MÉMOIRE D'HABILITATION DE l'UNIVERSITÉ de NANTES

Spécialité
Mathématiques Appliquées

École Doctorale Sciences & Technologies de l'Information et Mathématiques

Présentée par

Lise BELLANGER

Pour obtenir
l'HABILITATION à Diriger des Recherches

CONTRIBUTIONS À L'EXPLORATION DE DONNÉES ENVIRONNEMENTALES, ÉCOLOGIQUES, MÉDICALES ET ARCHÉOLOGIQUES

Soutenue le 6 février 2015

devant le jury composé de :

Natale Carlo LAURO, Professeur à l'Université de Naples, Italie	Rapporteur
Carlos MATRÁN BEA, Professeur à l'Université de Valladolid, Espagne	Rapporteur
Gilbert SAPORTA, Professeur émérite au CNAM, Paris	Rapporteur
David CAUSEUR, Professeur à l'Agrocampus Ouest, Rennes	Examinateur
Bruno FALISSARD, Professeur des Universités et Practicien Hospitalier à Paris-Sud	Examinateur
Anne PHILIPPE, Professeur à l'Université de Nantes	Examinateur
Véronique SÉBILLE, Professeur des Universités et Practicien Hospitalier à Nantes	Examinateur

A toutes celles et ceux avec qui je n'ai plus l'occasion d'échanger ;

mais que je n'oublie pas.

Deviens ce que tu es

Aphorisme attribué à Pindare (518 av J.-C. - 438 av. J.-C.),
repris par F. W. Nietzsche (1844 - 1900).

Remerciements

J'adresse mes plus vifs remerciements à Didier Dacunha-Castelle et Richard Tomassone, mes directeurs de thèse, qui m'ont constamment soutenue et encouragée. C'est toujours un plaisir d'échanger et de collaborer ensemble : la rédaction avec Richard d'un ouvrage, publié en 2014, sur l'*exploration de données* a été une très belle aventure !

J'exprime toute ma gratitude et mes remerciements respectueux à Natale Carlo Lauro, Carlos Matrán Bea et Gilbert Saporta qui m'ont fait l'honneur d'accepter de consacrer une partie de leur précieux temps en rapportant ce mémoire d'habilitation.

Je remercie vivement David Causeur, Bruno Falissard, Anne Philippe et Véronique Sébille d'avoir accepté d'examiner mon travail de recherche. Je n'ai jamais eu l'occasion de travailler avec David et Bruno ; mais j'admire le travail de recherche qu'ils mènent l'un comme l'autre avec une grande ouverture d'esprit dans des domaines bien différents. Je suis donc ravie qu'ils aient accepté d'évaluer le mien. Merci Anne de m'avoir encouragée à soutenir cette HDR ; pour les échanges très réguliers que nous avons autour de l'enseignement de la statistique à Nantes ainsi que ceux si passionnants et fréquents liés à nos travaux de recherche respectifs autour de la datation en archéologie. Enfin, je collabore avec Véronique dans le cadre de la recherche mais aussi de l'enseignement de la statistique. J'ai donc pu apprécier sa pugnacité, son dynamisme et sa joie de vivre. Je lui suis reconnaissante du soutien constant qu'elle m'apporte, encore ici en acceptant de faire partie de mon jury.

Merci à tous mes collaborateurs les plus proches, de plus ou moins longue date, sans qui ces travaux n'existeraient pas : Denis Baize, Anik Brind'Amour, Fanny feuillet, Jean-Benoit Hardouin, Philippe Husi, Pascale Jolliet, Pierre Legendre, Jean-Paul Lucas, Stéphanie Mahévas, Corinne Mandin, Véronique Sébille, Richard Tomassone, Verena Trenkel, Caroline Victorri-Vigneau ; ainsi que Christian Dina, Matilde Karakachoff, Soléna Le Scouarnec, Florianne Simonet, Richard Redon, Hervé Le Marec et tous les participants du programme VaCaRMe soutenu par la Région Pays de La Loire, et enfin plus récemment Brice Trouillet et tous les membres du GIS VALPENA. Merci aussi à tous

les autres collaborateurs que je ne nomme pas explicitement ; mais pour qui j'ai une pensée. Mes échanges avec vous tous constituent le sel du travail de recherche que je présente dans ce mémoire. Parfois, nos parcours scientifiques, très différents, ont nécessité un peu de temps avant de lever les *a priori*, se comprendre et aboutir à un langage commun nécessaire pour collaborer ; mais une fois ces étapes passées : quelle richesse !

Je remercie aussi tous les étudiants et les étudiantes qui m'ont fait confiance pour les encadrer, notamment Jean-Paul Lucas mais aussi très récemment Elodie Persyn pour le co-encadrement de leur thèse. Sans eux, une partie des travaux présentés ici n'existerait pas.

Je remercie également mes collègues du laboratoire de mathématiques Jean Leray UMR CNRS 6629 pour les discussions riches et les conseils qu'ils m'ont prodigués tout au long de ces années. Merci aussi aux membres de l'EA 4275 – SHERE dont je suis membre associé depuis sa création en 2008 : la confrontation de nos points de vue, lors du séminaire ou d'échanges plus informels, est toujours très fructueuse pour moi.

Enfin, je remercie mes proches, mes amis et plus particulièrement ma petite famille pour leur soutien, leur patience et leur affection.

Table des matières

Introduction

Ce manuscrit présente les principaux travaux que je conduis en tant que statisticienne, au sein du laboratoire de Mathématiques Jean Leray de Nantes (LMJL). Je suis arrivée au LMJL en 2000 et suis la première d'une longue série de probabilistes et statisticien(ne)s qui ont ensuite suivi mon recrutement (Philippe Carmona (Pr) en 2002, Anne Philippe (Pr) en 2005, Frédéric Lavancier (MC) en 2006, Nicolas Pétrellis (MC Chaire d'excellence Université-CNRS 2009-2014) en 2009 puis Paul Rochet (MC) en 2012).

A mon arrivée, tout était à faire ; et c'est ce qui m'a plu ! Prendre la responsabilité d'un certain nombre d'enseignements de Statistique, construire une formation dédiée aux Probabilités et à la Statistique au département de mathématiques ; mais aussi participer à la construction ou à l'élaboration des programmes de nouvelles formations, professionnalisantes ou à la recherche, dans lesquelles la Statistique avait toute sa place et puis enfin encadrer des étudiants. Cela m'a demandé beaucoup d'énergie ainsi qu'à mes collègues ; mais depuis mon recrutement :

- En 2004, un Master professionnel d'Ingénierie Mathématique, formation professionnelle de haut niveau en Statistique, Probabilités et en Calcul Scientifique, a vu le jour. J'ai été successivement responsable en Maîtrise d'Ingénierie Mathématique du module *Statistique inférentielle (2000-2001), P*robabilités (2000-2002), *Statistiques – apprentissage du logiciel de statistique R* (2001-2005), puis en Master 1 *Analyse des données* (depuis 2009) ; et en Master 2 du module *Modèle linéaire et extensions* (2004-2012), *Plans d'expérience* (2004-2006), *Data Mining* (2008-2012) et *Apprentissage* (depuis 2012).
- En 2006, un Master 2 « Modélisation en Pharmacologie Clinique et Epidémiologie » co-habilité avec les universités Angers-Brest-Nantes-Tours et Poitiers a été créé à l'Université de Rennes 1. Il a pour objectifs de former des chercheurs ou des professionnels de haut niveau capables non seulement de concevoir et d'analyser de façon approfondie des données de tous types de protocoles de recherche clinique (pharmacologie clinique, essai thérapeutique) et épidémiologique (recherche étiologique, évaluation de méthodes diagnostiques, recherche de facteurs

pronostiques), mais aussi de développer une recherche méthodologique adaptée à ces différents domaines. J'ai été successivement responsable du module *Régression linéaire multiple et plans d'expérience* (2006-2012) avec Jean-Louis Auget (Pr, UFR de Pharmacie, Nantes), puis depuis 2012 *Stratégie de modélisation et Modèle linéaire.* J'interviens aussi dans le module *Méthodes Statistiques avancées* dans lequel je présente les *Plans d'expérience* depuis 2012.

- En 2010, le parcours « Responsable de production en industrie bois » de la licence professionnelle « Bois et ameublement » spécialité construction et production bois était créé. Cette licence professionnelle est un partenariat entre partenariat entre l'UFR de Sciences et Techniques de l'Université de Nantes et l'Ecole .Supérieure du Bois (ESB). J'ai contribué à la création du parcours « Responsable de production en industrie bois » et en ai été responsable entre 2010et 2012. J'y enseigne une *première approche de la Statistique descriptive et inférentielle* depuis son ouverture.
- En 2012, un parcours de Master 1 « Bioinformatique – Biostatistique » a vu le jour dans le master Biologie Santé. Il permet de poursuivre ensuite en Master 2 "Modélisation en Pharmacologie Clinique et Epidémiologie - MPCE" ou "Bioinformatique". Je suis responsable du module *Introduction à l'analyse exploratoire multidimensionnelle* depuis sa création.
- J'encadre de nombreux étudiants de niveaux M1 et M2 sur des durées limitées (entre 4 et 6 mois). Je participe à des comités de suivi de thèse. J'ai co-dirigé avec Véronique Sébille (PU PH, UFR de Pharmacie, Nantes) un étudiant en thèse, Jean-Paul Lucas, entre 2009 et 2013. Enfin depuis octobre 2014, je co-encadre avec Christian Dina (IR CNRS, Institut du Thorax, Nantes) la thèse d'Elodie Persyn. Je présenterai plus amplement le travail de thèse de Jean-Paul Lucas et j'évoquerai les objectifs de la thèse d'Elodie Persyn dans la suite de ce document.

Mon travail de thèse, intitulé *Statistique de la pollution de l'air. Méthodes mathématiques. Applications au cas de la région parisienne* (Bellanger, 1999), comportait à la fois une partie statistique théorique sur les franchissements de niveaux de processus ponctuels non stationnaires ((Bellanger & Perera, 1999), (Bellanger & Perera, 2003)) et une partie appliquée à la pollution de l'air ((Bellanger & Tomassone, 1999), (Bellanger & Tomassone, 2000)).

Depuis, mon goût pour les applications de la statistique et pour le travail pluridisciplinaire a fait pencher mon travail de recherche vers la statistique appliquée. En fonction de mes appétences, j'ai donc fait le choix de poursuivre certaines collaborations de recherche débutées pendant ma thèse et d'en laisser certaines autres, liées aux travaux de statistique théorique, en sommeil. Mais, j'en ai aussi noué de nouvelles ; parfois d'ailleurs en lien directe avec les formations créées ; comme celle avec Véronique Sébille à l'origine de la création en 2008 de l'EA 4275-SPHERE "Biostatistique, Recherche Clinique et Mesures Subjectives en Santé" à Nantes.

Mon travail de recherche s'articule aujourd'hui principalement autour de l'*exploration de données* dans des domaines aussi variés que l'environnement, l'écologie, la santé et l'archéologie. J'y ai d'ailleurs consacré un ouvrage scientifique et de réflexion, écrit avec Richard Tomassone (Bellanger & Tomassone, Exploration de données et Méthodes statistiques : data analysis & data mining avec R, 2014). L'exploration de données est un travail de longue haleine qui nécessite de réelles collaborations pluridisciplinaires. C'est seulement à cette condition que le statisticien peut, après s'être imprégné de l'autre discipline, traduire les problématiques liées aux données traitées en langage mathématique, déterminer la ou les méthodes les plus adaptées pour résoudre le problème auquel il s'est attelé et enfin proposer des solutions, des réponses. Sans reprendre la démarche complète développée dans le chapitre 1 de (Bellanger & Tomassone, 2014) ; il me paraît nécessaire de souligner la nécessité de travailler par étapes et va et vient, entre les objectifs $\boldsymbol{\mathcal{O}}$, le modèle $\boldsymbol{\mathcal{M}}$ et les données $\boldsymbol{\mathcal{D}}$. (Figure 1).

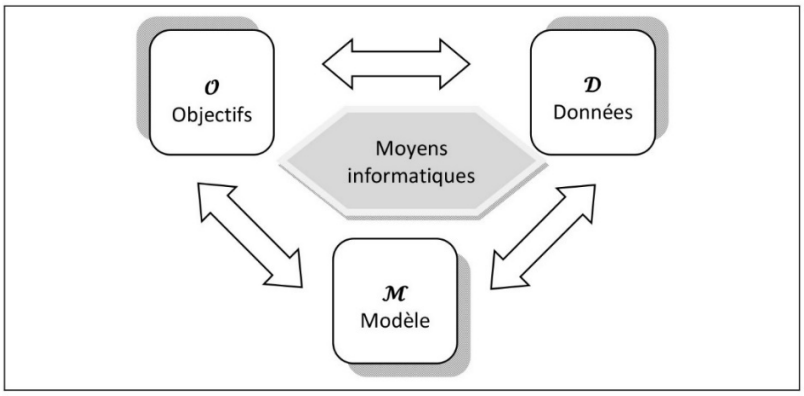


Figure 1 - Etapes fondamentales d'une analyse statistique.

Comme nous l'indiquons dans (Bellanger & Tomassone, 2014) :

La modélisation statistique est une forme d'art qui s'appuie sur un certain nombre d'outils mathématiques. C'est un instrument indispensable pour faire un apprentissage d'une réalité toujours complexe. Par sa capacité à simplifier la description d'une « situation » elle permet d'en déceler les traits essentiels. Il n'existe pas d'outil « toujours meilleur » pour n'importe quel utilisateur. Les outils les plus utiles sont ceux qui rendent aisé le plus grand nombre de tâches dans une exploration de données qu'un utilisateur doit réaliser.

Aussi ai-je toujours privilégié, dans les travaux qui seront présentés par la suite, des réponses statistiques, bien sûr adaptées et validées ; mais autant que possible, simples. Très souvent l'utilisation de critères statistiques permet de choisir parmi différents modèles (ou scénarios) possibles le « meilleur » ; mais parfois ces critères ne sont pas discriminants. L'avis d'un expert s'avère alors déterminant et indispensable pour décider du modèle à retenir sur des critères non-statistiques ; mais tout aussi pertinents comme nous le verrons dans l'étude des profils de consommations de médicaments psychotropes au chapitre 2. Ce manuscrit ayant pour vocation d'être une synthèse de mes travaux de recherche, les méthodes statistiques et les résultats obtenus sur les données explorées y sont décrits et mis en perspective. Le lecteur intéressé pourra par ailleurs consulter pour plus de détails les publications associées.

Ce mémoire se divise en trois parties correspondant aux trois grands domaines dans lesquels j'ai eu ou j'ai l'occasion d'explorer et traiter des données : l'environnement tout d'abord, puis la médecine et enfin l'archéologie.

Ces travaux font appel à des méthodes statistiques variées que j'ai essayé de regrouper dans le tableau ci-dessous :

Tableau 1 - Synthèse des méthodes statistiques.

Parties Méthodes statistiques	Environnement	Médecine	Archéologie
Méthodes factorielles : ACP, AFC, AFCM, MEM,…	x	x	x
Méthodes de classification et de classement	x	x	
Exploration de données spatiales	x		x
Modélisation : linéaire, linéaire généralisée, mixte	x	x	x
Théorie des sondages	x		
Théorie des valeurs extrêmes	x	x	
Ré-échantillonnage	x	x	x
Tests d'association par permutations		x	
Simulations	x	x	

Dans chacun de ces grands champs d'étude, j'ai indiqué en fin de chaque chapitre les perspectives et les projets de recherche que je souhaite mener à bien dans les années à venir.

Chapitre 1 : Exploration de données environnementales et écologiques (depuis 1995)

L'environnement a été le premier domaine dans lequel j'ai été amenée à travailler. Les problématiques auxquelles j'ai été et je suis encore confrontée sont très diverses : pollution (atmosphérique, des sols ; mais aussi à l'intérieur des logements français) et écologie marine.

1.1 LA POLLUTION (DEPUIS 1995)

1.1.1 Air extérieur : l'ozone troposphérique (O_3) (1995 - 2004)

Collaborateurs sur ce thème :

- AIRPARIF (surveillance de la qualité de l'air en Ile-de-France) : V. Bonneau, E. Gilibert ;
- AIR NORMAND (surveillance de la qualité de l'air en Haute Normandie) : M. Bobbia ;
- AgroParisTech/INRA (dpt MMIP, Paris) : R. Tomassone ;
- Université de Marseille (GREQAM) : C. Deniau, B. Ghattas ;
- Université Paris-Sud (Laboratoire de Mathématiques, Equipe de Probabilités, Statistique et Modélisation) : L. Bel, G. Ciuperca, D. Dacunha-Castelle, M. Misiti, Y. Misiti, G. Oppenheim, J.-M. Poggi.

J'ai tout d'abord commencé à explorer des données de pollution atmosphérique au cours de mon stage de DEA en 1995 intitulé *Étude de l'évolution du taux d'ozone troposphérique sur le site de Neuilly/Seine, pour la période 1989-1994* , puis de ma thèse (1996-1999) intitulée *Statistique de la pollution de l'air. Méthodes mathématiques. Applications au cas de la région parisienne* et enfin dans le cadre d'une collaboration entre AIRPARIF (organisme chargé de la surveillance de la qualité de l'air en région parisienne) et l'Université d'Orsay autour de la prévision des pointes de pollution en région parisienne.

Deux grandes questions m'ont intéressée : la prévision des pics d'ozone journalier et la compréhension des valeurs extrêmes de ce même polluant en région parisienne.

1.1.1.1 Prévision des pics d'ozone

Ce travail a porté sur la prévision à court terme, le matin pour l'après-midi, des concentrations maximales d'ozone des 8 stations du réseau d'alerte et du niveau d'alerte. En effet, les mesures effectuées sur ce réseau permettent de déclencher des procédures d'informations administratives et du public. A l'époque, pour chaque station, 3 niveaux avaient été définis journalièrement : 0 si le maximum d'Ozone était inférieur à 130 $\mu g/m^3$, 1 s'il est compris entre 180 et 360 $\mu g/m^3$ et 3 s'il est supérieur à 360 $\mu g/m^3$. Une alerte de niveau *i* était ensuite déclenchée si au moins deux stations de mesure enregistraient un niveau supérieur ou égal à *i*. Les niveaux 0 étaient de loin les plus nombreux, les niveaux 3 restant beaucoup plus rares. J'ai participé, en tant qu'ingénieur à AIRPARIF, à la réalisation de ce système de prévision des pointes de pollution à court terme (apprentissage des données par application de méthodes d'analyse des données, choix de modèles statistiques couvrant un large spectre de méthodes possibles, programmation (SAS, MATLAB) et à l'implémentation de procédures permettant l'automatisation de l'ensemble de la chaîne de calcul. Ces réalisations sont des rapports techniques, une publication internationale (Bel, et al., 1998) et une publication nationale (Bel, et al., 1999).

Les données utilisées étaient d'une part des données de pollution : O_3, NO, NO_2 mesurés sur les huit station du réseau d'alerte et d'autre part des données météorologiques : température (au sol, à 40m, à 100m), intensité et direction du vent (à 58m, à 110m). Le corpus de données, relativement volumineux, comprenait les jours d'été de 1992 à 1996. Le nettoyage des données pour ne conserver que les données fiables, l'imputation de celles qui pouvaient l'être et la suppression des biais de mesure liés au changement d'appareil au court de l'étude ont été autant d'étapes préliminaires nécessaires, longues et délicates.

L'absence d'éléments de référence nous avait incité à ajuster, valider en interne et externe puis comparer quatre modèles statistiques issus de méthodes variées : *non linéaire non paramétrique général* et une spécialisation *additive*, *classification-discrimination-régression* (par la suite noté *CDR*) et enfin *CART*[1].

[1] Acronyme de *Classification And Regression Trees*.

J'ai de mon côté travaillé avec Gabriela Ciuperca et Richard Tomassone sur la méthode *CDR* qui utilise des techniques classiques (Classification non supervisée, Analyse Discriminante et Régression linéaire). Cet assemblage de méthodes statistiques était et reste assez utilisé dans le domaine de la prévision de la pollution. Notre originalité a consisté à construire une partition des jours à partir des profils de concentration d'ozone journalier, puis se baser sur les types de temps déduits des variables météorologiques pour prévoir l'affectation d'un jour à un profil donné de concentration d'ozone journalier. Pour chacunes des huit stations du réseau d'alerte nous avions procédé comme suit :

1. Classification non supervisée des profils journaliers d'ozone
2. Analyse Factorielle Discriminante (*AFD*), en fonction des composantes principales construites à partir d'une centaine de variables disponibles jusqu'à 6h TU. L'*AFD* permettait de classer un jour donné en fonction de sa probabilité *a posteriori* d'appartenance à une classe déterminée dans l'étape précédente
3. Dans chaque classe construite à l'étape 1, Régression linéaire sur les composantes principales construites à partir des mêmes variables que celles ayant servi dans l'*AFD* et bien sur sélectionnées en fonction de leur significativité dans le modèle de régression.

L'été 1997, assez pollué (16 niveaux 1 et 4 niveaux 2 sur 138 jours étudiés) a aussi par la suite servi de période de validation externe. Les méthodes qui fournissaient les meilleurs résultats en terme d'alerte et de précision des valeurs d'ozone estimées étaient la méthode additive et *CDR*. Les prévisions à court terme obtenues par nos différents modèles ont constitué la base des informations qui étaient ensuite diffusées au public par AIRPARIF. Néanmoins de nombreuses améliorations devaient encore être apportée à ce travail. L'un des points principaux à travailler concernait les niveaux d'ozone élevés voir extrêmes, mal prévus par nos modèles. Je me suis donc tout naturellement concentrée sur cet aspect pendant ma thèse et quelques années après.

Je ne m'interesse plus directement à cette problématique de prévision journanlière des pics de pollution ; mais Charlotte Songeur, ancienne étudiante du Master professionnelle ingénierie mathématique de Nantes, que j'ai eu l'occasion d'encadrer à de multiples reprises a été embauchée en 2006 à la suite de son stage de Master en tant qu'

ingénieur d'étude à AIRPARIF. Depuis, quelques étudiants nantais sont allés effectuer leur stage de Master dans ce domaine.

Qu'en est-il aujourd'hui ? Depuis ce travail sur la prévision des épisodes de pollution, de nombreuses évolutions ont eu lieu. Le nombre de stations de mesures a augmenté : 18 stations font maintenant partie de la procédure d'alerte Ozone en Ile-de-France. Comme l'indique le Tableau 2 ci-dessous , la règlementation elle-même, en matière de procédure d'alerte, est différente : le seuil d'information est fixé par exemple à 180µg/m^3 en moyenne horaire sur au moins 3 stations simultanément alors que le premier niveau d'alerte est de 240µg/m^3 en moyenne horaire (toujours sur au moins 3 stations simultanément).

Tableau 2 - Ozone troposphérique : valeurs cibles, seuils règlementaires en 2014.

Polluant	Valeurs cibles	Objectifs à long terme	Seuil d'information	Seuils d'alerte
Ozone (O3)	**Pour la protection de la santé :** En moyenne sur 8 heures : 120 µg/m³, à ne pas dépasser plus de 25 jours par an (moyenne calculée sur 3 ans). **Pour la protection de la végétation :** AOT 40* de mai à juillet de 8h à 20h : 18 000 µg/m³.h (moyenne calculée sur 5 ans).	**Pour la protection de la santé :** En moyenne sur 8 heures : 120 µg/m³. **Pour la protection de la végétation :** AOT 40* de mai à juillet de 8h à 20h : 6 000 µg/m³.h.	**en moyenne horaire :** 180 µg/m³.	**Information :** 240 µg/m³ en moyenne horaire. **Actions à court terme obligatoires :** 240 µg/m³ pendant 3 heures consécutives.

* AOT 40 (exprimé en µg/m³.heure) signifie la somme des différences entre les concentrations horaires supérieures à 80 µg/m³ (= 40 ppb ou partie par milliard) et 80 µg/m³ durant une période donnée en utilisant uniquement les valeurs sur 1 heure mesurées quotidiennement entre 8 heures et 20 heures.

La prévision de la qualité de l'air utilise deux types de modèles :

- le modèle déterministe de chimie-transport CHIMERE qui comporte de nombreuses données d'entrée parmi lesquelles des données météorologiques, des inventaires d'émissions et des données aux limites du domaine étudié. Les prévisions sont déterminées pour chaque station pour les échéances de la veille, du jour même, du lendemain et du surlendemain, pour le dioxyde d'azote (NO2) et l'Ozone[2].
- En parallèle de l'Adaptation Statistique (*AS*) au niveau des stations pour J+0 et J+1 (*régressions linéaires multivariées*, *avec* ou *sans classes*). Ces prévisions sont ensuite utilisées pour réajuster le modèle CHIMERE avec les méthodes d'assimilation de données (pour la cartographie). Cependant, l'*AS* n'apportait pas d'amélioration notable dans le cas de l'ozone, donc à ce jour aucun **modèle de prévision statistique n'est implémenté pour l'ozone**. L'*AS* fonctionne pour le maximum horaire de NO2 et la moyenne des PM10.

De son côté, l'INERIS[3] a mis en place depuis 2003 le système *Prév'air*[4] qui diffuse quotidiennement des prévisions et des cartographies de qualité de l'air, issues de simulations numériques avec de l'*AS*, à différentes échelles spatiales (le Globe, l'Europe et la France pour l'ozone).

1.1.1.2 Hautes valeurs d'ozone

Au début de ma thèse, la notion de réchauffement climatique n'était pas encore une évidence et rencontrait de nombreux partisans ; mais aussi de nombreux détracteurs. Or comme nous le savons maintenant, avec le recul, l'impact d'un tel phénomène complexe a de nombreuses répercussions ; mais aussi de nombreuses causes. La question que je me suis posée était de déterminer si les événements de forte pollution par l'ozone troposphérique devenaient de plus en plus fréquents ou bien s'ils étaient seulement une conséquence de changements météorologiques affectant les conditions de formation

---

[2] Le lien suivant présente la plate-forme ESMERALDA ainsi que les prévisions : http://www.esmeralda-web.fr/?rubrique=esmeralda&article=index.

[3] Acronyme de *Institut National de l'Environnement Industriel et des Risques*.

[4] Voir le site internet associé : http://www.prevair.org/fr/general_prev.php.

d'ozone. En effet, les conditions météorologiques telles que la température journalière et la vitesse du vent jouent un grand rôle dans la survenue des pics de pollution. Les variations annuelles des conditions météorologiques peuvent donc masquer toute tendance à long (moyen)-terme de l'ozone à relier à des changements dans les émissions de précurseurs d'ozone. De plus, du point de vue de la santé publique, il était important d'essayer de dégager une tendance à moyen terme dans les épisodes de forte pollution ; ces résultats pouvant permettre une meilleure compréhension de la relation entre les épisodes de forte pollution et leurs effets à moyen terme sur la santé (augmentation/diminution du nombre de personnes atteintes d'allergies, d'insuffisances respiratoires, d'asthme...).

La théorie des valeurs extrêmes s'est très largement développée ces dernières décennies ; et de nombreux ouvrages de synthèse ont été publiés tels que (Leadbetter, Lindgren, & Rootzen, 1983), (Embrechts, Klüppelberg, & Mikoch, 1999; Falk, Hüsler, & Reiss, 2004) ou (Coles, 2001). Ceci a conduit au développement de nombreuses approches possibles permettant de modéliser les extrêmes dans des suites de données issues de domaines les plus variés (finance, assurance, environnement, …). Elles dépendent de la structure et de la complexité des données. Nous avons choisi une approche par processus ponctuels, fondée sur les dépassements d'un seuil élevé. Nous pouvons les classer de la manière suivante :

- *Etude et modélisation du maximum annuel ou d'une statistique d'ordre* : si les suites sont assez longues, une méthode classique consiste à modéliser le maximum annuel de périodes consécutives de taille égale (par exemple les années, mois ou jours) des séries (supposées *iid*[5]) par une des distributions des valeurs extrêmes comme dans (Gumbel, 1958). Mais cette méthode présente un défaut majeur : elle demande un nombre important de données, difficiles à obtenir, puisque caractéristiques de phénomènes rares. Une autre méthode d'analyse est basée sur un nombre fixé de statistiques d'ordre. L'estimation des paramètres est plus complexe, puisque les densités doivent prendre en compte la dépendance entre observations.

[5] Abréviation classique de *indépendantes et identiquement distribuées*.

- *Etude des pics dépassant un seuil :* la méthode POT (“ Peaks Over Threshold ”), très utilisée, a été développée pour la première fois dans le Flood Studies Report (NERC, 1975) ; elle a été décrite par exemple dans (Leadbetter, 1991). Elle est basée sur l’estimation des paramètres d’un modèle stochastique représentant les dépassements ou pics au-dessus d’un seuil fixé.
- *Etude et modélisation des jours et des tailles de dépassement de très haut niveau par un processus de Poisson non homogène* : les suites d’observations étudiées peuvent être indépendantes ou présenter une tendance, un phénomène de saisonnalité, une dépendance à long ou à court terme. C’est cette approche que nous avons retenue.

Le principal problème d’une approche par processus ponctuel est qu’elle doit tenir compte du regroupement des hautes valeurs, puisqu'elle est basée sur l'hypothèse d'indépendance des intervalles de temps entre deux dépassements de seuil *u* fixé élevé. Dans le cas de l’ozone, les modèles à une dimension paraissent peu réalistes. Ceci nous conduit à utiliser un modèle, *a priori* plus réaliste, prenant en compte simultanément fréquence et taille de dépassement. Le *processus de Poisson non-homogène* ((ultérieurement notée *PPNH*) dans le plan permet de modéliser le processus bi-dimensionnel des jours et des tailles de dépassement et ainsi détecter une tendance à moyen terme dans les épisodes de pollution aigüe par l’ozone, en tenant compte de la relation entre très hautes valeurs d’ozone et conditions météorologiques (Bellanger & Tomassone, 2000). L'approche statistique est basée sur le fait que l'on considère les dépassements d'un niveau élevé, se produisant dans le temps, comme des points d'un processus de Poisson. Des théorèmes limites pour de tels processus ont été développés par (Pickands, 1971) puis généralisés par (Leadbetter, Lindgren, & Rootzen, 1983). Ainsi, dans le cadre particulier de l'ozone troposphérique, (Smith R. , 1989), (Shively, 1991), et plus récemment (Smith & Shively, 1995) ont utilisé l'idée de considérer le nombre de dépassements de haut niveau comme généré suivant un *PPNH*, puisqu'une tendance peut exister.

Pour développer ce modèle, nous avons posé :

$$\Psi_i(y) = \begin{cases} P[Y > y \text{ le jour } i] \\ 0 \text{ si le jour } i \text{ est manquant} \end{cases}$$

La distribution de la variable aléatoire *Y* (maximum d'ozone journalier) le jour i, s'écrit donc $1 - \Psi_i(y)$, et on notera $\psi_i(y) = -\frac{d}{dy}[\Psi_i(y)]$, sa densité.

Si le processus est observé sur une période de temps]*0*,*T*[et si les pics d'ozone dépassant le seuil fixé u sont représentés par $(T_i; Y_i), 1 \leq i \leq N$, où T_i et Y_i sont supposées indépendantes $\forall i$. Le $i^{\text{ème}}$ pic se produit le jour T_i et prend la valeur $Y_i \geq u$. Le nombre total des *N* pics étant lui aussi une variable aléatoire, la densité conjointe des pics observés peut être approchée par :

$$L = \left[\left(\prod_{i=1}^{N} \Psi_{t_i}(u)\right) exp\left[-\int_0^T \Psi_t(u)dt\right]\right] \times \left[\prod_{i=1}^{N} \frac{\psi_{t_i}(y_i)}{\Psi_{t_i}(u)}\right] = A \times B$$

où :

- Le premier terme A entre crochets correspond à la densité d'un *PPNH* d'intensité $\Psi_i(u)$. Il correspond donc à la modélisation des jours de dépassement de niveau u (fréquence des dépassements).
- Le $i^{\text{ème}}$ terme du second terme *B* entre crochets correspond à la densité de Y_i sachant qu'un dépassement de seuil u a eu lieu.

Nous avons choisi de modéliser :

- l'intensité $\Psi_i(u)$du processus de Poisson à l'aide d'une *régression logistique* avec fonction de lien *logit* (cf. par exemple (Bellanger & Tomassone, 2014, pp. 373-395) ou (Hosmer & Lemeshow, 2000)), à cause de sa flexibilité et de l'interprétation relativement simple des paramètres. La fonction d'intensité du *PPNH* s'écrit donc :

$$\Psi_i(u) = \frac{exp(\alpha(i))}{1+exp(\alpha(i))} = P[Y > u \text{ le jour } i] \text{ où}$$

$$\alpha(i) = \alpha_0 + \alpha_1 t(i) + \sum_{j=2}^{p} \alpha_j w_j(i) + \sum_{j=2}^{p} \alpha_{1j} t(i) w_j(i) + \sum_{j,k=2}^{p} \alpha_{kj} w_j(i) w_k(i) \qquad \textbf{(1)}$$

$t(i)$ est un terme de tendance prenant la valeur k si le jour i appartient à l'année k (dans notre étude, $k \in \{1,2,\cdots,10\}$), $w_j(i)$ représente la valeur de la variable météorologique j le jour i. L'écriture de $\alpha(i)$ tient donc compte des interactions possibles entre covariables.

- La loi de la taille de dépassement par une loi exponentielle, cas particulier de la distribution de Pareto généralisée[6] (cf. (Pickands, 1975) et (Davison & Smith, 1990) pour les détails théoriques), loi très souvent utilisée dans les applications de la modélisation « *Peaks Over Threshold* » (*POT*).

  En effet, Les résultats théoriques dus à Pickands (1975) et Davison et Smith (1990) sur la distribution de la taille des dépassements d'un seuil élevé u (notée $X = Y - u$) permettent d'approcher la distribution de $X = Y - u$ sachant que $Y \geq u$, par une distribution de Pareto généralisée de paramètres ξ (*paramètre de forme*) et β (*paramètre d'échelle*) notée G :

$$P[Y_i > u + x \mid Y_i > u] \approx 1 - G\big(x; \beta(i), \xi(i)\big) = \left(1 + \xi(i)\frac{x}{\beta(i)}\right)^{-\frac{1}{\xi(i)}}$$

  avec $x = y - u$ où $\beta(i) > 0\ \forall i$; Y_i étant la variable aléatoire maximum d'ozone le jour i.
  On peut observer que si $\xi \approx 0$, la distribution de $X = Y - u$ sachant que $Y \geq u$ n'est autre qu'une distribution exponentielle de paramètre $1/\beta(i)$.

  Dans notre cas, nous avons donc, après vérification avec un test de Kolmogorov-Smirnov, supposé que la densité de la taille des dépassements d'un seuil élevé *u* (notée, $X = Y - u$), s'écrivait sous la forme la *densité exponentielle de paramètre* $1/\beta(i)$ où $\beta(i)$ prend la même forme analytique que $\alpha(i)$:

$$\beta(i) = \beta_0 + \beta_1 t(i) + \sum_{j=2}^{p} \beta_j w_j(i) + \sum_{j=2}^{p} \beta_{1j} t(i) w_j(i) + \sum_{j,k=2}^{p} \beta_{kj} w_j(i) w_k(i) \qquad \textbf{(2)}$$

[6] Notée ultérieurement en abrégé : *GPD*.

Différentes procédures permettent de déterminer le seuil u à partir duquel il est raisonnable d'appliquer cette modélisation. Le choix du seuil est complexe : trop faible nous ne pourrons pas utiliser les résultats asymptotiques, trop élevé nous aurons peu d'observations et une grande variabilité.

L'analyse de la vraisemblance est une première approche permettant de choisir les seuils raisonnables : les résultats asymptotiques suggèrent que le modèle est valide pour tous les seuils supérieurs à une certaine valeur inconnue ! Donc, le calcul des estimations des paramètres du modèle pour plusieurs seuils doit nous conduire à observer une stabilité de ces estimations. Il suffit d'examiner les graphes respectifs de chaque estimation de paramètre avec son intervalle de confiance correspondant (± 1.96 écart-type) par rapport au seuil. Ce graphe est souvent appelé *threshold choice plot*.

Une autre méthode graphique permet de sélectionner les seuils pour lesquels il est possible d'utiliser une distribution de Pareto généralisée à paramètres constants. En effet, si la distribution de $X = Y - u$ conditionnellement au seuil u, peut être approchée par distribution de Pareto généralisée définie précédemment, son espérance prend la forme linéaire en u suivante :

$$E[Y - u \mid Y > u] = \frac{\beta_0 + \xi u}{1 - \xi} \text{ où } \xi < 1$$

De plus, si le modèle *GPD* est valide pour un certain seuil u_0, alors il devra l'être pour tout seuil u supérieur. Par conséquent, le graphe de la taille moyenne observée des dépassements du seuil *u* par rapport à *u* devra être linéaire au-dessus d'une valeur du seuil pour laquelle les résultats asymptotiques sont valides. Ce graphe est généralement appelé *mean residual life plot*.

Une fois ce seuil déterminé, la méthode du maximum de vraisemblance permet d'estimer les paramètres, puis une procédure de sélection de type *stepwise* permet de ne conserver que les variables significatives. Enfin, le modèle est validé en interne. Il faut naturellement ensuite s'assurer que l'hypothèse d'indépendance mutuelle des dates (*resp.* tailles) est satisfaite, pour pouvoir utiliser un processus de Poisson non-homogène. Nous avons montré (Bellanger & Tomassone, 2000)

qu'un choix *a priori* du seuil u respectant ces hypothèses peut être effectué en tout début d'étude grâce à une procédure de *ré échantillonnage*.

Dans (Bellanger & Tomassone, 2000), les données disponibles étaient les valeurs des maxima journaliers d'ozone (fournis par AIRPARIF) mesurées sur quatre sites (Neuilly/Seine, Champs/Marne, Aubervilliers et Créteil), ainsi que les valeurs journalières de variables météorologiques (fournies par le mât du commissariat à l'Énergie Atomique de Saclay) décrites plus bas, durant les mois de mai à septembre de la période 1988-1997. Les mois de mai à septembre forment la période de l'année dans laquelle la majorité des hautes valeurs d'ozone sont enregistrées.

La valeur du maximum journalier (Y) utilisée dans notre analyse correspond à la valeur maximale prise entre 6h00 et 18h00 TU.

Les covariables utilisées dans l'analyse sont :

- la température maximale mesurée au sol (**`TMAX`**) : maximum des valeurs horaires entre 6h00 TU et 18h00 TU. (correspondant à *w3* dans le modèle) ;
- l'amplitude thermique (**`AMTEMP`**) : différence entre la valeur minimale et la valeur maximale mesurée entre 6h00 et 18h00. (correspondant à *w4* dans le modèle) ;
- la vitesse moyenne du vent mesurée à 58 mètres (**`VENT`**) : moyenne entre 6h00 et 18h00. (correspondant à *w5* dans le modèle) ;
- l'amplitude de vitesse de vent mesurée à 58 mètres (**`AMVENT`**): différence entre la valeur minimale et la valeur maximale de la vitesse du vent sur la période 6h00-18h00. (correspondant à *w6* dans le modèle) ;
- les variables dichotomiques **`t92`** et **`t93`** pour tenir compte des changements de capteurs intervenus en 1992 et 1993.

Là encore, un travail préliminaire de mise en forme du corpus a été nécessaire puisque les données comportaient des biais de mesure liés aux changements de capteurs mais aussi des données manquantes.

Le choix du seuil u permettant d'utiliser un *PPNH* est complexe : trop faible les résultats asymptotiques ne s'appliqueront pas, trop élevé trop d'observations seront conservés et les résultats seront sujet à de grandes variabilités. La stratégie classique du choix d'un seuil u raisonnable passe par la validation d'hypothèses concernant la

fréquence des dépassements (distribution des intervalles de temps S_i et indépendances mutuelle des S_i) et la taille des dépassements (distribution de la taille des dépassements X_i et indépendance mutuelle des X_i). Cette procédure est lourde, il paraît plus intéressant de choisir en début d'étude un seuil acceptable et de valider l'indépendance avant d'estimer les paramètres. Nous avons donc appliqué une procédure de ré-échantillonnage de type *bootstrap* ((Efron & Tibshirani, 1993) ; (Davison & Hinkley, 1997)), qui nous a permis pour un seuil u fixé, de calculer un intervalle de confiance à 95% basé sur les percentiles du coefficient de corrélation empirique entre intervalles adjacents. Si cet intervalle recouvrait la valeur 0, nous décidions que ce coefficient n'était pas significativement différent de 0. Après avoir déterminé la valeur raisonnable du seuil u permettant d'utiliser un *PPNH*, nous avons vérifié par une procédure *bootstrap* que les tailles de dépassements correspondantes n'étaient pas corrélées, en déterminant un intervalle de confiance à 95% du coefficient de corrélation uniquement pour les dépassements du seuil u se produisant des jours consécutifs. Ceci se justifie par le fait que, si les dépassements se produisant des jours successifs ne sont pas corrélés, alors les dépassements séparés par plus d'un jour ne le seront probablement pas aussi. Dans notre étude, les seuils u retenus pour modéliser les dépassements d'ozone et de comparer les résultats obtenus pour chacune des stations de mesure en région parisienne étaient (110, 120, 130, 140 et 150 μgm^{-3}).

La prise en compte de l'interaction entre l'année et la température dans le modèle logistique s'est avérée importante puisqu'elle améliorait beaucoup la qualité de l'ajustement ; mais de fait, créait une nouvelle difficulté inhérente à tout modèle avec interaction. Cet état de fait traduisait bien la complexité du phénomène physique étudié. Il n'a pas non plus été possible de conclure à une augmentation globale de la taille des dépassements dans la Région Parisienne, puisque la variable année n'était significative que sur un des sites modélisés.

Cette modélisation a cependant permis de mettre évidence des différences spatiales. Les tendances sur la période 1988-1997 se sont ainsi révélées complexes, fortement liées à la température, mais aussi à des phénomènes encore plus délicats à prendre en compte tels que l'évolution du trafic routier, du parc automobile et de la technologie des capteurs.

J'ai, après ma thèse, complété cette approche par *PPNH* décrite dans (Bellanger & Tomassone, 2000). La modélisation des jours et des tailles de dépassement pour chacune des quatre stations (Neuilly/Seine, Aubervilliers, Champs/Marne et Créteil) a été comparée à une modélisation globale regroupant ces différentes stations, pour les mois de mai à septembre de la période 1988-1997. Cette étude a abouti à (Bellanger, 2001).

Les méthodes de statistique spatiale étaient inappropriées à ce contexte, vu le faible nombre de stations à notre disposition. Nous avons donc contourné cette difficulté en utilisant une technique courante dans les applications du modèle linéaire : la transformation du facteur Station à quatre modalités en trois variables indicatrices représentant les stations dans les paramètres du *PPNH*. Nous avons ensuite comparé la modélisation par station à la modélisation globale :

- *Jours de dépassement* : les résultats obtenus n'ont pas permis d'obtenir un modèle unique pour les jours de dépassement du seuil 130. Cependant l'hypothèse d'un modèle partiel représentant les trois stations Aubervilliers, Champs/Marne et Créteil ne pouvaient être rejetée. Neuilly/Seine paraissait donc avoir un comportement très particulier, nous obligeant à conserver le modèle par station lui correspondant. Le modèle partiel a permis d'observer le caractère particulier de la station de Champs/Marne. Il est important de noter que la variable jour de dépassement contient moins d'informations que la variable taille de dépassement. Elle ne permet en effet pas toujours à elle seule de distinguer le comportement des stations d'observation (Créteil et Champs/Marne) de celui des stations appartenant au réseau de mesures de fond (réseau permettant de quantifier géographiquement la pollution atmosphérique ; Neuilly/Seine et Aubervilliers) : lors d'un épisode de pollution aigu, il est fort probable que toutes ces stations dépassent le seuil 130 ; mais ce dépassement a une durée et une amplitude différente suivant le type de stations.
- *Tailles de dépassement* : seuil pour le seuil 130, l'hypothèse d'un modèle unique n'a pas pu être rejetée. Ce modèle met en évidence les comportements similaires des stations de Neuilly/Seine et d'Aubervilliers. Il est important de souligner que cette étude, le choix des stations reposait sur la disponibilité des données, et non sur leur représentativité. Le modèle obtenu pour les tailles de dépassement du seuil

130 traduit donc principalement la différence de comportement entre les stations d'observation et stations de fond. Parmi les quatre stations étudiées, deux sont des sites d'observation, dont le comportement est difficile à prévoir du fait de l'influence ponctuelle, donc non systématique, d'axes routiers voisins et de parkings placés juste à côté du prélèvement. Il paraît donc raisonnable que les stations de Champs/Marne (av. J. Jaurès) et Créteil-Eglise se distinguent des deux stations urbaines de fond Neuilly/Seine et Aubervilliers.

Les résultats de cette modélisation traduisent bien la complexité des relations entre ces quatre stations ; relations qui diffèrent suivant que l'on étudie le processus des jours de dépassement ou celui des tailles de dépassement.

Toujours dans le but de détecter des tendances dans les très hautes valeurs d'ozone enregistrées en région parisienne, j'ai ensuite étendu la période d'étude (1988 à 2001 contre 1988-1997 précédemment) et augmenté le nombre de stations étudiées (sept stations (Neuilly/Seine, Aubervilliers, Champs/Marne, Créteil, Rambouillet, Paris (7ème) et Paris (13ème) contre quatre précédemment). Des *modèles non-paramétriques de type additifs généralisés* (Hastie & Tibshirani, 1990) ont été utilisés de manière exploratoire pour déceler des tendances plus complexes que celles prises en compte dans mes précédents travaux (uniquement linéaires). Ce travail a abouti à (Bellanger & Tomassone, 2004). La quantité de données traitées étaient importantes (10688 mesures d'ozone). Nous avons dans ce travail considéré les niveaux d'ozone dépassant le seuil u de 130 μgm^{-3}.

Nous avons choisi de modéliser :

- L'intensité $\Psi_i(u)$ du processus de Poisson comme indiqué dans **(1)** ; mais à l'aide d'une *régression logistique additive*. La fonction α du modèle additif dépendait aussi d'un facteur station et s'écrit :

$$\alpha(i, sta) = \alpha_0 + Station_{sta} + a_1(t) + \sum_{j=2}^{p} a_j\left(w_j(i)\right)$$

Où les $a_j(\,)$ représentent les fonctions de lissage estimées par splines de lissage et $Station_{sta} = \sum_{l=2}^{7} \gamma_{sta,l} D_{sta,l}$ .représente la combinaison linéaire des six variables indicatrices $D_{sta,l}$ représentant le facteur Station.

Une procédure de sélection de variable de type *stepwise* a ensuite été utilisée pour ne retenir que les variables significatives. Ce modèle emboîté, une fois validé, n'a été utilisé que pour détecter des non-linéarités et suggérer des transformations de variables qui ont ensuite été appliquées dans le cadre d'un modèle logistique paramétrique.

- La taille des dépassements (cf. **(2)**) à l'aide d'un modèle additif généralisé, pour détecter les non-linéarités et les incorporer ensuite dans un modèle linéaire généralisé. La fonction β du modèle additif dépend d'un facteur station et s'écrit :

$$\beta(i, sta) = \beta_0 + Station_{sta} + b_1(t) + \sum_{j=2}^{p} b_j\left(w_j(i)\right)$$

Où les $b_j(\,)$ représentent les fonctions de lissage estimées par splines de lissage.

L'étape de validation de nos modèles a aussi inclus l'utilisation de méthodes de ré-échantillonnage de type *bootstrap* pour confirmer la qualité des estimations obtenues. Nous avons alors pu mettre en évidence une évolution différente de la tendance sur la période d'étude 1988-1997 suivant que l'on s'intéresse aux jours ou la taille des dépassements du seuil 130 μgm^{-3}. Autant, une hausse importante a été révélée pour les jours de dépassement, autant une petite diminution a été détectée pour les tailles de dépassement. Dans les deux cas, des prévisions peuvent être obtenues. Une aire sous la courbe ROC de 0.9 indique de plus que la prévision des jours de dépassement du seuil 130 μgm^{-3} est excellente.

Dans ce dernier travail, les outils statistiques utilisés étaient donc clairement identifiés et comprenaient l'utilisation d'un *PPNH* associé à des modèles linéaires généralisés. Cependant, comme toujours quand on s'attelle à la modélisation statistique de données réelles, de nombreuses étapes ont été nécessaires pour aboutir à des résultats valides et interprétables :

- tout d'abord, une étape exploratoire et graphique qui a permis de déterminer le meilleur modèle ;
- d'autre part, une étape de validation interne et externe du modèle proposé.

L'ensemble de ces travaux a été présenté dans de nombreuses conférences.

1.1.2 Sols agricoles et culture du blé : les éléments traces métalliques (2006 - 2009)

Collaborateurs sur ce thème :

- INRA[7] (Science du Sol, Centre d'Orléans) : D. Baize ;
- AgroParisTech/INRA (dpt MMIP, Paris) : R. Tomassone.

Parmi les nombreux problèmes classiques de la Statistique, celui de l'étude de la relation entre variables est sans nul doute l'un des plus fréquents : on calcule le coefficient de corrélation entre deux variables quantitatives, on estime les paramètres d'un modèle de régression d'une variable à expliquer en fonction d'une ou plusieurs autres (les régresseurs ou variables explicatives) pour tenter d'"expliquer" cette variable et éventuellement de la prédire pour d'autres valeurs des régresseurs. Quand on dispose de deux groupes de variables une méthode, l'*Analyse des Corrélations Canoniques* souvent appelée *Analyse Canonique* (ultérieurement notée *ACC*), existe depuis bien longtemps (Hotelling, 1936). Bien que de nombreux logiciels offrent un programme pour réaliser les calculs : peu de publications avec des applications l'utilisent. Pour résumer, l'*ACC* est caractérisée par :

- une interprétation des résultats souvent délicate ;
- mais un intérêt théorique essentiel fournissant un cadre unificateur à un certain nombre d'autres méthodes.

Nous avons dans ce travail, à partir d'un exemple, montré que l'on pouvait tout de même exploiter les résultats fournis par une analyse canonique ; même si l'exploitation pouvait s'avérer complexe. Les données traitées proviennent d'une étude qui peut s'apparenter à un "cas d'école" pour l'analyse canonique : en 1998, le Ministère de Aménagement du Territoire et de l'Environnement avait lancé le programme GESSOL (Fonctions environnementales des sols et GEStion du patrimoine SOL). Une des questions fondamentales de ce programme était :

"Est-il possible de bâtir des modèles permettant de détecter par avance les cas de concentrations excessives en éléments traces métalliques (ETM) dans les grains de blé à partir de données pertinentes acquises sur des échantillons de sol ?".

[7] Acronyme de *Institut National de la Recherche Agronomique*.

Le problème était, et reste, d'une extrême importance pour de multiples raisons liées à l'évolution des pratiques agricoles ; en particulier celle liée à l'épandage de boues d'épuration riches en ETM et aux polémiques qui en découlent. En effet, dans la mesure où les métaux lourds peuvent s'avérer dangereux pour la santé, les problèmes de santé publique sont aussi très présents dans ce travail. Au moment de notre étude, les publications sur le sujet (Pinet, Lecomte, Vimont, & Auburtin, 2003), étaient principalement des compilations de résultats d'essais agronomiques sur de nombreuses plantes. Les seules méthodes d'analyse utilisées étaient la régression linéaire et l'analyse en composantes principales. Les résultats statistiques des régressions se limitaient à une équation, une valeur du coefficient de détermination (R^2), mais aucune analyse critique de la validité de ces régressions n'était faite.

Ce travail a abouti à plusieurs publications ((Bellanger, Baize, & Tomassone, 2006) ; (Baize, Bellanger, & Tomassone, 2009) ; (Bellanger & Tomassone, 2014, pp. 331-337) et a été présenté lors de la XXIII International Biometric Conference en juillet 2006 à Montréal (Canada). Le résultat essentiel a été la mise en évidence d'une relation forte entre concentration en ETM et les caractéristiques d'un sol.

1.1.3 Logements français : le plomb (2009 - 2013)

Travail de thèse de J.-P. Lucas, co-dirigé à 50% avec V. Sébille (PU-PH, EA 4275 - SPHERE[8] Nantes) (soutenu le 30 octobre 2013)

Collaborateurs sur ce thème :

- CSTB[9] (dpt « Energie, Santé, Environnement », Marne-La-Vallée) : J.-P. Lucas, S. Kirchner, C. Mandin ;
- InVs[10] (Paris): P. Bretin, Y. Le Strat, A. Le Tertre ;
- IRSET[11] (UMR 1085, Rennes) : P. Glorennec, B. Le Bot ;
- Groupe ISA[12] (EA 4515 Sols et Environnement, Université Lille Nord de France) : F. Douay ;
- Université de Nantes (EA 4275 - SPHERE) : V. Sébille.

Le travail de thèse de spécialité biostatistique de Jean-Paul Lucas (Lucas, 2013) s'inscrit dans le contexte de l'exposition au plomb en milieu résidentiel. Alors que des données relatives à l'exposition au plomb (Pb) sont collectées dans les logements français dans un cadre réglementaire, aucun état de la contamination par le Pb dans les logements n'avait été réalisé. A partir des données recueillies par sondage dans l'enquête environnementale *Plomb-Habitat*, les objectifs étaient :

- de fournir une estimation de la contamination par le Pb dans les logements français ;
- d'identifier les sources potentielles de contamination par le Pb des poussières à l'intérieure des logements à l'aide d'un modèle s'ajustant le mieux aux données disponibles en tenant compte de leurs spécificités.

Ce travail a donné lieu à plusieurs publications (Lucas, et al., 2012) ; (Lucas, Sébille, Le Tertre, Le Strat, & Bellanger, 2014) ; (Lucas, et al., 2014) et les résultats ont été présentations lors de plusieurs conférences. Cette partie s'inspire du manuscrit de thèse de Jean-Paul Lucas (Lucas, 2013) dans lequel de plus amples détails pourront être trouvés sur les données, les méthodes statistiques ainsi que les résultats complets.

---

[8] BioStatistique, Pharmacoépidémiologie Et Mesures Subjectives en Santé.

[9] Acronyme de *Centre Scientifique et Technique du Bâtiment.*

[10] Acronyme d'*Institut de Veille Sanitaire*.

[11] Acronyme d'*Institut de Recherche sur la Santé, l'Environnement et le Travail*.

[12] Institut Supérieur d'Agriculture de Lille.

Les données sur lesquelles repose le présent travail ne sont pas exhaustives ; mais sont des *données d'enquête*, obtenues à partir de l'investigation de logements. Les logements à enquêter figurent dans une liste construite à partir d'une procédure de tirage appelée *plan de sondage*. L'échantillon de logements correspond alors à une portion de la population finie (ou base de sondage). Les techniques de sondages utilisent une terminologie spécifique et possèdent des particularités propres. Nous commençons donc, dans un premier temps, par en rappeler les notions essentielles pour faciliter la lecture de la suite de cette partie. Pour les détails techniques liés à la théorie des sondages, nous renvoyons aux ouvrages de référence en français tels que (Ardilly, 2006), (Tillé, 2001) ou en anglais tels que (Cochran, 1977), (Lohr, 2009), (Lumley, 2010) ou (Särndal, Swensson, & Wretman, 2013).

Notions essentielles de théorie des sondages

Formalisation mathématique d'un sondage

La *théorie des sondages* se définit comme l'ensemble des outils statistiques permettant l'étude d'une population finie U de taille N, au moyen de l'examen d'une partie de celle-ci appelée *échantillon*. La méthode de sélection de l'échantillon d'individus[13] i $(i = 1, \dots, n)$, puis la formulation de l'estimateur associé à une fonction des valeurs y_i prises par la variable d'intérêt (non aléatoire) Y sur l'échantillon, définissent le *plan de sondage*. Il est donc important de souligner que dans le cadre de la théorie des sondages, on ne s'intéresse pas aux valeurs individuelles de la variable d'intérêt Y dans la population finie U ; mais à une fonction d'intérêt $f(y_1, \dots, y_n)$ liée à la nature de celle-ci : ce peut être une moyenne, une proportion, un total, un ratio, etc.

On appelle *échantillon* s de taille n de la population U, au sens de la théorie des sondage, un n-uple non-ordonné sans remise. L'ensemble $\mathbb{S}$ des échantillons de U est l'ensemble des parties non vides de U : $\mathbb{S} = \{s | s \subset U\} \backslash \emptyset$. L'objectif de l'*échantillonnage* est de construire un échantillon s, à partir d'une population U, en utilisant un plan de sondage. De manière à éviter, autant que possible, les biais de sélection et permettre

[13] Encore appelé unités.

d'associer un intervalle de confiance aux estimations des fonctions d'intérêt, la construction de l'échantillon doit reposer sur un tirage aléatoire des individus le composant. On parlera de *sondages probabilistes*, par opposition aux *sondages empiriques*, tels que les méthodes de quotas par exemple, qui ne permettent pas d'estimer correctement une variance et sont sujets à des biais de sélection non maitrisables. Nous n'évoquerons par la suite que les sondages probabilistes.

Pour disposer d'un échantillon s, on met en œuvre un *plan de sondage* permettant son tirage aléatoire non ordonné sans remise. Il se définit par une loi de probabilité $p(.)$ sur $\mathbb{S}$ telle que par définition $p(s) \in [0;1]$ pour tout $s \in \mathbb{S}$ et $\sum_{s\in\mathbb{S}} p(s) = 1$. Puisqu'un plan de sondage n'est rien d'autre qu'une loi de probabilité, nous pouvons définir l'*échantillon aléatoire* S, comme une variable aléatoire à valeurs dans $\mathbb{S}$. La loi de probabilité de S est donnée par $P(S = s) = p(s), s \in \mathbb{S}$.

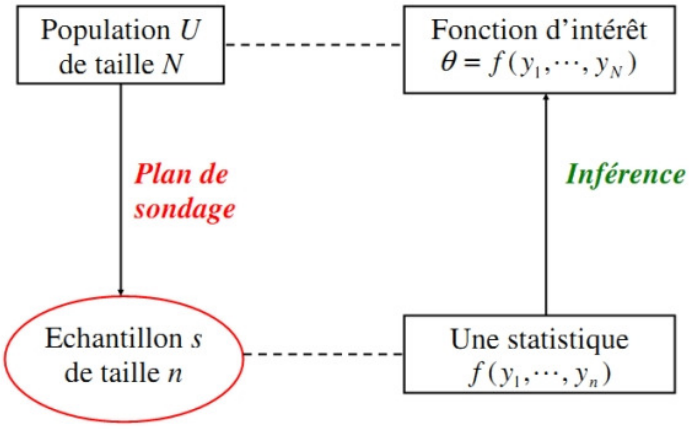


Figure 2 - Echantillonnage et inférence.

Comme l'illustre la Figure 2, une fois l'échantillon construit selon un plan de sondage à fixer, on peut réaliser une inférence rigoureuse et adaptée de la fonction d'intérêt.

Pour pouvoir sélectionner un échantillon s issu d'un tirage probabiliste dans lequel par définition chaque individu a une probabilité connue et fixée de faire partie de s, il faut disposer d'une liste appelée *base de sondage*. *La base de sondage* reproduit la population

U en en couvrant tous les éléments, sans doublon et en les identifiant sans ambiguïté afin de pouvoir les atteindre s'ils sont sélectionnés. Une base de sondage incomplète, c'est-à-dire non exhaustive ou contenant des non-réponses, donne lieu à un *défaut de couverture* de la population ; ce qui est problématique pour l'inférence.

La *probabilité d'inclusion d'ordre 1* de la $k^{\text{ème}}$ unité se définit comme la probabilité que cette même unité appartienne à l'échantillon :

$$\pi_k = P[k \in S] = \sum_{s \in \mathbb{S}} p(s) 1_{\{k \in s\}} = \sum_{s \ni k} p(s)\,;\ \forall k \in U$$

Par définition on a : $\pi_k = E\left(1_{\{k \in S\}}\right)$.

La *probabilité d'inclusion d'ordre 2* correspond quant à elle à la probabilité que deux unités distinctes appartiennent simultanément à un échantillon :

$$\pi_{kl} = P[k \in S, l \in S] = \sum_{s \ni k,l} p(s)\,;\ \forall k, l \in U \ \text{ avec } \ k \neq l$$

Par définition, on a : $\pi_{kl} = E\left(1_{\{k \in S\}} 1_{\{l \in S\}}\right)$ et donc on peut montrer que $var\left(1_{\{k \in S\}}\right) = \pi_k(1 - \pi_k)$ et $cov\left(1_{\{k \in S\}}, 1_{\{l \in S\}}\right) = \pi_{kl} - \pi_k \pi_l$.

Nous noterons par la suite : $\Delta_{kl} = \begin{cases} cov\left(1_{\{k \in S\}}, 1_{\{l \in S\}}\right); k \neq l \\ var\left(1_{\{k \in S\}}\right) sinon. \end{cases}$

Pour réaliser une enquête par sondage, la connaissance des probabilités d'inclusion de chaque unité appartenant à l'échantillon s est indispensable car elles vont pondérer la valeur de la variable d'intérêt recueillie sur chaque unité pour fournir une estimation de la fonction d'intérêt. Le *poids de sondage* est l'inverse de la probabilité d'inclusion :

$$w_k = \frac{1}{\pi_k}\,;\ \forall k \in U$$

Le poids de sondage indique le nombre d'individus représentés par k dans la population U.

