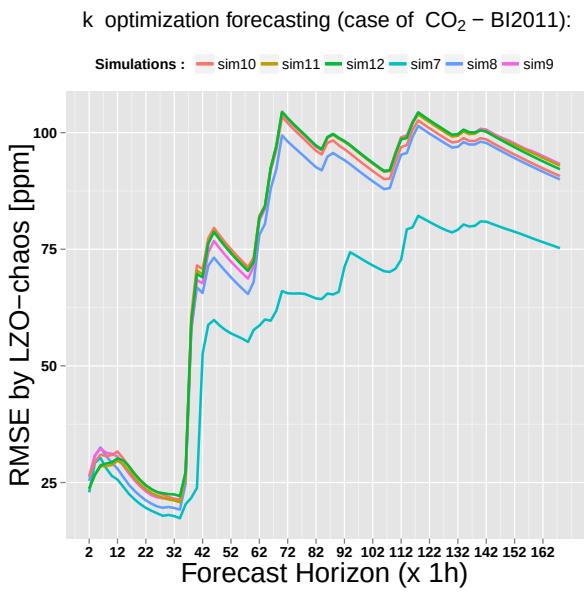
En effet, pour un pas de temps d'une minute à 10 minutes, la dimension de plongement m doit être le majorant de toutes les combinaisons et le délai τ leur minorant : *i.e.* m égal à la période principale et la fenêtre τ égale à 1. Pour un pas de temps supérieur à 10 minutes, nous modifierons légèrement cette considération, en augmentant la valeur de τ.

Pour l'exemple du CO_2, la période principale des fluctuations est d'une semaine (1008 valeurs) pour un pas de temps de 10 minutes ; on peut construire plusieurs combinaisons pour avoir cette valeur : $(m, \tau) = \{(1008, 1) ; (252, 4) ; (144, 7), \ldots\}$. La meilleure prévision fournie par la méthode **LZO** est réalisée pour la configuration $m = 1008$ et $\tau = 1$, mais de bonnes prévisions ont été enregistrées aussi pour la combinaison $m = 144$ et $\tau = 7$. Pour cette dernière, la complexité algorithmique pour la recherche des similarités est faible, ce qui implique un temps de calcul très réduit.

Nous présentons sur la Figure 7.6.2 la semaine de prévision par la meilleure combinaison des paramètres, *i.e.* $m = 1008$, $\tau = 1$, $k = 2$ et $r = 1.05$. La qualité de prévision fournie par la méthode **LZO** de la dynamique du chaos est très bonne, l'indice RMSE est toujours inférieur à 75 ppm et ceci quelque soit



(a) *Résultats des performances de prévision (RMSE) des concentrations de CO_2 sur un ensemble test d'une semaine avec les différentes valeurs de r pour un nombre de voisins $k = 2$ (bloc 1 du Tableau 7.6.1).*



(b) *Résultats des performances de prévision (RMSE) des concentrations de CO_2 sur un ensemble test d'une semaine avec les différentes valeurs de k pour un facteur d'accroissement $r = 1.05$ (bloc 2 du Tableau 7.6.1).*

FIGURE 7.6.1 – *Évolution de l'erreur quadratique moyenne (RMSE) dans l'étape de validation pour l'optimisation des paramètres de voisinage pour $m = 50$ et $\tau = 6$ dans le cas du CO_2 mesuré avec un pas de temps de 10 minutes dans le bureau individuel pendant la campagne 2011.*

l'horizon de prévision ; les valeurs du critère MAPE restent en deçà de 40 ppm. Par comparaison avec la méthode Holt-Winters (HW), la qualité de prévision fournie par le modèle local **LZO** est bien supérieure à celle obtenue par HW à court terme (inférieur à deux jours) ; les performances sont équivalentes à moyen terme (1-4 jours), mais pour un horizon de prévision total d'une semaine, la méthode de HOLT-WINTERS donne de meilleures prévisions en moyenne.

Les paramètres obtenus dans l'étape de validation ont été utilisés pour la prévision sur un ensemble de test d'une semaine. Sur la Figure 7.6.3, on présente les résultats de cette prévision ainsi que leurs performances par les deux critères RMSE et MAE. On remarque d'abord que l'échantillon test ne ressemble pas à une semaine type observée sur l'année, ainsi les concentrations durant le mardi (entre 3200-4640 minutes) présentent des fluctuations très faibles. Le modèle de prévision échoue dans la détection de ce changement, dû probablement à l'inoccupation ou à une forte aération. Rappelons que le profil de variabilité du mercredi diffère légèrement des autres jours de la semaine, le modèle est cependant capable de détecter cette caractéristique en validation et en test.

En résumé, les résultats importants qui ressortent de l'ensemble des simulations par la méthode **LZO-Chaos** peuvent être listés comme suit :

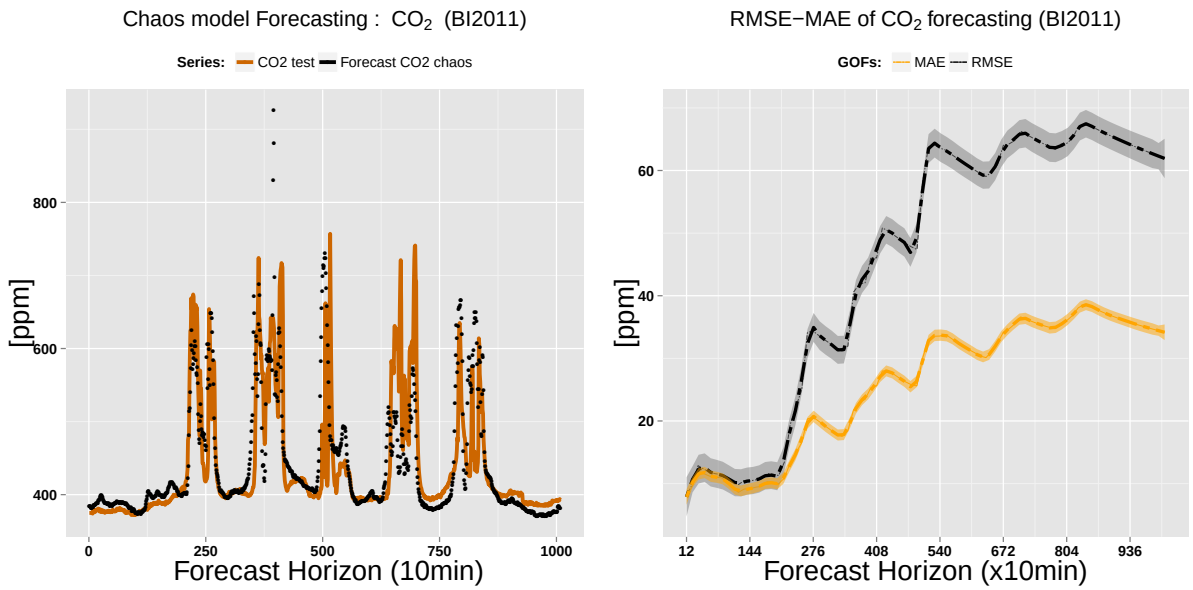


FIGURE 7.6.2 – *Prévision et performances (RMSE et MAE) par la méthode* ***LZO*** *des concentrations de CO_2 sur un horizon d'une semaine de* **validation**. *Les paramètres optimisés sont $m = 1008$, $\tau = 1$, $k = 2$ et $r = 1.01$. Les mesures de CO_2 sont au pas de temps de 10 minutes et issues de la campagne de 2011 dans le bureau individuel. Les courbes de performance ont été ajustées avec une régression Loess pour fournir aussi un intervalle de confiance (au seuil de 99%) de chaque indice.*

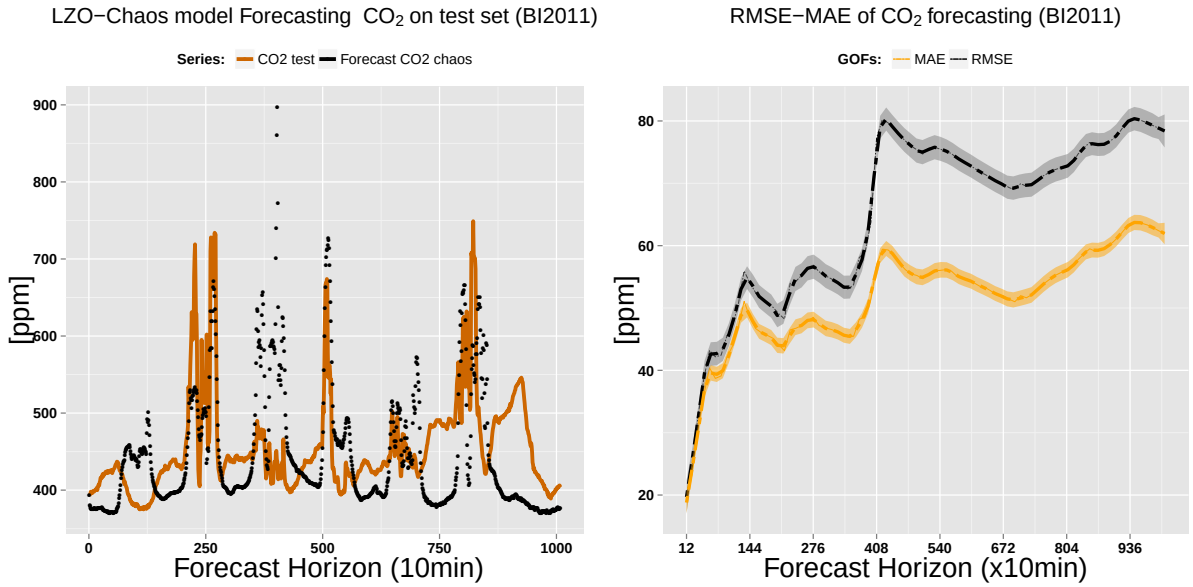


FIGURE 7.6.3 – *Prévision et performances (RMSE et MAE) par la méthode* ***LZO*** *des concentrations d CO_2 sur un horizon d'une semaine de* **test**. *Les paramètres optimisés sont $m = 1008$, $\tau = 1$, $k = 2$ et $r = 1.01$. Les mesures sont au pas de temps de 10 minutes et issues de la campagne de 2011 dans le bureau individuel. Les courbes de performance ont été ajustées avec une régression Loess pour fournir aussi un intervalle de confiance (au seuil de 99%) de chaque indice.*

1. pour des prévisions “acceptables”, les paramètres de voisinage doivent être petits par rapport au nombre d'observations ;
2. pour un modèle optimisé, les prévisions fournies par la méthode locale moyenne parvient à prévoir systématiquement les absences des mercredis après-midis et les week-ends, mais beaucoup moins l'inoccupation associée aux événements aléatoires ;
3. pour un τ élevé, les prévisions sont très fluctuantes sur un régime de faible variabilité (en l'absence de l'occupant) ;
4. plus la dimension de plongement m est grande, plus il faut réduire la fenêtre de délai τ, et inversement.

De manière plus globale, on propose le choix des paramètres de plongement comme ci-encadré :

Par des considérations empiriques, il faut choisir les paramètres m et τ de manière à avoir $m \times \tau = f_p^{-1}$, où f_p^{-1} est la période principale de la série temporelle. Pour le CO_2, $f_p^{-1} = 1$ semaine $= 1008$ valeurs de données au pas de temps de 10 minutes. Néanmoins, nous privilégions un m élevé et τ petit, car comme mentionné précédemment, un τ élevé génère des prévisions plus fluctuantes (bruitées).

7.6.2 Prévision des concentrations de HCHO

Nous retenons le dernier résultat de la section précédente et nous appliquons la procédure **LZO** de la théorie du chaos pour la prévision des fluctuations de formaldéhyde. Nous nous référons au Tableau 6.4.1 pour les informations relatives au découpage des séries temporelles des concentrations de HCHO en ensemble d'apprentissage et de test pour les campagnes de l'espace paysager.

7.6.2.1 Prévision de la série HCHO de la campagne de 2013

Un modèle de la dynamique du chaos a été appliqué aux données des concentrations de HCHO issues de la campagne de 2013. Les prévisions ainsi que les performances du modèle **LZO** sont présentées dans la Figure 7.6.4. Clairement, la prévision globale n'est pas très mauvaise, mais elle ne permet pas de détecter le changement abrupt survenu après 8 h de prévision. La qualité de la prévision du modèle reste acceptable sur les 11 premières heures de prévision, l'indice RMSE étant inférieur à 1 ppb sur cet horizon.

Notons que les performances globales exprimées en RMSE ou en MAE donnent uniquement le niveau moyen de prévision, indépendamment du signe. En effet, contrairement au modèle STL-ARIMA où la prévision sous-estime systématiquement le changement abrupt, le modèle **LZO** de la dynamique du chaos surestime légèrement la variabilité au début du changement abrupt, après exactement 8 h de prévision. Pour cette considération, la qualité de prévision du modèle **LZO** est un peu meilleure à celle obtenue par STL+ARIMA pour un horizon incluant le changement abrupt, *i.e.* après 11.5 h de prévision. En effet, alors que les erreurs en prévision par modèle le STL-ARIMA augmentaient très rapidement après 550 minutes, elles commencent à augmenter après 602 minutes pour le modèle **LZO**. Ce dernier modèle semble préférable au modèle basé sur la combinaison STL-ARIMA.

À présent, nous comparons ces résultats avec la méthode SSA, présentée à la fin du chapitre 6. Bien que la méthode SSA et le modèle **LZO-**chaos soient basés sur le même principe de plongement des séries temporelles, il y a une différence notable entre elles dans leur formulation en prévision.

Par comparaison avec le modèle SSA, la prévision à court terme par **LZO** est préférable à celle avec le modèle SSA, mais la qualité de prévision moyenne de ce dernier est supérieure au modèle **LZO** sur la totalité de la série de test.

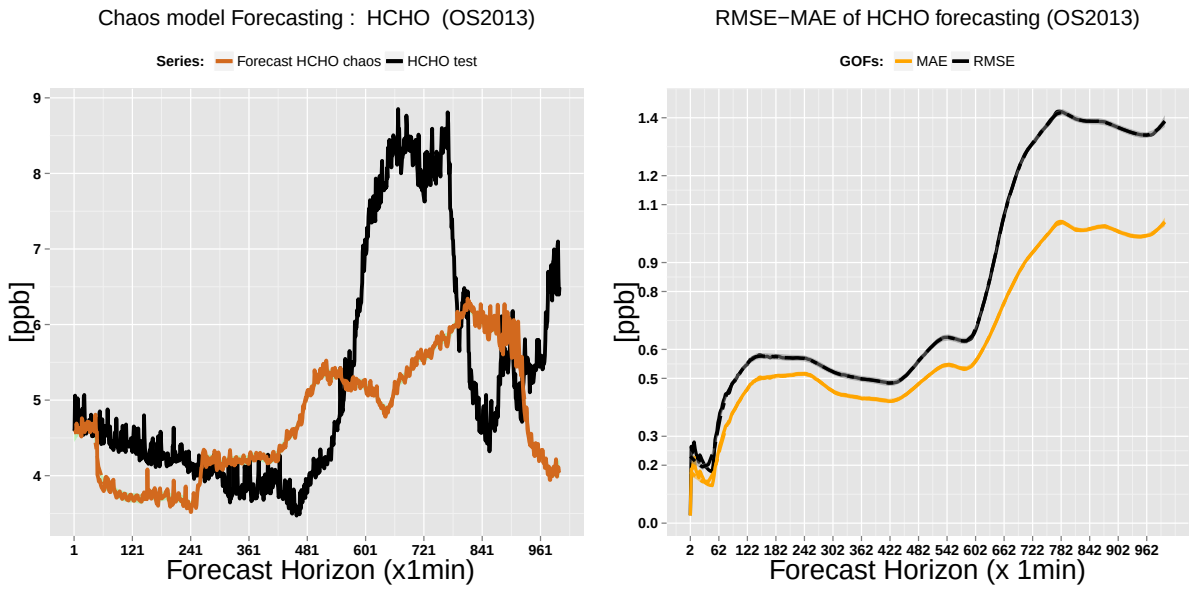


FIGURE 7.6.4 – *Prévision et performances (RMSE et MAE) par la méthode* ***LZO*** *des concentrations du HCHO sur un horizon de quatre jours de* **test**. *Les paramètres utilisés pour cette méthode sont* $m = 650$, $\tau = 1$, $k = 2$ *et* $r = 1.05$. *Les mesures sont au pas de temps d'une minute et elles sont issues de la campagne de 2013 dans l'espace paysager. Les courbes de performance ont été ajustées avec une régression Loess pour fournir aussi un intervalle de confiance (au seuil de 99%) de chaque indice.*

7.6.2.2 Prévision de la série HCHO de la campagne de 2015

Compte tenu des informations fournies par l'analyse des structures de variabilité temporelle, la dynamique des fluctuations de HCHO durant la campagne de 2015 était plus régulière par rapport à la campagne de 2013. Nous choisissons alors une combinaison des paramètres de plongement comme suit : $m = 72$ (un jour), et $\tau = 6$ (2 heures).

La Figure 7.6.5 montre la variabilité de la concentration pour les données tests leurs prévisions par **LZO** ainsi que les performances. Le modèle du chaos reste fidèle au niveau des fluctuations globale. Ainsi, il reproduit clairement l'oscillation journalière, mais échoue dans la détection des changements abruptes survenus après 1.5 jour de prévision. Au troisième jour de prévision, le modèle surestime largement (~ 4 ppb) les fluctuations. En ce qui qui concerne la qualité de prévision mesurée par le RMSE et MAE, elle inférieure à 1.26 ppb sur tout l'horizon de prévision pour le RMSE et inférieure à 1 ppb pour MAE.

En définitif, la règle selon laquelle un bon compromis entre les paramètres de plongement ($m \times \tau = f_p^{-1}$) aboutit à des prévisions acceptables est vérifiée.

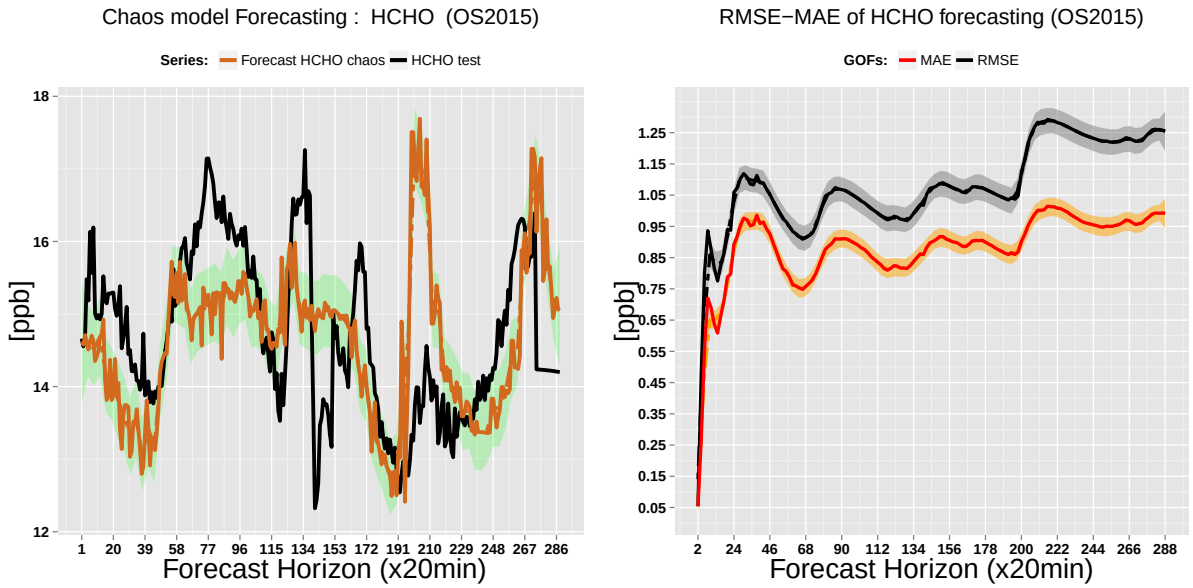


FIGURE 7.6.5 – *Prévision et performances (RMSE et MAE) par la méthode* ***LZO*** *des concentrations du HCHO sur un horizon de quatre jours de* **test**. *Les paramètres utilisés pour cette méthode sont* $m = 72$, $\tau = 6$, $k = 3$ *et* $r = 1.05$. *Les mesures sont au pas de temps de 20 minutes et elles sont issues de la campagne de 2015 dans l'espace paysager. Les courbes de performance ont été ajustées avec une régression Loess pour fournir aussi un intervalle de confiance (au seuil de 99%) de chaque indice.*

7.6.3 Prévision des concentrations de particules

Nous poursuivons la même procédure de prévision en respectant une paramétrisation adéquate pour la prévision des concentrations de particules de la campagne 2012. Rappelons que la résolution temporelle est de 20 minutes. Nous présentons uniquement les prévision des fractions 0.35 µm et 4.5 µm. Les résultats sont présentés sur la Figure 7.6.6.

Une première remarque générale peut être dégagée, la qualité de prévision obtenue par le modèle **LZO** dépend fortement du polluant mis en jeu. En effet, prévision est beaucoup plus meilleurs pour les particules moyennes que pour les particules fines, car l'évolution temporelle de celles-ci exhibent des variations brusques que le modèle n'arrive pas à intégrer en prévision. Par contre, les l'évolution des particules moyennes est beaucoup plus régulière et le modèle reproduit ce type de variation. À l'horizon 36 heures, la prévision des concentrations de particules 4.5 µm est très bonne (avec un RMSE $< 3.34\#.\text{cm}^{-1}$), alors que pour les particules de taille 0.35 µm, le modèle n'arrive pas à reproduire la non-linéarité (variation brusque) sur cet horizon.

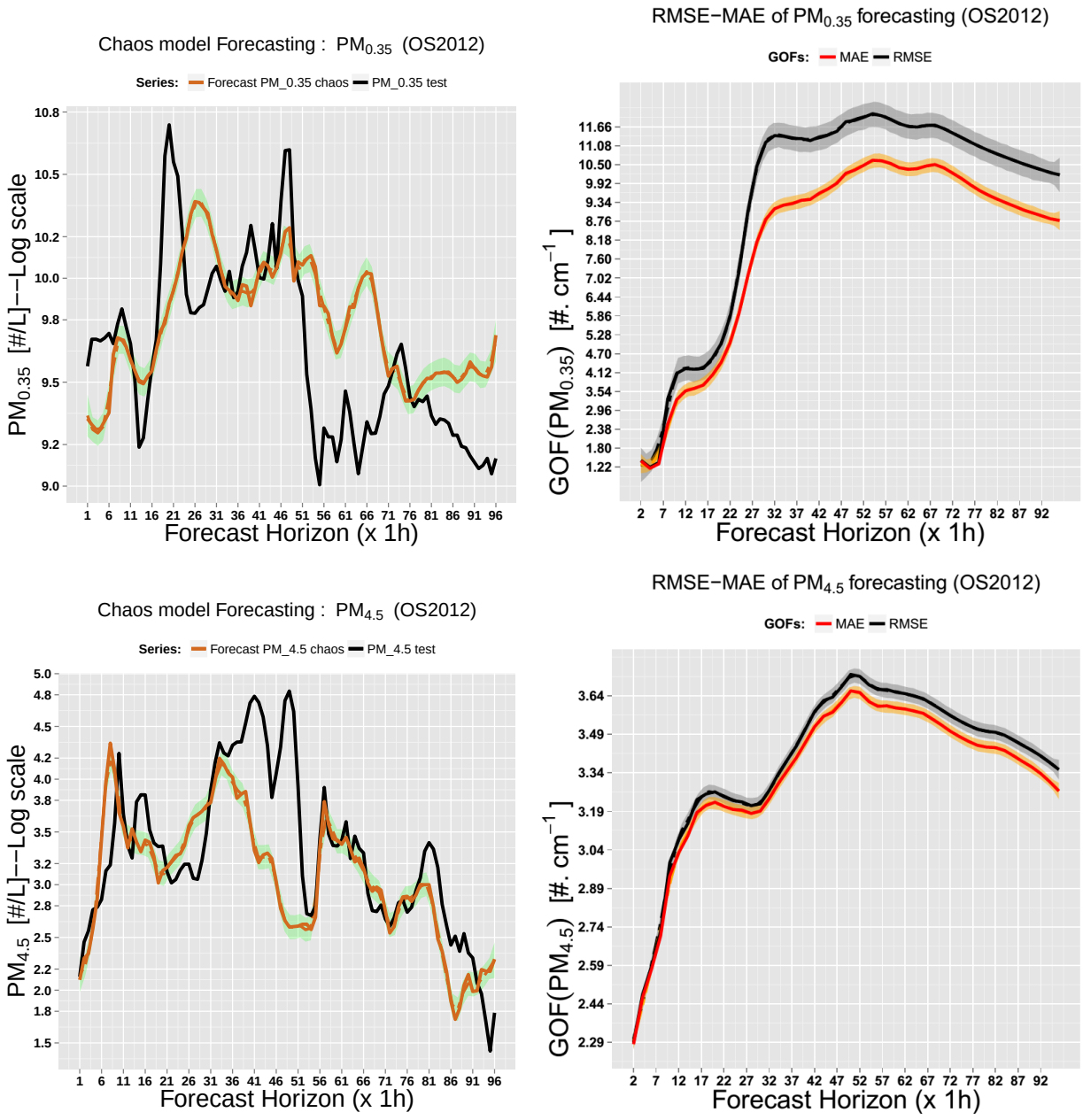


FIGURE 7.6.6 – *Prévision et performances (RMSE et MAE) par la méthode* ***LZO*** *des concentrations de* $PM_{0.35}$ *sur un horizon de quatre jours de* **test**. *Les paramètres utilisés pour cette méthode sont* $m = 168$, $\tau = 1$, $k = 2$ *et* $r = 1.05$. *Les mesures sont au pas de temps horaire et issues de la campagne de 2012 dans l'espace paysager. Les courbes de performance ont été ajustées avec une régression Loess pour fournir une grandeur sur l'étendu de l'intervalle de confiance (au seuil de 99%) de chaque indice.*

7.7 Modèles basés sur la décomposition en bandes spectrales

7.7.1 La procédure SBD-(SETAR/Chaos)

La Figure 7.7.1 montre les principales étapes du processus de prévision que nous avons adopté pour le modèle de décomposition en bandes spectrales couplé à deux types de modèles non-linéaires : SETAR ou Chaos. Le modèle de prévision proposé dans cette thèse se décline principalement sur trois étapes complémentaires :

Etape 1 : la décomposition en bandes spectrales (SBD).
Une décomposition en bandes spectrales est réalisée par transformée de Fourier rapide (FFT). Le signal est ensuite reconstruit après sélection d'une fréquence de coupure par FFT inverse. L'opération est ou non reconduite sur le signal reconstruit. La résolution finale du signal prétraité, c'est-à-dire le niveau de variations pris en compte, est contrôlé par la fréquence de coupure sélectionnée à chaque niveau de la décomposition et par le nombre de composantes FFT retenues. La somme de toutes les composantes FFT reconstruites correspond alors à la concentration finale filtrée des variations hautes fréquences. En résumé, durant cette première étape, pour m fréquences données, m composantes FFT peuvent être extraites.
Il est important de noter que cette étape est appliquée au jeu de données d'apprentissage et de validation pris ensemble de façon à pouvoir comparer les prévisions (avec des paramètres définis sur le seul jeu d'apprentissage) avec les données de validation brutes d'une part mais également avec les données lissées du jeu de validation.

Etape 2 : Apprentissage et validation
Lors de cette étape, un modèle autoregressif à seuil de type SETAR (Self-Exciting Threshold AutoRegressive model) ou une procédure de reconstruction de l'espace d'état (méthode de la dynamique du chaos) est appliquée à chacune des composantes FFT. Pour le modèle SETAR, les prévisions de chaque composante FFT est calculée par méthode de Monte-Carlo (MC) à partir d'un paramètre de retard p lié à l'autorégression et d'un délai d de la variable de transition. Pour la procédure chaos, la prévison de chaque composante FFT est déterminée à partir de l'approximation d'ordre zéro **LZO** (Hegger et al., 1999) en tenant compte de la dimension de plongement d_E et du délai temporel τ. Les prévisions sont ensuite additionnées et comparées à la concentration finale filtrée du jeu de validation et aux données brutes. La performance globale est calculée pour différents horizons de prévision h selon différents indicateurs d'ajustement (Goodness-Of-Fit). Cette opération est répétée pour différentes valeurs des paramètres du modèle SETAR et chaos jusqu'à l'optimisation de l'ajustement des prévisions aux données de validation. Selon ces critères, m jeux de paramètres optimisés sont obtenus correspondant aux m composantes FFT retenues. Le modèle global fournit alors une prévision finale qui est la somme de chacune des prévisions des m modèles retenus.

Etape 3 : Test
Le "meilleur" modèle global FFT-SETAR/Chaos est ensuite appliqué au jeu de données test en considérant un horizon de prévision égal à la longueur de la série test utilisée. Les prévisions optimisées précédemment sur le jeu de données de validation sont étendues au jeu de données test en utilisant les mêmes m paramètres retenus à l'étape 3. La prévision finale, qui résulte de la somme des prévisions de chacune des composantes à chaque pas de temps, est comparée aux données brutes du jeu de données test, ce qui permet d'évaluer la performance réelle du modèle

global. Le jeu de données test suit directement le jeu de données de validation. Ces données test ne sont bien entendu pas utilisées dans l'étape 1 ou l'étape 2.

L'utilisation de bandes spectrales étroites est recommandée pour chaque composante de façon à obtenir un modèle global plus flexible et plus robuste. Pour ces mêmes raisons, le nombre de composantes retenues est $m = 10$. Dans le cas du modèle SETAR, seuls des modèles à 2 régimes ont été considérés pour les raisons suivantes : *(i)* la variabilité du formaldéhyde révèle une distribution bimodale ; *(ii)* la stabilité des paramètres de régimes multiples nécessite plusieurs hypothèses statistiques difficiles à vérifier ; *(iii)* il n'existe pas de méthodologie de prévision bien établie pour des modèles à plus de deux régimes.

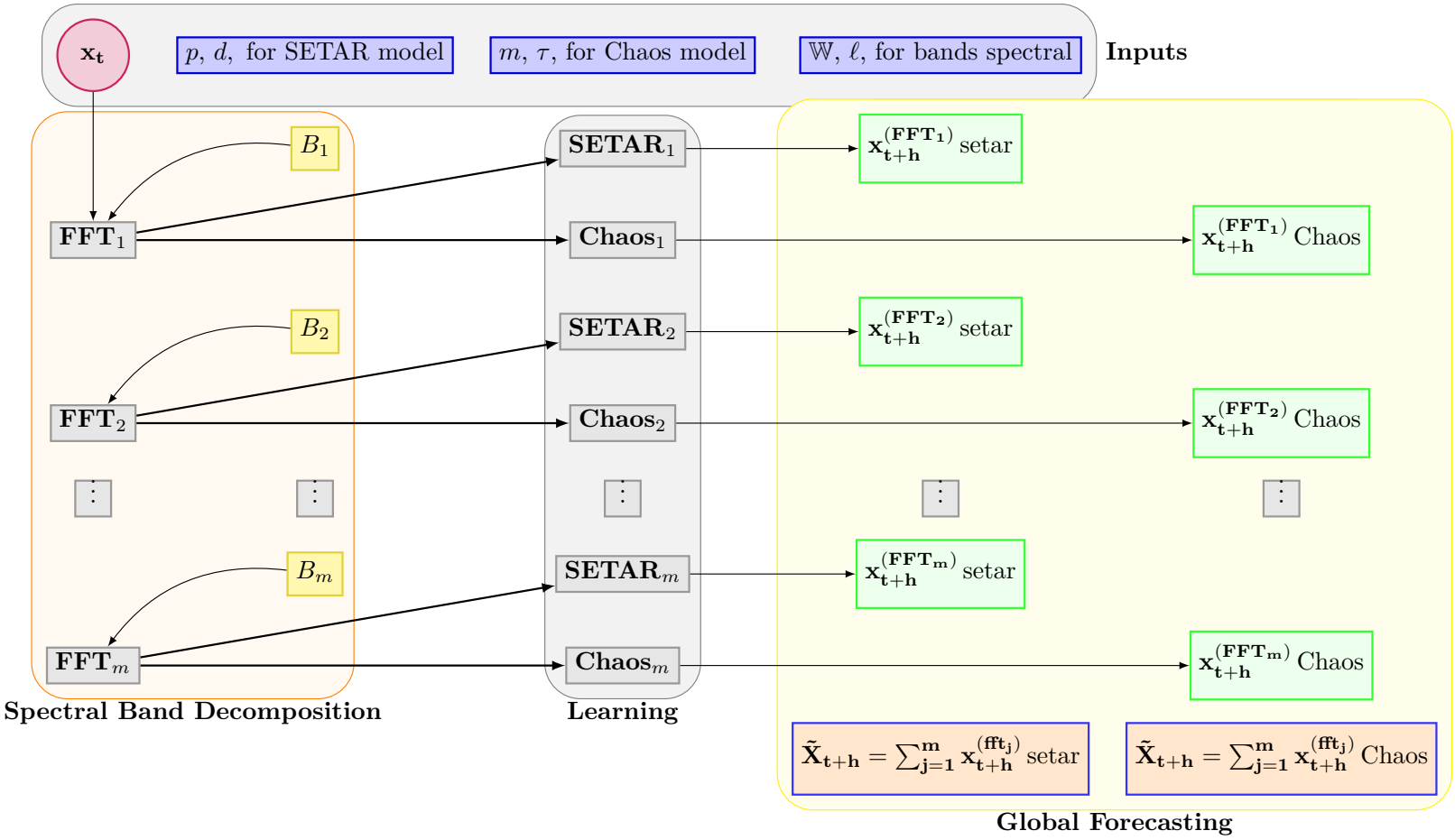


FIGURE 7.7.1 – *Principales étapes d'un processus de prévision par un modèle hybride FFT-TAR ou FFT-Chaos.*

7.7.2 Résultats de la prévision

Le modèle global est évalué au regard de deux objectifs différents : *(i)* l'influence de la décomposition du signal par de multiples FFT successives et *(ii)* la comparaison des performances de prévision entre FFT-SETAR et FFT-Chaos. Dans ce cadre, l'horizon de prévision h (h = 1 ; . . . ; 1000 min) est calculé selon 4 cas :

- un modèle SETAR est appliqué sur les données filtrées selon la procédure décrite ;
- un modèle SETAR est appliqué directement sur les données brutes ;
- un modèle chaos est appliqué sur les données filtrées selon la procédure décrite ;
- un modèle chaos est appliqué directement sur les données brutes.

La figure 7.7.2 montre l'évolution de la concentration pour les données de l'ensemble de test de HCHO et les prévisions par le modèle FFT-SETAR ou par le modèle FFT-Chaos.

D'importantes différences sont observées dans l'utilisation du modèle SETAR appliqué aux données filtrées par rapport aux données brutes. Le prétraitement par de multiples FFT améliore la capacité du modèle à reproduire les changements abrupts de concentration, phénomènes de variation non-linéaires. Les prévisions du modèle FFT-SETAR suivent la même tendance que les concentrations observées sur environ 12 heures (720 minutes) de la série de test. Par contre, le modèle SETAR appliqué directement sur les données brutes montre des variations beaucoup plus faibles.

Les mêmes remarques peuvent être formulées entre les modèles FFT-Chaos et chaos sur données brutes. Le modèle FFT-Chaos reproduit bien le changement abrupt de concentration et suit fidèlement la tendance d'évolution de la concentration sur une douzaine d'heures. Appliqué directement sur les données brutes, le modèle chaos présente plus de variabilité que le modèle SETAR mais sans reproduire les tendances et en sous-estimant le pic de variation.

La performance des modèles a été évaluée au moyen de différents critères d'ajustement et au regard des données brutes observées. Cette performance est discutée dans les différents points suivants.

7.7.2.1 Influence de la décomposition en bandes spectrales (SBD)

Déterminer cette influence revient à étudier le niveau de lissage des données brutes requis pour obtenir les meilleures prévisions. Le nombre de composantes FFT retenues et les fréquences de coupure utilisées vont définir ce point. Notre première approche pour améliorer les performances est d'utiliser un modèle non-linéaire pour chaque composante, soit SETAR ou chaos. Les modèles ont été comparés sur les mêmes composantes avec les mêmes fréquences de coupure. Leur performance a été évaluée à un horizon de 1000 min (1000 pas) par comparaison aux données tests.

Critères de performances des modèles SETAR et FFT-SETAR La Figure 7.7.3 montre les indicateurs de performances déterminés dans les deux cas (FFT-SETAR et SETAR seul). Les indicateurs utilisés sont l'erreur-type (racine carrée de l'erreur quadratique moyenne ou root mean square error RMSE), l'erreur moyenne absolue en pourcentage (MAPE), l'erreur moyenne absolue (MAE) et le coefficient de détermination R^2.

Tout d'abord, l'amplitude de variations de ces indicateurs est plus faible pour le modèle FFT-SETAR par rapport au SETAR seul. RMSE, MAE et MAPE augmentent avec l'horizon de prévision alors que

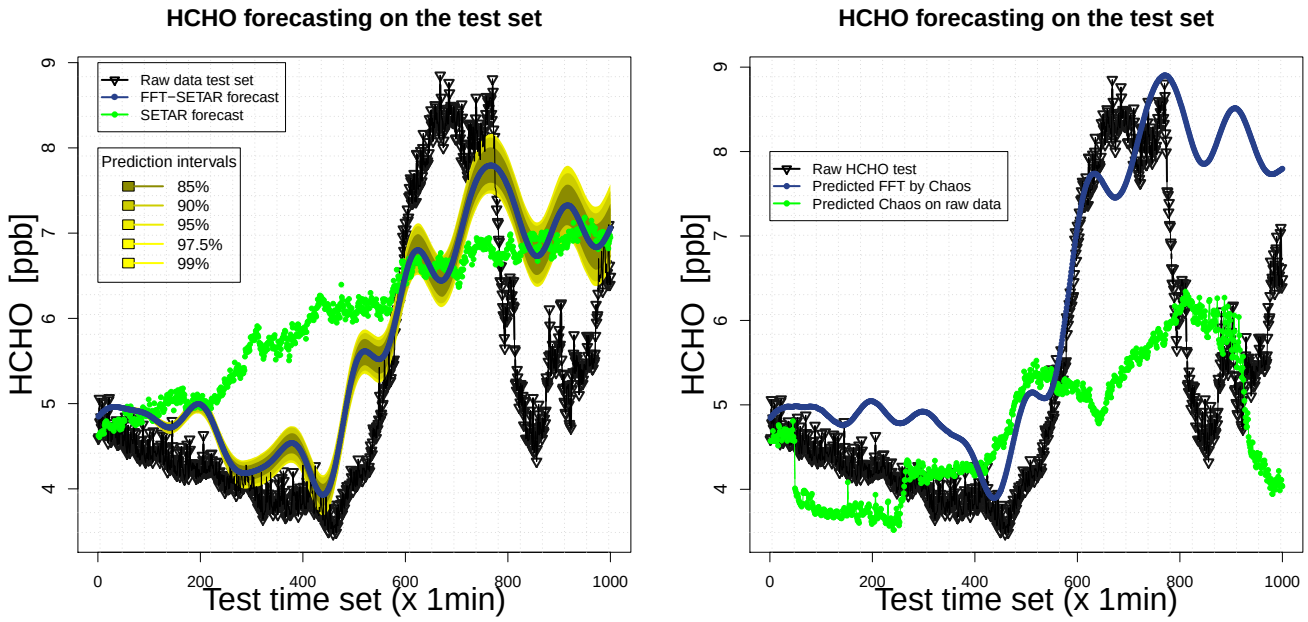


FIGURE 7.7.2 – *Prévisions de la concentration intérieure de formaldéhyde par la procédure SBD-(SETAR/Chaos). Les modèles FFT-SETAR (appliqué sur les composantes FFT) et SETAR (appliqué sur données brutes) sont représentés à gauche. Les modèles FFT-Chaos (appliqué sur les composantes FFT) et* ***LZO*** *du chaos (appliqué sur données brutes) sont représentés à droite.*

R^2 évolue de façon plus irrégulière, conséquence de l'intégration progressive de changements abrupts plus difficile à reproduire dans l'horizon de prévision qui font chuter R^2. Le RMSE ne dépasse pas 1 ppb et le MAPE est inférieur à 16.5 % pour le modèle FFT-SETAR. Pour le modèle SETAR seul, le RMSE est inférieur à 1.5 ppb et le MAPE est inférieur à 30 %. Le R^2 est optimal à 750 minutes pour le modèle FFT-SETAR et à 450 minutes pour le SETAR seul. Il est plus faible pour des horizons supérieurs ou inférieurs. Pour un horizon plus court jusqu'à 500 minutes, le RMSE reste inférieur à 0.5 ppb et le MAPE sous les 10 % pour le modèle FFT-SETAR et à respectivement 1.4 ppb et 30 % pour le modèle SETAR.