---

[14] La somme dans l'équation se fait sur les échantillons *s* (qui contiennent l'unité *k*) d'où la notation $s \ni k$.

Dans la suite, nous nous intéresserons uniquement aux plans de *taille fixe* *n*, c'est-à-dire que la taille de l'échantillon s est fixée préalablement à *n*. Les probabilités d'inclusion respectent alors des propriétés bien spécifiques :

$$\sum_{k\in U}\pi_k = n \; ; \sum_{\substack{k\in U \\ k\neq l}}\pi_{kl} = (n-1)\pi_l \text{ et enfin } \sum_{k\in U}\Delta_{kl} = 0 \text{ pour } l \in U \text{ fixé.}$$

(Horvitz & Thompson, 1952) ont introduit un estimateur linéaire sans biais d'un total t_y pour tout plan de sondage, si $\pi_k > 0 \; \forall k \in U$:

$$\hat{t}_{y,\pi} = \sum_{k\in S}\frac{y_k}{\pi_k}$$

Cet estimateur est appelé *π–estimateur*, *estimateur d'Horvitz-Thompson* ou encore *estimateur des valeurs dilatées*. Il admet pour variance :

$$var\left(\hat{t}_{y,\pi}\right) = -\frac{1}{2}\sum_{\substack{k,l\in U \\ k\neq l}}\left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l}\right)^2 \Delta_{kl}$$

Si le plan est à taille fixe *n*, $var\left(\hat{t}_{y,\pi}\right)$ admet comme estimateur sans biais, appelé *estimateur de Sen-Yates-Grundy* (Yates & Grundy, 1953) :

$$\widehat{var}\left(\hat{t}_{y,\pi}\right) = -\frac{1}{2}\sum_{\substack{k,l\in U \\ k\neq l}}\left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l}\right)^2 \frac{\Delta_{kl}}{\pi_{kl}} 1_{\{k\in S, l\in S\}}$$

On peut alors obtenir un intervalle de confiance de niveau de confiance $1-\alpha$ du total t_y en supposant que l'estimateur d'Horvitz-Thompson suit approximativement une loi Normale dès que la taille *n* de l'échantillon est suffisamment grande :

$$IC_{1-\alpha}(t_y) = \left[\hat{t}_{y,\pi} \pm u_{1-\alpha/2}\sqrt{\widehat{var}\left(\hat{t}_{y,\pi}\right)}\right]$$

où $u_{1-\alpha/2}$ représente le quantile d'ordre $1-\alpha/2$ d'une variable aléatoire Normale centrée réduite.

L'ensemble de ces résultats sur le total t_y s'étendent naturellement à toute fonction d'intérêt, fonction linéaire des y_i telle que la moyenne ou une proportion de la variable d'intérêt Y si les y_i sont dichotomiques.

Si l'on considère que la *taille de l'échantillon* est fixée à n ; il faudra donc déterminer à l'avance n pour répondre aux objectifs de l'enquête en tenant compte des contraintes liées à sa mise en œuvre telles que les limites budgétaires. Il faut donc se poser la question des incertitudes liées aux futures estimations que l'on est prêt à accepter. On cherchera n de manière à atteindre une précision absolue fixée b pour le paramètre étudié θ qui sera contenu dans un intervalle de confiance centré en $\hat{\theta}$ avec une probabilité d'au moins $1-\alpha$ *i.e.* $P\left[\theta \in \left[\hat{\theta}-b;\hat{\theta}+b\right]\right] \geq 1-\alpha$. En supposant que l'estimateur $\hat{\theta}$ suit approximativement une loi Normale, on sait que $P\left[\theta \in \left[\hat{\theta}-u_{1-\alpha/2}\sqrt{\widehat{var}(\hat{\theta})};\hat{\theta}+u_{1-\alpha/2}\sqrt{\widehat{var}(\hat{\theta})}\right]\right] \geq 1-\alpha$ où $u_{1-\alpha/2}$ est le quantile $1-\alpha/2$ de la loi Normale centrée réduite. Enfin, puisque $\sqrt{\widehat{var}(\hat{\theta})}$ dépend de la taille de l'échantillon n, on cherchera n_0 induisant la précision b requise.

Ainsi, par exemple, dans le cas du *π–estimateur* $\hat{t}_{y,\pi}$ d'un total t_y , la taille n de l'échantillon sera obtenue à l'aide de l'inégalité $u_{1-\alpha/2}\sqrt{\widehat{var}(\hat{t}_{y,\pi})} \leq b$, où la précision b aura été préalablement fixée.

A partir des bases de sondages disponibles et des informations auxiliaires[15] qu'elles contiennent, un plan va donc être choisi. Il en existe de nombreux types. Nous restreignons notre présentation aux principaux, utiles à la compréhension de ceux mis en place dans l'enquête *Plomb-Habitat* :

sondages aléatoires simples, *sondages à probabilités inégales*, *sondages stratifiés*, *sondages à plusieurs phases*, *sondages à plusieurs degrés.*

[15] La notion d'*information auxiliaire* regroupe toute information extérieure à l'enquête proprement dite permettant d'augmenter la précision des résultats d'un sondage.

Dans la suite de cette présentation, nous nous placerons toujours dans le cas des plans de taille n fixée c'est-à-dire que l'on désire tirer des échantillons dont les probabilités π_k d'inclusion d'ordre 1 sont fixées *a priori* et telles que $\sum_{k\in U} \pi_k = n$.

Plans sondage

Nous insistons sur le fait que nous ne présentons dans ce paragraphe que les plans de sondage à taille fixe utiles pour la suite. Il en existe de nombreux autres tels que par exemple les plans par grappes. Nous renvoyons le lecteur intéressé aux ouvrages spécialisés.

- Un plan de sondage sans remise est dit *aléatoire simple* si tous les échantillons de même taille n ont la même probabilité d'être sélectionnés. La loi de probabilité de S est donnée par :

$$p(s) = \begin{cases} 1/C_N^n & si\ s\ est\ de\ taille\ n \\ 0 & sinon \end{cases}$$

Les probabilités d'inclusion permettant d'obtenir le *π–estimateur* et sa variance se calculent facilement, on a pour tout $k, l \in U$ avec $k \neq l$:

$$\pi_k = \frac{n}{N} \text{ et } \pi_{kl} = \frac{n(n-1)}{N(N-1)}$$

Dans un sondage aléatoire simple, tous les individus de la population U ont la même probabilité d'être tirés au sort, égale à la *fraction de sondage* notée f, définie par :

$$f = \frac{n}{N}$$

Le Tableau 3 présente pour ce plan l'estimateur d'une moyenne, d'un total et d'une proportion obtenus en appliquant les résultats de (Horvitz & Thompson, 1952) :

Tableau 3 - Estimateurs pour un plan simple sans remise.

	Estimateur	Estimateur de la variance de l'estimateur
Moyenne	$\hat{\bar{y}}_\pi = \frac{1}{n}\sum_{k\in S} y_k$	$\widehat{var}(\hat{\bar{y}}_\pi) = (1-f)\frac{s_n^2}{n}$
Total	$\hat{t}_{y,\pi} = N\hat{\bar{y}}_\pi$	$\widehat{var}(\hat{t}_{y,\pi}) = N^2(1-f)\frac{s_n^2}{n}$
Proportion[16]	$\hat{p}_{y,\pi} = \frac{1}{n}\sum_{k\in S} y_k$	$\widehat{var}(\hat{p}_{y,\pi}) = \frac{N-n}{N(n-1)}\hat{p}_{y,\pi}(1-\hat{p}_{y,\pi})$

Où dans le cas d'un plan simple sans remise, $s_n^2 = \frac{1}{n-1}\sum_{k\in S}\left(y_k - \hat{\bar{y}}\right)^2$ estime sans biais la variance corrigée de la population $S_n^2 = \frac{1}{N-1}\sum_{k\in U}(y_k - \bar{y})^2$ avec $\bar{y} = \frac{1}{n}\sum_{k\in U} y_k 1_{\{k\in S\}}$.

Dans ce qui va suivre, nous ne donnerons pas les estimateurs correspondant aux autres plans de sondage exposés et renvoyons vers les ouvrages spécialisés. La démarche est toujours similaire : à partir des probabilités d'inclusion, le *π–estimateur* est utilisé pour obtenir un estimateur du total, de la moyenne ou d'une proportion et pour en déduire l'estimateur de sa variance.

- Parallèlement aux familles de plans à probabilités égales, il existe des plans de sondage où les individus ont des probabilités d'inclusion inégales. On les appelle *plans à probabilités inégales*. Ils permettent de pouvoir bénéficier de la connaissance d'une *variable auxiliaire* X, reliée à la variable d'intérêt Y, pour obtenir des estimations plus précises.

  Le principe consiste à définir des probabilités d'inclusion du premier ordre *proportionnelles* aux valeurs $x_k, k \in U$ prises par la variable auxiliaire X. Notons tout de même que cela implique donc sa connaissance sur toute la population ! Comme nous nous plaçons dans le cadre d'un plan à taille fixe n, on a $\sum_{k\in U}\pi_k = n$ et pour obtenir des probabilités d'inclusion proportionnelles aux x_k, *i.e.* $\pi_k = cx_k$, on doit alors choisir $c = \frac{n}{\sum_{l\in U} x_l} = \frac{n}{t_x}$.

[16] Le caractère Y est dans ce cas une variable dichotomique.

Dans cette famille de plans de sondage, il est important de souligner que les tirages sont effectués dans la population toute entière, telle qu'elle se présente à l'origine. Ceci ne sera pas le cas des plans que nous allons présenter par la suite ; même si les individus de la base de sondage n'auront eux-aussi pas tous la même probabilité d'inclusion : ceci peut donc troubler le lecteur…

- En cas de doute sur la relation linéaire entre la variable auxiliaire X et la variable d'intérêt Y, il sera préférable d'adopter un plan simple ou bien un *plan de sondage stratifié* . Le principe est identique à celui de l'analyse de variance : si les individus sont très différents les uns des autres vis-à-vis de Y, constituer des groupes homogènes au sein desquels on réalise des tirages devrait permettre de réduire la variabilité intra-groupe, augmenter la variabilité inter-groupe ; donc d'augmenter la précision des estimateurs.

La stratification consiste à partitionner la population de départ U en H groupes pouvant correspondre aux H modalités d'une variable auxiliaire qualitative. En pratique, cela revient à découper la base de sondage en plusieurs sous-bases de sondage. Un sondage est dit *stratifié aléatoire simple* si, comme illustré en Figure 3, pour chaque strate, on tire un échantillon selon un sondage aléatoire simple sans remise de taille fixe et que les tirages au sein de chaque strate sont mutuellement indépendants. La technique de stratification est employée par exemple pour couvrir des zones géographiques (administratives, définies selon des facteurs d'exposition, etc.).

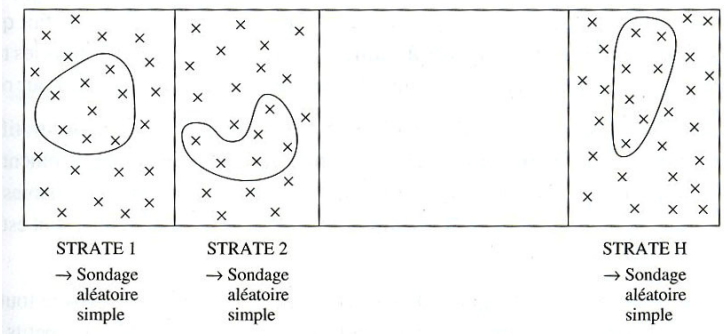


Figure 3 - Principe du plan de sondage stratifié aléatoire simple (figure tirée de (Ardilly, 2006, p. 89)).

Dans le cas général, un tirage stratifié est un tirage où les individus de la base de sondage n'ont pas tous la même probabilité d'être sélectionnés. La population U de taille N est préalablement découpée en H sous-ensembles $U_1, \dots, U_H$ appelés *strates* et tels que :

$$\bigcup_{h=1}^{H} U_h = U;\ U_{h'} \cap U_h = \emptyset, h' \neq h$$

Chaque strate U_h admet une taille N_h et l'on a bien évidemment $N = \sum_{h=1}^{H} N_h$. Soit S_h l'échantillon aléatoire tiré dans la strate U_h à l'aide d'un plan de sondage $p_h(\ \ )$.

L'échantillon aléatoire S de taille fixée à n s'écrit donc $S = \bigcup_{h=1}^{H} S_h$. Le plan de sondage associé $p(\ \ )$ n'est rien d'autre que :

$$p(s) = \prod_{h=1}^{H} p_h(s)\,;\, s = \bigcup_{h=1}^{H} s_h \text{ et } n = \sum_{h=1}^{H} n_h$$

Dans l'hypothèse d'un sondage aléatoire simple dans chaque strate, la probabilité d'inclusion d'ordre 1 d'un individu k appartenant à la strate U_h est égale à :

$$\pi_k = \frac{n_h}{N_h}$$

Pour les probabilités d'inclusion d'ordre 2, le résultat dépend du fait ou non que les individus k et l ` appartiennent à la même strate ou non :

$$\pi_{kl} = \begin{cases} \dfrac{n_h(1-n_h)}{N_h(1-N_h)}\ , si\ k\ et\ l\ \in U_h \\ \pi_k\pi_l = \dfrac{n_{h_1}}{N_{h_1}}\dfrac{n_{h_2}}{N_{h_2}}\ , si\ k\ \in U_{h_1} et\ l\ \in U_{h_2} \end{cases}$$

En sondage stratifié, il faut donc fixer *a priori* la taille n_h de chaque strate de l'échantillon. Lorsque la fraction de sondage $f_h = n_h/N_h$ dans la strate h est égale à la fraction de sondage globale $f = n/N$ et ceci pour toutes les strates. On parle alors de plan de sondage stratifié avec *allocation proportionnelle*. Ainsi, dans le cas d'un *sondage stratifié à allocation proportionnelle avec tirage aléatoire simple dans chaque strate*, chaque individu d'une même strate a la même probabilité d'inclusion

d'ordre 1 et l'effectif de la strate h est $n_h = n\,{}^{N_h}/_{N}$ supposé entier, ce qui est rarement le cas en pratique ! Il existe d'autres types d'allocations qui seront à envisager si l'on veut estimer une moyenne ou un total, nous renvoyons le lecteur intéressé au chapitre 7 de (Tillé, 2001) ou au chapitre 2 de (Ardilly, 2006).

En conclusion, la stratification est un principe simple qui apporte un gain de précision par rapport à un plan simple. C'est la raison statistique pour laquelle elle est largement utilisée dans les enquêtes. Il est cependant important de garder à l'esprit que pour réaliser une stratification, la base de sondage doit contenir, pour toutes ses unités, une information auxiliaire qualitative.

Lorsqu'une variable auxiliaire est présente pour chaque unité de la base de sondage, on a vu que l'on pouvait s'en servir pour réaliser un sondage à probabilités proportionnelles ou stratifié. Mais, quand elle n'est pas présente pour chaque unité de la base de sondage, c'est-à-dire pour chaque unité appartenant à la population d'intérêt, le *plan de sondage à plusieurs degrés* ou *le plan de sondage en deux phases* peuvent être de bonnes alternatives.

- L'information auxiliaire peut aussi servir à améliorer l'organisation d'une enquête et non pas à améliorer la précision des estimateurs comme dans le cas des plans de sondage stratifiés. Ainsi, il est parfois difficile pour certain type de populations, de construire une base de sondage listant la totalité des unités d'intérêt et il ne sera alors pas possible de les sélectionner directement. Par exemple, comme il n'existe pas de base de sondage identifiant les résidences principales en France métropolitaine (unités d'intérêt) avec le moyen de les atteindre facilement, il n'est pas possible d'utiliser les plans que nous venons de décrire précédemment pour réaliser une enquête. D'où l'intérêt d'utiliser un *plan de sondage à plusieurs degrés* pour tirer l'échantillon. On se limitera à présenter le plan *à deux degrés*, le raisonnement étant analogue pour les degrés supérieurs. Il consiste en un double échantillonnage : tout

d'abord sur les *unités primaires*[17], puis sur les *unités secondaires*. Dans notre exemple, un plan à deux degrés consisterait à échantillonner les régions (unités primaires) puis à prélever pour chaque région retenue un échantillon de résidences principales (unités secondaires).

On suppose que la population $U = \{1, \dots, N\}$ est subdivisée en M sous-populations $U_i, i = 1, \dots, M$, que l'on appellera *unités primaires*. Les unités primaires sont composées de N_i unités secondaires telles que l'on ait bien $\sum_{i=1}^{M} N_i = N$. La Figure 4 illustre le principe de fonctionnement d'un plan de sondage à deux degrés. L'échantillon S_1 du premier degré est de taille m et vaut $\bigcup_{j=1}^{m} S_{1,j}$ (réunion des parties en gris clair sur la Figure 4). L'échantillon final S obtenu avec un tel plan est $S = \bigcup_{j \in S_1} S_{2,j}$ et sa taille (aléatoire) est $n_S = \sum_{j \in S_1} n_j$ avec $n_j = card(S_{2,j})$ (réunion des parties ⭕ sur la Figure 4).

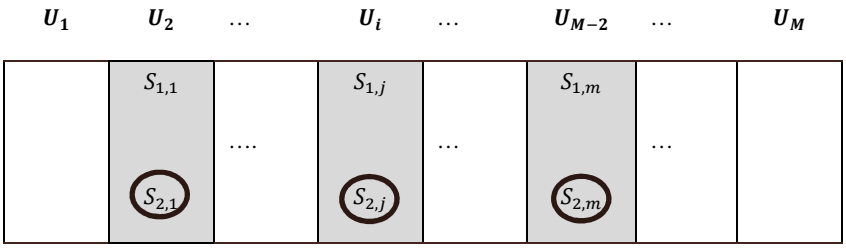


Figure 4 - Principe du plan à deux degrés.

Pour effectuer un plan à deux degrés, il faut donc :

- construire un échantillon S_1 d'unités primaires à partir d'un plan de sondage $p_1(\ \ )$ sur $\{1, \dots, M\}$;
- pour chaque unité primaire sélectionnée, construire un échantillon S_2 sur les unités secondaires à partir d'un plan de sondage $p_2(\ \ )$.

Il est de plus souhaitable qu'il possède les deux propriétés suivantes :

Invariance : le plan du second degré $p_2(\ \ )$ est indépendant du premier plan $p_1(\ \ )$, *i.e.*,

$P[S_2 = s_2 | S_1 = s_1] = P[S_2 = s_2]$;

Indépendance : les tirages du second degré sont mutuellement indépendants.

[17] En anglais : *primary sampling unit* ; d'où l'abréviation PSU.

Les probabilités d'inclusion d'ordre 1et 2 pour le premier degré sont notées respectivement par :

$$\pi_{1,i} = P[U_i \in S_1] \text{ et } \pi_{1,ij} = P\left[U_i \in S_1, U_j \in S_1\right] \; i = 1, \cdots, M \; ; \; j = 1, \cdots, M \text{ et } i \neq j$$

On note également $\pi_{k|i}$ la probabilité de sélectionner l'unité (secondaire) k sachant que l'unité (primaire) U_i a été choisie et d'une manière analogue $\pi_{kl|i}$ la probabilité d'inclusion d'ordre 2 sachant que U_i a été retenue.

La probabilité d'inclusion π_k pour $k \in U_i$ s'écrit alors :

$$\pi_k = P\left[k \in S_{2,i}, U_i \in S_1\right] = P\left[k \in S_{2,i} | U_i \in S_1\right] P[U_i \in S_1] = \pi_{k|i}\pi_{1,i}$$

Pour les probabilités d'inclusion d'ordre 2, en supposant que l'hypothèse d'indépendance est satisfaite, on obtient :

$$\pi_{kl} = \begin{cases} \pi_{kl|i}\pi_{1,i} \, , & k, l \in U_i \\ \pi_{k|i}\pi_{l|j}\pi_{1,ij} \, , & k \in U_i, l \in U_j, i \neq j \end{cases}$$

Le cas le plus simple et assez courant est le cas d'un *plan à 2 degrés avec tirage à probabilités égales*. Il consiste à sélectionner les unités primaires et les unités secondaires selon un plan de sondage aléatoire simple sans remise. Les probabilités d'inclusion d'ordre 1et 2 pour le premier degré valent alors : $\pi_{1,i} = \frac{m}{M}$ et $\pi_{1,ij} = \frac{m(m-1)}{M(M-1)} \; i = 1, \cdots, M \; ; \; j = 1, \cdots, M$ et $i \neq j$. Quant à la probabilité d'inclusion π_k pour $k \in U_i$, la taille des échantillons au sein des unité primaire étant fixée à n_i, elle vaut $\frac{mn_i}{MN_i}$.

- Le *plan de sondage en deux phases avec post-stratification* donne un cadre plus général au plan de sondage à plusieurs degrés. Son principe est illustré ci-dessous en Figure 5 :

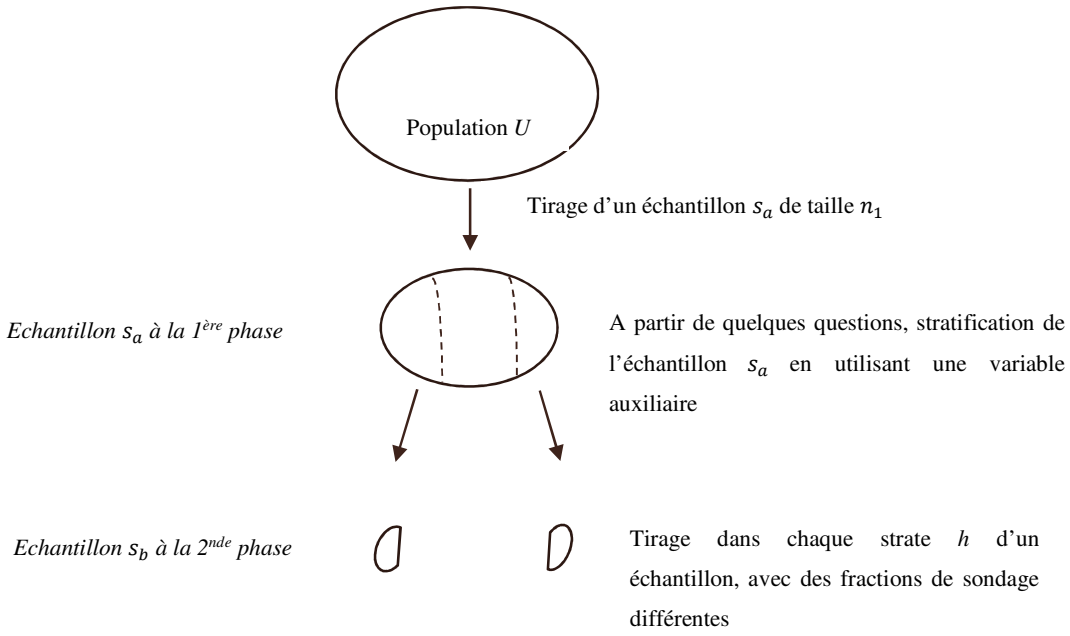


Figure 5 - Principe du sondage en deux phases.

Dans un sondage à deux phases, un échantillon s_a est tiré aléatoirement dans une population U à partir d'un plan de sondage (le plus souvent *simple*) $p_a(\ \ )$. Des informations auxiliaires sont alors collectées à l'aide de moyens peu coûteux sur les unités de s_a. Chaque individu de cet échantillon est ensuite affecté à une des H catégories pré-définies selon la variable auxiliaire. Puisque cette affectation s'effectue après le tirage de 1ère phase, les catégories sont appelées *post-strates*[18]. Puis, dans une seconde phase, des sous-échantillons issus de s_a sont tirés aléatoirement selon un plan de sondage $p(.|s_a)$ à partir des post-strates h du premier échantillon. L'échantillon s_b ainsi construit est la réunion de tous les sous-échantillons. La variable d'intérêt n'est observée que sur cet échantillon s_b.

[18] On parle donc de tirage en deux phases *avec post-stratification*.

La probabilité d'inclusion d'ordre 1 d'un individu k appartenant à la post-strate $h \in \{1, \dots, H\}$ s'écrit :

$$\pi_k = \frac{n_{2,h}}{n_{1,h}} \frac{n_1}{N}$$

Où $n_{1,h}$ est la taille de la post-strate h de l'échantillon s_a à la 1ère phase et $n_{2,h}$ la taille de la post-strate h de l'échantillon s_b à la 2ème phase.

Le plan de sondage en deux phases est un plan complexe, nous renvoyons par exemple à (Särndal, Swensson, & Wretman, 2013) ou (Lohr, 2009) pour l'écriture des probabilités π_{kl} d'inclusion d'ordre 2 ainsi que pour celle des estimateurs du total, de la moyenne ou d'une proportion et de la variance associée.

Redressement par post-stratification

Il est possible, pour améliorer la précision des estimateurs, d'utiliser dans leur expression une ou plusieurs variables auxiliaires ; c'est-à-dire uniquement à l'étape de l'estimation, sans que ces variables n'aient été partie prenante dans l'étape de tirage. C'est ce que l'on appelle le *redressement*. Il existe différentes méthodes de redressement, la plus connue et la plus utilisée est la *post-stratification.* Nous avons déjà évoquée cette notion précédemment dans le cadre particulier du plan de sondage en deux phases. Nous abordons ici le cadre plus général du *redressement par post-stratification*.

On suppose que la variable auxiliaire X est qualitative et peut prendre H modalités notées $\{1, \dots, H\}$. Il est alors possible de partitionner la population U de taille N en tenant compte de cette information :

$$U = \bigcup_{h=1}^{H} U_h \ \ avec\ U_h = \{i \in U\ tq\ x_i = h\} \text{ et } U_h \cap U_{h'} = \emptyset, h' \neq h$$

Puisque cette stratification intervient après le sondage, les U_h sont appelées *post-strates*. L'échantillon S de taille n est lui aussi partitionné en H sous-échantillons S_h de taille n_h. Le nombre d'unités N_h de la post-strate U_h est connu[19] et appelé la taille de la post-strate ; bien entendu $N = \sum_{h=1}^{H} N_h$ et $n = \sum_{h=1}^{H} n_h$.

[19] Les N_h constituent l'information auxiliaire.

Le *π–estimateur* du total est donné par :

$$\hat{t}_{y,\pi} = \sum_{k \in S} \frac{y_k}{\pi_k} = \sum_{h=1}^{H} \sum_{k \in S_h} \frac{y_k}{\pi_k}$$

L'*estimateur post-stratifié* correspondant s'écrit :

$$\hat{t}_{y,post} = \sum_{h=1}^{H} \sum_{k \in S_h} \frac{y_k}{\tilde{\pi}_k} = \sum_{h=1}^{H} \sum_{k \in S_h} \tilde{w}_k y_k$$

Où $\tilde{w}_k = 1/\tilde{\pi}_k$ et $\tilde{\pi}_k = \pi_k \times c_h$; $c_h = {N_h}/{\sum_{k \in S_h} \pi_k}$ est appelé *coefficient du redressement* de la post-strate h.

Ainsi, par exemple, dans le cas d'un plan aléatoire simple, comme $\pi_k = n/N$, $k \in U$, on peut montrer que le coefficient de redressement de chaque post-strate h, s'écrit $c_h = \frac{N_h \times n}{n_h \times N}$, $h = 1, \dots, H$.

Les propriétés de l'estimateur post-stratifié peuvent être vues dans (Tillé, 2001) ou (Särndal, Swensson, & Wretman, 2013). On retiendra que l'estimateur post-stratifié n'est pas sans biais ; mais l'est approximativement dès lors que les post-strates sont non vides. De plus, la précision des estimateurs sera toujours meilleure avec une stratification *a priori* plutôt qu'avec une post-stratification. Mais si la stratification *a priori* n'est pas envisageable alors la post-stratification sera intéressante dès lors qu'il existe une relation suffisamment forte entre la ou les variables de post-stratification et la variable d'intérêt.

Enfin, le redressement par post-stratification nécessite la connaissance des effectifs de tous les croisements entre les modalités des variables de redressement. Si ce n'est pas le cas, d'autres méthodes peuvent être envisagées telles que la technique de *calage sur marges*.

Données Plomb-Habitat

L'enquête environnementale *Plomb-Habitat* a été réalisée à partir d'un sous-échantillon d'enfants issu de l'enquête de prévalence du saturnisme infantile (6 mois-6 ans) *Saturn-Inf* menée par l'InVS entre 2007 et 2009 en France. L'enquête *Plomb-Habitat* a été pilotée par le Centre Scientifique et Technique du Bâtiment (CSTB) et réalisée entre octobre 2008 et août 2009. La population cible initiale était constituée du parc de résidences principales en France métropolitaine. Le sous-échantillonnage de l'enquête *Plomb-Habitat* ayant été réalisé à partir de l'échantillonnage des enfants de *Saturn-Inf*, seul le parc de résidences principales situées en France métropolitaine, où au moins un enfant âgé de 6 mois à 6 ans était présent, a pu être décrit. Le déroulement de l'enquête ainsi que les exploitations des données ont été suivies dans le cadre d'un comité de pilotage appelé COPIL regroupant les différents partenaires de l'enquête : le CSTB, l'École des Hautes Études en Santé Publique (EHESP), l'Institut de Veille Sanitaire (InVS), l'Hôpital Lariboisière AP-HP (Assistance Publique - Hôpitaux de Paris) et l'Institut Supérieur d'Agriculture de Lille (ISA).

Le *plan de sondage* mis en œuvre dans l'enquête *Plomb-Habitat* est un *plan à 2 phases*. Il est résumé sur Figure 6 qui s'interprète de la manière suivante :

- L'enquête *Saturn-Inf* a constitué la **première phase**[20]. Le plan de sondage mis en œuvre dans cette première phase est un *plan à 2 degrés stratifié au premier degré*.
    - *au premier degré*, 135 hôpitaux ont été tirés comme unités primaires en fonction d'une stratification construite à partir des régions administratives (22 en France métropolitaine) et du groupe à risque plomb des bassins de population auxquels appartenait chaque hôpital. Ces groupes à risque ont été construits et intégrés dans la base de sondage par l'InVS. La probabilité d'inclusion de l'hôpital k est notée π_k.

[20] L'exposant ^{a} sera utilisé pour désigner la 1ère phase de manière à éviter la confusion due à l'indexation par *j* appliquée aux logements de la 2ème phase et aux enfants dans la 1ère.

- *au second degré*, l'inclusion des enfants s'est déroulée aléatoirement, au cours d'une période variable selon les établissements, en fonction des disponibilités des médecins investigateurs.

  À ce stade, par strate, la probabilité d'inclusion de l'enfant j de l'hôpital k était :

$$\pi_j^a = \pi_k \pi_{j|k}$$

  où $\pi_{j|k}$ représente la probabilité conditionnelle de l'enfant j, estimée par le nombre d'enfants inclus dans l'hôpital k divisé par le nombre d'enfants hospitalisés dans le service pendant la période d'étude. Le poids de sondage de l'enfant j était donc $w_j^a = 1/\pi_j^a$.

  Un coefficient de redressement (post-stratification) noté $c_{h_a},\ h_a = 1, \dots, H_a$ a été appliqué au poids de sondage de l'enfant par l'InVS (Lucas, 2013, p. 69). Le poids post-stratifié de l'enfant j dans *Saturn-Inf* a été calculé à l'aide de l'expression :

$$\widetilde{w}_j^a = c_{h_a} \cdot w_j^a$$

- La **deuxième phase** est constituée d'un *plan de sondage stratifié aléatoire simple* qui a permis d'obtenir le sous-échantillon d'enfants, donc les logements à enquêter (à un enfant est associé une unique résidence principale). Les strates ont été construites à partir de la région administrative d'hospitalisation et du niveau de plombémie [21] : $\{< 30\ \mu g/L\}$: tirage aléatoire ; $\{[30; 100[\ \mu g/L\}$: inclusion systématique et enfin $\{\geq 100\ \mu g/L\}$: inclusion systématique.

  Étant donné l'échantillon obtenu en 1ère phase, la probabilité d'inclusion π_j^b de l'enfant j dans l'enquête Plomb-Habitat, a été estimée par le nombre d'enfants inclus dans l'enquête Plomb-Habitat divisé par le nombre d'enfants éligibles à l'enquête Plomb-Habitat. Le poids de l'enfant dans l'enquête Plomb-Habitat était donc $w_j^b = \widetilde{w}_j^a / \pi_j^b$. Un coefficient de redressement a été appliqué sur ce poids w_j^b de manière à ce que la somme des poids des enfants inclus dans une région d'habitation donnée

[21] Par définition, la *plombémie* est la mesure du taux de Plomb présent dans le sang. Elle permet de détecter le *saturnisme*, maladie induite par l'intoxication de l'organisme par le Plomb ou ses dérivés. Elle est généralement donnée en µg/L ou ppm.

soit égale au nombre d'enfants recensés dans cette région. Le poids post-stratifié correspondant au logement j appartenant à la post-strate h_b s'écrit:

$\tilde{w}_j^b = c_{h_b} w_j^b$ où c_{h_b} désigne le coefficient de redressement associé à la post-strate h_b

A partir de ces poids finaux, les niveaux en Pb contenus dans différents compartiments environnementaux ont pu être estimés.

Les pièces, indicées i, à investiguer dans un logement donné ont été automatiquement incluses dès lors qu'elles étaient du type recherché. Leur probabilité d'inclusion conditionnelle $\pi_{i|j}$, tout comme leur poids de sondage conditionnel, est donc égale à 1. La taille du sous-échantillon obtenu a été de 484 logements au lieu de 500[22].

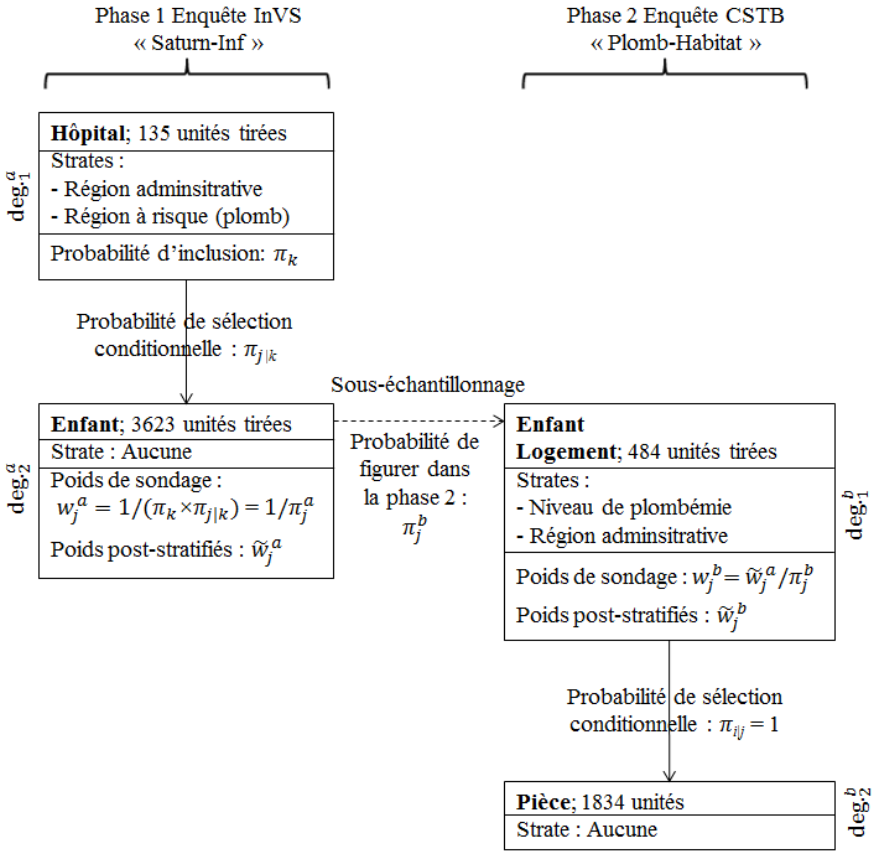


Figure 6 - Plan de sondage de l'enquête Plomb-Habitat.

[22] Plus de 500 enquêtes ont été réalisées; mais certaines d'entre elles n'ont pas été validées par la suite et ont donc été écartées.

Les nombreux objectifs de l'enquête *Plomb-Habitat* étaient :

- d'améliorer les connaissances sur les déterminants des plombémies ;
- d'identifier les sources et les compartiments environnementaux responsables des plombémies modérées (comprises entre 30 et 100 µg/L) ;
- de comparer la pertinence des analyses en plomb total et en plomb acido-soluble comme éléments explicatifs et/ou prédictifs des plombémies ;
- d'estimer la proportion de cas de saturnisme infantile (plombémie ≥ 100 µg/L) pour laquelle l'analyse des ratios isotopiques du plomb dans le sang et dans les compartiments environnement apportait une plus-value pour identifier la source ;
- d'établir un modèle empirique de prévision des plombémies en fonction des concentrations en plomb dans l'environnement ;
- de fournir un premier panorama de la contamination par le plomb dans le parc de logements français ;
- d'identifier les sources potentielles de contamination par le Pb des poussières à l'intérieure des logements.

Cette enquête a fait l'objet de 3 thèses de doctorat : Youssef Oulhote (EHESP/IRSET/INSERM U954), Jean-Paul Lucas (CSTB/EA4275 Univ. Nantes), Anne Etchevers (INSERM U1085/IRSET). Jean-Paul Lucas dans son travail de thèse s'est attelé à répondre aux deux derniers objectifs.

Estimation des niveaux en Pb en milieu résidentiel

À l'échelle métropolitaine, ***pour la première fois***, les distributions et la prévalence de dépassement de seuils des teneurs en Pb dans différents compartiments environnementaux des logements français ont été estimées. Les 5 compartiments environnementaux ciblés par l'enquête *Plomb-Habitat* ont été l'eau du robinet, la poussière déposée au sol à l'intérieur des logements, celle déposée au sol en parties communes, les revêtements intérieurs et le sol de l'aire de jeu extérieure de l'enfant. Cet état des lieux concerne les 3.6 millions résidences principales en France métropolitaine abritant au moins un enfant âgé de 6 mois à 6 ans. Pour cela, les outils de la *théorie des sondages* (voir par exemple (Lumley, 2010, pp. 28-37)) ont été appliqués pour obtenir les estimations

souhaitées et ainsi décrire les données de l'enquête *Plomb-habitat.* Afin d'améliorer les estimations, les poids de sondage w_j^b des logements ont tout d'abord été redressés par *post-stratification* sur des critères relatifs aux logements (Période de construction (<1949 ; ≥1949) ; Région administrative (22) et Type de logement (individuel ; collectif). Soit 88 $(2 \times 22 \times 2)$ post-strates initiales qui, pour certaines, ne possédaient aucun logement parmi l'échantillon des 484, pour d'autres qu'un nombre trop faible pour être prise en compte. Les post-strates ont donc été regroupées de telle sorte qu'une post-strate contienne au minimum 10 logements et que le regroupement soit cohérent. *In fine*, 24 post-strates ont construites qui ont permis de calculer les poids de sondage redressés $\tilde{w}_j^b$ des logements. Dès lors, une stratégie de description des niveaux en plomb a été mise en place, en particulier dans des sous-populations (domaines tels que par exemple la période construction), de manière à fournir un état de la contamination le plus informatif possible pour les pouvoirs publics. Les résultats détaillés se trouvent dans le manuscrit de thèse de Jean-Paul Lucas (Lucas, 2013, pp. 85-112). Un article (Lucas, et al., 2012) a aussi été publié.

Traduit en terme de pourcentage de logements incriminés, les niveaux en Pb ainsi que les prévalences de dépassement de seuils règlementaires européens ou américains indiquent que :

- le seuil règlementaire européen de 10 µg/L de Pb dans l'eau est dépassé pour 2.9% des logements ;
- environ 0.2% (*resp.* 4.1%) des logements (*resp.* parties communes d'immeubles collectifs) possèdent une charge moyenne en Pb dans les poussières supérieure au seuil règlementaire américain de 430 µg/m² ;
- 1.4% des aires de jeux extérieures ont une concentration en Pb supérieure au seuil américain de 400 mg/kg ;
- environ 24.5% (*resp.* 34.2 %) des logements (*resp.* des parties communes) possèdent un revêtement à base de Pb (≥ 1 mg/cm²) ;
- environ 4.7 % (*resp.* 7.1%) des logements (*resp.* parties communes) possèdent des revêtements dégradés à base de plomb (≥ 1 mg/cm²).

Ce travail a aussi permet de souligner que :

- Les niveaux en Pb mesurés dans la poussière intérieure déposée au sol sont approximativement de même ordre de grandeur dans les logements construits avant 1949, entre 1949 et 1974 et entre 1975 et 1993. Ils sont par contre plus faibles dans ceux construits après 1993.
- Dans les parties communes, les charges en Pb enregistrées dans la poussière déposée au sol sont 73 % plus élevées que celles à l'intérieur des logements. Elles ont de plus tendance à augmenter avec l'âge du bâtiment.

Cet état de la contamination servira dorénavant de référence pour les futures études évaluant les niveaux de contamination par le Pb en France en milieu résidentiel.

Des études ont par ailleurs montré que la poussière intérieure déposée au sol est le vecteur principal d'exposition au Pb chez l'enfant, donc potentiellement un bon prédicteur du niveau de plombémie infantile. Une autre grande partie de la thèse de Jean-Paul Lucas a consisté à étudier plus spécifiquement la provenance du Pb dans les poussières intérieures déposées au sol.

Modélisation des sources à contaminer la poussière intérieure

Modèle et résultats

Le Pb des poussières étant connu comme le principal prédicteur des plombémies infantiles (voir par exemple (Lanphear, et al., 1998) et (Lanphear B. , 2002)), la connaissance de la part attribuable à chaque source susceptible de contaminer en Pb la poussière pourra permettre aux pouvoirs publics de mettre en place des actions d'information et de réduction de ces sources de manière à réduire l'exposition au Pb de l'enfant.

Un *modèle multi-niveaux*[23] à 2 niveaux à *intercept*[24] aléatoire a été mis en œuvre sur les données de l'enquête *Plomb-Habitat*, afin d'expliquer les charges en Pb des poussières mesurées dans les 1834 pièces (niveau 1) investiguées dans les logements (niveau 2) en

[23] Aussi appelé *modèle hiérarchique* ou *modèle mixte*.

[24] Encore appelé en français : ordonnée à l'origine.

fonction des sources potentielles de contamination des poussières. Dans le cadre des données d'enquête, l'utilisation d'un modèle multi-niveaux se calque généralement sur un plan de sondage à plusieurs degrés. Les degrés du plan de sondage sont similaires aux niveaux du modèle ; à ceci près que la numérotation en est inversée. Dans notre cas, les unités de niveau 1 du modèle sont les pièces constituant le degré 2 du plan, et les logements, unités de niveau 2 du modèle correspondent aux unités du premier degré du plan. La conséquence importante est que les unités du niveau 1 du modèle sont les dernières unités échantillonnées suivant le plan de sondage.

La charge en Pb (en µg/m^2) de la poussière déposée au sol dans la pièce i au sein du logement j transformée[25] est notée y_{ij} et le modèle à 2 niveaux à *intercept* aléatoire se définit à l'aide des deux équations suivantes :

Niveau 1 (pièce *i* du logement *j*) : $y_{ij} = \beta_{0j} + \sum_{m=1}^{q_1} \varphi_m x_{ij}^{(m)} + \varepsilon_{ij}$ où $i = 1, \dots, n_j^{(1)}$

Niveau 2 (logement *j*) : $\beta_{0j} = \beta_0 + \sum_{r=1}^{q_2} \psi_r x_j^{(r)} + \zeta_j$ où $j = 1, \dots, n^{(2)}$

avec $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma_1^2)$, $\zeta_j \sim \mathcal{N}(0, \sigma_2^2)$ supposées indépendantes.

Pour ajuster un tel modèle sur des données d'enquête dans lequel le plan est *informatif*[26] pour la variable d'intérêt Y, les poids de sondage associés à chaque individu (ici la pièce *i* du logement *j*) de l'échantillon S doivent être introduits dans la vraisemblance sinon les estimateurs obtenus sont biaisés (Pfeffermann, Skinner, Holmes, Goldstein, & Rasbash, 1998). On introduit alors la pseudo-vraisemblance $\mathcal{L}(\boldsymbol{y})$ (Skinner, 1989), qui conduit au calcul de l'estimateur du pseudo-maximum de vraisemblance (PML), en associant à chaque unité un poids appelé *poids conditionnel* égal à l'inverse de sa probabilité d'inclusion conditionnelle. Dans le cas d'un modèle à 2 niveaux, sur données d'enquête issues d'un plan de sondage à plusieurs degrés, ils s'écrivent :

[25] La transformation classique *ln* a été appliquée pour atténuer la dissymétrie à droite de la distribution de la charge en Pb dans la poussière.

[26] Dans un plan *informatif*, la probabilité de tirage des échantillons est liée, d'une façon complexe, aux valeurs de la variable d'intérêt ; ce qui est le cas dans l'enquête *Plomb-Habitat*.

$$w_{i|j}^{(1)} = {}^{1}/_{\pi_{i|j}} \text{ et } w_j^{(2)} = {}^{1}/_{\pi_j}$$

où i indexe les unités du niveau 1 indiqué par $^{(1)}$ et j indexe celles du niveau 2 indiqué par $^{(2)}$.

La log-pseudo vraisemblance a alors pour expression :

$$\mathcal{L}(\boldsymbol{y}) = \sum_{j=1}^{n^{(2)}} w_j^{(2)} log\left[\int exp\left\{\sum_{i=1}^{n_j^{(1)}} w_{i|j}^{(1)} log\left(f(y_{ij}|\zeta_j)\right)\right\} g(\zeta_j) d\zeta_j\right]$$

où $\sum_{i=1}^{n_j^{(1)}} w_{i|j}^{(1)} log\left(f(y_{ij}|\zeta_j)\right)$ représente la contribution à la log-vraisemblance des unités de niveau 1 conditionnellement à l'existence de l'effet aléatoire ζ_j au niveau 2 et $g(\zeta_j)$ est la densité l'effet aléatoire ζ_j, *i.e.* une loi $\mathcal{N}(0, \sigma_2^2)$.

Pour un modèle à 3 niveaux ou plus, l'expression de la log-pseudo vraisemblance s'écrit sur le même principe ; mais devient vite compliquée à formaliser et nécessite une écriture par récurrence. Les variances des estimateurs des effets fixes et des estimateurs des paramètres de variance σ_1^1 et σ_2^1 sont ensuite obtenues par linéarisation de Taylor. On pourra se reporter pour plus détails à (Rabe-Hesketh & Skrondal, 2006). Cependant, malgré l'introduction de poids conditionnels les estimateurs ainsi obtenus peuvent s'avérer encore biaisés.

Dans l'enquête *Plomb-Habitat*, les pièces d'un logement (unités du niveau 1) ont été investiguées sans tirage aléatoire préalable, par conséquent leur probabilité de sélection conditionnelle $\pi_{i|j}$ (*resp.* le poids conditionnel associé $w_{i|j}^{(1)}$) vaut 1. La problématique relevée dans la littérature liée à leur probabilité de sélection inégale pouvant induire des biais n'a donc pas lieu d'être ici. En revanche, les logements (unités du niveau 2) ne correspondent pas aux entités du plus haut niveau possible. Dans le plan de sondage de Plomb-Habitat (voir description pages 29-30 et Figure 6 page 30), ce sont les hôpitaux en tant qu'unités hiérarchiquement supérieures qui constituent des unités hiérarchiquement supérieures aux logements. La question du choix de la pondération $w_j^{(2)}$ à associer aux logements dans le modèle se pose (Figure 7).

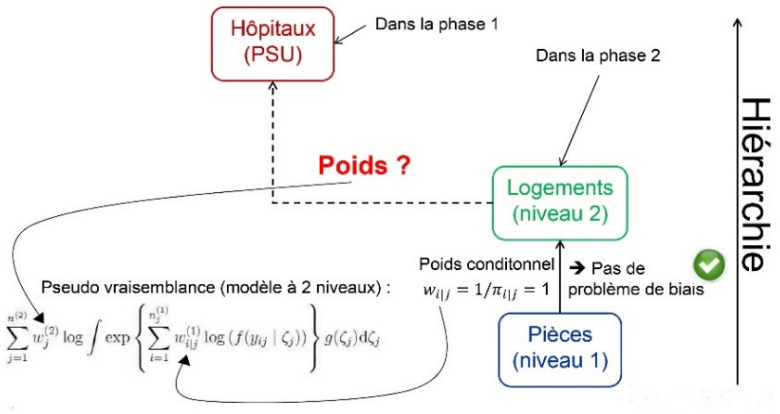


Figure 7 - Problème du choix des poids de niveaux 2 dans l'enquête Plomb-Habitat.

De plus, le choix du type de pondération pour le niveau 2, $w_j^{(2)}$, d'un modèle multi-niveau ne semblait pas avoir été abordé précédemment dans la littérature et les candidats potentiels, liés à la structure complexe de l'enquête *Plomb-Habitat*, sont nombreux. Voici, ci-dessous, les vecteurs de pondération qui ont été considérés :

1. $\mathbf{w_1}$: $\left\{1/\pi_j^b ; j = 1, \dots, n^{(2)}\right\}$; sorte de poids conditionnels ; même si π_j^b n'est pas une véritable probabilité conditionnelle puisqu'elle est située entre les deux phases du plan de sondage et non entre deux degrés.
2. $\mathbf{w_2}$: $\left\{w_j^b ; j = 1, \dots, n^{(2)}\right\}$; poids de sondage non stratifiés des logements ;
3. $\mathbf{w_3}$: $\left\{\widetilde{w}_j^b ; j = 1, \dots, n^{(2)}\right\}$; poids de sondage post-stratifiés sur critères logements (période de construction, région et type de logement).
4. $\mathbf{w_4}$: $\left\{1/\left(\pi_j^a \times \pi_j^b\right) ; j = 1, \dots, n^{(2)}\right\}$; poids de sondage sans aucune post-stratification, ni à l'étape de l'enfant, ni à celle du logement.
5. $\mathbf{w_5}$: $\left\{1 ; j = 1, \dots, n^{(2)}\right\}$; poids conduisant à une modélisation non pondérée.
6. $\mathbf{w_6}$: $\left\{1/\left(\pi_{j|k} \times \pi_j^b\right) ; j = 1, \dots, n^{(2)}\right\}$; poids correspondant aux produits de probabilités conditionnelles.

Afin d'évaluer l'impact que pourrait avoir la déclaration des hôpitaux dans la pseudo vraisemblance comme unités hiérarchiquement supérieures aux logements, un *modèle*

hiérarchique à 3 niveaux à *intercept* aléatoire a aussi été testé. On rappelle son écriture dans le contexte de nos données :

Niveau 1 (pièce *i* du logement *j* appartenant au bassin de population de l'hôpital *k*) :

$$y_{ijk} = \beta_{0jk} + \sum_{m=1}^{q_1} \varphi_m x_{ijk}^{(m)} + \varepsilon_{ijk} \text{ où } i = 1, \dots, n_{jk}^{(1)}$$

Niveau 2 (logement *j* appartenant au bassin de population de l'hôpital *k*) :

$$\beta_{0jk} = \beta_{0k} + \sum_{r=1}^{q_2} \psi_r x_{jk}^{(r)} + \zeta_{jk} \text{ où } j = 1, \dots, n_k^{(2)}$$

Niveau 3 (hôpital *k*) : $\beta_{0k} = \beta_0 + \sum_{p=1}^{q_3} \delta_p x_k^{(p)} + \xi_k$ où $j = 1, \dots, n^{(3)}$

avec $\varepsilon_{ijk} \sim \mathcal{N}(0, \sigma_1^2)$, $\zeta_{jk} \sim \mathcal{N}(0, \sigma_2^2)$ et $\xi_k \sim \mathcal{N}(0, \sigma_3^2)$ supposées indépendantes.

Il est important de remarquer que ce modèle ne correspond pas vraiment à la réalité des données recueillies puisque les 2 phases du plan de sondage (Figure 6) sont ignorées. Le plan est donc traité comme un plan de sondage à 3 degrés (hôpitaux, logements, pièces). Les hôpitaux formant alors les unités supérieures du plan de sondage, la seule pondération possible à leur associer dans la log-pseudo vraisemblance est $1/\pi_k ; k = 1, \dots, n^{(3)}$. Concernant le niveau 2 (logement), plusieurs possibilités de poids $w_j^{(2)}$ existent :

7. $\mathbf{w_7}$: $\left\{1/\pi_j^b ; j = 1, \dots, n_k^{(2)}\right\}$; identique à $\mathbf{w_1}$ dans le modèle à 2 niveaux.
8. $\mathbf{w_8}$: $\left\{1; j = 1, \dots, n_k^{(2)}\right\}$; pas de pondération pour le niveau 2.
9. $\mathbf{w_9}$: $\left\{1/\left(\pi_{j|k} \times \pi_j^b\right) ; j = 1, \dots, n_k^{(2)}\right\}$; identique à $\mathbf{w_6}$ dans le modèle à 2 niveaux.

Ces modèles ont été ajustés pour le Pb total et le Pb acido-soluble. Les données manquantes induisaient plus de 12% de perte d'observations. Les 9 scénarios ont été testés sur cas complets et sur données imputées à l'aide des équations chainées (ICE[27]) (White, Royston, & Wood, 2011) (1605 pièces (cas complets) et 1834 pièces réparties dans 429 logements). Les précisions des estimations obtenues pour les paramètres du modèle liés aux variables explicatives définies comme sources potentielles en Pb contaminant la

[27] Acronyme de *Imputation using Chained Equations*.

poussière intérieure ont été analysées. Les résultats sur cas complets et sur données imputées étaient très proches. Des écarts ont été observés selon le scénario choisi et il a donc été décidé de poursuivre le travail de modélisation sans introduire de pondération (scénario $\mathbf{w_5}$). Ce choix paraît être en effet le plus sage quand *i)* les informations *a priori* disponibles sur le type de plan de sondage utilisé sont insuffisantes, voire inexistantes ou, *ii)* ce qui est notre cas, quand le modèle multi-niveaux retenu n'est pas le reflet exact du plan sondage complexe ayant permis de générer les données. Une étude de simulation de Monte Carlo a ensuite été réalisée pour conforter le choix d'analyser les données de l'enquête à l'aide d'un modèle hiérarchique à 2 niveaux sans pondération ; nous reviendrons sur ce point plus loin.

Ce travail a permis d'étudier conjointement, pour la première fois, un grand nombre de sources pouvant contaminer la poussière intérieure. On a pu montrer que le Pb des poussières intérieures déposées au sol provenait majoritairement du Pb des poussières du palier. Le Pb provenant de l'extérieur, en particulier du sol et des poussières de l'aire de jeu extérieure de l'enfant étaient aussi des contributeurs de la contamination des poussières intérieures. Il a été montré en outre que les sites polluants, les démolitions d'anciens immeuble et le tabagisme à l'intérieur étaient des contributeurs substantiels. Les revêtements intérieurs à base de Pb ne sont plus à même de contribuer de manière importante à la contamination des poussières intérieures en population générale. La corrélation intra-classe, entre 2 charges en Pb de la poussière à l'intérieur d'un même logement, a été estimée approximativement égale à 0,60. Ces résultats, publiés (Lucas, et al., 2014), ont donc permis pour la première fois d'identifier des sources de la contamination des poussières intérieures et d'évaluer leur part d'importance relative. Ils peuvent à l'avenir permettre aux pouvoirs publics de prendre des décisions pour tenter de réduire les niveaux en plomb de ces sources.

Impact du choix des poids de niveau 2

Le traitement des données de l'enquête *Plomb-Habitat* à l'aide d'une modélisation à 2 niveaux adaptée aux données d'enquête, a soulevé le problème de la difficulté du choix de la pondération à inclure dans la log-pseudo vraisemblance. Alors que ce problème a été abordé dans la littérature pour les poids du niveau 1, pour le niveau 2, aucune référence n'aborde le sujet. Une *étude de simulation de Monte-Carlo* basée sur nos données a été réalisée afin de comparer les différentes pondérations à attribuer aux logements (niveau 2) dans le calcul de la pseudo-vraisemblance associée au modèle multi-niveaux. Il a fallu pour cela générer une population de logements la plus réaliste possible ; mais aussi sélectionner les individus à intégrer dans l'analyse avec la même procédure complexe de sondage en deux phases, estimer les paramètres du modèle multi-niveaux pour chacun des 9 scénarios et enfin réitérer le tout 500 fois.

Voici les différentes étapes de ce travail de comparaison par simulation :

1. construction d'un fichier à N=3581991 lignes correspondant aux logements et création d'un fichier associé dont les lignes sont les pièces des logements ;
2. création dans ces fichiers de colonnes correspondant à chaque covariable du modèle, en simulant les valeurs de ces covariables ;
3. génération de Y suivant un modèle multi-niveaux dont les paramètres estimés ont été fixés ;
4. tirage aléatoire d'un échantillon de 484 logements à partir du plan de sondage défini ;
5. estimations des paramètres du modèle multi-niveaux pour chacun des 9 scénarios ;
   Retourner à l'étape 1 et recommencer 500 fois.
6. Comparaison des estimations des paramètres du modèle à leur vraie valeur. Les scénarios ont été comparés en analysant le biais, le biais relatif et l'erreur quadratique moyenne associés à chacun des paramètres du modèle (les coefficients de régression ainsi que les deux paramètres de variances).

Ce travail a permis de conforter l'approche à l'aide d'un modèle à 2 niveaux non pondéré, adoptée pour expliquer les charges en Pb dans la poussière issues de l'enquête *Plomb-Habitat* dont le plan de sondage est complexe. Les différentes stratégies de pondération envisagées conduisent toutes à des estimateurs des paramètres dont le biais est proche de zéro. La variance du scénario $\mathbf{w_5}$, ne prenant pas en compte de poids de sondage, est parmi les plus faibles, ce qui conforte le choix que nous avions fait pour analyser les données d'enquête dans (Lucas, et al., 2014). Cette évaluation par simulation a aussi permis d'émettre un certain nombre de recommandations sur la mise en œuvre d'un modèle à 2 niveaux sur données d'enquête :

- Quand le plan de sondage utilisé pour récolter les données comporte des entités supérieures hiérarchiquement au niveau 2, il est préférable de ne pas utiliser de poids de sondage finaux tels que $\mathbf{w_2}$ ou $\mathbf{w_3}$; mais d'utiliser plutôt des poids conditionnels tels que $\mathbf{w_1}$ ou pas de poids avec $\mathbf{w_5}$.
- Quand les estimations sont très différentes entre les analyses pondérées et l'analyse non pondérée, il faut s'interroger sur l'adéquation entre le modèle choisi et le plan de sondage adopté pour générer les données et valider le choix du jeu de pondération par une étude de simulation.

Ce travail a été publié (Lucas, Sébille, Le Tertre, Le Strat, & Bellanger, 2014).

Perspectives

Ce travail pourrait donner lieu à de nombreuses perspectives tant pratiques que théoriques :

- définir le risque plomb global d'un logement à partir de l'agrégation des niveaux en Pb des compartiments environnementaux résidentiels ;
- proposer un nouveau Constat de Risque d'Exposition au Pb (CREP) ne mettant pas seulement en évidence le risque Pb lié aux revêtements ; mais qui permette de diagnostiquer le risque Pb global d'un logement ;
- poursuivre de manière plus théorique la modélisation multi-niveaux sur données d'enquête.

Mais, me concernant, à ce jour, ma contribution au domaine de la contamination par le Pb dans les logements s'arrête là ; par contre mon intérêt pour la théorie des sondages se

poursuit dans un autre domaine, la pêche, comme je l'évoquerai plus loin. Un projet que j'aimerais mener à bien, en lien direct avec ce travail, concerne la rédaction d'un ouvrage sur la planification d'expérience, thématique que j'enseigne et sur laquelle j'ai aussi un peu travaillé (Tomassone, Charles-Bajard, & Bellanger, 2000), et la théorie des sondages avec le logiciel *R*. En effet, ces deux domaines sont régulièrement confondus et de plus peu d'ouvrages pratiques existent dans ces deux domaines.

1.2 L'ECOLOGIE MARINE (depuis 2008)

Collaborateurs sur ce thème :

- IFREMER[28] (dpt EMH, IFREMER, Nantes) : A. Brind'Amour, S. Mahévas, V. Trenkel ;
- Université de Montréal (dpt des Sciences Biologiques) : P. Legendre.

1.2.1 Caractérisation des zones et saisons des activités de pêche

Depuis quelques décennies, les autorités ; mais aussi nous simple citoyen, prenons peu à peu conscience de l'état de santé alarmant de certains écosystèmes marins ; ainsi que de l'impact écologique des captures intensives et des captures accidentelles. Seule une approche pluridisciplinaire peut permettre d'aboutir à un état des lieux quantitatif et à la réalisation d'outils d'aide à la gestion et à l'évaluation de la viabilité des pêcheries tant sur le plan écologique que socio-économique.