Critères de performances des modèles SETAR et FFT-Chaos La figure suivante montre les critères d'ajustement des modèles FFT-Chaos et chaos sur données brutes. Pour le modèle FFT-Chaos, chaque composante est modélisée selon la reconstruction de son espace d'état pour une dimension de plongement donnée. La prévision est ensuite déterminée sur un horizon de 1000 pas (minutes). Sur les 600 premières minutes, il n'y a pas de différences marquées entre les deux modèles avec un RMSE inférieur à 0,6 ppb et un MAPE inférieur à 15 %. Passé ce temps, les performances se dégradent très vites pour le modèle chaos alors qu'elles restent acceptables (voire optimales selon le R^2 à 0,9) jusqu'à un horizon de 800 minutes (plus de 13 heures) pour le modèle FFT-Chaos. Ce dernier montre ainsi sa capacité à reproduire des variabilités faibles et fortes.

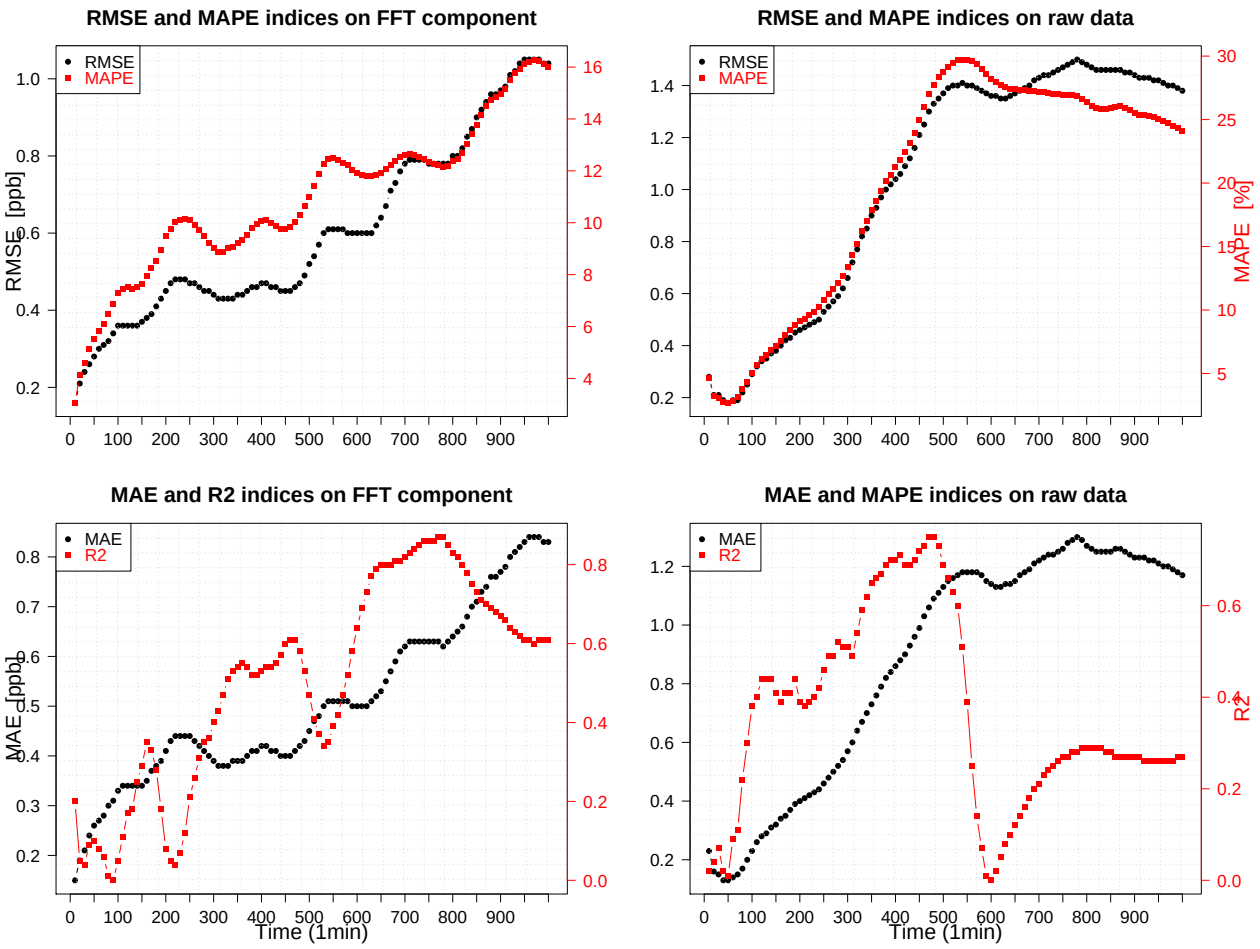


FIGURE 7.7.3 – *Critères d'ajustement des modèles FFT-SETAR (à gauche) et SETAR (à droite). RMSE et MAPE sont représentés sur les figures du haut et MAE et R^2 sur les figures du bas.*

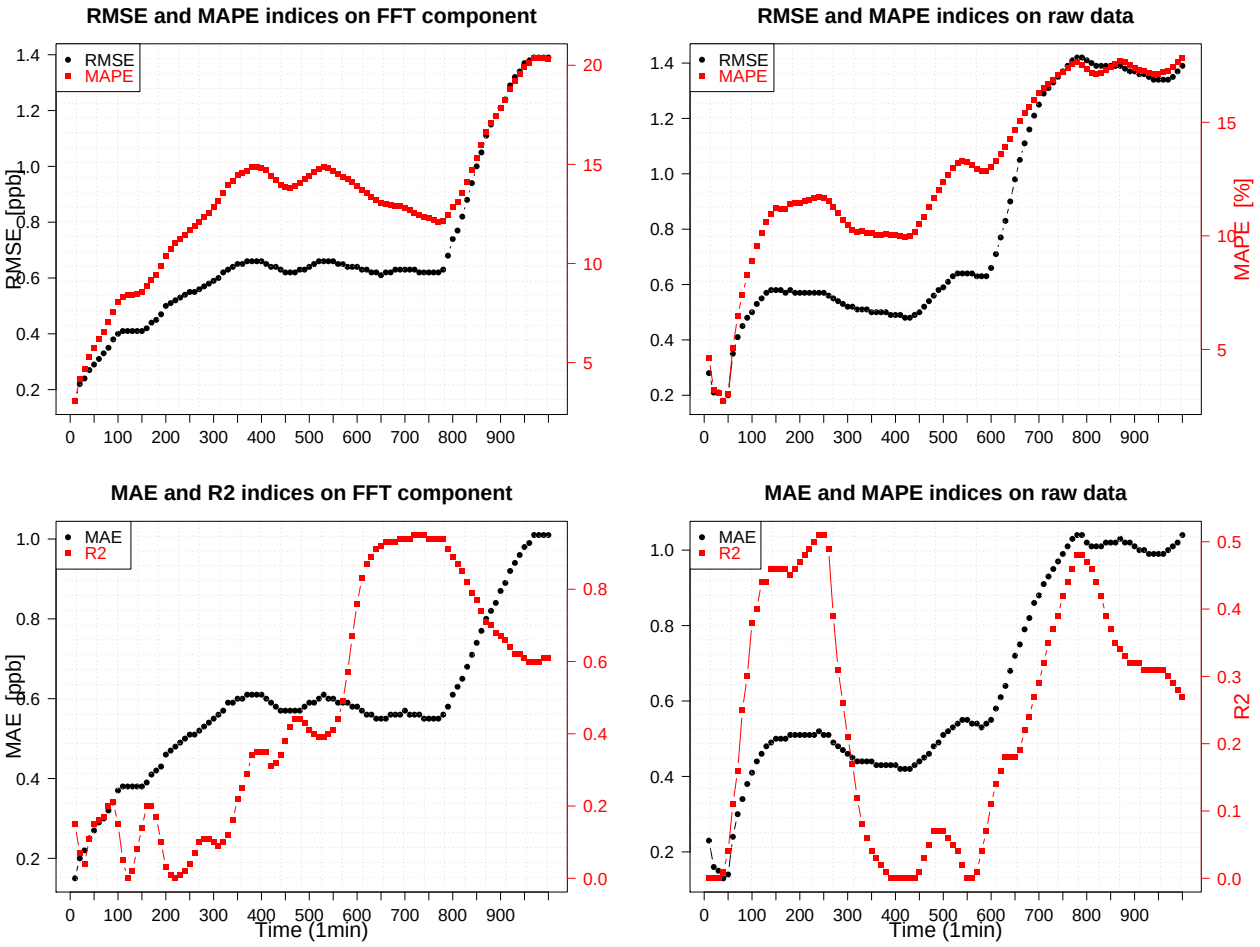


FIGURE 7.7.4 – *Critères d'ajustement des modèles FFT/chaos (à gauche) et chaos (à droite). RMSE et MAPE sont représentés sur les figures du haut et MAE et R^2 sur les figures du bas.*

7.7.2.2 Comparaison des modèles FFT/SETAR et FFT/chaos

Comparaison des modèles FFT/SETAR et FFT/chaosDans notre cas de figure, le modèle chaos est meilleur que le modèle SETAR dans la prise en compte du changement abrupt avec ou sans prétraitement. Avant ce changement, les performances des deux modèles apparaissent similaires. Cela reflète le fait que les composantes FFT qui représentent la partie déterministe de la variabilité de la concentration de formaldéhyde sont plus facilement modélisées par un modèle chaos qu'un modèle paramétrique stochastique de type SETAR.

Par contre, les performances à long terme (1000 minutes) mettent en avant le modèle FFT/SETAR devant FFT/chaos bien que ce dernier soit meilleur sur un horizon plus court (800 min). La différence entre les deux modèles au niveau du RMSE maximum observé,1,1 ppb pour FFT/SETAR et 1,4 ppb pour FFT/chaos, peut ne pas être considéré comme significative.

Sans de prétraitement particulier, le modèle chaos fournit des prévisions plus réalistes que le modèle SETAR. De fait, le modèle SETAR estime la moyenne générale des données alors que les prévisions par le modèle chaos reflètent plus le schéma de variation des données.

7.7.3 Conclusion et discussion

La décomposition spectrale permet d'améliorer les prévisions comparées à l'utilisation des données brutes. Elle permet d'atténuer l'effet des perturbation aléatoires nuisible au modèle mais écarte également la possibilité de reproduire des fluctuations très courtes de la concentration selon le niveau de lissage effectué. Elle permet également d'améliorer le temps de calcul et du coup fournit de meilleures performances, c'est ce que nous avons appélée l'efficience prévisionnelle : capacité du modèle à fournir des prévision utiles avant l'arrivée de l'échéance de prévision effective.

En règle générale, la performance du modèle hybride est un compromis entre le niveau de variabilié accordé aux données réalisé par la décomposition spectrale et la performance du modèle de prévision sur les données lissées. En effet, le prétraitement des données joue sur la qualité des prévisions, mais surtout elle permet de s'interroger sur l'importance que l'on veut donner aux variations observées et quelles variabilités sont utiles à la prévision. Les données prises en compte avec un pas de temps de 1 min présentent un bruit en partie liée aux fluctuations du signal de l'instrument et donc à l'incertitude de mesure qu'il est inutile de chercher à reproduire.

Dans notre cas, nous avons pris en compte des variations de période supérieures à 4.5 heures. Il serait certainement souhaitable de descendre à l'échelle de 1 heure, pour des prévisions de court terme. Les deux modèles utilisés reproduisent l'aspect non-linéaire de l'évolution de la concentration du formaldéhyde avec des performances légèrement supérieures pour le modèle basé sur la dynamique du chaos. Ces modèles hybrides permettent de prévoir la concentration du formaldéhyde avec de bonnes performances à un horizon de 12 heures.

En général, la prévision de la première composante FFT1, celle qui extrait la tendance générale de variation de la concentration détermine en grande partie la qualité de la prévision finale. Lorsqu'elle n'est pas performante, les prévisions sur les autres composantes FFT ne permettent pas d'améliorer la prévision finale. La décomposition spectrale n'est pas toujours nécessaire si les paramètres étudiés présentent une saisonnalité marquée. C'est le cas des particules les plus grosses, pour lesquelles une décomposition spectrale par FFT introduit des oscillations parasites lors des périodes monotones comme les week-ends. A titre de comparaison, la Figure 7.7.5 montre les résultats de la prévision sur des concentrations de particules mesurées dans l'espace de bureaux en 2012 par chaos.

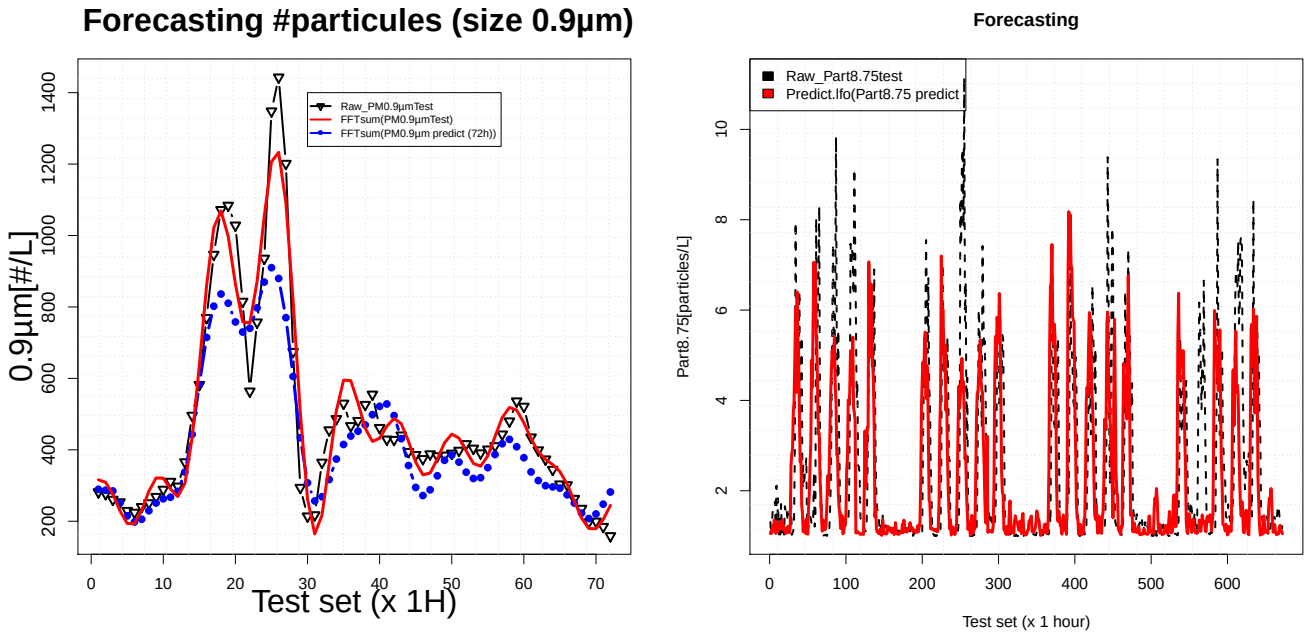


FIGURE 7.7.5 – *Prévisions obtenues sur les données tests par FFT-Chaos pour les particules de diamètre inférieur à 0,9 µm (série de validation) et directement par chaos pour les particules de diamètre inférieur à 8,75 µm (série de test).*

Les particules fines de 0.9 µm nécessitent une décomposition spectrale préalable de leur concentration alors qu'un résultat correct est observé pour les particules de 8.75 µm en appliquant directement le modèle sur les données brutes. Plus de détails sur ces modèles hybrides sont fournies dans une première version d'article qui a été soumis à la revue *international journal of forecasting*. Les travaux préalables à cet article notamment sur le développement du modèle FFT-SETAR ont fait l'objet de deux communications dans des conférences internationales (Ouaret et al., 2014d,b).

7.8 Discussion

En vertu de l'ensemble des résultats obtenus pour des différents modèles de prévision et pour les différentes séries temporelles de polluants, on peut dire :

1. qu'il n'existe pas de modèle immuable aux différents types de fluctuation des séries temporelles de polluants intérieurs ;
   (a) chaque série temporelle a sa propre structure de variabilité qui nécessite d'être explorée et la rendre utile en prévision ;
   (b) le modèle doit se plier au différentes caractéristiques de variabilité pour les différents types de polluants
2. tous les modèles aboutissent, à une paramétrisation près, à fournir des prévisions acceptables pour certaines classes de polluants, en particulier pour les concentrations de CO_2 et les particules moyennes ;

3. la dynamique des fluctuations de formaldéhyde et des particules fines est plus difficile à prédire.

Remarque 7.8.1. **le choix des paramètres, la méthode de prévision et le temps de calcul**

Nous avons remarqué l'importance du choix des paramètres pour le calcul des prévisions associé à chaque méthode. Différentes simulations de combinaisons des paramètres associées à la prévision des deux méthodes **LZO** et **LFO** et leur temps de calcul ont été réalisées. Globalement, la méthode par une moyenne au voisinage des proches voisins (**LZO**) est plus rapide que la méthode associée à la résolution d'un système d'équation linéaire (**LFO**). Cette différence est probablement due au fait que la méthode **LFO** nécessite une dimension de plongement grande afin d'éviter l'indétermination du système linéaire (matrice singulière) qui augmente le temps de calcul. Notons que la part de la complexité algorithmique la plus importante associée aux prévisions provient de la recherche des plus proches voisins. Ainsi pour trouver les proches voisins dans un espace $m-$dimensionnel d'une série temporelle de taille T, il faut calculer $T^2/2$ distances distinctes. Par conséquent, l'utilisation d'une norme particulière modifie considérablement les aspects calculatoires, nous renvoyons le lecteur intéressé aux pages 322-325 de la monographie de Kantz & Schreiber (2004) et les références citées dans ce livre pour plus de détails techniques.

CONCLUSION GÉNÉRALE ET PERSPECTIVES

"In the 1960s and 1970s, students frequently asked : Which kind of representation is best ? I usually replied that we'd need more research...But now I would reply : To solve really hard problems, we'll have to use several different representations. This is because each particular kind of data structure has its own virtues and deficiencies, and none by itself would seem adequate for all the different functions involved with what we call common sense." Marvin Minsky (1927-2016).

Les objectifs et outils de cette thèse

La qualité de l'air dans les environnements intérieurs réels est sujette à des fluctuations permanentes dans un système dynamique complexe. Cette complexité est due aux fluctuations des paramètres climatiques, aux conditions d'échanges entre l'air intérieur et l'air extérieur, mais surtout aux fluctuations aléatoires imposées par l'occupant et ses activités. L'occupant agit de manière plus ou moins aléatoire sur les différentes composantes du bâti, mais aussi réagit presque de la même manière à la variabilité de l'ambiance, altérant ainsi l'évolution régulière de la concentration en polluants dans l'air intérieur.

Ces variations se manifestent par des composantes non-linéaires et aléatoires qui se " greffent " sur l'évolution naturelle d'un système quasi-déterministe, comme un environnement réel non occupé. Dans cette thèse, il a été question d'étudier ces deux cas : deux environnements de type bureau (individuel et espace paysager) normalement occupés et une maison expérimentale (MARIA) occupée de manière ponctuelle.

L'analyse des structures de variabilité nécessite de faire appel à des outils statistiques très variés, menant ainsi à une recherche multidisciplinaire : statistique, traitement du signal, et en particulier la séparation aveugle des sources et l'analyse des séries temporelles. À notre connaissance, aucun des outils venant de ces domaines n'a été encore appliqué pour la qualité de l'air intérieur (QAI). Ce travail de recherche a pour but de tester leur utilité afin de mieux comprendre quels sont les déterminants des fluctuations des concentrations de polluants en air intérieur et prévoir leurs valeurs.

La recherche menée dans cette thèse avait un double objectif. Le premier objectif concerne la détermination de quelques facteurs permettant de comprendre et analyser les structures de variabilité temporelle des concentrations de polluants dans l'air intérieur, la variabilité de leurs sources, ainsi que les contributions de celles-ci. Le deuxième volet de cette thèse avait pour but de développer des modèles de prévision pour la QAI. Finalement, le problème soulevé dans le volet prévisionnel a été raffiné pour répondre à la question : quel modèle statistique de prévision pour quel type de polluant ?

Compte tenu de l'absence d'une littérature sur les applications des processus stochastiques dans le domaine de la QAI, et la non spécification d'une classe de modèles à utiliser dès le départ, cette thèse s'est positionnée dans le cadre d'une double exploration : identifier les principales classes de modèles susceptibles d'atteindre les objectifs tracés, et spécifier en quoi ces modèles sont utiles.

Conclusion

Dans cette thèse nous nous sommes intéressés à la modélisation de la qualité de l'air intérieur d'un point de vue statistique. Plus précisément, nous nous sommes focalisés sur les méthodes d'identification des sources et surtout sur les modèles de prévision stochastiques.

Pour le premier objectif, concernant l'identification des sources, plusieurs méthodes de séparation aveugle des sources ont été testées pour comparaison. Cette comparaison est effectuée, en outre, entre les résultats obtenus dans les différents environnements étudiés.

Pour l'objectif de prévision, en utilisant les méthodes d'analyse spectrale et les modèles des séries temporelles, nous avons proposé un nouveau modèle de décomposition de séries temporelles dans le cadre des données de haute fréquence. Ces modèles sont basés sur une Décomposition en Bandes Spectrale (SBD) de la série, combinée avec les modèles à changement de régime et les modèles issus de la théorie des systèmes dynamiques appliqués aux différentes composantes. Notre proposition s'est positionnée donc, à la fois par rapport aux méthodes de référence comme la Transformation de Fourier et par rapport aux méthodes plus récentes, comme les modèles non-paramétriques de type SSA (Singular Spectrum Analysis). Elles répondent en partie aux problèmes auxquels sont confrontés les modèles de prévision pour les données à pas de temps fin (dans notre cas, entre 1 minute et 1 heure).

Les structures de variabilité et conséquences de leurs traitements statistiques

La première partie de ces travaux de recherche était consacrée principalement à l'analyse des structures inhérentes aux fluctuations des séries temporelles des polluants de l'air intérieur. En effet, les modèles de prévision dépendent en particulier de la nature et des propriétés statistiques des séries temporelles traitées. La description des séries a été nécessaire.

Le premier chapitre (1) était une revue du contexte général lié à la problématique de la QAI, consistant, dans une perspective globale, à estimer l'exposition des populations aux différents polluants de l'air intérieur. Les questions sur la nécessité d'identifier les sources de fluctuations et la prévision des concentrations des polluants ont été abordées uniquement d'un point de vue statistique. Nous avons montré et justifié cette approche compte tenu de la difficulté des modèles physico-chimiques à répondre à ces objectifs dans un environnement réel inoccupé.

Le point le plus important qui ressort du deuxième chapitre (2) est celui de la quantité de l'information (désormais) disponible pour le suivi en continu de la qualité de l'air intérieur. Ce chapitre montre l'importance de l'instrumentation active mise en œuvre et les technologies associées, qui sont parfois

très délicates à maîtriser. Face à ces diverses technologies d'acquisition des données, nous avons mis en œuvre des programmes informatiques pour :

1. la synchronisation sur le même pas de temps des différents jeux de données (issues des différents capteurs) ;
2. la validation et la mise à jour des nouvelles données ;
3. le prétraitement statistique ;
4. la visualisation synthétique des résultats.

Cette étape de prétraitement des données était très coûteuse en temps, mais indispensable pour entamer toute réflexion sur un modèle statistique de prévision, voire même d'un prétraitement de type analyse spectrale ou mesure de prédictibilité. Tout au long de ce deuxième chapitre, bien que très schématique et descriptif, nous avons montré l'influence de quelques facteurs sur les concentrations de quelques types de polluants, en insistant sur quelques variables de contrôle comme l'occupation et l'ouverture des fenêtres.

En effet, l'impact de l'occupation sur la variabilité des concentrations du formaldéhyde était plus important en 2013 qu'en 2015, et ceci principalement au cours de la journée. Cette influence dépend des pratiques d'ouverture des fenêtres, celles-ci varient selon les mois et d'une année à une autre. L'ouverture des fenêtres permet de diminuer (de manière brusque) la concentration intérieure en formaldéhyde : de 20% pour une fenêtre ouverte jusqu'à 40% pour plus de 3 fenêtres ouvertes.

Ces analyses ont été faites avec la présence (inéluctable) des valeurs manquantes. L'annexe F veut montrer l'importance des méthodes d'imputation appliquées dans un cadre cohérent aux spécificités des données issues de la QAI. Cette partie a considéré le problème d'imputation de données manquantes comme condition nécessaire pour entamer l'analyse des séries temporelles. Nous avons montré que ces méthodes requièrent un contrôle à la sortie de l'algorithme afin de rendre les valeurs imputées plus cohérentes par rapport aux observations réelles.

Le dernier chapitre (3) de la première partie visait à identifier les principales caractéristiques des séries temporelles des mesures de polluants liés à la QAI par :

1. les structures de variabilité des concentrations des polluants étudiées par l'analyse spectrale classique ;
2. les structures de dépendance par la quantification des propriétés de type mémoire longue ;
3. la décomposition des séries temporelles en différentes composantes latentes.

Pour le premier point, ce type d'analyse nous a permis de mettre en évidence le type de décroissance de la densité spectrale. La plupart des fluctuations partagent une caractéristique commune : les spectres exhibent des pôles à la fréquence zéro, et ce, quelle que soit la résolution temporelle ou l'étendue de la série. Nous avons étudié en outre, la notion de prédictibilité ($\Omega_g - \textit{prédictibilité}$) selon Goerg (2013) de certains types de fluctuations. Sur cette dernière, nous avons remarqué une dépendance de la mesure Ω_g par rapport à la taille des séries, induisant une erreur systématique dans l'estimation et qui empêchait la comparaison. Pour palier à ce problème, une normalisation a été proposée afin de réduire cette erreur. Avec cette nouvelle mesure, Ω_g^*, on constate que les résultats sont cohérents à ce qu'on pourrait attendre de la mesure de la prédictibilité.

La dernière section du chapitre 3 s'est inscrite directement dans la continuité du travail engagé pour l'analyse des structures de variabilité. Son objectif a été de mettre en évidence les propriétés inhérentes des séries chronologiques par différentes méthodes de décomposition. Nous avons montré que pour certains polluants, comme le formaldéhyde dans l'espace paysager en 2013 (ou les particules fines pour

toutes les campagnes), la décomposition en composantes latentes était plus difficile que pour les fluctuations de 2015. Cette difficulté s'explique par le fait que la composante aléatoire est plus prononcée au sein de ce type de polluants que les composantes déterministes.

La méthode SSA (Single Spectrum Analysis) a été appliquée afin de mettre en évidence les caractéristiques qualitatives véhiculées par le processus générateur des données. Cette méthode, comme la méthode STL (Seasonal Trend Decomposition using Loess) avait une double fonction : la description par la décomposition et l'utilisation des composantes séparées pour la prévision. La méthode SSA, par la structure des vecteurs propres, offre une description qualitative riche et s'avère très utile pour la prévision, mais je pense que qu'une amélioration pourrait être apportée en introduisant un délai τ dans l'étape de plongement (la méthode SSA suppose que $\tau = 1$).

L'identification et contributions des sources

Le problème de l'identification des sources et de leurs contributions a été abordé dans cette thèse dans le cadre des méthodes inverses. Ces méthodes permettent de remonter à des "causes" ou à des grandeurs d'influence inconnues à partir de l'observation de leurs conséquences. Pour la problématique de la qualité de l'air intérieur, il s'agit de sources d'émission comme causes et les concentrations de polluants comme conséquences des transformations des polluants émis par les sources et qui se mélangent dans l'air avant d'être mesurés au niveau des récepteurs. Notre approche est de type récepteur.

Pour s'atteler à cette tâche, certaines approches sont basées sur des connaissances a priori "complètes", comme les modèles de bilan massique (Chemical Mass Balance- CMB) ou les modèles physico-chimiques. De ce point de vue, les informations disponibles pour la qualité de l'air intérieur sont trop limitées pour aborder ainsi le problème de séparation des sources dans la majorité des cas réels d'étude.

Pour s'affranchir de cette contrainte, une nouvelle famille d'approches consiste à introduire le moins d*'a priori* possible sur les données de départ, et à tenter de séparer les sources-facteurs "à l'aveugle ". Parmi celles-ci, les techniques classiques d'analyse statistiques multidimensionnelle (L'analyse en Composantes Principales ACP, l'Analyse Factorielle des Correspondances AFC, ...), l'Analyse en Composantes Indépendantes (ACI) et les techniques de factorisation en matrices non-négatives (PMF pour Positive Matrix Factorization, NMF pour Nonnegative Matrix Factorization).

Ces deux dernières classes de méthodes et leurs variantes font généralement deux types d'hypothèses :

- contrainte statistique d'indépendance des sources.
- contrainte de signe de ("non-négativité", voire positivité) sur les profils de sources et leurs contributions aux mélanges.

Nous avons examiné plus en détail le problème standard séparation aveugle des sources (SAS) et par l'approche la **NMF** sous l'angle de l'état de l'art et notamment de ses propriétés théoriques.

Bien qu'elles aient montré des résultats prometteurs dans d'autres domaines, notamment en traitement du signal, l'application de ces méthodes pour la qualité de l'air intérieur demeure inexistante. On peut trouver des applications dans la qualité de l'air extérieur uniquement. Dans le chapitre 4, nous avons montré l'intérêt de ces méthodes afin de répondre aux questions d'identifications des sources. Ce chapitre constitue une première tentative d'application de ces méthodes aux données de la QAI et pour l'instant il ouvre la voie à beaucoup de questions et un peu moins de réponses.

Plus particulièrement, dans cette partie nous avons examiné l'évaluation de ces méthodes pour différents jeux de données. Ces méthodes ont été principalement utilisées sur les concentrations de particules mesurées dans l'espace de bureaux, mais également dans d'autres environnements comme la maison

expérimentale ou le bureau individuel situés sur le même site. Les composantes ou facteurs identifiées représentent autant de sources de fluctuations séparées. Les composantes peuvent être interprétées en termes de contributions variables selon le diamètre des particules, mais également en termes de fluctuations ou d'occurrence au fil du temps.

L'importance d'une source intérieure de HCHO (ou groupement de sources) a été clairement séparée par rapport aux sources communes avec l'extérieur par l'application de la NMF. Sa contribution a été de l'ordre de 40%.

Pour autant, séparer n'est pas identifier. Sans informations extérieures, les facteurs extraits restent difficilement identifiables autrement que par l'observation de fluctuations caractéristiques d'un processus particulier. Pour apporter plus de renseignements sur les sources, nous avons étudié les profils temporels par une analyse de la variabilité polaire associée aux différents paramètres d'occupation et d'ouverture. Clairement, nous avons bien identifié la source "occupation" avec ce type d'analyse.

Il s'avère que la méthode NMF et PMF offrent un confort d'interprétation assez notable par rapport à l'ACI et l'ACP.

Le potentiel de ces méthodes et l'information contenue dans les différentes bases de données seront encore exploités, car ils n'ont pas été explorées de manière assez exhaustive.

Modèles de prévision

Le chapitre 5 de cette thèse passe d'abord en revue les fondamentaux de la prévision linéaire, indépendamment de la nature de la variable. Ensuite, un état de l'art sur les différentes approches de prévision appliquée à l'environnement a été entrepris. À l'issue de cette revue de littérature, nous avons remarqué une grande différence entre la nature des données issues de l'air extérieur et celles issues de l'air intérieur, notamment le pas temps utilisé et la régularité des séries chronologiques. Par ailleurs, à notre connaissance, très peu d'études de modélisation à des fins prévisionnelles pour la QAI ont été entreprises dans la littérature. Cette (mal-)heureuse contrainte nous oblige à regarder, tant au niveau théorique que pratique, le développement des autres domaines afin de transposer leurs expériences (parfois plus de 50 ans d'écart) aux exigences de nos données.

Le chapitre 6 est une conséquence directe des conclusions du chapitre 5, d'une "rareté" de travaux de recherche sur la prévision de la QAI, même avec des modèles très simples. Trois classes de modèles ont été appliquées pour fournir les conditions nécessaires de prévision. Un modèle de type lissage exponentiel a été appliqué avec un principe de parcimonie : optimisation des paramètres du modèle de prévision pour le CO_2 dans le bureau individuel et sa généralisation pour la prévision des concentrations des autres polluants (formaldéhyde et particules) issus de l'espace paysager.

Ensuite, le modèle STL-ARIMA à erreur GARCH a été estimé et utilisé pour la prévision :

- (i) des concentrations de formaldéhyde dans la maison expérimentale (pas de temps 1-min) et dans l'espace paysager de deux campagnes de 2013 (pas de temps 1-min) et de 2015 (pas de temps 20-min) ;
- (ii) des concentration de particules en nombre dans l'espace paysager durant la campagne de 2012 (pas de temps horaire).

La qualité de la prévision par décomposition SSA offre des perspectives très intéressantes. Pour la série de HCHO de la campagne de 2015, la prévision par la méthode SSA reste très précise à très court terme, mais diverge rapidement après 1.5 jours. Ainsi, le modèle arrive à prévoir environ 33 h avec une erreur inférieure à 0.5 ppb et les erreurs restent en deçà de 1 ppb sur les deux premiers jours.

Finalement, on s'est posé la question sur la possibilité de construire une méthode "propre" à l'environnemetrie intérieure. La chapitre 7 s'est positionné alors dans cette perspective en développant une classe de modèles basée sur l'hybridation entre un processus de décomposition en bandes spectrales avec les modèles non-linéaires de séries temporelles .

Les modèles de prévision non-linéaires abordés sont : les processus aléatoires à changement de régime (**TAR** et **MS-TAR**) et les modèles basés sur les systèmes dynamiques du chaos. Ces modèles ont été utilisés uniquement pour aborder les questions de prévision. Un modèle à changement de régime de type Markovien (transition probabiliste) a été l'objet d'une étude afin de mettre en évidence l'influence de quelques paramètres (température, humidité relative, ...) sur la variabilité du HCHO.

Les principaux résultats qui ressortent de cette étude peuvent être classés en deux catégories *:* premièrement, ceux relatifs à une étude plus générale sur la décomposition en bandes spectrale avec les modèles à changement de régime, et deuxièmement, les résultats relatifs à la prévision par la dynamique du chaos.

En ce qui concerne la méthode basée sur la décomposition en bandes spectrales, nous en examinons les conditions de "succès" de cette procédure.

Modèles FFT-TAR vs FFT-Chaos

Avec la procédure de décomposition SBD nous avons deux résultats principaux :

1. le pouvoir prédictif "global" des modèles par l'approche des systèmes dynamiques dépasse sensiblement les modèles à changement de régime ;
2. à très court terme, la qualité de la prévision ne dépend que très légèrement des paramètres du modèle et les modèles non-linéaires sont équivalents en prévision.

Modèles de Chaos sans prétraitement pour les séries régulières

Pour les modèles de chaos, de très bons résultats ont été obtenus pour certains classes de polluants. Ces modèles dépendent énormément des paramètres du modèle : la dimension de plongement m et le retard τ. Nous avons, à ce sujet, proposé une "recommandation" pour le choix ces deux paramètres de manière à ce que leur produit soit égal à la période principale : $m \times \tau = f_p^{-1}$.

Modèles de Chaos avec prétraitement SBD

La décomposition en bandes spectrales avec modélisation par la dynamique du chaos de chaque composante améliore considérablement la qualité de prévision. Ce résultat s'explique par le fait qu'au niveau des échelles très fines, des structures chaotiques se manifestent par des attracteurs étranges.

Quelques perspectives

Il reste un grand nombre de problèmes non résolus et des pistes non-exploitées qui laissent ouvertes des perspectives passionnantes de recherche et des améliorations potentielles sur les outils d'identification des sources, de leurs contributions et, surtout sur nos outils de prévision. En effet, si cette thèse espère apporter des améliorations, même préliminaires, sur notre compréhension de la variabilité de la QAI,

elle est loin de proposer une solution suffisante sur l'ensemble des problèmes liés à l'environnemétrie intérieure.

Nous évoquerons à présent quelques perspectives aux travaux menés pendant cette thèse.

La question des données

Le dispositif de mesures, la collecte et le prétraitement des données sont des étapes primordiales dans la construction de n'importe quel modèle statistique. Néanmoins, l'acquisition en temps réel des données de mesure sur une longue période est sujette à de nombreuses contraintes techniques, qui nécessitent un traitement statistique approprié ; ajoutons à ces considérations le pas de temps utilisé.

Sur la quantité d'informations et la complexité des données

L'instrumentation active de l'espace paysager pendant les trois ans de mesure a généré un flux d'informations très considérable, dépassant ainsi 5 millions (voire beaucoup plus) de points. Ces informations continuent à affluer les bases de données existantes et deviennent désormais très volumineuses. La question du pas de temps joue un rôle très important, tant au niveau théorique que pratique, alors "jusqu'où aller dans l'échelle temporelle" ?

Cette question peut ne pas sembler être de première importance. Pourtant, elle reste, à nos yeux très intrigante, car elle relève du choix de la classe des modèles à utiliser dans l'étape de prévision. Donc, le pas de temps et la "masse" des données disponibles soulèvent une question de méthodologie statistique. Ainsi, faut-il juste emprunter les méthodes déjà appliquées dans les autres domaines et les transposer sur les bases de données actuelles, ou bien construire en fonction des caractéristiques des données liées à la QAI, des modèles spécifiques ?

Bien que cette question fût soulevée par Granger (1998) dans le cadre de l'estimation statistique des données massives, très peu d'études ont souligné la prise en compte de la taille des données, et surtout le pas de temps. Il est intéressant de savoir si lors de l'analyse dans ce type de données, on doit prendre en compte la significativité statistique des coefficients estimés au niveau de 1%, au lieu du niveau de 5% (choix conventionnel). En effet, la communauté de statisticiens reconnaît (depuis un certain temps déjà) que lorsqu'on travaille sur des données de très longues durée, il devient "trop facile", voire systématique de rejeter l'hypothèse nulle (de la significativité des paramètres), car les intervalles de confiance sont d'une complexité de type $\mathcal{O}\left(T^{-1}\right)$ (Granger, 1998). Une possibilité pour résoudre ce problème consiste à laisser la valeur critique en fonction de la taille de l'échantillon.

Ces considérations, déjà de complexité assez importante, sont amplifiées par l'arrivée des nouvelles technologies de capteurs et la multiplication des objets connectés au sein du bâtiment : l'environnement numérique. Elles fournissent aussi des informations qui probablement vont être stockées (quelque part) et croisées avec d'autres jeux de données.

Un axe de recherche peut être identifié à partir de tous ces éléments, notamment celui des classes de problèmes où le volume et la complexité des données sont grands, qui sont issues ou à destination du calcul scientifique pour l'ingénierie de l'environnement intérieur. Cette thématique constituerait dans les années à venir un verrou majeur.

Je pense que même si les modèles de prévision pour la QAI sont loin (dans leurs aspects de conceptualisation) des modèles développés dans les autres domaines (cette thèse est une tentative de réduire cet écart), les problématiques liées au volume des données et à leur fréquence sont au même niveau

de difficulté dans tous les domaines, dont même le domaine de l'environnement intérieur. Donc aucun retard, en termes d'analyse de l'information, ne serait justifié à l'avenir, les données sont disponibles.

Sur la gestion des données manquantes

Nous avons partiellement traité le problème des données manquantes dans le cadre de l'analyse des séries temporelles (en annexe). À ce sujet, plusieurs perspectives nécessitent d'être éclairées, en particulier par rapport aux éléments soulevés dans le paragraphe précédent :

1. Le calcul d'incertitude liée à la méthode d'imputation sur les différents types de données (fortes autocorrélations, distributions polymodales...) ;
2. L'impact de cette incertitude sur l'intervalle de prévision. Plus précisément, il s'agit de savoir comment l'incertitude due à une méthode d'imputation se propage sur l'erreur en prévision.

Ces points n'ont pas été traités dans cette thèse, ils constituent en eux mêmes une thématique de recherche très intéressante, tant au niveau théorique, qu'applicatif. En effet, un nouvel objectif serait de proposer des méthodes d'imputation pour lesquelles l'erreur en prévision est moins influencée par l'incertitude liée à l'algorithme d'imputation.

Sur les outils de traitement statistique et logiciel

Nous avons souligné au cours de la présentation des données que la visualisation tient surtout aux aspects de synthèse et de clarté. Néanmoins, face à l'importance des données disponibles et celles à venir, le traitement statistique nécessite une gestion en amont de type algorithmique. Cette problématique et celle de la quantité de l'information amènent aussi à s'interroger sur la visualisation dynamique ou interactive des résultats.

La plupart des fonctions dans cette thèse ont été écrites sur R (2015), les graphiques ont été élaborés grâce au magnifique package créé par Wickham (2009). La plupart des bases de données ont été fragmentées afin de permettre une gestion de calcul plus efficiente. Le traitement statistique simultané de toutes les bases est nécessaire, donc la gestion du calcul est indispensable. Pour cela, les premiers tests avec un autre langage de programmation (données issues de la campagne de 2015), en l'occurrence `Julia`[16] montre une puissance considérable par rapport aux autres langages de programmation, et ceci tout en gardant les représentations graphiques typiques à `R` ou à `Python`.

Par conséquent, je pense que `Julia` en tant que langage de programmation combinant la puissance de calcul et les fonctionnalités de `R`, sera en premier plan pour la prochaine génération de traitement de données, car il vise le même champ des applications que `R` : la manipulation de données, les analyses statistiques et les systèmes de packages (convivialité de partage des connaissances).

La question de l'identification des sources

Les techniques de séparation des sources abordées dans ce manuscrit ont permis un retour d'expérience très enrichissant. Partant de cette expérience, plusieurs perspectives à ce travail peuvent être envisagées.

16. Écrit en C, mais est presque aussi rapide que C lui-même, plus rapide de dizaine de fois que Python, et de centaines de fois plus que R (pou en citer que les logiciels libre), voir l'introduction et la présentation du logiciel dans (Bezanson et al., 2014).

Une étape immédiate est l'application approfondie des méthodes de factorisation non-négatives, notamment mettre à disposition l'utilisation facile de ces méthodes. Sur le plan théorique, un problème ouvert est celui de l'estimation du nombre de sources. Donc, une question fondamentale est d'étudier si l'estimation du nombre de source est problème "démontrable" ; si oui déterminer un algorithme permettant de l'estimer.

Sur le plan du traitement des données réelles, une étude intéressante à mener concerne le traitement des séries temporelles issues de l'analyse de plusieurs environnements, donc une dimension spatiale s'ajoute aux deux dimensions temporelle et individuelle. La question concernant la fusion des données est immédiate.

Une orientation importante concerne l'extension de tous ces développements au cas multilinéaire : il s'agit des modèles PARAFAC. Je pense que ce type de modèle permettrait d'estimer la question fondamentale de la qualité de l'air intérieur, celle de la multi-expositions aux contaminants micro-environnementaux.

Prévision des modèles de prévision

Quel est l'avenir des modèles de prévision ? L'œil nu est le premier outil d'analyse statistique, donc tout dépend des données, de leurs structures et de leur volume, mais surtout tout dépend de l'horizon.

Sur le très court terme

Une extension immédiate de modèles appliqués ou développés dans cette thèse concerne la prévision des séries temporelles multidimensionnelles à seuil. Un modèle de type **LTDM** (pour Latent Threshold Dynamic Models) serait susceptible d'apporter des éclairages nouveaux sur deux questions simultanées : *(i)* l'identification des déterminants des fluctuations par l'analyse des facteurs latents et *(ii)* la prévision par des modèles dynamiques.

Les modèles de la dynamique du chaos sur des données multidimensionnelles devraient permettre d'améliorer la qualité de prévision sur les plusieurs séries définies à partir de plusieurs bandes spectrales.

L'analyse de causalité, de cointégration (éventuellement fractionnaire), sont des outils très puissants en analyse multivariée des séries temporelles : on devrait pouvoir analyser de façon conjointe les composantes latentes stochastiques des variables.

Sur le moyen terme

Les modèles prévision à moyen terme sont ceux qui devraient être intégrés dans un processus automatique de prévision. Une instrumentation active d'un environnement réel est alors envisageable avec des outils numériques actuels.

Pour finir, je pense que l'analyse topologique du chaos, à la Gilmore & Lefranc (2002), des attracteurs étranges que nous avons pu observer ouvrira une piste commune entre l'analyse des séries temporelles et la géométrie différentielle : le chaos s'installe pour unifier. Poincaré nous avait prévenus : *"Faire des mathématiques, c'est donner le même nom à des choses différentes"*. Aujourd'hui, le chaos signifie beaucoup de choses, bien plus que Poincaré ou Lorenz n'auraient pu l'imaginer (Ghys, 2007).

BIBLIOGRAPHIE

Abadie, M. O. & Blondeau, P. (2011). Pandora database : A compilation of indoor air pollutant emissions. *HVAC&R Research*, 17(4), 602–613. 1, 28

Abarbanel, H. (2012). *Analysis of Observed Chaotic Data*. Institute for Nonlinear Science. Springer New York. 266, 277

Abarbanel, H. D., Brown, R., Sidorowich, J. J., & Tsimring, L. S. (1993). The analysis of observed chaotic data in physical systems. *Reviews of modern physics*, 65(4), 1331. 277

Abraham-Frois, G. (1998). *Non-linear dynamics and endogenous cycles*, volume 463. Springer Science & Business Media. 264

Abraham-Frois, G., Lardic, S., & Mignon, V. (1998). Long-term memory and chaos : a note. In *Non-Linear Dynamics and Endogenous Cycles* (pp. 185–201). Springer. 109

Abt, E., Suh, H. H., Catalano, P., & Koutrakis, P. (2000). Relative contribution of outdoor and indoor particle sources to indoor concentrations. *Environmental science & technology*, 34(17), 3579–3587. 20, 23

Afshari, A., Matson, U., & Ekberg, L. E. (2005). Characterization of indoor sources of fine and ultrafine particles : a study conducted in a full-scale chamber. *Indoor air*, 15(2), 141–150. 23

Ailliot, P. (2004). *Modèles autorégressifs à changements de régimes markoviens. Applications aux séries temporelles de vent*. PhD thesis, Université de Rennes 1. 249, 250

AirParif (2008). Technical report, Airparif. 71

Akaike, H. (1974). A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, 19(6), 716–723. 195

Alzona, J., Cohen, B., Rudolph, H., Jow, H., & Frohliger, J. (1979). Indoor-outdoor relationships for airborne particulate matter of outdoor origin. *Atmospheric Environment (1967)*, 13(1), 55–60. 21

Alzueta, M. U. & Glarborg, P. (2003). Formation and destruction of ch2o in the exhaust system of a gas engine. *Environmental science & technology*, 37(19), 4512–4516. 17

Amari, S., Nagaoka, H., & Harada, D. (2007). *Methods of Information Geometry*. Translations of Mathematical Monographs. American Mathematical Society. 373

Amato, F., Rivas, I., Viana, M., Moreno, T., Bouso, L., Reche, C., Alvarez-Pedrerol, M., Alastuey, A., Sunyer, J., & Querol, X. (2014). Sources of indoor and outdoor pm2. 5 concentrations in primary schools. *Science of the Total Environment*, 490, 757–765. 24

Andersen, I., Lundqvist, G., & Mølhave, L. (1975). Indoor air pollution due to chipboard used as a construction material. *Atmospheric Environment (1967)*, 9(12), 1121–1127. 17, 18

Andrews, D. W. (1991). Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica : Journal of the Econometric Society*, (pp. 817–858). 114

Anh, V., Lam, K., Leung, Y., & Tieng, Q. (2000). Multifractal analysis of hong kong air quality data. *Environmetrics*, 11(2), 139–149. 106

Ardilly, P. (2006). *Les techniques de sondage*. Editions Technip. 384

Ashbaugh, L. L., Malm, W. C., & Sadeh, W. Z. (1985). A residence time probability analysis of sulfur concentrations at grand canyon national park. *Atmospheric Environment (1967)*, 19(8), 1263–1270. 57

Atkinson, R. & Arey, J. (2003). Gas-phase tropospheric chemistry of biogenic volatile organic compounds : a review. *Atmospheric Environment*, 37, 197–219. 17

Austin, J., Brimblecombe, P., & Sturges, W. (2002). *Air pollution science for the 21st century*, volume 1. Elsevier. 15

Azencott, R. & Dacunha-Castelle, D. (1984). *Séries d'observations irrégulières : modélisation et prévision*. Elsevier Masson. 86, 88, 191

Bachelier, L. (1900). Théorie de la spéculation. *Annales scientifiques de l'École Normale Supérieure*, 3(17), 21–86. 106

Bakó-Biró, Z., Wargocki, P., Weschler, C. J., & Fanger, P. O. (2004). Effects of pollution from personal computers on perceived air quality, sbs symptoms and productivity in offices. *Indoor air*, 14(3), 178–187. 25

Barahona, M. & Poon, C.-S. (1996). Detection of nonlinear dynamics in short, noisy time series. *Nature*, 381(6579), 215–217. 242

Barnsley, M. F. (1988). *Fractals every where*. Academic Press, San Diego. 106

Batschelet, E., Batschelet, E., Batschelet, E., & Batschelet, E. (1981). *Circular statistics in biology*, volume 111. Academic press London. 57

Baum, L. E., Eagon, J. A., et al. (1967). An inequality with applications to statistical estimation for probabilistic functions of markov processes and to a model for ecology. *Bull. Amer. Math. Soc*, 73(3), 360–363. 249

Baum, L. E. & Petrie, T. (1966). Statistical inference for probabilistic functions of finite state markov chains. *The annals of mathematical statistics*, 37(6), 1554–1563. 249

Baum, L. E., Petrie, T., Soules, G., & Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *The annals of mathematical statistics*, 41(1), 164–171. 249

Berge, A., Mellegaard, B., Hanetho, P., & Ormstad, E. (1980). Formaldehyde release from particleboard : Evaluation of a mathematical model. *European Journal of Wood and Wood Products*, 38(7), 251–255. 18

Berman, A. & Plemmons, R. J. (1979). Nonnegative matrices. *The Mathematical Sciences, Classics in Applied Mathematics*, 9. 163

Bezanson, J., Edelman, A., Karpinski, S., & Shah, V. B. (2014). Julia : A fresh approach to numerical computing. *arXiv preprint arXiv :1411.1607*. 306

Bohl, M. T. & Siklos, P. L. (2004). The present value model of us stock prices redux : a new testing strategy and some evidence. *The Quarterly Review of Economics and Finance*, 44(2), 208–223. 246

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of econometrics*, 31(3), 307–327. 233

Bollerslev, T., Russell, J., & Watson, M. (2010). *Volatility and time series econometrics : essays in honor of Robert Engle*. Oxford University Press. 233

Box, G. E. & Draper, N. R. (1987). *Empirical model-building and response surfaces*, volume 424. Wiley New York. 149

Box, G. E. & Jenkins, G. M. (1970). *Time series analysis : forecasting and control*. Holden-Day. 194

Box, G. E. & Jenkins, G. M. (1976). *Time series analysis : forecasting and control, revised ed*. Holden-Day. 86, 195, 382

Brillinger, D. (2001). *Time Series : Data Analysis and Theory*. Classics in Applied Mathematics. Society for Industrial and Applied Mathematics. 86, 92

Brinke, J. T., Selvin, S., Hodgson, A., Fisk, W., Mendell, M., Koshland, C., & Daisey, J. (1998). Development of new volatile organic compound (voc) exposure metrics and their relationship to "sick building syndrome" symptoms. *Indoor Air*, 8(3), 140–152. 26

Brockwell, P. J. & Davis, R. A. (1991). *Time series theory and methods*. Springer series in statistics. New York : Springer. 86, 87, 90, 91, 92, 94, 191, 192, 193

Brook, R. D., Franklin, B., Cascio, W., Hong, Y., Howard, G., Lipsett, M., Luepker, R., Mittleman, M., Samet, J., Smith, S. C., et al. (2004). Air pollution and cardiovascular disease a statement for healthcare professionals from the expert panel on population and prevention science of the american heart association. *Circulation*, 109(21), 2655–2671. 15

Broomhead, D. & King, G. P. (1986a). On the qualitative analysis of experimental dynamical systems. *Nonlinear Phenomena and Chaos*, 113, 114. 135, 136, 213

Broomhead, D. S. & King, G. P. (1986b). Extracting qualitative dynamics from experimental data. *Physica D : Nonlinear Phenomena*, 20(2), 217–236. 135, 136, 213, 260

Brown, B. W. & Mariano, R. S. (1989). Predictors in dynamic nonlinear models : large-sample behavior. *Econometric Theory*, 5(03), 430–452. 258

Brown, R. G. (2004). *Smoothing, forecasting and prediction of discrete time series*. Courier Corporation. 209

Brunekreef, B. & Holgate, S. T. (2002). Air pollution and health. *The lancet*, 360(9341), 1233–1242. 26

Brunelli, U., Piazza, V., Pignato, L., Sorbello, F., & Vitabile, S. (2007). Two-days ahead prediction of daily maximum concentrations of so 2, o 3, pm10, no 2, co in the urban area of palermo, italy. *Atmospheric Environment*, 41(14), 2967–2995. 201

Bruno, F., Cocchi, D., & Trivisano, C. (2004). Forecasting daily high ozone concentrations by classification trees. *Environmetrics*, 15(2), 141–153. 203

Bruno, R. & Raspa, G. (1989). Geostatistical characterization of fractal models of surfaces. In *Geostatistics* (pp. 77–89). Springer. 113

Caner, M. & Hansen, B. E. (2001). Threshold autoregression with a unit root. *Econometrica*, 69(6), 1555–1596. 246

Cao, L. (2002). Nonlinear prediction of time series using wavelet network method. In *Modelling and Forecasting Financial Data* (pp. 179–195). Springer. 260

Cappé, O., Moulines, E., & Rydén, T. (2009). *Inference in hidden markov models*. Springer Heidelberg. 249

Carslaw, D. C. & Beevers, S. D. (2013). Characterising and understanding emission sources using bivariate polar plots and k-means clustering. *Environmental Modelling & Software*, 40, 325–329. 57, 74

Carslaw, D. C., Beevers, S. D., Ropkins, K., & Bell, M. C. (2006). Detecting and quantifying aircraft and other on-airport contributions to ambient nitrogen oxides in the vicinity of a large international airport. *Atmospheric Environment*, 40(28), 5424–5434. 57, 74

Carslaw, N. (2007). A new detailed chemical model for indoor air pollution. *Atmospheric Environment*, 41(6), 1164–1179. 16, 28

Cartan, É. (1925). *La géométrie des espaces de Riemann*. Gauthier-Villars. 372

Casdagli, M. (1989). Nonlinear prediction of chaotic time series. *Physica D : Nonlinear Phenomena*, 35(3), 335–356. 277

Casdagli, M. (1992). Nonlinear forecasting, chaos and statistics. In *Modeling complex phenomena* (pp. 131–152). Springer. 260, 277

Casdagli, M., Eubank, S., Farmer, J. D., & Gibson, J. (1991). State space reconstruction in the presence of noise. *Physica D : Nonlinear Phenomena*, 51(1), 52–98. 260

Catelinois, O., Rogel, A., Laurier, D., Billon, S., Hemon, D., Verger, P., & Tirmarche, M. (2006). Lung cancer attributable to indoor radon exposure in france : impact of the risk models and uncertainty analysis. *Environmental health perspectives*, (pp. 1361–1366). 10

César, W. (2014). *Outils numériques et technologiques pour l'analyse de la qualité de l'air intérieur*. PhD thesis, Paris Est. 263

Chan, A. T. (2002). Indoor–outdoor relationships of particulate matter and nitrogen oxides under different outdoor meteorological conditions. *Atmospheric Environment*, 36(9), 1543–1551. 24

Chan, G. & Wood, A. T. (2004). Estimation of fractal dimension for a class of non-gaussian stationary processes and fields. *Annals of statistics*, (pp. 1222–1260). 113

Chan, K. S. & Tong, H. (1986). On estimating thresholds in autoregressive models. *Journal of Time Series Analysis*, 7(3), 179–190. 248

Chandler, R. & Scott, M. (2011). *Statistical methods for trend detection and analysis in the environmental sciences*. John Wiley & Sons. 86

Chao, C. Y. & Cheng, E. C. (2002). Source apportionment of indoor pm2. 5 and pm10 in homes. *Indoor and Built Environment*, 11(1), 27–37. 153

Chatoutsidou, S. E., Ondráček, J., Tesar, O., Tørseth, K., Zdimal, V., & Lazaridis, M. (2015). Indoor/outdoor particulate matter number and mass concentration in modern offices. *Building and Environment*. 22

Chelani, A. (2005). Predicting chaotic time series of pm10 concentration using artificial neural network. *International journal of environmental studies*, 62(2), 181–191. 202

Chelani, A. B., Singh, R., & Devotta, S. (2005). Nonlinear dynamical characterization and prediction of ambient nitrogen dioxide concentration. *Water, air, and soil pollution*, 166(1-4), 121–138. 202

Chen, C. & Zhao, B. (2011). Review of relationship between indoor and outdoor particles : I/o ratio, infiltration factor and penetration factor. *Atmospheric Environment*, 45(2), 275–288. 21

Chen, G., Abraham, B., & Bennett, G. W. (1997). Parametric and non-parametric modelling of time series - an empirical study. *Environmetrics*, 8(1), 63–74. 200

Chen, J.-L., Islam, S., & Biswas, P. (1998). Nonlinear dynamics of hourly ozone concentrations : nonparametric short term prediction. *Atmospheric environment*, 32(11), 1839–1848. 199, 202

Chikuse, Y. (2012). *Statistics on special manifolds*, volume 174. Springer Science & Business Media. 53

Chow, J. C. & Watson, J. G. (2002). Review of pm2. 5 and pm10 apportionment for fossil fuel combustion and other sources by the chemical mass balance receptor model. *Energy & Fuels*, 16(2), 222–260. 153

Chow, J. C., Watson, J. G., et al. (1998). *Guideline on speciated particulate monitoring*. Technical report, US Environmental Protection Agency. 19

Chow, S.-M. & Zhang, G. (2013). Nonlinear regime-switching state-space (rsss) models. *Psychometrika*, 78(4), 740–768. 248

Chreiky, R., Delmaire, G., Puigt, M., Roussel, G., Courcot, D., & Abche, A. (2015). Split gradient method for informed non-negative matrix factorization. In *Latent Variable Analysis and Signal Separation* (pp. 376–383). Springer. 163

Cichocki, A. & Amari, S.-i. (2002). *Adaptive blind signal and image processing : learning algorithms and applications*, volume 1. John Wiley & Sons. 153, 163

Cichocki, A., Zdunek, R., Phan, A. H., & Amari, S.-i. (2009). *Nonnegative matrix and tensor factorizations : applications to exploratory multi-way data analysis and blind source separation*. John Wiley & Sons. 153, 163, 165, 167, 168, 169

Clements, M. P. & Hendry, D. F. (2002). Explaining forecast failure in macroeconomics. *A companion to economic forecasting*, (pp. 539–571). 149

Cleveland, R. B., Cleveland, W. S., McRae, J. E., & Terpenning, I. (1990). Stl : A seasonal-trend decomposition procedure based on loess. *Journal of Official Statistics*, 6(1), 3–73. 131, 203

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American statistical association*, 74(368), 829–836. 131, 132, 203

Cogliano, V. J., Grosse, Y., Baan, R. A., Straif, K., Secretan, M. B., El Ghissassi, F., for Volume 88, W. G., et al. (2005). Meeting report : summary of iarc monographs on formaldehyde, 2-butoxyethanol, and 1-tert-butoxy-2-propanol. *Environmental health perspectives*, (pp. 1205–1208). 26

Colletaz, G. & Hurlin, C. (2007). *Modèles Non Linéaires et Prévisions*. Technical report. working paper or preprint. 248, 258

Comon, P. (1994). Independent component analysis, a new concept ? *Signal processing*, 36(3), 287–314. 157, 160

Comon, P. & Jutten, C. (2007). *Séparation de sources 1. Concepts de base et analyse en composantes indépendantes*. LAVOISIER. 152, 153, 161

Comon, P. & Jutten, C. (2010). *Handbook of Blind Source Separation : Independent Component Analysis and Applications*. Independent Component Analysis and Applications Series. Elsevier Science. 152, 153

Constantine, A. & Hall, P. (1994). Characterizing surface smoothness via estimation of effective fractal dimension. *Journal of the Royal Statistical Society. Series B (Methodological)*, (pp. 97–113). 113

Cover, T. M. & Thomas, J. A. (2012). *Elements of information theory*. John Wiley & Sons. 93, 159

Cox, D. R. (1961). Prediction by exponentially weighted moving averages and related methods. *Journal of the Royal Statistical Society. Series B (Methodological)*, (pp. 414–422). 209

Crone, S. F., Nikolopoulos, K., & Hibon, M. (2005). Automatic modelling and forecasting with artificial neural networks : A forecasting competition evaluation. *Final report for the IIF/SAS Grant*, 6. 201

Cutler, C. D. (1993). A review of the theory and estimation of fractal dimension. In H. Tong (Ed.), *Dimension estimation and models*, volume 1. World Scientific. 106

D'Agostino, R. (1986). *Goodness-of-Fit-Techniques*. Statistics : A Series of Textbooks and Monographs. Taylor & Francis. 242

Damian, A. M. (2003). *Modélisation zonale de la qualité de l'air à l'intérieur des bâtiments : application à l'évaluation de l'exposition des occupants*. PhD thesis, La Rochelle. 263

Dassonville, C., Demattei, C., Laurent, A.-M., Le Moullec, Y., Seta, N., & Momas, I. (2009). Assessment and predictor determination of indoor aldehyde levels in paris newborn babies' homes. *Indoor Air*, 19(4), 314–323. 16, 19

De Bruin, Y. B., Koistinen, K., Kephalopoulos, S., Geiss, O., Tirendi, S., & Kotzias, D. (2008). Characterisation of urban inhalation exposures to benzene, formaldehyde and acetaldehyde in the european union. *Environmental Science and Pollution Research*, 15(5), 417–430. 19

Demailly, J.-P. (2012). *Analyse numérique et équations différentielles*. EDP sciences. 378

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, (pp. 1–38). 250, 377, 378, 381, 385, 386

Destaillats, H., Maddalena, R. L., Singer, B. C., Hodgson, A. T., & McKone, T. E. (2008). Indoor pollutants emitted by office equipment : A review of reported data and information needs. *Atmospheric Environment*, 42(7), 1371–1388. 14, 15

Díaz-Robles, L. A., Ortega, J. C., Fu, J. S., Reed, G. D., Chow, J. C., Watson, J. G., & Moncada-Herrera, J. A. (2008). A hybrid arima and artificial neural networks model to forecast particulate matter in urban areas : The case of temuco, chile. *Atmospheric Environment*, 42(35), 8331–8340. 201

Diebold, F. X. & Inoue, A. (2001). Long memory and regime switching. *Journal of econometrics*, 105(1), 131–159. 144

Dieudonné, J. (2009). *A history of algebraic and differential topology, 1900-1960*. Springer Science &amp ; Business Media. 372

Dijk, D. v., Teräsvirta, T., & Franses, P. H. (2002). Smooth transition autoregressive models -a survey of recent developments. *Econometric Reviews*, 21(1), 1–47. 244, 248

Dingle, P. & Franklin, P. (2002). Formaldehyde levels and the factors affecting these levels in homes in perth, western australia. *Indoor and Built Environment*, 11(2), 111–116. 19

Douc, R., Matias, C., et al. (2001). Asymptotics of the maximum likelihood estimator for general hidden markov models. *Bernoulli*, 7(3), 381–420. 249

Douc, R., Moulines, E., Ryden, T., et al. (2004). Asymptotic properties of the maximum likelihood estimator in autoregressive models with markov regime. *The Annals of statistics*, 32(5), 2254–2304. 249, 250

Douc, R., Moulines, E., & Stoffer, D. (2014). *Nonlinear time series : Theory, methods and applications with R examples*. CRC Press. 92, 94, 191, 246, 247

Doukhan, P., Oppenheim, G., & Taqqu, M. S. (2003). *Theory and applications of long-range dependence*. Springer Science & Business Media. 86

Doz, C., Rabault, G., & Sobczak, N. (1995). Décomposition tendance-cycle : estimations par des méthodes statistiques univariées. *Économie & prévision*, 120(4), 73–93. 130

Dreyfus, G., Martinez, J.-M., Samuelides, M., Gordon, M. B., Badran, F., & Thiria, S. (2011). *Apprentissage statistique : Réseaux de neurones-Cartes topologiques-Machines à vecteurs supports*. Editions Eyrolles. 201

Dryden, I. L. & Kent, J. T. (2015). *Geometry Driven Statistics*. John Wiley & Sons. 53

Dubuc, B., Quiniou, J., Roques-Carmes, C., Tricot, C., & Zucker, S. (1989). Evaluating the fractal dimension of profiles. *Physical Review A*, 39(3), 1500. 111

Durdu, Ö. F. (2010). Stochastic approaches for time series forecasting of boron : a case study of western turkey. *Environmental monitoring and assessment*, 169(1-4), 687–701. 199

Edwards, R. D., Jurvelin, J., Koistinen, K., Saarela, K., & Jantunen, M. (2001). Voc source identification from personal and residential indoor, outdoor and workplace microenvironment samples in expolis-helsinki, finland. *Atmospheric Environment*, 35(28), 4829–4841. 16

El Raey, M., Shalaby, E., Ghatass, Z., & Marey, H. (2006). Time series analysis of ambient air concentrations in nile delta region. In *2nd Environmental Physics Conference*. 198

Elsner, J. & Tsonis, A. (1996). *Singular Spectrum Analysis : A New Tool in Time Series Analysis*. Springer Science & Business Media. 135

Elsner, P., Hatch, K. L., & Wigger-Alberti, W. (2003). *Textiles and the Skin*, volume 31. Karger Medical and Scientific Publishers. 16

Embrechts, P. (2009). *Selfsimilar processes*. Princeton University Press. 86

Enders, W. & Granger, C. W. J. (1998). Unit-root tests and asymmetric adjustment with an example using the term structure of interest rates. *Journal of Business & Economic Statistics*, 16(3), 304–311. 246

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica : Journal of the Econometric Society*, (pp. 987–1007). 231

Ezzati, M. (2005). Indoor air pollution and health in developing countries. *The Lancet*, 366(9480), 104–106. 26

Falconer, K. (2004). *Fractal geometry : mathematical foundations and applications*. John Wiley & Sons, 2 edition. 108, 109

Farmer, J. D. & Sidorowich, J. J. (1987). Predicting chaotic time series. *Physical review letters*, 59(8), 845. 190, 277

Fassò, A. & Negri, I. (2002a). Multi-step forecasting for nonlinear models of high frequency ground ozone data : a monte carlo approach. *Environmetrics*, 13(4), 365–378. 200

Fassò, A. & Negri, I. (2002b). Non-linear statistical modelling of high frequency ground ozone data. *Environmetrics*, 13(3), 225–241. 200

Feller, W. (1950). *An introduction to probability theory and its applications. Vol. I.* Wiley. 191

Fenech, A., Strlič, M., Cigić, I. K., Levart, A., Gibson, L. T., de Bruin, G., Ntanos, K., Kolar, J., & Cassar, M. (2010). Volatile aldehydes in libraries and archives. *Atmospheric Environment*, 44(17), 2067–2073. 15

Finlayson-Pitts, B. J. & Pitts Jr, J. N. (1999). *Chemistry of the upper and lower atmosphere : theory, experiments, and applications*. Academic press. 14, 15

Finz, G., Fronza, G., & Spirito, A. (1980). Multivariate stochastic models of sulphur dioxide pollution in an urban area. *Journal of the Air Pollution Control Association*, 30(11), 1212–1215. 198

Fisher, N. I. (1995). *Statistical analysis of circular data*. Cambridge University Press. 57

Fisk, W. & Seppanen, O. (2007). Providing better indoor environmental quality brings economic benefits. *Lawrence Berkeley National Laboratory*. 26

Fisk, W. J., Eliseeva, E. A., & Mendell, M. J. (2010). Association of residential dampness and mold with respiratory tract infections and bronchitis : a meta-analysis. *Environmental Health*, 9(72). 10

Fisk, W. J. & Rosenfeld, A. H. (1997). Estimates of improved productivity and health from better indoor environments. *Indoor air*, 7(3), 158–172. 26

Flanders, H. (1963). *Differential forms with applications to the physical sciences*. Math. Sci. Eng. New York, NY : Academic Press. Also as reprint ed. : New York, Dover, 1989. 374

Foata, D. & Fuchs, A. (2002). *Processus stochastiques : processus de Poisson, chaînes de Markov et martingales : cours et exercices corrigés*. Dunod. 250

Fraedrich, K. (1986). Estimating the dimensions of weather and climate attractors. *Journal of the atmospheric sciences*, 43(5), 419–432. 213

Francq, C. & Zakoian, J.-M. (2011). *GARCH models : structure, statistical inference and financial applications*. John Wiley & Sons. 233

Franses, P. H. & Van Dijk, D. (2000). *Non-linear time series models in empirical finance*. Cambridge University Press. 258, 259

Frausto, H. U. & Pieters, J. G. (2004). Modelling greenhouse temperature using system identification by means of neural networks. *Neurocomputing*, 56, 423–428. 204

Fuller, W. (2009). *Introduction to Statistical Time Series*. Wiley Series in Probability and Statistics. Wiley, 2 edition. 86, 191

Funaki, R., Tanabe, S.-i., Tanaka, H., & Nakagawa, T. (2003). Measurements of chemical emission rates from portable pc and electronic appliances. *Journal of Asian Architecture and Building Engineering*, 2(2), b55–b59. 17

Galka, A. (2000). *Topics in nonlinear time series analysis : with implications for EEG analysis*, volume 14. World Scientific. 264

Gardner, E. S. (1985). Exponential smoothing : The state of the art. *Journal of forecasting*, 4(1), 1–28. 209

Gardner, E. S. (2006). Exponential smoothing : The state of the art -part ii. *International journal of forecasting*, 22(4), 637–666. 209

Gardner, M. W. & Dorling, S. (1998). Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. *Atmospheric environment*, 32(14), 2627–2636. 201

Gaujoux, R. & Seoighe, C. (2010). A flexible r package for nonnegative matrix factorization. *BMC bioinformatics*, 11(1), 1. 163

Géhin, E., Ramalho, O., & Kirchner, S. (2008). Size distribution and emission rate measurement of fine and ultrafine particle from indoor human activities. *Atmospheric Environment*, 42(35), 8341–8352. 23

Geweke, J. & Porter-Hudak, S. (1983). The estimation and application of long memory time series models. *Journal of time series analysis*, 4(4), 221–238. 115, 116

Ghil, M., Allen, M., Dettinger, M., Ide, K., Kondrashov, D., Mann, M., Robertson, A. W., Saunders, A., Tian, Y., Varadi, F., et al. (2002). Advanced spectral methods for climatic time series. *Reviews of geophysics*, 40(1), 3–1. 86

Ghosh, S. (1996). A new graphical tool to detect non-normality. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(4), pp. 691–702. 242

Ghys, E. (2007). L'effet papillon. *Images des Mathématiques*. 307

Gilbert, N. L., Gauvin, D., Guay, M., Héroux, M.-È., Dupuis, G., Legris, M., Chan, C. C., Dietz, R. N., & Lévesque, B. (2006). Housing characteristics and indoor concentrations of nitrogen dioxide and formaldehyde in quebec city, canada. *Environmental Research*, 102(1), 1–8. 19

Gilmore, R. & Lefranc, M. (2002). The topology of chaos : Alice in stretch and squeezelan. *Hoboken : Wiley & sons inc*, (pp. 518). 239, 267, 272, 307

Giraitis, L., Koul, H. L., & Surgailis, D. (2012). *Large sample inference for long memory processes*, volume 201. World Scientific. 92

Gneiting, T., Ševčíková, H., Percival, D. B., et al. (2012). Estimators of fractal dimension : Assessing the roughness of time series and spatial data. *Statistical Science*, 27(2), 247–277. 110, 112

Goerg, G. (2013). Forecastable component analysis. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)* (pp. 64–72). 90, 93, 99, 103, 105, 145, 301

Goldfeld, S. M. & Quandt, R. E. (1973). A markov model for switching regressions. *Journal of econometrics*, 1(1), 3–15. 249

Golyandina, N. (2010). On the choice of parameters in singular spectrum analysis and related subspace-based methods. *Statistics and Its Interface*, 3(3), 259–279. 137, 140, 213

Golyandina, N., Nekrutkin, V., & Zhigljavsky, A. A. (2001). *Analysis of time series structure : SSA and related techniques*. CRC press. 135, 140, 213

Golyandina, N. & Zhigljavsky, A. (2013). *Singular Spectrum Analysis for time series*. Springer Science & Business Media. 213

Gourieroux, C. & Monfort, A. (1995). *Séries temporelles et modèles dynamiques*. Collection Economie et statistiques avancées. Série Ecole nationale de la statistique et de l'administration et du Centre d'études des programmes économiques. Economica. 86, 150, 191, 196, 209

Granger, C. W., Terasvirta, T., et al. (1993). *Modelling non-linear economic relationships*. Oxford University Press. 248

Granger, C. W. J. (1998). Extracting information from mega-panels and high-frequency data. *Statistica Neerlandica*, 52(3), 258–272. 305

Grassberger, P. & Procaccia, I. (1983). Characterization of strange attractors. *Physical review letters*, 50(5), 346. 202

Grimaldi, F. & Déoux, S. (2003). *L air et la santé*, chapter Chapitre 4 : polluants atmosphériques intérieurs, (pp. 35–53). Médecine-Sciences Flammarion. 15, 16

Grivas, G. & Chaloulakou, A. (2006). Artificial neural network models for prediction of pm 10 hourly concentrations, in the greater area of athens, greece. *Atmospheric Environment*, 40(7), 1216–1229. 204

Groth, A. & Ghil, M. (2011). Multivariate singular spectrum analysis and the road to phase synchronization. *Physical Review E*, 84(3), 036206. 135

Guégan, D. (1994). *Séries chronologiques non linéaires àtemps discret*. Presses Universitaires d'Aix-Marseille. 242

Guégan, D. (2003). *Les chaos en finance : approche statistique*. Economica. 144, 242, 260, 264

Guégan, D. (2008). Effect of noise filtering on predictions : on the routes of chaos. Documents de travail du Centre d'Economie de la Sorbonne 2008.08 - ISSN : 1955-611X. 260, 267

Gundel, L. A. & Sextro, R. G. (2005). *Aerosol physics and chemistry : indoor perspective, Chapter 10*, chapter Chapter, (pp. 189–224). CRC Press : Boca Raton. 20

Guo, H. (2011). Source apportionment of volatile organic compounds in hong kong homes. *Building and Environment*, 46(11), 2280–2286. 17, 153

Gupta, A. & Cheong, K. D. (2007). Physical characterization of particulate matter and ambient meteorological parameters at different indoor–outdoor locations in singapore. *Building and environment*, 42(1), 237–245. 22

Haghighat, F. & De Bellis, L. (1998). Material emission rates : literature review, and the impact of indoor air temperature and relative humidity. *Building and Environment*, 33(5), 261–277. 16

Hall, P. & Wood, A. (1993). On the performance of box-counting estimators of fractal dimension. *Biometrika*, 80(1), 246–251. 111, 112

Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica : Journal of the Econometric Society*, (pp. 357–384). 249

Hamilton, J. D. (1990). Analysis of time series subject to changes in regime. *Journal of econometrics*, 45(1), 39–70. 249, 250

Hamilton, J. D. (1994). *Time series analysis*, volume 2. Princeton university press Princeton. 250

Han, Y., Qi, M., Chen, Y., Shen, H., Liu, J., Huang, Y., Chen, H., Liu, W., Wang, X., Liu, J., et al. (2015). Influences of ambient air pm 2.5 concentration and meteorological condition on the indoor pm 2.5 concentrations in a residential apartment in beijing using a new approach. *Environmental Pollution*, 205, 307–314. 21

Hannan, E. (1966). *Notes on Time Series Analysis : Lectures, 1964-1965*. Number vol. 1 à 2 in Notes on Time Series Analysis : Lectures, 1964-1965. Department of Statistics, Johns Hopkins University. 86