La mer Celtique, pour laquelle un historique de données suffisant est disponible, est un terrain d'étude de premier choix pour analyser les caractéristiques spatio-temporelles des temps de pêche de chalutiers. Le temps de pêche d'un navire est défini comme le nombre d'heures en pêche, c'est à dire le nombre d'heures de présence en mer déduit du temps de route (le temps passé sur le chemin vers le lieu de pêche, entre différents lieux de pêche et au retour d'un lieu de pêche). Le temps consacré par les navires à la recherche du poisson est également considéré comme du temps de pêche. Les navires considérés, les chalutiers français, capturent principalement dans la partie centrale de la mer Celtique les espèces démersales-benthiques, c'est-à-dire les espèces vivant dans le fond de la mer

[28] Acronyme de *Institut Français de Recherche pour l'Exploration de la MER*.

(comme par exemple la langoustine ou le merlu). Tous les navires français inscrits au fichier de la Flotte de Pêche Communautaire (FPC) sont tenus de déclarer leur temps de pêche et leurs captures par rectangle statistique (unité spatiale internationale de découpage des océans) pour chaque sortie en mer. La partie centrale de la mer Celtique étudiée est constituée de 48 rectangles statistiques. Elle s'étend des côtes du nord de la Bretagne jusqu'aux côtes du sud de l'Irlande et de l'ouest du Pays de Galles. Certains rectangles statistiques de la zone étudiée ne sont pas entièrement composés de mer : pour les rectangles côtiers, une partie du continent réduit la surface « pêchable ». Pour prendre en compte ces différences dans nos analyses nous avons étudié les densités d'efforts de pêche plutôt que les temps de pêche bruts.

(Mahévas & Trenkel, 2002) ont utilisé un *modèle linéaire généralisé avec effets fixes et/ou aléatoires* prenant explicitement en compte les corrélations temporelles et spatiales pour modéliser le temps de pêche disponibles par bateau, par unité spatiale et temporelle sur la période 1991-1998. Ces données proviennent de 589 navires dont la taille varie entre 12 et 24 mètres. Elles sont disponibles par rectangle et par mois de la période d'étude. Au total, on disposait de 4536 temps de pêche. Les variables explicatives disponibles pour cette modélisation étaient l'unité spatiale, le mois, et l'année. Les auteures ont ainsi montré qu'il existait une corrélation temporelle d'ordre 1 (*i.e.* le mois précédent) et une corrélation spatiale du temps de pêche ; mais aussi une stabilité annuelle de l'activité de pêche sur la période. Ainsi, le temps de pêche dans une zone à un mois donné est stable sur les années étudiées ; il est dépendant *i)* du temps de pêche dans les zones adjacentes et *ii)* du temps de pêche au mois précèdent. Cependant, cette modélisation ne permettait pas de mettre en évidence des patterns d'unités spatiales et d'unités temporelles sur les données d'effort de pêche de la flottille française opérant en mer Celtique. J'ai donc débuté une collaboration avec S. Mahévas et V. Trenkel (dpt EMH, IFREMER, Nantes) dans le but de répondre à cette problématique. Une *méthode de classification avec contraintes de contiguïté* a été appliquée aux effets fixes spatiaux et temporels estimés à partir d'un modèle linéaire généralisé décrivant la variabilité de l'effort de pêche T en fonction d'effets fixes spatiaux (à l'échelle du rectangle statistique j) et temporels (à l'échelle du mois i et de l'année k) et en prenant en compte les corrélations spatiales et temporelles :

$T_{ijk}^{1/4} = m + \delta T_{(i-1)jk}^{1/4} + mois_i + rectangle_j + an_k + \varepsilon_{ijk}$, i=1,...,12 ; j=1,...,48; k=1,...,8

où

- T_{ijk} : temps de pêche au mois i dans le rectangle j pour l'année k, pour i=1,...,12; j=1,...,48; k=1,...,8
- $T_{(0)jk}^{1/4} = T_{(12)j(k-1)}^{1/4} ; k \geq 2$
- $T_{(i-1)jk}^{1/4}$: effet décrivant la persistance temporelle d'ordre 1
- $\boldsymbol{\varepsilon} \sim N_{4488}(\mathbf{0}; \boldsymbol{\Sigma})$ où $\boldsymbol{\Sigma} = \sigma^2 \mathbf{H}(\varphi) + \tau^2 \mathbf{I}$ où $\mathbf{H}(\varphi)_{jj'} = \rho(\varphi; d_{jj'})$, $d_{jj'}$ est la distance euclidienne entre les rectangles j et j', φ le paramètre de décroissance, ρ la fonction classique de covariance exponentielle des temps de pêche (voir par exemple (Cressie, 1993)).

Une fois le modèle validé. Une *classification ascendante hiérarchique* (*CAH*) *avec contraintes de contiguïté* a été appliquée à la matrice de dissemblance[29] représentée par les valeurs (1-*p values*) obtenues à partir des tests *F* de l'égalité deux à deux des effets $rectangle$ (*resp.* $mois$) ((Searle, 1997); (Rawlings, Pantula, & Dickey, 2001)). Les contraintes de contiguïté spatiales et temporelles sont imposées dans l'algorithme de classification pour s'assurer que seules les unités spatiales voisines et les unités temporelles successives sont bien agglomérées. Si η représente la dissemblance, une méthode simple pour définir l'index d'agrégation γ dans l'algorithme *CAH* utilisant la stratégie d'agrégation du lien simple (voir par exemple (Bellanger & Tomassone, 2014, p. 201) consiste à prendre :

$$\gamma(rectangle_i; rectangle_{i'}) = \eta(rectangle_i; rectangle_{i'}) + \kappa(rectangle_i; rectangle_{i'})$$

où κ est l'index de contiguïté qui prend la valeur 0 si les deux rectangles sont contigus et $+\infty$ sinon.

[29] Encore appelée matrice de *dissimilarité*.

Dans l'algorithme de classification, la dissemblance γ est alors utilisée à la place de η. Nous avons ainsi mis en évidence :

- 19 zones (Figure 8), de taille très variables mais généralement plus étendues au large qu'à proximité de la côte.

  La mer Celtique sud-ouest semble être plus homogène que les côtes. Ces résultats confirment des observations antérieures et pourraient s'expliquer par l'hétérogénéité plus prononcée des activités de pêche côtière par rapport à celles de pêche off-shore.
- 9 saisons caractérisées par des hétérogénéités plus marquées en hiver et au printemps.

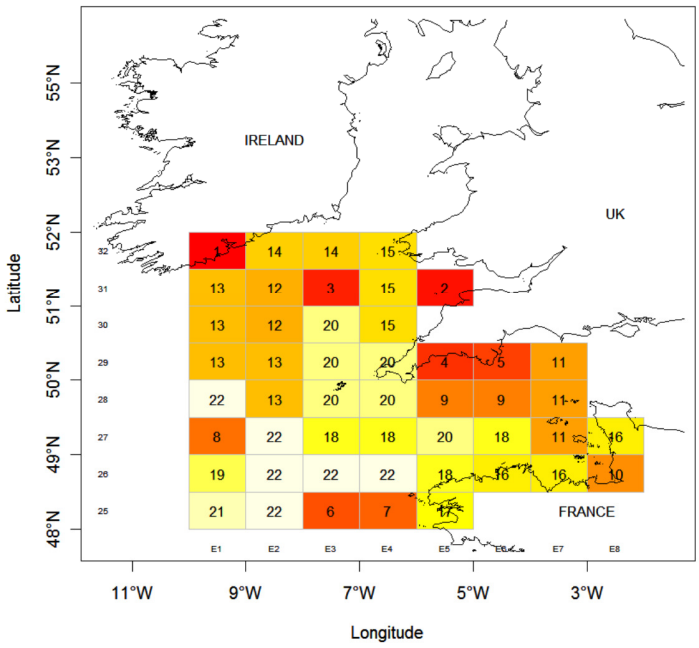


Figure 8 - Zones de pêche obtenue à partir d'une CAH avec contraintes de contiguïté pour la flottille francaise des chalutiers pêchant sur le plateau de la mer Celtique - période 1991-1998[30]. *(Mahévas, Bellanger, & Trenkel, 2008).*

Ce travail a donné lieu à un article (Mahévas, Bellanger, & Trenkel, 2008) et a été présenté lors de la XXIV International Biometric Conference de 2008 à Dublin(Irlande).

[30] Les rectangles appartenant à une même zone de pêche sont de même couleur et ont le même numéro.

L'approche adoptée dans cette étude est générique et peut être généralisée à toute variable de réponse. Cependant, il aurait été bon de tester la stabilité des partitions obtenues en utilisant d'autres stratégies d'agrégation que celle du lien simple ; la macro `contigclust` sous le logiciel SAS permettrait par exemple maintenant d'effectuer un tel travail de comparaison.

Je poursuis actuellement cette collaboration avec Stéphanie Mahévas et Anik Brind'Amour (Dpt EMH, IFREMER, Nantes) et Pierre Legendre (Pr Université de Montréal, Dept des Sciences Biologiques, Québec). Nous travaillons sur l'exploration de données spatiales à l'aide de méthodes multivariées telles que *Principal Coordinates of Neighbour Matrices* (*PCNM*) et sa généralisation *Moran's Eigenvector Map* (*MEM*). Ces méthodes forment en effet une nouvelle famille de méthodes multidimensionnelles pour l'analyse spatiale multi-échelle de données univariées ou multivariées. Elles sont une alternative aux modèles d'Auto-régression simultanée par exemple. Il est donc important de mieux comprendre leurs propriétés pour mieux appréhender les situations adaptées à leur mise en œuvre.

1.2.2 Analyse des structures spatiales à l'aide de la méthode *MEM* (depuis 2010)

Fondements théoriques de la méthode MEM

Moran's Eigenvector Map (*MEM*) est une méthode d'analyse de données multivariées reposant sur des objets mathématiques relativement simples à calculer. Elle s'applique à une grande variété de jeux de données. Elle permet d'obtenir des variables spatiales (*resp.* temporelles) orthogonales entre elles décrivant finement l'espace (*resp.* le temps) à différentes échelles. Ses nouvelles variables *MEM* peuvent ensuite facilement être intégrées en tant que variables explicatives dans un modèle linéaire généralisé par exemple. Cependant cette utilisation dans un cadre d'une modélisation doit se faire avec précaution puisque le plan d'échantillonnage a un impact sur le calcul des *MEM*. Notre travail doit permettre d'identifier un cadre de « bonne » utilisation.

Commençons par résumer le principe de cette méthode basée sur l'introduction des relations de voisinages dans l'analyse multivariée des structures spatiales (*resp.* temporelles) avant d'évoquer les questions que nous nous posons.

Variance totale, variance locale et variabilité globale

Notons :

- $\boldsymbol{y} = [y_i] \in \mathbb{R}^n$, un vecteur-colonne correspondant à la variable $\boldsymbol{y}$;
- $\mathbf{W} = \left[w_{ii'}\right] \in \mathcal{M}_{n\times n}$, matrice de pondération spatiale permettant de quantifier la proximité (géographique) entre les différents sites. La plus couramment utilisée est une matrice de contiguïté binaire **B** construite sur la base des distances entre les sites :

$$w_{ii'} = b_{ii'} = \begin{cases} 1 \; si \; le \; site \; i \; est \; voisin \; du \; site \; i' ; \; \forall i \neq i' \\ 0 \; sinon \end{cases}.$$

  Mais d'autres choix sont possibles pour **W** ;

- $\mathbf{P} = \left[p_{ii'}\right] \in \mathcal{M}_{n\times n}$, matrice obtenue à partir de **W** telle que $p_{ii'} = \frac{w_{ii'}}{\sum_{i,i'} w_{ii'}}$ ou $p_{ii'} = \frac{w_{ii'}}{2w}$ où, si **W** est une matrice de contiguïté, $w = \frac{1}{2}\sum_{i,i'} w_{ii'}$ correspond au nombre total de paires de voisins. Par construction $\sum_{i,i'} p_{ii'} = 1$;
- $\mathbf{D} = diag\left[p_{i+} = \sum_{i'} p_{ii'}\right] \in \mathcal{M}_{n\times n}$, matrice diagonale des poids des voisins. On parlera de poids uniformes si $\mathbf{D} = diag\left[\frac{1}{n}, i = 1, \dots, n\right] \in \mathcal{M}_{n\times n}$,

La description de la variable $\boldsymbol{y}$ peut alors être effectuée à deux échelles spatiales : une locale et une globale, la variabilité totale étant alors décomposée entre ses deux échelles.

Etant donné la matrice de pondération **D**, on définit la *moyenne* de la variable $\boldsymbol{y}$ par :

$$\bar{y}_{\mathbf{D}} = \sum_i p_{i+} y_i = (\boldsymbol{y})^{\mathrm{T}} \mathbf{D} \mathbf{1}_n$$

$\boldsymbol{y}$ est dit **D** –centré si sa moyenne connaissant **D** vaut 0. Il est noté $\boldsymbol{y}_{C_{\mathbf{D}}}$, défini par :

$$\boldsymbol{y}_{C_{\mathbf{D}}} = \boldsymbol{y} - \mathbf{1}_n \bar{y}_{\mathbf{D}}$$

Et donc la *variance totale* de $\boldsymbol{y}$ s'écrit : $Var(\boldsymbol{y}) = \left(\boldsymbol{y}_{C_{\mathbf{D}}}\right)^{\mathrm{T}} \mathbf{D} \boldsymbol{y}_{C_{\mathbf{D}}}$

La *variance locale* de $\boldsymbol{y}$ s'écrit : $LV(\boldsymbol{y}) = \sum_{i,i'} p_{ii'}(y_i - y_{i'})^2 = (\boldsymbol{y})^{\mathrm{T}}(\mathbf{D} - \mathbf{P})\boldsymbol{y}$

Elle représente la moyenne des écarts quadratiques entre sites.

Enfin, la *variabilité globale*[31] (ou auto-covariance spatiale) de la variable $\boldsymbol{y}$ entre les n sites est définie par :

$$GV(\boldsymbol{y}) = \sum_{i,i'} p_{ii'}(y_i - \bar{y}_{\mathbf{D}})(y_{i'} - \bar{y}_{\mathbf{D}}) = (\boldsymbol{y}_{C_{\mathbf{D}}})^{\mathrm{T}}\mathbf{P}\boldsymbol{y}_{C_{\mathbf{D}}}$$

On peut alors montrer que la variance totale se décompose comme suit :

$$Var(\boldsymbol{y}) = LV(\boldsymbol{y}) + GV(\boldsymbol{y})$$

Ainsi, la variance totale se décompose en une variabilité locale et une variabilité globale qui prennent en compte les relations de voisinage entre les sites. Pour plus de détails sur ces notions et leurs propriétés, nous renvoyons par exemple à (Thioulouse, Chessel, & Champely, 1995).

Relations avec l'indice de Moran

Indice d'autocorrélation spatiale de Moran des observations $y_1, \ldots, y_n$, noté $I(\boldsymbol{y})$, permet d'estimer l'importance de l'autocorrélation spatiale entre sites en tenant compte des relations de voisinage. Il se définit de la manière suivante :

$$I(\boldsymbol{y}) = \frac{n}{\sum_{i,i'} w_{ii'}} \frac{\sum_{i,i'} w_{ii'}\left(y_i - \bar{y}_{\mathbf{D}}\right)\left(y_{i'} - \bar{y}_{\mathbf{D}}\right)}{\sum_i \left(y_i - \bar{y}_{\mathbf{D}}\right)^2} = \frac{n}{(\mathbf{1}_n)^{\mathrm{T}}\mathbf{W}\mathbf{1}_n} \frac{\boldsymbol{y}^{\mathrm{T}}\left(\mathbf{I} - \frac{1}{n}\mathbf{1}_n(\mathbf{1}_n)^{\mathrm{T}}\right)\mathbf{W}\left(\mathbf{I} - \frac{1}{n}\mathbf{1}_n(\mathbf{1}_n)^{\mathrm{T}}\right)\boldsymbol{y}}{\boldsymbol{y}^{\mathrm{T}}\left(\mathbf{I} - \frac{1}{n}\mathbf{1}_n(\mathbf{1}_n)^{\mathrm{T}}\right)\boldsymbol{y}}$$

Et si on suppose $\boldsymbol{y}$ centré, alors :

$$I(\boldsymbol{y}) = I(\boldsymbol{y}_{C_{\mathbf{D}}}) = \frac{n}{(\mathbf{1}_n)^{\mathrm{T}}\mathbf{W}\mathbf{1}_n} \frac{\boldsymbol{y}_{C_{\mathbf{D}}}{}^{\mathrm{T}}\mathbf{W}\boldsymbol{y}_{C_{\mathbf{D}}}}{\boldsymbol{y}_{C_{\mathbf{D}}}{}^{\mathrm{T}}\boldsymbol{y}_{C_{\mathbf{D}}}} :$$

Où **W** vaut bien souvent **B** la matrice de contiguïté binaire.

---

[31] Comme $GV(\boldsymbol{y})$ n'est pas toujours positif, on ne peut pas parler de variance d'où la terminologie « variabilité ».

Comme une corrélation, l'indice d'autocorrélation spatiale de Moran mesure la ressemblance entre voisins, il varie généralement entre -1 et +1 ; mais peut parfois prendre des valeurs supérieures à +1 ou inférieures à -1 (Cliff & Ord, 1981) :

- Un indice positif traduit une autocorrélation spatiale positive : on parle d'attraction. Autrement dit, deux observations provenant de deux sites voisins auront tendance à prendre des valeurs proches les unes des autres.
- Un indice négatif indique la présence d'une autocorrélation spatiale négative : on parle de répulsion. Les valeurs de deux observations provenant de sites voisins auront tendance à être très différentes l'une de l'autre.
- Un indice proche de 0 traduit un pattern aléatoire.

Dans le cas où les poids sont uniformes *i.e.* $\mathbf{D} = diag\left[\frac{1}{n}, i = 1, \dots, n\right]$, l'indice de Moran vaut exactement :

$$I(\boldsymbol{y}) = \frac{GV(\boldsymbol{y})}{Var(\boldsymbol{y})} = 1 - \frac{LV(\boldsymbol{y})}{Var(\boldsymbol{y})}$$

On déduit que :

- Si $I(\boldsymbol{y}) \simeq 1$ alors $\frac{LV(\boldsymbol{y})}{Var(\boldsymbol{y})} \simeq 0$. Autrement dit, la part de variance locale contenue dans $Var(\boldsymbol{y})$ est très faible, ce qui traduit de faibles variations de $\boldsymbol{y}$ entre sites géographiques voisins. Dans cette situation, nous dirons que $\boldsymbol{y}$ possède une *structure spatiale à grande échelle*.
- Si $I(\boldsymbol{y}) \simeq -1$ alors $\frac{LV(\boldsymbol{y})}{Var(\boldsymbol{y})}$ est maximal. Autrement dit, la part de variance locale contenue dans $Var(\boldsymbol{y})$ est alors maximale, ce qui traduit une grande variabilité des données entre les sites voisins. Dans cette situation, nous dirons que $\boldsymbol{y}$ possède une *structure spatiale à fine échelle*.

L'espérance de l'Indice de Moran dans le cas d'aucune corrélation spatiale est :

$$E(I) = \frac{-1}{n-1}$$

Ainsi, les valeurs observées pour I inférieures à $E(I)$ indiquent une autocorrélation spatiale négative et celles supérieures une autocorrélation spatiale positive. Quand le

nombre d'observations n est grand $E(I)$ est proche de 0. Il est bien sûr possible de tester l'absence d'autocorrélation spatiale.

Il est cependant nécessaire d'insister sur l'impact du choix de la définition de la contiguïté dans la valeur de l'indice de Moran. En effet, lorsque l'on teste l'absence d'autocorrélation, on teste une autocorrélation spatiale particulière à partir d'une définition singulière de la contiguïté. Aussi est-il nécessaire d'en tenir compte, lorsqu'on rejette l'hypothèse d'absence d'autocorrélation.

Méthodes *PCNM* (Principal Coordinate analysis of Neighbour Matrices) et *MEM* (Moran's Eigenvectors Map)

PCNM consiste en une *analyse en coordonnées principales* (*PCoA*) d'une matrice de distances tronquées. Elle permet d'obtenir un ensemble de variables spatiales facilement utilisables comme descripteurs spatiaux dans un modèle linéaire. Elle repose initialement sur peu de résultats mathématiques ; les travaux de (Dray, Legendre, & Peres-Neto, 2006) ont permis de lier *PCNM* à une autre méthode (*MEM*) générant des variables qui maximisent l'indice d'autocorrélation spatiale de Moran. Les différentes étapes permettant d'obtenir les variables *PCNM* sont les suivantes :

1. Soit n observations provenant de n sites. Soit $\mathbf{D}$ la matrice des distances euclidiennes entre ces sites calculées à partir des coordonnées géographiques $(x_i; y_i)$ de ces sites :

$$\mathbf{D} = (d_{ii'})_{1\leq i,i'\leq n} \text{ avec } d_{ii'} = \sqrt{(x_i - x_{i'})^2 + (y_i - y_{i'})^2}$$

2. On se fixe un seuil de troncature τ et on définit la matrice de dissemblance $\mathbf{D}^*$ associée à $\mathbf{D}$, composée des distances tronquées *:*

$$\mathbf{D}^* = \left(d^*_{ii'}\right)_{1\leq i,i'\leq n} \text{ avec } d^*_{ii'} = \begin{cases} d_{ii'} \; si \; d_{ii'} \leq \tau \\ 4\tau \; si \; d_{ii'} > \tau \end{cases}$$

   Cette opération permet de mettre en valeur les sites voisins. $\mathbf{D}^*$ est appelée *matrice de voisinages entre les sites*. En pratique, on prendra comme valeur du seuil τ la plus grande distance entre deux sites voisins dans l'*arbre de longueur*

minimale[32] les reliant (voir par exemple (Hartigan, 1975) ou (Bellanger & Tomassone, 2014, pp. 202-203)).

3. *PCoA* de la matrice $\mathbf{D}^*$ (voir par exemple (Legendre & Legendre, 2012, pp. 424-444) ou (Bellanger & Tomassone, 2014, pp. 127-130)). Autrement dit, on diagonalise la matrice $\boldsymbol{\Delta}$ définie par la relation :

$$\boldsymbol{\Delta} = -\frac{1}{2}\left(\mathbf{I} - \frac{1}{n}\mathbf{1}_n(\mathbf{1}_n)^{\mathrm{T}}\right)\mathbf{D}_2^*\left(\mathbf{I} - \frac{1}{n}\mathbf{1}_n(\mathbf{1}_n)^{\mathrm{T}}\right)$$

$$= \frac{(4\tau)^2}{2}\left(\mathbf{I} - \frac{1}{n}\mathbf{1}_n(\mathbf{1}_n)^{\mathrm{T}}\right)\mathbf{S}^*\left(\mathbf{I} - \frac{1}{n}\mathbf{1}_n(\mathbf{1}_n)^{\mathrm{T}}\right)$$

$$\text{Où } \mathbf{D}_2^* = \left({d_{ii'}^*}^2\right)_{1\leq i,i'\leq n} \text{ et } \mathbf{S}^* = \left(1 - \frac{{d_{ii'}^*}^2}{(4\tau)^2}\right)_{1\leq i,i'\leq n}.$$

Suite à cette diagonalisation, on extrait uniquement les valeurs propres λ_k de $\boldsymbol{\Delta}$ strictement positives[33] et les vecteurs propres $\boldsymbol{u}^k$ associés dont la norme est égale à $\sqrt{\lambda_k}$ sont conservées. , où λ_k est la valeur propre de $\boldsymbol{\Delta}$ associée à $\boldsymbol{u}^k$. Le nombre de valeurs propres strictement positives fournit la dimension maximale de l'espace euclidien de représentation. Il est au plus égale à $(n-1)$. Les coordonnées euclidiennes de chaque site i sur chaque axe k sont exactement u_i^k.

En pratique, si le but est d'obtenir une représentation à l'aide de cartes des dissemblances entre sites, le nombre d'axes retenus doit être faible.

4. Les vecteurs propres ainsi calculés sont alors utilisés pour décrire les différentes structures spatiales liées aux données étudiées. Les vecteurs propres associés aux plus fortes valeurs propres seront utiles pour la description de structures spatiales à grande échelle. En revanche, les dernières valeurs propres positives permettront une description de structures spatiales à plus fine échelle. Dans la suite, nous désignerons par variables *PCNM* les coordonnées principales de $\mathbf{D}^*$.

---

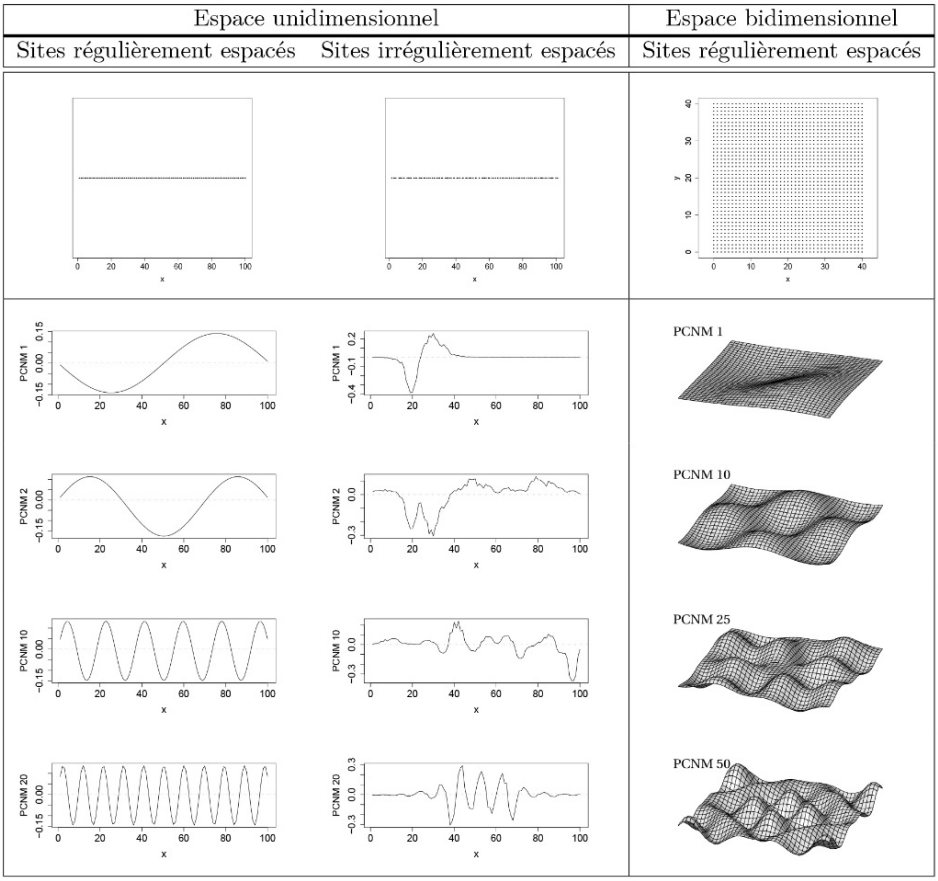
[32] En anglais : Minimum Spanning Tree (*MST*).

[33] Ce choix est dû au fait que l'on cherche à obtenir, à partir d'un tableau des dissemblances entre n sites, une représentation euclidienne de dimension au plus $(n-1)$.

Le Tableau 4, permet d'ores et déjà d'observer quelques propriétés des variables *PCNM* :

- Dans le cas de *sites régulièrement espacés* (*i.e.* on parlera de *grille régulière*), les variables *PCNM* ont une forme sinusoïdale dont la période augmente quand les valeurs propres associées diminuent. Le Tableau 4 donne les représentations de quelques *PCNM* pour des sites régulièrement espacés sur une droite (1ère colonne) et dans un espace à deux dimensions (3ème colonne).
- Dans le cas de sites géographiques irrégulièrement espacés (*i.e.* on parlera de *grille irrégulière*), les *PCNM* obtenues sont beaucoup moins lisses et régulières que précédemment (Tableau 4, 2ème colonne). Néanmoins, les variables *PCNM* obtenues restent classées suivant l'échelle des structures spatiales qu'elles décrivent : plus les valeurs propres associées aux *PCNM* sont faibles, plus les variables associées permettent une description de l'espace à fine échelle.

Tableau 4 - Exemples de variables PCNM pour différentes répartitions des sites d'observations.



PCNM est en fait un cas particulier ; celui dans lequel la matrice de pondération spatiale **W** est exactement la matrice de similarités $\mathbf{S}^*$ issue des distances géographiques tronquées.

La généralisation de cette approche a donné lieu à la méthode *MEM*[34]. Compte-tenu des informations contenues dans **W**, elle permet de définir de nouvelles variables spatiales décrivant l'espace à différentes échelles. Cette méthode consiste à diagonaliser $\mathbf{\Omega}$, la matrice de pondération spatiale **W** après centrage en lignes et en colonnes :

$$\mathbf{\Omega} = \left(\mathbf{I} - \frac{1}{n}\mathbf{1}_n(\mathbf{1}_n)^{\mathrm{T}}\right)\mathbf{W}\left(\mathbf{I} - \frac{1}{n}\mathbf{1}_n(\mathbf{1}_n)^{\mathrm{T}}\right)$$

Les travaux de (de Jong, Sprenger, & van Veen, 1984) ont permis d'interpréter les vecteurs propres obtenus en diagonalisant $\mathbf{\Omega}$. Ils ont ainsi montré que si **W** est symétrique, les vecteurs propres $\boldsymbol{u}^k$ de $\mathbf{\Omega}$, associés aux valeurs propres λ_k, maximisent l'indice d'autocorrélation de Moran projeté sur ces différents axes mutuellement orthogonaux par construction ; mais non normés à $\sqrt{\lambda_k}$. Quand cette méthode de filtrage est employée correctement, le premier vecteur propre $\boldsymbol{u}^1$, supposé centré, est l'ensemble de valeurs qui possède le plus grand indice de Moran $I(\boldsymbol{u}^1)$ que l'on puisse obtenir à partir de la matrice de pondération spatiale **W**. Il vérifie donc le problème classique en algèbre :

$$I(\boldsymbol{u}^1) = \max_{\boldsymbol{u}\in\mathbb{R}^n} I(u) = \max_{\boldsymbol{u}\in\mathbb{R}^n} \frac{\boldsymbol{u}^{\mathrm{T}}\mathbf{W}\boldsymbol{u}}{\boldsymbol{u}^{\mathrm{T}}\boldsymbol{u}} ;$$

Il vaut : $I(\boldsymbol{u}^1) = \frac{n}{\mathbf{1}^{\mathrm{T}}\mathbf{W}\mathbf{1}} \max_k \lambda_k$. Le second vecteur propre $\boldsymbol{u}^2$, orthogonal au premier, représente l'ensemble de valeurs qui possède le second indice de Moran et ainsi de suite, le plus petit indice de Moran étant quant-à-lui donné par $\frac{n}{\mathbf{1}^{\mathrm{T}}\mathbf{W}\mathbf{1}} \min_k \lambda_k$. Dans le cas où **W** n'est pas symétrique, il faut remplacer **W** par $(\mathbf{W} + \mathbf{W}^T)/2$. Ainsi quel que soit le vecteur $\boldsymbol{y}$ étudié sur un espace spatial caractérisé par **W**, on sait que :

$$\frac{n}{\mathbf{1}^{\mathrm{T}}\mathbf{W}\mathbf{1}} \min_k \lambda_k \leq I(\boldsymbol{y}) \leq \frac{n}{\mathbf{1}^{\mathrm{T}}\mathbf{W}\mathbf{1}} \max_k \lambda_k$$

[34] Par la suite, on n'évoquera donc plus que *MEM*.

(Griffith, 2000) a ainsi relié ce travail à l'*ACP* et interprété ces vecteurs propres comme une description, par au plus $(n-1)$[35] variables spatiales associées chacune à une valeur propre non nulle. Ces valeurs propres représentent les différentes autocorrélations latentes contenues dans la matrice de pondération spatiale $\mathbf{W}$. Les $\boldsymbol{u}^k$ associés à des λ_k fortement positives (*resp.* négatives) présentent des autocorrélations positives (*resp.* négatives) et décrivent des structures spatiales globales (*resp.* locales) en ordre décroissant (*resp.* croissant) d'importance. Comme les $\boldsymbol{u}^k$ forment une base orthonormale de $\mathbb{R}^n$, tout vecteur $\boldsymbol{y}_{C_\mathbf{D}}$ peut se décomposer dans cette base sous la forme : $\boldsymbol{y}_{C_\mathbf{D}} = \sum_{i=1}^{n} \boldsymbol{u}^i \left(\left(\boldsymbol{u}^i\right)^{\mathrm{T}} \boldsymbol{y}_{C_\mathbf{D}} \right)$. Les valeurs propres étant linéairement reliées au *I* de Moran, on peut tester H_0: "$I(\boldsymbol{u}^k) = 0$" par permutations pour chaque vecteur propre $\boldsymbol{u}^k$ et ne retenir que ceux qui représentent une autocorrélation significative. Dans la suite, les variables ou composantes *MEM* $\boldsymbol{u}^k$ seront notées $\boldsymbol{MEM}^k$.

Le choix de la matrice $\mathbf{W}$ est une étape cruciale et non sans conséquence sur les résultats des analyses spatiales (Tiefelsdorf, 2000). Dans le cas d'un échantillonnage régulier des sites (*i.e.* grille régulière), les structures définies par les valeurs et vecteurs propres sont proches quel que soit le choix de $\mathbf{W}$. Par contre, dans le cas d'échantillonnages irréguliers, les relations spatiales définissant $\mathbf{W}$ influencent grandement le nombre de valeurs propres positives et négatives ainsi que les structures spatiales détectées.

La matrice de pondération spatiale $\mathbf{W}$ peut, de manière assez générale, être vue comme le produit d'Hadamard d'une matrice de contiguïté binaire $\mathbf{B}$ précisant les sites connectés par une matrice de pondération $\mathbf{A}$ précisant l'intensité des connexions (cf. (Dray, Legendre, & Peres-Neto, 2006)). La matrice de contiguïté $\mathbf{B}$ peut être construite sur la base des distances entre les sites : la méthode la plus simple consiste à sélectionner une distance seuil en dessous de laquelle les sites seront définis comme en connexion ; mais il en existe de nombreuses autres (cf. (Legendre & Legendre, 2012), section 13.3)). La matrice $\mathbf{A}$ est utilisée pour pondérer les connexions définies dans $\mathbf{B}$ et rendre ainsi $\mathbf{W}$ plus réaliste. Il existe de nombreux choix possibles pour $\mathbf{A}$, le plus couramment utilisé est

[35] $(n-1)$ est le nombre maximum de valeurs propres non nulles associés à $\boldsymbol{\Omega}$.

basé sur la ressemblance géographique ($a_{ii'} = 1 - d_{ii'} / \max_i d_{ii'}$; $i, i' = 1, \dots n$). **W** peut donc s'écrire comme le produit matriciel de Hadamard des matrices **A** et **B** :

$$\mathbf{W} = [w_{ii'}] = [a_{ii'} b_{ii'}] \in \mathcal{M}_{n \times n}$$

Analyses multi-échelles et Régression

Détendançage des données

Les méthodes précédemment décrites permettent d'obtenir des variables orthogonales facilement utilisables pour modéliser le comportement spatial de données. Néanmoins, avant d'utiliser ces variables comme régresseurs, il est nécessaire de vérifier que la variable réponse $\boldsymbol{y}$ n'est pas linéairement liée aux coordonnées géographiques des sites. Pour cela, on régresse $\boldsymbol{y}$ sur les coordonnées géographiques des sites et si l'on est en présence d'une telle tendance, on utilisera par la suite le vecteur des résidus $\boldsymbol{r} = \boldsymbol{y} - \hat{\boldsymbol{y}}$ comme variable réponse.

Cette vérification est nécessaire car une tendance linéaire est caractéristique d'un phénomène se produisant à plus large échelle que l'étendue de la surface géographique traitée. Même si les variables *MEM* obtenues sont capables d'identifier ces tendances, elles ne sont pas appropriées pour traiter ce genre de comportement puisque ces variables sont des ondes sinusoïdales. Ainsi, utiliser les variables *MEM* sans détendancer la variable réponse conduira à utiliser trop de ces variables pour identifier la tendance et finalement, les structures spatiales à plus fines échelles seront décrites avec un nombre insuffisant de variables : ce qui dégradera la qualité du modèle.

Régression sur les variables *MEM*

La régression sur variables *MEM* est une alternative aux modèles d'auto-régression simultanée (« Simultaneous AutoRegressive Model » ou « SAR ») ; généralisation extrêmement consommatrice en ressource machine du modèle de régression linéaire, définie pour tenir compte de l'autocorrélation spatiale. En effet, les $(n-1)$ variables *MEM* décrivent la gamme complète de tous les patterns spatiaux possibles mutuellement orthogonaux. Comme en régression sur composantes principales, elles peuvent donc être utilisées comme variables explicatives dans un modèle de régression linéaire pour

expliquer une variable réponse $\boldsymbol{y}$ qui possède une certaine structure spatiale décrite par $\mathbf{W}$. Un des avantages est que tenir compte de l'autocorrélation spatiale au travers de variables *MEM* permet de ne pas avoir à supposer de structure d'autocorrélation particulière pour le terme d'erreurs du modèle (Griffith, 2000). Le vecteur $\boldsymbol{\beta}$ est donc estimé par la méthode des moindres carrés ordinaires. A l'aide d'une procédure de sélection de variables, on ne conserve dans le modèle final que les variables significatives.

Le modèle de décalage spatial dans lequel la dépendance spatiale est portée par la variable réponse peut donc s'écrire très simplement en fonction des variables *MEM* :

$$\boldsymbol{y} \quad = \rho \mathbf{W} \boldsymbol{y} + \boldsymbol{\varepsilon} \approx \rho \boldsymbol{\Psi} \boldsymbol{\Lambda} \boldsymbol{\Psi}^{\mathrm{T}} \boldsymbol{y} + \boldsymbol{\varepsilon} = \boldsymbol{\Psi} \boldsymbol{\beta} \quad + \boldsymbol{\varepsilon}$$

où

- $\boldsymbol{y}$ vecteur réponse centré ;
- $\boldsymbol{\Psi} \in \boldsymbol{\mathcal{M}}_{n\times(n-1)}$ matrice regroupant les variables *MEM*, à sélectionner, elles dépendent de la grille à n sites et du choix de la matrice de pondération spatiale $\mathbf{W} = \mathbf{B}$; matrice de contiguïté construite sur la base des distances entre les sites ;
- $\boldsymbol{\Lambda} \in \boldsymbol{\mathcal{M}}_{(n-1)\times(n-1)}$ matrice diagonale contenant les valeurs propres λ_k associées aux variables $\boldsymbol{MEM}^k \in \mathbb{R}^n$;
- ρ paramètre d'auto-régression ;
- $\boldsymbol{\beta} \quad \in \mathbb{R}^{n-1}$ vecteur des paramètres à estimer ;
- $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}; \sigma^2 \mathbf{I})$ vecteur des aléas supposé multinormal centré de matrice de variance-covariance $\sigma^2 \mathbf{I}$.

Souvent, en plus des coordonnées géographiques des sites d'observations, on dispose d'autres variables explicatives. On possède dans ce cas deux ensembles de variables descriptives :

- des *variables spatiales* formant une matrice $\boldsymbol{\Psi}$. Il s'agit des variables *MEM*, au plus au nombre de $(n-1)$;
- des *variables explicatives* formant une matrice notée $\mathbf{X}$.

En présence de ces deux types de descripteurs, on peut décomposer la variation de $\boldsymbol{y}$ en quatre composantes distinctes : (cf. (Legendre & Legendre, 2012, pp. 529-535))

- une fraction [a] expliquée par les variables explicatives $\mathbf{X}$ mais non structurée spatialement,
- une partie [b] spatialement structurée $\mathbf{\Psi}$ et également expliquée par les variables explicatives $\mathbf{X}$,
- une fraction [c] exclusivement expliquée par les variables spatiales $\mathbf{\Psi}$,
- une portion [d] expliquée par aucun de ces deux types de variables descriptives (variation résiduelle).

Cette décomposition est représentée sur la Figure 9 ci-dessous :

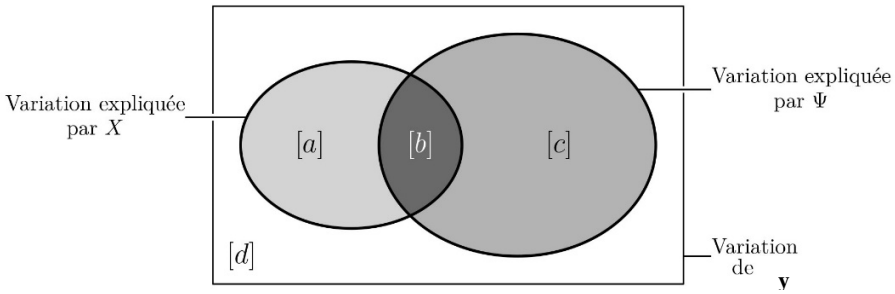


Figure 9 - Partitionnement de la variation de la variable $\boldsymbol{y}$ en présence de deux types de variables explicatives.

On veut modéliser les relations entre $\boldsymbol{y}$ et les variables explicatives $\mathbf{X}$ *;* tout en contrôlant les effets des variables spatiales $\mathbf{\Psi}$. Pour cela, on procède de la manière suivante :

- on régresse chacune des variables explicatives $\boldsymbol{x}^1, \boldsymbol{x}^2, \ldots, \boldsymbol{x}^p$ de la matrice $\mathbf{X}$ sur les variables spatiales $\mathbf{\Psi}$. On note alors $\mathbf{R_{X/\Psi}} = [\boldsymbol{r}^1 \quad \ldots \quad \boldsymbol{r}^p]$ la matrice contenant les vecteurs des résidus issus de chaque régression :

  $$\mathbf{R_{X/\Psi}} = \mathbf{X} - \mathbf{\Psi}(\mathbf{\Psi}^{\mathrm{T}}\mathbf{\Psi})^{-1}\mathbf{\Psi}^{\mathrm{T}}\mathbf{X}$$

  $\mathbf{R_{X/\Psi}}$ représente les parts de variation non structurées spatialement des variables $\boldsymbol{x}^1, \boldsymbol{x}^2, \ldots, \boldsymbol{x}^p$.
- on régresse le vecteur $\boldsymbol{y}$ sur la matrice des résidus $\mathbf{R_{X/\Psi}}$. La fraction [a] correspond au coefficient de détermination ajusté R_a^2 associé, c'est la portion de variation de $\boldsymbol{y}$ expliquée par $\mathbf{X}$ mais sans structure spatiale.

- la fraction $[c]$ est obtenue sur le même principe en régressant les variables spatiales $\boldsymbol{\Psi}$ sur les variables explicatives $\mathbf{X}$ puis en obtenant la matrice des résidus $\mathbf{R}_{\boldsymbol{\Psi}/\mathbf{X}}$ et en régressant le vecteur $\boldsymbol{y}$ sur la matrice des résidus $\mathbf{R}_{\boldsymbol{\Psi}/\mathbf{X}}$.

Les variables spatiales construites à partir de *MEM* ont l'avantage d'être orthogonales donc de ne pas poser de problème de multicolinéarité ; elles ont cependant le gros désavantage d'être nombreuses (nombre de sites -1) et de subir le fléau de la dimension ! L'utilisation en bloc de ces variables spatiales peut rendre le modèle de régression multiple retenu instable. Par conséquent, pour éviter ce problème, il est possible de séparer préalablement les variables spatiales en K groupes $\boldsymbol{\Psi} = [\boldsymbol{\Psi}^1 \vdots \ \ldots \ \vdots \boldsymbol{\Psi}^K]$.

La sélection des variables *MEM* dans un modèle est un point crucial ; (Bini & et al., 2009) ont montré que l'utilisation de différents critères pouvait fortement influencer l'interprétation des effets dans le modèle de régression.

Décomposition en sous-modèles – modélisation multi-échelle

Les variables spatiales *MEM* sont des variables orthogonales entre elles, ordonnées de telle sorte que les premières (*i.e.* associées aux plus fortes valeurs propres) décrivent des structures spatiales à large échelle alors que les dernières sont associées à des phénomènes ayant lieu à plus fine échelle.

Ainsi, le modèle complet peut-être décomposé en K sous-modèles permettant d'expliquer la variabilité de la variable réponse $\boldsymbol{y}$ à plusieurs échelles spatiales. $\boldsymbol{y}$ est alors régressé sur chacun des K groupes $\boldsymbol{\Psi}^j,\ j = 1, \ldots, K$ de *MEM* consécutives correspondant à une échelle spatiale précise. Par exemple, les premières *MEM* regroupées $\boldsymbol{\Psi}^1$ forment un modèle « à large échelle », les suivantes formeront un modèle à échelle un peu plus fine, etc. Le regroupement des variables spatiales en ensembles cohérents décrivant l'espace à des échelles précises distinctes est loin d'être simple puisqu'il n'existe pas de procédure standard permettant un « bon » découpage. Cette procédure dépend essentiellement des phénomènes que l'on veut modéliser. Néanmoins, faute d'information extérieure

satisfaisante, il est possible d'envisager un regroupement de ces variables spatiales en fonction des valeurs propres qui leur sont associées.

Le plus simple étant de séparer les variables *MEM* en 2 groupes regroupant d'un côté les variable *MEM* correspondant à des autocorrélations spatiales positives (*resp.* négatives) et donc à des valeurs propres positives (*resp.* négatives) puis d'analyser séparément les sous-groupes en utilisant par exemple une procédure de sélection de variables pas à pas de type *forward* pour obtenir un sous-modèle ne prenant en compte que les *MEM* significatives pour expliquer $\boldsymbol{y}$ (Blanchet , Legendre, & Borcard, 2008). Mais un découpage plus fin en $K > 2$ groupes formés en fonction des valeurs propres, permet de définir des structures à plusieurs échelles et de déterminer si les variables environnementales dont nous disposons sont responsables ou non de ces structures. Cependant la remarque suivante de (Dray, et al., 2012, p. 269) permet d'appréhender tout l'arbitraire de ces définitions de sous-modèles, aucune méthode consensuelle n'existe à ce jour :

"*In MEM-based methods, the definition of submodels using a potentially large number of synthetic spatial predictors is an arbitrary step and methodological developments are still required to refine an appropriate and rigorous approach to submodel selection*".

Travaux en cours en collaboration avec A. B*rind'Amour,* P. *Legendre et* S. *Mahévas*

Extending the spread of *MEM* using Negative *MEM*

Dans ce travail, nous cherchons à mieux interpréter les variables $\boldsymbol{MEM^k}$ dîtes négatives (i.e. telles que $\lambda_k < 0$) tout en développant une procédure de sélection des variables *MEM* à inclure dans un modèle. Nous étudierons ces questions à la fois grâce à une étude de simulation et sur des jeux de données réelles (effort de pêche et données écologiques).

Dans toutes les études utilisant les variable *MEM* comme variables explicatives d'une variable réponse mesurée spatialement, la première préoccupation est la sélection des variables *MEM* à introduire dans le modèle (Dray, Legendre, & Peres-Neto, 2006). En effet,

le nombre de *MEM* construites est lié au nombre de sites ; les considérer toutes dans le modèle créé donc un surajustement. La sélection des *MEM* significatives a donc souvent été effectuée à l'aide d'une procédure de sélection de type *forward* ((Borcard & Legendre, 2002), (Peres-Neto & Legendre, 2010)) qui a pour défaut majeure de retenir plus de variables que nécessaire (Platt & Denman, 1975). (Jombart, Devillard, Dufour, & Pontier, 2008) ont proposé de séparer les variables *MEM* en deux groupes (positives /négatives) ou davantage et d'effectuer une sélection de variables séparée sur chacun de ces sous-ensembles. Cependant, le nombre de variables explicatives conservées dans chaque modèle peut encore être grand.

Pour éviter ce problème, une approche alternative simple peut baser la sélection à sur le résultat d'un test de nullité du coefficient de corrélation linéaire dit de « Bravais-Pearson » entre chacune des $\tilde{n}$ variables *MEM* construites[36] et la variable réponse $\boldsymbol{y}$: H_0^i:$\rho(\mathbf{y}\,;\,\boldsymbol{MEM}^i)=0; i = 1, \dots, \tilde{n}$. Pour pouvoir effectuer ces comparaisons multiples, nous appliquons une correction de Bonferroni-Holm (Holm, 1979). Cette procédure se divise en plusieurs étapes, permettant ainsi un contrôle séquentiel de la probabilité de n'avoir aucun faux positif :

1. ordonner les *p*-values correspondant aux hypothèses nulles associées $H_0^{(1)}, \dots, H_0^{(\tilde{n})}$ par ordre croissant : $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(\tilde{n})}$
2. examiner, pour un seuil de significativité α fixé, les hypothèses dans l'ordre correspondant :
   - si $p_{(1)} > \frac{\alpha}{\tilde{n}}$ alors on ne peut rejeter aucune hypothèse nulle ;
   - sinon, on cherche le plus petit $k \in \{2, \dots, \tilde{n}\}$ tel que $p_{(k)} > \frac{\alpha}{\tilde{n}+1-k}$ et alors les hypothèses nulles $H_0^{(1)}, \dots, H_0^{(k-1)}$ sont rejetées tandis que toutes les autres $H_0^{(i)}$ $(i = k, \dots, \tilde{n})$ le sont pas.

Nous comparons cette approche à celle consistant à diviser l'ensemble des *MEM* en sous-groupes et à appliquer sur chacun une procédure de sélection des *MEM* de type pas à pas.

[36] D'après ce qui précède, $\tilde{n}$ est au plus égal à $n - 1$.

L'étude de simulation est basée sur un échantillonnage en 2D de n=100 sites situés sur une grille régulière 10×10. Nous cherchons à expliquer une variable réponse $\boldsymbol{y}$ à l'aide d'un modèle de régression linéaire multiple dont les variables explicatives sont les variables *MEM* calculées sur la grille régulière 10×10. Pour cela, nous étudions différents scénarios à l'aide de 8 variables aléatoires réponse $\{\boldsymbol{y}^c\}_{c=1,\dots,8}$ caractérisées par des structures d'autocorrélation bien précises, puis nous itérons cette procédure de simulation un grand nombre de fois (100, voire 500 fois). Le modèle complet s'écrit donc :

$$\boldsymbol{y}^c = \boldsymbol{\Psi}\boldsymbol{\beta}^c + \boldsymbol{\varepsilon}$$

où

- $c = 1, \cdots, 8$ cas simulés;
- $\boldsymbol{y}^c \in \mathbb{R}^{100}$ vecteur réponse dans le cas c (voir Tableau 5) ;
- $\boldsymbol{\Psi} = [\boldsymbol{\psi^1} \vdots \ \dots \ \vdots \boldsymbol{\psi^4}] \in \boldsymbol{\mathcal{M}}_{100\times\tilde{n}}$ matrice regroupant les variables *MEM*, à sélectionner, issues d'une grille régulière 2D à 100 sites et du choix suivant pour la matrice de pondération spatiale $\mathbf{W} = [w_{ii'}] = [a_{ii'}, b_{ii'}]$

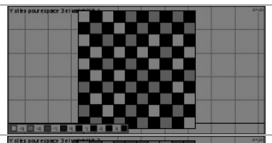
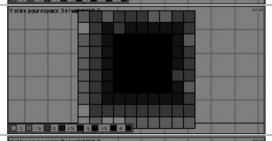
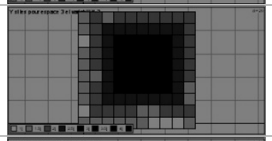
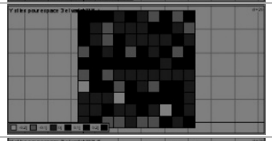
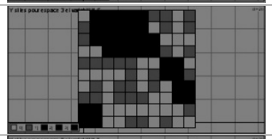
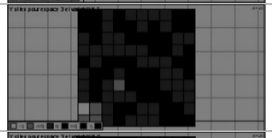
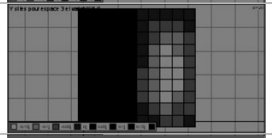
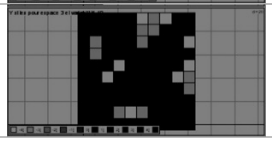
  la matrices de pondération $\mathbf{A} = [a_{ii'}] = \left[1 - d_{ii'}/\max_i d_{ii'}\right]$; $i, i' = 1, \dots 100$ et

  la matrice de contiguïté **B** est construite sur la base des distances entre les sites :

  tous les sites i, i' tels que $d_{ii'} \leq \sqrt{2}$ sont considérés comme en connexion.
- $\boldsymbol{\beta}^c \in \mathbb{R}^{\tilde{n}}$ vecteur des paramètres à estimer ;
- $\boldsymbol{\varepsilon} \sim N_{100}(\mathbf{0}; \sigma^2\mathbf{I})$ vecteur des aléas supposé multinormal centré de matrice de variance-covariance $\sigma^2\mathbf{I}$.

A l'aide de cette étude de simulation, nous voulons *i)* déterminer les structures d'autocorrélation imposées sur $\boldsymbol{y}$ qui rendent les *MEM* négatives (*resp.* postives) significatives dans ce modèle et *ii)* comparer les performances du test de corrélation proposé, à celles obtenues après un découpage en 4 sous-modèles de régression (large échelle $\boldsymbol{MEM^1}$ à $\boldsymbol{MEM^{22}}$, intermédiaire positive $\boldsymbol{MEM^{23}}$ à $\boldsymbol{MEM^{44}}$, intermédiaire négative $\boldsymbol{MEM^{45}}$ à $\boldsymbol{MEM^{66}}$, fine échelle $\boldsymbol{MEM^{67}}$ à $\boldsymbol{MEM^{\tilde{n}=89}}$) (Munoz, 2009) auxquels on applique une procédure de sélection de variables pas à pas de type *forward,* pour identifier les variables *MEM* significatives. Le Tableau 5 ci-dessous résume les 8 scénarios que nous avons choisi de simuler et les résultats auxquels nous nous attendons.

Tableau 5 - Simulated $\boldsymbol{y}$ with associated pattern and Moran indice.

$\boldsymbol{y}^c$	**Definition**	**Simulated pattern**	$I(\boldsymbol{y}^c)$	$\frac{GV(\boldsymbol{y})}{Var(\boldsymbol{y})}$	$\frac{LV(\boldsymbol{y})}{Var(\boldsymbol{y})}$	**Expected correlated *MEM* with $\boldsymbol{y}^c$** **Splitting *MEM* in 4 groups**
$\boldsymbol{y}^1$	Bernoulli (1,-1) + $N(0; 0.1^2)$		-1	-0.50	1.49	Negative *MEM* within the fourth group
$\boldsymbol{y}^2$	Linear decreasing function from the center to the bounds + $N(0; 0.1^2)$		0.83	0.41	0.52	Positive *MEM* within the first group
$\boldsymbol{y}^3$	$\boldsymbol{y}^1 + \boldsymbol{y}^2$		0.81	0.41	0.59	Positive and Negative *MEM* within the first and fourth groups
$\boldsymbol{y}^4$	$N(0; 0.1^2)$		-0.08	-0.04	1.04	None
$\boldsymbol{y}^5$	Uniform patch (value sampled in a uniform distribution) + $N(0; 0.1^2)$		0.54	0.27	0.73	Positive *MEM* within the first and second groups
$\boldsymbol{y}^6$	Patch increasing linearly on the diagonal+ $N(0; 0.1^2)$		0.48	0.26	0.74	Positive *MEM* within the second group
$\boldsymbol{y}^7$	$\boldsymbol{MEM}^1 + N(0; 0.1)$		1	0.47	0.53	Only $\boldsymbol{MEM}^1$
$\boldsymbol{y}^8$	Random patchy repulsive value within an uniform area		0.03	0.02	0.98	Positive *MEM* within the second group

Les variables $\boldsymbol{y}^4$ et $\boldsymbol{y}^7$ sont des cas de référence puisque nous connaissons par avance les variables explicatives qui doivent être sélectionnées : aucune dans le cas du bruit blanc $\boldsymbol{y}^4$ et uniquement $\boldsymbol{MEM}^1$ pour $\boldsymbol{y}^7$. L'approche développée dans les simulations sera ensuite appliquée à deux cas d'étude *i)* les temps de pêche en mer celtique sur une période que plus récente que dans (Mahévas, Bellanger, & Trenkel, 2008) *i.e.* 2001-2007 et *ii)* les

données issues d'une étude sur la distribution spatiale des oribates[37] dans les mousses (Borcard, Legendre, & Drapeau, 1992).

Un article est en préparation.

Extending the spread of *MEM*: The case of irregular sampling design.

Les écosystèmes et communautés marines sont de bons exemples de systèmes complexes. Ils sont composés d'une multitude d'unités interagissant à différentes échelles spatiales et temporelles. La caractérisation de ces échelles est une étape essentielle dans la compréhension et éventuellement la prévision des effets de changements dans les processus gouvernant ces systèmes. Ceci nécessite l'utilisation de méthodes statistiques permettant de caractériser quantitativement les patrons spatiaux, temporels et les interactions spatio-temporelles ; robustes aux changements et aux différences d'échantillonnage des observations disponibles. Ces changements ou différences sont notamment liés à la régularité des distances entre les sites d'échantillonnage et à la couverture complète ou non (*i.e.* sous-échantillonnage) de la zone d'étude.

Dans ce second travail, nous nous proposons d'étudier le comportement et la sensibilité de la méthode multivariée *MEM* pour différents scénarios de plans d'échantillonnage (zones complète, réduite, irrégulière) via l'utilisation de simulations numériques. Ces différents cas sont répétés 100 fois chacun. Nous tentons de répondre aux questions suivantes :

- L'approche *MEM* est-elle pertinente dans le cas de plans d'échantillonnage irréguliers ?
- Est-ce que les variables *MEM* calculées à partir d'un échantillonnage incomplet capturent les patterns spatiaux qu'elles seraient supposées capturées si l'échantillonnage était complet ?
- Y a-t-il un seuil d'irrégularité au-delà duquel cette méthode est à proscrire ?

[37] Groupe d'acariens caractérisé par la présence d'une carapace qui recouvre le corps. Les oribates sont minuscules (< 1mm) et vivent dans le sol, la litière, la matière organique et les tapis de mousses et de lichens où ils peuvent atteindre de très grandes densités de population.

Dans une première étape, nous nous concentrons sur une sélection de quatre scénarios de simulation caractérisant des plans d'échantillonnage fréquemment rencontrés dans les études écologiques. Ils sont présentés dans le Tableau 6 ci-dessous. Trois éléments nous ont guidé : *i)* l'irrégularité des sites d'échantillonnage induite par le sous-échantillonnage de la zone d'échantillonnage (*i.e.* échantillonnage aléatoire), *ii)* l'irrégularité des sites d'échantillonnage générée par la couverture partielle de la zone d'échantillonnage (*i.e.* blocs de sites non échantillonnés), et *iii)* l'adéquation entre processus d'observation et processus écologique étudié. Nous testons donc la capacité de la méthode *MEM* à détecter correctement les structures spatiales « prédéfinies » dans les scénarios. L'étude de simulation est basée sur un échantillonnage en 2D de sites (n=400 sites situés sur un carré 20×20), sauf pour le scénario 4 où 53% de la grille 20×20 est supprimée pour créer des blocs de sites manquants. Dans les trois premiers scénarios, 5 variables *MEM* (notées $\boldsymbol{MEM}_{scale}^{(j)}$; j=1,…,5)[38], sur les 399, sont alors sélectionnées et vont être étudiées. Elles représentent un gradient des différentes structures spatiales (de la plus large à la plus fine).

Pour chacun des quatre scénarios étudiés (*A)*, *B)*, *C)* et *D)* du Tableau 6), nous appliquons deux calculs possibles des variables *MEM*:

- une méthode dite *complète* pour laquelle les variables *MEM* sont calculées sur la grille régulière 20×20 puis les coordonnées de sites non échantillonnés sont supprimés (notées MEM_{comp}) ;
- une méthode dite *réduite* pour laquelle les variables *MEM* sont directement calculées à partir des sites échantillonnés sur la grille régulière 20×20 (notées MEM_{red}).

[38] Pour les trois premiers scénarios : les *MEM* 1, 10, 150, 200 et 350 sont choisies alors que pour le scénario 4 le nombre de sites étant plus faible 4 *MEM* sont étudiées : les numéros 10, 60, 110 et 150.