Hänninen, O. O., Alm, S., Katsouyanni, K., Künzli, N., Maroni, M., Nieuwenhuijsen, M. J., Saarela, K., Srám, R. J., Zmirou, D., & Jantunen, M. J. (2004). The expolis study : implications for exposure research and environmental policy in europe. *Journal of Exposure Science and Environmental Epidemiology*, 14(6), 440–456. 15

Hanrahan, G. (2011). *Artificial neural networks in biological and environmental analysis*. CRC Press. 201

Hansen, B. E. (1997). Inference in tar models. *Studies in nonlinear dynamics & econometrics*, 2(1). 247

Harvey, A. C. (1984). A unified view of statistical forecasting procedures. *Journal of Forecasting*, 3(3), 245–275. 209

Hassanvand, M. S., Naddafi, K., Faridi, S., Arhami, M., Nabizadeh, R., Sowlat, M. H., Pourpak, Z., Rastkari, N., Momeniha, F., Kashani, H., et al. (2014). Indoor/outdoor relationships of pm 10, pm 2.5, and pm 1 mass concentrations and their water-soluble ions in a retirement home and a school dormitory. *Atmospheric Environment*, 82, 375–382. 24

Hastie, T. J. & Tibshirani, R. J. (1990). *Generalized additive models*, volume 43. CRC Press. 382

He, C., Morawska, L., & Taplin, L. (2007). Particle emission characteristics of office printers. *Environmental science & technology*, 41(17), 6039–6045. 24

He, C., Morawska, L., Wang, H., Jayaratne, R., McGarry, P., Johnson, G. R., Bostrom, T., Gonthier, J., Authemayou, S., & Ayoko, G. (2010). Quantification of the relationship between fuser roller temperature and laser printer emissions. *Journal of Aerosol Science*, 41(6), 523–530. 14

Hegger, R., Kantz, H., & Schreiber, T. (1999). Practical implementation of nonlinear time series methods : The tisean package. *Chaos : An Interdisciplinary Journal of Nonlinear Science*, 9(2), 413–435. 264, 275, 278, 288

Higuchi, T. (1988). Approach to an irregular time series on the basis of the fractal theory. *Physica D : Nonlinear Phenomena*, 31(2), 277–283. 126

Ho, K., Cao, J., Harrison, R. M., Lee, S., & Bau, K. (2004). Indoor/outdoor relationships of organic carbon (oc) and elemental carbon (ec) in pm 2.5 in roadside environment of hong kong. *Atmospheric Environment*, 38(37), 6327–6335. 22

Hodgson, A., Beal, D., & McIlvaine, J. (2002). Sources of formaldehyde, other aldehydes and terpenes in a new manufactured house. *Indoor Air*, 12(4), 235–242. 16

Holt, C. C. (1957). *Forecasting Trends and Seasonals by Exponentially Weighted Averages*. Pittsburgh Office of Naval Research memorandum, 5 edition. 209, 210

Holt, C. C. (2004). Forecasting seasonals and trends by exponentially weighted moving averages. *International Journal of Forecasting*, 20(1), 5–10. 209

Honaker, J. & King, G. (2010). What to do about missing values in time-series cross-section data. *American Journal of Political Science*, 54(2), 561–581. 388

Honaker, J., King, G., Blackwell, M., et al. (2011). Amelia ii : A program for missing data. *Journal of Statistical Software*, 45(7), 1–47. 388

Honerkamp, J. (1993). *Stochastic Dynamical Systems : Concepts, Numerical Methods, Data Analysis*. Wiley. 261

Hooyberghs, J., Mensink, C., Dumont, G., Fierens, F., & Brasseur, O. (2005). A neural network forecast for daily average pm 10 concentrations in belgium. *Atmospheric Environment*, 39(18), 3279–3289. 204

Hopke, P. K. (1991). *Receptor modeling for air quality management*, volume 7. Elsevier. 161

Hopke, P. K. (2010). The application of receptor modeling to air quality data. *Pollution atmosphérique*, (SEP), 91–109. 161

Hopke, P. K., Ramadan, Z., Paatero, P., Norris, G. A., Landis, M. S., Williams, R. W., & Lewis, C. W. (2003). Receptor modeling of ambient and personal exposure samples : 1998 baltimore particulate matter epidemiology-exposure study. *Atmospheric Environment*, 37(23), 3289–3302. 154

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5), 359–366. 201

Hornik, K., Stinchcombe, M., & White, H. (1990). Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks. *Neural networks*, 3(5), 551–560. 201

Hosking, J. R. (1981). Fractional differencing. *Biometrika*, 68(1), 165–176. 107

Hoskins, J. A. (2003). Health effects due to indoor air pollution. *Indoor and Built Environment*, 12(6), 427–433. 10, 26

Hunter, J. S. (1994). Environmetrics : An emerging science. In *Environmental Statistics*, volume 12 of *Handbook of Statistics* (pp. 1 – 7). Elsevier. 7

Hurst, H. (1951). Long-term storage capacity of reservoirs. *Transactions of the American Society of Civil Engineers*, 116(1), 770–799. 106, 113

Hurvich, C. M., Deo, R., & Brodsky, J. (1998). The mean squared error of geweke and porter-hudak's estimator of the memory parameter of a long-memory time series. *Journal of Time Series Analysis*, 19(1), 19–46. 115, 144

Hyndman, R. & Khandakar, Y. (2008). Automatic time series forecasting : the forecast package for r. *Journal of Statistical Software*, 27(1). 230

Hyndman, R., Koehler, A. B., Ord, J. K., & Snyder, R. D. (2008). *Forecasting with exponential smoothing : the state space approach*. Springer Science &amp ; Business Media. 209, 210, 212

Hyndman, R. J. (1995). Highest-density forecast regions for nonlinear and non-normal time series models. *Journal of Forecasting*, 14(5), 431–441. 257

Hyndman, R. J. (1996). Computing and graphing highest density regions. *The American Statistician*, 50(2), 120–126. 257

Hyndman, R. J. & Athanasopoulos, G. (2014). *Forecasting : principles and practice*. OTexts. 230

Hyndman, R. J., Koehler, A. B., Snyder, R. D., & Grose, S. (2002). A state space framework for automatic forecasting using exponential smoothing methods. *International Journal of Forecasting*, 18(3), 439–454. 209, 210, 212, 213

Hyvarinen, A. (1999). Survey on independent component analysis. *Neural computing surveys*, 2(4), 94–128. iv, 152, 157, 160, 161

Hyvärinen, A., Karhunen, J., & Oja, E. (2004). *Independent component analysis*, volume 46. John Wiley & Sons. 153, 157, 160

Hyvärinen, A. & Oja, E. (1997). A fast fixed-point algorithm for independent component analysis. *Neural computation*, 9(7), 1483–1492. 159

Hyvärinen, A. & Oja, E. (1998). Independent component analysis by general nonlinear hebbian-like learning rules. *Signal Processing*, 64(3), 301–313. 160

IARC (2006). *Formaldehyde, 2-Butoxyethanol and 1-tert-Butoxypropan-2-ol*, volume 88. World Health Organization. 16

Ilacqua, V., Dawson, J., Breen, M., Singer, S., & Berg, A. (2015). Effects of climate change on residential infiltration and air pollution exposure. *Journal of Exposure Science and Environmental Epidemiology*. 26

Inoue, T., Taguri, M., & Hoshi, M. (1986). Prediction of nitrogen oxide concentration by a regression model. *Atmospheric Environment (1967)*, 20(12), 2325–2337. 198

Institute of Medicine, U. (2011). *Climate Change, the Indoor Environment, and Health*. National Academies Press. 15, 24

Ionescu, A. (2010). Retour aux sources de pollution atmosphérique : point de vue des scientifiques français. *Pollution Atmospherique*, 5. 152

Islam, S., Bras, R. L., & Rodriguez-Iturbe, I. (1993). A possible explanation for low correlation dimension estimates for the atmosphere. *Journal of applied meteorology*, 32(2), 203–208. 202

Jammalamadaka, S. R. & Sengupta, A. (2001). *Topics in circular statistics*, volume 5. World Scientific. 57

Jayawardena, A. (2014). *Environmental and hydrological systems modelling*. CRC Press. 106

Jensen, R. P., Luo, W., Pankow, J. F., Strongin, R. M., & Peyton, D. H. (2015). Hidden formaldehyde in e-cigarette aerosols. *New England Journal of Medicine*, 372(4), 392–394. 16

Jian, L., Zhao, Y., Zhu, Y.-P., Zhang, M.-B., & Bertolatti, D. (2012). An application of arima model to predict submicron particle concentrations from meteorological factors at a busy roadside in hangzhou, china. *Science of the Total Environment*, 426, 336–345. 198, 199

Jolliffe, I. (2002). *Principal component analysis*. Wiley Online Library. 153

Jolliffe, I. T. & Stephenson, D. B. (2011). *Forecast verification : a practitioner's guide in atmospheric science*. John Wiley & Sons, 2 edition. 149

Jones, A. P. (1999). Indoor air quality and health. *Atmospheric environment*, 33(28), 4535–4564. 10, 11

Jones, M. C. & Sibson, R. (1987). What is projection pursuit ? *Journal of the Royal Statistical Society. Series A (General)*, (pp. 1–37). 160

JORF (2011). Décret n ° 2011-1727 du 2 décembre 2011 relatif aux valeurs-guides pour l'air intérieur pour le formaldéhyde et le benzène. 16

Junger, W. & de Leon, A. P. (2015). Imputation of missing data in time series for air pollutants. *Atmospheric Environment*, 102, 96–104. 381, 382

Jutten, C. & Herault, J. (1991). Blind separation of sources, part i : An adaptive algorithm based on neuromimetic architecture. *Signal processing*, 24(1), 1–10. 157

Kagi, N., Fujii, S., Horiba, Y., Namiki, N., Ohtani, Y., Emi, H., Tamura, H., & Kim, Y. S. (2007). Indoor air quality for chemical and ultrafine particle contaminants from printers. *Building and Environment*, 42(5), 1949–1954. 14

Kantz, H. & Schreiber, T. (2004). *Nonlinear time series analysis*, volume 7. Cambridge university press. 260, 264, 275, 278, 297

Kfoury, A., Ledoux, F., Limem, A., Delmaire, G., Roussel, G., & Courcot, D. (2014). The use of a non negative matrix factorization method combined to pm2. 5 chemical data for a source apportionment study in different environments. In *Air Pollution Modeling and its Application XXIII* (pp. 79–84). Springer. 163

Kfoury, A., Ledoux, F., Roche, C., Delmaire, G., Roussel, G., & Courcot, D. (2016). Pm 2.5 source apportionment in a french urban coastal site under steelworks emission influences using constrained non-negative matrix factorization receptor model. *Journal of Environmental Sciences*. 163

Khlaifi, A. (2007). *Estimation des sources de pollution atmosphérique par modélisation inversée*. PhD thesis, Paris 12. 163

Khokhlov, V. N., Glushkov, A. V., Loboda, N. S., & Bunyakova, Y. Y. (2008). Short-range forecast of atmospheric pollutants using non-linear prediction method. *Atmospheric Environment*, 42(31), 7284–7292. 202

Kim, S. E. & Kumar, A. (2005). Accounting seasonal nonstationarity in time series models for short-term ozone level forecast. *Stochastic Environmental Research and Risk Assessment*, 19(4), 241–248. 200

Kirchner, S., Arenes, J., Cochet, C., Derbez, M., Duboudin, C., Elias, P., Gregoire, A., Jedor, B., Lucas, J., Pasquier, N., et al. (2007a). *Campagne Nationale Logements : état de la qualité de l'air dans les logements français*. Technical report, Observatoire de la Qualité de l'Air Intérieur. 12, 24, 184

Kirchner, S., Arenes, J.-F., Cochet, C., Derbez, M., Duboudin, C., Elias, P., Gregoire, A., Jédor, B., Lucas, J.-P., Pasquier, N., et al. (2007b). État de la qualité de l'air dans les logements français. *Environnement, Risques & Santé*, 6(4), 259–269. 12, 17, 24, 30, 184

Kirchner, S., Buchmann, A., Cochet, C., Dassonville, C., Derbez, M., Leers, Y., Lucas, J.-P., Mandin, C., Ouattara, M., Ramalho, O., et al. (2011). *Qualité d'air intérieur, qualité de vie. 10 ans de recherche pour mieux respirer*. Centre Scientifique et Technique du Bâtiment (CSTB). 26

Kiss, P., Müller, R., & Jánosi, I. (2007). Long-range correlations of extrapolar total ozone are determined by the global atmospheric circulation. *Nonlinear Processes in Geophysics*, 14(4), 435–442. 106

Klepeis, N. E., Nelson, W. C., Ott, W. R., Robinson, J. P., Tsang, A. M., Switzer, P., Behar, J. V., Hern, S. C., Engelmann, W. H., et al. (2001). The national human activity pattern survey (nhaps) : a resource for assessing exposure to environmental pollutants. *Journal of exposure analysis and environmental epidemiology*, 11(3), 231–252. 11

Klinkenberg, B. & Goodchild, M. (1992). The fractal properties of topography : a comparison of methods. *Earth Surface Processes and Landforms*, 17(3), 217–234. 106

Koçak, K., Şaylan, L., & Şen, O. (2000). Nonlinear time series prediction of o 3 concentration in istanbul. *Atmospheric Environment*, 34(8), 1267–1271. 202

Kock, A. B., Teräsvirta, T., et al. (2011). Forecasting with nonlinear time series models. *Oxford Handbook of Economic Forecasting*, (pp. 61–87). 248

Kolmogorov, A. (1956). *Foundations of the Theory of Probability.*, volume Translation Edited by Nathan Morrison. With an Added Bibliography by A.T. Bharucha-Reid. New York, second english edition edition. 87

Kolmogorov, A. N. (1941). The local structure of turbulence in incompressible viscous fluid for very large reynolds numbers. In *Dokl. Akad. Nauk SSSR*, volume 30 (pp. 299–303). 106

Koopmans, L. H. (1995). *The spectral analysis of time series*. Academic press. 86

Koponen, I. K., Asmi, A., Keronen, P., Puhto, K., & Kulmala, M. (2001). Indoor air measurement campaign in helsinki, finland 1999–the effect of outdoor air pollution on indoor air. *Atmospheric Environment*, 35(8), 1465–1477. 22

Krasnopolsky, V. M. & Chevallier, F. (2003a). Some neural network applications in environmental sciences. part i : forward and inverse problems in geophysical remote measurements. *Neural Networks*, 16(3), 321–334. 201

Krasnopolsky, V. M. & Chevallier, F. (2003b). Some neural network applications in environmental sciences. part ii :advancing computational efficiency of environmental numerical models. *Neural Networks*, 16(3), 335–348. 201

Kříž, R. (2014). Chaos in nitrogen dioxide concentration time series and its prediction. In *Nostradamus 2014 : Prediction, Modeling and Analysis of Complex Systems* (pp. 365–376). Springer. 202

Krolzig, H.-M. (2013). *Markov-switching vector autoregressions : Modelling, statistical inference, and application to business cycle analysis*, volume 454. Springer Science & Business Media. 249

Kukkonen, J., Partanen, L., Karppinen, A., Ruuskanen, J., Junninen, H., Kolehmainen, M., Niska, H., Dorling, S., Chatterton, T., Foxall, R., et al. (2003). Extensive evaluation of neural network models for the prediction of no 2 and pm 10 concentrations, compared with a deterministic modelling system and measurements in central helsinki. *Atmospheric Environment*, 37(32), 4539–4550. 203

Kumar, K., Yadav, A., Singh, M., Hassan, H., & Jain, V. (2004). Forecasting daily maximum surface ozone concentrations in brunei darussalam—an arima modeling approach. *Journal of the Air & Waste Management Association*, 54(7), 809–814. 198, 199

Kumar, R., Aggarwal, R., & Sharma, J. (2013). Energy analysis of a building using artificial neural network : A review. *Energy and Buildings*, 65, 352–358. 204

Kumar, U. (2015). An integrated ssa-arima approach to make multiple day ahead forecasts for the daily maximum ambient o 3 concentration. *Aerosol and Air Quality Research*, 15(1), 208–219. 203

Kumar, U. & De Ridder, K. (2010). Garch modelling in association with fft-arima to forecast ozone episodes. *Atmospheric Environment*, 44(34), 4252–4265. 200, 203

Kumar, U. & Jain, V. (2010). Arima forecasting of ambient air pollutants (o3, no, no2 and co). *Stochastic Environmental Research and Risk Assessment*, 24(5), 751–760. 199

Kusiak, A., Li, M., & Zheng, H. (2010). Virtual models of indoor-air-quality sensors. *Applied Energy*, 87(6), 2087–2094. 204

Kuswanto, H. & Sibbertsen, P. (2008). *A study on spurious long memory in nonlinear time series models*. Technical report, Discussion papers//School of Economics and Management of the Hanover Leibniz University. 144

Langer, S., Ramalho, O., Derbez, M., Ribéron, J., Kirchner, S., & Mandin, C. (2016). Indoor environmental quality in french dwellings and building characteristics. *Atmospheric Environment*, 128, 82–91. 19

Lardic, S. & Mignon, V. (1999). La mémoire longue en économie : une revue de la littérature. *Journal de la société française de statistique*, 140(2), 5–48. 116

Lardic, S. & Mignon, V. (2002). *Econométrie des séries temporelles macroéconomiques et financières*. Economica. 116

Larson, T., Gould, T., Simpson, C., Liu, L.-J. S., Claiborn, C., & Lewtas, J. (2004). Source apportionment of indoor, outdoor, and personal pm2. 5 in seattle, washington, using positive matrix factorization. *Journal of the Air & Waste Management Association*, 54(9), 1175–1187. 154

Leconte, G. (2013). Le problème des prédictions dans les sciences expérimentales. *Philonsorbonne*, (7), 81–99. 149

Lee, C.-K. & Lin, S.-C. (2008). Chaos in air pollutant concentration (apc) time series. *Aerosol and Air Quality Research*, 8(4), 381–391. 202

Lee, C.-W. & Hsu, D.-J. (2007). Measurements of fine and ultrafine particles formation in photocopy centers in taiwan. *Atmospheric Environment*, 41(31), 6598–6609. 14

Lee, D. D. & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755), 788–791. 153, 163

Lee, M. H., Rahman, N. H. A., Latif, M. T., Nor, M. E., Kamisan, N. A. B., et al. (2012). Seasonal arima for forecasting air pollution index. *American Journal of Applied Sciences*, 9(4), 570–578. 198

Lee, S., Lam, S., & Fai, H. K. (2001). Characterization of vocs, ozone, and pm 10 emissions from office equipment in an environmental chamber. *Building and Environment*, 36(7), 837–842. 10, 14

Leech, J., Wilby, K., McMullen, E., & Laporte, K. (1996). Enquête sur les profils d'activité humaine au canada : description de la méthodologie et de la population étudiée. *Santé*, 17(3-2000). 11

Letellier, C. & Gilmore, R. (2013). *Topology and Dynamics of Chaos : In Celebration of Robert Gilmore's 70th Birthday*, volume 84. World Scientific. 272

Lévesque, B., Auger, P. L., Bourbeau, J., Duchesne, J.-F., Lajoie, P., & Menzies, D. (2003). Qualité de l'air intérieur. *Environnement et santé publique : fondements et pratiques Volume Chapitre*, 12. 16

Liang, W., Lv, M., & Yang, X. (2016). The effect of humidity on formaldehyde emission parameters of a medium-density fiberboard : Experimental observations and correlations. *Building and Environment*, 101, 110–115. 19

Liebovitch, L. S. & Toth, T. (1989). A fast algorithm to determine fractal dimensions by box counting. *Physics Letters A*, 141(8), 386–390. 111

Little, R. J. & Rubin, D. B. (2014). *Statistical analysis with missing data*. John Wiley & Sons. 377, 378, 384, 386

Liu, W., Zhang, J., Zhang, L., Turpin, B., Weisel, C., Morandi, M., Stock, T., Colome, S., & Korn, L. (2006). Estimating contributions of indoor and outdoor sources to indoor carbonyl concentrations in three urban areas of the united states. *Atmospheric Environment*, 40(12), 2202–2214. 16, 19

Lo, A. W. (1991). Long-term memory in stock market prices. *Econometrica*, 59(5), 1279–1313. 114

Long, C. M., Suh, H. H., Catalano, P. J., & Koutrakis, P. (2001). Using time-and size-resolved particulate data to quantify indoor penetration and deposition behavior. *Environmental Science & Technology*, 35(10), 2089–2099. 23

Lorenz, E. N. (1969). Atmospheric predictability as revealed by naturally occurring analogues. *Journal of the Atmospheric sciences*, 26(4), 636–646. 262, 275

Luukkonen, R., Saikkonen, P., & Teräsvirta, T. (1988). Testing linearity against smooth transition autoregressive models. *Biometrika*, 75(3), 491–499. 248

Malmgren-Hansen, B., Olesen, S., Pommer, K., Winther Funch, L., Pedersen, E., Willum, O., et al. (2011). Emission and evaluation of chemical substances from selected electrical and electronic products. survey of chemical substances in consumer products. survey no. 32-2003. danish environmental protection agency. 17

Mandelbrot, B. (1963). The variation of certain speculative prices. *The Journal of Business*, 36(4), pp. 394–419. 106, 113

Mandelbrot, B. (1972). Statistical methodology for nonperiodic cycles : from the covariance to r/s analysis. In *Annals of Economic and Social Measurement, Volume 1, number 3* (pp. 259–290). NBER. 106, 113, 124

Mandelbrot, B. (1975). *Les objets fractals : forme, hasard, et dimension*. Flammarion. 113

Mandelbrot, B. B. (1965). Une classe de processus stochastiques homothétiques à soi ; application à la loi climatologique de he hurst. *Comptes Rendus hebdomadaires des seances de l'academie des sciences*, 260(12), 3274. 106, 108, 113

Mandelbrot, B. B. (1983). *The fractal geometry of nature*, volume 173. Macmillan. 106, 113, 263

Mandelbrot, B. B. & Van Ness, J. W. (1968). Fractional brownian motions, fractional noises and applications. *SIAM review*, 10(4), 422–437. 106, 108, 113, 371

Mandelbrot, B. B. & Wallis, J. R. (1969). Some long-run properties of geophysical records. *Water resources research*, 5(2), 321–340. 113

Mandin, C., Derbez, M., Lucas, J.-p., Ramalho, O., Gregoire, A., Lethrosne, M., Riberon, J., & Kirchner, S. (2009). Campagne nationale «logements» de l'observatoire de la qualité de l'air intérieur : de la description de la pollution intérieure à sa compréhension. *Pollution atmosphérique*, 51(204), 389–393. 11, 24, 184

Mañé, R. (1981). On the dimension of the compact invariant sets of certain non-linear maps. In *Dynamical systems and turbulence, Warwick 1980* (pp. 230–242). Springer. 135, 260, 266

Manneville, P. (2004). *Instabilités, chaos et turbulence*. Editions Ecole Polytechnique. 260, 264, 267

Marchand, D., Chaventré, F., Ramalho, O., Laffitte, J.-D., Collignan, B., & Weiss, K. (2013). De l'évaluation du risque à la gestion de la crise : le cas du syndrome des bâtiments malsains. *Environnement, Risques & Santé*, 12(4), 325–329. 26

Mardia, K. V. & Jupp, P. E. (2009). *Directional statistics*, volume 494. John Wiley & Sons. 57

Marks-Tarlow, T. (1999). The self as a dynamical system. *Nonlinear Dynamics, Psychology, and Life Sciences*, 3(4), 311–345. 126

Maroni, M., Seifert, B., & Lindvall, T. (1995). *Indoor air quality : a comprehensive reference book*. Elsevier. 10, 13, 15

Martuzevicius, D., Grinshpun, S. A., Lee, T., Hu, S., Biswas, P., Reponen, T., & LeMasters, G. (2008). Traffic-related pm 2.5 aerosol in residential houses located near major highways : indoor versus outdoor concentrations. *Atmospheric Environment*, 42(27), 6575–6585. 154

Massey, D., Kulshrestha, A., Masih, J., & Taneja, A. (2012). Seasonal trends of pm 10, pm 5.0, pm 2.5 & pm 1.0 in indoor and outdoor environments of residential homes located in north-central india. *Building and Environment*, 47, 223–231. 24

Mawhin, J. & Rouche, N. (1973). *Equations differentielles ordinaires, Tome 1 : théorie générale*. Masson et Cie, Paris. 261

Mendell, M. J., Fisk, W. J., Kreiss, K., Levin, H., Alexander, D., Cain, W. S., Girman, J. R., Hines, C. J., Jensen, P. A., Milton, D. K., et al. (2002). Improving the health of workers in indoor environments : priority research needs for a national occupational research agenda. *American journal of public health*, 92(9), 1430–1440. 26

Mendez, M., Blond, N., Blondeau, P., Schoemaecker, C., & Hauglustaine, D. A. (2015). Assessment of the impact of oxidation processes on indoor air pollution using the new time-resolved inca-indoor model. *Atmospheric Environment*, 122, 521–530. 28

Menzies, D. & Bourbeau, J. (1997). Building-related illnesses. *New England Journal of Medicine*, 337(21), 1524–1531. 26

Michalowicz, J. V., Nichols, J. M., & Bucholtz, F. (2013). *Handbook of differential entropy*. CRC Press. 93

Michelot, N., Marchand, C., Ramalho, O., Delmas, V., & Carrega, M. (2013). Monitoring indoor air quality in french schools and day-care centers. *HVAC&R Research*, 19(8), 1083–1089. 30

Mignon, V. (1998). Méthodes d'estimation de l'exposant de hurst. application aux rentabilités boursières. *Économie & prévision*, 132(1-2), 193–214. 113, 114, 115

Milionis, A. & Davies, T. (1994). Regression and stochastic models for air pollution. review, comments and suggestions. *Atmospheric Environment*, 28(17), 2801–2810. 199

Molenberghs, G., Fitzmaurice, G., Kenward, M. G., Tsiatis, A., & Verbeke, G. (2014). *Handbook of Missing Data Methodology*. CRC Press. 377, 378, 387, 388

Molnár, P., Johannesson, S., & Quass, U. (2014). Source apportionment of pm2. 5 using positive matrix factorization (pmf) and pmf with factor selection. *Aerosol Air Quality Res*, 14, 725–733. 155

Morawska, L., He, C., Johnson, G., Jayaratne, R., Salthammer, T., Wang, H., Uhde, E., Bostrom, T., Modini, R., Ayoko, G., et al. (2009). An investigation into the characteristics and formation mechanisms of particles originating from the operation of laser printers. *Environmental science & technology*, 43(4), 1015–1022. 14, 24

Morawska, L., He, C., Wang, H., McGarry, P. D., Salthammer, T., Jayaratne, R., Johnson, G. R., Bostrom, T. E., Modini, R., Uhde, E., et al. (2008). Particle emission from laser printers. In *11th International Conference on Indoor Air Quality and Climate* : 11th International Conference on Indoor Air Quality and Climate. 24

Mosqueron, L., Momas, I., & Le Moullec, Y. (2001). Personal exposure to fine particle in parisian office workers. In *12th world Clean Air and Environment* : 12th world Clean Air and Environment. 24

Mustafaraj, G., Lowry, G., & Chen, J. (2011). Prediction of room temperature and relative humidity by autoregressive linear and nonlinear neural network models for an open office. *Energy and Buildings*, 43(6), 1452–1460. 204

Muth, J. F. (1960). Optimal properties of exponentially weighted forecasts. *Journal of the american statistical association*, 55(290), 299–306. 209

Nadadur, S. S. & Hollingsworth, J. W. (2015). *Air Pollution and Health Effects*. Springer. 26

Nazaroff, W. W. & Singer, B. C. (2004). Inhalation of hazardous air pollutants from environmental tobacco smoke in us residences. *Journal of Exposure Science and Environmental Epidemiology*, 14, S71–S77. 10

Nazaroff, W. W. & Weschler, C. J. (2004). Cleaning products and air fresheners : exposure to primary and secondary air pollutants. *Atmospheric Environment*, 38(18), 2841–2865. 10, 16

Nazaroff, W. W., Weschler, C. J., & Corsi, R. L. (2003). Indoor air chemistry and physics. *Atmospheric Environment*, 37(39), 5451–5453. 11, 12, 345

Neveu, J. (1965). *Bases mathematiques du calcul des probabilites.* Masson et Cie, John Wiley and sons. 191

Newey, W. & West, K. (1987). A simple positive definite, heteroscedasticity and autocor0 relation consistent covariance matrix estimator. *Econometrica*, 55, 103080. 114

Norris, G., Vedantham, R., Duvall, R., Wade, K., Brown, S., Prouty, J., Bai, S., DeWinter, J., & Foley, C. (2015). *EPA Positive Matrix Factorization (PMF) 5.0.* Technical report, US Environmental Protection Agency, Office of Research and Development, Washington, DC. 180

Nuttall, A. H. & Carter, G. C. (1982). Spectral estimation using combined time and lag weighting. *Proceedings of the IEEE*, 70(9), 1115–1125. 100

Ord, J. K., Koehler, A., & Snyder, R. D. (1997). Estimation and prediction for a class of dynamic nonlinear statistical models. *Journal of the American Statistical Association*, 92(440), 1621–1629. 209

Orey, S. (1970). Gaussian sample functions and the hausdorff dimension of level crossings. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 15(3), 249–256. 108

Ouaret, R., Ionescu, A., Petrehus, V., Candau, Y., & Ramalho, O. (2016). Particulate matter variability sources in an open-plan office : comparison of two monitoring campaigns. In *European Aerosol Conference* (pp.1̃). Tours, France. 177

Ouaret, R., Ionescu, A., Ramalho, O., Candau, Y., Gehin, E., & Petrehus, V. (2014a). Analysis of the temporal variability of indoor particulate matter concentrations using Blind Source Separation methods : a comparative study. In *International Aerosol Conference* (pp.1̃). Busan, South Korea. 171, 177, 179

Ouaret, R., Ionescu, A., Ramalho, O., Candau, Y., Petrehus, V., & Labat, L. (2014b). Modelling the time fluctuation of indoor air formaldehyde concentrations : variability structure identification and forecasting using non linear models. In *Indoor Air 2014* (pp. 321–328). 296

Ouaret, R., Ionescu, A., Ramalho, O., Petrehus, V., & Candau, Y. (2014c). Caractérisation et identification de la variabilité temporelle des sources de particules dans un environnement intérieur : approche statistique. In *29ème Congrès Français sur les Aérosols*, volume 1 (pp. 6 pages). Paris, France. 171, 177

Ouaret, R., Ionescu, A., Ramalho, O., Petrehus, V., & Candau, Y. (2014d). Forecasting indoor pollutants concentrations using fast fourier transform (fft) and regime switching models. In *International work-conference on Time-Series analysis*, volume 1 (pp. 52–63). 296

Paatero, P. (1997a). Least squares formulation of robust non-negative factor analysis. *Chemometrics and intelligent laboratory systems*, 37(1), 23–35. 161, 162

Paatero, P. (1997b). A weighted non-negative least squares algorithm for three-way 'parafac' factor analysis. *Chemometrics and Intelligent Laboratory Systems*, 38(2), 223–242. 161, 162

Paatero, P. (1999). The multilinear engine, a table-driven, least squares program for solving multilinear problems, including the n-way parallel factor analysis model. *Journal of Computational and Graphical Statistics*, 8(4), 854–888. 163

Paatero, P. (2000). *User's guide for positive matrix factorization programs PMF2 and PMF3.* Technical report, U.S. Environmental Protection Agency. 162

Paatero, P. & Tapper, U. (1994). Positive matrix factorization : A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5(2), 111–126. 153, 156, 161, 162, 163

Paatero, P., Tapper, U., Aalto, P., & Kulmala, M. (1991). Matrix factorization methods for analysing diffusion battery data. *Journal of Aerosol Science*, 22, S273–S276. 153, 161

Page, J. (2007). *Simulating occupant presence and behaviour in buildings*. PhD thesis. 250

Palma, W. (2007). *Long-memory time series : theory and methods*, volume 662. John Wiley & Sons. 86

Panzhauser, E., Mahdavi, A., & Nagda, N. (1993). A computational model for the prediction and evaluation of formaldehyde concentration in residential buildings. *ASTM SPECIAL TECHNICAL PUBLICATION*, 1205, 197–210. 18

Parkin, T. & Robinson, J. (1992). Analysis of lognormal data. In *Advances in soil science* (pp. 193–235). Springer. 197

Parnell, A. C. (2013). *Climate Time Series Analysis : Classical Statistical and Bootstrap Methods*. Wiley Online Library. 86

Parzen, E. (1975). *Multiple time series : Determining the order of approximating autoregressive schemes*. Technical report, DTIC Document. 195

Pasero, E. G. A. & Mesin, L. (2010). Artificial neural networks to forecast air pollution. 201

Paudel, S., Elmtiri, M., Kling, W. L., Le Corre, O., & Lacarriere, B. (2014). Pseudo dynamic transitional modeling of building heating energy demand using artificial neural network. *Energy and Buildings*, 70, 81–93. 204

Peng, C.-K., Buldyrev, S. V., Havlin, S., Simons, M., Stanley, H. E., & Goldberger, A. L. (1994). Mosaic organization of dna nucleotides. *Physical Review E*, 49(2), 1685. 117

Peng, C.-K., Havlin, S., Stanley, H. E., & Goldberger, A. L. (1995). Quantification of scaling exponents and crossover phenomena in nonstationary heartbeat time series. *Chaos : An Interdisciplinary Journal of Nonlinear Science*, 5(1), 82–87. 117

Percival, D. B. & Walden, A. T. (1993). *Spectral analysis for physical applications*. Cambridge University Press. 100

Perdrix, A., Parat, S., Liaudy, S., & Maître, A. (2005). Syndrome des batiments malsains (sbm). *Revue francophone des laboratoires*, 2005(373), 67–72. 26

Peters, E. E. (1994). *Fractal market analysis : applying chaos theory to investment and economics*, volume 24. John Wiley & Sons. 144

Peters, E. E. (1996). *Chaos and order in the capital markets : a new view of cycles, prices, and market volatility*, volume 1. John Wiley & Sons. 144

Pewsey, A., Neuhäuser, M., & Ruxton, G. D. (2013). *Circular statistics in R*. Oxford University Press. 57

Pisoni, E., Farina, M., Carnevale, C., & Piroddi, L. (2009). Forecasting peak air pollution levels using narx models. *Engineering Applications of Artificial Intelligence*, 22(4), 593–602. 200

Plouvin, M., Limem, A., Puigt, M., Delmaire, G., Roussel, G., & Courcot, D. (2014). Enhanced nmf initialization using a physical model for pollution source apportionment. In *ESANN*. 163

Poincaré, H. (1928). *Oeuvres complètes de Henri Poincaré*. Gaultier-Villars, Paris. 372

Priestley, M. (1982). *Spectral analysis and time series*. Number vol. 1 à 2 in Probability and mathematical statistics. Academic Press. 86, 92

Program, N. T. (2010). *Final report on carcinogens background document for formaldehyde.* Technical Report 10-5981, U.S. Department of Health and Human Services. Public Health Service. National Toxicology Program. 26

Quarteroni, A. M., Sacco, R., & Saleri, F. (2008). *Méthodes numériques : algorithmes, analyse et applications.* Springer Science & Business Media. 378, 379

R (2015). *R : A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. 163, 278, 306, 388

Raga, G. & Le Moyne, L. (1996). On the nature of air pollution dynamics in mexico city i. nonlinear analysis. *Atmospheric Environment*, 30(23), 3987–3993. 199, 201

Ramalho, . O. (2004). Technical report. 13

Ramalho, O., Lucas, J.-P., Mandin, C., Derbez, M., & Kirchner, S. (2012). Niveaux de particules dans les environnements intérieurs en france. *Pollution Atmosphérique*, (spécial), 37–42. 24

Rangarajan, G. & Ding, M. (2003). *Processes with long-range correlations : Theory and applications*, volume 621. Springer Science & Business Media. 86

Rao, C. & Patil, G. (1994). *Handbook of Statistics : Environmental statistics ; edited by G.P. Patil and C.R. Rao.* Number vol. 12. North-Holland Publishing Company. 86

Ribéron, J. & O'Kelly, P. (2002). Maria an experimental tool at the service of indoor air quality in housing sector. In *Proceedings of the 9th International Conference of Indoor Air Quality and Climate*, volume 3 (pp. 191–195). 33

Richardson, L. F. (1922). *Forms Whereon to Write the Numerical Calculations Described in Weather Prediction by Numerical Process.* Nabu Public Domain Reprints. 190

Robeson, S. & Steyn, D. (1990). Evaluation and comparison of statistical forecast models for daily maximum ozone concentrations. *Atmospheric Environment. Part B. Urban Atmosphere*, 24(2), 303–312. 198

Robinson, J. & Nelson, W. (1995). National human activity pattern survey data base. *USEPA, Research Triangle Park, NC.* 11

Robinson, P. M. (2003). *Time series with long memory.* Oxford University Press. 86

Roll, J. (1981). *Contribution à la proprioception musculaire, à la perception et au contrôle du mouvement chez l'homme.* Thesis, Th. Sci. nat. Aix-Marseille 1. 152

Roulet, C.-A. (2004). *Santé et qualité de l'environnement intérieur dans les bâtiments*, volume 22. PPUR presses polytechniques. 15

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3), 581–592. 384

Rubin, D. B. (2004). *Multiple imputation for nonresponse in surveys*, volume 81. John Wiley & Sons. 377, 378, 384, 386, 388

Ruelle, D. (1990). The claude bernard lecture, 1989. deterministic chaos : the science and the fiction. In *Proceedings of the Royal Society of London A : Mathematical, Physical and Engineering Sciences*, volume 427 (pp. 241–248). : The Royal Society. 202

Salcedo, R., Ferraz, M. A., Alves, C., & Martins, F. (1999). Time-series analysis of air pollution data. *Atmospheric Environment*, 33(15), 2361–2372. 199

Salthammer, T., Fuhrmann, F., Kaufhold, S., Meyer, B., & Schwarz, A. (1995). Effects of climatic parameters on formaldehyde concentrations in indoor air. *Indoor Air*, 5(2), 120–128. 19

Salthammer, T., Mentese, S., & Marutzky, R. (2010). Formaldehyde in the indoor environment. *Chemical Reviews*, 110(4), 2536–2572. 14, 16, 17, 18, 19, 335

Sami, M., Waseem, A., Jafri, Y. Z., Shah, S. H., Khan, M. A., Akbar, S., Siddiqui, M. A., & Murtaza, G. (2012). Prediction of the rate of dust fall in quetta city, pakistan using seasonal arima (sarima) modeling. *International Journal of Physical Sciences*, 7(10), 1713–1725. 199

Samorodnitsky, G. (2007). Long range dependence. *Foundations and Trends® in Stochastic Systems*, 1(3), 163–257. 115

Saporta, G. (1990). *Probabilités, analyse des données et statistique*. Technip. 153

Saraga, D., Pateraki, S., Papadopoulos, A., Vasilakos, C., & Maggos, T. (2011). Studying the indoor air quality in three non-residential environments of different use : A museum, a printery industry and an office. *Building and Environment*, 46(11), 2333–2341. 14

Sarwar, G., Corsi, R., Kimura, Y., Allen, D., & Weschler, C. J. (2002). Hydroxyl radicals in indoor environments. *Atmospheric Environment*, 36(24), 3973–3988. 28