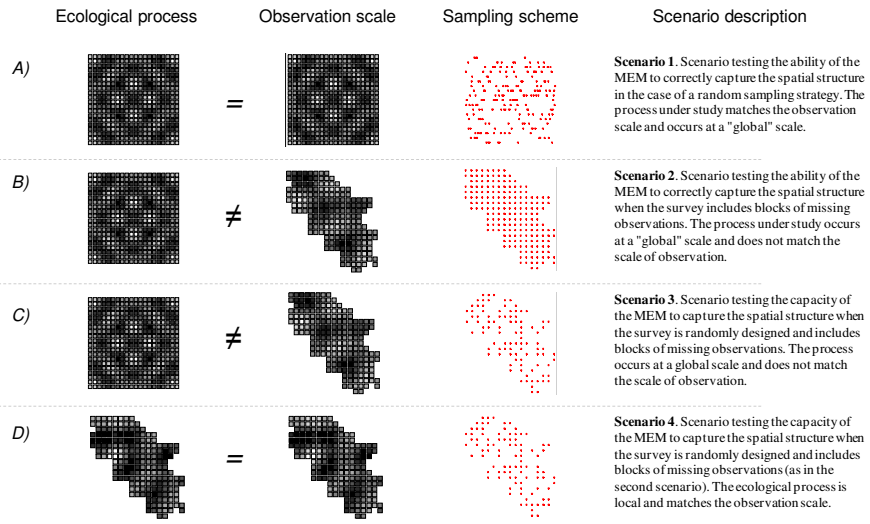
Voici une courte description des 5 scénarios testés, présentés dans le Tableau 6 ci-après:

1. Le scénario *Random sampling design* (voir Tableau 6 ligne *A)*) teste la capacité des variables *MEM* à saisir correctement la structure spatiale dans le cas d'une stratégie d'échantillonnage aléatoire évaluée à 7 seuils différents de couverture de la zone étudiée (10%, 25%, 50%, 60%, 80%, 90%, 100%). L'irrégularité est aléatoire et crée des distances inégales entre les sites échantillonnés. L'échelle du processus écologique concorde avec celle d'observation, c'est une échelle « globale ». Les MEM_{scale} sont calculées à partir des coordonnées des sites régulièrement espacés sur la grille régulière 20×20. Pour chaque seuil de couverture retenu, les MEM_{comp} sont obtenues en supprimant les sites manquants des MEM_{scale} alors que les MEM_{red} sont directement calculées à partir des seuls sites échantillonnés.
2. Le scénario *Blocks of missing data* (voir Tableau 6 ligne *B)*) teste la capacité des variables *MEM* à capturer précisément la structure spatiale lorsque le recueil de données comporte des blocs d'observations manquantes, blocs correspondant par exemple à des zones inaccessibles. Les sites échantillonnés couvrent 53% de la zone du processus écologique et sont régulièrement espacés ; mais comme la zone d'observation est réduite, les distances entre sites deviennent irrégulières. Le processus écologique se produit à une échelle «globale» qui ne correspond pas à l'échelle d'observation. Les MEM_{scale} sont calculées à partir des coordonnées des sites régulièrement espacés sur la grille régulière 20×20. Les MEM_{comp} sont obtenues en ne conservant que les coordonnées dans les MEM_{scale} des sites de la zone d'observation (*i.e.* 53% de la zone totale) alors que les MEM_{red} sont directement calculées à partir des sites de la zone d'observation.
3. Le scénario *Random sampling design and blocks of missing data on a global structure* (voir Tableau 6 ligne *C)*) teste la capacité des variables *MEM* à capturer la structure spatiale lorsque le recueil de données est conçu de manière aléatoire et comprend des blocs d'observations manquantes (comme dans le second scénario). La zone d'observation couvre 53% de la zone du processus écologique ; la stratégie d'échantillonnage aléatoire est évaluée à 7 seuils différents de couverture de la zone d'observation (10%, 25%, 50%, 60%, 80%, 90%, 100%). L'échelle du processus

écologique ne concorde pas avec celle d'observation plus petite. Le calcul des variables MEM est similaire au scénario 1. Les MEM_{scale} sont calculées à partir des coordonnées des sites régulièrement espacés sur la grille régulière 20×20. Les MEM_{comp} sont obtenues en ne conservant que les coordonnées dans les MEM_{scale} des sites de la zone d'observation (*i.e.* 53% de la zone totale) au seuil de couverture fixé. Les MEM_{red} sont directement calculées à partir des seuls sites échantillonnés.

4. Le scénario *Random sampling design and blocks of missing data on a local structure* (voir Tableau 6 ligne *D)*) teste la capacité des variables *MEM* à capturer la structure spatiale lorsque le recueil de données est conçu de manière aléatoire et comprend des blocs d'observations manquantes (comme dans le second scénario). Il est similaire au troisième scénario, mais dans ce cas, l'échelle du processus d'observation est identique à celle du processus écologique, *i.e.* plutôt locale. La stratégie d'échantillonnage aléatoire est évaluée à 7 seuils différents de couverture de la zone étudiée (10%, 25%, 50%, 60%, 80%, 90%, 100%). Dans ce scénario, le processus écologique se déroule à une échelle différente des 3 précédents. Les MEM_{scale} sont calculées à partir des coordonnées des sites régulièrement espacés sur la grille après suppression des blocs de sites manquants (*i.e.* 53% de la zone totale 20×20). Le même processus de calcul que pour le scénario 3 est appliqué pour obtenir les MEM_{comp} et les MEM_{red}.

Tableau 6 - Scénarios simulés.



La capacité des variables MEM_{comp} (*resp.* MEM_{red}) à expliquer les variations de chacune des cinq structures spatiales prédéfinies $MEM_{scale}^{(j)}$ est évaluée à l'aide des modèles de régression linéaire suivants :

$$\boldsymbol{MEM}_{scale}^{(j)} = \boldsymbol{\Psi}_{comp}\boldsymbol{\beta}_{comp}^{(j)} + \boldsymbol{\varepsilon}^{(j)} \text{ et } \boldsymbol{MEM}_{scale}^{(j)} = \boldsymbol{\Psi}_{red}\boldsymbol{\beta}_{red}^{(j)} + \boldsymbol{\varepsilon}^{(j)};\ j = 1,2,..,5^{39}$$

où

- $\boldsymbol{MEM}_{scale}^{(j)} \in \mathbb{R}^{\tilde{n}}$ vecteur réponse comportant les coordonnées des $\tilde{n} \leq n^{40}$ sites échantillonnés sur la $j^{\text{ème}}$ variable *MEM* sélectionnée issue de la grille régulière 2D à n sites associé au processus écologique ;
- $\boldsymbol{\Psi}_{comp}$ (*resp.* $\boldsymbol{\Psi}_{red}$) matrice regroupant les variables MEM_{comp} (*resp.* MEM_{red}),
- $\boldsymbol{\beta}_{comp}$ (*resp.* $\boldsymbol{\beta}_{red}$) vecteur des paramètres à estimer ;
- $\boldsymbol{\varepsilon} \sim N_{\tilde{n}}(\mathbf{0}; \sigma^2\mathbf{I})$ vecteur des aléas supposé multinormal centré de matrice de variance-covariance $\sigma^2\mathbf{I}$.

Le comportement des *MEM* pour les différents scénarios est ensuite évalué à l'aide des critères suivants :

- le nombre de variables *MEM* sélectionnées dans le modèle de régression linéaire ;
- la concordance entre la variable *MEM* à décrire et celle(s) servant de variable(s) explicative(s) ;
- le coefficient de détermination ajusté R_{aj}^2 et enfin
- la muticolinéarité entre variables *MEM* (uniquement pour les MEM_{comp}).

La meilleure approche à retenir sera bien sûr celle correspondant au modèle de régression dans lequel les variables explicatives *MEM* sont peu nombreuses et cohérentes avec celle que l'on cherche à expliquer ; la qualité d'ajustement est la meilleure possible et enfin la multicolinéarité entre variables explicatives *MEM* est la plus faible possible. Nos résultats montrent que, les variables *MEM* peuvent être utilisées dans un contexte d'échantillonnage irrégulier ; mais avec précaution. En effet, dès que moins de 75% des

---

[39] $j = 1,2,..,4$ pour le scénario 4.

[40] $n = 400$ si la grille est régulière 20×20 ; moins dans le dernier scénario. $\tilde{n}$, le nombre de sites échantillonnés, varie de 10, 25, 50, 60, 70, 80, 90 à 100% du nombre total de sites n.

sites sont préservés, les variables *MEM* deviennent sensibles à l'irrégularité de la grille ; cette variabilité s'accroît encore lorsque l'on modélise des structures spatiales à échelles fines. Sous de fortes conditions d'irrégularité, l'utilisation de *MEM* construites à partir (des coordonnées) des sites irréguliers (*i.e.* MEM_{red}) peut même aboutir à la création de structures spatiales inexistantes sur la grille initiale.

Nous proposons; afin de pallier ces problèmes, de reconstruire une grille régulière à partir des sites disponibles échantillonnés de manière irrégulière (Figure 10) :

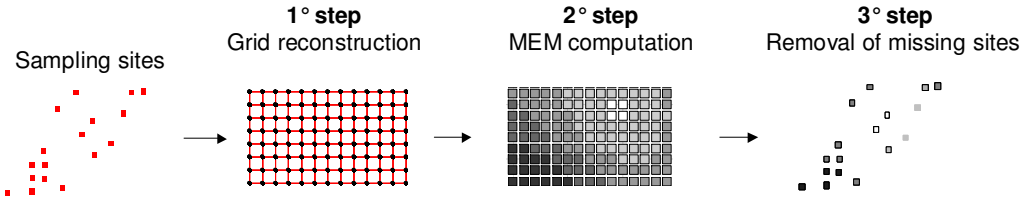


Figure 10 - Reconstruction d'une grille régulière : une solution au problème de variabilité des MEM dans le cas d'un échantillonnage irrégulier.

Cette solution est aussi testée par simulations et évaluée à l'aide du nombre de *MEM* significatives obtenues à l'aide du test de nullité du coefficient de corrélation linéaire entre chacune des variables *MEM* et la variable réponse avec une correction de Bonferroni-Holm ; mais aussi à l'aide du coefficient de détermination ajusté. Une ANOVA non paramétrique a permis de s'assurer qu'elle était adaptée à tout type d'échelle (fine ou large).

Enfin, il nous reste à appliquer cette approche à deux cas d'étude très différents (temps de pêche en mer celtique sur la période 2001-2007 et l'abondance de poissons juvéniles dans les nourriceries côtières de la baie de Vilaine (sud Bretagne) sur la période 2008-2010.

Un article est en cours d'écriture.

1.2.3 Autre perspective : exploration des données du GIS VALPENA (depuis 2014)

Collaborateurs sur ce thème :

- Comités régionaux des pêches (Pays de la Loire, Bretagne, Nord-Pas-de-Calais-Picardie) ;
- IGARUN[41] (UMR 6554 LETG[42]-Géolittomer UMR CNRS 6554Nantes) : B. Trouillet et E. Plissonneau (Stagiaire M2 professionnel Ingénierie Mathématique, Nantes).

La spatialisation des activités de pêche est devenue un enjeu mondial. Jusqu'à très récemment, il n'existait pas de base de données fiable permettant tant la défense des intérêts des professionnels que de servir d'outil de gestion et d'aide à la prise de décisions interne à la profession. La création d'un état des lieux « pêche » robuste et régulier était indispensable. A cet effet, le projet VALPENA (éVALuation des activités de PEche au regard des Nouvelles Activités)[43] a vu le jour en 2010 et est devenu un Groupement d'Intérêt Scientifique (GIS) en 2014. Il est conçu comme une plateforme collaborative structurant des observatoires régionaux des pratiques de pêche dans une optique d'aménagement de l'espace maritime. Aujourd'hui, il réunit les Comité Régional des Pêches Maritimes et des Élevages Marins (CRPMEM) du Nord Pas de Calais Picardie, de Basse- Normandie, de Haute-Normandie, de Bretagne et des Pays de la Loire ; soit 5 sur les 10 à l'échelle nationale. D'autres sont susceptibles de rejoindre cette plateforme. Il a pour objectif de combler un manque d'informations sur les pratiques de pêche, exacerbé par un accroissement des besoins de connaissance sur les pratiques spatiales en lien avec l'essor de la demande d'espace en mer particulièrement fort ces dernières années. En effet, l'intensification des activités existantes ; mais aussi la mise en place de nouveaux projets tels que l'éolien offshore ou l'aquaculture, obligent les pêcheurs à modifier leurs pratiques. Ces activités créent une pression de plus en plus importante sur la pêche maritime professionnelle et sont à l'origine de conflits de plus en plus nombreux. VALPENA permet donc des réflexions mutilpartenariales et interdisciplinaires sur ces sujets très sensibles.

---

41 Acronyme de *Institut de Géographie et d'Aménagement Régional de l'Université de Nantes*.

42 Acronyme de *Littoral, Environnement, Télédétection, Géomatique*.

43 Voir par exemple http://www.msh.univ-nantes.fr/94461849/0/fiche___article/&RH=1326210943749 ou http://www.lemarin.fr/articles/detail/items/valpena-cartographier-la-peche-pour-mieux-la-defendre.html.

Ces CRPMEM collectent, sur la base d'entretiens, des données permettant de décrire finement l'activité de pêche, tant spatialement que temporellement. Ainsi, les données à traiter concernent aujourd'hui près de 2700 navires, c'est-à-dire l'ensemble des flottilles situées de la frontière belge jusqu'au sud de la Vendée. Avec une clé qui est le navire, les bases de données VALPENA combinent :

- des données d'enquête localisant, à l'échelle mensuelle, les activités de pêche en fonction d'un référentiel spatial constitué d'un maillage géométrique de 3 milles nautiques de côté environ, en distinguant les engins de pêche mise en œuvre et les espèces ciblées. L'unité obtenue pour chaque navire est donc « maille x mois x engin utilisé x espèce ciblée » qu'il est possible d'agréger et de ré agréger en fonction des besoins de l'analyse ;
- des données techniques et administratives : port d'attache, caractéristiques techniques (longueur, puissance, etc.), etc.

Ce système permet aujourd'hui de fournir des données capitales tant pour que les pêcheurs puissent exprimer, voire défendre leurs intérêts en lien avec le développement de nouveaux usages en mer (énergies marines renouvelables, parcs marins, extraction de granulats, etc.), qu'à des fins de recherche. Ces données sont totalement inédites et, compte tenu de leur caractère sensible et stratégique, sont confidentielles. Elles permettent de compléter, en les précisant, d'autres données détenues par le Ministère en charge de la pêche. Le laboratoire LETG-Géolittomer continue de développer des outils informatiques facilitant le recueil de ces données (interface de saisie de données d'enquêtes,etc.) et leur exploitation cartographique. Cependant, la masse de données disponibles ainsi que la qualité hétérogène des données recueillies nécessitent un *travail exploratoire*, au niveau *spatial* ; mais aussi au niveau de l'*échantillonnage* à adopter pour éviter dans les années à venir d'enquêter tous les navires tous les ans. Une collaboration a vu le jour très récemment; impliquant dans un premier temps Elodie Plissonneau, une stagiaire étudiante du Master 2 professionnel Ingénierie Mathématique de Nantes embauchée en CDD[44] de 6 mois en septembre 2014, ainsi que ma participation au conseil scientifique pluridisciplinaire (géographie, économie, droit, statistiques…) de VALPENA.

[44] Acronyme de *Contrat à Durée Déterminée*.

Chapitre 2 : Exploration de données médicales (depuis 2008)

2.1 LA PHARMACO-EPIDEMIOLOGIE (2008 - 2014)

Collaborateurs sur ce thème :

- CEIP[45] (CHU, Nantes) et Université de Nantes (EA 4275 - SPHERE) : P. Jolliet, C. Victorri-Vigneau ;
- Université de Nantes (EA 4275 - SPHERE) : F. Feuillet, J.-B. Hardouin, V. Sébille.

Initialement limités aux troubles les plus sévères, l'utilisation des médicaments psychotropes a rapidement été élargie à des troubles de moindre gravité et leur usage s'est peu à peu banalisé. Comparativement aux autres pays européens, la consommation de médicaments psychotropes est plus élevée en France. Les principales classes de médicaments consommés sont, en premier lieu les anxiolytiques (utilisés contre l'anxiété) et les hypnotiques (somnifères), en second lieu, les antidépresseurs. Les co-prescriptions sont aussi fréquentes, notamment les associations anxiolytiques-hypnotiques et anxiolytiques-antidépresseurs.

Les situations cliniques liées à la surconsommation médicamenteuses sont diverses : il peut s'agir d'une part d'une surconsommation dans le cadre d'une prescription pour des troubles associés à des critères de gravité ou pour lesquels il existe une inefficacité ou une résistance au traitement, d'autre part d'une surconsommation volontaire du sujet dans le cadre d'un usage compulsif du médicament. Or l'ensemble des travaux menés à ce jour ne permettent cependant pas de caractériser finement :

- les médicaments psychotropes pouvant conduire à des abus ou des situations de mésusage ;
- le profil des consommateurs de ces médicaments, notamment les sujets présentant une surconsommation.

Il est donc nécessaire de disposer d'informations quantitatives détaillées sur les doses et co-prescriptions au sein d'un échantillon de consommateurs, afin d'identifier pour un médicament donné un niveau de consommation au-delà duquel le sujet aura une

[45] Acronyme de *Centre d'Evaluation et d'Information sur la Pharmacodépendance*.

consommation extrême et aussi de déterminer des groupes homogènes de consommateurs. Ces informations sont indispensables pour évaluer l'impact de mesures visant à modifier l'utilisation des médicaments psychotropes dans notre pays.

L'accès aux données des Caisses Primaires d'Assurance Maladie[46] contribue, à travers la réalisation d'études pharmaco-épidémiologiques, à améliorer le niveau de connaissance sur l'usage des médicaments psychotropes en France. Même si ces données reflètent la délivrance par la pharmacie et non la consommation réelle ; elles permettent tout de même d'évaluer le potentiel d'abus et de dépendance des médicaments.

2.1.1 Surconsommation médicamenteuse : médicaments psychotropes à risque

Dans ce travail, nous avons développé des outils d'identification et d'évaluation de la surconsommation médicamenteuse à partir de données issues des bases de données de la Caisse Régionale d'Assurance Maladie (CRAM) des Pays de La Loire. A partir de ces bases de données, le CEIP[47] de Nantes a conçu un indicateur de surconsommation, appelé facteur F, permettant de rendre compte, pour tout médicament psychotrope, de l'importance du phénomène. Ce facteur F est un critère maintenant couramment utilisé pour étudier le dépassement de la posologie maximale recommandée. Il se définit comme le rapport entre la posologie moyenne quotidienne du patient et la dose maximale recommandée par le RCP[48]. Ainsi, pour un patient donné, si F est inférieur ou égal à 1, tout va bien, le médicament est bien utilisé ; par contre si F est supérieur à 1 alors on peut suspecter un problème de pharmacodépendance, de mésusage de prescription, d'abus ou de détournement, c'est-à-dire d'utilisation du médicament pour une fin autre que le but médical ou psychiatrique visé. À peu près tous les médicaments d'ordonnance peuvent être employés à des fins autres que celles prévues, mais les cas d'abus concernent habituellement les médicaments ayant des propriétés psychotropes.

---

[46] En abrégé : CPAM.

[47] Les Centres d'Évaluation et d'Information sur la Pharmacodépendance (CEIP) sont implantés au sein de Centres Hospitalo-Universitaires (CHU) et sont spécialisés en pharmacologie clinique ou expérimentale, en toxicologie analytique ou en épidémiologie. Ils sont chargés d'évaluer le potentiel de dépendance et d'abus des substances psychotropes, notamment les médicaments.

[48] Acronyme de *Résumé des Caractéristiques du Produit*. Les RCP sont publiés par l'Agence Nationale de Sécurité du Médicament et des produits de santé (ANSM, ex-AFSSAPS). Tous les produits pharmaceutiques ayant obtenu une autorisation de mise sur le marché (AMM) ont un RCP et une notice, disponibles sur le site internet de l'ANSM.

La valeur seuil de 1 peut pour certains médicaments être remise en cause ; c'est le cas de certains médicaments antidouleurs, calmants, pour lesquels la prescription peut être régulièrement supérieure à la posologie recommandée ; mais aussi de médicaments pouvant entrainer des phénomènes d'accoutumance ou de dépendance. Nous avons donc tenté dans ce travail de répondre aux questions suivantes :

- Comment déterminer le seuil u à partir duquel, si F$> u$, le patient peut être qualifié de surconsommant ?
- Comment discriminer au mieux la population des patients en 2 groupes (surconsommants/normaux), en fonction de variables explicatives ?
- Comment prévoir l'appartenance à l'un ou l'autre des groupes pour un patient donné ?

L'approche statistique que nous avons adoptée est une combinaison de plusieurs méthodes:

1. La *théorie des valeurs extrêmes* au travers du modèle *Peak Over Threshold (POT)* pour déterminer, pour la substance étudiée, un seuil u, au-delà duquel un patient peut être considéré comme surconsommant (comportement extrême). De plus amples détails ont déjà été donnés sur ce modèle au paragraphe 1.1.1.2 du chapitre 1. Une fois le seuil u fixé, et les paramètres de la distribution de Pareto généralisée (GPD) estimés, nous nous sommes assurés que l'hypothèse distributionnelle effectuée sur les tailles de dépassement, sachant qu'un dépassement du seuil u avait eu lieu, était respectée à l'aide d'un P-P ou d'un Q-Q plot.

   Le seuil u retenu grâce au modèle POT permet de scinder la population étudiée en deux groupes : les surconsommants avec un facteur de F$> u$ et les autres ayant un facteur F $\leq u$.
2. La *régression logistique* pour obtenir le profil des surconsommants en fonction des variables explicatives disponibles. Nous avons choisi la régression logistique avec fonction de lien *logit* à cause de sa flexibilité et de l'interprétation simple des paramètres en terme d'*odds ratio*.
3. La *courbe ROC* et les indices associés (sensibilité et spécificité) pour étudier la capacité du modèle de régression retenu à bien discriminer la population étudiée.

C'est la première fois qu'une telle approche était mise en œuvre dans ce domaine.

Les données provenaient de la base de données régionale de l'assurance maladie des Pays de La Loire. La population sélectionnée comprenait l'ensemble des patients ayant eu au moins une délivrance du médicament étudié entre le 1er juillet 2005 et le 31 décembre 2005. La consommation de psychotropes en France est la plus importante de l'Union Européenne. Parmi les médicaments psychotropes, certains sont susceptibles d'un détournement d'usage de par leurs propriétés psychoactives. Nous avons choisi d'étudier deux d'entre eux :

- la tianeptine (Stablon®) : un antidépresseur (ATD) avec des cas d'abus et de dépendance clairement identifiés. 7263 patients ont eu au moins deux délivrances de tianeptine entre le 1er juillet 2005 et le 31 décembre 2005 dans la région des Pays de La Loire ;
- le zolpidem (Stilnox®) : un hypnotique pour lequel des phénomènes de tolérance et de dépendance physique avec syndromes de sevrage existent et donc pour lequel les indications doivent être soigneusement pesées ; notamment chez les patients ayant des antécédents psychiatriques. 33584 patients ont eu au moins deux délivrances de zolpidem entre le 1er juillet 2005 et le 31 décembre 2005 dans la région des Pays de La Loire.

Les variables disponibles comprenaient celles relatives :

- au médicament étudié: facteur **F** (rapport entre la posologie moyenne quotidienne du patient et la dose maximale recommandée);
- au patient: sexe (SEX) et âge (AGE) ;
- à la prescription : nombre de prescripteurs (NOMAD.M); catégorie du prescripteur (CATPRES) ;
- relatives à la délivrance : nombre de pharmacies (NOMAD.P); rapport entre le nombre de délivrances observé et celui théorique lié à l'ordonnance (RAPDELB) ;
- aux traitements associés : nombre total; par classe; par indication (CONS.AS, N.BZO.T, N.BZO.1, N.BZO.2, N.BZO.3, N.BZO.4, N.BZO.5, NBIN.ATD, NBIN.NL, MORPH.2CAT).

Le Tableau 7 ci-dessous résume les caractéristiques générales des patients ayant eu au moins deux délivrances de tianeptine (*resp.* zolpidem) entre le 1er juillet 2005 et le 31 décembre 2005 dans la région des Pays de La Loire. Comme attendu pour ce type de médicaments, notre échantillon comportait une majorité de femmes et les âges médians étaient de plus de 60 ans pour les deux médicaments. Le nombre maximum de médicaments psychotropes (**CONS.AS**) délivrés pendant la période d'étude était important pour les deux médicaments (14 et 16 pour la tianeptine de zolpidem); ce qui pouvait suggérer que certains patients étaient plus gravement malades parmi les consommateurs de ces médicaments. Cette hypothèse est renforcée par le fait qu'environ 20% des patients avaient également consultés un psychiatre (**CATPRES**). De manière similaire, le facteur F atteignait des valeurs maximales très élevées pour les deux médicaments (environ 11 pour tianeptine et 30 pour zolpidem). Ces niveaux extrêmes sont souvent caractéristiques de situations d'abus et/ou de de pharmacodépendance.

Tableau 7 - Caractéristiques des patients pour tianeptine et zolpidem.

Variable (*name of variable*)	**Tianeptine** N = 7263	**Zolpidem** N = 33584
Age, years (**AGE**)	66; 29 [15 – 102]	63; 24 [6 – 106]
Gender: male (0) vs Female (1) (**SEX**)	2301 / 4962	10344 / 23240
More than 3 prescribing doctors (**NOMAD.M**): Yes (1) vs. no (0)	32 (0.4%)	300 (0.9%)
Call for psychiatrist (**CATPRES**): 1 yes vs. 0 no	1529 (21.1%)	6098 (18.2%)
More than 3 pharmacies who delivered the studied drug (**NOMAD.P**): Yes vs. no	78 (1.1%)	438 (1.3%)
Nb of psychotropic drugs delivered during the same period (**CONS.AS**)	1; 1 [0 – 14]	1; 2 [0 – 16]
- Nb of BZD drugs (**N.BZO.T**)	1; 1 [0 – 10]	1; 1 [0 – 10]
- More than one anxiolytic BZD drug (**N.BZO.1**)	4121 (56.7%)	14578 (43.4%)
- More than one other anxiolytic drug (**N.BZO.2**)	412 (5.7%)	930 (2.8%)
- More than one hypnotic BZD drug (**N.BZO.3**)	1074 (14.8%)	3565 (10.6%)
- More than one other hypnotic drug (**N.BZO.4**)	2183 (30.1%)	2846 (8.5%)
- Nb of patients with Rivotril drug (**N.BZO.5**)	354 (4.9%)	1596 (4.8%)
- More than one other ATD drugs (**NBIN.ATD**)	1518 (20.9%)	13359 (39.8%)
- More than one neuroleptic drugs (**NBIN.NL**)	1043 (14.4%)	3388 (10.1%)
- At least one morphine drugs (**MORPH.2CAT**)	126 (1.7%)	839 (2.5%)
R ratio (nb of dispensations/28 days)> 1 (**RAPDELB**) :Yes vs. no	2041 (28.1%)	7120 (21.2%)
Consumption factor (**F**)	[0.06 – 10.95]	[0.03 – 30.07]

Notes: BZD: benzodiazepine, ATD: antidepressant, *R* ratio: number of dispensations/28 days.
Data summaries are median; interquartile range (IQR) [minimum–maximum] for continuous variables or numbers of patients (percentages) for categorical variables.

Nous avons ensuite utilisé, pour chaque médicament, le modèle *POT* de manière à partitionner l'échantillon en deux groupes en fonction du seuil u retenu : les surconsommants avec un facteur de F> u et les autres ayant un facteur F ≤ u. Les seuils obtenus correspondent à une réalité clinique et sont conformes à la littérature (Tableau 8).

Tableau 8 - Estimations par maximum de vraisemblance des paramètres de la GPD pour tianeptine et zolpidem.

	Tianeptine	**Zolpidem**
F Threshold	**1.1**	**2.0**
Nexc (N)	524 (7263)	318 (33 584)
nllh	-339.135	176.665
Maximum Likelihood Parameter Estimates		
Shape ξ (SE)	0.307 (0.052)	0.607 (0.092)
Scale β (SE)	0.142 (0.009)	0.350 (0.036)

Notes: nexc = the number of data points above the threshold,
nllh = the negative logarithm of the likelihood evaluated at the maximum likelihood estimates, SE = standard error.

Pour tianeptine, le seuil de 1.1 correspond à une utilisation normale du médicament dans son indication : pas de dépassement du seuil fixé dans le résumé des caractéristiques du produit. Sur la période d'étude, les patients dont le facteur F dépassaient 1.1 étaient rares (7%) et présentaient :

- un comportement addictif associé à des consommations très importantes, ou
- des critères de gravité de leur pathologie dépressive nécessitant une augmentation de la posologie à finalité thérapeutique (associée ou non à l'utilisation d'autres médicaments antidépresseurs), décidée par le prescripteur.

Ces observations ont ensuite été confortées par les résultats de la régression logistique (Tableau 9).

Pour zolpidem (Tableau 8 et Tableau 9), le seuil de 2.0 nous a tout d'abord questionné puisque les recommandations précisent que la dose maximale utilisée ne doit pas excéder 1 comprimé par jour (*i.e.* un facteur F égal à 1). Ce résultat trouve cependant toute sa justification dans l'analyse fine des propriétés de ce médicament. Tout d'abord, zolpidem est un hypnotique se caractérisant par une demi-vie courte : les patients peuvent reprendre un comprimé au cours de la nuit. Il a de plus été montré qu'il existait des phénomènes de tolérance et de dépendance physique avec syndrome de sevrage : des sujets augmentent le nombre de comprimés pris le soir afin de ressentir l'effet hypnotique. Ces effets sont connus, si bien qu'en pratique l'utilisation de deux comprimés par jour est trop souvent

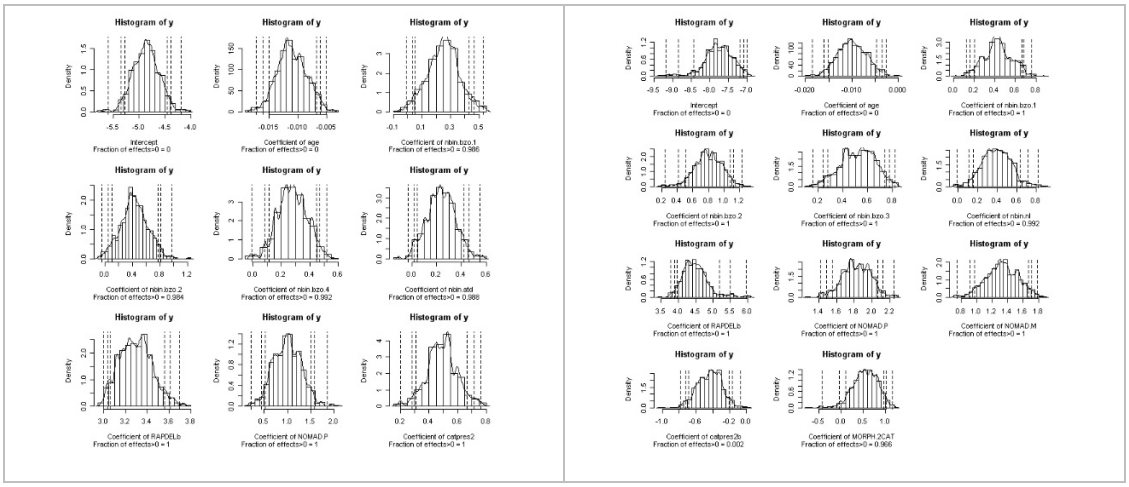
acceptée. Enfin, cette diminution d'efficacité peut expliquer l'utilisation d'autres hypnotiques simultanément. Dans le corpus de données étudié, les patients dont le facteur F dépassait 2 étaient cependant très rares (1%) et correspondaient à des sujets ayant des antécédents psychiatriques et/ou addicts (à la recherche « d'effets psychiques positifs ») utilisant des doses extrêmement élevées et ayant recours à des comportements frauduleux pour se procurer cette quantité.

Tableau 9 - Multivariate logistic regression analysis of over consumption risk for tianeptine and zolpidem; results of stepwise selection procedure. P-value<5%.

	Tianeptine	**Zolpidem**
F Threshold	**1.1**	**2.0**
Logistic Variable	OR [95% CI]	
Age, 10 years increase	0.90 [0.85 – 0.95]	0.91 [0.84 – 0.98]
NOMAD.M		3.733 [2.451– 5.688]
CATPRES	**1.63 [1.32 – 2.01]**	**0.64 [0.48– 0.85]**
NOMAD.P	2.69 [1.42 – 5.05]	6.18 [4.34 – 8.79]
N.BZO.1	1.28 [1.03 –1.59]	1.55 [1.17 – 2.03]
N.BZO.2	1.55 [1.08 – 2.21]	2.22 [1.50 – 3.27]
N.BZO.3		1.68 [1.28 – 2.21]
N.BZO.4	1.33 [1.08 – 1.63]	
NBIN.ATD	1.26 [1.01 – 1.56]	
NBIN.NL		1.49 [1.11 – 1.99]
MORPH.2CAT		1.78 [1.06 – 2.97]
RAPDELB	26.81 [19.97 – 35.98]	84.66 [43.34 – 165.33]
ROC area	0.87	0.93

Afin de valider le modèle logistique retenu et d'analyser la sensibilité des estimations des paramètres à de faibles variations dans les données, nous avons utilisé la méthode *bootstrap non paramétrique* (Tableau 10). Nous avons ainsi pu vérifier que les densités estimées étaient proches de gaussiennes. Les résultats obtenus en utilisant la méthode du maximum de vraisemblance dans le Tableau 9 étaient donc robustes et la relation détectée susceptible d'être vraie.

Tableau 10 - Bootstrap Distribution for logistic Regression Coefficients for tianeptine (left) and zolpidem (right).



D'après le Tableau 9, pour les deux médicaments étudiés, les facteurs augmentant le risque d'être classé « surconsommant » sont l'âge (AGE) plutôt jeune, le nomadisme pharmaceutique (NOMAD.P) important, la consommation associé au traitement d'au moins un anxiolytique (N.BZO.1) ou un hypnotique (N.BZO.2) et enfin le nombre élevé de délivrances (RAPDELB). Par exemple, le nomadisme pharmaceutique augmente considérablement le risque d'être surconsommant. Ce risque est multiplié par 2.69 pour tianeptine (*resp.* par 6.18 pour zolpidem). A l'inverse, le fait de faire appel à un psychiatre (CATPRES) a un effet opposé sur le risque de surconsommation selon le médicament considéré : pour tianeptine, le risque est multiplié par 1.63 alors que pour zolpidem il est divisé par 1.56. Les autres variables explicatives sélectionnées ont un effet significatif sur l'un des deux médicaments ; mais pas les deux simultanément : la prise d'au moins un autre médicament hypnotique (N.BZO.4) ou un antidépresseur (NBIN.ATD) est uniquement associé au risque de surconsommation de la tianeptine alors que le nomadisme médical (NOMAD.M), la prise d'au moins un benzodiazépine hypnotique (N.BZO.3), un neuroleptique (NBIN.NL)ou de la morphine (MORPH.2CAT) sont associés au risque de surconsommation de zolpidem.

La qualité des résultats obtenus a été analysée à l'aide des outils classiques : courbe ROC et son aire (AUC) et table de confusion. Dans les deux cas, la capacité du modèle logistique à discriminer les deux groupes de patients déterminés à l'aide du modèle POT est excellente (Figure 11 et Tableau 11) avec une AUC de 0.87 pour tianeptine et 0.93 pour zolpidem.

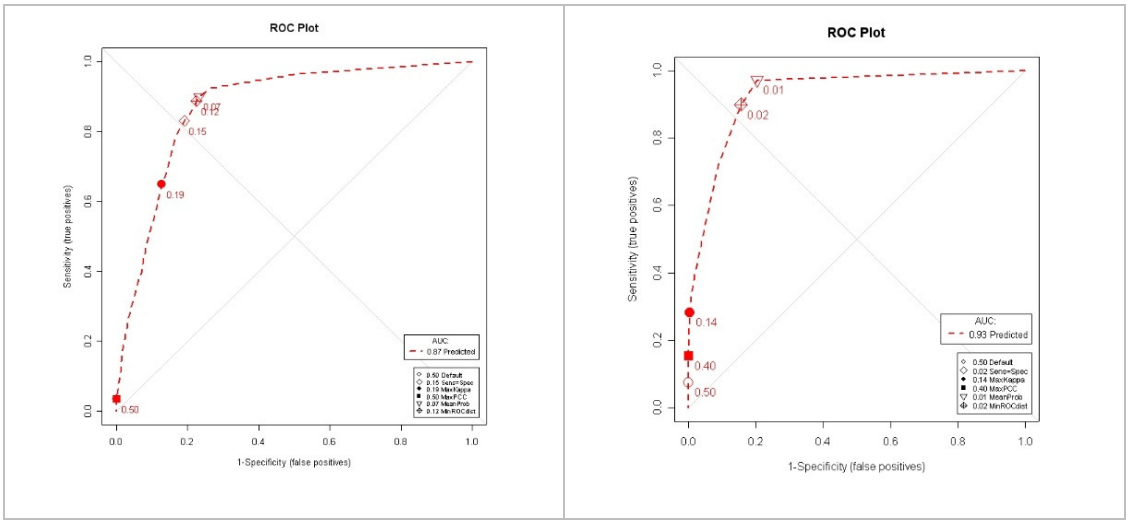


Figure 11 - ROC curve for tianeptine (left), zolpidem (right).

Tableau 11 - Classification Table Based on the Logistic Regression Model in Tableau 9 using a Cutpoint of 0.15 (sens=spec) for tianeptine (top) (resp. 0.02 for zolpidem (bottom)).

Classified	***Observed***		**Total**
	1	**0**	
1	435	1287	1722
0	89	5432	5521
Total	524	6719	7243

Sensitivity=435/524=83%; Specificity=5432/6719=80.8%;

PCC=Percent Correctly Classified=81.1%

Classified	***Observed***		**Total**
	1	**0**	
1	286	5183	5469
0	32	28083	28115
Total	318	33266	33584

Sensitivity=286/318=89.9%; Specificity=5432/6719=84.4%;

PCC=Percent Correctly Classified=84.5%.

Dans ce travail, pour deux médicaments particuliers, tianeptine et zolpidem, nous avons :

- déterminé un seuil u au-delà duquel le comportement des patients peut être considéré comme extrême ;
- étudié ce comportement extrême des patients ;
- construit un modèle permettant de classer et prédire le comportement d'un nouveau patient en fonction des variables explicatives introduites dans la régression logistique.

La procédure statistique mise en œuvre est tout à fait originale dans le domaine de la pharmaco-épidémiologie et elle est généralisable à n'importe quel médicament. Elle peut permettre de détecter les médicaments pour lesquels la dose maximale recommandée par le RCP est régulièrement dépassée ; ainsi que l'ampleur de ce dépassement. Cette procédure peut être une aide à la détection de cas d'abus. Elle peut aussi permettre de réévaluer la dose maximale en vigueur. Ce travail a donné lieu à un article publié en 2013 (Bellanger, Vigneau, Pivette, Jolliet, & Sébille, 2013) dont sont issus les tableaux 9 à 11 ainsi que la figure 10. Il a été présenté lors de deux conférences.

2.1.2 Surconsommation médicamenteuses : profils des consommateurs

Comme nous l'avons vu dans le paragraphe précédent, le facteur F est un indicateur de surconsommation permettant de rendre compte, pour tout médicament psychotrope, de l'importance du phénomène. Grâce à cet indicateur nous avons pu quantifier l'importance du phénomène de surconsommation ; mais aussi différencier les sujets surconsommants, des sujets non surconsommants (cf. § 2.1.1). Cependant, ce seul facteur de surconsommation n'est pas suffisant pour discriminer et caractériser les différents profils de consommations de médicaments psychotropes, puisqu'il limite les recherches à deux groupes, surconsommants *versus* non surconsommants. Or ces groupes peuvent recouvrir des profils de consommateurs différents. De plus, cet indicateur ne prend en compte pour constituer les groupes qu'un seul aspect relatif à la consommation ; en laissant de côté des informations importantes, telles que le nomadisme.

Nous avons, dans ce travail, comparé les partitions obtenues sur des données pharmaco-épidémiologiques, à l'aide de deux méthodes de classification : la *CAH*[49] sur les premières composantes d'une *AFCM*[50] et *l'Analyse en Classes Latentes*[51] (en abrégé par la suite *ACL*), introduite par Lazarsfeld vers 1950, qui permet d'identifier des sous-groupes d'individus (des classes latentes) à partir de données qualitatives. Les critères utilisés pour comparer ces méthodes comprenaient : le nombre de classes, la concordance entre partitions, l'interprétation des groupes construits et la stabilité dans le temps. Pour cette étude, notre intérêt s'est porté sur les caractéristiques de consommation du bromazépam (Lexomil®), l'un des psychotropes les plus prescrits et consommés en France. Nous avons choisi ce médicament car un certain nombre d'études ont montré l'existence d'un usage détourné, comme pour le zolpidem. Le corpus de données comprenait l'ensemble des assurés enregistrés dans les bases de données des CPAM de la région Pays de la Loire, âgés de 18 ans et plus et ayant eu au moins deux délivrances de bromazépam à deux dates différentes, au cours du premier semestre 2008 et du premier semestre 2009. Les deux fichiers de données étaient composés de 40644 assurés pour 2008 et 44756 pour 2009.

Les partitions obtenues à l'aide de la *CAH* et l'*ACL* ont été construites à partir de 6 variables dichotomiques. Des seuils permettant de transformer les variables d'intérêt en variables binaires ont donc été préalablement établis en collaboration avec les collègues pharmacologues de l'EA 4275 (Wainstein, et al., 2011). Les variables retenues permettent de caractériser un comportement de consommation : facteur F, nomadisme médical et pharmaceutique, spécialité du prescripteur, adéquation de la prescription avec les recommandations liées et non liées à la classe thérapeutique du bromazépam. Le sexe a été considéré comme une variable illustrative. Le Tableau 12 ci-après indique le codage des variables binaires retenues :

---

[49] Acronyme de *Classification Ascendante Hiérarchique.*

[50] Acronyme de *Analyse Factorielle des Correspondances Multiples*.

[51] Encore appelée dans la littérature *modèle en classes latentes*.

Tableau 12 - Codage des variables binaires.

Variable	Codage	
Sexe	0	Homme
	1	Femme
Surconsommant (facteur F > 1)	0	Non surconsommant
	1	Surconsommant
Nomadisme médical	0	Non nomade
	1	Nomade
Nomadisme pharmaceutique	0	Non nomade
	1	Nomade
Spécialité du prescripteur	0	Spécialiste
	1	Généraliste
Recommandations liées	0	Conformité
	1	Non-conformité
Recommandations non liées	0	Conformité
	1	Non-conformité

Les individus ont été regroupés en « profils » en fonction des combinaisons de valeurs possibles prises par les variables binaires. Sur les 64 possibilités de profils réponse (2^6) ; seuls 53 ont été observés dans les données du premier semestre 2008 (*resp.* 58 pour le premier semestre 2009) auxquels nous avons associé un poids correspondant au nombre d'individus présentant chaque profil. Pour comparer ces deux périodes, seuls les profils présents dans les bases de données des deux années ont été conservés ; soit 53 profils d'individus. Le Tableau 13 décrit le profil des consommateurs de Bromazépam pour les deux années 2008 et 2009. En 2009, les consommateurs étaient majoritairement des consommatrices (environ 74% de femmes) et l'âge moyen était de 62 ans. Plus de quatre fois sur cinq, la prescription était faite par un médecin généraliste. La proportion de surconsommants était faible (1,2% des utilisateurs avait un facteur F supérieur à 1), tout comme celle liée au nomadisme médical (0,4%) et pharmaceutique (1,3%). Près de 40% des prescriptions n'étaient pas en conformité avec les recommandations liées à la classe thérapeutique du bromazépam et 6% n'étaient pas en conformité avec les recommandations non liées au bromazépam. Les proportions observées en 2008 sont très similaires. La prévalence des comportements de consommation est donc très stable entre les deux années étudiées.

Tableau 13 - Description des caractéristiques et des comportements de consommation chez les usagers de bromazépam en 2008 et 2009.

	2008 (*n*=40644)	2009 (*n* =44756)
Age (mean ± std)	62.0 ± 15.2	62.4 ± 15.2
Sexe : Femme	73.7%	73.6%
Surconsommant : 1	1.1%	1.2%
Nomadisme médical : 1	0.4%	0.4%
Nomadisme pharmaceutique : 1	1.2%	1.3%
Spécialité du prescripteur : 1	85.9%	86.0%
Recommandations liées : 1	39.2%	38.1%
Recommandations non liées : 1	7.2%	6.1%

Ce travail a fait suite au stage de Master 2 de Fanny Feuillet (Feuillet, 2009), réalisé au sein de l'équipe d'accueil 4275 de Nantes, à partir des bases de données de la CRAM des Pays de la Loire. F. Feuillet avait alors mis en évidence l'intérêt de l'*ACL* pour caractériser la consommation de médicaments psychotropes (Wainstein, et al., 2011). La pharmacodépendance peut en effet être considérée comme une variable latente, inobservable directement ; mais dont les effets se traduisent par des comportements de consommation spécifiques et identifiables. L'*ACL* avait ainsi permis d'obtenir une partition des consommateurs de médicaments psychotropes présentant des comportements de consommation similaires. Nous allons tout d'abord décrire succinctement le principe de l'*ACL* dans le cas dichotomique : on en trouvera un exposé plus détaillé dans l'ouvrage de (Droesbeke, Lejeune, & Saporta, 2005, pp. 71-82) ainsi que dans (Everitt, 1984) (Bartholomew & Knott, 1999) et (Hagenaars & McCutcheon, 2002).

Le modèle en classes latentes ou *ACL* fait partie de la famille des modèles de mélange fini ancrés dans un cadre probabiliste qui permet de tester la qualité d'ajustement du modèle et de calculer des mesures de qualité d'ajustement. Le corpus étant décrit par des variables qualitatives (ou catégorielles), un modèle log-linéaire suffisamment contraint pour le rendre identifiable est retenu pour modéliser chaque composante du mélange. L'hypothèse fondamentale la plus souvent faite est que conditionnellement à l'appartenance à une classe, les variables qualitatives sont indépendantes *i.e.* les associations observées entre les variables sont dues au fait que les individus appartiennent à des classes latentes différentes. On obtient alors l'*ACL* en considérant des lois

multinomiales multivariées avec indépendance conditionnellement aux composantes du mélange. L'hypothèse d'indépendance conditionnelle[52] est cependant souvent peu réaliste car il est difficile de justifier l'intérêt de rechercher des composantes ayant cette propriété outre par la simplicité de mise en œuvre : des extensions peuvent permettent d'en tenir compte comme nous le verrons plus loin. L'*ACL* est considérée comme l'équivalent de l'Analyse Factorielle (cf. par exemple (Bellanger & Tomassone, 2014, pp. 159-183)) dans le cas où les variables observées sont qualitatives. En tant que cas particulier des modèles de mélange de distributions, l'*ACL* peut également apporter une réponse à un problème de classification supervisée ou non (Biernacki, 2009) d'où l'intérêt de comparer les partitions obtenues avec l'*ACL* et la *CAH*.

En se limitant au cas des variables dichotomiques, qui sera le nôtre par la suite, supposons que les données sont constituées de n individus et p variables dichotomiques $\mathrm{X}^1, \mathrm{X}^2, \dots, \mathrm{X}^p$. Soient Y la variable latente à C classes latentes, $\boldsymbol{x}_i = \begin{pmatrix} x_i^1 & \dots & x_i^p \end{pmatrix}^{\mathrm{T}} \in \{0;1\}^p$ le vecteur de réponses binaires de l'individu i $(i = 1, \dots, n)$; $p_{jm} = P\left[\mathrm{X}^j = 1 \,\middle|\, \mathrm{Y} = m\right]$ la probabilité que $\mathrm{X}^j = 1$ pour un individu de la classe latente m et $\pi_m = P[\mathrm{Y} = m]$ la probabilité *a priori* d'appartenir à la classe latente m ($0 < \pi_m < 1$ et $\sum_{m=1}^{C} \pi_m = 1$). L'*ACL*, en terme de modèle de mélange, revient à supposer que $\boldsymbol{x}_i$ est issu de manière indépendante d'un mélange de lois de Bernoulli multivariées $\mathcal{B}(p_{1m}, \dots, p_{pm})$ d'où l'expression de la probabilité marginale de réponse :

$$P[(\mathrm{X}^1, \mathrm{X}^2, \dots, \mathrm{X}^p) = \boldsymbol{x}_i] = \sum_{m=1}^{C} \pi_m \prod_{j=1}^{p} (p_{jm})^{x_i^j} (1 - p_{jm})^{1 - x_i^j} = \sum_{m=1}^{C} \pi_m \, p_m(\boldsymbol{x}_i)$$

Dans un modèle de mélange, π_m est appelée proportion du mélange et $p_m(\boldsymbol{x}_i)$, expression de la loi de chaque classe latente m, est appelée composante du mélange.

[52] Encore appelée dans la littérature *hypothèse d'indépendance locale*.

On déduit alors la probabilité *a posteriori*, qu'un individu i $(i = 1, \dots, n)$ de vecteur réponse $\boldsymbol{x}_i$ appartienne à la classe latente m. Cette probabilité permet d'affecter tout individu à la classe m la plus probable i.e. celle maximisant :

$$P[Y = m \mid (\mathrm{X}^1, \mathrm{X}^2, \dots, \mathrm{X}^p) = \boldsymbol{x}_i] = \frac{\pi_m \prod_{j=1}^{p} (p_{jm})^{x_i^j} (1 - p_{jm})^{1 - x_i^j}}{\sum_{m=1}^{C} \pi_m \prod_{j=1}^{p} (p_{jm})^{x_i^j} (1 - p_{jm})^{1 - x_i^j}}$$

Pour un n-échantillon donné, la méthode du maximum de vraisemblance est alors utilisée pour estimer p_{jm} et π_m , puis en déduire une estimation de $P[Y = m \mid (\mathrm{X}^1, \mathrm{X}^2, \dots, \mathrm{X}^p) = \boldsymbol{x}]$. La log-vraisemblance observée de l'échantillon s'écrit :

$$l(\boldsymbol{\theta}; \boldsymbol{x}_1, \dots, \boldsymbol{x}_n) = \sum_{i=1}^{n} ln \left(\sum_{m=1}^{C} \pi_m \prod_{j=1}^{p} (p_{jm})^{x_i^j} (1 - p_{jm})^{1 - x_i^j} \right)$$

où $\boldsymbol{\theta} = \left\{ \left((p_{1m}, \dots, p_{pm}); \pi_m \right)_{m=1,\dots,C} \right\}$ représente l'ensemble des paramètres inconnus à estimer à partir du n-échantillon.

Cependant, l'optimisation directe sur $\boldsymbol{\theta}$ étant difficile, la maximisation de l s'effectue à l'aide de l'algorithme EM (*Espérance-Maximisation*) qui conduit à de bons résultats en pratique ; toutefois la procédure peut parfois converger vers des extrema locaux. Pour éviter cette convergence vers des maxima locaux dans l'algorithme *EM*, chaque modèle a été testé avec 100 valeurs initiales différentes aléatoirement choisies.

Pour comparer et choisir parmi plusieurs modèles, le critère *AIC*[53] ou le critère *BIC*[54] sont utilisés. Le « meilleur » modèle est celui minimisant l'un de ces deux critères. C'est un compromis entre qualité de l'ajustement et parcimonie (nombre de paramètres à estimer). Dans notre cas, nous avons choisi de minimiser le critère *BIC*, critère conseillé par (Nylund, Asparouhov, & Muthén, 2007). Le nombre de classes latentes a été déterminée en testant successivement l'ajustement des modèles emboîtés à 1, 2,…,

[53] Acronyme de *An Information Criterion*.

[54] Acronyme de *Bayesian Information Criteron*.

jusqu'au plus grand nombre plausible de classes latentes. Tenant compte de l'avis d'experts pharmacologues, nous avons fixé ce nombre maximum à 8.

Pour identifier les problèmes de dépendance locale entre variables au sein des classes latentes, nous avons utilisé un indice diagnostic appelé *bivariate residual* (*BVR*) qui correspond à la statistique du test du Chi-deux D^2 divisée par le degré de liberté associé.

$$BVR = \frac{D^2}{ddl} \text{ avec } D^2 = \sum_{(j,j')} \frac{(O_{jj'} - np_{jj'})^2}{np_{jj'}} \text{ où}$$

- $O_{jj'}$ représentent les effectifs observés dans la table de contingence croisant les items des variables *j* et *j'*;
- $np_{jj'}$ représentent les effectifs théoriques correspondants, prédits par le modèle, sous l'hypothèse d'indépendance locale :

$$p_{jj'} = \sum_{m=1}^{C} \pi_m (p_{jm})^{x^j} (1 - p_{jm})^{1-x^j} (p_{j'm})^{x^{j'}} (1 - p_{j'm})^{1-x^{j'}}$$

Lorsque que *BVR* est supérieurs à 1, on considère que les deux variables concernées ne sont pas indépendantes dans chaque classe latente. Une extension du modèle de classes latentes classique permettant de prendre en compte cette dépendance locale, consiste à modifier le modèle en incluant une nouvelle variable appelée *effet direct* qui combine et remplace les deux variables observées dépendantes (X^1 et X^2) en une seule à quatre modalités (X^{12}) comme indiqué dans le Tableau 14 ci-après.

Tableau 14 - Principe pour la construction d'un effet direct.

Nouvelle variable (effet direct)	Variables observées	
X^{12}	X^1	X^2
1	0	0
2	0	1
3	1	0
4	1	1

Dans notre cas, le modèle ne sera plus alors basé uniquement sur des variables dichotomiques ; on y inclura aussi des effets directs (variables à 4 modalités puisque les variables initiales sont dichotomiques) qui remplaceront les variables initiales détectées comme liées. Pour plus détails sur le sujet, nous renvoyons à l'ouvrage de (Skrondal &

Rabe-Hesketh, 2004) ainsi qu'au site internet[55] de John Uebersax évoquant de manière pratique la détection et les solutions pour prendre en compte le problème de la dépendance locale en *LCA* à l'aide du logiciel Latent Gold.

Une fois le modèle *ACL* validé, la dernière étape consiste à affecter chaque individu i à la classe latente la plus probable à l'aide de l'estimation de la probabilité *a posteriori* que cet individu appartienne à chacune des classes latentes :

$$\hat{P}[Y = m|(\mathrm{X}^1, \mathrm{X}^2, \dots, \mathrm{X}^p) = \boldsymbol{x}_i] = \frac{\hat{\pi}_m \prod_{j=1}^{p} (\hat{p}_{jm})^{x_i^j} (1 - \hat{p}_{jm})^{1-x_i^j}}{\sum_{m=1}^{C} \hat{\pi}_m \prod_{j=1}^{p} (\hat{p}_{jm})^{x_i^j} (1 - \hat{p}_{jm})^{1-x_i^j}}; m = 1, \dots, C$$

Les calculs ont été réalisés à l'aide du logiciel Latent Gold 4.5[56], spécialisé dans ce type de modèles à variables latentes.

Le but de notre travail étant de comparer, sur un corpus de données volumineux issu des bases de données de la CRAM des Pays de la Loire, la partition obtenue à l'aide de l'*ACL* à celle obtenues avec une méthode de classification non supervisée de type euclidien. Cependant, les données étudiées étant de type qualitatif, nous avons tout d'abord effectué une *AFCM* (pondérée), sur les composantes de laquelle une *CAH avec critère d'agrégation de Ward* a ensuite été appliquée. Comme l'*ACP,* l'*AFCM* peut être utilisée pour synthétiser l'information et réduire la dimension du problème. Nous avons décidé de reconstruire partiellement notre corpus de données à l'aide des premières composantes factorielles représentant au moins 60 % de l'inertie totale, en considérant donc que la part d'inertie négligée était un bruit similaire au terme aléatoire d'un modèle de reconstitution d'une matrice de données. Une *CAH* a alors été appliquée sur les composantes retenues. Le nombre optimal de classes a ensuite été déterminé visuellement à partir du dendrogramme en analysant les indices d'agrégation ; mais aussi à l'aide du graphique des silhouettes (Bellanger & Tomassone, 2014, pp. 194-195). L'*AFCM* et la *CAH* ont été réalisées avec le logiciel SAS 9.2. Enfin, l'homogénéité de chaque partition *ACL* et *CAH* a été évaluée à l'aide de l'inertie intra-classe.

---

55 http://www.john-uebersax.com/stat/condep.htm

56 Le logiciel Latent Gold est distribué par Statistical Innovations: http://www.statisticalinnovations.com/products/latentgold_v4.html

Nous disposions donc de partitions, construites sur le même jeu de données, à deux périodes différentes (1er semestre 2008 et 1er semestre 2009) à partir de deux méthodes de classification différentes (*ACL* et *CAH*). Nous voulions répondre aux questions suivantes :

- Pour une période donnée, les partitions obtenues à l'aide de ces deux méthodes sont-elles comparables ?
- Les partitions obtenues évoluent-elles au cours des deux périodes d'étude ? Existe-t-il une méthode plus stable que l'autre ?

Un certain nombre d'articles et quelques ouvrages traitent des mesures de comparaison ou indices de similarités entre deux partitions, on peut citer : (Youness & Saporta, 2004), (Nakache & Confais, 2005, pp. 201-205) et (Hubert & Arabie, 1985). Aucun consensus ne semble cependant se dégager quant à l'utilisation d'un indice en particulier. Pour mesurer la concordance, nous avons construit la matrice de confusion[57] à partir des coïncidences d'appartenance dans chacune des deux partitions, des paires d'individus (*resp.* profils). Notant $\mathcal{P}_{ACL}$ (*resp.* $\mathcal{P}_{CAH}$), la partition en C_{ACL} (*resp.* C_{CAH}) classes obtenue à partir d'un ensemble de n observations en utilisant l'*ACL* (*resp. CAH*). **N** désigne la matrice de confusion croisant les classes des deux partitions $\mathcal{P}_{ACL}$ et $\mathcal{P}_{CAH}$ (Tableau 15).

Tableau 15 - Matrice de confusion croisant les classes des deux partitions $\mathcal{P}_{ACL}$ et $\mathcal{P}_{CAH}$.

Partition	$\mathcal{P}_{CAH}$					
	Classes	$\mathcal{P}_{CAH}^{(1)}$	$\mathcal{P}_{CAH}^{(2)}$	…	$\mathcal{P}_{CAH}^{(C_{CAH})}$	Total
$\mathcal{P}_{ACL}$	$\mathcal{P}_{ACL}^{(1)}$	n_{11}	n_{12}	…	$n_{1C_{CAH}}$	n_{1+}
	$\mathcal{P}_{ACL}^{(2)}$	n_{21}	n_{22}		$n_{2C_{CAH}}$	n_{2+}
				⋮		
	⋮	⋮	⋮		⋮	⋮
	$\mathcal{P}_{ACl}^{(C_{ACL})}$	$n_{C_{ACL}1}$	$n_{C_{ACL}2}$	…	$n_{C_{ACL}C_{CAH}}$	$n_{C_{ACL}+}$
	Total	n_{+1}	n_{+2}	…	$n_{+C_{CAH}}$	$n_{++} = n$

[57] Autrement appelée dans ce cas *matrice de coïncidence-paires*.

La matrice de confusion **N** peut s'obtenir à partir des tableaux disjonctifs **K** associés à chacune des partitions ; on a $\mathbf{N} = (\mathbf{K_{ACL}})^{\mathrm{T}}\mathbf{K}_{CAH}$. Chaque partition k peut être représentée par un tableau relationnel $\mathbf{T}^k$ tel que $\mathbf{T}^k = (\mathbf{K_k})^{\mathrm{T}}\mathbf{K}_k$, de terme général défini par :

$$t_{ii'}^k = \begin{cases} 1 \text{ si } i \text{ et } i' \text{sont 2 individus dans la même classe de } \mathcal{P}_k \\ 0 \text{ sinon} \end{cases}$$

Lorsque l'on croise deux partitions, on s'intéresse aux paires d'individus qui sont affectées ou non dans les mêmes classes. Elles sont au total $\mathrm{C}_n^2 = \frac{n(n-1)}{2}$ et peuvent être représentées par les 4 types suivants :

- a : le nombre de paires de points dans la même classe pour chacune des deux partitions ;
- b : le nombre de paires de points dans la même classe de $\mathcal{P}_{ACL}$ et séparées dans $\mathcal{P}_{CAH}$;
- c : le nombre de paires de points dans la même classe de $\mathcal{P}_{CAH}$ et séparées dans $\mathcal{P}_{ACL}$;
- d : le nombre de paires de points séparées dans $\mathcal{P}_{ACL}$ et dans $\mathcal{P}_{ACL}$.

Nous avons, à partir des notations précédentes, calculé les indices suivants :

- l'*indice symétrique de Rand* R et l'*indice asymétrique de Jaccard* J qui mesurent le pourcentage de concordance entre les deux partitions. Ils prennent des valeurs comprises entre 0 et 1; une valeur proche de 1 indique une forte ressemblance. Ils sont définis par :

$$R(\mathcal{P}_{ACL}\,;\mathcal{P}_{CAH}) = \frac{a+d}{a+b+c+d} \text{ et } J(\mathcal{P}_{ACL}\,;\mathcal{P}_{CAH}) = \frac{a}{a+b+c+d}$$

- le *coefficient de Kappa* K qui mesure l'accord entre 2 partitions ayant le même nombre de classes C, en tenant compte de la part de concordance due au hasard (Cohen, 1960). Il prend des valeurs entre -1 et 1. La coïncidence sera d'autant plus élevée que la valeur de Kappa est proche de 1. Il s'écrit :

$$K(\mathcal{P}_{ACL}\,;\mathcal{P}_{CAH}) = \frac{c_{obs} - c_{theo}}{1 - c_{theo}}$$

Où $c_{obs} = \frac{\sum_{i=1}^{C} n_{ii}}{n}$ représente la proportion de concordance observée et $c_{theo} = \frac{1}{n^2}\sum_{i=1}^{C} n_{i+}n_{+i}$ représente la proportion de concordance attendue sous l'hypothèse d'indépendance des partitions (*i.e.* due au seul hasard).