Sauer, T., Yorke, J., & Casdagli, M. (1991). Embedology. *Journal of Statistical Physics*, 65(3-4), 579–616. 260, 266

Sawaragi, Y., Soeda, T., Tamura, H., Yoshimura, T., Ohe, S., Chujo, Y., & Ishihara, H. (1979). Statistical prediction of air pollution levels using non-physical models. *Automatica*, 15(4), 441–451. 198

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. CRC press. 377, 384

Schafer, J. L. (1999). Multiple imputation : a primer. *Statistical methods in medical research*, 8(1), 3–15. 377

Schafer, J. L. (2003). Multiple imputation in multivariate problems when the imputation and analysis models differ. *Statistica Neerlandica*, 57(1), 19–35. 377

Schafer, J. L. & Graham, J. W. (2002). Missing data : our view of the state of the art. *Psychological methods*, 7(2), 147. 377, 378

Schepers, H. E., Van Beek, J. H., & Bassingthwaighte, J. B. (1992). Four methods to estimate the fractal dimension from self-affine signals (medical application). *Engineering in Medicine and Biology Magazine, IEEE*, 11(2), 57–64. 106

Schmittbuhl, J., Vilotte, J.-P., & Roux, S. (1995). Reliability of self-affine measurements. *Physical Review E*, 51(1), 131. 106

Schripp, T., Mulakampilly, S., Delius, W., Uhde, E., Wensing, M., Salthammer, T., Kreuzig, R., Bahadir, M., Wang, L., & Morawska, L. (2009). Comparison of ultrafine particle release from hardcopy devices in emission test chambers and office rooms. *Gefahrstoffe–Reinhaltung der Luft*, 69(3), 71–76. 14, 24

Schwarz, G. et al. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2), 461–464. 195

Seifert, B. & Ullrich, D. (1987). Methodologies for evaluating sources of volatile organic chemicals (voc) in homes. *Atmospheric Environment (1967)*, 21(2), 395–404. 12

Seinfeld, J. H. & Pandis, S. N. (2012). *Atmospheric chemistry and physics : from air pollution to climate change*. John Wiley & Sons. 19

Seuront, L. (2009). *Fractals and multifractals in ecology and aquatic science*. CRC Press. 106

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27(3), 379–423. 93

Shi, J. P. & Harrison, R. M. (1997). Regression modelling of hourly no x and no 2 concentrations in urban air in london. *Atmospheric Environment*, 31(24), 4081–4094. 198

Shumway, R. H. & Stoffer, D. S. (2013). *Time series analysis and its applications*. Springer Science & Business Media. 230

Simpson, R. & Layton, A. (1983). Forecasting peak ozone levels. *Atmospheric Environment (1967)*, 17(9), 1649–1654. 198

Sinclair, J., Psota-Kelty, L., Weschler, C., & Shields, H. (1990). Measurement and modeling of airbone concentrations and indoor surface accumulation rates of ionic substances at neenah, wisconsin. *Atmospheric Environment. Part A. General Topics*, 24(3), 627–638. 22

Slini, T., Kaprara, A., Karatzas, K., & Moussiopoulos, N. (2006). Pm 10 forecasting for thessaloniki, greece. *Environmental Modelling & Software*, 21(4), 559–565. 204

Slini, T., Karatzas, K., & Moussiopoulos, N. (2002). Statistical analysis of environmental data as the basis of forecasting : an air quality application. *Science of the total environment*, 288(3), 227–237. 199

Small, M. (2005). *Applied nonlinear time series analysis : applications in physics, physiology and finance*, volume 52. World Scientific. 264

Sousa, S., Martins, F., Alvim-Ferraz, M., & Pereira, M. C. (2007). Multiple linear regression and artificial neural networks based on principal components to predict ozone concentrations. *Environmental Modelling & Software*, 22(1), 97–103. 204

Sousa, S., Pires, J., Martins, F., Pereira, M., & Alvim-Ferraz, M. (2009). Potentialities of quantile regression to predict ozone concentrations. *Environmetrics*, 20(2), 147–158. 204

Spengler, J., McCarthy, J., & Samet, J. (2001). *Indoor Air Quality Handbook*. McGraw-Hill Education. 16, 26

Spengler, J. D. & Sexton, K. (1983). Indoor air pollution : a public health perspective. *Science*, 221(4605), 9–17. 10

Spivak, M. (1970-1975). *A comprehensive introduction to differential geometry*, volume I-V. Berkeley : Publish or Perish Inc, 3 edition. 372, 373

Spivak, M. (1979). *Differential Geometry*. Publish or Perish. 372

Stolwijk, J. A. (1992). Risk assessment of acute health and comfort effects of indoor air pollution. *Annals of the New York Academy of Sciences*, 641(1), 56–62. 10

Sundell, J. (2004). On the history of indoor air quality and health. *Indoor air*, 14(s7), 51–58. 10

Takens, F. (1981). Detecting strange attractors in turbulence. In D. Rand & L.-S. Young (Eds.), *Dynamical Systems and Turbulence, Warwick 1980*, volume 898 of *Lecture Notes in Mathematics* (pp. 366–381). Springer Berlin Heidelberg. 135, 260, 265, 266, 275

Tang, X., Bai, Y., Duong, A., Smith, M. T., Li, L., & Zhang, L. (2009). Formaldehyde in china : Production, consumption, exposure levels, and health effects. *Environment international*, 35(8), 1210–1224. 19

Taqqu, M. S. (1988). Self-similar processes. *Encyclopedia of Statistical Sciences*. 86

Taqqu, M. S., Teverovsky, V., & Willinger, W. (1995). Estimators for long-range dependence : an empirical study. *Fractals*, 3(04), 785–798. 113, 126

Taylor, C. C. & Taylor, S. J. (1991). Estimating the dimension of a fractal. *Journal of the Royal Statistical Society. Series B (Methodological)*, (pp. 353–364). 111

Teräsvirta, T. (1994). Specification, estimation, and evaluation of smooth transition autoregressive models. *Journal of the american Statistical association*, 89(425), 208–218. 244, 246, 248

Teräsvirta, T., Van Dijk, D., & Medeiros, M. C. (2005). Linear models, smooth transition autoregressions, and neural networks for forecasting macroeconomic time series : A reexamination. *International Journal of Forecasting*, 21(4), 755–774. 201

Thatcher, T. L. & Layton, D. W. (1995). Deposition, resuspension, and penetration of particles within a residence. *Atmospheric Environment*, 29(13), 1487–1497. 23, 335

Thatcher, T. L., Lunden, M. M., Revzan, K. L., Sextro, R. G., & Brown, N. J. (2003). A concentration rebound method for measuring particle penetration and deposition in the indoor environment. *Aerosol Science & Technology*, 37(11), 847–864. 20, 335

Theiler, J. (1990). Estimating fractal dimension. *JOSA A*, 7(6), 1055–1073. 106

Tol, R. S. (1996). Autoregressive conditional heteroscedasticity in daily temperature measurements. *Environmetrics*, 7(1), 67–75. 200, 205

Tong, H. (1977). Some comments on the canadian lynx data. *Journal of the Royal Statistical Society. Series A (General)*, 140(4), pp. 432–436. 242

Tong, H. (1978). On a threshold model. In S. . Noordhoff (Ed.), *Pattern Recognition and Signal Processing*, number 29 in Series E : Applied Sc. (pp. 575–586). : NATO ASI. 244

Tong, H. (1983). *Threshold models in non-linear time series analysis*. Lecture notes in statistics. Springer-Verlag. 242, 246

Tong, H. (1993). *Non-linear Time Series : A Dynamical System Approach*. Non-Linear Time Series. Clarendon Press. 242, 247, 264

Tong, H. & Lim, K. S. (1980). Threshold autoregression, limit cycles and cyclical data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 42(3), pp. 245–292. 242, 244, 247

Tovalin-Ahumada, H., Whitehead, L., & Blanco, S. (2007). Personal exposure to pm 2.5 and element composition–a comparison between outdoor and indoor workers from two mexican cities. *Atmospheric Environment*, 41(35), 7401–7413. 22

Treut, H. L. (2009). Qui fait la pluie et le beau temps ? *La revue pour l'histoire du CNRS*, (24). 190

Tricot, C. (1999). *Courbes et dimension fractale*. Springer Science & Business Media. 109

Triebig, G., Schaller, K.-H., Beyer, B., Müller, J., & Valentin, H. (1989). Formaldehyde exposure at various workplaces. *Science of the total environment*, 79(2), 191–195. 18

Uhde, E., He, C., & Wensing, M. (2006). Characterization of ultra-fine particle emission from a laser printer. In *Proc. Int. Conf. Healthy Building*, volume 2 (pp. 479–482). : Citeseer. 14, 24

Ulens, T., Millet, S., Van Ransbeeck, N., Van Weyenberg, S., Van Langenhove, H., & Demeyer, P. (2014). The effect of different pen cleaning techniques and housing systems on indoor concentrations of particulate matter, ammonia and greenhouse gases (co 2, ch 4, n 2 o). *Livestock Science*, 159, 123–132. 23

Uria-Tellaetxe, I. & Carslaw, D. C. (2014). Conditional bivariate probability function for source identification. *Environmental Modelling & Software*, 59, 1–9. 74

van Netten, C., Shirtliffe, C., & Svec, J. (1989). Temperature and humidity dependence of formaldehyde release from selected building materials. *Bulletin of Environmental Contamination and Toxicology*, 42(4), 558–565. 18

Varotsos, C. & Kirk-Davidoff, D. (2006). Long-memory processes in ozone and temperature variations at the region 60 s–60 n. *Atmospheric Chemistry and Physics*, 6(12), 4093–4100. 106

Vautard, R. & Ghil, M. (1989). Singular spectrum analysis in nonlinear dynamics, with applications to paleoclimatic time series. *Physica D : Nonlinear Phenomena*, 35(3), 395–424. 135, 213

Vautard, R., Yiou, P., & Ghil, M. (1992). Singular-spectrum analysis : A toolkit for short, noisy chaotic signals. *Physica D : Nonlinear Phenomena*, 58(1), 95–126. 135, 213, 260

Vialar, T. (2005). *Dynamiques non linéaires chaotiques en finance et économie*. Economica. 262, 264, 266

Viana, M., Kuhlbusch, T., Querol, X., Alastuey, A., Harrison, R., Hopke, P., Winiwarter, W., Vallius, M., Szidat, S., Prévôt, A., et al. (2008). Source apportionment of particulate matter in europe : a review of methods and results. *Journal of Aerosol Science*, 39(10), 827–849. 154, 339

Viljoen, H. & Nel, D. (2010). Common singular spectrum analysis of several time series. *Journal of Statistical Planning and Inference*, 140(1), 260–267. 135

Vincent, D., Annesi, I., Festy, B., & Lambrozo, J. (1997). Ventilation system, indoor air quality, and health outcomes in parisian modern office workers. *Environmental research*, 75(2), 100–112. 24

Vingarzan, R. & Taylor, B. (2003). Trend analysis of ground level ozone in the greater vancouver/fraser valley area of british columbia. *Atmospheric Environment*, 37(16), 2159–2171. 199

Walden, A. (1989). Accurate approximation of a 0th order discrete prolate spheroidal sequence for filtering and data tapering. *Signal Processing*, 18(3), 341–348. 100

Wallace, L. A., Emmerich, S. J., & Howard-Reed, C. (2004). Source strengths of ultrafine and fine particles due to cooking with a gas stove. *Environmental Science & Technology*, 38(8), 2304–2311. 10

Wallace, L. A., Mitchell, H., T O'Connor, G., Neas, L., Lippmann, M., Kattan, M., Koenig, J., Stout, J. W., Vaughn, B. J., Wallace, D., et al. (2003). Particle concentrations in inner-city homes of children with asthma : the effect of smoking, cooking, and outdoor pollution. *Environmental health perspectives*, 111(9), 1265. 20

Wang, H.-K., Huang, C.-H., Chen, K.-S., & Peng, Y.-P. (2010). Seasonal variation and source apportionment of atmospheric carbonyl compounds in urban kaohsiung, taiwan. *Aerosol and Air Quality Resarch*, 10(6), 559–570. 19

Wang, X.-K. & Lu, W.-Z. (2006). Seasonal variation of air pollution index : Hong kong case study. *Chemosphere*, 63(8), 1261–1272. 198

Wasserstein, R. L. & Lazar, N. A. (2016). The asa's statement on p-values : context, process, and purpose. *The American Statistician*, (just-accepted), 00–00. 254

Watanabe, M. & Yamaguchi, K. (2003). *The EM algorithm and related statistical models*. CRC Press. 377, 378

Weisel, C. P., Zhang, J., Turpin, B., Morandi, M., Colome, S., Stock, T., Spektor, D., Korn, L., Winer, A., Kwon, J., et al. (2005). Relationships of indoor, outdoor, and personal air (riopa). part i. collection methods and descriptive analyses. *Research Report (Health Effects Institute)*, (130 Pt 1), 1–107. 81

Weng, Y.-C., Chang, N.-B., & Lee, T. (2008). Nonlinear time series analysis of ground-level ozone dynamics in southern taiwan. *Journal of environmental management*, 87(3), 405–414. 199

Wensing, M., Pinz, G., Bednarek, M., Schripp, T., Uhde, E., & Salthammer, T. (2006). Particle measurement of hardcopy devices. In *Proceedings of the Healthy Building 2006 Conference, Lisbon, Portugal*, volume 2 (pp. 4–8). 14

Wensing, M., Schripp, T., Uhde, E., & Salthammer, T. (2008). Ultra-fine particles release from hardcopy devices : sources, real-room measurements and efficiency of filter accessories. *Science of the Total Environment*, 407(1), 418–427. 24

Westmoreland, E. J., Carslaw, N., Carslaw, D. C., Gillah, A., & Bates, E. (2007). Analysis of air quality within a street canyon using statistical and dispersion modelling techniques. *Atmospheric Environment*, 41(39), 9195–9205. 57

WHO-Europe (2000). *Air quality guidelines for Europe*. Technical report, World Health Organization. 15

Wickham, H. (2009). *ggplot2 : Elegant Graphics for Data Analysis*. Springer-Verlag New York. 306

Winters, P. R. (1960). Forecasting sales by exponentially weighted moving averages. *Management Science*, 6(3), 324–342. 209

Wisthaler, A., Tamás, G., Wyon, D. P., Strøm-Tejsen, P., Space, D., Beauchamp, J., Hansel, A., Märk, T. D., & Weschler, C. J. (2005). Products of ozone-initiated chemistry in a simulated aircraft environment. *Environmental Science & Technology*, 39(13), 4823–4832. 17

Wold, S., Esbensen, K., & Geladi, P. (1987). Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3), 37–52. 153

Wolkoff, P. (2013). Indoor air pollutants in office environments : assessment of comfort, health, and performance. *International journal of hygiene and environmental health*, 216(4), 371–394. 13, 14

Wolkoff, P. & Nielsen, G. D. (2010). Non-cancer effects of formaldehyde and relevance for setting an indoor air guideline. *Environment international*, 36(7), 788–799. 16, 18, 19

Wolkoff, P., Wilkins, C., Clausen, P., & Nielsen, G. (2006). Organic compounds in office environments–sensory irritation, odor, measurements and the role of reactive chemistry. *Indoor air*, 16(1), 7–19. 14

Wolkoff, P., Wilkins, C. K., Clausen, P. A., & Larsen, K. (1993). Comparison of volatile organic compounds from processed paper and toners from office copiers and printers : methods, emission rates, and modeled concentrations. *Indoor air*, 3(2), 113–123. 14

Wood, S. (2006). *Generalized additive models : an introduction with R*. CRC press. 57

Yakovleva, E., Hopke, P. K., & Wallace, L. (1999). Receptor modeling assessment of particle total exposure assessment methodology data. *Environmental Science & Technology*, 33(20), 3645–3652. 154

Yiou, P., Baert, E., & Loutre, M. (1996). Spectral analysis of climate data. *Surveys in Geophysics*, 17(6), 619–663. 86

Yu, K., Cheung, Y., Cheung, T., & Henry, R. C. (2004). Identifying the impact of large urban airports on local air quality by nonparametric regression. *Atmospheric Environment*, 38(27), 4501–4507. 57

Zeghnoun, A., Dor, F., & Grégoire, A. (2010). *Description du budget espace temps et estimation de l'exposition de la population française dans son logement*. Technical report, Institut de veille sanitaire, Observatoire de la qualité de l'air intérieur. 11

Zhang, Y., Luo, X., Wang, X., Qian, K., & Zhao, R. (2007). Influence of temperature on formaldehyde emission parameters of dry building materials. *Atmospheric Environment*, 41(15), 3203–3216. 18

Zhao, W., Hopke, P. K., Gelfand, E. W., & Rabinovitch, N. (2007). Use of an expanded receptor model for personal exposure analysis in schoolchildren with asthma. *Atmospheric Environment*, 41(19), 4084–4096. 154

Zhong, W. (2013). *An introduction to healthcare and medical textiles*. DEStech Publications, Inc. 16

Zhu, Y., Yang, L., Meng, C., Yuan, Q., Yan, C., Dong, C., Sui, X., Yao, L., Yang, F., Lu, Y., et al. (2015). Indoor/outdoor relationships and diurnal/nocturnal variations in water-soluble ion and pah concentrations in the atmospheric pm 2.5 of a business office area in jinan, a heavily polluted city in china. *Atmospheric Research*, 153, 276–285. 24

TABLE DES FIGURES

LISTE DES TABLEAUX

ANNEXE A

CONTRIBUTIONS

Communications

OUARET R., IONESCU A., RAMALHO O., PETREHUS V., CANDAU Y. (2014), *Caractérisation et identification de la variabilité temporelle des sources de particules dans un environnement intérieur : approche statistique*, 29ème Congrès Français sur les Aérosols, Paris, 22-23 janvier 2014, 6 p.

OUARET R., IONESCU A., RAMALHO O., PETREHUS V., CANDAU Y. (2014), *Identification des sources de variabilité des niveaux de polluants par analyse en composantes indépendantes et par estimation de mélange des données de l'air intérieur*, Journées Interdisciplinaires de la Qualité de l'Air, Villeneuve d'Ascq, 10-11 février 2014, Paris, 18 p.

OUARET R., IONESCU A., RAMALHO O., CANDAU Y., PETREHUS V. (2014), *Forecasting indoor pollutants concentrations using Fast Fourier Transform (FFT) and Regime Switching Models*, ITISE 2014, International work-conference on Time-Series analysis, June 25-27 2014, Granada (Spain), Vol. 1, 52-63.

OUARET R., IONESCU A., RAMALHO O., CANDAU Y., PETREHUS V., LABAT L. (2014), *Modelling the time fluctuation of indoor air formaldehyde concentrations : variability structure identification and forecasting using non-linear models*, Proc. Indoor Air 2014, July 7-12 2014, Hong Kong, Vol. V, 321-328.

OUARET R., IONESCU A., RAMALHO O., CANDAU Y., GEHIN E., PETREHUS V. (2014), *Analysis of the temporal variability of indoor particulate matter concentrations using Blind Source Separation methods : a comparative study*, International Aerosol Conference 2014, Aug. 28 – Sep. 2, 2014, Busan, South Korea, poster.

OUARET R., IONESCU A., PETREHUS V., CANDAU Y., RAMALHO O. (2016), *Forecasting particulate matter concentrations in an indoor environment*, 22nd European Aerosol Conference, Tours (France), September 4th - 9th 2016.

OUARET R., IONESCU A., PETREHUS V., CANDAU Y., RAMALHO O. (2016), *Particulate matter variability sources in an open-plan office : comparison of two monitoring campaigns*, 22nd European Aerosol Conference, Tours (France), September 4th - 9th 2016.

OUARET R., IONESCU A., RAMALHO O, CANDAU Y. (2017), *Indoor air pollutant sources using blind source separation methods,* 25th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, Bruges (Belgium), April 26th - 28th 2017.

Rapports

RAMALHO O., IONESCU A., OUARET R., LABAT L., CANDAU Y. **TRIBU**. *Suivi dynamique en temps réel de la qualité de l'air intérieur dans un environnement de bureaux - Contributions des sources et Modèle prévisionnel*, CSTB, Mai 2014, 122 pages. [DSC-Obs-QAI/2014-042R].

RAMALHO O., IONESCU A., OUARET R., LE PONNER E., CANDAU Y. **TRIBU**. *Suivi dynamique en temps réel de la qualité de l'air intérieur dans un environnement de bureaux - Contributions des sources et Modèle prévisionnel*, CSTB, Mars 2016, 151 pages. [DSC-OBS-QAI/2016-017].

Article

OUARET R., IONESCU A., PETREHUS V., CANDAU Y., RAMALHO O. (2016), Spectral Band Decomposition Coupled with Nonlinear Models for Indoor Formaldehyde Concentration Forecasting, *Stochastic Environmental Research and Risk Assessment*, soumis.

ANNEXE B

VUE GLOBALES DES *MICRO*-ENVIRONNEMENTS

B.1 Espace de bureaux

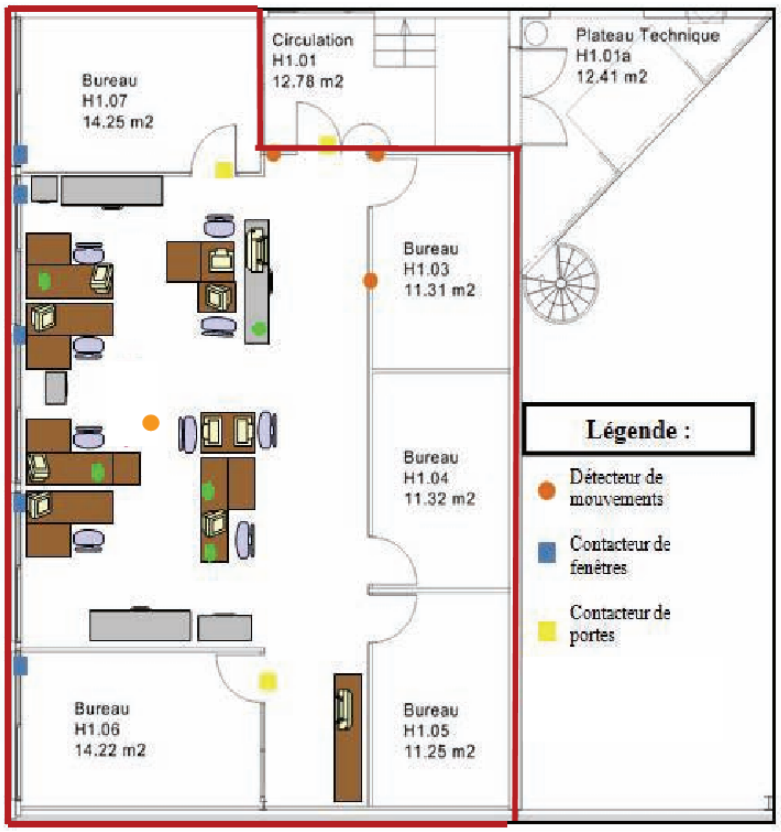


FIGURE B.1.1 – Plan de l'espace de bureaux ($132m^2$, 364 m^3)

FIGURE B.1.2 – Plan de l'espace de bureaux ($132m^2$, 364 m^3)

B.2 Bureau individuel

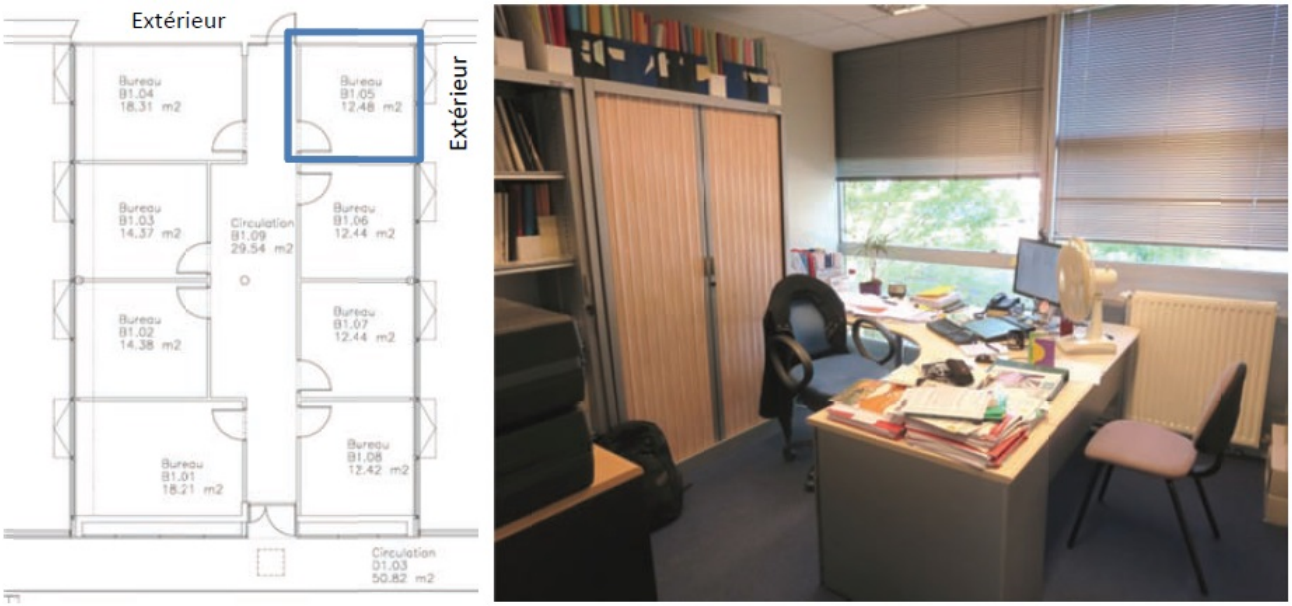


FIGURE B.2.1 – Plan et vue du bureau individuel depuis la porte d'entrée.

B.3 Maison expérimentale (MARIA)

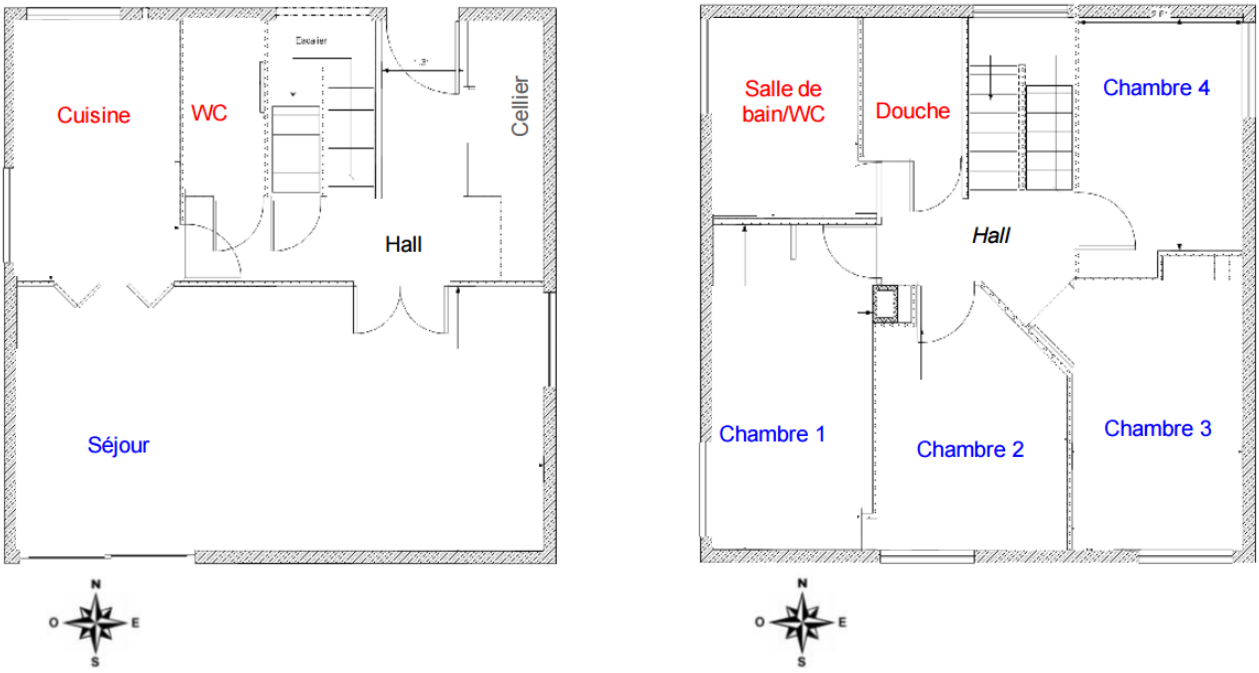


FIGURE B.3.1 – Plan de la maison expérimentale MARIA.

ANNEXE C

STATISTIQUES SUPPLÉMENTAIRES DES PARAMÈTRES DE LA QAI

C.1 Bureau individuel

C.1.1 Les PM

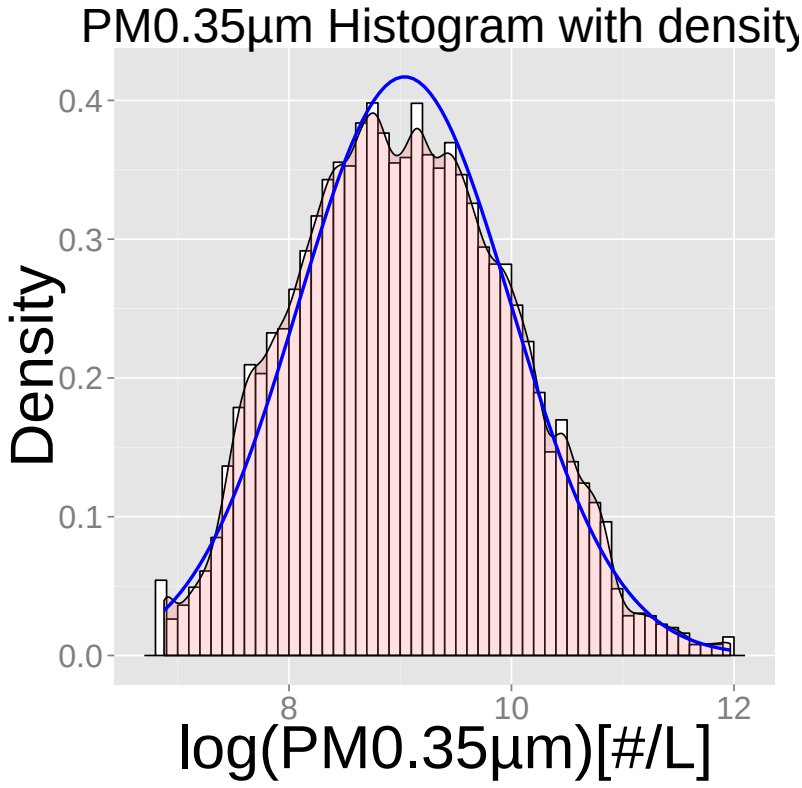


FIGURE C.1.1 – *Distribution de probabilité de la concentration en nombre de particules ($\#L^{-1}$) pour le diamètre 0.35 μm. Les valeurs sont exprimée en logarithme. La courbe lisse (en bleu) est la densité de probabilité de la loi normal ayant les mêmes paramètres que les concentrations de cette fraction.*

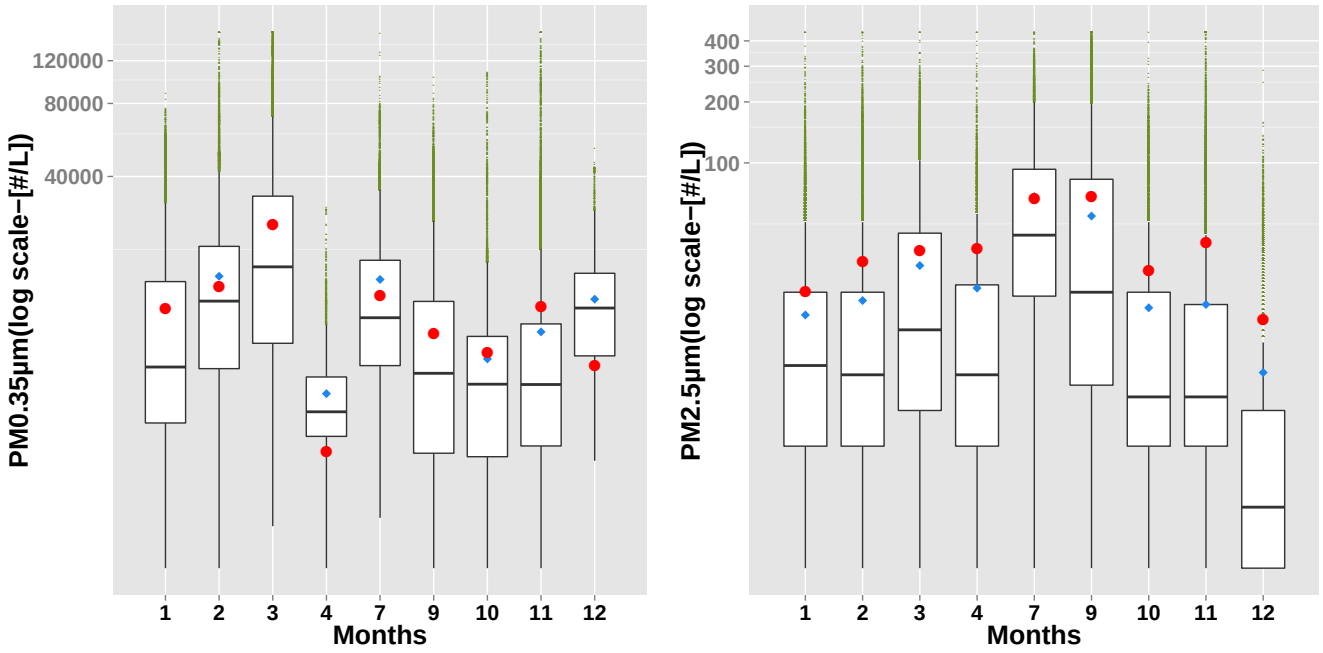


FIGURE C.1.2 – Distributions mensuelles de la concentration de particules dans le bureau individuel. Exemples de particules de diamètre médian 0.35 et 2.5 μm. Les valeurs sont exprimées en nombre de particules par litre ($\#L^{-1}$). Le point rond rouge est l'écart type calculé pour chaque mois. Le point losange bleu est la moyenne mensuelle.

C.1.2 Variabilité des Hydrocarbures Aromatiques Polycycliques (HAP)

La concentration totale en Hydrocarbures Aromatiques Polycycliques (HAP) est estimée grâce à un module supplémentaire associé au compteur de particules. Les mesures ont été réalisées toutes les minutes entre juillet 2010 et juillet 2011, mais nous considérons comme valides les données entre le 06 octobre 2010 et le 04 avril 2011. La mesure fournit uniquement la concentration moyenne des HAPs totaux particulaires et elle est exprimée en $ng \cdot m^{-3}$. Bien que les HAPs soient majoritairement d'origine extérieure et liés au trafic, les mesures extérieures n'ont pas été effectuées durant cette période.

En tout, nous disposons de 235822 observations (dont 9.5% des données manquantes) de la concentration moyenne en HAP totaux. Le Tableau C.1.1 fournit les premiers indicateurs statistiques estimés durant toute la période de mesure (06/10/2010 -04/04/2011).

On constate à première vue que la variabilité du HAP est dominée par des valeurs faibles ($Q_1 = 1.08\,ng{\cdot}m^{-3}$ et $Q_3 = 5.38\,ng{\cdot}m^{-3}$) avec plusieurs épisodes de fortes variabilités (écart-type=$7.04\,ng{\cdot}m^{-3}$). La densité de probabilité présente une queue de distribution leptokurtique (Kurtosis=38.41) et et étalée vers la droite (Skewness =4.9). Au regard de ces caractéristiques, on peut penser que la distribution serait de type log-normale. Cette variabilité exprimée en termes de coefficient de variation géométrique est de l'ordre de 60% (C_v pour la variable log (HAP)).

TABLE C.1.1 – Statistiques des concentrations des HAPs totaux durant la période de 06 octobre 2010 au 04 avril 2011 avec un pas de temps d'une minute. Les mesures sont issues de la campagne du bureau individuel.

x	n	Mini	Max	Q_1	Q_3	$\bar{x}$	Médiane	C_v	σ_{hap}
HAP	235822	0	109.28	1.08	5.37	4.64	2.38	1.51	7.04
Log(HAP)	235822	0	4.7	7.32	1.85	1.34	1.21	0.6	0.8

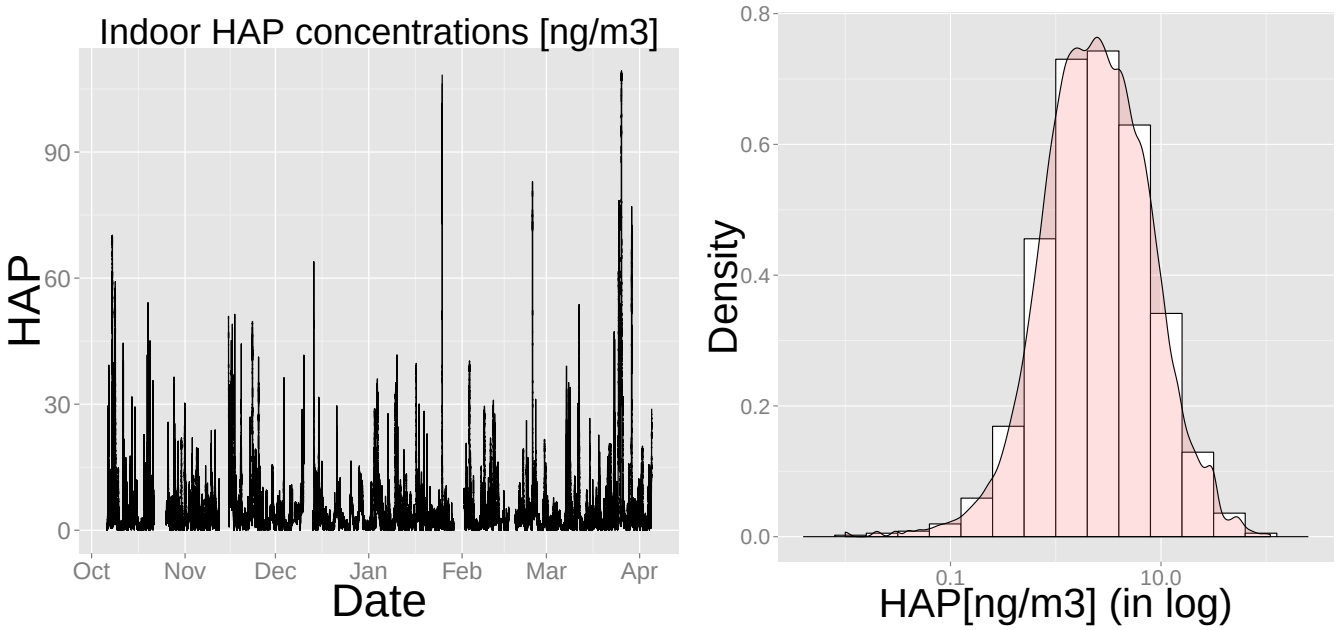
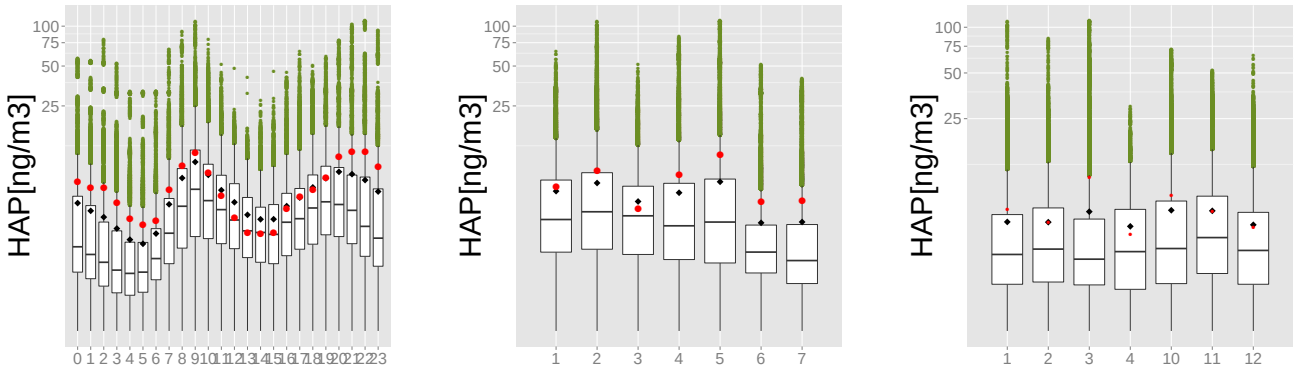


FIGURE C.1.3 – *Fluctuations de la concentration en HAP totaux dans le bureau individuel (à gauche), son histogramme et l'estimation de la densité de probabilité (échelle logarithmique décimale) (à droite). Les données représentées sont du 06 octobre 2010 au 04 avril 2011 au pas de temps d'une minute* $\left[ng \cdot m^{-3}\right]$.