Ces indices ont été calculés en prenant en compte le poids de chaque profil. Ce qui revenait à faire le calcul sur l'ensemble des individus. Pour faciliter les interprétations, la matrice de confusion a aussi été présentée en croisant les effectifs liés aux profils de consommateurs et non seulement ceux associés aux individus. Les directives de prescription n'ayant pas évolué entre 2008 et 2009, nous nous attendions à une certaine stabilité des résultats. Pour le vérifier, la stabilité de chaque méthode de classification au cours des deux périodes a été étudiée en calculant pour chacune d'elle la matrice de confusion croisant les classes des deux partitions obtenues aux deux périodes ainsi que le coefficient de Kappa associé (calculé sur le nombre de profils de consommateurs).

Sur nos données, l'*ACL* avec 4 classes latentes est celle qui s'est avérée le mieux s'ajuster aux données parmi les modèles d'*ACL* testés. Pour les deux années, des dépendances locales ont été détectées entre 4 paires de variables : Spécialité du prescripteur et nomadisme médical , Recommandations liées et facteur F, Recommandations liées et nomadisme médical, Recommandations non liées et facteur F. Nous avons alors introduit dans le modèle les quatre effets directs correspondants. Les probabilités d'appartenance aux classes latentes ($\hat{\pi}_m, m = 1, \dots, 4$) et les probabilités conditionnelles correspondantes ($\hat{p}_{jm} = P\left[\mathrm{X}^j = 1 \middle| \mathrm{Y} = m\right], m = 1, \dots, 4$) ont été estimées ; puis chaque individu i a ensuite été affecté à la classe pour laquelle il possédait une probabilité estimée *a posteriori* $\hat{P}[Y = m|(\mathrm{X}^1, \mathrm{X}^2, \dots, \mathrm{X}^p) = \boldsymbol{x}_i]$ maximum. Une description des classes latentes ainsi formées est fournie dans le Tableau 16 ci-après.

Tableau 16 - Description of Latent Class Models after modal assignment – 2008 and 2009.

	2008				2009			
	Class 1	*Class 2*	*Class 3*	*Class 4*	*Class 1*	*Class 2*	*Class 3*	*Class 4*
Class repartition (%)	58.0	33.1	8.6	0.4	61.0	30.6	8.1	0.4
Overconsumption	0.0	1.4	4.4	76.6	0.2	1.3	4.0	79.8
« Doctor shopping »	0.0	0.0	1.8	55.8	0.0	0.0	2.4	51.4
« Pharmacy shopping »	0.5	0.7	3.6	96.8	0.5	0.9	3.9	95.6
General practitioner prescription	90.6	100.0	1.7	51.3	90.3	100.0	1.2	57.9
Not in agreement to the therapeutic class	0.0	95.0	88.0	73.4	0.0	100.0	88.7	79.8
Not in agreement to other classes	0.0	12.8	34.0	24.0	2.2	7.3	29.4	21.3

Le Tableau 16 montre les 4 classes de consommateurs de bromazépam cliniquement distincts identifiées avec ACL. Les prévalences étaient similaires entre les deux années. Avec l'aide des pharmacologues, voici une interprétation des résultats de 2009, similaires à ceux de 2008 :

- la classe 1 « *Consommateurs non problématiques* » correspond au groupe majoritaire (61%). Elle est caractérisée par une absence de nomadisme médical ou pharmaceutique et de surconsommation. Les prescriptions de bromazépam sont réalisées majoritairement par des médecins généralistes (90%), en accord avec les recommandations ;
- la classe 2 « *Consommateurs limites* » forme également un groupe important (31%). Les prescriptions de bromazépam émanent très largement d'un médecin généraliste et, dans deux tiers des cas, ne sont pas conformes aux recommandations liées aux benzodiazépines. Cette classe pourrait comporter des consommateurs ayant développé une tolérance induite par une consommation chronique susceptible d'entrainer des d'abus ou des phénomènes de dépendance ou de surconsommation ;
- la classe 3 « *Consommateurs présentant des critères de sévérité d'un trouble de la santé mentale* » (8%) est composée de patients recevant une association de psychotropes dont la prescription est non conforme aux recommandations (89%)

Toutefois ces prescriptions émanent très largement de spécialistes (99%). Cette classe comprend des patients pouvant souffrir de troubles mentaux persistants ou graves nécessitant une plus grande association de médicaments psychotropes, association pas toujours en accord avec les recommandations ;

- la classe 4 « *Consommateurs présentant un comportement de transgression* » est minoritaire (0.4%). Une proportion importante de patients a un comportement de nomadisme pharmaceutique (96%) et médical (51%). Ils sont polyconsommants (80%) et surconsommants (80%). Ces comportements peuvent être considérés comme frauduleux et suggèrent un usage compulsif de bromazépam (Wainstein, et al., 2011).

Sur les même données, une *CAH* a été réalisée à partir des coordonnées des profils des patients sur les trois 1ers axes factoriels de l'*AFCM* parmi les 12 $(2 \times 6 - 6)$ possibles ; ceux-ci représentant environ 60% de l'inertie totale. La méthode d'agrégation choisie était le critère de Ward, méthode classiquement adoptée lors d'un enchaînement *AFCM – CAH*. La partition obtenue comporte, elle aussi, 4 classes décrites dans le Tableau 17 :

Tableau 17 - Description of clusters by Agglomerative Hierarchical Clustering – 2008 and 2009.

	2008				2009			
	Class 1	*Class 2*	*Class 3*	*Class 4*	*Class 1*	*Class 2*	*Class 3*	*Class 4*
Class repartition (%)	52.3	28.1	17.4	2.2	53.3	27.7	16.7	2.4
Overconsumption	0.0	0.0	0.0	51.4	0.0	0.0	0.0	50.2
« Doctor shopping »	0.0	0.0	0.0	17.0	0.0	0.0	0.0	18.1
« Pharmacy shopping »	0.0	0.0	0.0	53.8	0.0	0.0	0.0	53.9
General practitioner prescription	100.0	100.0	24.2	60.0	100.0	100.0	21.2	61.7
Not in agreement to the therapeutic class	0.0	100.0	55.3	66.7	0.0	100.0	52.2	69.4
Not in agreement to other classes	0.0	0.0	38.8	21.8	0.0	0.0	33.6	19.5

On retrouve globalement la même interprétation que pour les classes définies par l'*ACL ;* mais les classes 1 et 2 sont plus homogènes et les classes 3 et 4 plus hétérogènes. Pour 2009[58], on observe que :

- la classe 1 correspond au groupe majoritaire (53%). Elle est caractérisée par une absence de nomadisme médical ou pharmaceutique et de surconsommation et des prescriptions de bromazépam réalisées par des médecins généralistes (100%), en accord avec les recommandations ;
- la classe 2 forme également un groupe important (28%). Les prescriptions de bromazépam émanent en totalité d'un médecin généraliste et ne sont pas conformes aux recommandations liées à la classe thérapeutique du bromazépam ;
- la classe 3 (17%) est composée de patients recevant une association de psychotropes dont la prescription n'est pas toujours conforme aux recommandations liées. Les prescriptions émanent le plus souvent de spécialistes (79%). Comme pour les classes 1 et 2, cette classe ne comporte aucun patient développant un comportement de nomadisme pharmaceutique ou médical ou de surconsommation ;
- la classe 4 est minoritaire (2%). Une proportion importante de patients a un comportement surconsommant (54%) et de nomadisme pharmaceutique (54%).

On a donc obtenu deux partitions en quatre classes effectuées sur les mêmes individus et engendrées par deux méthodes différentes (*ACL* et *CAH*), pour deux années de l'étude. La question s'est donc naturellement posée de les comparer. La première différence entre méthodes est apparue en termes d'inertie intra-classe. En effet, l'inertie intra-classe des partitions obtenues avec *CAH* s'est avérée plus faible (1.2) que pour *ACL* (2.7) : la partition *CAH* est plus homogène. Ceci s'observe surtout pour les classes 1 et 2 obtenues avec *CAH* (Tableau 17) pour lesquelles l'affectation est exclusive : la prévalence des comportements de consommation est soit 0%, soit 100%. En revanche, la partition *LCA* a fourni des profils plus nuancés. Le Tableau 18 et le Tableau 19 présentent les matrices de confusions entre *CAH* et *LCA*. Sur la diagonale, on trouve les profils (*resp.* individus) concordants. Les profils (*resp.* d'individus) classifiés différemment se situent en dehors de

[58] Une interprétation similaire peut être faite pour 2008.

la diagonale. Ils représentent peu d'individus (environs 12% en 2009) ; mais un nombre non négligeable de profils (50% environ en 2009). La plupart des différences se situent entre la classe 3 *ACL* et la classe 4 *CAH* : cela représente 18 profils qu'il est difficile de distinguer sans données supplémentaires. En effet, la frontière entre un patient atteint de troubles mentaux sévères (classe 3 *ACL*) et un consommateur compulsif (classe 4 *ACL*) est ténue ! Il ne paraît donc pas étonnant de voir apparaître des différences selon la méthode de classification utilisée.

Tableau 18 - Confusion matrix between LCA partition and AHC partition (data from 2008) with number of responses profiles and number of users.

		LCA clusters				
		Class 1	Class 2	Class 3	Class 4	Total
AHC clusters	Class 1	1 *21236*	0 *0*	0 *0*	0 *0*	1 *21236*
	Class 2	0 *0*	1 *11435*	0 *0*	0 *0*	1 *11435*
	Class 3	1 *2218*	2 *1708*	3 *3145*	0 *0*	6 *7071*
	Class 4	2 *112*	4 *291*	**18** *345*	21 *154*	45 *902*
	Total	4 *23566*	7 *13434*	21 *3490*	21 *154*	53 *40644*

Kappa coefficient = 0.80 ; Rand indice = 0.89 ; Jaccard indice = 0.77

Tableau 19 - Confusion matrix between LCA partition and AHC partition (data from 2009) with number of responses profiles and number of users.

		LCA clusters				
		Class 1	Class 2	Class 3	Class 4	Total
AHC clusters	Class 1	1 *23838*	0 *0*	0 *0*	0 *0*	1 *23838*
	Class 2	0 *0*	1 *12414*	0 *0*	0 *0*	1 *12414*
	Class 3	2 *3246*	1 *970*	3 *3234*	0 *0*	6 *7450*
	Class 4	3 *197*	3 *305*	**18** *369*	21 *183*	45 *1054*
	Total	6 *27281*	5 *13689*	21 *3603*	21 *183*	53 *44756*

Kappa coefficient = 0.80 ; Rand indice = 0.88 ; Jaccard indice = 0.76

Ainsi en 2009 (Tableau 19), les 4 classes *ACL* contiennent respectivement 6, 5, 21 et 21 profils alors que les classes *CAH* en contiennent comme pour 2008 respectivement 1, 1, 6 et 45. On retrouve là aussi le fait que les classes 1 et 2 formées à l'aide la *CAH* sont plus homogènes, alors que la classe 4 est quant-à-elle très hétérogène avec 45 profils différents. Les valeurs prises par l'indice de Rand, de Jaccard et le coefficient de Kappa calculés sur le nombre d'individus confirment la bonne concordance entre partitions.

Les résultats concernant la stabilité sur les deux périodes des deux méthodes sont présentés dans le Tableau 20 et le Tableau 21. Les partitions engendrées par *CAH* sont identiques, traduisant une stabilité parfaite de la méthode. Pour l'*ACL*, la concordance est excellente avec 4 profils sur les 53 changeant de classe d'une année sur l'autre.

Tableau 20 - Confusion matrix between LCA partition in 2008 and LCA partition in 2009 with number of responses profiles.

		LCA clusters 2009				
		Class 1	Class 2	Class 3	Class 4	Total
LCA clusters 2008	Class 1	4	0	0	0	4
	Class 2	2	4	1	0	7
	Class 3	0	1	20	0	21
	Class 4	0	0	0	21	21
	Total	6	5	21	21	53

Kappa coefficient = 0.89

Tableau 21 - Confusion matrix between AHC partition in 2008 and AHC partition in 2009 with number of responses profiles.

		AHC clusters 2009				
		Class 1	Class 2	Class 3	Class 4	Total
AHC clusters 2008	Class 1	1	0	0	0	1
	Class 2	0	1	0	0	1
	Class 3	0	0	6	0	6
	Class 4	0	0	0	45	45
	Total	1	1	6	45	53

Kappa coefficient = 1.00

Ce travail nous a permis de prendre toute la mesure des difficultés liées à une approche pluridisciplinaire. Il montre bien que les choix finaux peuvent être différents selon l'angle d'attaque que l'on prend. Dans ce cas, le statisticien conservera sans hésiter la partition issue de la *CAH,* plus homogène et stable dans le temps, alors que le pharmacologue préfèrera, même si elle est moins robuste, la partition générée par l'*ACL,* plus représentative de la réalité complexe qu'il côtoie. Le choix de l'une ou l'autre des partitions à retenir dépendra des objectifs que l'on s'est fixé : est-ce obtenir une partition la plus stable et homogène possible au sens statistique ? Ou bien, est-ce obtenir une partition la plus réaliste possible prenant en compte autant que ce peut la complexité des données d'un point de vue pharmacologique ?

Ce travail a donné lieu à un article (Feuillet, Bellanger, Hardouin, Vigneau, & Sébille, à paraître 2014) dont sont issus les tableaux 16 à 21 présentés précédemment.

2.2 L'EPIDEMIOLOGIE GENETIQUE (depuis 2013)

Collaborateurs sur ce thème :

- Institut du Thorax (INSERM UMR 1087/ CNRS UMR 6291, Nantes) : C. Dina, M. Karakachoff,
  S. Le Scouarnec, R. Redon, F. Simonet, J.-J. Schott et E. Persyn (stagiaire M2, AgroCampus Rennes).

Ce travail collaboratif s'intègre dans un programme plus vaste, soutenu par la Région Pays de La Loire, intitulé VaCaRMe[59]. VaCaRMe a pour objectif de comparer l'ADN des individus malades à celui des habitants de la Région et ainsi mettre en évidence les mutations génétiques responsables des maladies cardiovasculaires, respiratoires et métaboliques. Cette collaboration avec des collègues de l'Institut du thorax a débuté en 2012 par un encadrement de stage de Master 2 (Tang, 2013). Elle s'est poursuivie en 2014 avec le travail de stage de Master 2 d'Elodie Persyn (Persyn, 2014) puis le co-encadrement de sa thèse débutée en octobre 2014.

[59] Vaincre les maladies Cardiovasculaires, Respiratoires et Métaboliques. http://www.vacarme-project.org/.

2.2.1 Détection de variants rares

L'identification de facteurs génétiques de risque pour les maladies est un des enjeux majeurs en génétique humaine. De nombreuses recherches portent donc sur le développement de méthodes d'analyses du déterminisme génétique des maladies. Les études d'associations pangénomiques ont identifié de nombreuses associations entre variants fréquents et maladies complexes grâce entre autre à la méthode GWAS[60] qui se résume à une étude d'association à très grande échelle (autant d'analyse d'association cas-témoin que de variants génétiques analysés). Cependant, la méthode GWAS a tendance à ne pouvoir détecter que les effets génétiques importants puisqu'elle repose implicitement sur le postulat « common disease – common variant ». Elle fait donc l'impasse sur les variants génétiques rares qui jouent un rôle non négligeable de susceptibilité génétique dans le cas de maladies multifactorielles. Néanmoins, les GWAS ont beaucoup amélioré la connaissance des facteurs génétiques de susceptibilité aux maladies multifactorielles pour une maladie donnée ; cependant l'ensemble des variants ainsi identifiés n'explique qu'une faible partie de la variabilité du phénotype[61] (héritabilité). L'approche récente de Whole Exome Sequencing reposant sur de nouvelles avancées technologiques de séquençage permet maintenant la caractérisation des variants rares dont on pense qu'une localisation accrue dans un gène, ou une région génomique, conférerait un risque modéré à important de développer la maladie étudiée. De nombreux *tests d'association* fondés sur l'agrégation de multiples variants ont été proposés pour identifier les variants génétiques rares associés à une pathologie donnée. C'est un domaine de recherche récent et très actif (voir par exemple (Morgenthaler & Thilly, 2007), (Li & Leal, 2008), (Madsen & Browning, 2009), (Price, et al., 2010), (Liu & Leal, 2010), (Neale, et al., 2011), (Wu, et al., 2011) et (Chen, et al., 2013)). Il est donc apparu nécessaire d'évaluer les méthodes statistiques existantes à l'aide d'études de simulations prenant en compte divers scénarios génétiques et sur des données réelles ; mais aussi d'élaborer des mises en garde sur leur utilisation et de mettre

60 Acronyme de : *Genome Wide Association Study*.

61 En génétique, le *phénotype* se définit comme l'ensemble des caractéristiques observables d'un individu ou d'un organisme. Il dépend du génotype (ensemble des caractéristiques génétiques), de la possession de gènes de prédisposition et de facteurs environnementaux.

en place une stratégie à suivre pour détecter au mieux les variants rares liés au phénotype étudié dans le cadre de données réelles.

Les premiers résultats de ce travail ont été présentés lors de la XXVII International Biometric Conference (IBC) à Florence (Italie) en juillet 2014. Elodie Persyn a présenté son stage lors de la journée annuelle du groupe Biopharmacie & Santé de la *SFdS*, le 27 novembre 2014. Un article sur la détection de variants rares dans le cadre de l'étude génétique du syndrome de Brugada vient aussi d'être soumis au journal HMG[62] en novembre 2014.

2.2.2 Perspectives : encadrement de Thèse (depuis octobre 2014)

Elodie Persyn poursuit depuis peu ce travail dans le cadre d'une thèse intitulée « Analyse d'association de variants génétiques rares dans une population démographiquement stable ». Cette thèse est financée par un contrat doctoral de la région Pays de Loire pour une durée de 3 ans, contrat lié au projet VaCaRMe. Elodie Persyn travaillera sous la direction de Richard Redon. Je la co-encadrerai avec Christian Dina. Il s'agit donc d'une collaboration interdisciplinaire entre deux laboratoires : génomique et mathématiques.

Description du sujet

L'institut du thorax a initié une étude de recherche de l'effet de variants génétiques rares dans la survenue de maladies dégénératives à étiologie complexe. Dans ce contexte, le projet VaCaRMe vise à identifier des gènes responsables de pathologies cardio-vasculaires. La stratégie de recherche de VaCaRMe repose sur l'étude de populations démographiquement stables (rurales) du grand Ouest de la France, populations qui ont eu peu d'apport génétique extérieur à même de « diluer » la susceptibilité aux maladies étudiées. L'un de ses axes de recherche s'intéresse à l'Epidémiologie génétique et à la recherche de variants génétiques pathogènes par études de liaison et d'association dans ce type de populations. A terme, le génotypage de 5000 individus appartenant à la population générale des Pays de la Loire, avec un sous-ensemble de 100 individus séquencés sur tout

62 Acronyme de *Human Molecular Genetics*.

le génome, permettra la création d'un corpus de données exploitable. Par ailleurs, les informations liées à un nombre important de patients souffrants de pathologies cardiaques et issus de la même région seront aussi disponibles. La problématique de cette thèse porte sur le développement et la mise en œuvre de méthodes statistiques permettant l'identification de variants génétiques rares associés à une pathologie donnée à l'aide des données générées par le projet VaCaRMe ; mais aussi par simulations mimant des scénarios démographiques complexes.

L'implication de variants génétiques fréquents dans les pathologies humaines a été confirmée par un grand nombre d'études d'association génome entier. Il apparaît cependant qu'une partie importante de l'héritabilité soit due à la présence de variants rares, qui auraient des effets génétiques plus forts. Tester l'association de ces variants est problématique du fait de leur faible fréquence dans la population générale. En effet, malgré un effet potentiellement important, le petit nombre d'observations (dû à une basse fréquence des allèles) ainsi que le grand nombre de variants rares (nécessitant des comparaisons multiples d'où l'utilisation d'une correction de type Bonferroni ou Benjamini-Hochberg pour les plus couramment employées) ne permet pas d'obtenir une puissance suffisante pour discriminer les vrais des faux positifs. Plusieurs solutions, non mutuellement exclusives, ont été envisagées, la plus classique consiste à combiner les variants rares (tests dits *poolés*) dans une unité fonctionnelle, l'unité la plus commune étant le gène. Au cours de cette dernière décennie, un grand nombre de tests statistiques *poolés* ou non ont été développés, chacun étant plus ou moins puissant puissant selon le scénario étudié ; mais aucun ne s'avérant être uniformément plus puissant. Après un travail de synthèse bibliographique sur les tests d'association existants et de leurs propriétés, il s'agira pour l'étudiant(e) sélectionné(e) de développer une approche statistique originale permettant de tester simultanément un enrichissement de variants rares délétères ou protecteurs chez les individus malades (par rapport aux témoins).

Par ailleurs, il est démontré que les analyses d'association de variants rares sont sensibles à la différence d'origine démographique des individus et que par conséquent des différences de fréquences peuvent refléter une différence d'origine démographique plutôt qu'un effet d'un gène sur la maladie. On ne peut donc pas se dispenser de prendre en compte les déterminants environnementaux dans les analyses ! Or, la correction

habituellement appliquée dans le cadre classique des études d'association pour pallier ce problème (*ACP* sur le génome) semble insuffisante (Jombart, Pontier, & Dufour, 2009). Il s'agira donc aussi, dans ce travail de thèse, de déterminer comment éviter ce facteur de confusion, soit à l'aide de méthodes statistiques multivariées prenant en compte l'information spatiale telles que l'*ACP* spatiale (Jombart, Devillard, Dufour, & Pontier, 2008) ou *MEM*, soit en intégrant cette information dans un modèle statistique sous forme d'interactions gène-environnement.

Enfin, la détection de facteurs génétiques de susceptibilité à la maladie peut aussi être envisagée en commençant par orienter la sélection des cas et des témoins sur des groupes géographiques (communes rurales) stables *i.e.* dans lesquels on observe une forte prévalence de la pathologie étudiée et les sujets apparentés sont plus nombreux que dans la population générale. La fréquence des variants génétiques de susceptibilité sera alors plus élevée que dans la population générale et permettra de ce fait plus facilement la détection des mutations en cause. Le dernier objectif de la thèse concernera la mise en place de procédures de simulations de populations présentant les caractéristiques démographiques de populations rurales stables. Ces procédures permettront d'évaluer la capacité des méthodes existantes et de l'approche développée à détecter des variants rares, selon différents scénarios démographiques complexes.

Chapitre 3 : Exploration de données archéologiques (depuis 2000)

Collaborateurs sur ce thème :

- LAT[63] (UMR CITERES, Tours) : P. Husi ;
- LUNAM, Université de Nantes (UMR 6566 CreAAH, Laboratoire de Recherches archéologiques) : Y. Henigfeld ;
- INRAP[64] : F. Ravoire ;
- Institut régional du Cancer (Unité de biostatistique, Montpellier) : Y. Laghzali ;
- AgroParisTech/INRA (dpt MMIP, Paris) : R. Tomassone.

3.1 LA DATATION DE CONTEXTES PAR LA CERAMIQUE

La question du temps, omniprésente en archéologie, passe par le rapport à l'objet et au contexte archéologique[65]. Elle s'appréhende par la succession ou encore l'évolution qui sont à relier à la chronologie relative ; mais aussi par la date ou datation de contextes archéologiques à relier à la chronologie absolue. Si on examine plus en détails les méthodes de datation utilisant le mobilier archéologique, un certain nombre de problèmes méthodologiques se posent. En effet, ces méthodes sont fondées sur deux types d'objets principaux : ceux – rares – qui portent leur propre date, ce qui ne les rend pas pour autant totalement fiables puisqu'ils s'inscrivent comme les autres dans un système chronologique comportant ses propres faiblesses et contradictions (monnaies ou documents épigraphiques…) ; ceux qui ne sont datés que par référence à une chrono-typologie (céramique et tout autre mobilier) (voir par exemple (Ferdière, 2007)). Chaque système de datation ayant ses propres biais méthodologiques, il est donc essentiel de confronter les différentes sources utilisées pour examiner la cohérence des dates proposées, tout en ne perdant pas du vue ce que l'on cherche à dater !

Dans nos travaux (((Bellanger, Husi, & Tomassone, 2006), (Bellanger, Husi, & Tomassone, 2006), (Bellanger, Husi, & Tomassone, 2008), (Bellanger & Husi, 2012) et (Bellanger & Husi, 2013)), fruits d'une collaboration pluridisciplinaire entre statisticiens

63 Acronyme de *Laboratoire Archéologie et Territoires*.

64 Acronyme de *Institut National de Recherches Archéologiques Préventives*.

65 Seront utilisés indistinctement les termes de « contextes archéologiques » et « ensembles stratigraphiques ». Ils se définissent comme les entités chrono-fonctionnelles constitutives du site archéologique (niveaux d'occupation, ensembles clos sous la forme de dépotoirs ou de latrines…). Nous n'avons retenu dans ces travaux que ceux dont la chronologie est considérée comme fiable.

et archéologues, nous avons développé une modélisation permettant pour un contexte archéologique donné :

- d'obtenir une datation absolue cohérente, tirée de la relation entre date d'émission de la monnaie ou d'un autre élément datant et faciès céramique ;
- de tenter d'appréhender sa durée ou intensité de vie ;
- d'utiliser en retour les profils chronologiques ainsi construits pour mieux le caractériser fonctionnellement.

Nous comparons datation absolue et accumulation pour un contexte donné à l'aide de deux courbes de densité probabilité :

- la première est une courbe de densité gaussienne issue d'un modèle de régression linéaire. Elle permet d'estimer la date moyenne d'émission d'une monnaie retrouvée dans un contexte en fonction de son faciès céramique. D'un point de vue archéologique, cette estimation d'un *terminus post quem*[66], avec tous les biais liés à la datation par les monnaies, représente un point d'ancrage chronologique dans le temps événementiel et calendaire ;
- la seconde est celle d'un mélange de gaussiennes, obtenue à partir du modèle précédent. La date d'un ensemble stratigraphique est estimée par la moyenne pondérée des dates estimées des productions céramiques (groupes techniques) qu'il contient. En supposant l'indépendance des productions, la densité de probabilité associée est une somme pondérée de gaussiennes. D'un point de vue archéologique, cette estimation de la datation représente mieux le temps de l'accumulation inscrit dans la matière (Olivier, 2001). Elle peut s'interpréter, au mieux comme un processus de formation reflétant une succession à l'échelle du temps archéologique si la qualité du contexte archéologique le permet ; au pire, comme une imprécision dans la datation, révélant une forte pollution du contexte par la présence de matériel redéposé ou intrusif.

66 *Terminus post quem* (TPQ), ou date plancher se définit comme la date avant laquelle un ensemble archéologique n'a pas pu se former. En principe, cette date est donnée par l'élément daté le plus récent contenu dans l'ensemble.

La confrontation des courbes de densité permet ensuite de :

- valider la méthode d'un point de vue chronologique ; en explorant l'intensité de l'occupation de chaque contexte ;
- mieux appréhender les questions d'ordre chrono-fonctionnel, par une meilleure interprétation de la nature des contextes archéologiques.

3.1.1 Mobilier archéologique et modélisation statistique

3.1.1.1 Le corpus des données archéologiques

Le choix du corpus est une étape importante puisqu'elle détermine fortement les interprétations archéologiques ultérieures. Il est indispensable, dans la phase de construction du modèle, de ne considérer que les contextes les moins perturbés, ceux susceptibles de nous révéler la plus grande quantité de matériel contemporain de l'action interprétée par l'archéologue. Dans notre cas, le corpus était composé de 278 ensembles stratigraphiques, répartis en 95 ensembles de référence et 183 supplémentaires[67]. Les ensembles de référence correspondent aux ensembles de Tours datés ou non par des monnaies ; mais sélectionnés pour la précision de leur chrono-stratigraphie et la qualité de leur corpus céramique.

La céramique n'ayant intrinsèquement aucune prise sur le temps calendaire, nous avons dû nous référer à des objets datés par ailleurs : dans notre cas les monnaies en contexte, c'est-à-dire contemporaines de l'action archéologique. Parmi les 95 ensembles de référence de Tours, seuls 25 sont datés de cette manière. Par soucis de cohérence, parmi les monnaies retrouvées dans chaque ensemble, la plus récente a été retenue : elle définit alors un *terminus post quem*. Comme toute datation, celle fondée sur les monnaies comporte certains biais : d'une part la durée de circulation avant démonétisation, qu'il est difficile d'estimer pour une étude qui couvre presque quinze siècles ; d'autre part, l'histoire propre à chaque monnaie *in situ*, perdue rapidement ou thésaurisée. Tous ces facteurs peuvent provoquer un décalage potentiel entre date de la monnaie et chronologie d'un ensemble stratigraphique ; un décalage de quelques décennies, n'est pas impossible entre

[67] Ensembles ne participant pas à la construction du modèle.

la date d'émission de la monnaie, plus haute chronologiquement, et la datation effective du contexte archéologique. Tous les ensembles retenus dans ces travaux ont été choisis car ils correspondent à des niveaux d'occupation ou des ensembles clos appartenant à de longues séquences stratigraphiques urbaines, pour lesquels les biais liés à un tel décalage sont minimes. En outre, ce choix vient également de la bonne répartition des contextes datés par des monnaies jalonnant la totalité de la fourchette chronologique entre le IV^e^ et le XVII^e^ siècle.

Les ensembles supplémentaires sont stratigraphiquement isolés et chronologiquement peu documentés pour la ville de Tours (37 ensembles), ou proviennent de sites du centre-ouest de la France, extérieurs à cette ville (146 ensembles).

Les productions céramiques, aussi nommées groupes techniques, dont la discrimination se fait en fonction de la nature de l'argile (taille, fréquence, nature des inclusions) et du type de couverte (sans traitement de surface, engobe, glaçure…) sont inventoriées et classées selon leur appartenance à un atelier lorsque ce dernier est connu et, le cas échéant, rattachées à une tradition de fabrication commune. On recense donc pour cette étude un corpus d'environ 15 000 individus ventilés dans 200 groupes techniques. Cette approche a nécessité de longs efforts d'analyse de la céramique pour harmoniser les méthodes de traitement élaborées depuis plus de 15 ans par un petit groupe de chercheurs et ainsi mettre en place des outils typologiques et des techniques de quantification communs à ce vaste espace d'étude couvrant l'ouest de la France ((Husi, 2003) et (Husi, 2013)). Afin d'élargir le champ d'investigation, un réseau de chercheurs couvrant pour l'instant l'espace européen francophone, a été fédéré autour d'un site internet ICERAMM[68] conçu comme une base de données localisée mettant en ligne les outils typologiques régionaux communs et les notices de sites.

[68] Pour *Information sur la CERamique Médiévale et Moderne* : http://iceramm.univ-tours.fr/.

3.1.1.2 La modélisation statistique

Toute proposition de datation d'un contexte archéologique s'avère délicate et nécessite de ne pas perdre de vue ce que l'on cherche à dater : l'action ponctuelle inscrite dans un temps événementiel ou le processus d'accumulation inscrit dans la durée. L'approche statistique adoptée conduit à l'estimation de deux courbes de densité de probabilité pour dater chaque contexte archéologique ; très visuelle cette approche permet un regard critique immédiat de l'archéologue. Deux étapes ont donc été nécessaires. La première nous a permis d'estimer une date correspondant au *terminus post quem* du contexte, un curseur reflétant le temps événementiel. La seconde, qui utilise les résultats de la première étape, permet alors d'estimer le profil chronologique du contexte, image plus proche du temps archéologique, de l'accumulation.

Étape 1 : modélisation du temps événementiel (*dateEv)*

Il s'agit d'estimer la date d'un contexte archéologique, en fonction de l'assemblage céramique qui le compose. La méthode retenue ici est d'ajuster un modèle de régression linéaire multiple reliant une date connue dans le temps calendaire (ici celle de l'émission d'une monnaie) à son faciès céramique. À la suite de cette phase d'ajustement, nous calculons une prévision ponctuelle et par intervalle de confiance à 95 %, dans un premier temps pour les ensembles de Tours datés ou non par des monnaies composant le corpus de référence (soit 95 ensembles). Puis, nous utilisons le modèle construit pour prévoir la date des ensembles supplémentaires. Les ensembles extérieurs à la ville de Tours ne participent jamais à la construction du modèle. Ce choix permet d'éviter les imprécisions liées à la variation potentielle du faciès céramique due l'éloignement géographique de l'ensemble retenu par rapport au point de référence représenté par Tours. La démarche statistique se décompose en différentes étapes.

- *Synthèse de l'information contenue dans le corpus de données de référence*. On effectue une *Analyse factorielle des correspondances* (*AFC*) avec comme matrice de données **N** les ensembles stratigraphiques en colonne et en ligne les groupes techniques quantifiés en nombre minimum d'individus (soit 95 ensembles dont le faciès céramique est composé de 200 groupes techniques). La matrice de données **N** peut s'écrire sous la forme :

$$\mathbf{N} = \begin{bmatrix} n_{11} & n_{12} & \cdots & n_{1J} \\ n_{21} & & & \\ \vdots & & \ddots & \vdots \\ n_{I1} & & \cdots & n_{IJ} \end{bmatrix} = \left[\mathbf{N}^{(1)} \middle| \quad \mathbf{N}^{(2)}\right] \text{ où } J = 95 \text{ et } I = 200$$

$\mathbf{N}^{(1)} \in \mathcal{M}_{200\times25}$ contient les informations liées aux ensembles datés et $\mathbf{N}^{(2)} \in \mathcal{M}_{200\times70}$ celles liées aux ensembles non datés ou possédant une date peu fiable. Une *AFC* est réalisée à partir des 95 ensembles stratigraphiques de référence de Tours comprenant 25 ensembles datés par au moins une monnaie et 70 ensembles non datés, sélectionnés uniquement pour la qualité de leur chrono-stratigraphie et de leur corpus céramique. Puis, nous conservons uniquement les 10 premiers axes factoriels qui expliquent environ 64 % de l'inertie totale du nuage de points (ou information initiale). Ainsi notre tableau de contingence, croisant 95 ensembles et 200 groupes techniques, se réduit à un tableau 95×10, *reconstitution incomplète du corpus de données*. Ce principe commun aux techniques d'analyse factorielle permet de réduire la dimension du problème étudié et donc dans une étape de modélisation de réduire le nombre de variables explicatives prises en compte dans le modèle (dans notre cas 10 au lieu de 200 !).

- *Estimation d'une date pour chaque ensemble stratigraphique* étudié en fonction de son faciès céramique. Comme nous disposons de peu d'ensembles datés (25), il est nécessaire de sélectionner un nombre limité de facteurs qui serviront de variables explicatives. Le choix de ne retenir que 10 facteurs a été fixé après avoir étudié la variabilité des assemblages céramiques à l'aide de *méthodes de ré-échantillonnage*. Un *modèle de régression linéaire multiple gaussien* est alors ajusté sur ces 25 ensembles de référence datés par les monnaies en fonction des composantes factorielles de l'*AFC* significatives parmi les 10 conservées, soit les 8 premières uniquement notées F^k (k=1,…,8). Il peut s'écrire sous la forme :

$$dateEv^j = \beta_0 + \sum_{k=1}^{8} \beta_k (F^k)_j + \varepsilon_j \quad \forall j = 1, \cdots, 25$$

où ε_j sont Normalement distribués suivant une N(0 ;σ^2) et ${F^k}_j$ est la $j^{\text{ième}}$ coordonnée de la $k^{\text{ième}}$ composante factorielle (*j=1,…,25* et *k=1,…,8*).

Après avoir estimé les paramètres β_k et σ^2 du modèle, en utilisant les résultats classiques de la régression linéaire multiple et vérifié que le modèle s'ajustait bien aux données (on obtient un coefficient de détermination ajusté R^2_{aj} d'environ 0.98 et un écart-type résiduel de 23 ans environ !), nous pouvons inférer sur la datation d'un ensemble ; mais aussi prévoir celle d'un autre dans lequel aucune monnaie n'a été retrouvée, donc non daté. Cette estimation ou prévision [69] se construit uniquement à partir de la part d'information contenue dans le faciès céramique du contexte étudié, c'est-à-dire à partir de ses coordonnées sur les 8 premières composantes factorielles de l'*AFC* significatives. On obtient alors une estimation ponctuelle et par intervalle de confiance à 95% du temps événementiel (*dateEv*) des ensembles de référence ainsi que des ensembles supplémentaires. La qualité des prévisions obtenues pour les ensembles supplémentaires est analysée à l'aide d'arguments externes tels que la présence de monnaies non prise en compte dans le modèle ainsi que la cohérence avec des arguments stratigraphiques connus.

Quelle que soit la nature de l'ensemble, de référence ou supplémentaire, le modèle linéaire gaussien étant validé, il est possible de déterminer et tracer la densité de probabilité gaussienne estimée de $\widehat{dateEv}$ correspondant à chaque ensemble. Elle sera représentée par une courbe grise[70] sur les figures à suivre.

Étape 2 : du temps événementiel au temps de l'accumulation (*dateAc)*

On utilise le modèle de régression construit à la première étape et les formules de transition de l'*AFC* reliant les coordonnées des points-lignes (groupes techniques) aux coordonnées des points-colonnes (ensembles) pour obtenir une estimation ponctuelle de la datation de chaque groupe technique ($dateEv_i; i = 1, \dots, 200$) ainsi qu'un intervalle de confiance à 95 %. On peut alors définir le temps archéologique (noté $dateAc$), autrement dit, le temps de l'accumulation d'un ensemble, comme la somme pondérée des datations des groupes techniques le constituant : les poids étant définis comme les proportions d'individus de chaque groupe technique présent dans l'ensemble. Soit, $dateAc^j = \sum_{i=1}^{I} \pi_{i/j}\, dateEv_i$.

---

[69] On parle de prévision quand aucune monnaie n'a été retrouvée dans l'ensemble stratigraphique.

[70] En orange pour la version couleur.

En supposant l'indépendance des variables aléatoires $dateEv_i$[71], on peut approcher la loi du temps de l'accumulation de tout ensemble *j* par le *mélange de lois gaussiennes* :

$$\widehat{dateAc}^j = \sum_{i=1}^{I} \hat{\pi}_{i/j}\, \widehat{dateEv}_i \approx \sum_{i=1}^{I} \hat{\pi}_{i/j}\, N\left(\widehat{dateEv}_i; s^2\left(\widehat{dateEv}_i\right)\right)$$

$$\text{où } \hat{\pi}_{i/j} = \frac{n_{ij}}{n_{i+}} \text{ tq } \sum_{i=1}^{I} \hat{\pi}_{i/j} = 1$$

On obtient, pour chaque ensemble, une courbe plurimodale représentant l'estimation de la loi du temps de l'accumulation, fondée sur le mélange de densités unimodales (datation de chaque groupe technique). Le tracé de cette loi correspond à la courbe noire sur les différentes figures présentées.

3.1.2 Interprétation et validation des résultats

L'interprétation de ces deux courbes de densité s'effectue à l'aide d'un graphique les représentant simultanément pour un ensemble donné. La Figure 12 présente les résultats relatifs à un niveau de jardin utilisé pour rejeter des déchets domestiques, action datée dans sa phase principale par une monnaie de 1341. On remarque que le pic de la courbe de densité de $\widehat{dateAc}$ (courbe grise) se superpose au pic principal de la courbe de densité de $\widehat{dateEv}$ (courbe noire). Sur cette dernière, les ondulations mineures visibles de part et d'autre du mode s'interprètent suivant la nature de l'ensemble comme la présence d'un matériel redéposé ou intrusif et/ou comme l'existence d'une occupation dont l'intensité a fluctué dans le temps. L'activité principale correspond au moment où les modes des deux courbes de densité se juxtaposent.

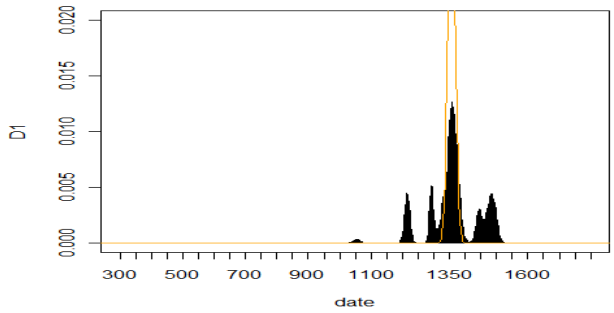


Figure 12 - Juxtaposition des deux courbes (dateEv et dateAc) pour un contexte archéologique interprété comme une zone de rejets domestiques (XIV^e siècle, Tours, Site 8, Ensemble D1, LAT).

[71] Par construction des groupes techniques, cette hypothèse est réaliste.

Une superposition presque parfaite des courbes (Figure 13) révèle quant à elle, un contexte archéologique homogène, avec une part infime de matériel redéposé ou intrusif.

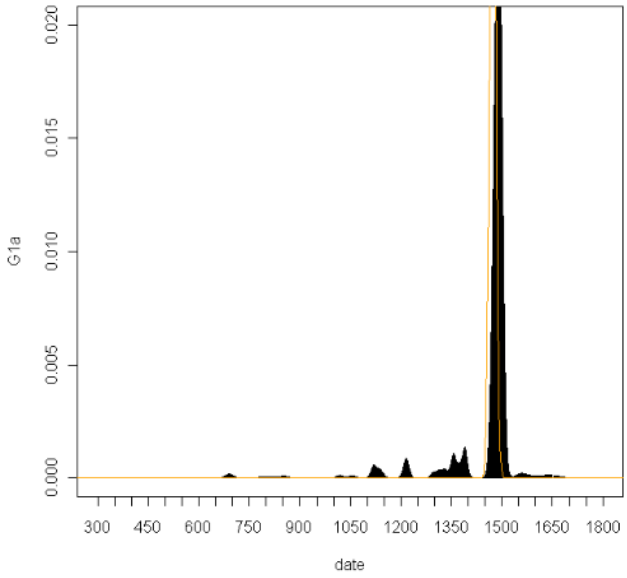


Figure 13 - Dépotoir extérieur utilisant une structure domestique maçonnée abandonnée (XVI^e siècle, Tours, Site 3, Ensemble G1a, LAT).

La validation externe de notre démarche de modélisation est indispensable pour s'assurer de la robustesse des résultats. Elle se fait par validation croisée, à l'aide de données issues d'ensembles stratigraphiques datés ; mais n'ayant pas participé à la construction du modèle. Comme, l'archéologue cherche à confronter des *terminus post quem*, les dates externes sont comparées aux prévisions $\widehat{dateEv}$. Dans la plupart des cas, les dates externes sont proches ou incluses dans l'intervalle de confiance de prévision de $dateEv$ (IC à 95 %). Si décalage il y a, il est toujours dans le même sens : la date externe est un peu antérieure à la fourchette chronologique proposée. Le temps qui peut s'écouler entre la date d'émission et le moment où la monnaie est perdue, voire démonétisée est une des causes du décalage qui peut être observé. La lecture des résultats doit se faire d'une manière critique, en replaçant l'ensemble daté dans l'histoire chrono-stratigraphique du site ; mais aussi en comprenant la réalité de l'objet daté en fonction de la source et des méthodes mises en œuvre.

L'exemple présenté dans le Tableau 22 concerne trois fours domestiques mis au jour sur une fouille à Fondettes (commune voisine de Tours). Il est révélateur de l'importance d'interpréter les diverses datations disponibles en les replaçant dans l'histoire du site : qui date quoi ? Alors que l'archéomagnétisme date la dernière utilisation des fours, la céramique date l'utilisation des structures comme cendrier durant les quelques années qui suivent l'abandon des fours. Au mieux, il peut exister un léger décalage entre ces deux datations, la première devant être logiquement légèrement antérieure à la seconde ou contemporaine. C'est le cas ici puisque, la datation archéomagnétique fournit un intervalle de confiance compris entre 515 et 645 et celle du modèle statistique à l'aide de la céramique (*dateEv)* un intervalle compris entre 631 et 690. Ainsi, s'il existait une monnaie, elle aurait donc 95 % de chance d'être comprise dans cet intervalle de confiance : Le *terminus post quem* du dernier usage des fours, comme celui de l'abandon des structures, a donc une probabilité très faible d'être antérieur à 631 puisque ces deux événements ont de fortes chances d'être quasi contemporains.

Tableau 22 - Validation externe à partir d'ensembles stratigraphiques datés n'ayant pas participé pas à la construction du modèle (sites de Chinon, Rigny et Fondettes).

Ensembles stratigraphiques	**Datations**	**Estimations du modèle céramique** (Intervalles de confiance à 95%)	
		Borne inf	Borne sup
Rigny : bâtiment 1 (LAT)	C*. 662-776	707	851
Rigny : bâtiment 13 (LAT)	M**. 1100	1105	1216
Rigny : bâtiment 17 (LAT)	M**. 1050	1060	1164
Rigny : bâtiment 22.2 (LAT)	M**. 1388	1300	1398
Chinon : fosse 57 (SADIL)	M**. 1423	1382	1497
Fondettes : fours (SADIL)	AM***. 3 dates (515-645)	631	690
LAT = Laboratoire Archéologie et Territoires ; SADIL = Service archéologique départemental d'Indre et Loire			
Datation : *C14 ; **Monnaies ; *** Archéomagnétisme			

Cette recherche a largement contribué à préciser les datations dans l'espace étudié.

3.2 LES APPORTS DE LA MODELISATION

3.2.1 Caractérisation fonctionnelle des contextes archéologiques

Le premier apport est de permettre de mieux caractériser la nature fonctionnelle des contextes archéologiques étudiés, à partir de la distribution de la céramique dans le temps (voir par exemple (Bellanger & Husi, 2012) et (Bellanger & Husi, 2013)). Notre démarche, fondée sur un assemblage approprié de méthodes statistiques et sur les connaissances de l'archéologue, consiste à analyser, comparer et classifier empiriquement les formes des courbes de densité sous l'angle fonctionnel afin d'établir des groupes de formes de courbes. On utilise surtout ici *dateAc* qui permet une lecture de l'accumulation du dépôt dans le temps. En effet, le profil chronologique du dépôt confronté à son interprétation archéologique informe l'archéologue sur son processus de formation et lui permet d'identifier les facteurs susceptibles d'expliquer les similitudes entre profils de contextes. On trouvera des interprétations plus détaillées dans (Husi, 2013).

3.2.1 Caractérisation socio-économique des contextes archéologiques à l'échelle spatiale

Le deuxième apport est de permettre de mieux appréhender l'impact de la variation spatiale des contextes sur les datations obtenues (voir (Bellanger, Husi, & Laghzali, A paraître 2014)). Comme nous l'avons évoqué précédemment, un ensemble pour lequel la datation ne présentera pas de problème majeur présente une excellente juxtaposition des deux courbes de densité de probabilité liées à *dateEv* et *dateAc (cf.* (Figure 13). Nous pouvons donc étudier, au travers de la comparaison des résultats obtenus pour *dateEv* et *dateA,* la capacité du modèle à bien prévoir un ensemble daté comme celle d'un ensemble non daté. En effet, pour un ensemble non daté par une monnaie, la confrontation des résultats fournira un indice de fiabilité de la modélisation : plus la différence sera grande, plus on pourra remettre en cause la capacité intrinsèque de notre modélisation à le dater. Les causes peuvent être multiples : qualité de l'assemblage céramique (redéposition), aire géographique de construction du modèle, aire temporelle de construction du modèle (ensemble trop récent ou trop ancien ; pas de référence). L''étude spatiale de la différence absolue entre ces deux dates estimées $abs\left(\widehat{dateEV} - \widehat{dateAC}\right)$, permet de définir des

aires socio-économiques fondées sur la plus ou moins grande proximité au référentiel (Tours) des faciès céramiques des contextes étudiés.

Différents *indicateurs spatiaux* peuvent être utilisés pour décrire, de façon simple, à partir de données collectées sur les sites disponibles du Centre-Ouest de la France, la distribution spatiale de $abs\left(\widehat{dateEV} - \widehat{dateAC}\right)$. Nous avons choisi d'utiliser des indicateurs ne dépendant pas de délimitations arbitraires de l'aire géographique étudiée. Ils caractérisent la position (centre de gravité et patchs spatiaux), l'occupation de l'espace (inertie, isotropie). (Cotter, et al., 2007) présente en détails, dans le cadre d'une population halieutique, les indicateurs spatiaux disponibles. Ces indicateurs seront des outils utiles pour détecter des tendances :

- De grandes différences de datation pourront traduire l'incapacité du modèle à fournir une date estimée consensuelle ; la distance importante au site de référence (Tours) peut en être la cause.
- La détermination de patchs spatiaux permettra de mettre en évidence et d'interpréter des aires chronologiquement homogènes de faciès céramiques.

Commençons par rappeler brièvement la définition des indicateurs que nous avons utilisés avant de présenter les résultats obtenus sur nos données.

3.2.1.1 *Définition des indicateurs spatiaux*

Le *centre de gravité* (CG) (représente la position géographique moyenne d'une population, *i.e.* la moyenne des positions de ses individus, tandis que l'*inertie* (I) représente la dispersion spatiale autour de (CG). En pratique, ces statistiques sont estimées à partir des données par des sommes discrètes sur la position des échantillons. Dans le cas d'un échantillonnage irrégulier, des surfaces d'influence affectées aux échantillons sont utilisées comme pondérateurs. En pratique, pour N points d'observation $s_i = (x_i, y_i)$, des poids ou aires d'influence[72] α_i et $z(s_i)$ la valeur de la variable régionalisée z en s_i, on définit :

$$GC = \frac{\sum_{i=1}^{N} \alpha_i s_i z(s_i)}{\sum_{i=1}^{N} \alpha_i z(s_i)} \text{ et } I = \frac{\sum_{i=1}^{N} (s_i - GC)^2 \alpha_i z(s_i)}{\sum_{i=1}^{N} \alpha_i z(s_i)}$$

[72] L'introduction de poids est optionnelle, dans notre cas nous avons supposé $\alpha_i = 1; i = 1, \dots, N$.

Une distribution spatiale est dite *isotrope* si elle possède la même dispersion autour (CG) dans toutes les directions. En général, ce n'est pas vrai, et la distribution spatiale est dite *anisotrope*. Dans le plan, en suivant le même principe que l'*ACP*, l'inertie totale (I) d'une population peut être décomposée selon ses deux axes principaux, orthogonaux entre eux, et expliquant respectivement les parts maximum et minimum de l'inertie totale. Les axes principaux et leur inertie associée *(*(I_{max}) et (I_{min})*)* sont obtenus comme les vecteurs propres et les valeurs propres de la matrice d'inertie $\mathbf{S}_I$ de l'*ACP* usuelle du triplet $\left(\mathbf{X}_c, \mathbf{Q} = \mathbf{I}_p, \mathbf{D} = \mathbf{diag}(\alpha_i z(s_i); i = 1, \cdots, N)\right)$ où $\mathbf{X}_c \in \mathcal{M}_{N\times 2}$ est la matrice centrée des coordonnées géographiques $(x_i, y_i)_{i=1,\ldots,N}$. En utilisant les propriétés classiques de l'*ACP*, on peut donc interpréter la racine carrée de l'inertie le long d'un des deux axes donnés comme l'écart type de la projection des positions des individus de la population le long de cet axe. Cette décomposition se représente sur une carte à l'aide d'une croix illustrant les deux directions principales (Figure 16 a et b). Une anisotropie existe quand il y a une différence d'inertie entre les deux directions. L'*indice d'anisotropie* $(Aniso)$ se définit comme la racine carrée du rapport entre l'inertie maximale et l'inertie minimale : plus l'indice est grand devant 1, plus le contraste est marqué entre les directions à cause de l'anisotropie. De même, l'*indice d'isotropie* (Iso) se définit comme l'inverse de l'anisotropie. Il prend des valeurs comprises entre 0 et 1. Soit :

$$Iso = \sqrt{\frac{I_{min}}{I_{max}}} \in [0;1] \text{ et } Aniso = Iso^{-1} = \sqrt{\frac{I_{max}}{I_{min}}} \geq 1$$

$$\text{où } I = I_{max} + I_{min}$$

La distribution spatiale du phénomène étudié dans une aire donnée peut présenter des agrégations hétérogènes appelées patchs. Un algorithme, proposé par P. Petitgas, permet de les identifier (Figure 14): un point d'observation s_i est attribué à un patch selon sa valeur de la variable régionalisée z et sa distance aux autres patchs existants. La position d'un patch est alors déterminée par son centre de gravité.

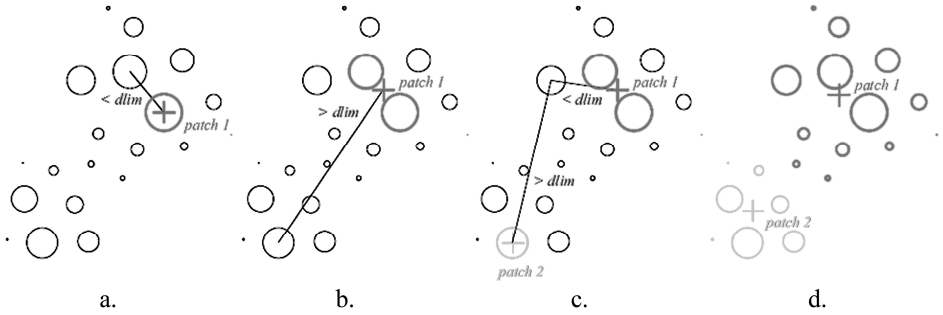


Figure 14 - Etapes permettant la détermination du nombre de patchs d'une population spatialement distribuée.

L'algorithme[73] commence par déterminer la plus grande valeur $z(s_i)$ et considère ensuite chaque point s_i par ordre décroissant de $z(s_i)$. La plus forte valeur initie le premier patch (Figure 14a.). Ensuite, chaque point est affecté à ce premier patch si sa distance au centre de gravité du patch en construction est plus petite qu'une valeur *dlim* fixée en début d'analyse. Si ce n'est pas le cas, il définit un nouveau patch (Figure 14b, c et d.). Ne sont retenus à la fin que les patchs contenant au moins 10 % de l'effectif total N. L'indice créé est alors le *nombre de patchs* (NP).

Pour plus de détails, nous renvoyons par exemple à (Cotter, et al., 2007).

3.2.1.2 Application à la différence absolue entre datations d'un même contexte

Le corpus étudié est composé de 147 ensembles situés sur 25 sites archéologiques différents correspondant à 10 lieux (villes et sites ruraux) (Figure 15). L'hypothèse retenue est que, plus les contextes archéologiques sont distants du point de référence (Tours), plus les dates estimées lors de chacune des deux étapes du modèle sont susceptibles de diverger (Tableau 23). Une petite différence absolue indique que les deux dates estimées sont proches, donc que le modèle fonctionne bien (Tours et Blois) ; même si la date réelle est inconnue. A l'opposé, une différence importante est interprétée comme un manque de fiabilité du modèle qui s'explique par une mauvaise adéquation du modèle aux données traitées (Châtellerault et Poitiers).

[73] Présenté dans le cas où la variable régionalisée prend des valeurs positives ; à adapter sinon.

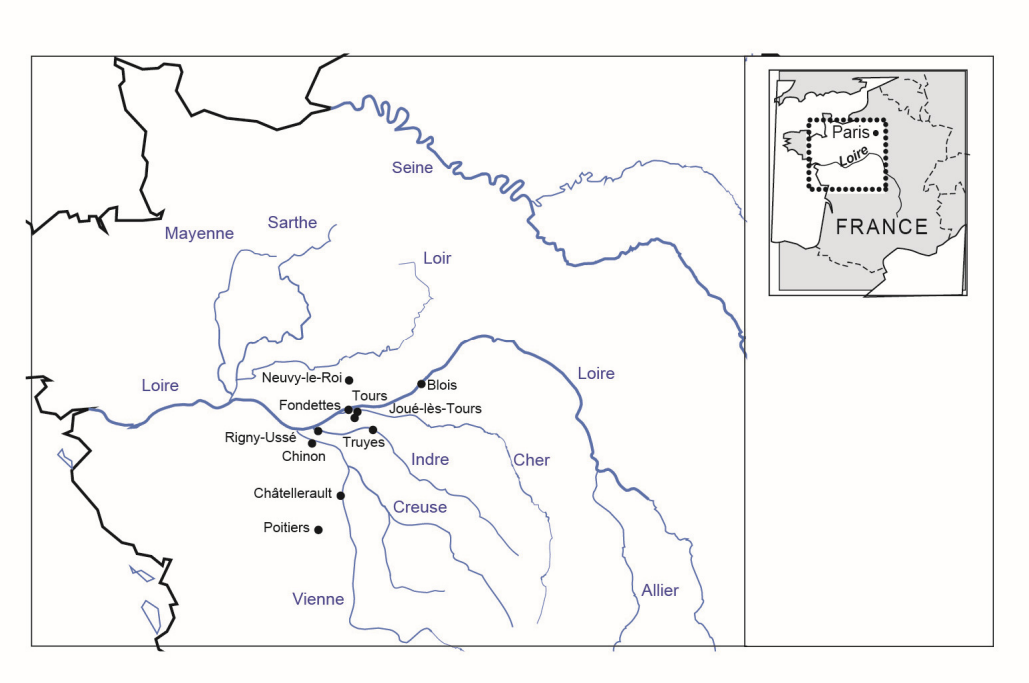


Figure 15 - Représentation des 10 lieux du Centre-Ouest de la France sélectionnés.

Tableau 23 - Extrait du corpus de données et des valeurs obtenues
(en grisée valeurs importantes pour Châtellerault et Poitiers).

Villes et sites	Ensembles	Monnaies	Modèle étape 1 *(dateEV)*	Modèle étape 2 (*dateAC*)	abs(*dateEV – dateAC*)
Tours : Exemples d'ensembles actifs de Tours (avec monnaie) utilisés pour la construction des deux étapes du modèle					
Tours	AA000034	354	355	370	**15**
Tours	AA00016e	814	846	813	**33**
Tours	AA000020	1100	1092	1107	**15**
Tours	D1	1341	1357	1360	**3**
Tours	J2	1476	1475	1486	**11**
Tours	G1b	1488	1489	1489	**0**
Tours.....	R	1631	1635	1611	**24**
Autres villes : exemples d'ensembles supplémentaires (sans monnaie) datés par les estimations du modèle					
Blois (65 km of Tours)	Z001	?	859	853	**6**
Blois.....	Z003	?	849	852	**3**
Chatellerault	Z021	?	1104	1208	**104**
Chatellerault.....	Z022	?	957	808	**149**
Poitiers (100 km of Tours)	Z094	?	1062	1203	**141**
Poitiers	Z096	?	994	1203	**209**
Poitiers.....	Z097a	?	1037	1205	**168**

Pour mieux appréhender l'impact des variations de la variable régionalisée sur les indicateurs spatiaux, on compare un ensemble de données distribué spatialement de manière constante (Figure 16a) aux résultats obtenus sur nos données. La Figure 16b permet d'observer que le centre de gravité (CG) est sensible aux valeurs élevées de $abs\left(\widehat{dateEV} - \widehat{dateAC}\right)$ obtenues pour Poitiers et Châtellerault et qu'il existe une anisotropie marquée de direction préférentielle nord-sud. Cette orientation correspond à

une forte hétérogénéité entre au sud-ouest (Poitiers, Châtellerault) et au nord (Tours, Fondettes, Joué-Lès-Tours, Truyes,…) ; révélant un usage de récipients de même nature ou de même tradition de fabrication dans les deux espaces, mais pour des périodes légèrement différentes. Ce décalage chronologique dans l'utilisation de ces produits révèle très vraisemblablement l'existence d'espaces socio-économiques, ou du moins de réseaux d'approvisionnement, distincts.

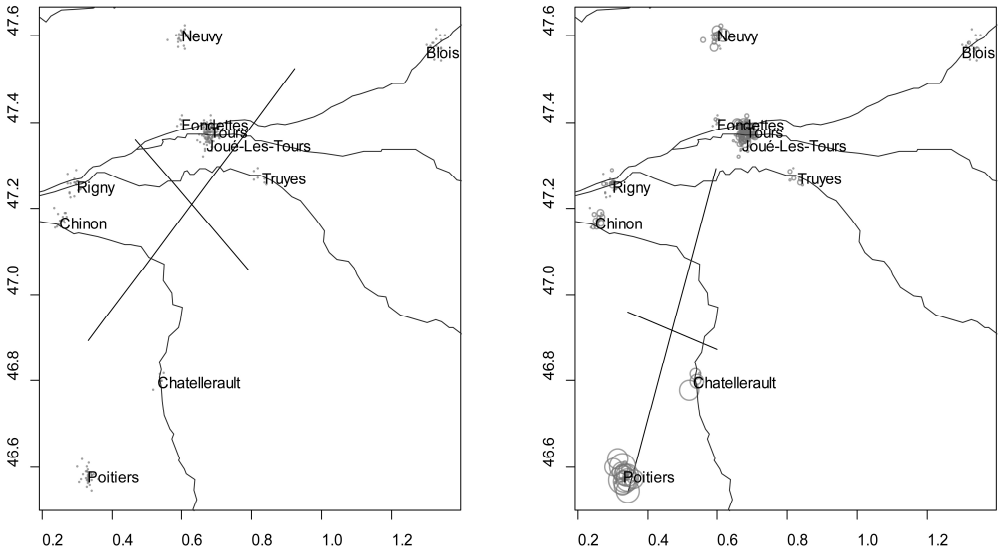


Figure 16 a et b - Exemple de distribution spatiale des ensembles qui montre l'impact entre une distribution constante (à gauche) et une distribution fondée sur les valeurs $abs(\widehat{dateEV} - \widehat{dateAC})$ (à droite). La croix est positionnée au niveau de CG , à partir duquel est représentée la racine carrée de l'inertie selon les deux directions principales.