Un fait marquant dans les fluctuations des HAP est la distribution des valeurs extrêmes. En effet, sur l'ensemble de la série, seul 1% des données ont des valeurs entre 33 et 109.3 $\mathrm{ng \cdot m^{-3}}$, 5% entre 16 $\mathrm{ng \cdot m^{-3}}$; et 80% des données sont inférieures à 8 $\mathrm{ng \cdot m^{-3}}$. Ceci indique qu'en termes de la variabilité des sources, plusieurs peuvent y contribuer pour former les 80% des valeurs inférieures à 8 $\mathrm{ng \cdot m^{-3}}$, mais seules des sources épisodiques ($\sim$ 1 h/Jour) contribuent pour constituer les événements de faible probabilité. Afin de voir si l'existence d'un profil temporel serait en mesure d'expliquer de telles observations, la Figure C.1.4 montre les distributions diurne, hebdomadaire et mensuelle de la concentration en HAP.



(a) *Profil journalier HAP*

(b) *Profil hebdomadaire de HAP. Le 1 représente lundi, le 7 représente Dimanche*

(c) *Profil mensuel de HAP avec 1 pour Janvier et le 12 pour décembre.*

FIGURE C.1.4 – *Distribution de la concentration en HAP totaux dans le bureau individuel pendant la période* du 06/10/2010 au 04/04/2011. *Les concentrations ont été exprimées en logarithme.*

Clairement, l'aspect quasi-sinusoïdal de la distribution diurne montre l'importance de l'émission des HAPs durant certaines heures de la journée. En particulier, on peut observer deux pics à 10 h et 23 h correspondant aux valeurs les plus élevées et aux fortes variabilités ($\sigma_{hap} \approx 11.3\,\mathrm{ng}\cdot\mathrm{m}^{-3}$). Le niveau de fluctuation le plus bas se produit entre 13 h et 15 h avec une médiane de $1.3\,\mathrm{ng}\cdot\mathrm{m}^{-3}$ et un maximum de $45\,\mathrm{ng}\cdot\mathrm{m}^{-3}$. En ce qui concerne la variation hebdomadaire, la médiane et l'écart-type observés durant les jours ouvrés sont en moyenne plus élevés que ceux observés les week-ends et les mercredis. En effet, les maximums observés les week-ends et les mercredis ne dépassent les $50\,\mathrm{ng}\cdot\mathrm{m}^{-3}$, alors qu'ils varient entre 70 et $109\,\mathrm{ng}\cdot\mathrm{m}^{-3}$ pour les autres jours. En absence des sources de combustions dans le bureau, la source extérieure, notamment le trafic routier est tenue responsable de ces émissions.

Pour la distribution mensuelle, il est très difficile d'en tirer les conclusions du fait qu'on a observé uniquement 7 mois (de octobre à décembre 2010 et de janvier à avril 2011). On peut dire que les différences observées entre les mois n'est pas très significative. Cette constatation pourrait être due à l'influence du rythme du trafic routier extérieur qui se répercute à l'intérieur du local.

C.2 Variabilité de la concentration en polluants extérieur (Station Lognes 2009-2013)

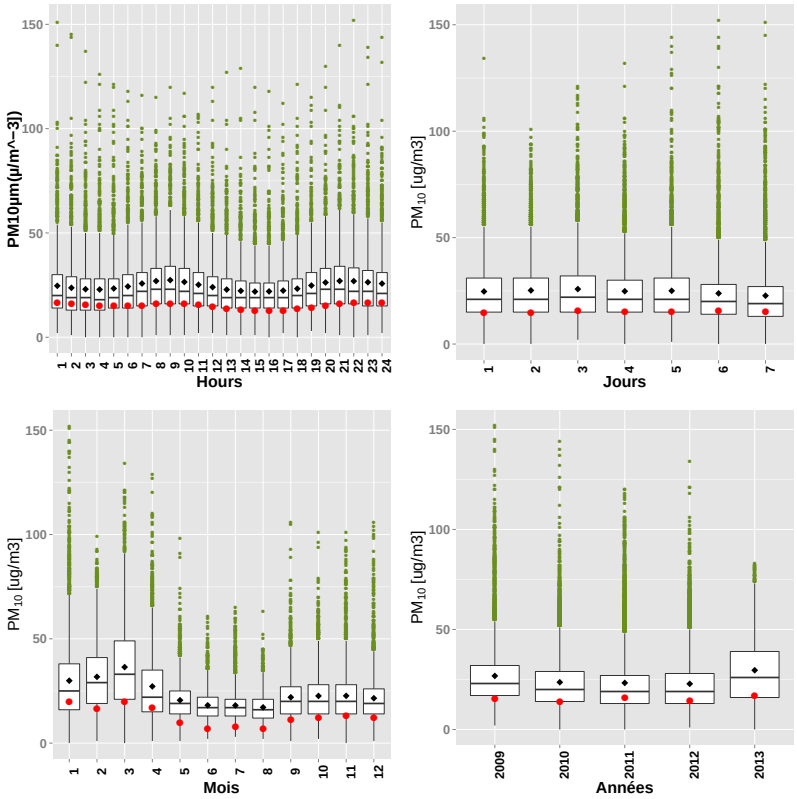


FIGURE C.2.1 – Distribution des concentrations de $PM_{10}(\mu g.m^{-3})$

C.2.1 Conditions climatiques

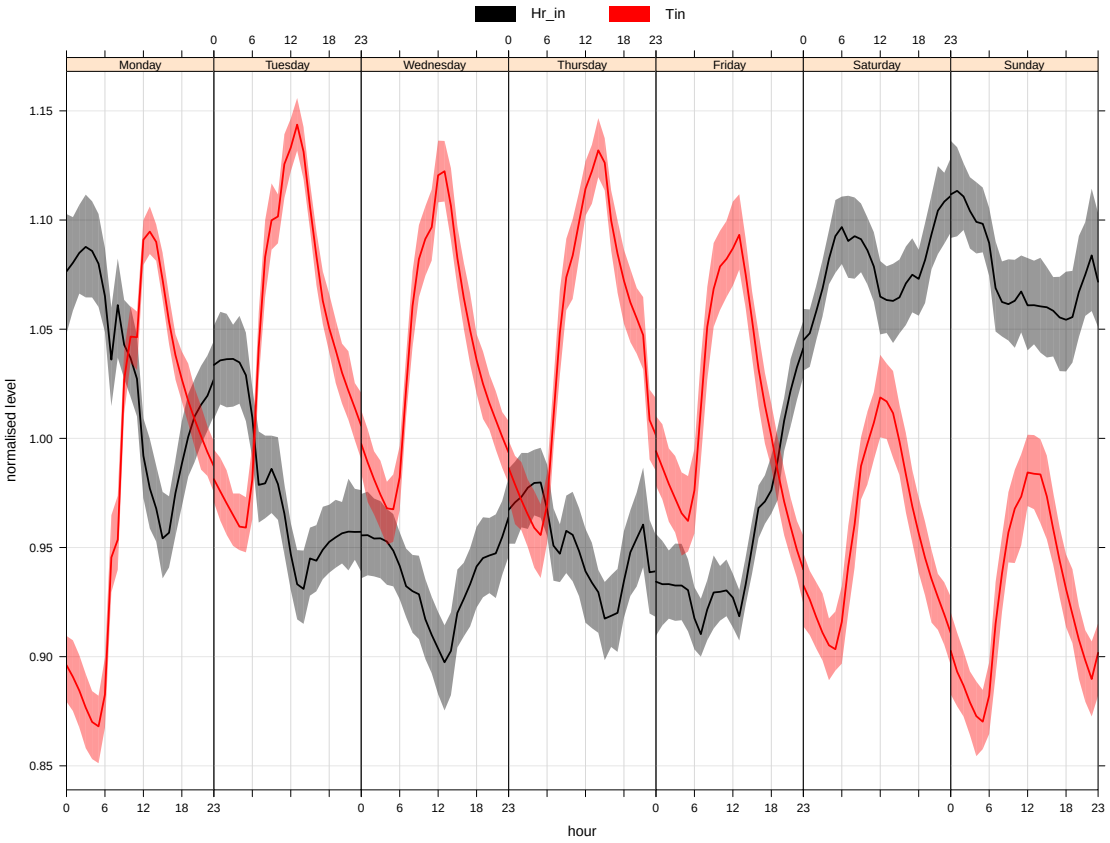


FIGURE C.2.2 – *Semaine de variabilité-type des paramètres température intérieur (**Tin**) et humidité relative intérieure (**Hr_in**) dans la maison expérimentale.*

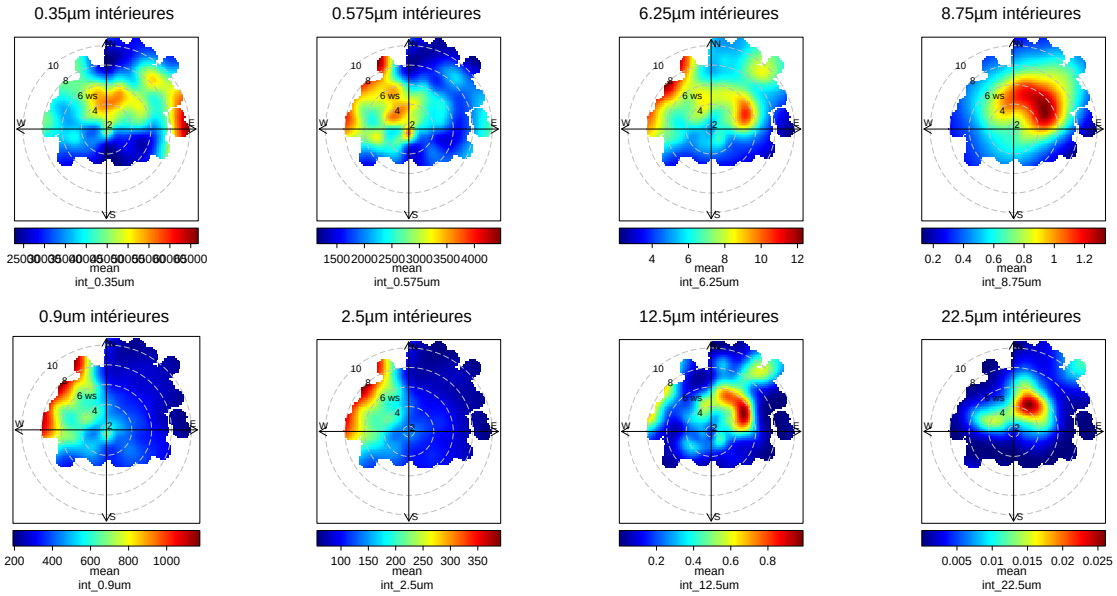


FIGURE C.2.3 – *Distribution polaire des la concentration en nombre de particules intérieures en fonction de la direction et de la vitesse du vent.*

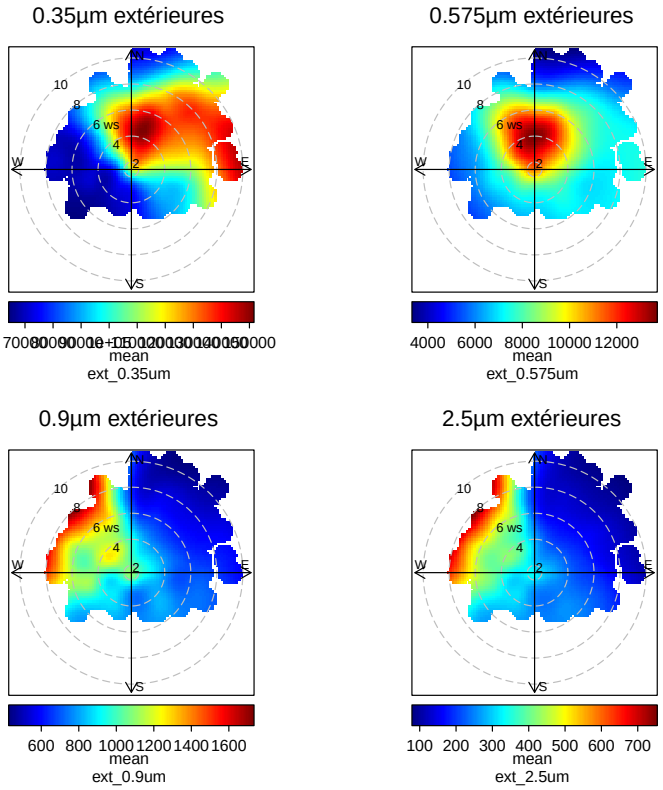


FIGURE C.2.4 – *Distribution polaire des la concentration en nombre de particules intérieures en fonction de la direction et de la vitesse du vent.*

C.2.2 Campagne 2013

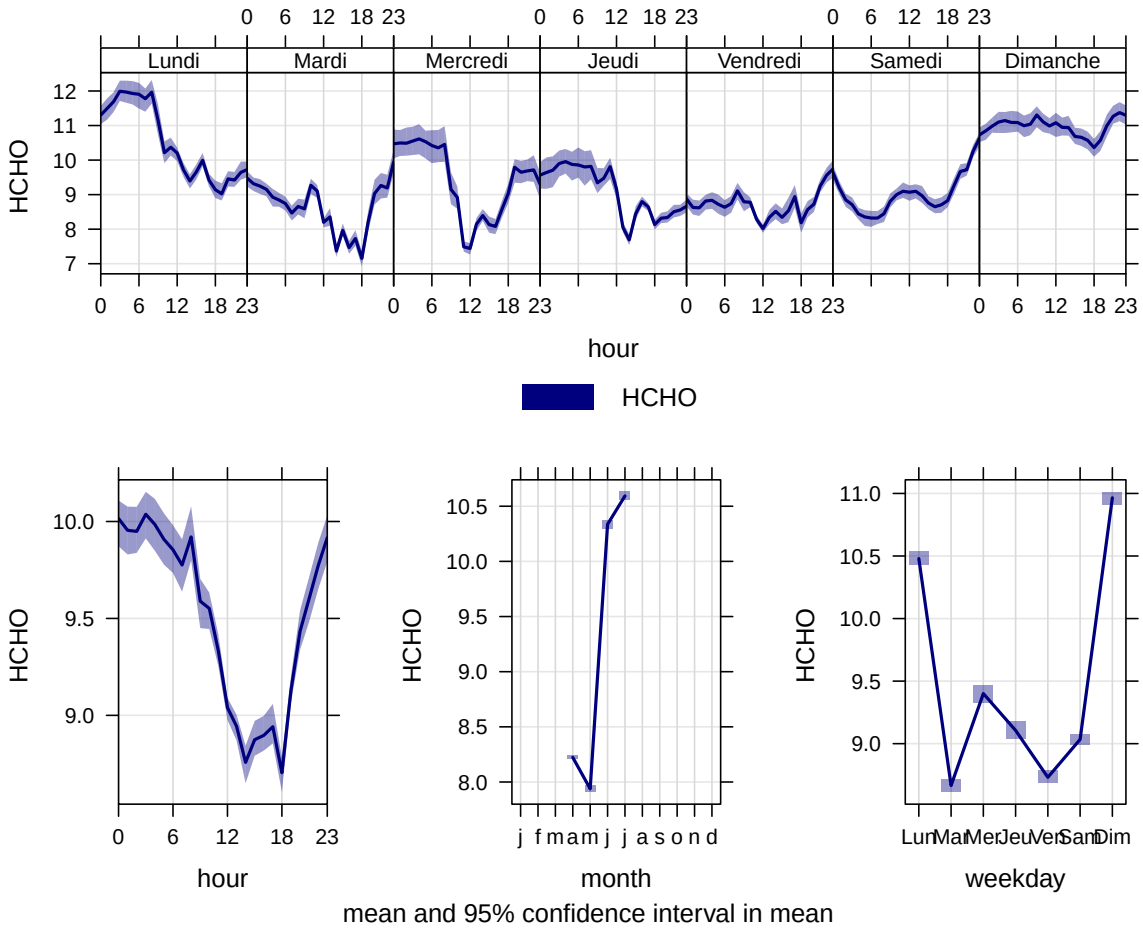


FIGURE C.2.5 – *Récapitulatif de la variabilité des concentrations du HCHO dans l'espace paysager lors de la campagne 2013.*

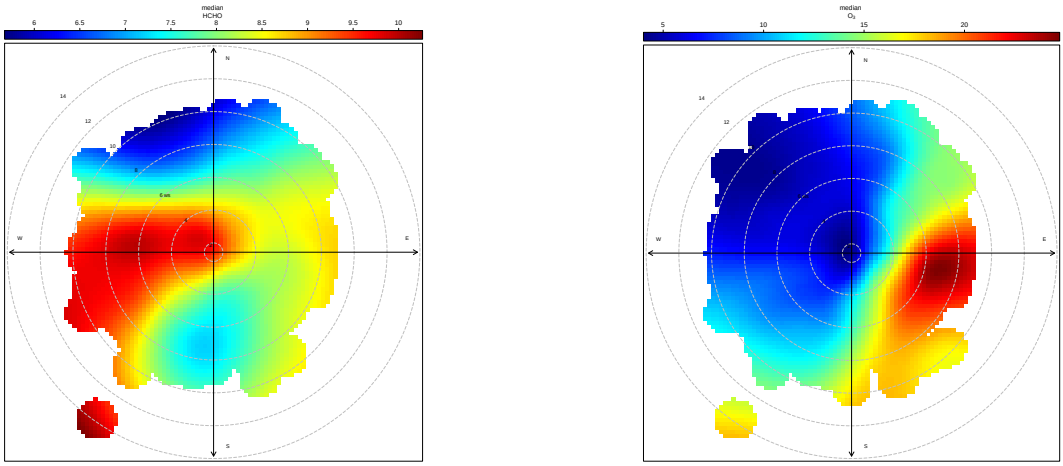


FIGURE C.2.6 – *Influence des conditions extérieure sur le nivaux des concentrations intérieure du formaldéhyde et de l'ozone.*

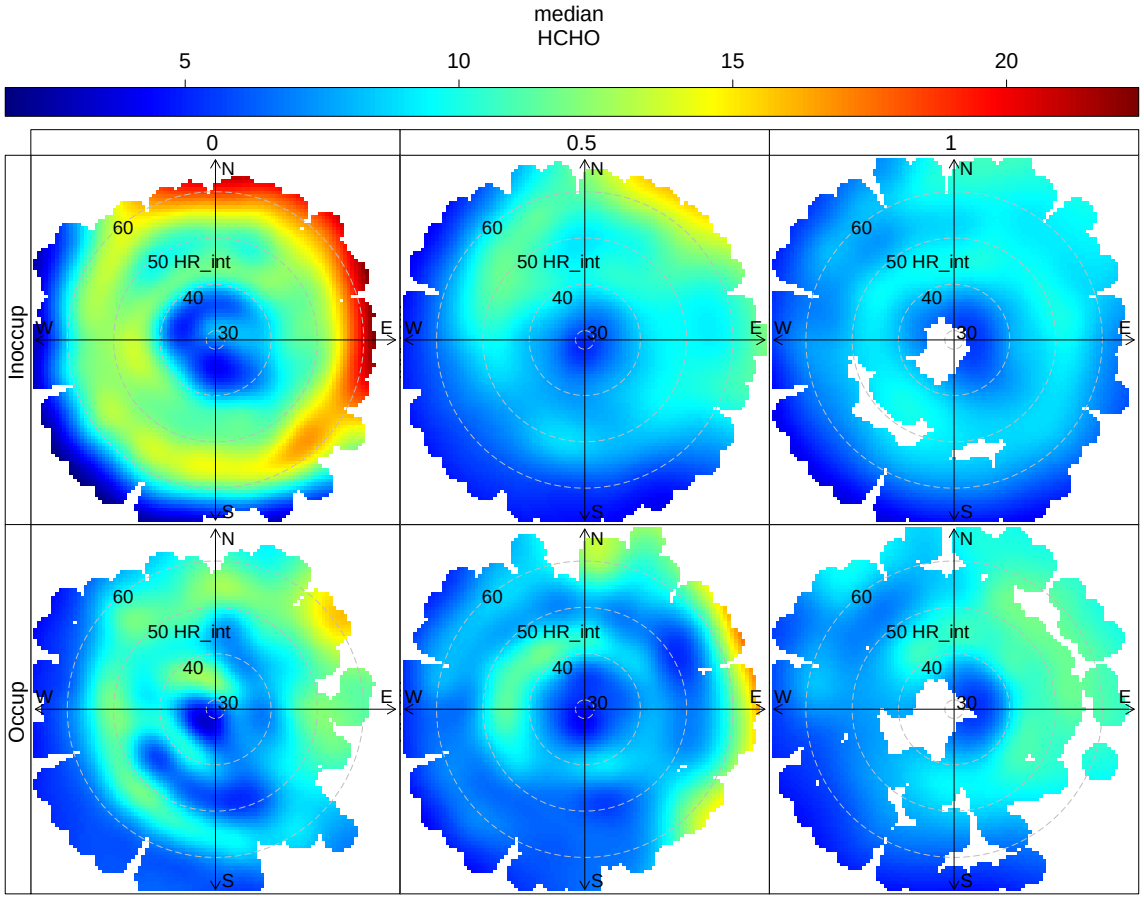


FIGURE C.2.7 – *Variabilité du HCHO avec la variation de la température intérieure et de l'humidité spécifique intérieure associée à l'état d'occupation et du ratio des fenêtres ouvertes. Les mesures couve la période du 27/04/2013 au 31/07/2013 en pas de temps d'une minute. Deux capteurs de l'état des fenêtres sur cinq instrumentés n'ont pas été pris en compte durant la période de mesure.*

ANNEXE D

STRUCTURE DE VARIABILITÉ POUR LES SÉRIES DE LA QAI

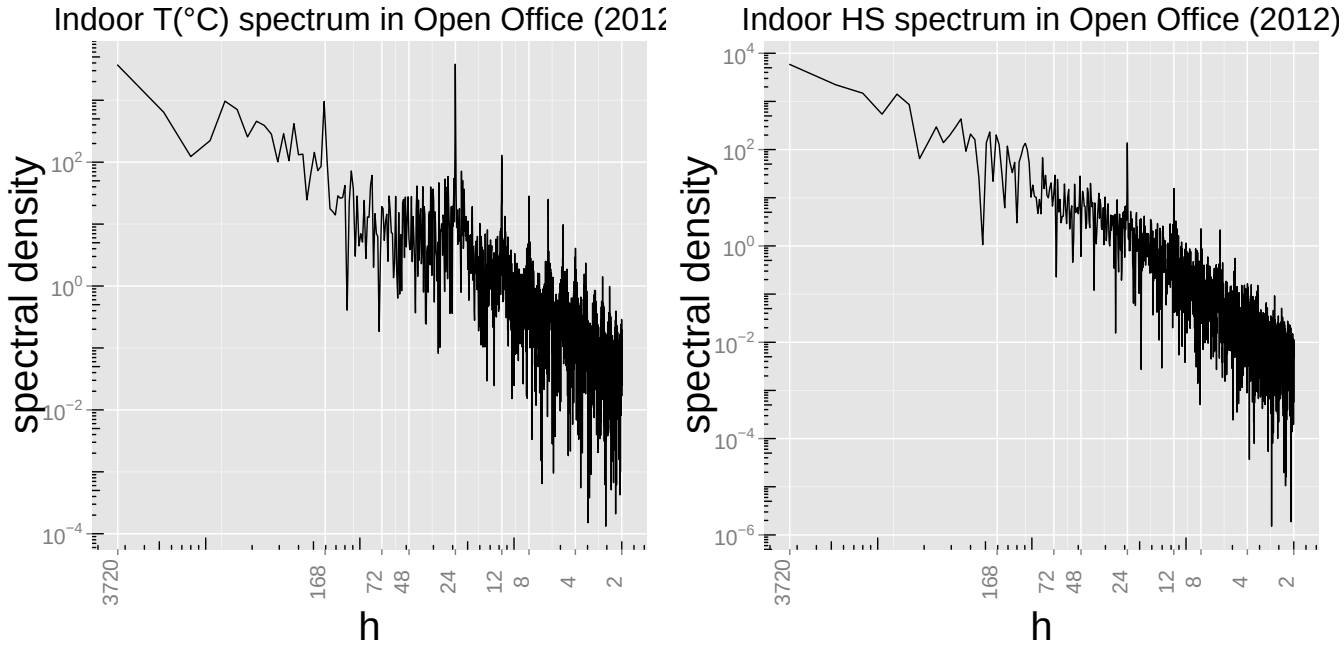


FIGURE D.0.1 – *Densités spectrales de la température et de l'humidité spécifique. Campagne de mesure (horaire) dans l'espace paysager entre février 2012 et juillet 2012.*

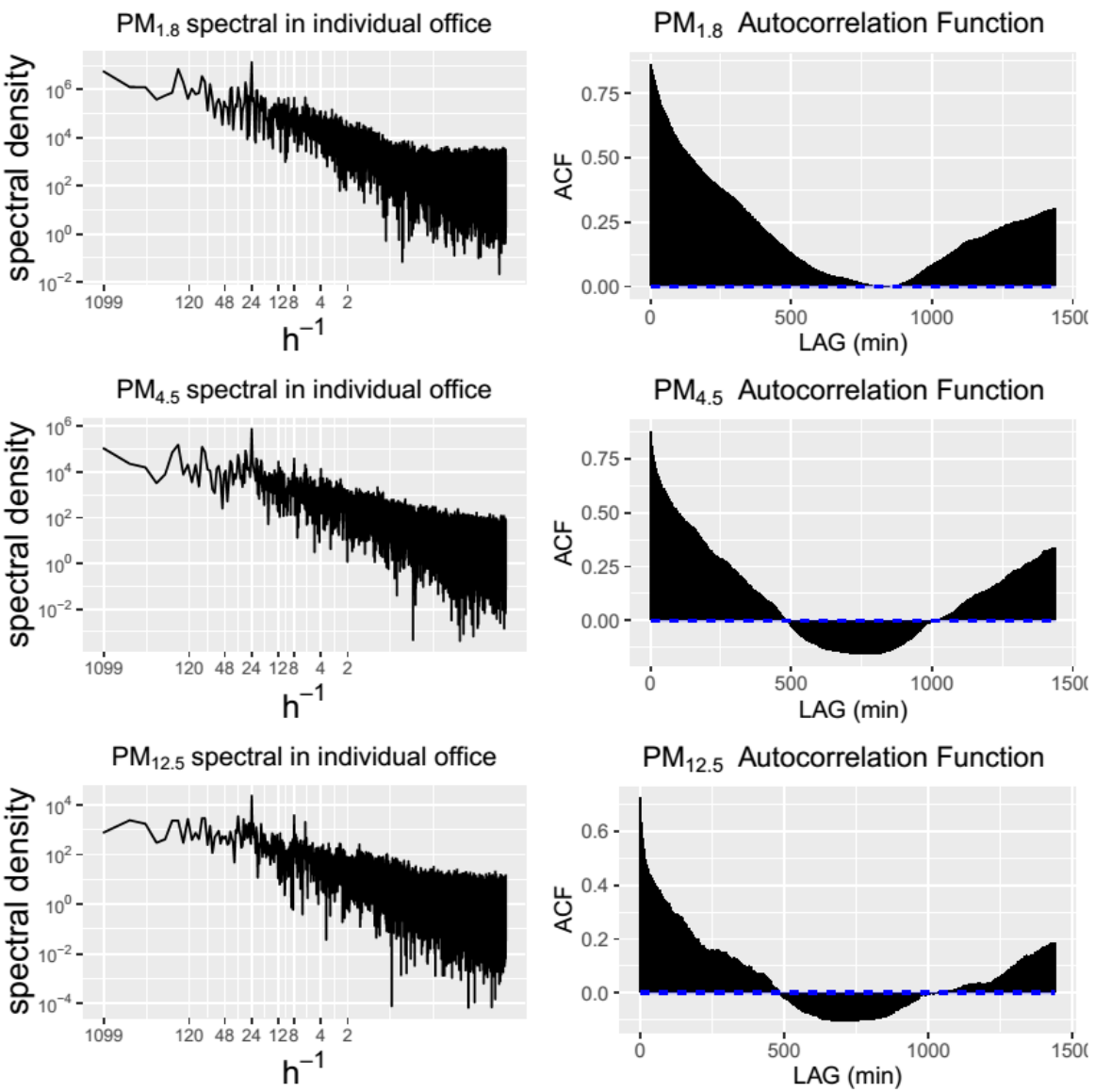


FIGURE D.0.2 – *Propriétés spectrales de différentes taille de particules observées dans le bureau individuel.*

ANNEXE E

MATHÉMATIQUE

E.1 Mouvements Browniens

Un mouvement brownien est un processus stochastique continue donnée par :

$$B(t,\lambda) = \int_{\infty}^{t} W(s)\, ds, \tag{E.1.1}$$

tels que $W(s)$ est un processus bruit blanc Gaussien. Le mouvement Brownien est caractérisé par l'indépendance de ses incrémentations et sa densité spectrale est de type λ^{-2}. Autrement dit :

- $B(t_2,\lambda) - B(t_1,\lambda)$ est de moyenne nulle et sa variance est proportionnelle à $|t_2 - t_1|$,
- $B(t_2,\lambda) - B(t_1,\lambda)$ et $B(t_4,\lambda) - B(t_3,\lambda)$ sont indépendants.

Une généralisation de ces processus est donnée par Mandelbrot & Van Ness (1968) par les mouvements Browniens fractionnaires $B_{\mathbf{H}}(t,\lambda)$:

$$B_{\mathbf{H}}(t,\lambda) = \frac{1}{\Gamma\left(\mathbf{H}+\frac{1}{2}\right)} \left\{ \int_{-\infty}^{0} \left[(t-s)^{\mathbf{H}-\frac{1}{2}} - (-s)^{\mathbf{H}-\frac{1}{2}} \right] dB(s,\lambda) + \int_{0}^{t} (t-s)^{\mathbf{H}-\frac{1}{2}}\, dB(s,\lambda) \right\}. \tag{E.1.2}$$

E.2 Mesure spectrale et théorème de Herglotz

Soit $\mathbb{T} = [-\pi, \pi]$ et $\mathcal{B}(\mathbb{T})$ la tribu de borélienne associée. Le théorème de Herglotz établit l'équivalence entre la fonction d'autocovariance et une mesure définie sur l'intervalle $\{\mathbb{T}, \mathcal{B}(\mathbb{T})\}$. Cette mesure est appelée *mesure spectrale du processus.*

Théorème E.2.1. *(Herglotz) Une suite $\{\gamma(h)\}_{h\in\mathbb{Z}}$ est de type positif si et seulement si il existe une mesure positive sur $\{\mathbb{T}, \mathcal{B}(\mathbb{T})\}$ telle que*

$$\gamma(h) = \int_{\mathbb{T}} e^{ih\lambda} \nu(d\lambda). \tag{E.2.1}$$

Si la suite $\{\gamma(h)\}_{h\in\mathbb{Z}}$ *est sommable (i.e.* $\sum_h |\gamma(h)| < +\infty$*), la mesure* ν *possède une densité* S *(fonction positive) par rapport à la mesure de Lebesgue sur* $\{\mathbb{T}, \mathcal{B}(\mathbb{T})\}$*, donnée par la série entière uniformément convergente :*

$$S(\lambda) = \frac{1}{2} \sum_{h\in\mathbb{Z}} \gamma(h) e^{ih\lambda} \geq 0. \tag{E.2.2}$$

Lorsque γ *est la fonction d'autocovariance d'un processus stationnaire au second ordre, la mesure* ν *est appelée la mesure spectrale et la fonction* S*, lorsque'elle existe, est dite densité spectrale de puissance.*

E.3 Éléments de la géométrie différentielle

E.3.1 Outils de la géométrie différentielle

Cette section est consacrée à l'étude des outils de base de la géométrie différentielle, utilisés comme ingrédients nécessaires pour une construction robuste de la théorie des systèmes dynamiques. La géométrie différentielle repose (principalement) sur l'étude des *variétés*[1] : espace $\mathbb{M}$ de dimension n dans lequel chaque point peut être localisé par n réel au moyen d'un système de coordonnées locales. Les variétés sont des objets mathématiques qui peuvent être appréhender (intuitivement) comme la généralisation des espaces euclidiens $\mathbb{R}^n$.

Cette section est largement inspirée de Spivak (1975). Pour des explications historiques sur l'invention de la notion de variété et ses motivations, nous renvoyons à l'ouvrage d'histoire de Dieudonné (2009) et aux textes originaux comme la leçon inaugurale de Riemann, traduite et commentée dans Spivak (1979) ; développée dans Poincaré (1928); Cartan (1925).

E.3.2 Notion de variété différentielle et difféomorphisme

Un objet fondamental pour l'étude de la théorie des systèmes dynamiques est la variété ; qui généralise la notion d'espace euclidien[2]. Une variété $\mathbb{M}^n$ est un espace $n-$dimensionnel qui est localement un espace euclidien $\mathbb{R}^n$, mais n'est pas nécessairement l'espace $\mathbb{R}^n$ lui-même[3]. Une sphère unité de dimension

1. L'idée des notions de variétés remonte au mathématicien Bernhard Riemann (1826-1866) : dans son article "sur les hypothèses qui servent de base à la géométrie" 1854 (ref). Il fut le premier à étendre de façon systématique la notion de surface à des objets de dimension plus grande, qu'il baptisa *« Mannigfaltigkeié »*. Aussi il généralisa des propriétés métriques et différentielles des surfaces "usuelles" de l'espace euclidien à des espaces "courbés".

il introduisit le concept de variété comme généralisation des propriétés métriques et différentielles des surfaces "usuelles" de l'espace euclidien à des espaces "courbés" considérés comme espaces de référence et non plus comme plongés dans un espace plus vaste de dimension supérieure. Ainsi par exemple une courbe plane (parabole ou sinusoïde) dans un espace de 3D est une variété de dimension 1.

L'interaction entre la dynamique du système et la topologie de $\mathcal{V}$ est souvent mise en avant par les mathématiciens.

2. L'utilisation d'un minuscule pour **e**uclidien est là juste pour se conformer eux écrits de la plupart des auteurs.

3. L'espace euclidien $\mathbb{R}^n$ doté d'un système de coordonnées globales $(x^1, \ldots, x^n)$ et est fondamentalement une variété très importante.

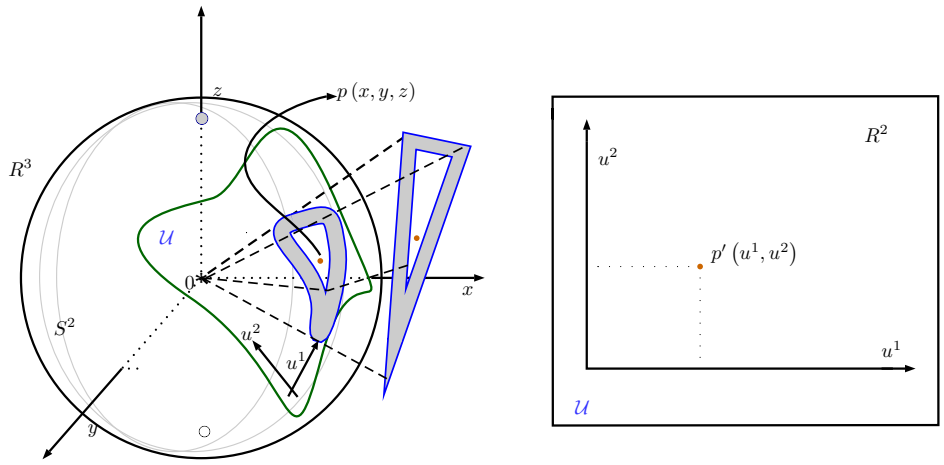


FIGURE E.3.1 – 2-Sphere $\mathbb{S}^2$ et coordonnées locales .

n ; noté $\mathbb{S}^n$, dans l'espace euclidien $\mathbb{R}$ de dimension $(n+1)$ est un exemple typique d'une variété à n dimensions $(\mathbb{M}^n)$.

En effet, si $x \in \mathbb{M}$, alors il existe un voisinage $\mathcal{U}$ de x et un entier $n \geq 0$ tel que $\mathcal{U}$ est homéomorphe à $\mathbb{R}^n$. De façon plus précise : $\forall x \in \mathbb{M}$, il existe un voisinage ouvert $\mathcal{U}_x$ et un homéomorphisme $\varphi_x : \mathcal{U}_x \to \varphi_x(\mathcal{U}_x) \subset \mathbb{R}^n$.

De fait, la variété la plus simple à construire est l'espace $\mathbb{R}^n$ lui même ; ainsi pour tout $x \in \mathbb{R}^n$ on peut prendre un voisinage U être tout $\mathbb{R}^n$ Spivak (1975).

Définition E.3.1. *Une variété topologique*[4] *de dimension n est un espace topologique séparé*[5] *dont tout point est contenu dans un ouvert homéomorphe a un ouvert de* $\mathbb{R}^n$.

Considérons une $2-$sphère unité $\mathbb{S}^2$ (variété à deux dimensions) plongée dans l'espace à trois dimensions $\mathbb{R}^3$ (Fig :E.3.1). Soit un point $p = (x, y, z)$ dans $\mathbb{R}^3$, la $2-$sphère $\mathbb{S}^2$ est définie par tout point p satisfaisant $\|p\|^2 = x^2 + y^2 + z^2 = 1$, avec $\|.\|$ la norme euclidienne usuelle. La $2-$sphère $\mathbb{S}^2$ n'est pas une partie de l'espace euclidien $\mathbb{R}^2$; mais on peut distinguer immédiatement un voisinage qui peut être décrit par deux coordonnées d'un domaine de l'espace $\mathbb{R}^2$. Un point p' d'un ouvert de $\mathcal{U}$ de la $2-$sphère $\left(p' \in U \subset \mathbb{S}^2\right)$ est représenté par les coordonnées locales $\left(u^1, u^2\right)$.