Afin de mieux comprendre l'imbrication de ces espaces, la construction de patchs spatiaux a été appliquée de manière hiérarchique (Figure 17) en faisant varier la distance maximale acceptable entre points d'observation (ici les sites archéologiques) et centre de gravité d'un patch (*dlim*). Cette approche s'avère, dans sa philosophie, très similaire à celle d'une classification descendante hiérarchique (*CDH*) : on commence tout d'abord par rechercher deux patchs, puis trois puis quatre ainsi de suite. Elle a permis de classifier les sites, dans des divisions de plus en plus fines de patchs. Nous avons ainsi pu déterminer l'échelle d'analyse (nombre de patchs) qui prenait le plus de sens d'un point de vue archéologique.

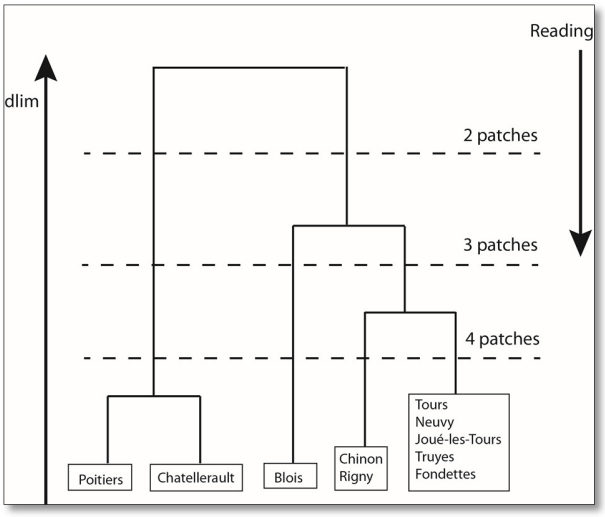


Figure 17 - Nombre de patchs spatiaux divisés suivant la classification hiérarchique.

La partition en trois patchs est celle qui semble archéologiquement la plus intéressante (Figure 18). Elle révèle trois espaces socio-économiques fondés sur un commerce de la poterie qui ne semble pas dépasser un rayon de 30km autour des principaux centres de consommation (Tours, Blois, Poitiers). Ces trois espaces se distinguent soit par des variations techniques et esthétiques dans la réalisation des récipients peints et glaçurés pour une même période ; soit par des traditions de fabrication identiques mais avec un décalage d'usage dans le temps.

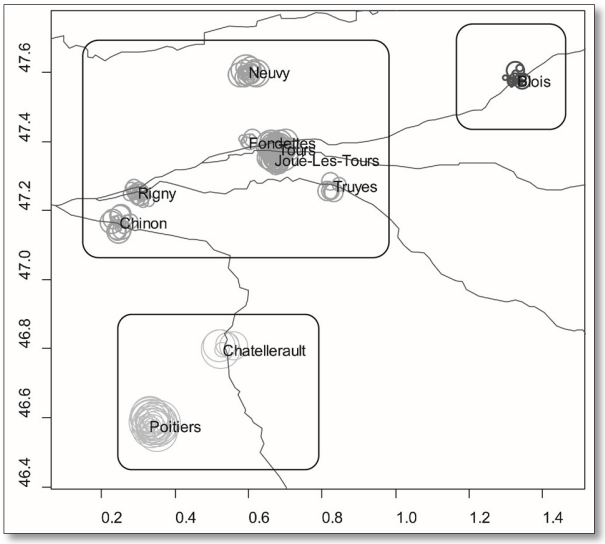


Figure 18 - Répartition en 3 patchs spatiaux (la différence des résultats s'observe dans l'intensité des grisés).

Ces premiers résultats traduisent bien l'importance de poursuivre et de développer une telle recherche. L'analyse spatiale de l'information chronologique est d'une grande aide pour construire ou préciser les contours des aires culturelles. En effet, la présence du même type de poterie à différents endroits et / ou à des moments différents, reflète la rapidité de la circulation des produits, des savoir-faire, des modes et des concurrences.

3.3 PERSPECTIVES

La grande force de notre démarche statistique est la simplicité de sa mise en œuvre qui n'altère en rien l'interprétation archéologique (Baxter, 2008). Les perspectives sont nombreuses :

- La construction d'un référentiel fonctionnel des courbes, à partir des contextes archéologiques dont l'interprétation ne laisse aucun doute, permettra de préciser la nature de contextes dont l'interprétation est plus hypothétique. Plusieurs méthodes statistiques semblent envisageables : classification non supervisée des courbes par analyse fonctionnelle ((Ramsay & Silverman, 2002) et (Ferraty & Vieu, 2006)) ou par analyse procustéenne généralisée (Gower & Dijksterhuis, 2004), puis classification supervisée pour affecter une courbe à un groupe du référentiel.
- L'intégration de nouveaux sites afin de mieux cerner l'organisation sociale et économique du Centre-Ouest de la France ; mais aussi d'étendre la fourchette chronologique aux périodes plus récentes afin d'observer le phénomène dans la plus longue durée.
- La poursuite du travail de comparaison débuté dans (Henigfeld, Husi, Ravoire, & Bellanger, 2013) sur les systèmes d'approvisionnement des centres urbains pour tenter de mesurer les rapports qu'entretiennent les sociétés urbaines médiévales avec les produits céramiques. Nous avons utilisé pour le moment une méthode du type des « méthodes k-tableaux » (voir par exemple (Dazy, Le Barzic, & Saporta, 1996)) adaptée au type de données à traiter (tableaux de contingence) : l'Analyse Factorielle Multiple (en abrégé *AFM* ; voir par exemple (Escofier & Pagès, 1994)).
- La création d'un outil statistique d'aide à la datation des contextes par la céramique, utilisable en ligne par les archéologues. Le développement se fera sous le logiciel libre R pour les analyses statistiques et sous le système ArSol[74] pour l'exploitation des données.

[74] http://citeres.univ-tours.fr/spip.php?article505

Bibliographie

Ardilly, P. (2006). *Les techniques de Sondage.* Paris: Editions Technip.

Baize, D., Bellanger, L., & Tomassone, R. (2009). Relationships between concentrations of trace metals in wheat grains and soil. *Agronomy for Sustainable Development, 29*(2), 297-312.

Bartholomew, D., & Knott, M. (1999). *Latent Variable Models and Factor Analysis.* Londres: Edward Arnold.

Baxter, M. J. (2008). Mathematics, Statistics and Archaeometry: the past 50 years or so. *Archaeometry, 50*, 968-982.

Bel, L., Bellanger, L., Bobbia, M., Ciuperca, G., Dacunha-Castelle, D., Gilibert, E., . . . Oppenheim, G. (1998). On forecasting Ozone Episodes in the Paris area. *Listy Biometryczne-Biometrical Letters, 35*(1), 37-66.

Bel, L., Bellanger, L., Bonneau, V., Ciuperca, G., Dacunha-Castelle, D., Deniau, C., . . . Tomassone, R. (1999). Eléments de comparaison de prévisions statistiques des pics d'ozone. *Revue de Statistique Appliquée, XLVII*(3), 7-25.

Bellanger, L. (1999). *Statistique de la pollution de l'air. Méthodes mathématiques. Applications au cas de la région parisienne.* Doctorat de Sciences de l'Université Paris-Sud, Spécialité : Mathématiques et applications des mathématiques.

Bellanger, L. (2001). Une analyse globale de la tendance dans les hautes valeurs d'ozone mesurées en région parisienne. *Revue de Statistique Appliquée, XLIX*(3), 73-92.

Bellanger, L., & Husi, P. (2012). Statistical Tool for Dating and interpreting archaeological contexts using pottery. *Journal of Archaeology Science, 39*(4), 777-790.

Bellanger, L., & Husi, P. (2013). Mesurer et modéliser le temps inscrit dans la matière à partir d'une source matérielle : la céramique médiévale. *Mesure et Histoire Médiévale, XLIII Congrès National de la Société des Historiens Médiévistes de*

l'enseignement Supérieur Public (SHMESP) (pp. 119-134). Paris: Publication de la Sorbonne.

Bellanger, L., & Perera, G. (1999). High-level exceedances of non-stationary processes and irregular sets. *C.R.Acad.Sci. Paris, 328*(Serie I), 337-342.

Bellanger, L., & Perera, G. (2003). Compound Poisson limit theorems for high-level exceedances of some nonstationary processes. *Bernoulli, 9*(3), 497-515.

Bellanger, L., & Tomassone, R. (1999). Wind direction and maximum pollutants concentration. A case-study with simple statistical tools. Dans B. a. ed., *Advances in Environmental and Ecological Modelling* (pp. 205-214). Paris: Elsevier.

Bellanger, L., & Tomassone, R. (2000). La pollution de l'air dans la région parisienne : étude de la tendance dans les hautes valeurs d'ozone. *Revue de Statistique Appliquée, XLVIII*(1), 5-24.

Bellanger, L., & Tomassone, R. (2004). Trend in High Tropospheric ozone levels: application to Paris Monitoring Site. *Statistics, 38*, 217-241.

Bellanger, L., & Tomassone, R. (2014). *Exploration de données et Méthodes statistiques : data analysis & data mining avec R*. Paris: Ellipses. Collection Références Sciences.

Bellanger, L., Baize, D., & Tomassone, R. (2006). L'analyse des corrélations canoniques appliquée à des données environnmentales. *Revue de Statistique Appliquée, LIV*(4), 7-40.

Bellanger, L., Husi, P., & Laghzali, Y. (A paraître 2014). Spatial statistic analysis of dating using pottery: an aid to the characterization of cultural areas in West Central France. *XXXX Across Space and Time, Proceedings of the 41th International Conference on Computer Applications and Quantitative Methods in Archaeology (CAA-2013).* Perth (Australie).

Bellanger, L., Husi, P., & Tomassone, R. (2006). Statistical aspects of pottery quantification for dating some archaeological contexts. *Archaeometry, 48*, 169-183.

Bellanger, L., Husi, P., & Tomassone, R. (2006). Une approche statistique pour la datation de contextes archéologiques. *Revue de Statistique Appliquée, LIV*(2), 65-81.

Bellanger, L., Husi, P., & Tomassone, R. (2008). A statistical approach for dating archaeological contexts. *Journal of Data Science, 6*(2), 135-154.

Bellanger, L., Vigneau, C., Pivette, J., Jolliet, P., & Sébille, V. (2013). Discrimination of psychotropic drugs over-consumers using a threshold exceedance based approach. *Statistical Analysis and Data-Mining, 6*(2), 91-101.

Biernacki, C. (2009). Pourquoi les modèles de mélange pour la classification ? *La Revue de Modulad, 40*, 1-22.

Bini, M., & et al. (2009). Coefficient shifts in geographical ecology: an empirical evaluation of spatial and non-spatial regression. *Ecography, 32*, 193-204.

Blanchet , F., Legendre, P., & Borcard, D. (2008). Forward selection of explanatory variables. *Ecology, 89*, 2623–2632.

Borcard, D., & Legendre, P. (2002). All-scale spatial analysis of ecological data by means of principal coordinates of neighbour matrices. *Ecological Modelling, 153*, 51–68.

Borcard, D., Legendre, P., & Drapeau, P. (1992). Partialling out the spatial component of ecological variation. *Ecology, 73*, 1045-1055.

Brocard, D., Gillet, F., & Legendre, P. (2011). *Numerical ecology with R.* New York: Springer Science.

Carle, A. (2009). Fitting multilevel models in complex survey data with design weights : Recommendations. *BMC medical research methodology, 9*(1), 49. doi:10.1186/1471-2288-9-49

Chen, Y.-C., Carter, H., Parla, J., Kramer, M., Goes, F., Pirooznia, M., . . . Karchin, R. (2013). A Hybrid Likelihood Model for Sequence-Based Disease Association Studies. *PLoS Genetics, 9*(1).

Cliff, A., & Ord, K. (1981). *Spatial processes. Models and applications.* Londres: Pion.

Cochran, W. (1977). *Sampling Techniques, third ed.* New York: Wiley.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational Psychology Measurement, 20*, 37-46.

Coles, S. (2001). *An Introduction to Statistical Modeling of Extremes Values.* London: Springer Series in Statistics, Springer-Verlag.

Coles, S., & Tawn, J. (1996). Modelling Extremes of the Area Rainfall Process. *Journal of the Royal Statistical Society: series B (statistical methodology), 2*, 329-347.

Cotter, J., Petitgas, P., Mesnil, B., Trenkel, V., Rochet, M.-J., Woilez, M., . . . Lembo, G. (2007). FISBOAT manual of indicators and methods for assessing fish stocks using only fishery. *ICES CM*, *27.*

Cressie, N. (1993). *Statistics for Spatial Data.* New York: John Wiley and Sons Inc.: Wiley Series in Probability and Mathematical.

Davison, A. C., & Smith, R. L. (1990). Models for exceedances over high thresholds (with discussion). *Journal of the Royal Statistical Society: series B (statistical methodology), 62*, 191-208.

Davison, A., & Hinkley, D. (1997). *Bootstrap Methods and their Application.* New york: Cambridge University press.

Dazy, F., Le Barzic, F., & Saporta, G. &. (1996). *L'analyse des données évolutives : méthodes et applications.* Paris: Editions Technip.

de Jong, P., Sprenger, C., & van Veen, F. (1984). On extreme values of Moran's I and Geary's c. *Geographical analysis, 16*, 17–24.

Dray, S., Legendre, P., & Peres-Neto, P. (2006). Spatial modelling : a comprehensive framework for principal coordinate analysis of neighbour matrices (PCNM). *Ecological Modelling, 196*, 483–493.

Dray, S., Pelissier, R., Couteron, P., Fortin, M.-J., Legendre, P., Peres-Neto, P., . . . Wagner, H. (2012). Community ecology in the age of multivariate multiscale spatial analysis. *Ecological Monographs, 82*(3), 257–275.

Droesbeke, J.-J., Lejeune, M., & Saporta, G. (. (2005). *Modèles Statistiques pour Données Qualitatives.* Paris: Editions Technip.

Efron, B., & Tibshirani, R. J. (1993). *An Introduction to the Bootstrap.* New York: Chapman and Hall.

Embrechts, P., Klüppelberg, C., & Mikoch, T. (1999). *Modelling Extremal Events for Insurance and Finance.* New York, 2nd ed.: Springer Verlag.

Escofier, B., & Pagès, J. (1994). Multiple Factor Analysis (AFMULT package). *Computational Statistics and data Analysis, 18*, 121-140.

Everitt, B. (1984). *An introduction to latent variable models.* Londres: Chapman & Hall.

Falk, M., Hüsler, J., & Reiss, R. (2004). *Laws of small numbers : Extremes and rare events.* Birkhaüser.

Ferdière, A. (2007). Le temps des archéologues, le temps des céramologues. *SFECAG, actes du congrès de Langres*, (pp. 15-24).

Ferraty, F., & Vieu, P. (2006). *Nonparametric Functional Data Analysis : theory and Practice.* New York: Springer.

Feuillet, F. (2009). *Caractéristaion de la consommation de médicaments psychotropes à partir des bases de données de l'Assurance Maladie : apport des modèles à classes latentes.* Rapport de stage de Master 2 Professionnel Biostatistique (Bdx 2).

Feuillet, F., Bellanger, L., Hardouin, J.-B., Vigneau, C., & Sébille, V. (à paraître 2014). On comparison of clustering methods for pharmacoepidemiological data. *Journal of Biopharmaceutical Statistics*.

Gower, J., & Dijksterhuis, G. (2004). *Procrustes Problems* (Vol. 30). New York: Oxford University Press, Statistical Science Serie.

Griffith, D. A. (2000). A linear regression solution to the spatial autocorrelation problem. *Journal of Geographical Systems, 2*, 141–156.

Gumbel, E. (1958). *Statistics of Extremes.* New York: Columbia Univ. Press.

Hagenaars, J. A., & McCutcheon, A. L. (2002). *Applied latent class analysis.* Cambridge: Cambridge University Press.

Hartigan, J. (1975). *Clustering Algorithms.* New York: Wiley and Sons.

Hastie, T., & Tibshirani, R. (1990). *Generalized Additive Models.* London: Chapman & Hall.

Henigfeld, Y., Husi, P., Ravoire, F., & Bellanger, L. (2013). L'approvisionnement des villes médiévales (XIIe-XVIe siècles) dans le nord de la France à partir de l'étude de la céramique. Dans E. Lorans, & X. Rodier (Éd.), *Colloque « Archéologie Urbaine », 137eme congrès du Comité des Travaux Historiques et Scientifiques (CTHS)* (pp. 419-431). Tours: PUFR.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6*, 65-70.

Horvitz, D., & Thompson, D. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association, 47*, 663-685.

Hosmer, D., & Lemeshow, S. (2000). *Applied Logistic Regression* (éd. 2nd ed.). New York: John Wiley.

Hotelling, H. (1936). Relations between two sets of variables. *Biometrika, 28*, 321-377.

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification, 2*(1), 193-218.

Husi, P. (2003). *La céramique médiévale et moderne du Centre-Ouest de la France (11e-17e siècle), chrono-typologie de la céramique et approvisionnement de la vallée de la Loire moyenne, supplément à la Revue Archéologique du Centre de la France.* Tours: 20e supplément à la Revue Archéologique du centre de la France, FERAC.

Husi, P. (2013). *La céramique du haut Moyen Age (6e – 10e s.) dans le bassin de la Loire moyenne : de la chrono-typologie aux faciès culturels.* Tours: 49e Supplément à la Revue Archéologique du Centre de la France, ARCHEA/FERACF.

Jombart, T., Devillard, S., Dufour, A.-B., & Pontier, D. (2008). Revealing cryptic spatial patterns in genetic variability by a new. *Heredity, 101*, 92–103.

Jombart, T., Pontier, D., & Dufour, A.-B. (2009). Genetic markers in the playground of multivariate analysis. *Heredity, 102*, 330-341.

Lanphear, B. (2002). Environmental lead exposure during early childhood. *The Journal of Pediatrics, 140*(1), 40-47.

Lanphear, B., Matte, T., Rogers, J., Clickner, R., Dietz, B., Bornschein, R., . . . Jacobs, D. (1998). The contribution of lead-contaminated house dust and residential soil to children's blood lead levels : A pooled analysis of 12 epidemiologic studies. *Environmental Research, 79*(1), 51-68.

Leadbetter, M. (1991). On a basis for “Peaks over Threshold” modeling. *Statistics and Probability Letters, 12*, 357-362.

Leadbetter, M., Lindgren, G., & Rootzen, N. (1983). *Extremes and Related Properties of random Sequences and Series.* New York: Springer Verlag.

Legendre, P., & Legendre, L. (2012). *Numerical ecology* (éd. 3rd). Amsterdam: Elsevier Science BV.

Li, B., & Leal, S. (2008). Methods for Detecting Associations with Rare Variants for Common Diseases: Application to Analysis of Sequence Data. *The American Journal of Human Genetics, 83*, 311-321.

Liu, D., & Leal, S. (2010).). A Novel Adaptive Method for the Analysis of Next-Generation Sequencing Data to Detect Complex Trait Associations with Rare Variants Due to Gene Main Effects and Interactions. *PLoS Genetics, 6*(10).

Lohr, S. (2009). *Sampling: Design and Analysis, 2nd ed.* Boston: Brooks/Cole, Cengage Learning.

Lucas, J.-P. (2013). *Contamination des logements par le plomb : Prévalence des logements à risque et Identification des déterminants de la contamination.* Thèse de doctorat de l'Université de Nantes.

Lucas, J.-P., Bellanger, L., Le Strat, Y., Le Tertre, A., Glorennec, P., Le Bot, B., . . . Sébille, V. (2014). Sources Contribution of Lead in Residential Floor Dust and Within-Home Variability of Dust Lead Loading. *Science of The Total Environment (STOTEN), 470-471*(1 Feb 2014), 768-779.

Lucas, J.-P., Le Bot, B., GlorennecP., Etchevers, A., Bretin, P., Douay, F., . . . Mandin, C. (2012). Lead Contamination in French Children's Homes and Environment. *Environmental Research*(116), 58-65.

Lucas, J.-P., Sébille, V., Le Tertre, A., Le Strat, Y., & Bellanger, L. (2014). Multilevel modelling of survey data: impact of the 2-level weights used in the pseudolikelihood. *Journal of Applied Statistics, 41*(4), 716-732. doi:10.1080/02664763.2013.847404

Lumley, T. (2010). *Complex surveys : A Guide to Analysis Using R.* Hoboken, NJ, USA: Wiley.

Madsen, B., & Browning, S. (2009). A Groupwise Association Test for Rare Mutations Using a Weighted Sum Statistic. *PLoS Genetics, 5*(2).

Mahévas, S., & Trenkel, V. (2002). Utilisation de modèles mixtes pour décrire la distribution spatio-temporelle du temps de pêche de la flotille française en mer Celtique. *Journal de la SFdS, 143*, 177-186.

Mahévas, S., Bellanger, L., & Trenkel, V. (2008). Cluster analysis of linear model coefficients under contiguity constraints for identifying spatial and temporal fishing effort patterns. *Fisheries Research, 93*(1-2), 29-38.

Morgenthaler, S., & Thilly, W. (2007). A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: A cohort allelic sums test (CAST). *Mutation Research, 615*, 28-56.

Munoz, F. (2009). Distance-based eigenvector maps (DBEM) to analyse metapopulation structure with irregular sampling. *Ecological Modelling, 220*, 2683–2689.

Nakache, J.-P., & Confais, J. (2005). *Approche pragmatique de la Classification.* Paris: Editions Technip.

Neale, B., Rivas, M., Voight, B., Altshuler, D., Devlin, B., Ortho-Melander, M., . . . Daly, M. (2011). Testing for an Unusual Distribution of Rare Variants. *PLoS Genetics, 7*(3).

NERC. (1975). *Flood Studies Reports, Vol. 1.* London: National Environmental Research Council.

Nylund, K., Asparouhov, T., & Muthén, B. (2007). Deciding on the Number of Classes in Latent Class Analysis and Growth Mixture Modeling: A monte Carlo Simulation Study. *Structural Equation Modeling, 14*(4), 535-569.

Olivier, L. (2001). Temps de l'histoire et temporalités des matériaux arcjhéologiques : à propos de la nature chronologique des vestiges matériels. *Antiquités Nationales, 33*, 189-201.

Peres-Neto, P., & Legendre, P. (2010). Estimating and controlling for spatial structure in the study of ecological communities. *Global Ecology and Biogeography, 19*(2), 174–184.

Persyn, E. (2014). *Comparaison de méthodes statistiques liées à l'identification de variants génétiques rares associés à une pathologie donnée.* Rennes: Rapport de M2 «

Statistique pour les sciences agronomiques et agroalimentaires », Agrocampus Ouest.

Pfeffermann, D., Skinner, C., Holmes, D., Goldstein, H., & Rasbash, J. (1998). Weighting for unequal selection probabilities in multilevel models. *Journal of the Royal Statistical Society: series B (statistical methodology), 60*(1), 23-40.

Pickands, J. (1971). The two-dimensional Poisson process and extremal processes. *J. Appl. Prob., 8*, 745-756.

Pickands, J. (1975). Statistical inference using extreme order statistics. *Ann. Statist., 3*, 119–131.

Pinet, C., Lecomte, J., Vimont, V., & Auburtin, G. (2003). *Teneurs des plantes à vocation agronomique en éléments traces suite à l'épandage de déchets organiques.* Angers: ADEME.

Platt, T., & Denman, K. (1975). Spectral analysis in ecology. *Annual Review of Ecology and Systematics, 6*, 189–210.

Price, A., Kryukov, G., de Bakker, P., Purcell, S., Staples, J., Wei, L.-J., & Sunyaev, S. (2010). Pooled Association Tests for Rare Variants in Exon-Sequencing Studies. *The American Journal of Human Genetics, 86*, 832-838.

Rabe-Hesketh, S., & Skrondal, A. (2006). Multilevel modelling of complex survey data. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 169*(4), 805-827.

Ramsay, J., & Silverman, B. (2002). *Applied Functional Data Analysis.* New York: Springer.

Rawlings, J., Pantula, S., & Dickey, D. (2001). *Applied Regression Analysis: A Research Tool* (éd. 2nd ed.). New York: Springer and Verlag.

Särndal, C.-E., Swensson, B., & Wretman, J. (2013). *Model Assisted Survey Sampling.* New-York: Springer.

Searle, S. R. (1997). *Linear Models.* New York: Wiley Classics Library.

Shively, T. S. (1991). An analysis of the trend in ground-level ozone using nonhomogeneous Poisson processes. *Atmospheric Environment, 25B*(4), 387-396.

Skinner, C. (1989). Domain means, regression and multivariate analysis. Dans A. o. surveys, & D. H. C. J. Skinner (Éd.). Chichester: Wiley.

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models.* Boca Raton, FL: Chapman & Hall/CRC.

Smith, R. (1989). Extreme values analysis of environmental time series: An application to trend detection in ground-level ozone (with discussion). *Statistical Sciences, 4*, 367–393.

Smith, R., & Shively, T. (1995). Point process approach to modelling trends in tropospheric ozone based on exceedances of a high threshold. *Atmospheric Environment, 29*(3), 3489-3499.

Tang, Y. (2013). *Statistical Tests for the detection of Associations of Rare Variants.* Vannes: Rapport de M2 Professionnel Mathématiques, Informatique, Statistiques.

Thioulouse, J., Chessel, D., & Champely, S. (1995). Multivariate analysis of spatial patterns: a unified approach to local and global structures. *Environmental and Ecological Statistics, 2*(1), 1-14.

Tiefelsdorf, M. (2000). *Modelling Spatial Processes - The Identification and Analysis of Spatial Relationships in Regression Residuals by Means of Moran's I.* Berlin: Springer verlag.

Tillé, Y. (2001). *Théorie des sondages : Echantillonnage et estimation en populations finies.* Paris: Dunod.

Tomassone, R., Charles-Bajard, S., & Bellanger, L. (2000). Discussion sur l'article : La planification des expériences : choix des traitements et dispositif expérimental de Pierre Dagnelie. *SFdS, 141*(1-2), 59-64.

Wainstein, L., Victorri-Vigneau, C., Sébille, V., Hardouin, J.-B., Feuillet, F., Pivette, J., . . . Jolliet, P. (2011). Pharmacoepidemiological characterization of psychotic drugs consumption using a latent class analysis. *International Clinical Psychopharmacology, 26*(1), 54-62.

White, I. R., Royston, P., & Wood, A. M. (2011). Multiple imputation using chained equations: issues and guidance for practice. *Stat. Med., 30*, 377-399.

Wu, M., Lee, S., Cai, T., Li, Y., Boehnke, M., & Lin, X. (2011). Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test. *The American Journal of Human Genetics, 89*(1), 82-93.

Yates, F., & Grundy, P. (1953). Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society. Series B (Methodological), 15*(2), 253-261.

Youness, G., & Saporta, G. (2004). Une méthodologie pour la comparaison de partitions. *Revue de statistique appliquée, 52*(1), 97-120.

Glossaire des acronymes importants

ACC : Analyse des Corrélations Canoniques

ACL : Analyse en Classes Latentes

ACP : Analyse en Composantes Principales

AFC : Analyse Factorielles des Correspondances

AFCM : Analyse Factorielles des Correspondances Multiples

AFD : Analyse Factorielle Discriminante

AFM : Analyse Factorielle Multiple

CART: Classification And Regression Trees

CAH : Classification Ascendante Hiérarchique

GPD : Distribution de Pareto généralisée

MEM : Moran's Eigenvector Map

MST : Minimum Spanning Tree

PCNM : Principal Coordinates of Neighbour Matrices

PCoA : Analyse en Coordonnées Principales

POT : Peaks Over Threshold

PPNH : Processus de Poisson Non-Homogène

ROC : Receiver Operating Characteristic

SAR : Simultaneous AutoRegressive Model

Table des figures

Table des tableaux

Annexe 1 : Production scientifique

Articles avec comité de lecture

1. Feuillet F., Bellanger L., Hardouin J.B., Vigneau C., Sébille V. (2014). On comparison of clustering methods for pharmacoepidemiological data. *Journal of Biopharmaceutical Statistics*. (under press).
2. Lucas J.-P; Sébille V., Le Tertre A., Le Strat Y.; Bellanger L. (2014). Multilevel modelling of survey data: impact of the 2-level weights used in the pseudolikelihood. *Journal of Applied Statistics*, **41**(4): 716-732. DOI:10.1080/02664763.2013.847404. http://dx.doi.org/10.1080/02664763.2013.847404.
3. Bellanger L., Vigneau C., Pivette J., Jolliet P. and Sébille V. (April 2013). Discrimination of psychotropic drugs over-consumers using a threshold exceedance based approach. *Statistical Analysis and Data-Mining*, **6**(2): 91-101.DOI: 10.1002/sam.11165.
4. Bellanger L., Husi P., Tomassone R. (2008). A statistical approach for dating archaeological contexts. *Journal of Data Science*, **6**(2): 135-154. Revue en ligne, http://www.sinica.edu.tw/~jds/
5. Bellanger L., Baize D., Tomassone R. (2006). L'analyse des corrélations canoniques appliquée à des données environnementales. *Revue de Statistique Appliquée*, LIV(4): 7-40.
6. Bellanger L., Husi P., Tomassone R. (2006). Une approche statistique pour la datation de contextes archéologiques. *Revue de Statistique Appliquée*, LIV(2): 65-81.
7. Bellanger L., Tomassone R. (2004). Trend in High Tropospheric ozone Levels: application to Paris Monitoring Site. *Statistics*, **38**(3): 217-241. DOI: 10.1080/02331880410001696116
8. Bellanger L., Perera G. (2003). Compound Poisson limit theorems for high-level exceedances of some non-stationary processes. *Bernoulli*; **9**(3): 497-515.
9. Bellanger L. (2001). Une analyse globale de la tendance dans les hautes valeurs d'ozone mesurées en région parisienne. *Revue de Statistique Appliquée*, XLIX(3): 73-92.
10. Tomassone R., Charles-Bajard S., Bellanger L. (2000). Discussion sur l'article : La planification des expériences : choix des traitements et dispositif expérimental de Pierre DAGNELIE ; publié dans le numéro du Journal de la SFdS contenant l'article de Pierre Dagnélie ainsi que la contribution d'autres participants à la discussion : Azaïs J.-M. et Monod H., Cheroute G., Demonsant J., Duby C., Finney D. J., Kobilinski A., Sado M.-C. et Sado G. *SFdS*, 141 (n°1-2) : 59-64.
11. Bellanger L., Tomassone R. (2000). La pollution de l'air dans la région parisienne : étude de la tendance dans les hautes valeurs d'ozone. *Revue de Statistique Appliquée*, XLVIII(1): 5-24.
12. Bel L., Bellanger L., Bonneau V., Ciuperca G., Dacunha-Castelle D., Deniau C., Ghattas B., Misiti M., Misiti Y., Oppenheim G., Poggi J.-M, Tomassone R. (1999). Eléments de comparaison de prévisions statistiques des pics d'ozone. *Revue de Statistique Appliquée*, XLVII(3): 7-25.
13. Bel L., Bellanger L., Bobbia M., Ciuperca G., Dacunha-Castelle D., Gilibert E., Jackubowicz P., Oppenheim G., Tomassone R. (1998). On forecasting Ozone Episodes in the Paris Area. *Listy Biometryczne-Biometrical Letters*, **35**(1): 37-66.

Notes et articles pluridisciplinaires

14. Lucas J.-P., Bellanger L., Le StratY., Le Tertre A., Glorennec P., Le Bot B., Etchevers A., Mandin C., Sébille V. (2014) Sources Contribution of Lead in Residential Floor Dust and Within-Home Variability of Dust Lead Loading; *STOTEN (Science of the Total Environment)*. 470–471, 1 Feb 2014: 768–779. http://dx.doi.org/10.1016/j.scitotenv.2013.10.028
15. Lucas J.-P., Le Bot B., Glorennec P., Etchevers A., Bretin P., Douay F., Sébille V., Bellanger L., Mandin C. (2012). Lead Contamination in French Children's Homes and Environment. *Environmental Research*, 116: 58-65.
16. Bellanger L., Husi P. (2012). Statistical Tool for Dating and interpreting archaeological contexts using pottery. *Journal of Archaeological Science*, **39**(4): 777-790. http://dx.doi.org.gate3.inist.fr/10.1016/j.jas.2011.06.031
17. Baize D., Bellanger L., Tomassone R. (2009). Relationships between concentrations of trace metals in wheat grains and soil. *Agronomy for Sustainable Development*, **29**(2): 297-312.
18. Mahévas S., Bellanger L., Trenkel V. (2008). Cluster analysis of linear model coefficients under contiguity constraints for identifying spatial and temporal fishing time patterns. *Fisheries Research*, **93**(1-2): 29-38.
19. Bellanger L., Husi P., Tomassone R. (2006). Statistical aspects of pottery quantification for dating some archaeological contexts. *Archaeometry*, **48**(1): 169-183. . DOI: 10.1111/j.1475-4754.2006.00249.x
20. Bellanger L., Perera G. (1999). High-level exceedances of non-stationary processes and irregular sets. *C.R.Acad. Sci. Paris*, **328**, Série I: 337-342.

Actes de conférences Nationales et internationnales avec comité de relecture

21. Bellanger L., Husi P., Laghzali Y. (à paraître 2014). Spatial statistic analysis of dating using pottery: an aid to the characterization of cultural areas in West Central France. *In.* XXXX Across Space and Time, Proceedings of the 41th International Conference on Computer Applications and Quantitative Methods in Archaeology (CAA-2013)**,** Perth (Australie), March 25 - 28 2013.
22. Henigfeld Y., Husi P., Ravoire F., Bellanger L. (2013). L'approvisionnement des villes médiévales (XIIe-XVIe siècles) dans le nord de la France à partir de l'étude de la céramique. /In /: Lorans E., Rodier X. dir. – Colloque « archéologie urbaine », 137eme congrès du Comité des Travaux Historiques et Scientifiques (CTHS) Composition (s) urbaine (s), PUFR, Tours (France) : 419-431.
23. Bellanger L., Husi P. (2013). Mesurer et modéliser le temps inscrit dans la matière à partir d'une source matérielle : la céramique médiévale. *In* : Mesure et Histoire Médiévale, XLIII^{e} Congrès National de la Société des Historiens Médiévistes de l'enseignement Supérieur Public (SHMESP), Publication de la Sorbonne : 119-134.

Chapitres d'ouvrage

24. Husi P., Bellanger L. (juillet 2014). De la modélisation à la datation : le Tableau Général des Ensembles (TGE) du site 3, modalités d'établissement du Tableau et grille de lecture. *In* : Galinié H., Husi P., Motteau J. (dir.), *Recherche sur Tours 9. Des thermes de l'Est de Caesarodunum au château de Tours : Le site 3*, volume papier et en ligne (http://citeres.univ-tours.fr/rt9/). 50^{e} supplément à la Revue Archéologique du Centre de la France. ed. FERACF Tours : 101-107.

25. Husi P., Bellanger L. (2013a). De la modélisation chronologique des données à l'identification des espaces économiques. *In* : Husi P. (dir.) - *La céramique du haut Moyen Age (6e – 10e s.) dans le bassin de la Loire moyenne : de la chrono-typologie aux faciès culturels*. 49e supplément à la Revue Archéologique du Centre de la France. ed. ARCHEA/ FERACF.
    Husi P., Bellanger L. (2013b). La typologie des récipients : un indicateur qui confirme une tendance générale, ibidem.
    Husi P., Bellanger L. (2013c). Flux et échanges de produits : une aide à la définition d'aires culturelles à la fin du haut Moyen Âge, ibidem.
26. Bellanger L., Tomassone R. (1999). Wind direction and maximum pollutants concentration. A case-study with simple statistical tools. In Blasco and Weill ed., *Advances in Environmental and Ecological Modelling*. Elsevier, Paris: 205-214.

Ouvrages

27. Bellanger L., Tomassone R. (mars 2014), Exploration de données et méthodes statistiques : Data analysis & Data mining avec *R*. *Collection Références Sciences, Editions* Ellipses, Paris. 480 pages.

A paraître

- Husi P., Bellanger L. (à paraître 2014). De la modélisation à la datation du site de Rigny, In. Zadora-Rio, Galinie dir *La fouille du site de Rigny (7e-19e s.). De la colonia de Saint-Martin de Tours au centre paroissial,* collection Référentiels, co-éditée par la MSH-Paris et les Editions Epistèmes.

Soumis

- Le Scouarnec S.,Karakachoff M., Gourraud J.-B., Lindenbaum P., Bonnaud S., Portero V., Duboscq-Bidot L., Daumy X., Simonet F., Teusan R., Baron E., Violleau J., Persyn E., Bellanger L., Barc J., Chatel S., Martins R., Mabo P., Sacher F., Haïssaguerre M., Kyndt F., Schmitt S., Bézieau S., Le Marec H., Dina C., Schott J.-J., Probst V., Redon R. (soumis novembre 2014 au *Human Molecular Genetics* (*HMG*)). Testing the burden of rare variation in arrhythmia-susceptibility genes provides new insights into molecular diagnosis for Brugada syndrome.

Articles en cours d'écriture

- Mahévas S., Brind'Amour A., Doray M., Legendre P., Bellanger L. Extending the spread of Moran Eigenvector Maps (MEM) using Negative MEM.
- Brind'Amour A., Mahévas S., Legendre P., Bellanger L. Extending the spread of Moran Eigenvector Maps (MEM): The case of irregular sampling design.

Autres

28. Bellanger L. (1999). *Statistique de la pollution de l'air. Méthodes mathématiques. Applications au cas de la région parisienne*. Thèse de doctorat de l'Université Paris XI – Orsay, dirigée par D. Dacunha-Castelle et R. Tomassone.
29. Bel L., Bellanger L., Bobbia M., Bonneau V., Ciuperca G., Coursol J., Dacunha-Castelle D., Deniau C., Ghattas B., Misiti M., Misiti Y., Oppenheim G., Poggi J.-M., Tomassone R. (Octobre 1997). *Prévision des épisodes de pollution dans la région parisienne, O_3 et NO_2 : phase opérationnelle*. Rapport de contrat de recherche AIRPARIF, 193 pages.

30. Bel L., Bellanger L., Bobbia M., Chapalain V., Ciuperca G., Coursol J., Dacunha-Castelle D., Jackubowicz P., Oppenheim G., Tomassone R. (Octobre 1996). *Prévision des épisodes de pollution dans la région parisienne : O_3 et première étude sur le NO_2.* Rapport de contrat de recherche AIRPARIF, 118 pages.
31. Bellanger L. (1995). *Étude de l'évolution du taux d'ozone troposphérique sur le site de Neuilly/Seine, pour la période 1989-1994.* Mémoire de DEA, dirigé par R. Tomassone.

Posters, Communications orales dans des conférences nationales et internationales

- Bellanger L., Persyn E., Simonet F., Redon R., Schott J.-J., Le Scouarnec S., Karakachoff M., Dina C. (6 – 11 Juillet 2014). Rare variants in human genetic diseases: comparison of association statistical tests. XXVII International Biometric Conference, Florence (Italie).
- Le Scouarnec S., Portero V., Daumy X., Bonnaud S., Duboscq-Bidot L., Teusan R., Lindenbaum P., Karakachoff M., Simonet F., Bellanger L., Gourraud J.-B., Sacher F., Barc J., Chatel S., Dina C., Kyndt F., Beziau S., Schott J.-J., Probst V., Redon R. (29-31 janvier 2014). Criblage systématique du spectre et de la prévalence des variations génétiques touchant les gènes de susceptibilité aux arythmies cardiaques héréditaires. 7èmes Assises de Génétique humaine et médicale, Bordeaux.
- Bellanger L., Husi P., Laghzali Y. (25-28 mars 2013). Spatial statistics analysis of dating using pottery: characterization of socio-economic spaces in West Central France. 41st Computer Applications and Quantitative Methods in Archaeology Conference (CAA 2013), Perth, Western Australia.
- Lucas J.-P; Le Bot B.; Glorennec P.; Etchevers A.; Bretin P.; Douay F.; Mandin C; Sébille V. ., Bellanger L (5-7 novembre 2012). Modélisation multi-niveaux de données d'enquête : impact des poids de sondage de niveau 2 introduits dans la pseudo-vraisemblance sur les estimations. 7e colloque francophone sur les sondages ENSAI.
- Lucas J.-P; Le Bot B.; Glorennec P.; Etchevers A.; Bretin P.; Douay F.; Mandin C., Bellanger L ; Sébille V. (5-7 novembre 2012). Etat de la contamination par le plomb des logements français. 7e colloque francophone sur les sondages ENSAI.
- Husi P., Bellanger L. (31 Mai – 3 Juin 2012). Mesurer et modéliser le temps inscrit dans la matière à partir d'une source matérielle : la céramique médiévale. 43e Congrès National de la Société des Historiens Médiévistes de l'enseignement Supérieur Public (SHMESP), Tours.
- Henigfeld Y., Husi P., Ravoire F., Bellanger L. (23 – 27 avril 2012). L'approvisionnement des villes médiévales (XIIe-XVIe siècles) dans le nord de la France à partir de l'étude de la céramique. 137ème congrès du Comité des Travaux Historiques et Scientifiques (CTHS) Composition (s) urbaine (s), Colloque « Archéologie Urbaine », Tours.
- Lucas J.-P; Le Bot B.; Glorennec P.; Etchevers A.; Bretin P.; Douay F.; Sébille V.; Bellanger L.; Mandin C. (24-26 octobre 2011) Lead Contamination in French Housing. 21st annual International Society of Exposure Science (ISES) Conference 2011, Baltimore (Etats-Unis).
- Brind'amour A., Mahévas S., Doray M., Bellanger L., Legendre P. (29 juin – 1er juillet 2011) Considérations méthodologiques de l'analyse des structures spatiales à l'aide des « MEM ». 10 Forum AFH Boulogne-sur-Mer.

- Mahévas S., Brind'amour A., Bellanger L., Legendre P. and Doray M. (20-24 septembre 2010) Investigating spatial and temporal relationships in fisheries and ecology field using rigorously Moran's eigenvector maps. ICES Annual Science Conference, Nantes.
- Bellanger L. (13 au 18 juillet 2008). Identification of spatial and temporal fishing using cluster analysis of linear model coefficients under contiguity constraints

  et

  Determination of an overconsumption's threshold for the evaluation of abuse and dependence potential of drugs. XXIV International Biometric Conference, Dublin (Irlande).
- Bellanger L. (16 juillet- 21 juillet 2006). Canonical Correlation Analysis Applied to Environmental Data. XXIII International Biometric Conference, Montréal (Canada).
- Bellanger L. (29 juin- 2 juillet 2003). Statistical aspects of pottery quantification for dating archaeological contexts in the Tours city. CARME 2003 (International conference on Correspondence Analysis and related Methods), Barcelone (Espagne).
- Bellanger L., Tomassone R. (13-17 mai 2002). Les hautes valeurs d'ozone : nombre, taille et validation. XXXIV Journées de Statistique, organisées par la Société Française de Statistique (SFdS), Bruxelles Louvain-la-Neuve (Belgique).
- Bellanger L., Charles-Bajard S., Tomassone R. (12-16 novembre 2001). Algunos conceptos y instrumentos estadísticos en ciencias de la vida y otros campos. Journées du CLAPEM, La Havane (Cuba).
- Bellanger L., Husi P. et Tomassone R. (21-25 août 2000). Statistical Tools for Ceramics Dating. STAT'2000, International Conference on Mathematical Statistics, Conférence organisée par l'Institut de Mathématiques de Wroclaw, l'Université de technologie et l'Institut de Mathématiques de l'Académie des Sciences de Pologne. Szklarka Poreba (Pologne).
- Bellanger L. (17-21 mai 1999). Utilisation d'un processus de Poisson non-homogène pour étudier la tendance dans les hautes valeurs d'ozone. XXXIème Journées de Statistique, organisées par la Société Française de Statistique (SFdS), Grenoble.

Communications aux principaux séminaires et journées d'études

- Tissier T., Bellanger L. (15 Novembre 2012). Analyse statistique de gènes différentiellement exprimés dans le cadre d'une étude par méta-analyse : application à l'analyse de la tolérance opérationnelle en transplantation rénale. Journée scientifique du groupe Biopharmacie et Santé de la SFdS.
- Bellanger L., Husi P. (15-16 mars 2012). Chronologie et stratigraphie : modélisation statistique à partir du mobilier archéologiques- Séminaire du réseau inter-MSH Information Spatiale et Archéologie (ISA) « Statistiques spatiales et géostatistiques appliquée à l'archéologie » – MSH Val de Loire, Tours.
- Bellanger L., Vigneau C., Pivette J., Jolliet P., Sébille V. (8 Juin 2009) Surconsommation médicamenteuse: création et validation statistique d'un outil épidémiologique adapté aux bases de données de l'assurance maladie. *Journées Scientifiques de l'Université de Nantes, colloque "Recherche en Pharmacoépidémiologie"*, Nantes.
- Bellanger L., Husi P., mai 2006. Une approche statistique pour la datation de contextes archéologiques à partir des données stratigraphiques et mobilières des fouilles de Tours, Rigny-Ussé et Chinon : résultats et perspectives. Séminaire du Laboratoire Archéologie et Territoires (UMR 6173 CITERES), Tours.

Annexe 2 : Cinq publications représentatives du travail de recherche

1. TREND IN HIGH TROPOSPHERIC OZONE LEVELS: APPLICATION TO PARIS MONITORING SITE.

Bellanger L., Tomassone R. (2004). *Statistics*, **38**(3)

Statistics, June 2004, Vol. 38(3), pp. 217–241

TREND IN HIGH TROPOSPHERIC OZONE LEVELS. APPLICATION TO PARIS MONITORING SITES

LISE BELLANGER[a,*] and RICHARD TOMASSONE[b,†]

[a]*Laboratoire de Mathématiques Jean Leray, Université de Nantes, 2 rue de la Houssinière, BP 92208, 44322 Nantes Cedex 3, France;* [b]*Département de Mathématique et Informatique, Institut National Agronomique, 16 rue Claude Bernard, 75231 Paris Cedex 05, France*

(Received 27 August 2002; Revised 14 July 2003; In final form 19 December 2003)

This article describes the extreme value analysis of tropospheric ozone level exceedances collected at seven monitoring sites in the Paris metropolitan area during May–September over the 14-year period 1988–2001. The purpose of the study is to establish whether observed trends over time in ozone exceedances of a high threshold are real or if they are the result of meteorological changes affecting the conditions under which ozone is generated. A non-homogeneous Poisson process (NHPP), with parameters depending on meteorological covariates, temporal trend and sites factor, is used to model regionally the exceedance times and sizes of daily maxima of ozone over a high threshold. We highlight the importance of non-linear methods to detect the non-linearities.

Keywords: Extreme value theory; Non-homogeneous Poisson processes; Tropospheric ozone; Temporal trend; Logistic regression; Generalized linear models; Generalized additive models; Experimental design

1 INTRODUCTION

Since the beginning of the industrial era the quantity of pollutants rejected into the ambient air has considerably increased. This pollution, because of its local and planetary impact, explains the need for individual and collective preventative actions. Because the Parisian area, with more than ten million inhabitants, represents one of the largest urban agglomerations of Europe, early attention was turned to air pollution problems. As occurs in all great cities, Paris has a serious photochemical tropospheric ozone (O_3) air pollution problem. As the urban emission pattern of O_3-forming pollutants is fairly uniform, it seems that the variations of O_3 concentrations are controlled by meteorological factors: principally, exposure to sunshine, temperature and wind speed (ventilation and transport), and by a series of atmospheric reactions involving precursor pollutants. These precursors include a variety of volatile organic compounds, including for the most part non-methane hydrocarbons (HC), nitric oxide (NO) and nitrogen dioxide (NO_2). The HC and NO_2 are both emitted from transportation and industrial processes, whereas other volatile organic compounds are emitted by cars and, principally, facilities using chemical solvents.

* E-mail: lise.bellanger@math.univ.nantes.fr
† Corresponding author. 8, rue de l'Eglise, Le Bourg, 45210 Chevry-sous-Le Bignon, France; E-mail: rr.tomassone@wanadoo.fr

ISSN 0233-1888 print; ISSN 1029-4910 online
DOI: 10.1080/02331880410001696116

It has long been known that tropospheric ozone is an air pollutant that can have health and environmental effects (Michaelis, 1997), including respiratory problems, forest retardation and crop injury. In addition, during the last decade one has seen a growing diversity of statistical literature on ozone pollutant topics with the application of a wide range of statistical methodologies categorized under three broad statistical approaches: regression-based modelling, extreme value approaches and space-time models (see for example Thompson *et al.*, 2001 and Nychka *et al.*, 1998 for a review of statistical methods and problems). This article describes the extreme value analysis of tropospheric ozone level exceedances collected at seven monitoring sites in the Paris metropolitan area over the period May–September 1988–2001. The purpose of the study is to establish whether observed trends over time in ozone exceedances of a high threshold are real or if they are the result of meteorological changes affecting the conditions under which ozone is generated.

Extreme value theory has developed rapidly in recent years and become widely used in many disciplines. Many monographs have been published; for example, Leadbetter *et al.* (1983), Falk *et al.* (1994), Embrechts *et al.* (1999), or more recently Reiss and Thomas (2001) and Coles (2001). A number of possible approaches to modelling extreme values in a population may be possible, depending on the structure and complexity of the data. They could be clustered in three broad categories:

- *Classical extreme value models (study of the annual maxima or of the kth largest maxima).* If the sequences are fairly long, the classical method consists of treating maxima of consecutive periods of equal length (for example years) of the series as independently and identically distributed belonging to one of the extreme value distribution families (Gumbel, Fréchet and Weibull) that can be combined into a single family of models called the generalized extreme value (GEV) of distributions. But this method has one big drawback: it is a wasteful method if other data on extremes are available.
- *Threshold models.* The 'peaks over threshold' (POT) method consists of choosing a high threshold and modelling the process of exceedances of it, conditionally upon the chosen level. The natural parametric family of distributions for such exceedances is the generalized Pareto family (Pickands, 1975).
- *Point process approach of extremes.* Point process techniques give insight into the structure of limit processes that occur in extreme value theory. It leads to nothing new, but enables a more natural formulation of non-stationarity and dependence, for example. However, as no general theory is currently established for non-stationary processes, it is usual, in practical extremes problems, to use standard extreme value models depending on time.

The article is organized as follows. In Section 2 the data sources and the framework for the statistical analysis are described. In Section 3 we describe models for extreme values, and more specifically, previous models for air pollution data. We present our model, details of our analysis and conclusions in Sections 4 and 5.

2 THE DATA

The data considered in this article were collected at seven monitoring sites among the leading Paris area network of monitors, located at a certain distance from the emission pollution sources, quantifying geographically air pollution and provided to us by AIRPARIF (organization in charge of supervision of air pollution in the Paris area). The data consist of O_3 exceedances over $130\,\mu g\,m^{-3}$ observed on these sites, during the months of May–September over the 14-year period 1988–2001, together with daily values of meteorological variables described

TABLE I Monitoring sites status for tropospheric ozone (O_3).

Name of site	*Code*	*Observation*	*Altitude of measurements (m)*
Neuilly	NE	Suburban (westward)	3–5
Aubervilliers	AU	Suburban (northward)	3–5
Champs-sur-Marne	CH	Suburban (eastward)	3–5
Créteil	CR	Suburban (eastward)	3–5
Rambouillet	RA	Country (westward)	18
Paris (district 7)	P07	Urban	3–5
Paris (district 13)	P13	Urban	13

below. May–September months are considered to be the period of the year where most of the high values of O_3 occur ('high ozone' season). Separate modelling of the association between ozone level and meteorological variables is the simplest one, but it ignores any information on regional dynamics of ozone and meteorology available in the analysis. Consequently, to find a balance between interpretation, simplicity of approach, and increasing of statistical power, the monitoring sites (described in Table I) will be used as control variables. The monitoring sites are of two types:

- Urban or suburban sites (NE, AU, CH, CR, P07 and P13) permit the measurement of basic pollution and are representative of a surrounding area. Their location is far from direct pollution sources.
- Country sites (only RA) are used to gain some insight about pollution transfer due to wind.

The altitude of measurement differs from one monitoring site to another (for example, ozone is registered at 18 m for RA, at 13 m for P13, but only at 3–5 m for the other sites). These differences may introduce a confounding effect between sites and measurement altitudes, as we shall see later.[1]

For our study, we dispose of, in all, 10,688 measurements of O_3.

A good graphical representation of site differences is given by a boxplot representation (Fig. 1).

Ozone (O_3) is a secondary pollutant, photochemically produced. The variability in surface ozone concentration levels is affected by the strengths of sources and precursor emissions, as well as by meteorological conditions, which are inherently variable. To assess that part of trend in ozone exceedances over a high threshold that cannot be accounted for by meteorology, and would indeed perhaps characterize trend over time, we need to build a model which relates ozone to meteorology.

The variables used to perform the analysis, provided by Saclay Observatory, are given in Table II.

We choose a small number of meteorological variables known to have a great influence on O_3 generation. In fact, when all values of ozone are taken into account, effects of these covariates on ozone level are clearly visible, for example:

– the expected effect of higher Tmax is to increase ozone levels;
– the expected effect of higher Trange is to increase ozone levels;

[1] Further information on measurements may be obtained on AIRPARIF site http://www.airparif.assu.fr

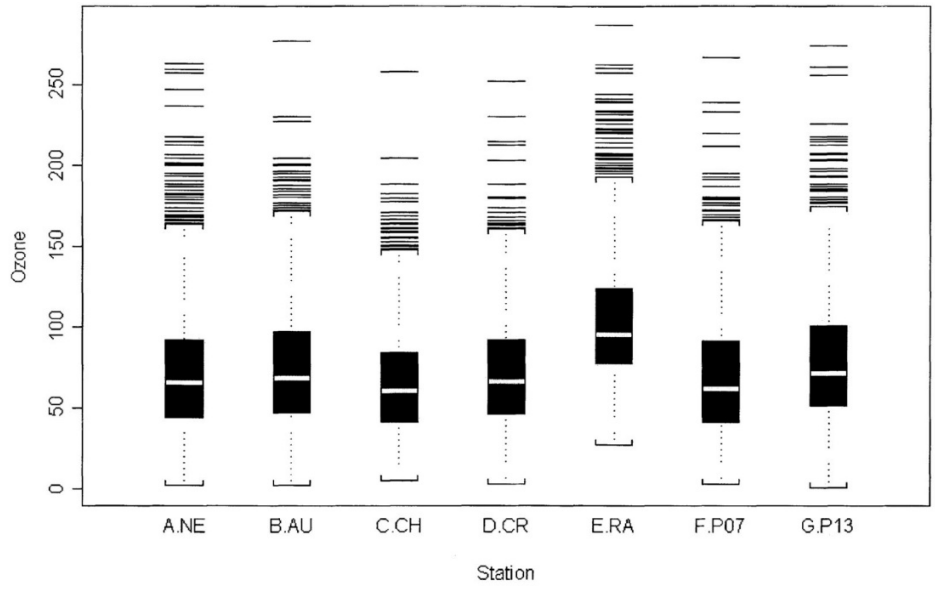


FIGURE 1 Boxplot graphs of ozone values for the seven sites.

– the expected effect of increased Wind is to reduce ozone levels because higher wind speed tends to disperse pollutants present in the air.

But when we consider only the subset of data corresponding to ozone values exceeding $130\,\mu g\,m^{-3}$, it is more difficult to detect clear relationships. Some scatterplots of the data are contained in Figure 2.

Figure 3 shows the curve obtained by fitting a local regression model with one predictor t (temporal trend) to ozone data using non-parametric regression procedures. The form of a possible trend seems more apparent than in Figure 2 and the appearance of the data subset corresponding to ozone values exceeding $130\,\mu g\,m^{-3}$ suggests the following: to assess a temporal trend, it may be important to introduce a more complex temporal change than a linear one.

Figure 4 shows that using local regression models for fitting one predictor (Tmax, Wind and Trange) at a time to ozone data leads to the same observations, and also brings to light the classic co-relationship between these covariates and ozone concentration. In the following, surely non-linear effects and/or interaction between predictors have to be taken into consideration, specifically for wind speed.

TABLE II Variables (dependent, control, predictor) used in the study.

Variable	*Code*	*Status*	*Unit*
Daily maximum ozone level	Ozone	Dependent	$\mu g\,m^3$
Daily maximum temperature	Tmax	Predictor	°C
Daily minimum temperature	Tmin	Predictor	°C
Daily temperature range (maximum–minimum)	Trange	Predictor	°C
Daily average wind speed	Wind	Predictor	$m\,s^{-1}$
Year	t	Predictor	From 1 (1988) to 14 (2001)
Monitoring site	Station	Control	

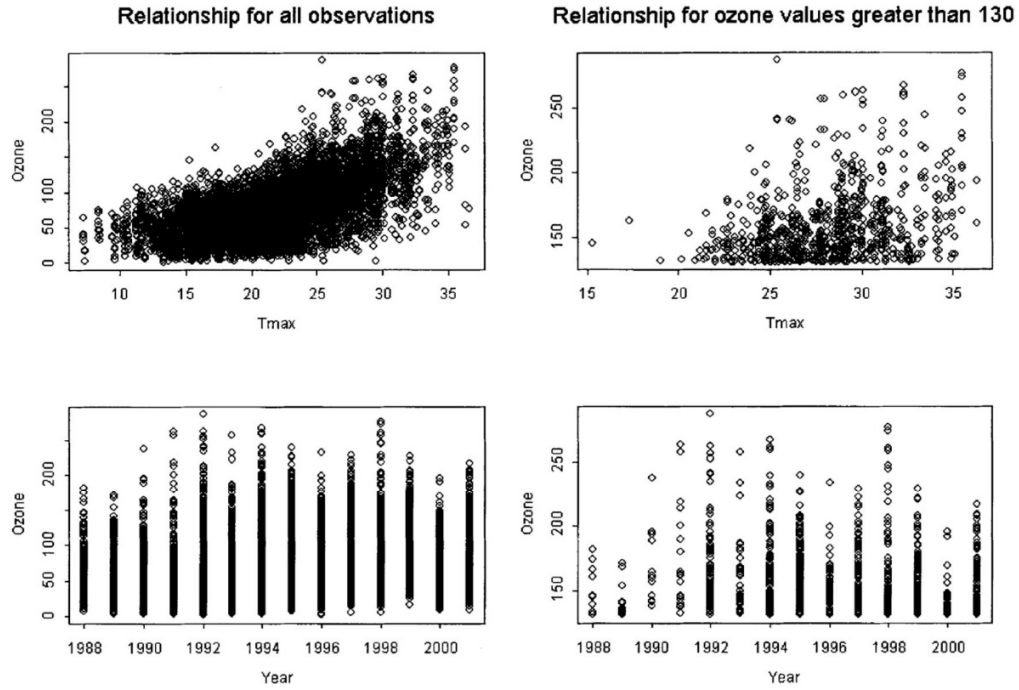


FIGURE 2 O_3 *vs*. Tmax (upper row) and t (lower row) all observations, and ozone levels higher than the chosen threshold.

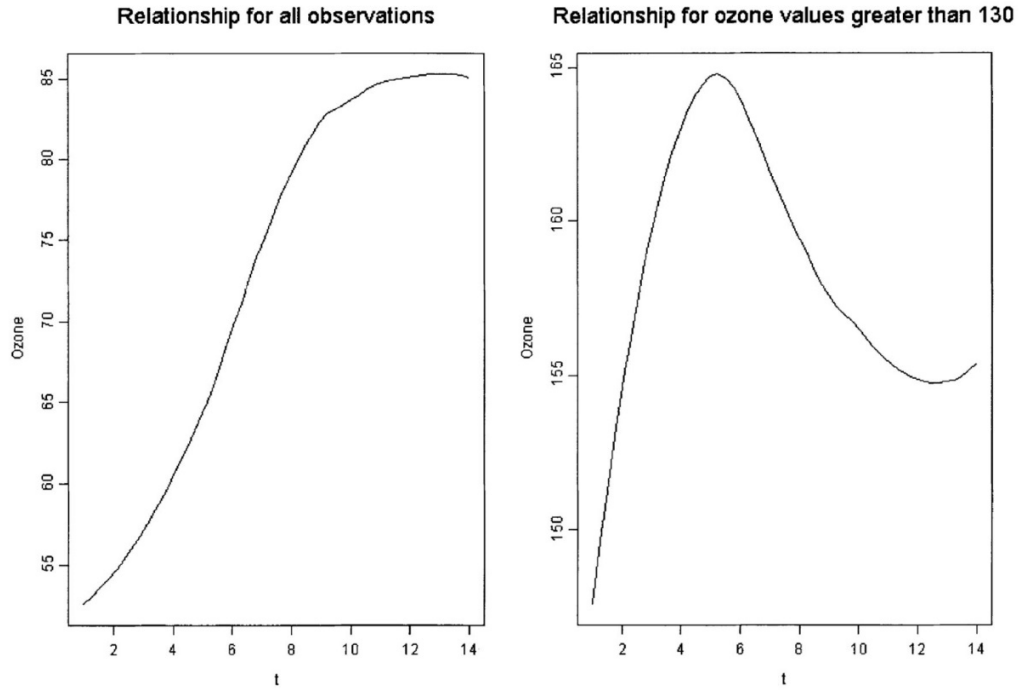


FIGURE 3 Local regression model with one predictor t for ozone levels higher than the chosen threshold.

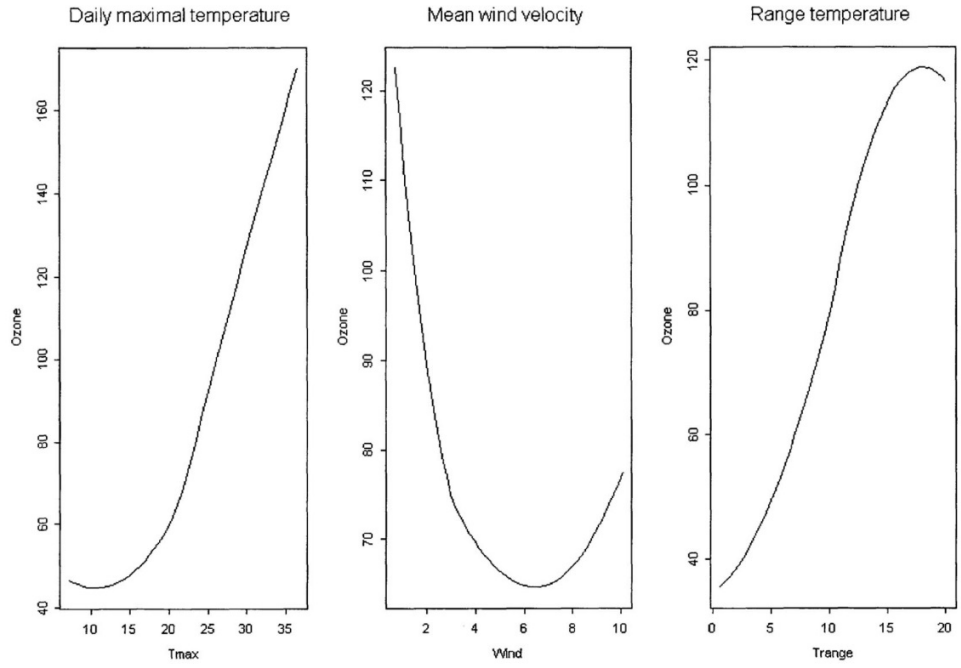


FIGURE 4 Local regression models with one predictor Tmax, Wind and Trange.

3 MODELS FOR EXTREME VALUES

In this section, we first review the foundations of classical extreme value theory and models, then we describe more general models, taking into account seasonality or dependence or covariates. The reader seeking more details on extreme values theory may refer to specialized books, as indicated in the Introduction.

3.1 Models in the Independent and Identically Distributed Case

When $Y_1, \ldots, Y_n$, is a sequence of n independent random variables having a common distribution function F, the model focuses on the statistical behaviour of $M_n = \max(Y_1, \ldots, Y_n)$. The following result (see for example Embrechts *et al.*, 1999) is the key of the extreme value analysis.

If there exists two sequences of constants $\{a_n > 0\}$ and $\{b_n\}$ such that:

$$P\left[\frac{(M_n - b_n)}{a_n} \leq y\right] = (F(a_n y + b_n))^n \underset{n}{\longrightarrow} H(y), \tag{1}$$

where H is a non-degenerate distribution function, then H belongs to the GEV family of distributions, having the form:

$$H(y; \mu, \xi, \sigma) = \exp\left\{-\left[1 + \xi\left(\frac{y - \mu}{\sigma}\right)\right]^{-1/\xi}\right\} \tag{2}$$

defined on the set $\{y\colon 1 + \xi(y - \mu)/\sigma > 0\}$, where $\mu \in R$, $\sigma > 0$ and $\xi \in R$. The case $\xi = 0$ has to be treated separately.

Many techniques exist to estimate the three parameters (μ is a location parameter, σ a scale one and ξ a shape one). But likelihood-based techniques are the most flexible ones, particularly in the presence of covariates. Maximum likelihood estimators exist in large samples, provided that $-1 < \xi \leq -0.5$ and are asymptotically normal if $\xi > -0.5$ (Smith, 1985). The case $\xi \leq -0.5$ rarely arises in practice (Davison and Smith, 1990), so this theoretical limitation does not seem to be an obstacle. Finally, this model could be extended to other extreme order statistics, and more generally, to the k-largest order statistics: Smith (1984; 1986) used this approach to study hydrologic data, keeping in mind the five largest order statistics.

Threshold methods allow one to take into account all available data on extremes and is also potentially a better alternative to the previous method. Let $Y_1, \ldots, Y_n$ be a sequence of n independent random variables having a common distribution function F. It is also possible to consider as extreme events the events $\{Y_i > u\}$, where u is some fixed high threshold. The stochastic form of these events is given by:

$$P[Y > u + x / Y > u] = \frac{1 - F(u + x)}{1 - F(u)} > 0. \tag{3}$$

If the common distribution function F were known, the distribution of threshold exceedances in Eq. (3) would also be known. But in practice, this is generally not the case, so we have to use a result (Pickands, 1975) allowing approximation for high values of the threshold. This result says that for u sufficiently high, the distribution of $\{Y - u\}$ given that $\{Y > u\}$ is approximately

$$G(x; \tilde{\sigma}, \xi) = 1 - \left[1 + \frac{\xi x}{\tilde{\sigma}}\right]^{-1/\xi} \tag{4}$$

defined on the set $\{x: x > 0 \text{ and } 1 + \xi x/\tilde{\sigma} > 0\}$, where $\tilde{\sigma} = \sigma + \xi(u - \mu)$.

G belongs to the *family of generalized Pareto distribution* (GPD). The case $\xi = 0$, interpreted by taking $\xi \to 0$ in Eq. (4) corresponds to an exponential distribution with parameter $1/\tilde{\sigma}$, widely used in studies using the POT method. Using likelihood-based techniques, we get analogous results to those for the GEV distribution concerning the regularity conditions that are required for the usual asymptotic properties associated with the maximum likelihood estimator to be valid.

There are many papers concerning this approach. Davison and Smith (1990) and Leadbetter (1991) describe it in detail; it is applied to a wide variety of domains such as hydrology or air pollution (see for example Smith, 1984; Davison, 1984; Hosking and Wallis, 1987; Heffernan and Tawn, 2002).

It is also possible to characterize extremes by a *point process*. For details on point process theory, we refer for example to Cox and Isham (1992) and Kallenberg (1983). Point process techniques give insight into the structure of limit processes that occur in extreme value theory; it leads to nothing new, but enables a more natural formulation of non-stationarity and dependence. If we assume $Y_1, \ldots, Y_n$ to be a sequence of n independent random variables having a common distribution function F and satisfying Eq. (1), then the fundamental result (Pickands, 1971) states the following.

Letting z_- and z_+ be the lower and the upper endpoints of H, respectively, the sequence of point processes N_n defined on R^2 by $N_n = \{(i/(n+1),\ (Y_i - b_n)/a_n : i = 1, \ldots, n\}$ converges in distribution on regions of the form $]0, 1[\times]u, \propto[$ for $u > z_-$, to a *Poisson process* whose intensity measure is the product of Lebesgue measure and of that defined by $\ln(-H(\cdot))$ such

that for any region of the form $A = [t_1, t_2] \times [z, z_+[$ with $= [t_1, t_2] \subset]0, 1[$, it is given by

$$\Lambda(A) = (t_2 - t_1)\left[1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right]^{1/\xi}.$$

As we shall also see in more detail in Section 3.3, early versions of the POT method assumed a non-homogeneous Poisson process (NHPP) to model the times of exceedances over high thresholds, combined with independent generalized Pareto random variables to represent the sizes of the exceedance, assuming that the threshold is exceeded.

3.2 More General Models Taking into Account Seasonality, Dependence or Covariates

To assume that the underlying process consists of a sequence of independent random variables is usually unrealistic. The data could exhibit seasonality or short-range dependence leading to the clustering of high-level exceedances (for example hot days tending to occur together). For the handling of seasonal data, there are two broad approaches:

- The 'separate seasons' approach, in which the year is partitioned into a finite number of independent seasons with separate models for each one (Smith, 1989 for ozone data).
- The pre-whitening model, in which known seasonal components are removed before carrying out the extreme value analysis.

While the theory of extreme values in dependent stochastic processes is not developed here, a precise development is given in Leadbetter *et al.* (1983). We will only present here a few results important for understanding the methods used to resolve practical problems. For stationary processes, most of the results of this theory are of the form that, under suitable mixing conditions (see Leadbetter *et al.*, 1983), which cover a very wide range of processes, the high level excesses behave asymptotically as a clustered Poisson process. For the POT method, the usual practical solution is to fit the point process model to the *cluster* maxima. The following general relation defines a key parameter θ for such models, which is a measure of the amount of clustering in the process and is called the *extremal index*:

$$P[M_n \leq y] \approx [F(y)]^{n\theta},$$

where $0 < \theta \leq 1$ ($\theta = 1$ being the case of no clustering and $1/\theta$ is the limiting mean cluster size).

As we may find in Coles (2001, p. 177–178):

> 'The declustering is an ad hoc device to circumvent the difficulty that the joint distribution of successive threshold excesses is unspecified by the general theory. [$\cdots$]. But there are disadvantages to this approach: cluster identification is often arbitrary; information on extremes is discarded; and the opportunity for modelling within-cluster behaviour is lost'.

This is a very important aspect in our application since we are interested in trend over time in ozone exceedances of a high threshold. An alternative to declustering could be found by strengthening the assumptions on the process compared to the stationarity. If one had considered $Y_1, \ldots, Y_n$ as a stationary first-order Markov chain, this assumption would have led to models that could better represent temporal trend, but it would always be a simplification of data dependence. For further reading on this topic, see Smith *et al.* (1997) or the book of Coles (2001) containing references (end of chapter 9).

In the case of non-stationary series, there is no general theory, or the results are very difficult to apply to real data (see for example, Hüsler, 1986; Coles, 2001 (Chapter 6 where

other references are given in 6.4); Bellanger and Perera, to appear). In practical problems, variations over time in the observed process are often expressed in terms of a linear trend in the location parameter of the appropriate extreme value model, or as more complex changes, such as a quadratic model or a change-point one, but they may also be expressed in terms of the other extreme value parameters.

A major aspect of statistics is the explanation of systematic variation in a response variable in terms of covariates. Combining several data sets using covariates improves the fit of the model too. The most successful and widely used technique for this is the generalized linear model as presented in Dobson (2002) and McCullagh and Nelder (1989).

3.3 Threshold Selection: An Important Point is the Choice of the Threshold u

In the case of the threshold model, the issue of threshold choice implies a balance between bias and variance. In practice, a threshold as low as possible, subject to the limit model providing a reasonable approximation, is generally adopted. Two methods are available: the first one is an exploratory one based on the mean of the GPD, and uses the well-known *Mean excess plot*, also termed the *mean residual life plot*; the second one is an assessment of the stability of parameters estimates, called *threshold invariance property* (if the distribution function of $\{Y - u\}$ given that $\{Y > u\}$ is a GPD then for any threshold $v \geq u$ the distribution function of $\{Y - v\}$ given that $\{Y > v\}$ is also a GPD) (see for more details Smith, 1989; Davison and Smith, 1990; Embrechts *et al.*, 1999; Coles, 2001).

Other important topics related to extreme value theory but not relevant to our study are not tackled in this article. These topics include the extension to multivariate and spatial cases. Extensions of the POT model (Davison and Smith, 1990) and of the point process model (Pickands, 1971; Smith, 1989) have been proposed for multivariate extremes. Reviews of the statistical aspects of multivariate extreme value modelling are given by Coles and Tawn (1991) and Joe *et al.* (1992), for example. As with the univariate approaches, application of these methods assumes that asymptotic results hold exactly in some suitable joint tail region. Therefore, these models are less fully prescribed by the general theory. But the main issue arising from their use is the problem of dimensionality creating difficulties for computation and model validation. Our modelling approach in the study of the temporal trend in high ozone levels in the Paris area is to apply standard univariate techniques to model extremes of the ozone series registered at the seven monitoring sites.

4 MODELLING APPROACH

4.1 Previous Extreme Value Models in Ozone Time Series Used to Study Temporal Trend in High Levels

To detect a temporal trend, two main approaches exist:

- The first one considers the temporal trend in the absolute size of the annual maximum value of ozone (or of the k-largest values that occur in a year). If the size of the considered k-largest order statistics is decreasing through time, then one concludes a downward trend in the extreme values of ozone (Smith, 1989; Shively, 1990). But as stated before, this approach is wasteful in data and previously quoted papers do not take into account the effects of meteorological conditions on ozone levels registered.

– The alternative approach uses the characterization of extremes with a point process (Shively, 1991; Smith and Shively, 1995). Because of the problem of clustering of high-level ozone exceedances, the most successful methodology models the bi-dimensional point process of exceedance times and sizes of a high threshold, occurring through time as an NHPP whose parameters depend on both time and meteorology. The management of this model implies:
  - choice of a reasonable threshold;
  - selection and estimation of the parameters corresponding to the covariates included in it;
  - diagnostic tests to check that the model assumptions are satisfied. There are two types of assumptions to check: those concerning the use of NHPP, and also those regarding the distribution and independence of the exceedance sizes.

But this process tends to be cumbersome because it begins by choosing a seemingly reasonable threshold that could turn out to be inappropriate after the step of checking independence and distribution assumptions.

Consequently, we decided to adapt this second approach by developing an original process that was less wasteful in time and took into account trends more complex than a linear one such as the change-point model.

4.2 Our Approach

4.2.1 The Model

In this section, we develop the NHPP used in modelling trends in tropospheric ozone based on exceedances of a high threshold u, whose value must be fixed (Bellanger and Tomassone, 2000). Here, we will only raise the specifics of our methodology. We begin by setting:

$$\Psi_i(y) = \begin{cases} P(Y_i > y) & \text{if day } i \text{ is not missing} \\ 0 & \text{if day } i \text{ is missing.} \end{cases}$$

The probability distribution of the random variable Y_i (daily maximum ozone day i) is $1 - \Psi_i(y)$ and its density function takes the following expression: $f_i(y) = -(\mathrm{d}/\mathrm{d}y)[\Psi_i(y)]$.

Let us consider that the process is observed over a time period $]0, T[$. Peaks over threshold u are represented by (T_i, Y_i), $1 \leq i \leq N$, where T_i and Y_i are supposed to be independent and N, the total number of peaks, is also a random variable. That is, the ith observed peak occurs on day t_i and takes the value $y_i \geq u$, $1 \leq i \leq n$.

We supposed that $t_1, \ldots, t_n$ are the realizations of a Poisson process on $]0, T[$ with intensity function $\Psi_t(u)$. If the density function of the exceedance size (setting $X_i = Y_i - u$), given that $Y_i \geq u$, is $f_{t_i}(y)/\Psi_{t_i}(u)$, using the Poisson property, then the approximate joint density of the observed peaks is:

$$\begin{aligned} L(t_1, \ldots, t_n, y_1, \ldots, y_n) &= L_1(t_1, \ldots, t_n) L_2(y_1, \ldots, y_n) \\ &= \left[\exp\left\{ -\int_0^{t_1} \Psi_t(u)\, \mathrm{d}t \right\} \left(\prod_{i=1}^{n-1} \Psi_{t_i}(u) \exp\left\{ -\int_{t_i}^{t_{i+1}} \Psi_t(u)\, \mathrm{d}t \right\} \right) \right. \\ &\quad \left. \times \Psi_{t_n}(u) \exp\left\{ -\int_{t_n}^{T} \Psi_t(u)\, \mathrm{d}t \right\} \right] \left[\prod_{i=1}^{n} \frac{f_{t_i}(y_i)}{\Psi_{t_i}(u)} \right]. \end{aligned}$$

From which we get,

$$L(t_1, \ldots, t_n, y_1, \ldots, y_n) = \left[\left(\prod_{i=1}^{n} \Psi_{t_i}(u) \right) \exp\left(- \int_0^T \Psi_t(u)\, \mathrm{d}t \right) \right] \left[\prod_{i=1}^{n} \frac{f_{t_i}(y_i)}{\Psi_{t_i}(u)} \right], \quad (5)$$

where the first term in square brackets models the exceedance times of the threshold level u and the second one the sizes when an exceedance has occurred.

In the following, for notational purposes t_i is replaced by i.

Model for the Exceedance Dates Over Threshold u. We consider that the number of exceedances is distributed as an NHPP. There exist many possibilities to model the intensity function $\Psi_i(u)$ of the Poisson process that take into account the relationship between meteorological conditions and the frequency of the exceedances (see for an overview of parametric and non-parametric models for the intensity function of an NHPP to test for a trend Vaquera-Huerta *et al.*, 1997). For example, Shively (1991), Smith and Shively (1995) and Chavez-Demoulin (1999) use the exponential model for the intensity introduced by Cox (1955). In Chavez-Demoulin (1999), tests and interval estimates for parameters are then obtained by generalized linear model techniques with Poisson errors (p. 78). In our study, we consider a logistic regression model for threshold exceedance. A mathematical survey and practical examples of the logistic regression could be found in Hosmer and Lemeshow (2000); application to modelling high threshold exceedances of ozone with a logit model in for example Smith and Huang (1993) or Bellanger and Tomassone (2000). We decided to choose a logistic regression model because we need a simple one to describe a regular growth able to detect a trend. Other concurrent models, when we tried them, did not give better results. The logistic one presents a great flexibility; the interpretation of its parameters is easy; the ease in its parametrization also permits us to detect differences in station effects; its implementation in many systems including S-Plus is also a practical advantage (Venables and Ripley, 1999; Chambers and Hastie, 1992). Even if it is clear that it has no physical justification, it is just the simplest able to describe a complex situation. Indeed, the data consist of a sequence of ones and zeros, corresponding to exceedance or non-exceedance, respectively; and we suppose that the probability of exceeding the threshold u on day i is given by

$$\Psi_i(u) = \frac{\exp(\alpha(i))}{1 + \exp(\alpha(i))} = P[Y_i > u, \text{ day } i]. \quad (6)$$

The simple functional form of the logit link $\alpha(\cdot)$, that defines the classical multiple logistic regression model considered, is given by:

$$\alpha(i, \text{sta}) = \alpha_0 + \alpha_1 t(i) + \text{Station}_{\text{sta}} + \sum_{j=2}^{p} \alpha_j w_j(i) + \sum_{j=2}^{p} \alpha_{1j} t(i) w_j(i) + \sum_{k,j=2}^{p} \alpha_{kj} w_k(i) w_j(i), \quad (7)$$

where

$$\text{Station}_{\text{sta}} = \sum_{l=2}^{7} \gamma_{\text{sta},l} D_{\text{sta},1}.$$

Where $t(i)$ is the year – 1987 – in which observation i occurred, $w_j(i)$, $j = 2, \ldots, p$ are the meteorological covariates, $D_{\text{sta},1}$ denotes the design variables associated to the polychotomous independent variable monitoring sites and $\gamma_{\text{sta},1}$ denotes the coefficient for these design variables (sta = NE, AU, CH, CR, RA, P07 or P13). Interactions are also added by creating variables that are equal to the product of the value of the considered covariates (for example t

and w_j, or w_k and w_j. Furthermore, we decided to specify the design variables for monitoring sites using reference cell coding with NE as the reference group (see Tab. III):

The purpose of our study is to detect if a temporal trend is present, either a decreasing or an increasing one. In model formulation (7) the term t was introduced to detect it. In previous papers (Bellanger, 2001; Bellanger and Tomassone, 2000) we found that associating t and the interaction t^* Tmax was interesting. But this procedure prevents us from detecting an easily interpretable temporal trend. This result is due to the fact that we had only considered a generalized linear model. Consequently, to improve the modelling, we try to model the intensity function $\Psi_t(u)$ as given in Eq. (6) by a logistic additive model (Hastie and Tibshirani, 1990). Indeed, we replace the generalized linear model by an additive one:

$$\alpha(i, \text{sta}) = \alpha_0 + \text{Station}_{\text{sta}} + a_1(t) + \sum_{j=2}^{p} a_j(w_j(i)). \tag{8}$$

Equation (8) specifies a semi-parametric model in which the factor Station is to appear linearly and the meteorological and temporal covariates are to be modelled by a non-parametric smooth term. The $a_j(\,)$ functions represent smooth terms to be fitted using smoothing splines as smoothers. Their use permits a direct visual description of trends in the resulting of plot of response vs. predictors. Concerning statistical inference, we deal with a sufficiently large sample to use the deviance as a tool to assess models and to compare different ones. For much more discussion and appropriate definitions of deviance and degrees of freedom for the general semi-parametric models, see for example Hastie and Tibshirani (1990), Davison and Hinkley (1997) and Chavez-Demoulin (1999). Considering that data themselves must choose smoothing parameter, we use an automatic procedure (as a default amount of smoothing proposed in the $s(\,)$ function in S language). The stepwise model selection procedure, widely used in the linear model, is performed to build the model. But as one of our goals is an exploratory one, associated with extrapolation for future, this additive model will be used as an exploratory tool to detect more easily non-linearity and to suggest relationships for terms in the model (parametric transformations or alternative forms). Then we will build a new linear model, with simpler interpretation, and perform tests to compare these two previous models, by analysing the change in deviance relative to the change in degree of freedom using the likelihood ratio test.

Finally, assuming that the limiting Poisson process on $]0, T[$ of intensity function given by Eq. (6) is an acceptable approximation of the process of the times of exceedances over u, and given the occurrence of one event, the expected waiting time until the next event could be calculated using the density function for the time to the next event (from r) given by:

$$k_r(z) = \exp[-h(r+z) - h(r)]\Psi_{r+z}(u), \tag{9}$$

TABLE III Specification of the design variables for monitoring sites (Site), using reference cell coding with NE as the reference group.

	Design variables					
Station	D_{AU}	D_{CH}	D_{CR}	D_{RA}	D_{P07}	D_{P13}
NE (1)	0	0	0	0	0	0
AU (2)	1	0	0	0	0	0
CH (3)	0	1	0	0	0	0
CR (4)	0	0	1	0	0	0
RA (5)	0	0	0	1	0	0
P07 (6)	0	0	0	0	1	0
P13 (7)	0	0	0	0	0	1

where

$$h(r) = \int_0^r \Psi_t(u)\,\mathrm{d}t.$$

Model for the Exceedance Sizes over Threshold u. Theoretical results on the probability distribution of the exceedances sizes over u (noted $X = Y - u$), summarized in Section 3, allow us to approach for u sufficiently large, the probability distribution of $X = Y - u$ given that $\{Y > u\}$ by a GPD and:

$$P[X_i > x/Y_i > u] \approx 1 - G\left(x; \beta\left(i\right), \xi\left(i\right)\right) = \left[1 + \xi\left(i\right)\beta\left(i\right)x\right]^{-1/\xi(i)}$$

with $x = y - u$ and the parameter $\beta(i)$ squares to $1/\tilde{\sigma}\ (i)$ in Section 3.

As for the parameter $\alpha(\cdot)$ of the model for exceedance dates defined in Eq. (7), we assume that $\beta(\cdot)$ the first time takes the following simplest functional form, that is the linear one:

$$\beta(i, \text{sta}) = \beta_0 + \beta_1 t(i) + \text{Station}_{\text{sta}} + \sum_{j=2}^{p} \beta_j w_j(i) + \sum_{j=2}^{p} \beta_{1j} t(i) w_j(i) + \sum_{k,j=2}^{p} \beta_{kj} w_k(i) w_j(i). \tag{10}$$

Moreover, to simplify we assume that only $\beta\ (\cdot)$ takes into account meteorological changes and temporal trends; $\xi\ (\cdot)$ is also supposed to be constant. Then, for u sufficiently large, the density function for X can be approached by:

$$g\left(x; \beta\left(i\right), \xi\right) = \beta\left(i\right)\left(1 + \xi\beta\left(i\right)x\right)^{-((1/\xi)+1)}. \tag{11}$$

In addition, as discussed in Section 3, the case where $\xi = 0$ corresponds here to the exponential distribution with parameter $\beta(i)$, and the expected size of an exceedance given that an exceedance occurring day i is $E\left[X_i/Y_i > u\right] = 1/\beta\left(i\right)$.

As for the exceedance dates, we build a generalized additive model as an exploratory tool that allows us to keep the non-linearities and to incorporate those relationships suggested by the tool in a generalized linear model of the form given by Eq. (10).

4.2.2 Modelling Procedure

We must keep in mind that the basic assumptions needed to apply the point process approach in our case are the following:

- Exceedance sizes and times are independent of each other.
- The excesses over u occur at the times $T_i, i = 1, \ldots, n$ modelled by an NHPP with intensity function $\Psi_{i(u)}$ given by Eqs. (6) and (7). Consequently, if this assumption is satisfied, the inter-event times $S_i = T_i - T_{i-1}$ are independent and have the distribution implied by Eq. (9).
- The corresponding sizes of exceedance are independent and have a GPD distribution, the density function of which is given by Eq. (11).

So we adopt the following modelling procedure, divided into three main stages:

Threshold Selection. We use the bootstrap to test hypotheses that allow us to obtain reasonable threshold u that checks the independence hypotheses of the inter-event times S_i and also of the exceedance sizes $X_i = Y_i - u$. For this, we test whether the correlation between adjacent inter-event times is null and also whether the correlation between exceedance sizes that occur on consecutive days is null. Using the relationship between confidence intervals and hypothesis tests (the 95% percentile confidence intervals are the sets of the plausible values of the respective correlations, having observed the estimated correlations (Efron and Tibshirani, 1993)), we calculate an estimate and a 95% percentile confidence interval based on percentiles of the bootstrap distribution of the statistic (Efron and Tibshirani, 1993; Davison and Hinkley, 1997) at different fixed thresholds u, for the two previous correlations. In this way, for a given threshold u, if the two confidence intervals contain the value 0, we decide that we cannot reject the hypothesis that correlations are 0 due to independence. In this case, we consider the previously considered threshold as sufficiently large to continue the extreme value analysis. For details on resampling techniques, we refer for example to Shao and Tu (1996) where a more mathematical survey is given and to Efron and Tibshirani (1993) or Davison and Hinkley (1997) where further details and many examples could be found.

There are some points to make in this regard. First of all, in practice it is difficult to apply more sophisticated time-series tests for independence because of the complicated nature of the model for the exceedance times and inter-event times. Second, we compute the correlation between exceedance sizes that occur on consecutive days because if these excesses are unrelated we could infer that excesses more than one day apart will also be unrelated (Smith and Shively, 1995).

Estimation and Selection of the Parameters of the Models. In terms of fitting statistical methodologies, several recent contributions have been made: Coles and Dixon (1999), for example, argue the superiority of the fitting procedure based on the probability-weighted moments for the small sample; Davison and Ramesh (2000) outline a semi-parametric approach based on local polynomial fitting; Hall and Wellner (1981) develop a non-parametric approach to estimating trends when fitting parametric models to extreme values from windstorm severity and maximum temperature series; Chavez-Demoulin (1999) combines the point process for exceedances with smoothing methods based on penalized log-likelihood criteria to model changes in large values for minimal temperatures registered at different sites in Switzerland from 1971 to 1997.

But in our case, likelihood-based methods have many practical and theoretical advantages. They are in most applications of extreme value modelling regular in the usual sense of likelihood theory, meaning that the standard asymptotic likelihood results are applicable (Smith, 1985). Moreover, as we stated before, likelihood functions can be constructed for complex modelling situations, enabling regression modelling. In addition, for both multivariate modelling approaches described previously, generalized linear theory (McCullagh and Nelder, 1989; Dobson, 2002; Hosmer and Lemeshow, 2000) was explored considering only estimation by maximum likelihood produced by iteratively re-weighted least squares (IRLS). The selection or deletion of covariates from the models is based on the stepwise procedure widely used in linear regression. The t-values (estimated coefficients divided by their asymptotic standard errors) are used to test whether the coefficients are zero and therefore to check the partial importance of each variable included in the models. The likelihood ratio test is used to compare some models and to fit a reduced model containing only those variables thought to be significant. We obtain at the end of this step a *preliminary main effects linear model*. Prior to this step, since we want to look more closely at the continuous covariates in the models by checking the assumption of linearity in the link function, we build generalized additive models (GAM) of the type discussed by Hastie and Tibshirani (1990) to model possible non-linear changes. The plots of the fits of the generalized additive model suggest the analytical form of

the relationships that are useful to be refitted, without the loss of too much precision, as a generalized linear one. Once we ascertain that each of the continuous covariates is scaled correctly and obtain the final generalized linear models, after comparing these new models to the old one using the likelihood ratio test, we then assess the adequacy and check the fit of the final generalized models using classical measures of goodness-of-fit (McCullagh and Nelder, 1989; Dobson, 2002; Hosmer and Lemeshow, 2000) and bootstrap procedures. Several strategies are possible, depending on the context. Shao and Tu (1996) suggest for the generalized linear model resampling the residuals, *i.e.*, a *model-based resampling*. Indeed, Davison and Ramesh (2000) base the bootstrap assessment of a local polynomial fit on Studentized quantities. These are parametric approaches. But Davison and Hinkley (1997, Section 6.2.4) wrote that

> ... parametric simulation for a generalized linear model involves simulating new sets of data from the fitted parametric model. It has the usual disadvantage of the parametric bootstrap, that datasets generated from a poorly fitting model may not have the statistical properties of the general data. This applies particularly when count data are overdispersed relative to a Poisson or binomial model, unless the overdispersion has been modelled successfully.

Indeed, in our case (a large data set with lack of precision for some covariates) we adopt a completely non-parametric approach (*case-based resampling*) corresponding to a randomly weighted generalized linear model.

Diagnostic Checks. After fitting the parameters of the models, we need to check assumptions regarding the distributions hypothesis previously assumed (Section 4.2.2). For this we used probability plots (Snedecor and Cochran, 1971; Lawless, 1982; Draper and Smith, 1981) and the Kolmogorov–Smirnov test. The distribution function of the inter-event times $S(s = (t + s) - t)$ is computed using Smith and Shively (1995), which leads to the following approximation:

$$F_t(s) = P[S \leq s] \approx 1 - \exp\left[-\sum_{k=1}^{s} \Psi_{t+k}(u)\right]. \tag{12}$$

5 DATA ANALYSIS

Using the threshold selection procedure described previously (for more details see Bellanger and Tomassone, 2000; Bellanger, 2001), the data we consider in this article consist of ozone exceedances over $130\,\mu g\,m^{-3}$. To model the exceedance times and sizes, we determine which covariates should be included in the expression of $\alpha(i, \text{Station})$ given by Eq. (7) and $\beta(i, \text{Station})$ given by Eq. (10). To take into account the possible non-linear parameters in the covariates, we first use a non-parametric approach that might be viewed as an exploratory tool, and follow the procedure explained in the previous section.

5.1 Exceedance Times

To model the exceedance times, we first use a logistic additive model, introducing t, Tmax, Wind and Trange in Eq. (8), but also Station in order to detect possible monitoring site effects. Station is introduced as linear effects coded at seven levels using reference cell coding with NE as the reference group. Partial residual plots (Chambers and Hastie, 1992) allow us to detect clearly that the effect of t (the temporal trend) is more complicated than a simple linear one. The analysis of curvature in the pattern of the partial residuals (plotted against each covariate) (Fig. 5) suggests that non-linear transformation of the variables might improve the fit.

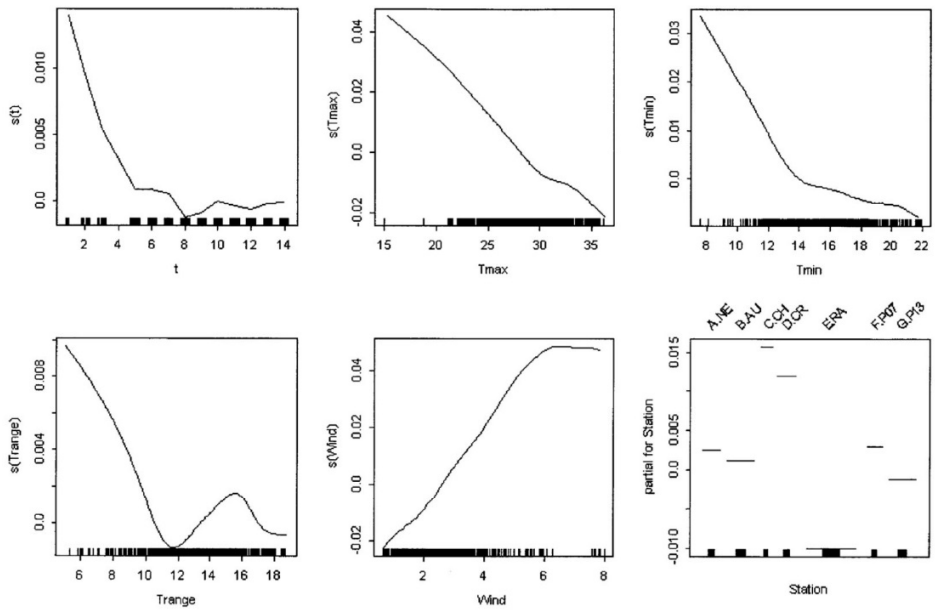


FIGURE 5 Partial residual plots for the exceedance times.

This result convinces us to build a classical logistic regression model with logit link, but introducing a change-point for t ($t = 3$ corresponding to 1990), Wind (Wind $= 6\,\mathrm{m\,s^{-1}}$) and Trange (Trange $= 10\,^{\circ}\mathrm{C}$); where we use a log transformation for temperature.

The best model, obtained with the stepwise selection of variables, includes the following covariates: $t^{*}1_{(t>3)}$, log(Tmax), Wind$^{*}1_{(\mathrm{Wind}>6)}$, Wind$^{*}1_{(\mathrm{Wind}>6)}$, Station, Trange$^{*}1_{(\mathrm{Trange}<10)}$. Estimated logistic regression coefficients are in Table IV. Thus, the estimated probability of an exceedance on day i of the threshold $130\,\mu\mathrm{g\,m^{-3}}$ takes the form (6), where the expression for

TABLE IV Estimated coefficients for the exceedance times model (period 1988–2001).

Coefficients	*Value*	*Std. error*	*t-Value*
Intercept	−43.682	1.820	−24.107
$t^{*}1_{(t>3)}$	0.163	0.017	9.411
log(Tmax)	12.807	0.535	23.927
Wind$^{*}1_{(\mathrm{Wind}<6)}$	−0.599	0.055	−10.826
Wind$^{*}1_{(\mathrm{Wind}\geq 6)}$	−0.340	0.050	−6.779
Station AU	0.461	0.202	2.285
Station CH	−0.460	0.238	−1.933
Station CR	0.128	0.216	0.594
Station RA	2.191	0.199	11.031
Station P07	0.009	0.226	0.041
Station P13	0.836	0.206	4.061
Trange$^{*}1_{(\mathrm{Trange}<10)}$	−0.083	0.020	−4.217
Null deviance		4395.377 at 7730 df	
Residual deviance		2243.237 at 7719 df	

α (i,sta) is given by Eq. (8) replacing the estimated coefficients by those in Table IV. They can be interpreted as follows:

- After allowing for the confounding effects of meteorological conditions, there is a significant increasing trend through time after 1990 ($t > 3$) in the frequency of the exceedances over $130\,\mu g\,m^{-3}$.
- Daily maximum temperature, through its logarithm, has a significant increasing effect on the probability of an exceedance.
- Daily average wind speed has a significant decreasing effect on the probability of an exceedance, with varying importance for low ($<6\,m\,s^{-1}$) and high values ($\geq 6\,m\,s^{-1}$).
- The range of temperatures down to 10 °C during the day has a significant decreasing effect on the probability of an exceedance.
- The estimated coefficients of the design variables for Station are quite different compared to the chosen control site NE. The frequency of exceedances is lower at CH, equal at CR and P07, higher at AU, P13 and greatest at the rural monitoring site RA.

Since an assessment of the temporal trend is a major current topic of public health interest, we analyse the previous results more precisely.

To interpret our results we introduce the so-called measure of association, the odds ratio. This statistic, widely used (notably in epidemiology), has an interest for continuous variables such as t, because it approximates the relative risk of an exceedance time on a given period (for example, increase of one, two or more years) as the logit difference between two years. In our case, considering that all other covariates in the model are taken to be fixed, the estimated odds ratio for a change of five years in t (where $t > 3$) is $\exp(5 * 0.163) = 2.26$, with an estimated standard error of $5 * 0.017$. This means that the risk of an exceedance time over $130\,\mu g\,m^{-3}$ increases 2.3 times for an increase of five years. The endpoints of a 95% confidence interval for this odds ratio are (2.18, 2.34).

It is also interesting to summarize the results of the fitted logistic regression model via a classification table. The table is the result of a cross-classification of the outcome variable (the exceedance time over 130) with a dichotomous variable whose value is derived from the estimated logistic probability. To obtain this value, we must define a cutpoint probability c and compare each estimated probability to c: if the estimated probability exceeds c the derived variable is 1 (or True); otherwise, it is equal to 0 (or False). With the estimated model and the commonly used cutpoint value 0.5, we obtain the results of Table V.

Due to the fact that classification is sensitive to the relative sizes of the groups and always favours classification into the larger group, we obtain a large amount of misclassifications for exceedance time with $c = 0.5$. To choose the optimal cutpoint for the purposes of classification, one might select the cutpoint that maximizes both sensitivity and specificity as defined in Table V; Figure 6 shows that $c = 0.11$ is an optimal choice, whereas that is approximately the

TABLE V Classification table based on the logistic regression model using a cutpoint of 0.5. Sensitivity = 6974/7095 = 98.3%; specificity = 288/636 = 45.3%.

	Observed		
Classified	*<130*	*≥130*	*Total*
<130	6,974	348	7,322
≥130	121	288	409
Total	7,095	636	7,731

TABLE VI Classification table based on the logistic regression model using a cutpoint of 0.11. Sensitivity = 6246/7095 = 88.0%; specificity = 558/636 = 87.74%.

	Observed		
Classified	*<130*	*≥130*	*Total*
<130	6,246	78	6,324
≥130	849	558	1,407
Total	7,095	636	7,731

point where the sensitivity and the specificity curves cross. We obtain the classification results of Table VI.

A more complete description of classification accuracy is given by the area under the receiver operating characteristic (ROC) curve. Used in signal detection theory, this curve describes the performance of a receiver detecting the existence of a signal in the presence of noise: it plots the probability of detecting true signals (sensitivity) and false signals (1−specificity) for an entire range of possible cutpoints. The area under the ROC curve provides a measure of the model's ability to discriminate between exceedance or not of the level $130\,\mu g\,m^{-3}$. We obtain an area equal to 0.93 which is considered as an outstanding discrimination (Fig. 6). In fact, the area under the ROC curve corresponds to the value of the Mann–Whitney U statistic for our data, and is equal to the number of times the 7095 days observed as True have a higher probability than the 636 days observed as False.

As we said before, we need to check assumptions regarding the distribution hypothesis of the inter-event times given by Eq. (9). Using the approximation given by Eq. (12), we obtain for each site the results given in Table VII. We cannot reject the distributional hypothesis for five sites, but for RA and P13 we reject it. One possible explanation might be found in the

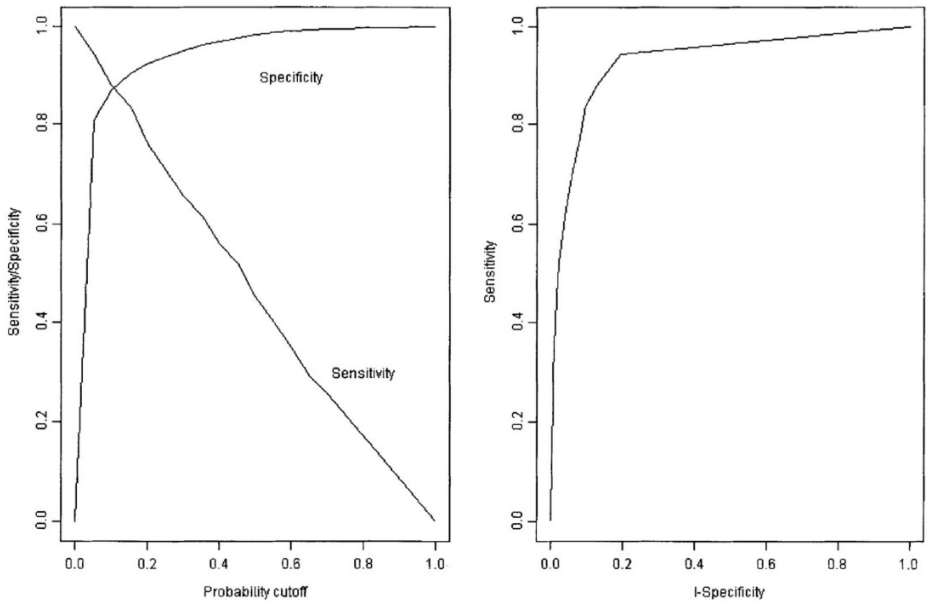


FIGURE 6 ROC curve: plot of sensitivity *vs.* 1−specificity for all possible cutpoints.

TABLE VII Validation of the distribution hypothesis of the inter-event times for threshold 130 (Kolmogorov–Smirnov test).

Station	*NE*	*AU*	*CH*	*CR*	*RA*	*P07*	*P13*
KS	0.1283	0.1276	0.1729	0.0828	0.1843	0.2189	0.2115
p-value	0.5263	0.3277	0.4168	0.9249	0.0026	0.0884	0.0151

differences in measurement altitudes for RA and P13, described in Section 2. It is clear that we are not able to separate site effect and altitude effect, and the confounding in these two sites introduces a deficiency in our interpretation!

A final validation was made by resampling, using 1000 replications by bootstrap. The results are given in Table VIII and Figure 7. The 95% percentile confidence interval for the odds ratio for a change of five years in t is (1.9, 2.3) slightly larger than the previous one; but the estimated risk of exceedance times still remains the same (2.3 times for a change of five years).

5.2 Exceedance Sizes

As described in Section 3, the natural distribution for the exceedance sizes over 130 $\mu g\,m^{-3}$ is the GPD, the exponential distribution being a limiting form of it, corresponding to $\xi \to 0$. Initially, we check that the exponential distribution is appropriate to our case in applying a one-sample Kolmogorov–Smirnov test of composite exponentiality data (KS = 0.037, p-value = 0.34) allowing us to conclude that we cannot reject the hypothesis that the true distribution is the exponential one. Therefore, to obtain an expression of β (i,sta) given by Eq. (10), we have to fit an exponential distribution to the exceedances sizes over 130.

We proceed as for the exceedance times. We first use a generalized additive model with the inverse link, the canonical one in the case of the exponential distribution. This exploratory

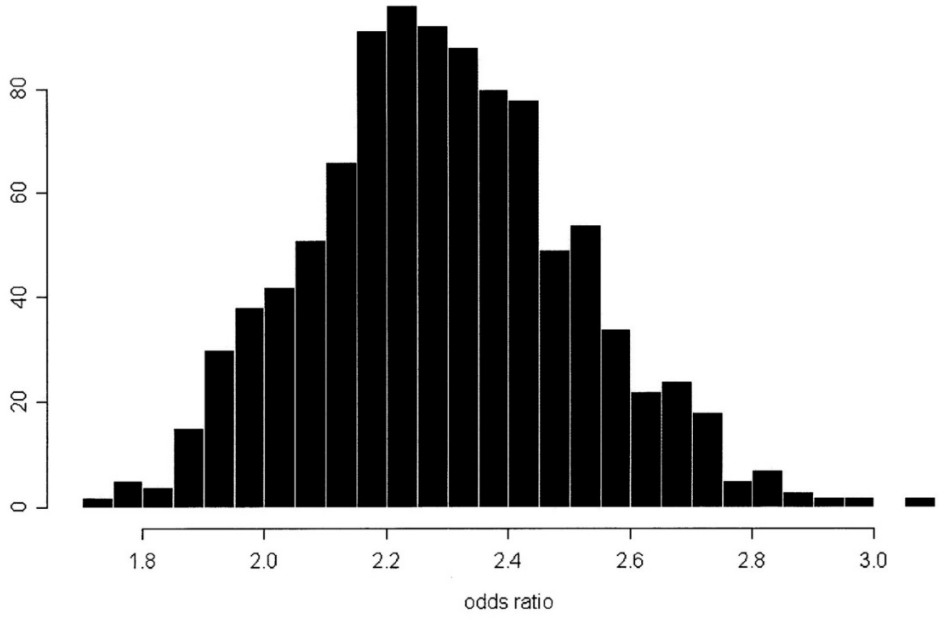


FIGURE 7 Odds ratio histogram for $B = 1000$ replicates by bootstrap.

TABLE VIII Exceedance times comparison of direct estimation of the model parameter with bootstrap resampling (quantiles: $Q_{0.025}$: 0.025; $Q_{0.250}$: 0.250; $Q_{0.750}$: 0.750; $Q_{0.975}$: 0.975) over threshold 130 (first row) and all observations (second row).

Predictor	*Estimate*	*Std. error*	*t-value*	*Min*	$Q_{0.025}$	$Q_{0.250}$	*Median*	*Mean*	$Q_{0.750}$	$Q_{0.975}$	*Max*
Intercept	−43.68	1.81	24.11	−50.88	−47.57	−45.20	−44.06	−44.05	−42.86	−40.56	−37.99
log(Tmax)	12.81	0.54	23.93	11.22	11.92	12.57	12.89	12.92	13.25	13.96	15.03
Wind<6	−0.599	0.55	10.83	−0.787	−0.727	−0.643	−0.603	−0.604	−0.562	−0.495	−0.432
Wind>= 6	−0.340	0.051	6.78	−0.578	−0.444	−0.379	−0.346	−0.347	−0.314	−0.259	−0.198
$t > 3$	0.163	0.017	9.41	0.111	0.128	0.153	0.165	0.165	0.178	0.202	0.225
Station AU	0.461	0.202	2.28	−0.172	0.060	0.323	0.459	0.462	0.599	0.859	1.074
Station CH	−0.460	0.238	1.93	−1.36	−0.941	−0.62	−0.46	−0.468	−0.312	−0.018	0.304
Station CR	0.128	0.216	0.59	−0.562	−0.333	−0.028	0.121	0.118	0.266	0.543	0.864
Station RA	2.191	0.199	11.03	1.619	1.782	2.055	2.191	2.194	2.327	2.606	2.845
Station P07	0.009	0.226	0.04	−0.773	−0.436	−0.162	−0.010	−0.005	0.151	0.428	0.660
Station P13	0.836	0.206	4.06	0.149	0.454	0.697	0.843	0.837	0.974	1.215	1.417
Trange< 10	−0.083	0.020	4.22	−0.143	−0.126	−0.097	−0.083	−0.083	−0.070	−0.046	−0.022
Odds-ratio (five years)	2.3			1.7	1.9	2.1	2.3	2.3	2.4	2.7	3.1

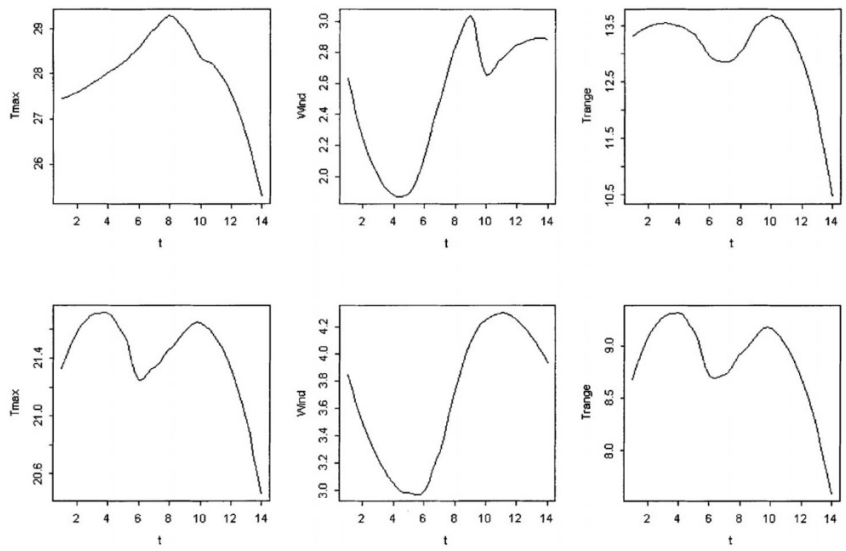


FIGURE 8 Partial residual plots for the exceedances over threshold 130. Ozone levels higher than the chosen threshold (upper row), all observations (lower row).

phase suggests a non-linear temporal trend (Fig. 8). It convinces us to fit a classical generalized linear model (exponential regression in our case) with inverse link, introducing non-linearity in the temporal trend as $\beta_{11}/t(i) + \beta_{12}t(i)$.

The results of fitting this model are shown in Table IX.

Thus, the expected size of an exceedance given that an exceedance of the threshold $130\,\mu g\,m^{-3}$ occurs on day i takes the form $1/\hat{\beta}(i,\text{sta})$, where the expression for $\hat{\beta}(i,\text{sta})$ given by Eq. (10) is replaced by:

$$\hat{\beta}(i,\text{sta}) = -0.01825 + \frac{0.09046}{t(i)} + 0.00217t(i) + 0.0273\text{Wind}(i) - 0.00119\text{Wind:Trange}(i) + \hat{\text{S}}\text{tation}_{\text{sta}},$$

where Wind:Trange represents the interaction between Wind and Trange

TABLE IX Estimated coefficients for exceedance sizes model (period 1988–2001).

Term	*Value*	*Std. error*	*t-value*
Intercept	−0.01825	0.01208	−1.51
1/t	0.09046	0.03541	2.55
t	0.00217	0.00082	2.64
Wind	0.02730	0.00343	7.97
Wind : Trange	−0.00119	0.00023	−5.15
Station AU	−0.00001	0.00486	−0.00
Station CH	0.01507	0.00716	2.10
Station CR	0.01297	0.00623	2.09
Station RA	−0.00787	0.00420	−1.87
Station P07	0.002207	0.00558	0.40
Station P13	−0.00059	0.00479	−0.12

Note: Residual deviance: 565.9165 on 625 df.

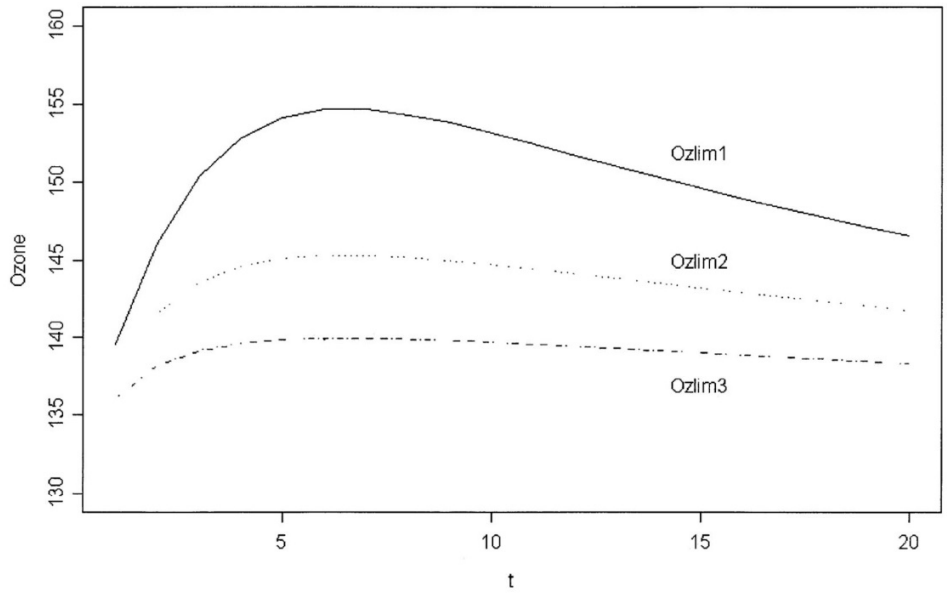


FIGURE 9 Expected ozone values over 130 *vs.* temporal trend.

These coefficients can be interpreted as follows. The result of the temporal trend is thus modelled as effecting the intercept by adding a hyperbolic term and Figure 9 shows the curvature in the pattern of the expected ozone values for different values of the retained meteorological covariates Wind and Trange (mean values (3.733 m s^{-1} and 8.9 °C) for Ozlim1, third quartile values (2.54 m s^{-1} and 11.4 °C) for Ozlim2, and near extremal values (2.0 m s^{-1} and 20.0 °C) for Ozlim3).

No important trend exists, but a small decreasing effect began several years ago. As seen for exceedance times, high Wind values have a decreasing effect on the expected size of an exceedance, whereas Trange (nested in Wind in the model) has an increasing one. The estimated coefficients of the design variables for Station, coded at seven levels using NE as the reference site, allow us to see that RA tends to show higher exceedance sizes but not very significant

TABLE X Estimated ozone values over threshold 130 for three sites using model in Table VIII.

		CH				*NE*				*RA*			
Site, year ahead		$t = 1$		$t = 5$		$t = 1$		$t = 5$		$t = 1$		$t = 5$	
Wind	*Trange*	*fit*	*se*	*fit*	*se*	*fit*	*se*	*fit*	*se*	*fit*	*se*	*fit*	*se*
1.0	10	150	3	147	3	158	4	153	4	166	6	158	6
1.0	15	152	4	149	3	164	6	157	6	176	10	164	9
1.0	20	156	5	152	5	172	10	162	8	193	21	173	15
1.5	10	147	2	145	2	153	3	150	3	158	4	153	4
1.5	15	150	3	148	3	159	5	154	4	168	7	159	6
1.5	20	155	5	151	4	169	9	160	7	187	17	170	13
2.0	10	145	2	144	2	150	2	147	2	153	2	150	3
2.0	15	148	2	146	2	156	3	151	3	162	5	156	5
2.0	20	154	4	150	4	167	8	159	7	182	15	167	11

Note: fit, estimated value over 130; se, standard error.

(p-value of 0.0620). CH and CR both have lower ones with p-values respectively 0.0360 and 0.0370. These results are a confirmation of what was detected on Box plot (Fig. 1). Therefore, after allowing for the confounding effects of meteorological conditions, there is a different but small downward trend over time in the exceedance sizes for NE, RA, CH and CR, the other monitoring sites having the same behaviour as NE (Table X).

To check the assumption regarding the distribution of the exceedance sizes, we use a Kolmogorov–Smirnov test (KS $= 0.044$ and p-value $= 0.5362$), and conclude that the exponential distribution hypothesis cannot be rejected.

A final validation was made by resampling, using 1000 replications in bootstrap. The results (not given here) are similar to estimated ones.

6 CONCLUSIONS

This work is a continuation of some previous ones (Bellanger and Tomassone, 2000; Bellanger, 2001). The use of NHPP, whose parameters depend on both time and meteorology, appears to be an improvement when compared to other models. The simultaneous analysis of exceedance times and sizes seems to be a good compromise, even if within our data, confounding between sites and measurement altitudes may introduce some doubt in conclusions. The distribution of exceedance sizes is well approximated by an exponential law, that is the limited model case in which shape parameter $\xi = 0$ in the GPD model. The method hereby exposed is limited to the case where ξ is supposed to be constant. Indeed, because the shape parameter ξ is dominant in determining the qualitative behaviour of the GPD and it can also allow it to vary with time. It will be interesting to test much more complicated models for the exceedance sizes, even if in our case a test does not reject the exponential model.

Both exceedance times and sizes are quite well predicted by classical meteorological variables. The importance of site measurements must not be neglected, although confounding has to be analysed in more detail. An important increasing trend through time was revealed for exceedance times and a small decreasing one for exceedance sizes. In both cases predictions can be made; but the quantitative aspect must be taken with caution.

These results confirm the complexity of relations between ozone and meteorological variables, but these relations are clearly estimated. Validation, using bootstrap resampling, gives good confirmation of estimation quality.

We may argue that the tools used in our analysis are well identified: the use of NHPP associated with generalized linear models are the basic ones. But, as always in statistical applications, various other steps are essential in the modelling approach:

- first, in an exploratory and graphical step, choose the best (if possible) model and,
- secondly, validate the proposed model after having estimated the different important parameters.

Some further improvements will be necessary to update results with new data.

Acknowledgements

We have to thank AIRPARIF organization for the use of their data, Professor Dacunha-Castelle from Orsay University for his constant help in conducting this work and Mary Beth Loup and Frank Parrotta for reading, helpful comments and suggestions.

References

Bellanger, L. (2001). Une analyse globale de la tendance dans les hautes valeurs d'ozone mesurées en région parisienne. *Revue de Statistiques Appliquées*, **XLIX**(3), 73–92.

Bellanger, L. and Tomassone, R. (2000). La pollution de l'air dans la région parisienne. Étude de la tendance dans les hautes valeurs d'ozone. *Revue de Statistiques Appliquées*, **XLVIII**(1), 5–24.

Bellanger, L. and Perera, G. Compound Poisson limit theorems for high-level exceedances of some non-stationary processes. *Revue Bernoulli* (to appear).

Chambers, J. M. and Hastie, T. J. (1992). *Statistical Models*. S. Wadsworth and Brooks Cole Advanced Books and Software, Pacific Grove, CA.

Chavez-Demoulin, V. (1999). Two problems in environmental statistics: Capture-recapture models and smooth extremal models. *PhD thesis*, EPFL, Lausanne, Switzerland.

Coles, S. (2001). *An Introduction to Statistical Modeling of Extreme Values*. Springer Series in Statistics, New-York.

Coles, S. and Dixon, M. (1999). Likelihood-based inference for extreme value models. *Extremes*, **2**(1), 5–23.

Coles, S. and Tawn, J. A. (1991). Modelling extreme multivariate events. *J. R. Statist. Soc. B*, **53**, 377–392.

Cox, D. R. (1955). Some statistical methods connected with series of events. *JRSS*, **XVII**(No. 2), 129–157, 176.

Cox, D. R. and Isham, V. (1992). *Point Processes*. Chapman & Hall, London, 2iéme éd.

Davison, A. C. (1984). Modelling excesses over high thresholds, with an application. In: Tiago de Oliveira, J. (Ed.), *Statistical Extremes and Applications*, Dordrecht, Reidel, pp. 424–434.

Davison, A. C. and Smith, R. L. (1990). Models for exceedances over high thresholds (with discussion). *J. R. Statist. Soc.*, **52**, 393–442.

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and their Application*. Cambridge University Press. (statwww.epfl.ch/Davison/BMA).

Davison, A. C. and Ramesh, N. I. (2000). Local likelihood smoothing of sample extremes. *J. R. Statist. Soc. B*, **62**, Part 1, pp. 191–208.

Dobson, A. J. (2002). *An Introduction to Generalized Linear Models*, 2nd ed. Chapman & Hall, London.

Draper, N. R. and Smith, H. (1981). *Applied Regression Analysis*, 2nd ed. John Wiley & Sons, New York, pp. 177–183.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. New York: Chapman & Hall.

Embrechts, P., Kluppelberg, C. and Mikosch, T. (1999). *Modelling Extremal Events for Insurance and Finance*, 2nd ed. Springer Verlag, New York.

Falk, M., Husler, J. and Reiss, R.-D. (1994). *Law of Small Numbers: Extremes and Rare Events*. DMV Seminar 23, Birkhäuser-Verlag.

Hall, W. J. and Wellner, J. (1981). Mean residual life. In: Csörgo, M., Dawson, D. A., Rao, J. N. K. and Md, A. K., Saleh, E. (Eds.), *Statistics and Related Topics*, North-Holland, Amsterdam, pp. 169–184.

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*. Chapman & Hall, London.

Heffernan, J. E. and Tawn, J. A. (2002). An extreme value analysis for the investigation into the sinking of the M. V. Derbyshire. Submitted to *Appl. Statist.*

Hosking, J. M. R. and Wallis, J. R. (1987). Parameter and quantile estimation for the generalized Pareto distribution. *Technometrics*, **29**, 339–349.

Hosmer, D. W. and Lemeshow, S. (2000). *Applied Logistic Regression*, 2nd ed., John Wiley, New York.

Hüsler, J. (1986). Extreme values of non-stationary random sequences. *J. Appl. Probab.*, **23**, 937–950.

Joe, H., Smith, R. L. and Weissman, I. (1992). Bivariate threshold methods for extremes. *J. R. Statist. Soc. B*, **54**, 171–183.

Kallenberg, O. (1983). *Random Measures*. Academic Press, New York.

Lawless, J. F. (1982). *Statistical Models and Methods for Lifetime Data*. John Wiley & Sons, New York, pp. 84–88.

Leadbetter, M. R. (1991). On a basis for 'Peaks over Threshold' modeling. *Statistics Probability Lett.*, **12**, 357–362.