Typiquement, la représentation de la surface de la terre, une bonne approximation d'une 2-sphère $\mathbb{S}^2$ se fait naturellement par un ensemble de cartes qui définissent un atlas géographique. Donc, chaque carte est, dans un sens mathématique, une bijection bicontinue (homéomorphisme) d'une région de la terre sur une partie de l'espace $\mathbb{R}^2$. On dit alors que $(\mathcal{U}_x, \varphi_{\mathcal{U}_x})$ est une carte[6] locale de $\mathbb{M}$ ou un système de coordonnées locales (Figure E.3.2). Pour tout point[7] $p \in \mathcal{U} \subset \mathbb{M}$, il est possible d'assigner les n

4. *De façon générale, on peut définir la variété par l'espace métrique. Cette nouvelle définition s'affranchit de quelques aspects « pathologiques » des espaces non-métrisables ; voir Spivak (1975).*

5. *Aussi appelé espace de Hausdorff ; deux points distincts quelconques dans l'espace topologique admettent toujours des voisinages disjoints. Tout espace métrique est séparé.*

6. *Une carte définit « coordinate patch » sur* $\mathbb{M}$.

7. Peut importe la nature des éléments de $\mathbb{M}$, ils sont considérés comme des *points* de la variété. D'ailleurs, dans Amari et al. (2007) (Page 1) ; les points de la variété sont considérés comme des distributions de probabilités.

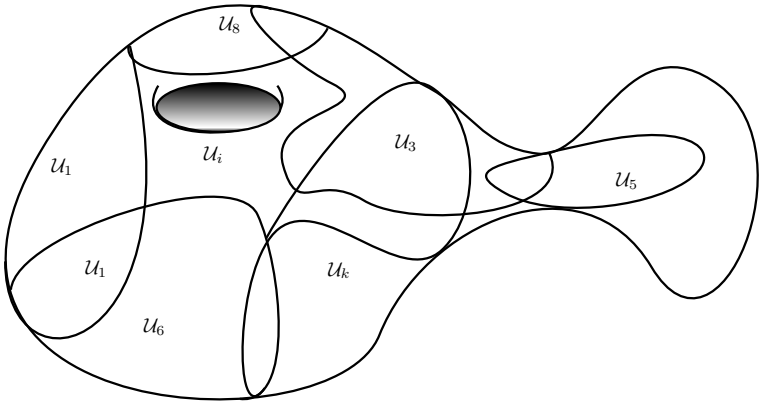


FIGURE E.3.2 – Atlas d'une variété $\mathcal{M}^n$

coordonnées du point $\varphi_{\mathcal{U}}(p)$ dans $\mathbb{R}^n$; l'homéomorphisme φ est appelé une application de coordonnées locales avec le $k-ème$ composante $x^k_{\mathcal{U}}$. Un point p sur la carte $\mathcal{U}$ est représenté par les coordonnées locales $p\left(x^1_p,\ldots,x^n_p\right)$ et l'ensemble des cartes vont constituer un atlas. Autrement dit, un atlas de la variété est la famille (pas nécessairement finie) de cartes $\left\{(\mathcal{U}_i,\varphi_i)_{i\in\mathcal{I}}\right\}$ qui recouvre entièrement $\mathbb{M}$ (Figure E.3.2).

Si deux cartes $(\mathcal{U},\varphi)$ et $(\mathcal{U}',\varphi')$ sont telle que $\mathcal{U}\cap\mathcal{U}'\neq\emptyset$, alors l'application

$$\varphi'\circ\varphi^{-1}:\varphi\left(\mathcal{U}\cap\mathcal{U}'\right)\to\varphi'\left(\mathcal{U}\cup\mathcal{U}'\right)$$

est un homéomorphisme. Considérons un chevauchement de deux cartes $(\mathcal{O},\Phi_1)$ et $(\mathcal{U},\Phi_2)$ en un domaine d'intersection contenant le point p. Soient les coordonnée locales $p=x=\left(x^1,x^2,\ldots,x^n\right)$ et $p=y=\left(y^1,y^2,\ldots,y^n\right)$ des cartes $(\mathcal{O},\Phi_1)$ et $(\mathcal{U},\Phi_2)$ respectivement (FigureE.3.3). Le point p peut être représenté par les deux systèmes de x et de y ; en particulièrement, y^i est exprimé en termes de x comme $y^i=y^i\left(x^1,\ldots,x^n\right),\ (i=1,\ldots,n)$. Ainsi, pour Φ_1 et Φ_2 suffisamment lisses et différentiables, le déterminant du Jacobien de forme

$$|J|=\frac{\partial(y)}{\partial(x)}=\frac{\partial\left(y^1,\ldots,y^n\right)}{\partial\left(x^1,\ldots,x^n\right)} \tag{E.3.1}$$

est non nul pour tout point $p\in\mathcal{O}\cap U$ Flanders (1963).

Soient une application $F:\mathbb{M}^n\to\mathbb{W}^r$ lisse d'une variété $\mathbb{M}^n$ sur $\mathbb{W}^r$, un système de coordonnée local $x=\left(x^1,x^2,\ldots,x^n\right)$ dans le voisinage du point $p\in\mathbb{M}^n$ et un système $w=\left(w^1,w^2,\ldots,w^n\right)$ dans le voisinage de $F(p)$ sur $\mathbb{W}^r$. L'application F est décrite par r fonctions $F^i(x),(i=1,\ldots,r)$ de n variables ; avec $F^i(x)$ sont différentiables sur $x^j,(j=1,\ldots,n)$. Pour $n=r$, F est un difféomorphisme : F et F^{-1} sont différentiables.

Définition E.3.2. *Soient $\mathcal{O}$ et $\mathcal{U}$ deux ouverts d'espaces affines E et F de dimension finie et Φ une application de classe $\mathcal{C}^r$ $(1\leq r\leq+\infty)$ définie de $\mathcal{O}$ à valeurs dans $\mathcal{U}$.*

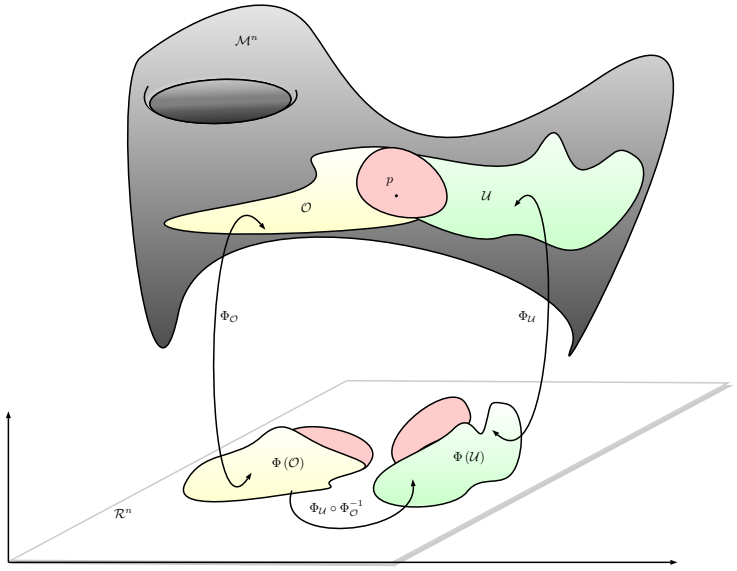


FIGURE E.3.3 – Fonction de transition (coordinate maps)

1. *Un difféomorphisme (de classe $\mathcal{C}^r$) d'un ouvert $\mathcal{O}$ de $\mathbb{R}^n$ dans un ouvert $\mathcal{U}$ de $\mathbb{R}^n$ est une application bijective $\Phi\colon \mathcal{O} \to \mathcal{U}$ qui est différentiable (de classe $\mathcal{C}^r$) de même que son application réciproque Φ^{-1} ; on dit que Φ est un $\mathcal{C}^r$-difféomorphisme.*
2. *Une partie $\mathcal{M}$ d'une variété $\mathbb{M}$ de dimension n est une sous-variété de dimension p si pour tout élément x de $\mathcal{M}$, il existe des voisinages $\mathcal{O}$ et $\mathcal{U}$ de x dans $\mathbb{M}$ et de 0 dans $\mathbb{R}^n$ respectivement, et un difféomorphisme $\Phi\colon \mathcal{O} \to \mathcal{U}$ tel que $\Phi(\mathcal{O} \cap \mathcal{M}) = \mathcal{U} \cap (\mathbb{R}^p \times \{0\})$.*
3. *Une immersion en un point x d'un ouvert $\mathcal{O}$ ($x \in \mathcal{O}$) est une application injective $\Phi\colon \mathcal{O} \to F$ différentielle en x (i.e. si sa matrice jacobienne est de rang $n = dim(E)$). Si E et F deux variétés topologiques de classe $\mathcal{C}^r$, l'application $\Phi\colon E \to F$ est une immersion si l'application linéaire tangente $Tf(x)$ injective de dimension (le rang de $Tf(x)$=dim (F)).*
4. *Une submersion en un point x d'un ouvert $\mathcal{O}$ est une application surjective $\Phi\colon \mathcal{O} \to F$ différentielle en x (i.e. si sa matrice jacobienne est de rang $n = dim(F)$). De façon similaire, si on considère les variétés topologiques ; l'application linéaire tangente est surjective.*

E.3.3 Plongement

Dans ce paragraphe, nous discuterons les conditions nécessaires pour lesquelles l'application Φ doit remplir afin qu'elle soit utilisable dans la reconstitution de l'espace d'états. Si nous disposons de toute l'information du système dynamique déterministe $S(t)$, nous serons en mesure de prédire complètement l'évolution de ce système. Ce pouvoir prédictif remarquable peut être préservé pour les variables $X(t)$ sous l'hypothèse d'injectivité de Φ et d'un bon *plongement*[8]. Un bon plongement implique que l'image $\Phi(\mathbb{M})$ de la variété contenant l'évolution du système n'a pas d'auto-intersection (ou auto-jonction) et conserve les structures topologiques et analytiques de la variété. Considérons une série sinusoïdale sans bruit que constitue un système dynamique, une reconstitution de l'espace d'états forme un cycle limite.

8. Le concept de plongement est basé uniquement sur les ensembles, espaces et les applications entre elles ; et ne requière pas la présence d'un système dynamique dans ces espaces.

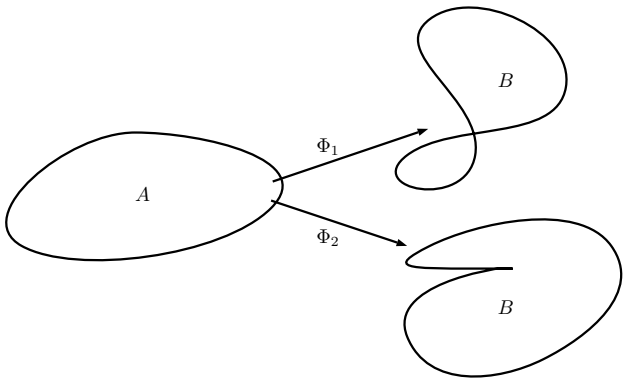


FIGURE E.3.4 – *Plongement d'une sous-variété. Pour distinguer si un objet qui se présente en dimension 2 sous l'aspect d'un chiffre 8 (courbe B par Φ_1) est en fait une boucle sans point double (courbe A), il faut se placer dans un espace au moins tridimensionnel pour pouvoir changer d'angle de perspective.*

Un plongement consiste alors en l'application d'une transformation sur ce cycle limite qui conserve les propriétés topologiques et la structure différentielle de la variété.

La Figure E.3.4 montre un exemple simple de la reconstitution d'un cycle limite (la sou-variété A) plongée sur deux sous-variétés par deux applications Φ_1 et Φ_2. Clairement, l'image de $\Phi_1(A) = B$ produit un point d'intersection ; ce point ayant deux antécédents par Φ_1, l'application Φ_1 n'est pas un plongement. Même si l'image $\Phi_2(A) = C$ n'a aucune auto-jonction, (tout point de B a exactement un seul antécédent par Φ_2) l'application Φ_2 sur A ne préserve pas la condition de différentiabilité de la variété A.

De façon générale, la première condition pour qu'une application Φ soit un plongement est que l'image par Φ de la variété $\mathbb{M}$ du système dynamique ne possède pas une auto-intersection. La seconde étant la préservation des structures différentielles : l'application dérivée $D\Phi$, une matrice de dimension $m \times d$ pour tout $s \in \mathbb{M}$ doit aussi être injective (matrice plein rang). Un système dynamique f^t sur la $\mathbb{M}$ implique une correspondance du système dynamique F^t sur $\Phi(\mathbb{M})$, si Φ est un plongement :

$$X(t) = F^t(X(0)) = \Phi f^t \Phi^{-1}(X(0)) \tag{E.3.2}$$

Définition E.3.3. *Une application Φ de $\mathcal{U}$ dans $\mathcal{O}$ est un plongement si $\Phi(\mathcal{U})$ est une sous-variété de $\mathcal{O}$ et si Φ est un difféomorphisme de $\mathcal{U}$ sur $\Phi(\mathcal{U})$.*

ANNEXE F

GESTION DES DONNÉES MANQUANTES

Cette thèse s'inscrit dans une problématique réelle de la qualité de l'air intérieur et tente de fournir des éléments pour l'analyse statistique des séries temporelles. Néanmoins la mesure en continu des différents paramètres souffre d'un mal commun à de nombreuses disciplines : la présence des données manquantes. L'analyse des séries temporelles dans cette situation est un problème embarrassant, en particulier pour la prévision par les modèles paramétriques. Ce chapitre offre quelques bribes de méthodes qui pourraient aider à contourner ce problème pour l'analyse spectrale et la prévision statistique.

F.1 Introduction

Le mode de prise en compte des données manquantes est un problème omniprésent en traitement statistique et les méthodes d'analyse et d'imputation sous-jacentes sont extrêmement nombreuses. Une description, même synthétique de la revue de littérature traitant ce sujet est hors de la portée de cette thèse, on se propose donc juste de l'aborder comme une étape préliminaire avant d'entamer la modélisation.

Ce chapitre est inspiré par l'article de Dempster et al. (1977) et les ouvrages de référence qui synthétisent la plus grande partie des méthodes proposées depuis quarante ans (Molenberghs et al., 2014; Schafer, 1997; Watanabe & Yamaguchi, 2003; Little & Rubin, 2014; Rubin, 2004).

Une méthode simple, mais naïve d'obtenir un tableau de données complet sur lequel toute analyse statistique peut être mise en œuvre est l'imputation de la valeur manquante par la moyenne de la variable. Encore plus simple, rendre le tableau complet par suppression des individus ayant des valeurs manquantes. Cependant, ces méthodes peuvent non seulement déformer les distributions et les relations entre variables (Schafer & Graham, 2002), mais aussi la structure d'autocorrélation dans le cas des séries temporelles. En outre, ces approches "simplistes" altèrent l'estimation des paramètres par le fait qu'elle ne tient pas compte de l'incertitude due aux données manquantes.

Plusieurs alternatives ont été proposées dans la littérature, mais on distingue deux méthodes avancées de gestion des données manquantes. La première, appelée imputation multiple (Schafer, 1999, 2003) consiste

à générer plusieurs jeux de données artificiels complets à partir d'une estimation de la distribution des données manquantes, et de réaliser l'inférence sur ces jeux de données. Une synthèse sur les recherches est décrite dans les ouvrages (Rubin, 2004; Little & Rubin, 2014). La deuxième est basée sur le principe de la Maximisation de la Vraisemblance (MV) des données complètes à partir des données observées ; dans ce cas, plusieurs approches sont possibles : l'algorithme EM (Dempster et al., 1977; Watanabe & Yamaguchi, 2003) et les versions Bayésiennes (Schafer & Graham, 2002; Molenberghs et al., 2014).

On note par ailleurs que la plupart des méthodes développées et appliquées sont sur les données de type coupes transversales (en Anglais : *cross-sectional* data). L'effet d'une valeur manquante dans les coupes transversales est moindre par rapport à son effet sur les séries temporelles, et ce d'autant plus en présence d'autocorrélation forte. En effet, un enregistrement à un instant donné est unique pour une série temporelle et son phénomène générateur est irréversible. La continuité des observations est fondamentale pour plusieurs raisons : l'analyse fréquentielle ne permet pas de révéler le contenu spectral de la chronique (l'application n'est pas bijective) et les modèles de prévision sont inadaptés pour ce type de données.

F.2 Notations

Formellement, soit $\mathcal{D}_T$ un tableau de n séries temporelles $(X_1, X_2, \ldots, X_n)$ scindées en deux sous-ensembles $\mathcal{D}_T = \left(\mathcal{D}_T^{\text{Obs}}, \mathcal{D}_T^{\text{Miss}}\right)$, où $\mathcal{D}_T^{\text{Obs}}$ représente les séquences observées suivies par des séquences de valeurs manquantes $\mathcal{D}_T^{\text{Miss}}$ (de l'anglais missing data), on note par "NA" (Not Available) toutes les valeurs de $\mathcal{D}_T^{\text{Miss}}$.

La distribution de $\mathcal{D}_T$ est caractérisée par un vecteur de paramètres $\boldsymbol{\theta} \in \Theta$, où Θ représente l'espace des paramètres. On se place donc dans le cadre d'un modèle statistique paramétrique $\left(\mathbb{E}_x, \mathcal{E}, (\mathbf{P}_{\boldsymbol{\theta}})_{\boldsymbol{\theta}\in\Theta}\right)$: chaque X_i, $i = 1, \ldots, n$ suit une loi gouvernée par le vecteur de paramètres $\boldsymbol{\theta} \in \mathbb{R}^d$ (pour une loi Gaussienne, $\Theta = \mathbb{R}^2$).

F.3 Interpolation par des fonctions splines

Dans toute la suite, on désignera par $\mathbb{P}_n$ l'espace vectoriel des fonctions polynômes sur le corps $\mathbb{R}$ à coefficients réels, de degré inférieur ou égal à n. On a donc, $\dim \mathbb{P}_n = n + 1$. La notation $\mathcal{C}([a, b])$ désignera l'espace des fonctions continues sur l'intervalle $[a, b]$ avec des valeurs dans $\mathbb{R}$.

F.3.1 Généralités sur l'interpolation

Soient $a = t_0 \leq \ldots \leq t_i \leq \ldots \leq t_n = b$, $n + 1$ points d'une subdivision de l'intervalle $[a, b]$ et x_i, $i = 1, \ldots, n$ les observations correspondantes. L'interpolation polynomiale consiste à trouver un polynôme d'interpolation $\Pi_m \in \mathbb{P}_m$ de degré inférieur ou égale à m, tel que

$$\Pi_m(t_i) = a_m t_i^m + a_{m-1} t_i^{m-1} + \ldots + a_1 t_i + a_0 = x_i. \tag{F.3.1}$$

Posé sous cette forme, ce problème est sur- ou sous-déterminé pour $m \neq n$. Toutefois, dans l'espace $\mathbb{P}_m$ où $m = n$, il y a une seule solution et une, tel que $\Pi_n(t_i) = x_i$, pour $i = 0, \ldots, n$, (voir (Quarteroni et al., 2008) page 260 et (Demailly, 2012) chapitre II pour une démonstration complète). En effet, les polynômes caractéristiques de Lagrange $\ell_i \in \mathbb{P}_n$ associés aux nœuds t_i, $0 \leq i \leq n$ tel que

$$\ell_i(t) = \prod_{\substack{j=0 \\ j \neq i}}^{n} \frac{t - t_j}{t_i - t_j} \quad i = 0, \ldots, n, \tag{F.3.2}$$

forment une base de $\mathbb{P}_n$. En décomposant Π_n sur cette base, on a $\Pi_n(t) = \sum_{j=0}^{n} b_j l_j(t)$, d'où la relation suivante :

$$\Pi_n(t_i) = \sum_{j=0}^{n} b_j l_j(t_i) = x_i, \quad i = 0, \ldots, n, \tag{F.3.3}$$

Comme $l_j(t_i) = \delta_{ij}$ $(l_j(t_i) = 0 \, si \, j \neq i \, l_i(t_i) = 1)$, on en déduit immédiatement que $b_i = x_i$. Le problème ci-dessus admet donc au moins une solution ; le polynôme d'interpolation existe et s'écrit sous la forme suivante

$$\Pi_n(t) = \sum_{j=0}^{n} x_j l_j(t) \tag{F.3.4}$$

La relation F.3.3 est appelée formule d'interpolation de Lagrange, et les polynômes $\ell_i(t)$ sont des polynômes caractéristiques de Lagrange.

F.3.2 Fonctions splines

Une spline est une courbe régulière, polynomiale par morceaux et qui peut avoir des discontinuités pour une dérivée supérieure. Cette section est largement inspirée de (Quarteroni et al., 2008).

Définition F.3.1. Soient $a = t_0, \ldots, t_i, \ldots t_n$, $n+1$ nœuds distincts de l'intervalle $[a, b]$, avec $a = t_0 \leq t_1 \leq \ldots \leq t_n = b$. La fonction $s_k(t)$ sur l'intervalle $[a, b]$ est une spline de degré k relative aux nœuds t_j si, $s_k \in \mathcal{C}^{k-1}([a, b])$, on a

$$s_{k|[t_j, t_{j+1}]} \in \mathbb{P}_k, \quad j = 0, 1, \ldots, n-1. \tag{F.3.5}$$

En pratique, tout polynôme de degré k sur $[a, b]$ est une spline. Il peut y avoir des discontinuités de la dérivée $k-$ième aux points internes $t_1, \ldots, t_{n-1}$. Dans la plupart des applications, la spline utilisée est cubique.

L'interpolation par des splines cubiques est particulièrement importante car elles sont de degrés inférieures permettant une approximation $\mathcal{C}^2$. La spline cubique étant de degré 3 par morceaux, sa dérivée seconde doit être continue. Soient les notations suivantes : $f_i = s_3(t_i)$ et $\mathcal{M}_i = s_3''(t_i)$, $0 \leq i \leq n$, pour $i = 1, \ldots, n$ et $s_3(t) = s_{3,i-1}(t)$ pour $t \in [t_{i-1}, t_i]$. La dérivée seconde $s''_{3,i-1}$ est linéaire et

$$s''_{3,i-1}(t) = s_3''(t_{i-1}) \frac{t_i - t}{h_i} + s_3''(t_i) \frac{t - t_{i-1}}{h_i}, \tag{F.3.6}$$

où $h_i = t_i - t_{i-1}$, $(1 \leq i \leq n)$. En intégrant deux fois F.3.6, on obtient

$$s_{3,i-1}(t) = \mathcal{M}_{i-1}\frac{(t_i - t)^3}{6h_i} + \mathcal{M}_i\frac{(t - t_{i-1})^3}{6h_i} + c_{i-1}(t - t_{i-1}) + \widetilde{c}_{i-1}, \tag{F.3.7}$$

les constantes d'intégration étant déterminées en imposant les valeurs aux extrémités $s_3(t_{i-1}) = f_{i-1}$ et $s_3(t_i) = f_i$. Ceci donne, pour $1 \leq i \leq n-1$

$$\widetilde{c}_{i-1} = s_3(t_{i-1}) - s_3''(t_{i-1})\frac{h_i^2}{6}, \tag{F.3.8}$$

$$c_{i-1} = \frac{s_3(t_i) - s_3(t_{i-1})}{h_i} - \frac{h_i}{6}\left(s_3''(t_i) - s_3''(t_{i-1})\right). \tag{F.3.9}$$

Introduisons les notations suivantes : $\mu_i = \frac{h_i}{h_i+h_{i+1}}$, $\lambda_i = \frac{h_{i+1}}{h_i+h_{i+1}}$, $d_i = \frac{6}{h_i+h_{i+1}}\left(\frac{f_{i+1}-f_i}{h_{i+1}} - \frac{f_i-f_{i-1}}{h_i}\right)$. Pour $i = 1, \ldots, n-1$ les relations précédentes conduisent au système linéaire suivant :

$$\mu_i\mathcal{M}_{i-1} + 2\mathcal{M}_i + \lambda_i\mathcal{M}_{i+1} = d_i \tag{F.3.10}$$

Le système F.3.10 a $n+1$ inconnues et $n-1$ équations ; deux conditions $(k-1)$ restent à fixer. Généralement, on pose :

$$2\mathcal{M}_0 + \lambda_0\mathcal{M}_1 = d_0,\ 0 \leq \lambda_0, \tag{F.3.11}$$

$$\mu_n\mathcal{M}_{n-1} + 2\mathcal{M}_n = d_n,\ \mu_n \leq 1, \tag{F.3.12}$$

où d_0 et d_n sont des valeurs données. Pour obtenir des splines naturelles (satisfaisant $a = s_3''(t_0) = s_3''(t_n) = b = 0$), on doit annuler les coefficients ci-dessus. Ce qui revient à prolonger la spline au-delà des points extrêmes de l'intervalle $[a, b]$ et à traiter a et b comme des points internes. Cette stratégie donne une spline à comportement "régulier".

En général, le système F.3.10 est tridiagonal de la forme

$$\underbrace{\begin{bmatrix} 2 & \lambda_0 & 0 & \ldots & 0 \\ \mu_1 & 2 & \lambda_1 & & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & & \mu_{n-1} & 2 & \lambda_{n-1} \\ 0 & \ldots & 0 & \mu_n & 2 \end{bmatrix}}_{\boldsymbol{A}} \underbrace{\begin{bmatrix} \mathcal{M}_0 \\ \mathcal{M}_1 \\ \vdots \\ \mathcal{M}_{n-1} \\ \mathcal{M}_n \end{bmatrix}}_{\boldsymbol{\mathcal{M}}} = \underbrace{\begin{bmatrix} d_0 \\ d_1 \\ \vdots \\ d_{n-1} \\ d_n \end{bmatrix}}_{\boldsymbol{B}}, \tag{F.3.13}$$

et il peut être facilement résolu par factorisation de type $\boldsymbol{LU}$ de la matrice $\boldsymbol{A}$ définie dansF.3.13. Les matrices bidiagonales $\boldsymbol{L}$ et $\boldsymbol{U}$ sont de la forme :

$$\boldsymbol{L} = \begin{bmatrix} 1 & 0 & 0 & \ldots & 0 \\ \beta_1 & 1 & 0 & & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & & \beta_{n-1} & 1 & 0 \\ 0 & \ldots & 0 & \beta_n & 1 \end{bmatrix},\ \boldsymbol{U} = \begin{bmatrix} \alpha_1 & \gamma_1 & 0 & \ldots & 0 \\ 0 & \alpha_2 & & & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & & 0 & \alpha_{n-1} & \gamma_{n-1} \\ 0 & \ldots & 0 & 0 & \alpha_n \end{bmatrix},$$

avec $a_1 = 2$, $\beta_i = \frac{\mu_i}{\alpha_{i-1}}$, $\alpha_i = 2 - \beta_i\gamma_{i-1}$, $\gamma_i = \lambda_i$, $i = 2, \ldots, n$. Ces formules sont connues sous le nom d'*algorithme de Thomas.*

F.3.3 La procédure MTSDI

Dans nos applications, la spline est ajustée autour de la donnée manquante par la méthode MTSDI (Missing Data Imputation in Multivariate Time Series via EM Algorithm) proposée dans Junger & de Leon (2015). Cette méthode est appliquée aux séries temporelles multivariées des polluants et nous l'utilisons pour cette raison. En outre, pour des raisons de représentation, nous gardons les mêmes notations que dans l'article de Junger & de Leon (2015), car ici, il s'agit d'un traitement multidimensionnel.

Soit $\boldsymbol{x}_t, (t = 1, \ldots, n)$ la réalisation de p vecteurs aléatoires $\boldsymbol{X}$ normalement distribués avec m données manquantes. Le vecteur $\boldsymbol{x_t}$ peut être rangé de manière à avoir les éléments manquants en première position, *i.e.* $\boldsymbol{x_t} = \left(\underbrace{x_{t1}, x_{t2}, \ldots, x_{tm}}_{manquantes}, \underbrace{x_{t(m+1)}, \ldots, x_{tp}}_{observées} \right)^\top$, noté $\boldsymbol{x}_t = (\boldsymbol{x}_{t1}, \boldsymbol{x}_{t2})^\top$. Par ailleurs, on considère que la période observée peut être cindée sur B fenêtres temporelles (soit $b = 1, 2, \ldots, B$) et chaque fenêtre comporte au cours du temps différents régimes sous-jacents de covariances. Par conséquent, l'estimation du vecteur moyen au temps t et la matrice de covariance de la fenêtre temporelle b peuvent être positionnés comme la configuration $\boldsymbol{x_t}$, *i.e.* :

$$\tilde{\mu}_t \begin{bmatrix} \tilde{\mu}_{t1} \\ \tilde{\mu}_{t2} \end{bmatrix} \text{ et } \tilde{\boldsymbol{\Sigma}}_b = \begin{bmatrix} \tilde{\boldsymbol{\Sigma}}_{b11} & \tilde{\boldsymbol{\Sigma}}_{b12} \\ \tilde{\boldsymbol{\Sigma}}_{b21} & \tilde{\boldsymbol{\Sigma}}_{b22} \end{bmatrix}.$$

La méthode de Junger & de Leon (2015) consiste en la modification de l'algorithme EM pour l'estimation du vecteur moyen et des matrices de covariances d'une distribution multidimensionnelle de la loi Gaussienne en présence des valeurs manquantes (Dempster et al., 1977). Pour la partie EM, il s'agit de trouver la valeur espérée de la log-vraisemblance (vecteurs de paramètres) de l'ensemble complet de données par rapport aux données manquantes, connaissant les données observées.

En général, les valeurs initiales de $\tilde{\mu}_0$ et $\tilde{\boldsymbol{\Sigma}}_0$ sont estimées à partir des données incomplètes observées. À la $(k+1)$-ième itération de l'étape E (estimation) de l'algorithme EM, la valeur manquante est imputée avec l'espérance conditionnelle aux valeurs observées et à l'estimation des paramètres à l'étape précédente (k), comme suit :

$$\tilde{\boldsymbol{x}}_{\boldsymbol{t1}}^{(k+1)} = \mathbb{E}\left[\boldsymbol{X}_{t1} \mid \boldsymbol{x}_{\boldsymbol{t2}}, \tilde{\mu}_t^{(k)}, \tilde{\boldsymbol{\Sigma}}_b^{(k)}\right] = \tilde{\mu}_{\boldsymbol{t1}}^{(k)} + \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{b12}}^{(k)} \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{b22}}^{(k)-1} \left(\boldsymbol{x}_{\boldsymbol{t2}} - \tilde{\mu}_{\boldsymbol{t2}}^{(k)}\right). \tag{F.3.14}$$

Les contributions sur les matrices de covariances sont données par :

$$\begin{aligned} \widetilde{\boldsymbol{x}_{t1}\boldsymbol{x}_{t1}^\top}^{(k+1)} &= \mathbb{E}\left[\boldsymbol{X}_{t1}\boldsymbol{X}_{t1}^\top \mid \boldsymbol{x}_{t2}, \tilde{\mu}_t^{(k)}, \tilde{\boldsymbol{\Sigma}}_b^{(k)}\right] \\ &= \tilde{\boldsymbol{\Sigma}}_{b11}^{(k)} - \tilde{\boldsymbol{\Sigma}}_{b12}^{(k)} \tilde{\boldsymbol{\Sigma}}_{b22}^{(k)-1} \tilde{\boldsymbol{\Sigma}}_{b21}^{(k)} + \tilde{\boldsymbol{x}}_{t1}\tilde{\boldsymbol{x}}_{t1}^\top \end{aligned} \tag{F.3.15}$$

et

$$\widetilde{\boldsymbol{x}_{t1}\boldsymbol{x}_{t2}^\top}^{(k+1)} = \mathbb{E}\left[\boldsymbol{X}_{t1}\boldsymbol{X}_{t2}^\top \mid \boldsymbol{x}_{t2}, \tilde{\mu}_t^{(k)}, \tilde{\boldsymbol{\Sigma}}_b^{(k)}\right] = \tilde{\boldsymbol{x}}_{t1}\tilde{\boldsymbol{x}}_{t2}^\top. \tag{F.3.16}$$

Dans l'étape M (Maximisation), les estimations du maximum de vraisemblance révisées de u_b et $\boldsymbol{\Sigma}_b$ sont calculées. À l'itération $(k+1)$ et en laissant du côté les indices,

$$\tilde{\mu}_b = \sum_{t=1}^{n_b} \frac{\tilde{\boldsymbol{x}}_{bt}}{n_b} \text{ et } \tilde{\boldsymbol{\Sigma}}_b = \sum_{t=1}^{n_b} \frac{\widetilde{\boldsymbol{x}_{tb}\boldsymbol{x}_{tb}^\top}}{n_b} - \tilde{\mu}_b\tilde{\mu}_b^\top.$$

Dans la version de Junger & de Leon (2015), la contribution temporelle au niveau de chaque série temporelle μ_t est estimée indépendamment en utilisant une méthode *ad hoc*. Typiquement, l'utilisation de n'importe qu'elle méthode susceptible de modéliser μ_t est envisageable à ce niveau. Plusieurs possibilités peuvent être utilisées : un filtre de type ARIMA (Box & Jenkins, 1976), un modèle additif généralisé (GAM, Generalized Additive Model) (Hastie & Tibshirani, 1990) ou une interpolation DE type spline.

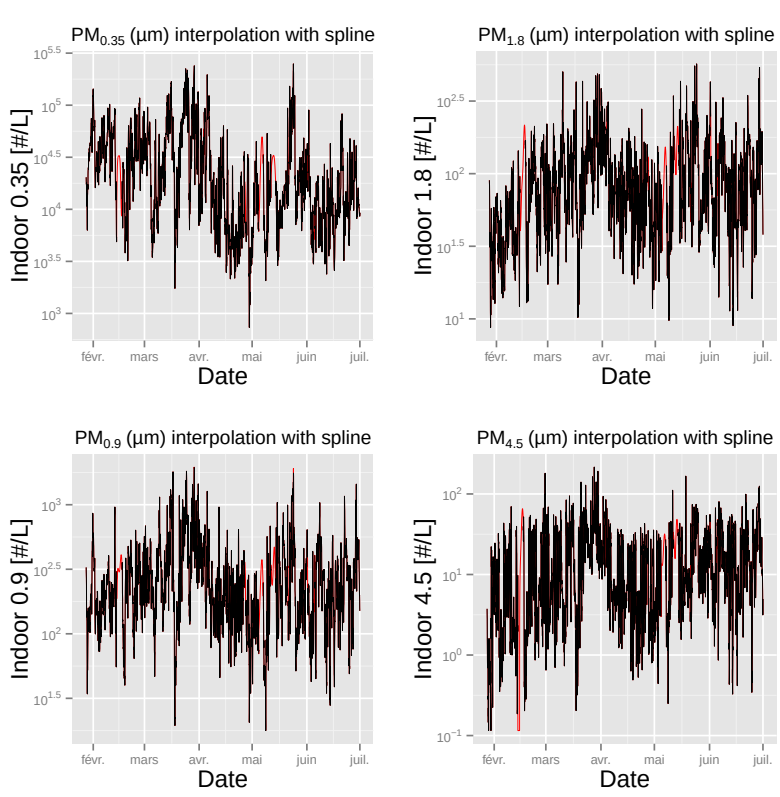
En ce qui concerne les applications de cette méthode aux différents jeux de données de la QAI, nous optons pour l'utilisation de la spline cubique $s^{(j)}$ pour chaque $\mu_t^{(j)}$ $(j = 1, \ldots, p)$. Cette procédure consiste en la minimisation de la fonctionnelle

$$g\left(s^{(j)}\right) = \sum_{k=1}^{K} \left\{X_t - s_k^{(j)}\left(\nu_k\right)\right\}^2 + \lambda \int_a^b \left\{s''\right\}^2 dx, \tag{F.3.17}$$

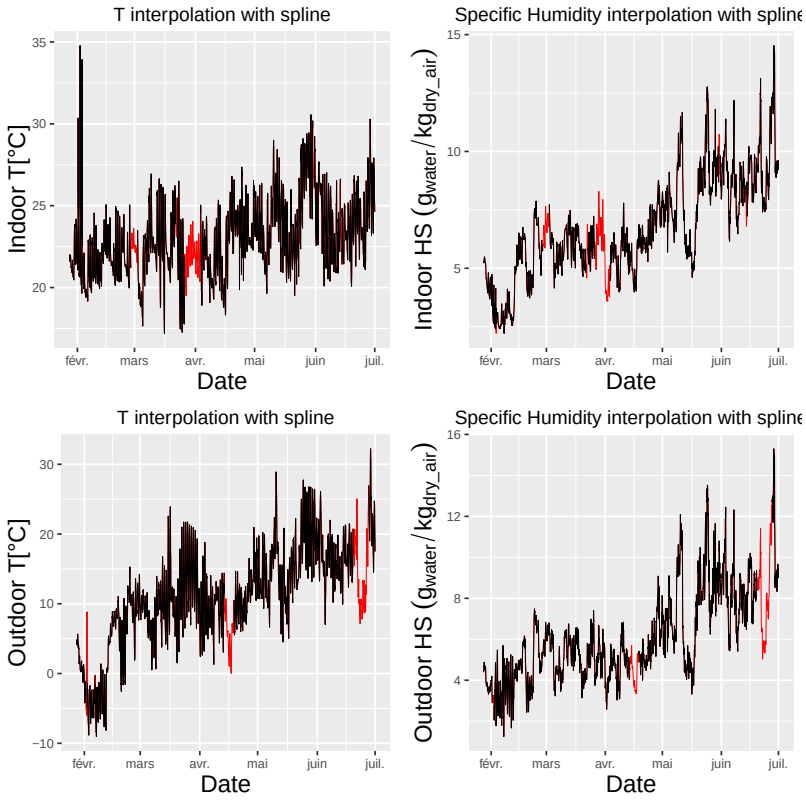
où s'' est la dérivée seconde de s, les nœuds $\nu_1, \nu_2, \ldots, \nu_K$ sont ordonnés sur l'intervalle $[a, b]$ et λ est un paramètre de lissage. La solution à ce problème d'optimisation est une spline cubique et pour chaque covariable $X^{(j)}$, le niveau est donné par $\mu_t^{(j)} = s(x_t^{(j)})$. Le paramètre du lissage λ est exprimé en termes des nombre de degré de liberté associé au nombre de nœuds dans la subdivision de l'intervalle $[a, b]$.

Le paramétrage de la spline est important pour une bonne estimation de la valeur manquante. Au risque de lisser les fluctuations des séries, il est conseillé d'éviter de sur-ajuster (sur-apprentissage) les points au voisinage de la valeur manquante. Aussi, la position des valeurs manquantes au sein de la série temporelle peut aussi altérer leur estimation ; notamment lorsque ces dernières se trouvent au voisinage des bords et à proximité de valeurs extrêmes. Il est donc nécessaire de contrôler les sorties des algorithmes en imposant quelques conditions afin d'avoir une estimation cohérente à l'observation. Ainsi, pour une valeur estimée inférieure au minimum des données sans valeurs manquantes, alors l'estimation de la donnée manquante vas prendre le $5^{ème}$ centile. À l'inverse, si elle est supérieure au maximum, alors la valeur manquante prend le le $95^{ème}$ centile.

La Figure F.3.1 illustre des exemples d'imputation des données manquantes par interpolation spline de l'algorithme de Junger & de Leon (2015).



(a) *Interpolation des données manquantes des concentrations en nombre des PM intérieurs par la méthode MTSDI.*



(b) *Interpolation des données manquantes des paramètres climatiques : température intérieur et extérieure et l'humidité spécifique intérieure et extérieure.*

FIGURE F.3.1 – *Exemple d'imputation des données manquantes par la méthode MTSDI avec la spline cubique. Les données brutes sont issues de la campagne de 2012 dans l'espace paysager.*

F.4 Méthodes basées sur le Maximum de Vraisemblance (MV)

F.4.1 L'hypothèse MAR (Missing At Random) et vraisemblance

Lorsque l'on souhaite analyser plusieurs séries temporelles présentant des données manquantes, il est nécessaire d'appréhender le mécanisme de réponse associé à ces données. Si la succession des séquences $\mathcal{D}_T^{\text{Obs}}$ et $\mathcal{D}_T^{\text{Miss}}$ est une variable aléatoire de loi de paramètre de nuisance $\boldsymbol{\xi}$, alors le mécanisme de réponse pour une série entière peut être représenté par une matrice indicatrice (Rubin, 2004) :

$$\mathcal{R}[i,j] = \begin{cases} 0 & \text{si } \mathcal{D}_T[i,j] = \text{NA} \\ 1 & \text{sinon} \end{cases}. \tag{F.4.1}$$

Sous cette forme, la loi du processus générateur de données $\mathcal{D}_T$ fait intervenir non seulement le vecteur de paramètres $\boldsymbol{\theta}$, mais également le mécanisme de réponse $\mathcal{R}$. Par conséquent, toute procédure d'estimation ou de test est fondée sur la loi $\mathcal{L}_{\mathcal{D},\mathcal{R}}$ du couple $(\mathcal{D}_T, \mathcal{R})$:

$$\mathbf{P}(\mathcal{D}_T, \mathcal{R}) = \mathbf{P}\left(\mathcal{D}_T^{\text{Obs}}, \mathcal{D}_T^{\text{Miss}}, \mathcal{R}\right). \tag{F.4.2}$$

Si la probabilité qu'une valeur soit manquante ne dépend pas de la valeur non-observée, l'information contenue dans $\mathcal{D}_T$ suffit entièrement à caractériser le mécanisme de réponse $\mathcal{R}$, dans ce cas $\mathbf{P}(\mathcal{R} \mid \mathcal{D}_T, \boldsymbol{\xi}) = \mathbf{P}\left(\mathcal{R} \mid \mathcal{D}_T^{\text{Obs}}, \boldsymbol{\xi}\right)$ et on parle alors des données manquantes au hasard.

Étant donné que dans F.4.2 le processus $\mathcal{R}$ est aléatoire, la difficulté de définir un modèle réaliste de la distribution des données $\mathbf{P}(\mathcal{D}_T, \mathcal{R})$ est double. Suivant les arguments de Rubin (1976) et sous les conditions *d'ignorabilité*[1] du mécanisme de réponse $\mathcal{R}$, l'inférence sur le vecteur $\boldsymbol{\theta}$ (MV ou Bayesien) peut être réalisée uniquement à partir des données observées. L'hypothèse MAR permet d'estimer la vraisemblance des données observées, on montre alors (Schafer, 1997) :

$$\begin{aligned} \mathbf{P}\left(\mathcal{R}, \mathcal{D}_T^{\text{Obs}} \mid \boldsymbol{\theta}, \boldsymbol{\xi}\right) &= \int \mathbf{P}(\mathcal{R}, \mathcal{D}_T \mid \boldsymbol{\theta}, \boldsymbol{\xi})\, d\mathcal{D}_T^{\text{Miss}} \\ &= \int \mathbf{P}(\mathcal{R} \mid \mathcal{D}_T, \boldsymbol{\xi})\, \mathbf{P}(\mathcal{D}_T \mid \boldsymbol{\theta})\, d\mathcal{D}_T^{\text{Miss}} \\ &= \mathbf{P}\left(\mathcal{R} \mid \mathcal{D}_T^{\text{Obs}}, \boldsymbol{\xi}\right) \int \mathbf{P}(\mathcal{D}_T \mid \boldsymbol{\theta})\, d\mathcal{D}_T^{\text{Miss}} \\ &= \mathbf{P}\left(\mathcal{R} \mid \mathcal{D}_T^{\text{Obs}}, \boldsymbol{\xi}\right) \mathbf{P}\left(\mathcal{D}_T^{\text{Obs}} \mid \boldsymbol{\theta}\right). \end{aligned} \tag{F.4.3}$$

La vraisemblance des données observées sous l'hypothèse MAR peut être factorisée en deux portions, la première concerne le paramètre d'intérêt $\boldsymbol{\theta}$ et la seconde est relative au paramètre de nuisance $\boldsymbol{\xi}$. Si les paramètres $\boldsymbol{\theta}$ et $\boldsymbol{\xi}$ sont distincts, le mécanisme de réponse est "*ignorable*" : la vraisemblance d'ignorer le mécanisme $\mathcal{R}$ des données manquantes est donnée par la relation (Little & Rubin, 2014) :

$$L\left(\boldsymbol{\theta} \mid \mathcal{D}_T^{\text{Obs}}\right) \propto \boldsymbol{P}\left(\mathcal{D}_T^{\text{Obs}} \mid \boldsymbol{\theta}\right). \tag{F.4.4}$$

Nous appelons, sans ambigüité, la relation F.4.4 la vraisemblance des données observées, bien qu'en réalité, c'est la formule F.4.3 qui offre l'estimation complète de la vraisemblance. Car nous supposons, comme le suggère (Little & Rubin, 2014), que le critère *d'ignorabilité* est satisfait partout, donc il n'est pas nécessaire de travailler directement sur la relation F.4.3.

1. C'est-à-dire que la loi de la variable d'intérêt ne dépend pas du fait qu'il y a observation ou pas. Cette hypothèse permet de manipuler la variable d'intérêt de manière aveugle. On peut consulter l'ouvrage d'Ardilly (2006) pour plus de détails sur ces concepts qui sont essentiellement tirés de la théorie de sondage.

F.4.2 L'algorithme Expectation-Maximisation (EM)

L'algorithme EM (Expectation-Maximisation) est un algorithme itératif dû à Dempster, Laird et Rubin (1977) (**DLR**-théorie). Il s'agit d'une méthode d'estimation paramétrique s'inscrivant dans le cadre général du MV. Il permet de trouver les paramètres de vraisemblance maximum d'un modèle probabiliste lorsque ce dernier dépend de variables latentes (non observables).

Généralement, on se sert de l'algorithme EM comme solution plausible, lorsque les seules données dont on dispose ne permettent pas l'estimation des paramètres et/ou l'expression de la vraisemblance est analytiquement très difficile à maximiser. Typiquement, cette difficulté est souvent rencontrée dans les cas de la modélisation des variables latentes (mélange Gaussien) où dans le cas de la présence des valeurs manquantes. L'intérêt principal de l'approche EM sous l'hypothèse MAR est de permettre de contourner cette difficulté en factorisant la vraisemblance en deux parties "faciles" à manipuler.

Soient $\boldsymbol{x} \in \mathcal{D}_T^{\text{Obs}}$ et $\boldsymbol{y} \in \mathcal{D}_T^{\text{Miss}}$ deux sous-ensembles de séries temporelles qui représentent, respectivement, les données complètes et manquantes de $\mathcal{D}_T$. On note par $f(\boldsymbol{x} \mid \boldsymbol{\theta})$ la fonction de densité de probabilité des données complètes et par $g(\boldsymbol{y} \mid \boldsymbol{\theta})$ la fonction de densité de probabilité des données incomplètes. On suppose par ailleurs qu'il existe une fonction $y \to y(x)$ de Ω_X sur Ω_Y. Alors, la fonction de densité de probabilité de $\boldsymbol{y}$, $g(\boldsymbol{y} \mid \boldsymbol{\theta})$ est :

$$g(\boldsymbol{y} \mid \boldsymbol{\theta}) = \int_{\Omega_Y(\boldsymbol{y})} f(\boldsymbol{y} \mid \boldsymbol{\theta})\, dx, \tag{F.4.5}$$

avec $\Omega_Y(\boldsymbol{y})$ le sous-espace de l'espace Ω_X déterminé par l'équation $\boldsymbol{y} = y(\boldsymbol{x})$. Pour le problème des données manquantes, la **DLR**-théorie suppose que *(i)* les paramètres à estimer sont indépendants du mécanisme générateur des données manquantes, et *(ii)* les données manquantes sont de type MAR.
Soit $LL_c(\boldsymbol{\theta}) = \log f(\boldsymbol{x} \mid \boldsymbol{\theta})$, le logarithme de la fonction de vraisemblance des données complètes et $LL(\boldsymbol{\theta}) = \log g(\boldsymbol{y} \mid \boldsymbol{\theta})$, le logarithme de la fonction de vraisemblance des données manquantes. Le but de l'algorithme EM est de trouver une estimation de maximum de vraisemblance de $\boldsymbol{\theta}$, le point atteint par le maximum de $LL(\boldsymbol{\theta})$. Cette approche aborde le problème d'optimisation de $LL(\boldsymbol{\theta})$ des données incomplètes par un processus itératif de $LL_c(\boldsymbol{\theta})$ (des données complètes). Puisqu'elle est non-observable, (inobservable dans le cas de trajectoires), cette quantité est remplacée par l'espérance conditionnelle aux observations et par des paramètres ayant des valeurs temporaires.

$$\boldsymbol{\theta}^{(k)} = \arg\max_{\boldsymbol{\theta} \in \Theta} \mathbb{E}\left[LL_c(\boldsymbol{\theta}) \mid \boldsymbol{y}, \boldsymbol{\theta}^{(k-1)}\right]; \tag{F.4.6}$$

L'estimation récursive de La formule F.4.6 peut être décomposée en deux étapes : E-step et M-step. Partant d'une valeur initiale $\boldsymbol{\theta}^{(0)}$ du paramètre défini dans Θ, une suite finie de valeurs du paramètre $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(K)}$ correspondant à une suite croissante de valeurs de la vraisemblance : $LL\left(\boldsymbol{\theta}^{(k)}\right) \leq LL\left(\boldsymbol{\theta}^{(k+1)}\right)$. L'itération $k-1$ de EM, avec $k = 0, \ldots K$, est constituée dans cet ordre, basée sur les deux étapes suivantes :

E-step (Expectation) : pour évaluer l'espérance conditionnelle de la vraisemblance des données complètes conditionnellement à $\boldsymbol{y}$ et à la valeur temporaire $k-1^{ème}$ du paramètre $\boldsymbol{\theta}^{(k-1)}$:

$$\mathcal{Q}\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k-1)}\right) = \mathbb{E}\left[LL_c(\boldsymbol{\theta}) \mid \boldsymbol{y}, \boldsymbol{\theta}^{(k-1)}\right], \tag{F.4.7}$$

M-step (Maximisation) : pour trouver le $\boldsymbol{\theta}^{(k)}$ afin de maximiser $\mathcal{Q}\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(k-1)}\right)$ calculée dans l'étape E :

$$\mathcal{Q}\left(\boldsymbol{\theta}^{(k)};\boldsymbol{\theta}^{(k-1)}\right) \geq \mathcal{Q}\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(k-1)}\right). \tag{F.4.8}$$

Ces deux étapes sont répétées jusqu'à ce qu'elles convergent par rapport à un certain critère vers l'un de ses maxima locaux :

$$LL\left(\boldsymbol{\theta}^{(k)}\right) \geq LL\left(\boldsymbol{\theta}^{(k-1)}\right). \tag{F.4.9}$$

Remarque F.4.1. Posé tel quel, l'algorithme garantit que la vraisemblance augmente à chaque itération, ce qui conduit donc à des estimateurs de plus en plus corrects. Ainsi, $LL\left(\boldsymbol{\theta}^{(k)}\right)$ est une suite croissante de valeurs de log-vraisemblance. Pour une preuve de la croissance de la vraisemblance d'une itération à l'autre, on se réfère à l'article original de Dempster et al. (1977). Notons aussi que le mécanisme itératif de l'algorithme est très astucieux, et débouche sur une amélioration progressive et réciproque des données manquantes et de la valeur du vecteur de paramètres $\boldsymbol{\theta}$.

F.4.3 Méthode Bayésienne

Avant d'aborder les fondements de l'imputation *via* l'approche Bayésienne, rappelons rapidement le cadre probabiliste du raisonnement Bayésien : toute inférence est basée sur la distribution a postériori des paramètres inconnus conditionnellement aux quantités observées. Pour nous, les paramètres inconnus sont $(\boldsymbol{\theta},\xi)$ et les quantités observées sont $\mathcal{D}_T^{\text{Obs}}$ et R. Par le biais du théorème de Bayes, la distribution *a postériori* peut être écrite comme

$$\boldsymbol{P}\left(\boldsymbol{\theta},\xi \mid \mathcal{D}_T^{\text{Obs}},R\right) = \frac{\boldsymbol{P}\left(R,\mathcal{D}_T^{\text{Obs}} \mid \boldsymbol{\theta},\xi\right)\pi\left(\boldsymbol{\theta},\xi\right)}{\int\int \boldsymbol{P}\left(R,\mathcal{D}_T^{\text{Obs}} \mid \boldsymbol{\theta},\xi\right)\pi\left(\boldsymbol{\theta},\xi\right)d\boldsymbol{\theta}d\xi}, \tag{F.4.10}$$

où $\pi(\bullet)$ est la distribution *a priori* appliquée aux paramètres $(\boldsymbol{\theta},\xi)$. Sous l'hypothèse MAR, en replaçant la dernière relation de la formule F.4.3 dans F.4.10, on obtient

$$\boldsymbol{P}\left(\boldsymbol{\theta},\xi \mid \mathcal{D}_T^{\text{Obs}},R\right) \propto \boldsymbol{P}\left(R \mid \mathcal{D}_T^{\text{Obs}},\xi\right)\boldsymbol{P}\left(\mathcal{D}_T^{\text{Obs}} \mid \boldsymbol{\theta}\right)\pi\left(\boldsymbol{\theta},\xi\right). \tag{F.4.11}$$

L'inférence Bayésienne uniquement sur $\boldsymbol{\theta}$ est basée sur la marginale *a postériori* obtenue en intégrant cette fonction sur le paramètre de nuisance ξ. Sous la condition d'ignorabilité, telle que définie dans (Little & Rubin, 2014; Rubin, 2004), la distribution *a priori* se factorise comme

$$\pi\left(\boldsymbol{\theta},\xi\right) = \pi_\theta\left(\boldsymbol{\theta}\right)\pi_\xi\left(\xi\right), \tag{F.4.12}$$

par conséquent, la marginale *a postériori* du $\boldsymbol{\theta}$ est

$$\begin{aligned}\boldsymbol{P}\left(\boldsymbol{\theta} \mid \mathcal{D}_T^{\text{Obs}},R\right) &= \int \boldsymbol{P}\left(\boldsymbol{\theta},\xi \mid \mathcal{D}_T^{\text{Obs}},R\right)d\xi \\ &\propto \boldsymbol{P}\left(\mathcal{D}_T^{\text{Obs}} \mid \boldsymbol{\theta}\right)\pi_\theta\left(\boldsymbol{\theta}\right)\int \boldsymbol{P}\left(R \mid \mathcal{D}_T^{\text{Obs}},\xi\right)\pi_\xi\left(\xi\right)d\xi \\ &\propto L\left(\boldsymbol{\theta} \mid \mathcal{D}_T^{\text{Obs}}\right)\pi_\theta\left(\boldsymbol{\theta}\right),\end{aligned} \tag{F.4.13}$$

telle que la proportionnalité est à un facteur multiplicatif qui ne comporte pas $\boldsymbol{\theta}$. Notons aussi que R n'apparaît pas dans la dernière relation de F.4.13 et donc, $\boldsymbol{P}\left(\boldsymbol{\theta} \mid \mathcal{D}_T^{\text{Obs}}, R\right) = \boldsymbol{P}\left(\boldsymbol{\theta} \mid \mathcal{D}_T^{\text{Obs}}\right)$: toute l'information sur les valeurs du paramètre $\boldsymbol{\theta}$ est encapsulée dans la marginale *a postériori* qui s'affranchit du mécanisme de réponse des données manquantes,

$$\boldsymbol{P}\left(\boldsymbol{\theta} \mid \mathcal{D}_T^{\text{Obs}}\right) \propto L\left(\boldsymbol{\theta} \mid \mathcal{D}_T^{\text{Obs}}\right) \pi_\theta\left(\boldsymbol{\theta}\right). \tag{F.4.14}$$

Soit $\boldsymbol{\theta}^{(0)}$ un tirage initial obtenu à partir d'une approximation de la distribution *a postériori* de $\boldsymbol{\theta}$. Pour une valeur $\boldsymbol{\theta}^{(k)}$ de $\boldsymbol{\theta}$ au $k-$ème tirage :

Imputation : estimer $\tilde{\mathcal{D}}_T^{\text{Miss}(k+1)}$, la valeur manquante avec une densité $\boldsymbol{P}\left(\mathcal{D}^{\text{Miss}} \mid \mathcal{D}^{\text{Obs}}, \boldsymbol{\theta}^{(k)}\right)$

Estimat Postérieure : estimer $\boldsymbol{\theta}^{k+1}$ avec une densité $\boldsymbol{P}\left(\boldsymbol{\theta} \mid \mathcal{D}^{\text{Obs}}, \tilde{\mathcal{D}}^{\text{Miss}(k)}\right)$.

F.5 Imputation multiple (MI)

F.5.1 Survol théorique

La méthode d'imputation multiple (Mutiple Imputation, en anglais) décrite en termes algorithmiques, se déroule généralement en trois phases :

1. Générer m copies de jeux de données du tableau incomplet, dans lesquels les valeurs manquantes sont tirées aléatoirement *via* la distribution de probabilités, puis imputer les m tableaux par une procédure appropriée (Molenberghs et al., 2014).
2. Estimer les paramètres d'intérêt Q de chaque tableau.
3. Agréger les résultats correspondant à l'étape précédente.

Cette méthode, dans le processus d'estimation des paramètres est équivalente aux méthodes basées sur le maximum de vraisemblance. En effet, la démarche repose sur une approche Bayésienne du modèle d'inférence : estimer l'espérance de la loi *a postériori* d'une certaine quantité Q. Cette quantité peut représenter le paramètre $\boldsymbol{\theta}$ qui régit la distribution, ou une proportion, une variance etc. La loi *a postériori* de Q est de la forme

$$\boldsymbol{P}\left(Q \mid \mathcal{D}_T^{\text{Obs}}, R\right) = \int \boldsymbol{P}\left(Q \mid \mathcal{D}_T^{\text{Obs}}, \mathcal{D}_T^{\text{Miss}}\right) f\left(\mathcal{D}_T^{\text{Miss}} \mid \mathcal{D}_T^{\text{Obs}}, R\right) d\mathcal{D}_T^{\text{Miss}}, \tag{F.5.1}$$

et sous l'hypothèse MAR, on obtient

$$\boldsymbol{P}\left(Q \mid \mathcal{D}_T^{\text{Obs}}, R\right) = \int \boldsymbol{P}\left(Q \mid \mathcal{D}_T^{\text{Obs}}, \mathcal{D}_T^{\text{Miss}}\right) f\left(\mathcal{D}_T^{\text{Miss}} \mid \mathcal{D}_T^{\text{Obs}}\right) d\mathcal{D}_T^{\text{Miss}}. \tag{F.5.2}$$

La factorisation de $\boldsymbol{P}\left(Q \mid \mathcal{D}_T^{\text{Obs}}, R\right)$ en deux composantes de deux lois *a postériori* permet d'effectuer plus facilement des tirages aléatoires. Ainsi, la première composante traite la loi de distribution du paramètre d'intérêt Q conditionnellement au jeu de données complet : la succession des valeurs observées $\mathcal{D}_T^{\text{Obs}}$ et les blocs de données manquantes $\mathcal{D}_T^{\text{Miss}}$. Le second terme de cette intégrale concerne la distribution des données manquantes conditionnellement aux données observées.

L'imputation des données manquantes par $f\left(\mathcal{D}_T^{\text{Miss}} \mid \mathcal{D}_T^{\text{Obs}}\right)$ permet d'approximer l'intégrale de la relation F.5.2 par la moyenne des lois *a postériori* évaluées sur les données générées $\left(\mathcal{D}_{T_m}^{\text{Miss}}\right)_{1\leq m\leq M}$ de m-copies du tableau $\mathcal{D}_T$:

$$\boldsymbol{P}\left(Q \mid \mathcal{D}_T^{\text{Obs}}, R\right) \approx \frac{1}{M}\sum_{m=1}^{M} \boldsymbol{P}\left(Q \mid \mathcal{D}_{T_m}^{\text{Miss}}, \mathcal{D}_T^{\text{Obs}}\right). \tag{F.5.3}$$

L'espérance Q conditionnellement aux observations $\mathcal{D}_T^{\text{Obs}}$ est alors approchée par

$$\begin{aligned} \mathbb{E}\left(Q \mid \mathcal{D}_T^{\text{Obs}}\right) &\approx \int Q \tfrac{1}{M}\sum_{m=1}^{M} \boldsymbol{P}\left(Q \mid \mathcal{D}_{T_m}^{\text{Miss}}, \mathcal{D}_T^{\text{Obs}}\right) dQ \\ &\approx \tfrac{1}{M}\sum_{m=1}^{M} \mathbb{E}\left(Q \mid \mathcal{D}_{T_m}^{\text{Miss}}, \mathcal{D}_T^{\text{Obs}}\right) = \tfrac{1}{M}\sum_{m=1}^{M} \hat{Q} \\ &= \bar{Q} \end{aligned} \tag{F.5.4}$$

où $\hat{Q}$ est l'estimation de Q pour le jeu imputé et $\bar{Q}$ est l'estimateur agrégé à partir des estimateurs $\left(\hat{Q}_m\right)_{1\leq m\leq M}$ de chaque tableau. Clairement, ces estimations doivent être complétées par une analyse de la variabilité due à l'échantillonnage (*intra-imputation*) et la variabilité liée à la présence de données manquantes (*inter-imputation*). La mesure de la variance de la loi *a postériori* peut être approximée par la décomposition des différentes variances :

$$\begin{aligned} \mathbb{V}\left(Q \mid \mathcal{D}_T^{\text{Obs}}\right) &\approx \left(\tfrac{1}{M}\sum_{m=1}^{M} \underbrace{\mathbb{V}\left(Q \mid \mathcal{D}_{T_m}^{\text{Miss}}, \mathcal{D}_T^{\text{Obs}}\right)}_{\approx \bar{U}_m}\right) + \\ &\qquad \underbrace{\left(\frac{1}{M-1}\sum_{m=1}^{M}\left(\hat{Q}_m - \bar{Q}\right)^2\right)}_{=B} \\ &= \bar{U}_{\text{Intra}} + B_{\text{Inter}} \end{aligned} \tag{F.5.5}$$

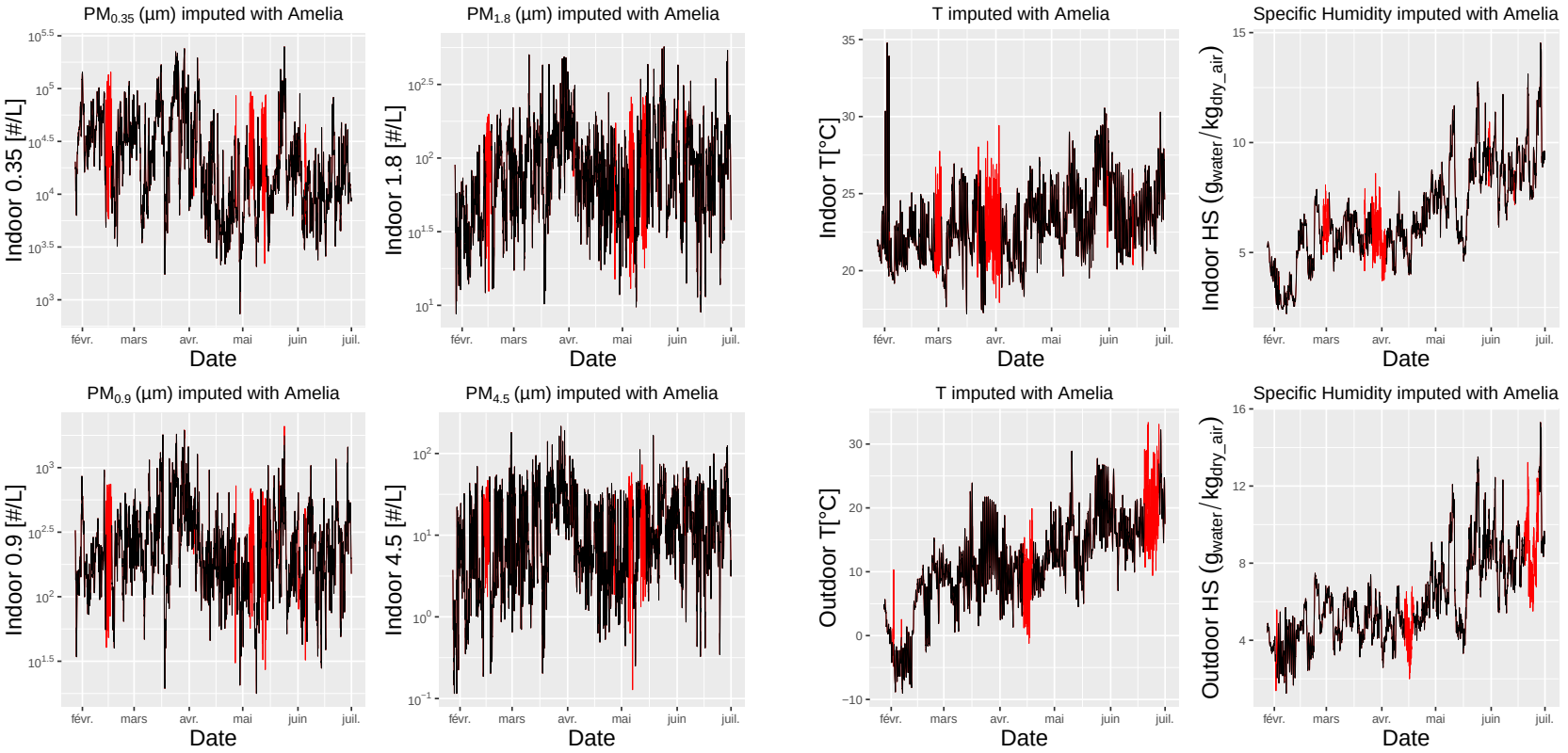
où $\bar{U}_{\text{Intra}}$ est l'estimateur de la variance *intra-imputation*, et B_{Inter} est l'estimation de la variabilité *inter-imputation*. Pour plus de discussion sur le raffinement de cette estimation et plusieurs autres détails techniques, notamment sur les méthodes MCMC, Bootstrape d'échantillonnage et le choix de M, on peut consulter les ouvrage suivants : Molenberghs et al. (2014), chapitre 12 et Rubin (2004), les chapitres 3 et 4.

F.5.2 Le programme Amelia II

La méthode d'imputation multiple Amelia II a été développé en 2011 par Honaker et al. (2011) ; elle s'appuie sur une combinaison de l'algorithme EM avec avec une approche Bootstrap (EMB). L'incertitude sur l'estimation des paramètres se propage en utilisant la simulation bootstrap (Honaker & King, 2010). Plus précisément, la fonction `amelia` implémentée dans le logiciel R (2015) par Honaker et al. (2011), génère M tables de données incomplètes par bootstrap, et sur chaque table, la matrice de covariance est estimée par l'algorithme EM. Ensuite, toutes les matrices de covariance sont utilisées pour imputer les

M tables. Le modèle repose sur une hypothèse de normalité des données, alors une transformation de type logarithmique est nécessaire pour réduire le biais crée par cette contrainte.

Le programme Amelia II donne généralement de très bons résultats pour les tables de grandes dimensions et pour les données de type coupes transversales. En revanche, pour les séries temporelles, notamment de haute résolution, Amelia II amplifie la variance globale, et ce d'autant plus l'autocorrélation est forte.



(a) *Interpolation des données manquantes des concentrations en nombre des PM intérieurs par la méthode* Amelia II.

(b) *Interpolation des données manquantes des paramètres climatiques : température intérieur et extérieure et l'humidité spécifique intérieure et extérieure.*

FIGURE F.5.1 – *Exemple d'imputation des données manquantes par le programme* Amelia II. *Les données brutes sont issues de la campagne de 2012 dans l'espace paysager.*

La Figure F.5.1 illustre une estimation des données manquantes par le programme 𝔸melia𝕀𝕀, pour les séries issues de la campagne 2012 de l'espace paysager. Clairement, les séries estimées sont beaucoup plus fluctuantes avec cette méthode que les séries imputées avec la méthode basée sur les splines (qui effectuent un lissage). Pour certains polluants, cette amplification de variabilité peut être utile, notamment pour le CO_2 ou les particules de tailles moyennes. En revanche, pour les paramètres climatiques, l'estimation par 𝔸melia 𝕀𝕀 présente plus de fluctuations de nuisances, mais qui reste néanmoins acceptable.

Dans le contexte de cette thèse, les modèles d'imputation appliqués ont été traités de manière à avoir des séries complètes cohérentes aux observations, notamment pour les données au pas de temps d'une minute. Le traitement était fastidieux surtout lorsqu'une méthode est adaptée pour un polluant mais pas pour un autre.

La conséquence directe du problème de données manquantes sur le modèle de prévision est la propagation de l'erreur liée à l'imputation sur les valeurs prédites, notamment lorsque les données manquantes se situent au voisinage de l'extrémité gauche de la série. Cette section a montré néanmoins que les méthodes développées dans la littérature, et que nous avons modifié pour certaines, sont applicables sur des séries incomplètes de la QAI. Il n'en demeure pas moins qu'on ne peut pas attendre une imputation sans biais. Ceci est lié en partie à la quantité de données manquantes et aux structures de variabilité des séries.

En définitive, le problème d'imputation des séries temporelles est loin d'être simple et résolu aujourd'hui. Il est malaisé de savoir avec précision la réalité "physique" de l'environnement qui a généré les "trous", les causes de leur présence et leurs impacts sur un modèle de prévision, au point que certaines séries ont été jusqu'à remettre en cause leur validité.

F.6 Conclusion et perspectives

Ce chapitre s'est intéressé à l'imputation des données manquantes. En montrant des applications sur différents jeux de données, les méthodes d'imputation peuvent être considérées comme utiles à condition d'analyser en amont chaque série. Ces modèles nécessitent des corrections afin d'avoir une estimation acceptable (physiquement interprétable) de la valeur manquante. Ainsi, chaque sortie des algorithmes d'imputation a été contrôlée afin d'éviter les valeurs aberrantes. En effet, aucune valeur d'imputation inférieure ou supérieure au minimum ou au maximum de l'ensemble de la série n'est acceptée dans les algorithmes de gestion des valeurs manquantes. Cette procédure est effectuée afin de minimiser la contribution des sources d'incertitudes liées à l'imputation sur les erreurs générées par les modèle de prévision.

D'autres problèmes restent ouverts, notamment dans le cadre des séries temporelles. La corrélation croisée des variables et l'autocorrélation au sein de la même variable sont rarement évoquées dans la littérature. La gestion de ces contraintes en présence des données manquantes et leur influence pour un modèle de prévision est particulièrement complexe.

ANNEXE G

RÉSULTATS SUPPLÉMENTAIRES : LES PRÉVISIONS

G.1 PACF des résidus des modèles STL-ARIMA

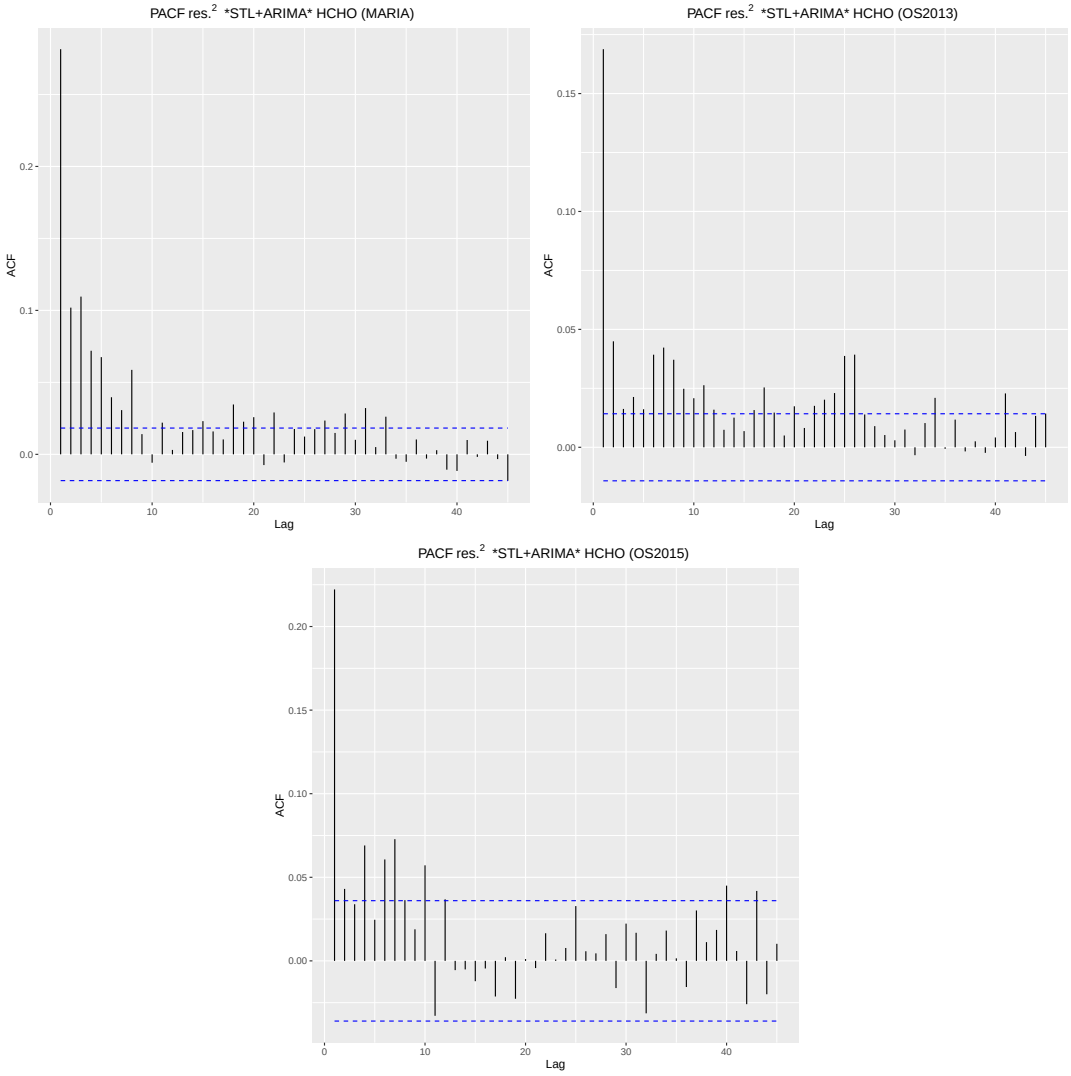


FIGURE G.1.1 – *PACF des résidus des modèles .*

RÉSUMÉ

Les caractéristiques des fluctuations des concentrations de polluants dans un environnement intérieur normalement occupé dépendent fortement de plusieurs paramètres, en particulier de l'occupation et de l'activité des occupants, qui altèrent de manière considérable la nature statistique de leur variabilité temporelle. Le travail de cette thèse concerne l'analyse d'un cas réel d'environnement intérieur de type bureau (individuel ou espace paysager). Les polluants cibles formaldéhyde et particules ont été enregistrées sur une longue période et avec un pas de temps fin. Cette thèse s'articule autour de trois axes de recherche : *(i)* la caractérisation des fluctuations des concentrations des polluants cible ; *(ii)* la détermination de la variabilité des sources de ces fluctuations et *(iii)* la prévision des concentrations de ces polluants. Cette recherche a été abordée uniquement par une approche statistique. Le premier axe concerne la détermination des caractéristiques communes partagées par les différents polluants, i.e. l'extraction et la quantification des invariants statistiques des fluctuations. Le deuxième axe porte, à l'aide des approches de type séparation aveugle des sources, sur l'estimation les déterminants des sources de variabilité. Enfin, le troisième axe est consacré à la prévision des fluctuations de la concentration des polluants. Ce dernier découle directement des résultats obtenus dans le premier axe. L'analyse des séries temporelles pour ce type de données (hautes fréquences) doit prendre en compte l'échelle de temps sur laquelle évoluent plusieurs microstructures. Les travaux menés dans cette thèse ont utilisé plusieurs outils, en l'occurrence l'analyse spectrale (propriétés de dépendance à long terme par la mesure fractale et la statistique R/S), la mesure de $\Omega-$prédictibilité, ainsi que la décomposition des séries en composantes latentes (STL, SSA et SBD). Ces outils ont permis d'identifier les classes de modèles appropriés pour l'étape de prévision.

Les non-linéarités apparaissent surtout sous forme de changements abrupts qui se greffent au sein de l'évolution régulière du système dynamique. Les modèles autorégressifs à changement de régime ont été abordés dans cette perspective et les modèles des systèmes dynamiques (dynamique du chaos) peuvent mettre en évidence des propriétés qualitatives des séries temporelles. Un nouveau type de modèles de prévision a été proposé pour répondre aux exigences de la nature des données hautes fréquences. Ce modèle consiste en l'introduction d'une étape de décomposition des séries en bandes spectrales (SBD) couplée avec une étape de modélisation par des modèles autorégressifs à seuil (TAR) ou par la dynamique du chaos : FFT-(TAR/Chaos). Les résultats montrent que le prétraitement par décomposition en bandes spectrales des séries temporelles améliore sensiblement la prévision des concentrations de certains polluants (HCHO et particules fines). L'analyse de la variabilité des sources ainsi que de leurs contributions ont été abordés par les méthodes de séparation aveugle des sources qui sont basées sur une factorisation matricielle sous contrainte statistique d'indépendance (ACI) ou sous contrainte de non-négativité. Ces méthodes ont été utilisées pour les concentrations de particules. Les composantes peuvent être interprétées en termes de contributions variables selon le diamètre des particules, mais également en termes de fluctuations ou d'occurrence au fil du temps.

Mots-clés : Fluctuations, structures de variabilité, décomposition en bandes spectrales, modèles à changement de régime, dynamique du chaos, prévision non-linéaire, prédictibilité, qualité de l'air intérieur.