Leadbetter, M. R. (1993). On exceedance based environmental criteria. *Technical Report 9*, National Institute of Statistical Sciences, P.O. Box 14162, Research Triangle Park, N.C. 27709.

Leadbetter, M. R., Lindgren, G. and Rootzen, H. (1983). *Extremes and Related Properties of Random Sequences and Series*. Springer Verlag, New York.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, 2nd ed. Chapman & Hall, London.

Michaelis, W. (1997). *Air Pollution: Dimensions, Trends and Interactions with a Forest Ecosystem*. Springer Verlag, New York.

Nychka, D., Piegorsch, W. W. and Cox, L. H. (Editors) (1998). Case studies in environmental statistics. *Lecture Notes in Statistics*. Springer, New York.

Pickands, J. (1971). The two-dimensional Poisson process and extremal processes. *J. Appl. Prob.*, **8**, 745–756.

Pickands, J. (1975). Statistical inference using extreme order statistics. *Ann. Statist.*, **3**, 119–131.

Reiss, R. D. and Thomas, M. (2001). *Statistical Analysis of Extreme Values with Applications to Insurance, Finance, Hydrology and Other Fields*, 2nd ed., Birkhauser Verlag AG.

Shao, J. and Tu, D. (1996). *The Jackknife and Bootstrap*. Springer Verlag, New York.

Shively, T. S. (1990). An analysis of the long-term trend in ozone data from two Houston, Texas monitoring sites. *Atmospheric Environment*, **24B**(4), 293–301.

Shively, T. S. (1991). An analysis of the trend in ground-level ozone using nonhomogeneous Poisson processes. *Atmospheric Environment*, **25B**(4), 387–396.

Smith, R. L. (1984). Threshold methods for sample extremes. In: Tiago de Oliveira, J. (Ed.), *Statistical Extremes and Applications*. Dordrecht, Reidel, pp. 621–638.

Smith, R. L. (1985). Maximum likelihood estimation in a class of nonregular cases. *Biometrika*, **72**(1), 67–90.

Smith, R. L. (1986). Extreme value theory based on the *r* largest annual events. *J. Hydrol.*, **86**, 27–43.
Smith, R. L. (1989). Extreme values analysis of environmental time series: An application to trend detection in ground-level ozone (with discussion). *Statistical Sciences*, **4**, 367–393.
Smith, R. L. and Huang, L. (1993). Modeling high threshold exceedances of urban ozone. *National Institute for Statistical Science Technical Report # 6*.
Smith, R. L. and Shively, T. S. (1995). Point process approach to modelling trends in tropospheric ozone based on exceedances of a high threshold. *Atmospheric Environment*, **29**(3), 3489–3499.
Smith, R. L., Tawn, J. A., and Coles, S. G. (1997). Markov chain models for threshold exceedances. *Biometrika*, **84**, 249–268.
Snedecor, G. W. and Cochran, W. G. (1971). *Méthodes statistiques*. Association de Coordination Technique Agricole, Paris.
Thompson, M. L., Reynolds, J., Cox, L. H., Guttorp, P. and Sampson, P. D. (2001). A review of statistical methods for meteorological adjustment of tropospheric ozone. *Atmospheric Environ.*, **35**, 617–630.
Vaquera-Huerta, H., Villasenor, J. A. and Hughes, J. (1997). Statistical analysis of trends in urban ozone. In Barnett, V. and Turkman, K. F. (Eds.), *Statisitcs for the Environment 3: Pollution Assessment and Control*. John Wiley, New York, pp. 175–183.
Venable, W. N. and Ripley, B. D. (1997). Modern applied statistics with S-Plus. In: Chambers, J., Eddy, W., Härdle, W., Sheather, S., Tierney, L. (Eds.), *Statistics and Computing*, 2nd ed., Springler-Verlag, New York.

2. Cluster analysis of linear model coefficients under contiguity constraints for identifying spatial and temporal fishing effort patterns.

Mahévas S., Bellanger L., Trenkel V. (2008). *Fisheries Research*, **93**(1-2): 29-38

Fisheries Research 93 (2008) 29–38

Contents lists available at ScienceDirect

Fisheries Research

journal homepage: www.elsevier.com/locate/fishres

Cluster analysis of linear model coefficients under contiguity constraints for identifying spatial and temporal fishing effort patterns

Stéphanie Mahévas[a,*], Lise Bellanger[b,1], Verena M. Trenkel[a]

[a] IFREMER, Département EMH, BP 21105, 44311 Nantes Cedex 03, France
[b] Université de Nantes, Laboratoire de Mathématiques Jean Leray, UMR CNRS 6629, BP 92208, 44322 Nantes Cedex 03, France

ARTICLE INFO

Article history:
Received 5 June 2007
Received in revised form 5 February 2008
Accepted 11 February 2008

Keywords:
Generalised linear model
Cluster analysis under contiguity constraints
Statistics for spatial data
Spatial and seasonal pattern
Allocation of Fishing effort
Fleet dynamics

ABSTRACT

For fisheries management purposes, it is essential to take into account spatial and seasonal characteristics of fishing activities to allow a reliable assessment of fishing impact on resource. This paper presents a novel technique for describing spatial and temporal patterns in fishing effort. The spatial and seasonal fishing activity patterns of the French trawler fleet in the Celtic Sea during the period 1991–1998 were analysed by modelling fishing effort (fishing time) with generalised linear models. The linear model for fishing effort included fixed effects for both spatial (statistical rectangles) and temporal units (months). In addition, spatial correlations in any given month were modelled by an exponentially decreasing function. Temporal correlations were included using the previous month's fishing effort for a given spatial unit as predictor. A method based on cluster analysis of estimated model coefficients of spatial or temporal fixed effects is proposed for identifying groups of similar spatial and temporal units. A contiguity constraint is imposed in the clustering algorithm, ensuring that only neighbouring spatial units or consecutive temporal units are grouped. The cluster analysis identified 22 spatial and 9 temporal groups. Winter and spring months stood out as being more variable than the remaining months. Spatial groups were of varying size, and generally larger offshore. The proposed method is generic and could for example be used to analyse temporal and spatial patterns in catch or catch rate data.

Résumé: Dans un objectif de gestion des pêcheries, pour établir un diagnostic fiable de l'impact de la pêche sur la ressource, il est nécessaire d'intégrer les spécificités spatiales et temporelles de l'activité de pêche. Ce papier présente une nouvelle méthode pour décrire des structures spatiales et temporelles de l'effort de pêche. La distribution spatiale et saisonnière des chalutiers français pêchant en mer Celtique entre 1991 et 1998 est analysée en modélisant l'effort de pêche (temps de pêche) à l'aide d'un modèle linéaire généralisé. Le modèle décrivant la variabilité de l'effort de pêche incluait des effets fixes spatiaux (à l'échelle du rectangle statistique) et temporels (à l'échelle du mois). Les corrélations spatiales à un mois donné étaient modélisées par une fonction exponentielle décroissante de la distance et pour tenir compte des corrélations temporelles nous avons introduit, pour une unité spatiale donnée, l'effort de pêche du mois précédent comme variable explicative dans le modèle. Une méthode de classification des effets fixes spatiaux (respectivement temporels) du modèle statistique est alors proposée pour construire des groupes d'unités spatiales (respectivement des groupes d'unités temporelles). Des contraintes de contiguïté spatiale et temporelle sont imposées dans l'algorithme de classification pour s'assurer que seules les unités spatiales voisines et que seules les unités temporelles successives soient groupées. L'application de cette méthode de classification a permis d'identifier 22 zones et 9 saisons. Les mois d'hiver et de printemps ressortent comme étant plus hétérogènes que les autres. La taille des zones est très variable et généralement plus grande au large qu'à la côte. La méthode proposée est générique et pourrait être par exemple utilisée pour identifier des structures spatiales et temporelles des données de capture ou de taux de capture.

1. Introduction

Fishing fleet dynamics are characterized by the choice of fishing location and the set of target species at a given time of the year (Hilborn and Ledbetter, 1985). Seasonal species migrations (Biseau, 1998; Vignaux, 1996a), economic changes and weather conditions

* Corresponding author. Tel.: +33 240374181; fax: +33 240374075.
E-mail addresses: Stephanie.Mahevas@ifremer.fr (S. Mahévas), Lise.Bellanger@univ-nantes.fr (L. Bellanger).
[1] Tel.: +33 251125900; fax: +33 251125912.

0165-7836/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.fishres.2008.02.008

(Holland and Sutinen, 1999; Sampson, 1991) make fishing activities variable in both time and space. In order to reliably evaluate the impact of a given fishing fleet on a particular resource, it has been argued that taking account of spatial and seasonal characteristics of fishing activities is essential for reliable stock assessments and realistic forecasting models for management purposes (Booth, 2000). This leads first to a decomposition of fishing effort by métier which is defined by season, location, target species and fishing gear (Biseau and Gondeau, 1988), and commonly accepted as a fundamental feature of fishing activities (ICES, 2004). The exploration of alternative management measures is another field of application of these spatial and temporal patterns. Babcock and Pikitch (2000) underlined the importance of spatial and seasonal knowledge of populations and fleets to design appropriate marine protected areas and successful management measures. Several simulation tools for management scenario testing have been developed that require definition of distinct spatial and temporal fishing activity units (e.g. Sparre, 2003; Mahévas and Pelletier, 2004; Pelletier and Mahévas, 2005).

In fisheries science, hierarchical cluster analysis is commonly employed for grouping observation units, such as observation years for scientific surveys (Poulard, 2001), catch composition for identifying métiers and strategies (Pelletier and Ferraris, 2000) or species spatial distributions (Verdoit et al., 2003). The general aim is to group sampling units that show common patterns. Here we propose a model-based cluster analysis for grouping variables, such as a temporal or spatial effects. These variables are the coefficients of a linear model, and hence assess the average features of variability of observations conditional on model formulation. Clustering estimated model effects instead of raw observations allows us to ignore local fluctuations of observations not explained by spatial or temporal factors, for instance due to autocorrelation structures in observations.

Cluster analysis is an algorithmic procedure providing partitions of the initial population. Two broad clustering families have been developed (Lebart et al., 1997; Gordon, 1996): mobile centroid clustering methods (Hartigan and Wong, 1979) and hierarchical clustering approaches (Gordon, 1987). The first family partitions directly units into disjoint groups. Allocation of units is iterative in order to minimize the distance of each unit to the estimated centroids of the clusters. Hierarchical clustering is based on an agglomerative technique grouping units two by two (or divisive technique splitting the group into two groups). Some additional constraints, usually contiguity constraints, are often required in the classification to take neighbourhoods into account. Clustering with contiguity constraints requires first the definition of a neighbourhood relationship (e.g. horizontal, spherical adjacencies) and second performing a clustering algorithm modified to take into account the neighbourhood constraints. Gordon (1996) provides a review of constrained classification methods. Different linkage methods can be used to decide whether objects are similar enough to be grouped. The most commonly used methods are complete linkage and single linkage. Complete linkage is often used in ecology when one wishes to delineate clusters with clear discontinuities (Legendre and Legendre, 1998). The single linkage method has the advantage over the other methods to only use rank distance and consequently, to be rather similar to non-parametric methods. It is also the only linkage method allowing hierarchical clustering with contiguity constraints (Everitt et al., 2001). Unfortunately, for noisy data, this linkage method is also well known to cause chaining in the dendrogam. Several studies have analysed and characterized the chaining phenomenon, see for instance Hartigan (1975) and Everitt et al. (2001). More details of these linkage techniques can be found in Gordon (1981) and Lebart et al. (1997).

In this study, we developed a clustering method with contiguity constraints based on a modified dissimilarity matrix and using a minimum linkage method independently on spatial and temporal units. Each unit is characterized by an estimated parameter value provided by a generalised linear model. The modified dissimilarity matrix is computed using (1) the $1-p$-value derived from a Fisher statistical test applied to estimated values to assess the null-hypothesis of equality of pairs of parameters for temporal units (or spatial units) and (2) the neighbourhood constraint. The temporal neighbourhood relationship is assumed horizontal: sorting the units in sequential order, a temporal unit can only be grouped with the previous and following temporal unit. Spatial units are located on a regular grid and the eight neighbours of a spatial unit define the spatial neighbourhood. We apply the proposed model-based clustering algorithm for determining spatial and seasonal patterns in fishing effort for the French trawler fleet in the Celtic Sea (Fig. 1). In the following description we assume that the temporal unit corresponds to calendar months and the spatial units to statistical rectangles (1° longitude by 0.5° latitude).

2. Material and method

2.1. Data

The data come from the French trawler fleet operating in the central part of the Celtic Sea during the period 1991–1998. The fleet consists of 589 trawlers between 12 and 24 meters in length. For each vessel-trip, total trawling time was available per statistical rectangle. We modelled total fleet fishing time per statistical rectangle and per month for each year (Mahévas and Trenkel, 2002).

2.2. Model

The approach has three steps (Fig. 2): (1) conducting an exploratory analysis of fishing time data to investigate statistical data distributions, autocorrelation structures, etc.; (2) fitting an appropriate statistical model to fishing time data to estimate spatial and temporal effects; (3) separately clustering the spatial and temporal effects estimated in the previous step to provide fishing zones and seasons.

Based on the exploratory data analysis (step 1, descriptive analysis and plots), a set of models for describing the spatial and temporal distributions of fishing time was defined (Table 1). The factors for statistical rectangles, months and years were modelled as fixed effects. A strongly right-skewed distribution was found for monthly fishing times per rectangle. We used the Box–Cox method relying on a maximum likelihood estimation to estimate the best power transformation of fishing time that would achieve normality (Draper and Smith, 1998). Consequently, fishing time is normalised by a fourth-root transformation. The full model for fleet fishing time T_{ijk} in month i, rectangle j and year k is defined as

$$T_{ijk}^{1/4} = m + \delta T_{(i-1)jk}^{1/4} + \text{month}_i + \text{rectangle}_j + \text{year}_k + \varepsilon_{ijk} \quad (1)$$

for $i = 1,\ldots,12$; $j = 1,\ldots,48$; $k = 1,\ldots,8$, assuming $\delta T_{(0)jk}^{1/4} = \delta T_{(12)jk}^{1/4}$, and where $\varepsilon \sim N_{n=4488}(0, \Sigma)$.

To take into account that fishing time in a rectangle might be correlated with fishing time in neighbouring rectangles, we include a spatial covariance structure with a nugget effect, and specify the covariance matrix as $\Sigma = \sigma^2 H(\varphi) + \tau^2 I$ where $(H(\varphi))_{jj'} = \rho(\varphi; d_{jj'})$, $d_{jj'}$ is the Euclidean distance between rectangle j and j', φ is the decay parameter, ρ is chosen as the classical exponential covariance function (see for example Cressie, 1993) of fishing times in

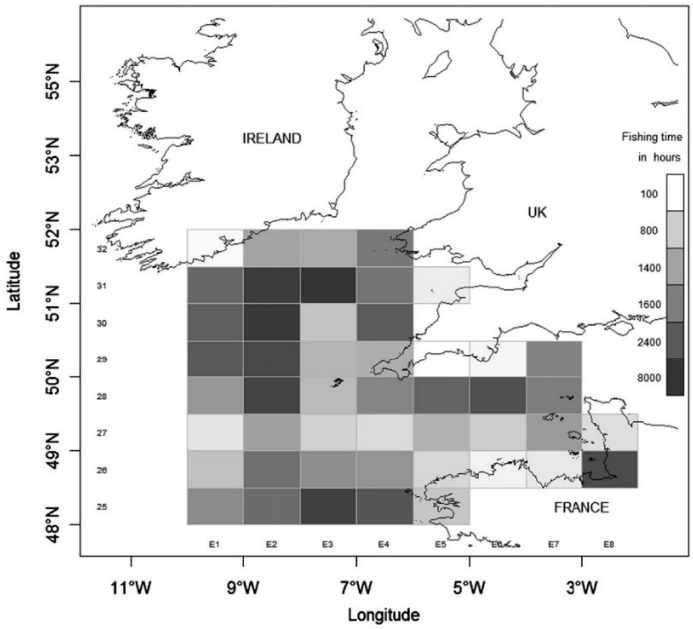


Fig. 1. Monthly fishing times of the French bottom trawlers in the Celtic Sea averaged over the years 1991–1998. For each statistical rectangle, the grey level is proportional to the monthly average. The rectangle name is the combination of a number stated on the left and a letter-number stated at the bottom of the graphic.

neighbouring rectangles j' in the same month i and year k and τ^2 is the nugget effect variance. The Euclidean distance between two rectangles is calculated using the centre of the rectangles identified by its geographical coordinates in degree (longitude, latitude). The model (1) includes the term $\delta T^{1/4}_{(i-1)jk}$ for describing the dependence of fishing time in a given rectangle j on the previous month's fishing time in the same rectangle (including the transition between December ($i = 12$) and January ($i = 1$)).

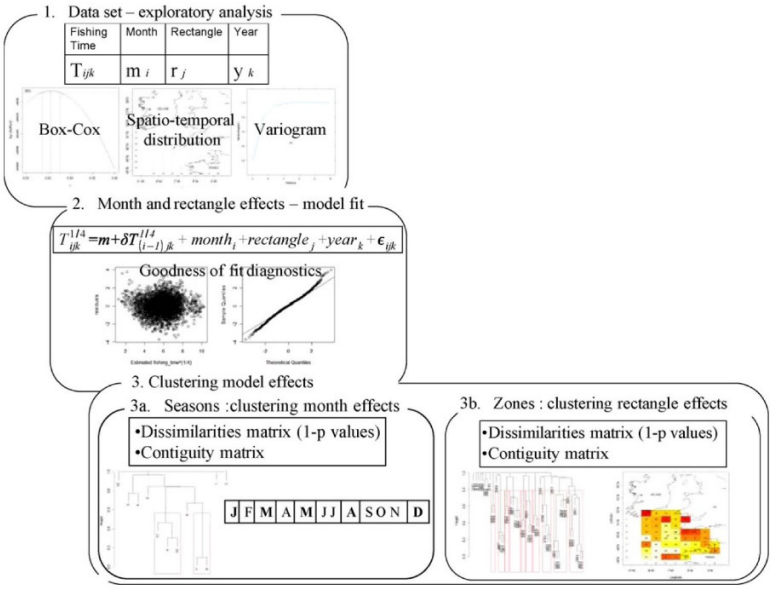


Fig. 2. Flowchart of the proposed modelling approach: (1) exploratory analysis of fishing time data, (2) model fitting to estimate month and ICES-rectangle effects and (3) cluster analysis to provide fishing seasons and zones.

Table 1
Comparison of different model fits (AIC) for total fishing time ($T^{1/4}$) allocated by month and rectangle in the Celtic Sea by the French trawler fleet during 1991–1998; d.f.: degrees of freedom

Model	Explanatory variables	d.f.	AIC
Basic	Month + rectangle + year	67	13547.95
AR (month-1)	(Previous month's fishing time)$^{1/4}$ + basic	68	12051.66
Spatial correlation	Basic + exp(neighbouring rectangle time)	69	12892.28
Full	Basic + (previous month's fishing time)$^{1/4}$ + exp(neighbouring rectangle time)	70	11582.09

The model is parametrized using classical treatment contrasts for coding of factors. In the following, the first level is set to 0 for each factor (January for month, 25E1 for rectangle and 1991 for year) and thus each coefficient represents the difference between that level and level one.

Model comparison and selection was carried out using Akaike's information criteria (AIC) (Akaike, 1974; Pinheiro and Bates, 2000). Residual plots were used to check model assumptions (McCullagh and Nelder, 1989). All models were fitted by maximum likelihood using R 2.5.1 (http://www.r-project.org).

2.3. Clustering algorithm with contiguity constraints

A hierarchical cluster analysis (HCA) (Lebart et al., 1997) is performed for grouping levels of spatial and temporal variables, using the set of dissimilarities for each pair of *spatial variables* (or *temporal variables*). In addition, we impose contiguity constraints on the set of allowable classification solutions: the objects in a class are required not only to be similar to one another, but also to comprise a spatial (or temporal) contiguous set of objects. For this, neighbours of a statistical rectangle are the eight adjacent rectangles and neighbours of a given month are the previous and following month. The only simple appropriate clustering method using contiguity constraints is the single linkage method. In practice, we implement a crude version of the *single linkage* clustering method with seasonal (or spatial) constraint using a classical hierarchical clustering algorithm (function `hclust` in R), setting the dissimilarities between non-adjacent months (or rectangles) to high values.

If η is the dissimilarity in the HCA, we define γ as the aggregate (joining) index in the usual HCA for rectangle (respectively month) by:

$$\gamma(\text{rectangle}_i, \text{rectangle}_{i'}) = \eta(\text{rectangle}_i, \text{rectangle}_{i'}) + \kappa(\text{rectangle}_i, \text{rectangle}_{i'})$$

where κ is the contiguity index defined by $\kappa(\text{rectangle}_i, \text{rectangle}_{i'}) = 0$ if the contiguity constraint is satisfied for rectangle_i and $\text{rectangle}_{i'}$, else $\kappa(\text{rectangle}_i, \text{rectangle}_{i'}) = +\infty$. In the HC algorithm, γ is then used as set of dissimilarities. The clustering results are not sensitive to the actual value. We detail below the computation of the dissimilarities η.

2.4. Raw data clustering

Clustering methods are classically applied to raw data. To demonstrate the necessity of using a based-model clustering approach, we first applied the hierarchical cluster method with imposed spatial contiguity constraints directly to the raw fishing effort data. For identifying fishing zones (similarly fishing seasons), averages of fishing times per ICES-rectangle (similarly, per month) over the study period were calculated and the dissimilarity η between two rectangles (or 2 months) was calculated as the squared differences between respective averages of the fishing times.

2.5. Model-based clustering

To identify homogeneous fishing time areas and seasons, clustering of model coefficients is carried out using dissimilarities calculated using the $1-p$ values of statistical tests on the estimated coefficients. Let us consider a special case of the general method for constructing tests for general linear models for hypotheses involving linear functions of parameters. We denote β the vector of parameters:

$$\beta = [m, \delta, \text{month}_2, \ldots, \text{month}_1, \text{rectangle}_2, \ldots, \text{rectangle}_4, \text{year}_2, \ldots, \text{year}]^T \in \Re^{67=1+1+11+47+7}$$

We use a F-test to test the equality of pairs of coefficients for factor month (or rectangle) (Searle, 1997; Rawlings et al., 2001).

We express model (1) as a "classic" general linear model $T^{*1/4} = X^*\beta + \varepsilon^*$ where $\varepsilon^* \sim N(0, \sigma^2 I)$ and maximum likelihood (ML) estimates of the model parameter vector $\beta \in \Re^{67}$ is obtained by solving an ordinary least-squares problem. For example, the single null hypothesis that two coefficients month_i and $\text{month}_{i'}$ (or rectangle_j and $\text{rectangle}_{j'}$) are equal is:

$$H_0 : K^T\beta = 0 \text{ against } H_1 : K^T\beta \neq 0$$

In our case, for example, to test H_0: $\text{month}_i = \text{month}_{i'}$ against H_1: $\text{month}_i \neq \text{month}_{i'}$, K is a row vector of length 67 with the ith element $K_i = 1$, the i'th element $K_{i'} = -1$ and zeros elsewhere.

The sum of squares for the hypothesis can be written $Q = (K^T\hat{\beta})^2/(K^T(X^{*T}X^*)^{-1}K)$ and has 1 degree of freedom. Thus, using classical notions of general linear models, the F-ratio, equivalent to a t-test is:

$$F = (Q/1)/(s^2) = (K^T\hat{\beta})^2/(K^T(X^{*T}X^*)^{-1}K)s^2) \sim_{H_0} F(1, 4421)$$

where $s^2 = \text{SC}_{\text{res}}/4421$ and SC_{res} is the residual sum of squares of the model and $4421 = 4488 - 67$ corresponding to the number of degrees of freedom.

A hierarchical cluster analysis (HCA) (Lebart et al., 1997) is then performed for grouping levels of spatial and temporal variables, using the set of dissimilarities produced by the $(1-p)$-values of the previous F-tests for each pair of factor levels month (or rectangle) and the contiguity constraints.

As we are not interested in the complete hierarchy but only partitions, we select one of the solutions in the nested sequence of clusterings that comprise the hierarchy in cutting the dendrogram at a particular height (sometimes termed the *best cut* indicated by large changes in fusion levels in the appearance of the dendrogram based on visual inspection (Everitt et al., 2001)).

3. Results

Exploratory data analysis showed that average monthly fleet fishing times varied considerably in space suggesting a strong rectangle effect (Fig. 1). In contrast, monthly averages (respectively annual monthly averages) did not show any specific month (respectively year) patterns. For each statistical rectangle, temporal autocorrelation was analysed using the Durbin Watson statistic

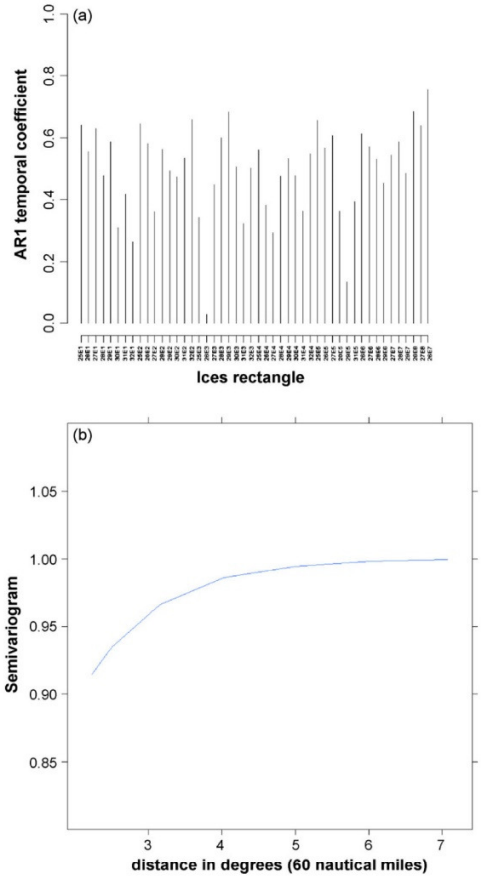


Fig. 3. Exploratory analysis of autocorrelation structures. (a) For each rectangle, the coefficient estimates of temporal autocorrelation of order 1 calculated using monthly fishing time series. (b) Semi-variogram values corresponding to the exponential correlation model calculated using monthly fishing times per rectangle and the Euclidean distance between rectangles. The selected neighbourhood distance is equal to 3.16°.

and we concluded that there existed a temporal autocorrelation of order one for all but two rectangles, 26E3 and 29E5 (Fig. 3a). These two rectangles were characterized by random fishing times. We explored the spatial autocorrelation structure plotting the semi-variogram. Fig. 3b shows an exponential spatial autocorrelation structure. The neighbourhood retained was 3.162278° (190 nautical miles), that is the eight adjacent rectangles.

Several models nested within the full model (Eq. (1)) were fitted to the fishing time data. Residual plots did not show any strong trend, indicating that all the models had reasonable fits (Fig. 4). The comparison of AIC values for the different models showed that overall temporal variations were more important than spatial variations (Table 1). The best model (smallest AIC) was the full model, which included both temporal and spatial correlations. The correlations between explanatory variables were examined and found

Table 2
Analysis of variance table for full model for fishing time per rectangle and month by the French bottom trawlers operating in the Celtic Sea during 1991–1998 (see Table 1 for model definition)

Effect	d.f.	*F*-value	*p*-Value
Intercept	1	61682.91	<0.0001
(Previous month's fishing time)$^{1/4}$	1	8246.55	<0.0001
Month	11	8.30	<0.0001
Rectangle	47	18.53	<0.0001
Year	7	1.56	0.1415
Corr struct	Lower	Est.	Upper
Spatial range	0.718	0.784	0.855
Nugget	5.967365E−43	2.872111E−08	1.0000000

to be less than 0.01. The temporal relationship with fishing time in the previous month was highly significant and estimated as $\delta = 0.47$ (95% confidence interval [0.45; 0.5]) (Table 2). The estimated spatial correlation coefficient was $\varphi = 0.78$ (95% confidence interval [0.71; 0.86]) demonstrating a strong spatial pattern in fishing time allocation. Year was not a significant factor (Table 2), which indicates that spatial and temporal exploitation patterns were stable over the study period.

We applied the clustering algorithm with contiguity constraints to raw fishing time. We only present the results for the spatial study but a similar conclusion was obtained for the temporal study. The dendrogram for the single linkage clustering with spatial constraints of the mean fishing effort showed a strong chaining effect (Fig. 5). Thus it did not provide interpretable results and the choice of a particular partition is not obvious.

The application of the clustering algorithm to the estimated month and rectangle coefficients of the full model resulted in 22 fishing areas by grouping statistical rectangles (Fig. 6) and 9 fishing periods by grouping months (Fig. 7). The number of clusters was determined by the best cut on the dendrogram. These fishing areas and periods exhibit similar fishing exploitation patterns.

The identified seasons were: (1) January, (2) February, (3) March, (4) April, (5) May, (6) June, (7) July and August, (8) September, October and November and finally (9) December. January is the month during which the fishing activity is the most intense, closely followed by March and then July, August and April (Table 3). Given that January and March are not consecutive months, the clustering analysis did not group them in the same cluster. The fishing activity of the studied fleet was the least in December followed by February, November, May, October, September and June. Differences in month effects were larger in winter and spring than in summer (Table 3). All winter and spring months therefore appeared in separate clusters leading to seven distinct seasons in winter and spring

Table 3
Fishing seasons and estimated month effects with 95% confidence intervals using model (1)

Season number	Month	Lower bound	Month effect	Upper bound
1	January	0	0	0
2	February	−0.41	−0.64	−0.18
3	March	−0.01	−0.24	0.22
4	April	−0.15	−0.38	0.08
5	May	−0.39	−0.62	−0.16
6	June	−0.23	−0.45	0.004
7	July	−0.11	−0.34	0.11
	August	−0.13	−0.36	0.10
8	September	−0.29	−0.52	−0.06
	October	−0.33	−0.56	−0.10
	November	−0.40	−0.63	−0.17
9	December	−0.78	−1.01	−0.55

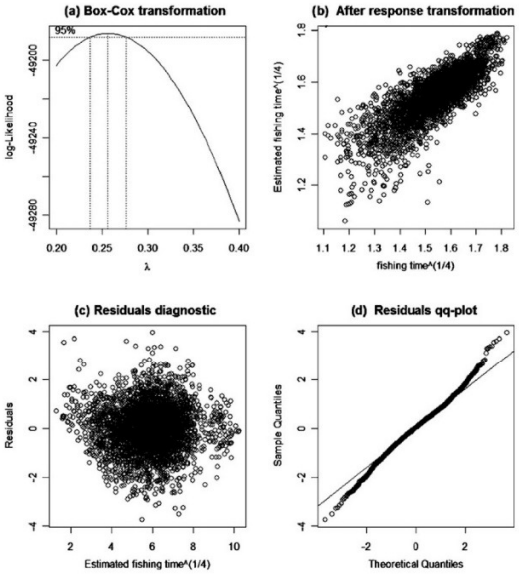


Fig. 4. Diagnostic plots: (a) log-likelihood plot for the Box–Cox transformation of fishing time. 95% confidence interval for the transformation indicates that 1/4 is a reasonable choice for the sake of interpretability, (b) fitted transformed fishing time versus observed transformed fishing time, (c) deviance residuals versus fitted transformed fishing time and (d) qq-plots of the deviance residuals. (b), (c) and (d) were generated using model (1).

and two in summer. December, January and February stood out as the months that were grouped last in the clustering process. With respect to the estimated month effects and the number of seasons, we concluded that fishing activity was less stable in winter than in summer. The longest season was obtained in autumn (September to November) and this season was characterized by an intermediate fishing activity.

The spatial clustering partitioned the rectangles into 22 areas of similar fishing effort, shown with the same colour with respect to their fishing effort level (Fig. 8). The number of rectangles per fishing area varied from one to six. The coastal clusters were smaller (for example, areas 1, 2, 4, 5) than the off-shore clusters (for instance, areas 12, 13, 22). Five off-shore fishing areas (areas 3, 6, 8, 24, 21) were however defined by a single rectangle. Area 3 (31E3 rectangle) and area 6 (25E3 rectangle) were the most visited fishing zones, whereas area 8 (27E1 rectangle) was among the least visited fishing zones. Regarding their estimated effects, fishing times in these ICES-rectangles contrast with fishing times in the neighbouring rectangles (Table 4). Areas 19 and 21 are located on the shelf break and probably are subject to both shelf and deep water fishing activities. This might explain their position within single-rectangle clusters. The fish-

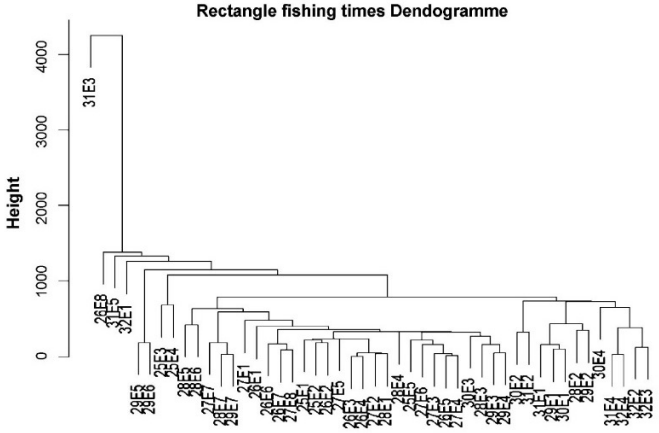


Fig. 5. Dendrogram showing single linkage clustering of the Euclidian distance between raw fishing times per rectangle (average fishing time per ICES-rectangle over the study period).

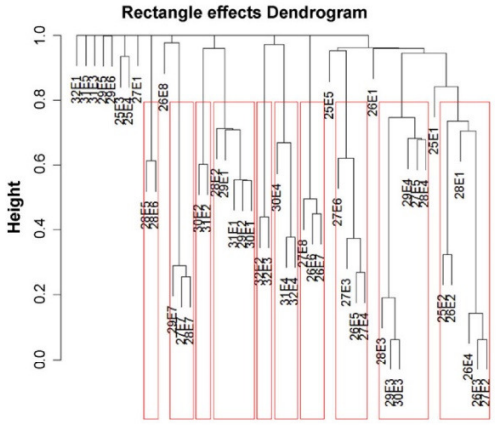


Fig. 6. Dendrogram showing single linkage clustering of $1-p$ values of tests on estimated coefficients of rectangles for which levels are different. The boxes characterize the grouped rectangles by cutting the dendrogram at height $1-p$ equals 0.75.

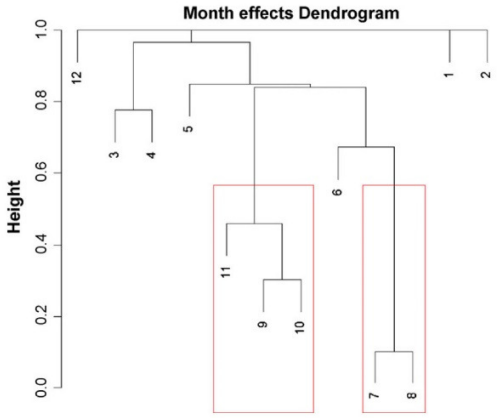


Fig. 7. Dendrogram showing single linkage clustering of $1-p$ values of tests on difference between levels of estimated model coefficients for *month*. The boxes characterize the grouped months by cutting the dendrogram at height p equals 0.5.

ing zones most visited by the French trawler fleet were located South of Ireland (areas 3, 12, 13), off Cornwall (area 9) and off the West of France (areas 6, 7). The waters close to the Irish and English cost were the least visited areas by the French fleet (areas 1, 2, 4, 5, 16). The largest homogeneous fishing area (6 rectangles, area 22) was located in the South-Western part of the Celtic Sea.

4. Discussion

4.1. Model-based clustering

We propose a method for characterizing spatial and temporal patterns in fishing effort based on a hierarchical cluster analysis of coefficients from a linear model with imposed constraints of spatial and temporal contiguity.

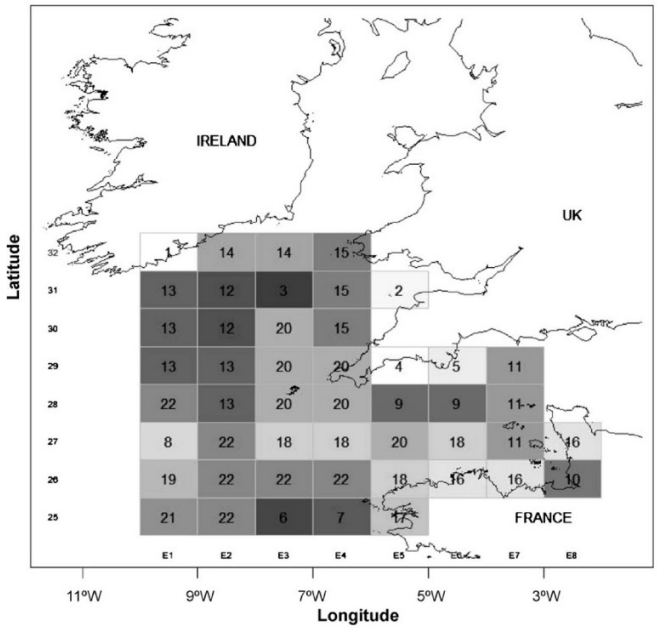


Fig. 8. Map of fishing areas resulting from the cluster analysis on estimated model coefficients for model of total fishing time by French trawler fleet in the Celtic Sea. All statistical rectangles belonging to the same fishing area have the same shading and the same number. The grey level is proportional to the estimated spatial fishing effort (average rectangle effect).

Table 4
Fishing zones and estimated rectangle effects with a 95% confidence intervals using model (1)

Zone number	ICES-rectangle	Lower bound	Rectangle effect	Upper bound
1	32E1	−1.61	−1.34	−1.07
2	31E5	−1.28	−1.01	−0.73
3	31E3	1.46	1.73	2.01
4	29E5	−1.62	−1.34	−1.06
5	29E6	−1.23	−0.96	−0.69
6	25E3	0.49	0.74	0.99
7	25E4	0.28	0.54	0.79
8	27E1	−0.94	−0.69	−0.44
9	28E5	0.10	0.35	0.61
	28E6	0.19	0.45	0.71
10	26E8	−0.03	0.23	0.49
11	27E7	−0.30	−0.04	0.21
	28E7	−0.27	−0.01	0.25
	29E7	−0.22	0.03	0.29
12	30E2	0.51	0.77	1.03
	31E2	0.42	0.68	0.94
13	28E2	0.39	0.64	0.90
	29E1	0.06	0.31	0.57
	29E2	0.27	0.52	0.78
	30E1	0.17	0.43	0.69
	31E1	0.089	0.35	0.60
14	32E2	−0.51	−0.25	0.01
	32E3	−0.44	−0.19	0.07
15	30E4	−0.03	0.22	0.48
	31E4	−0.14	0.11	0.37
	32E4	−0.20	0.06	0.32
16	26E6	−1.20	−0.93	−0.66
	26E7	−1.12	−0.86	−0.59
	27E8	−1.04	−0.77	−0.51
17	25E5	−0.58	−0.33	−0.07
18	26E5	−0.80	−0.55	−0.29
	27E4	−0.79	−0.54	−0.28
	27E4	−0.85	−0.59	−0.33
	27E6	−0.70	−0.44	−0.18
19	26E1	−0.45	−0.23	−0.01
20	27E5	−0.43	−0.17	0.08
	28E3	−0.56	−0.30	−0.05
	28E4	−0.31	−0.05	0.20
	29E3	−0.59	−0.33	−0.07
	29E4	−0.43	−0.17	0.09
	30E3	−0.60	−0.34	−0.08
21	25E1	0	0	0
22	25E2	−0.002	0.21	0.44
	26E2	−0.06	0.17	0.41
	26E3	−0.21	0.03	0.29
	26E4	−0.20	0.06	0.31
	27E2	−0.20	0.04	0.30
	28E1	−0.34	−0.08	0.17

Clustering methods are classically applied to raw data. However, this could be not appropriate when contiguity constraints are used. The single linkage criterion applied to the raw (noisy) data induced a strong chaining structure and the derived dendrogram did not allow valid clustering. In contrast, applying the proposed method to model coefficients, no chaining effect occurred in the dendrogram (Fig. 6) and a valid and interpretable partition could easily be performed (best cut at 22 clusters). Indeed, fitting a model accounting for month, rectangle and year explanatory variables allowed to estimate jointly seasonal and spatial effects while filtering the data from inter-annual variations. As the exploratory analysis showed the presence of autocorrelation structures in the fishing time data, the statistical model explicitly took them into account. Consequently, our model-based approach overcame the problem of chaining which is characteristic of single linkage clustering.

The method was illustrated by an application to fishing time data for the French trawler fleet fishing in the central Celtic Sea during the period 1991–1998. In the example, spatial correlations were modelled by an exponential function and temporal correlations by introducing previous month's fishing time in a given rectangle as predictor in the model. In this study, due to the absence of fishing in some rectangles during some months there was an unequal number of observations in each cell. The total number of observations is 4488 instead of 4608(=12 × 48 × 8) in the case of a balanced design, representing less than 3% of missing values. As our factorial design is unbalanced, the orthogonality property of main effects (and interactions also) present in balanced data is no longer valid (Montgomery, 2005). This means that changing the order of the factors in the model could lead to differences in estimated effects. Fortunately, in our case the imbalance is too small to impact the results of the model. Moreover, as the total number of observations compared to the number of missing values is large, the results of the hierarchical cluster analysis performed on $1 - p$ values from the F-tests are not affected by this problem of missing observations (the critical region of the F-test with 5% significance level is nearly the same $f95\%; 1{,}4421 \approx f95\%; 1{,}4541 \approx 3.84$). But, it is important to stress that our method is suitable only for balanced design or nearly balanced design for a large data set. In other cases, users can encounter problems with fitting the model or/and performing the cluster analysis.

In this paper, we selected as the full model a model with only main effects rather than a model with interactions for which model parameters would have been more difficult to interpret. In the Celtic Sea case study no clear interaction between month and rectangle was detected by exploratory analyses. Nevertheless, the introduction of a month × rectangle interaction into the model could improve its goodness of fit but would above all increase the difficulty of interpretation. If interactions were significant, the method might be adjusted to provide season-areas with similar patterns. Thus instead of clustering the coefficients of rectangles and months separately, the resulting estimated coefficients month × rectangle would be clustered in two steps. First fixing the month, it would lead to maps of fishing areas per month. Second, fixing the rectangle, it would lead to one seasonal year-split per rectangle. This approach will be considered in a future analyses using indicators to quantify temporal stability and spatial heterogeneity of fishing areas. Other recently proposed spatio-temporal models (Banerjee et al., 2004) such as STAR (spatio-temporal autoregressive) models or Bayesian spatial models might be another approach to deal with interactions.

In the context of mixed fisheries, management advice needs to be based on integrated approach accounting for fleets and species, spatial and temporal features of the fishery. Fishing time data inform on where, when and how long fishermen fish and hence this data is suitable for describing fishing zones and seasons. However, spatial management requires additional information to these raw definitions of zones and seasons to regulate fishing access. For instance it would be necessary to characterize each zone and season by the set of targeted species, using catch data available in commercial log-books as additional explanatory variables in the model. In this study, we have not used catch data given that reported catches often have a poor spatial resolution and might bias the perception of the catch process. Indeed, the onboard sorting process leads to discards which are not reported and still not-sufficient understood to provide an accurate esti-

mate of real catch (Rochet et al., 2002). Considering the fishery as a group of vessels sharing a similar fishing activity (ICES, 2004), an alternative approach integrating target species information might be to split the studied fleet into fisheries and then to apply the proposed modelling scheme to each fishery separately.

Finally, the approach used in this study is generic, and can be generalised to any response variable. For instance, the method could be applied for analysing spatial and temporal patterns in monthly catches.

4.2. Temporal and spatial fishing patterns in Celtic Sea

French trawlers had a tendency to return to the same statistical rectangles in subsequent months. This was shown by a positive temporal autocorrelation. Given that total fleet fishing times per rectangle were analysed, this pattern could have been produced by the same or by different vessels. Similar effects have been reported for the New England trawl fishery (Holland and Sutinen, 1999) and the New Zealand hoki fishery (Vignaux, 1996b). In Holland and Sutinen (1999), variables describing behaviour types, i.e., fishing areas visited during the previous 10 days, were the most important explanatory variables when modelling fishing revenues. Vignaux (1996b) found that the previous day's fishing location significantly explained the choice of the fishing area on a given day. In this study, spatial correlations of fishing times between neighbouring rectangles in a given month were found. This may be explained by assuming similar fish habitats and consequently similar population abundances and community structures. It must be noted that statistical rectangles are arbitrary spatial units. If independent information concerning the relevant ecological factors were available, these could be used in the model.

The cluster analysis performed on the model coefficients for months and rectangles resulted in 22 fishing areas and 9 fishing periods. Seasonal and spatial patterns, either in CPUE data (Vignaux, 1996b; Silvano and Begossi, 2001) or in fishing times (Greenstreet et al., 1999; Jenning et al., 1999; Béné and Tewfik, 2000), have been identified for many fisheries.

The activity of the French fleet was found to be stable during summer and autumn months (two seasons over 5 months). The same seasonal pattern was also pointed out by Greenstreet et al. (1999) for the UK trawler fleet in North Sea. In contrast, during winter and spring the fishing activity appeared to be much more variable from 1 month to another as each month formed a separate cluster (seven seasons over 7 months). In this analysis, December was not surprisingly identified as the month characterized by the lowest fishing time: the second part of December is known to be a period off, specially for French vessels going out for 1 or 2 weeks-trips.

The size of the estimated fishing areas was rather variable: one area consisted of more than six statistical rectangles (area 22) whereas others were made up of only one rectangle (area 1 to area 8, area 10, area 16, area 17, area 19, area 21). This suggests that the latter set of rectangles had their own particular fishing dynamics despite significant overall spatial correlations of fishing times. The South-Western Celtic Sea seems to be more homogeneous than the borders, which often consisted of single rectangle clusters such as cluster 1 close to the Irish coast and cluster 2 close to Cornwall. These results confirm earlier observations and might be explained by more pronounced heterogeneity in coastal fishing activities compared to off-shore fishing (Biseau et al., 1999; Greenstreet et al., 1999). However, three off-shore areas (8, 19 and 21) are clusters constituted by a single rectangle. Areas 8 and 19 are rectangles characterized by average monthly fishing efforts smaller than their six neighbouring rectangles, but no other obvious factors can explain this differences. By contrast, area 21 is a rather atypical fishing area. This rectangle is situated on the continental slope with large variations in bathymetry. As also indicated by the exploratory analysis (Fig. 1), the statistical rectangle 31E3 formed a cluster on its own characterized by consistently the largest fishing times (Table 4). This large discrepancy between this fishing area and the others has already been observed by Pelletier and Ferraris (2000) and is still observed in 2005 (SIH-Ifremer, 2007). This area is an attractive fishing zone because of the availability of valuable species. It is known to be visited by vessels targeting Nephrops (*Nephrops norvegicus*) (Coull et al., 1998) and has also been characterized by a large abundance of whiting (*Merlangius merlangus*) (Verdoit et al., 2003). The large fishing times in areas 13 and 12 might be similarly explained (Biseau et al., 1999). Regarding area 6 and area 7, large CPUEs of megrim (*Lepidorhombus wiffiagonis*) and monkfish (*Lophius piscatorius and L. budegassa*) (Petitgas et al., 2003; Biseau et al., 1999) but also the proximity to fishing ports might explained their high fishing times. These two areas are crossed on the way back, and consequently might be used for last fishing operations.

The fishing areas and periods obtained by our method can be used in several ways, for example as the basic units in a spatial management model (Pelletier et al., 2001; Mahévas and Pelletier, 2004; Pelletier and Mahévas, 2005) or in a spatial stock assessment model (Stefansson and Palsson, 1997) or to apportion global fishing mortality in space and time.

Acknowledgments

We would like to thank Ludger Evers and Céline Metaireau for carrying out preliminary analyses, Steven Juggins for suggestions for clustering algorithms, Richard Tomassone and two anonymous referees for comments on the manuscript. This study was performed using logbook data registered by the French Fishery ministry (DPMA) and extracted from Harmonie, the database containing the French Fisheries Information System managed by Ifremer. This work was funded by the European Commission, project contract CAFE (Contract no 022644).

References

Akaike, H., 1974. A new look at the statistical model identification. IEEE Transactions on Automatic Control 19, 716–725.
Babcock, E.A., Pikitch, E.K., 2000. A dynamic model of fishing strategy choice in a multispecies trawl fishery with trip limits. Canadian Journal of Fisheries and Aquatic Sciences 57, 357–370.
Banerjee, S., Carlin, B.P., Gelfond, A.E., 2004. Hierarchical Modeling and Analysis for Spatial Data. Chapman & Hall/CRC, New York, USA.
Béné, C., Tewfik, A., 2000. Analysis of fishing effort allocation and fishermen behaviour through a system approach. CEMARE No. 155.
Biseau, A., Gondeau, O., 1988. Apport des méthodes d'ordination en typologie des flottilles. Journal du Conseil International pour l'Exploration de la Mer 44, 286–296.
Biseau, A., 1998. Definition of a directed fishing effort in a mixed-species trawl fishery and its impact on stock assessments. Aquatic Living Resources 11 (3), 119–136.
Biseau, A., Maguer, C., Sanz-Aparicio, C., 1999. Pêcheries bigoudènes. Bilan des connaissances, Contrat CE (DG XIV) No. 97/0028.
Booth, A.J., 2000. Incorporating the spatial component of fisheries data into stock assessment models. ICES Journal of Marine Science 57, 858–865.
Cressie, N.A.C., 1993. *Statistics for Spatial Data*, Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley and Sons Inc., New York.
Coull, K.A., Johnstone, R., Rogers, S.I., 1998. Fisheries Sensitive Maps in British Waters. UKOOA Ltd, 58 pp.
Draper, N.R., Smith, H., 1998. Applied Regression Analysis. John Wiley and Sons, New York, USA.
Everitt, B.S., Landau, S., Leese, M., 2001. Cluster Analysis. Arnold, London, UK.
Gordon, A.D., 1981. Classification. Chapman & Hall/CRC, New York, USA.
Gordon, A.D., 1987. A review of hierarchical classification. Journal of the Royal Statistical Society 150 (2), 119–137.
Gordon, A.D., 1996. A survey of constrained classification. Computational Statistics and Data Analysis 21, 17–29.

Greenstreet, S.P.R., Spence, F.B., Shanks, A.M., McMillan, J.A., 1999. Fishing effects in north-east Atlantic shelf seas: patterns in fishing effort, diversity and community structure. II. Trends in fishing effort in the North Sea by UK registered vessels landing in Scotland. Fisheries Research 40, 107–124.
Hartigan, J.A., 1975. Clustering Algorithms. Wiley, New York.
Hartigan, J.A., Wong, M.A., 1979. A *K*-means clustering algorithm. Applied Statistics 28, 100–108.
Hilborn, R., Ledbetter, M., 1985. Determinants of catching power in the British Columbia salmon purse seine fleet. Canadian Journal of Fisheries and Aquatic Sciences 42, 51–56.
Holland, D.S., Sutinen, J.G., 1999. An empirical model of fleet dynamics in New England trawl fisheries. Canadian Journal of Fisheries and Aquatic Sciences 56, 253–264.
ICES, 2004. Report of the Study Group on the Development of Fishery-based Forecasts. CM/ACFM 11.
Jenning, S., Alvsvag, J., Cotter, A.J.R., Ehrich, S., Greenstreet, S.P.R., Jarre-Teichmann, A., Mergardt, N., Rijnsdorp, A.D., Smedstad, O., 1999. Fishing effects in north-east Atlantic shelf seas: patterns in fishing effort, diversity and community structure. III. International trawling effort in the North Sea: an analysing of spatial and temporal trends. Fisheries Research 40, 125–134.
Lebart, L., Morineau, A., Piron, M., 1997. Statistique Exploratoire Mutidimensionnelle, 2nd ed. Dunod, Paris, France.
Legendre, P., Legendre, L., 1998. Numerical Ecology. Developments in Environmental Modelling, vol. 20. Elsevier.
Mahévas, S., Trenkel, V., 2002. Utilisation de modèles mixtes pour décrire la distribution spatio-temporelle du temps de pêche de la flottille française en mer Celtique. Journal de la Société Française de Statistiques – Modèles Mixtes et Biométrie 143, 177–186.
Mahévas, S., Pelletier, D., 2004. ISIS-FISH, a generic and spatially explicit simulation tool for evaluating the impact of management measures on fisheries dynamics. Ecological Modelling 171, 65–84.
McCullagh, P., Nelder, J.A., 1989. Generalized Linear Models. Chapman and Hall, New York, USA.
Montgomery, D.C., 2005. Design and Analysis of Experiments, 6th ed. John Wiley and Sons Inc., New York.
Pelletier, D., Ferraris, J., 2000. A multivariate approach for defining fishing tactics from commercial catch and effort data. Canadian Journal of Fisheries and Aquatic Sciences 57, 51–65.
Pelletier, D., Mahévas, S., Poussin, B., Bayon, J., Andre, P., Royer, J.T., 2001. A Conceptual Model for Evaluating the Impact of Spatial Management Measures on the Dynamics of a Mixed Fishery. Spatial Processes and Management of Marine Populations Alaska Sea Grant College Program, AK-SG-01-02, pp. 54–66.
Pelletier, D., Mahévas, S., 2005. Fisheries simulation models for evaluating the impact of management policies, with emphasis on marine protected areas. Fish and Fisheries 6, 307–349.
Petitgas, P., Poulard, J.-C., Biseau, A., 2003. Comparing commercial and research survey catch per unit of effort: megrimin the Celtic Sea. ICES Journal of Marine Science 60, 66–76.
Pinheiro, J.C., Bates, D.M., 2000. Mixed-Effects Models in S and S-Plus. Statistics and Computing. Springer and Verlag, New York, USA.
Poulard, J.C., 2001. Distribution of hake (*Merluccius merluccius*, Linnaeus, 1758) in the Bay of Biscay and the Celtic Sea from the analysis of French commercial data. Fisheries Research 50, 173–187.
Rawlings, J.O., Pantula, S.G., Dickey, D.A., 2001. Applied Regression Analysis: A Research Tool, 2nd ed. Springer and Verlag, New York, USA.
Rochet, M., Péronnet, I., Trenkel, V., 2002. An analysis of discards from the French trawler fleet in the Celtic Sea. ICES Journal of Marine Science 59, 538–552.
Sampson, D.B., 1991. Fishing tactics and fish abundance, and their influence on catch rates. ICES Journal of Marine Science 48, 291–301.
Searle, S.R., 1997. Linear Models. Wiley Classics Library, New York, USA.
SIH-Ifremer, 2007. Synthèse des flottilles de pêche 2005. Flottle mer du Nord-Manche-Altantique. Système d'Informations Halieutiques de l'Ifremer, 55 pp.
Silvano, R.A.M., Begossi, A., 2001. Seasonal dynamics of fishery at the Piracicaba River. Fisheries Research 51, 69–86.
Sparre, P., 2003. An EXCEL-based software toolbox for stochastic fleet-based forecast. ICES Annual Conference, CM 2003/V:07.
Stefansson, G., Palsson, O.K., 1997. BORMICON: A Boreal Migration and Consumption model. Marine Research Institute No. 59. Reykjavik, Iceland.
Verdoit, M., Pelletier, D., Bellail, R., 2003. Are commercial logbook and scientific data useful for characterizing the spatial and seasonal distribution of exploited populations? The case of the Celtic Sea whiting. Aquatic Living Resources 16, 467–485.
Vignaux, M., 1996a. Analysis of spatial structure in fish distribution using commercial catch and effort data from New Zealand hoki fishery. Canadian Journal of Fisheries and Aquatic Sciences 53, 963–973.
Vignaux, M., 1996b. Analysis of vessel movements and strategies using commercial catch and effort data from the New Zealand hoki fishery. Canadian Journal of Fisheries and Aquatic Sciences 53, 2126–2136.

3. Statistical Tool for Dating and interpreting archaeological contexts using pottery.

Bellanger L., Husi P. (2012). *Journal of Archaeology Science*, **39**(4): 777-790

Journal of Archaeological Science 39 (2012) 777–790

Contents lists available at ScienceDirect

Journal of Archaeological Science

journal homepage: http://www.elsevier.com/locate/jas

Statistical tool for dating and interpreting archaeological contexts using pottery

Lise Bellanger [a], Philippe Husi [b,*]

[a] Université de Nantes, Laboratoire de Mathématiques Jean Leray – UMR 6629, CNRS/Université de Nantes, France
[b] CNRS UMR 6173 CITERES, Laboratoire Archéologie et Territoires, CNRS/Université de Tours, France

ARTICLE INFO

Article history:
Received 30 March 2011
Received in revised form
20 June 2011
Accepted 26 June 2011

Keywords:
Pottery analysis
Chronology
Probability density curves for dating archaeological contexts
Archaeological tool for socio-economic and functional analysis
Centre-west of France
Medieval period

ABSTRACT

Analyzing chronological patterns is one of the major issues in archaeology. How can the date of a specific context be estimated? Is it possible to identify residual and intrusive material in it at the same time? Numerous statistical methodological approaches have been developed and implemented to estimate dates but have less often addressed the issue of socio-economic area or the functional interpretation of contexts. This article deals with the construction and analysis of two different probability estimate density curves of context dates using pottery. By contrasting the two curves we can define the boundaries of the socio-economic area and make a chrono-functional interpretation of a context. This statistical tool allows the archaeologist to visualize and analyze chronological patterns easily. The method is applied to the analysis of contexts in the town of Tours in particular and more generally in the centre-west of France, based on collected pottery finds.

1. Introduction

The issue of time, omnipresent in archaeology, which includes the relationship to the object and to the archaeological context,[1] can be approached in three ways:

- The time-span linked to the pattern and thus to the intensity of occupation
- The succession of events or changes over time, linked to relative chronology
- The date or dating linked to absolute chronology

For the archaeologist, these three aspects of time are often merged or closely linked (Ferdière, 2007: 15).

There are three broad categories of dating methods in archaeology, which are frequently interrelated:

- The most common laboratory methods are 14C, dendrochronology, thermoluminescence and archaeomagnetism, which raise a number of problems for archaeological dating (Seigne, 2007);
- Those that refer to historical events, which must be firmly connected to the archaeological contexts, the sources used acting at very different chronological scales;
- "Traditional" methods, based on the artefacts found in the archaeological context; in practice these are the ones most commonly used by archaeologists.

In this article, we focus on dating methods using archaeological artefacts. Detailed examination of these methods reveals a number of methodological problems. They are based on two types of object: those – rare – which bear their own date (coins, epigraphic documents, etc.), and those which are only dated with reference to a chrono-typology (pottery and other artefacts). In these cases, as in methods based on historical events, it is important to examine the relationship of the object to the context in order to assess its use for dating, which could vary according to the nature of the context (building, midden, domestic occupation, rubble etc.). The object selected for dating provides the context with a *terminus post quem* (earliest possible date). For a given context, the relative chronology of the stratigraphic deposit then enables the possible dating interval to be narrowed down by providing the *terminus ante quem* (latest possible date) (Ferdière, 2007: 16).

* Corresponding author.
E-mail addresses: lise.bellanger@univ-nantes.fr (L. Bellanger), philippe.husi@univ-tours.fr (P. Husi).

[1] An archaeological context is composed of one or a series of levels (stratigraphic units) that are temporally and functionally interpreted (rubbish pits, levels of construction or home occupation of a house, etc.). Sometimes a context corresponds to several archaeological levels, but the reverse is not true.

0305-4403/$ – see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jas.2011.06.031

Consequently, it appears that the common problem of these three methods arises from the type of dating object: in general, it is the cutting, manufacture, issue (coins), or production (pottery) of the object that is dated, or the intensity of its use, its abandonment, or re-use, and not the context. Thus, the date range gives a more or less precise overall dating for the context, but provides no information about the actual duration or intensity of occupation. It is this duration which is often very difficult to ascertain.

For pottery, the emergence of production types can generally be fairly well determined chronologically; their disappearance is more difficult to date precisely, due in particular to problems related to usage, residuality and the geographical area concerned (Kulpa, 1997; Van de Weghe et al., 2007). This explains the interest of using common typological criteria to constitute the pottery profiles of archaeological contexts, and then of referring to dating elements such as non-residual coins. Each dating system has its own methodological bias, and it is thus important to compare the sources used in order to ensure the consistency of the proposed dates, while keeping in mind exactly what is being dated.

It is then possible to:

- Obtain a coherent absolute dating, from the relationship between the date of the coins or other dating element and the pottery profiles;
- Attempt to determine the time-span of an archaeological context, often reduced to a succession of events;
- Use the chronological profiles of the archaeological contexts based on the dated pottery profiles to gain a better understanding of socio-economic and functional mechanisms.

With this objective, a group of archaeologists and statisticians recently joined forces to work on the question of chronological modelling. In line with Baxter (2008), who considered that questions raised by archaeologists could be resolved by statistics, our aim was to apply statistical analysis to archaeological data. The purpose of this article is not to provide precise details of the pottery model constructed to establish the absolute date of archaeological contexts; these details can be found in Bellanger et al. (2006a, 2006b, 2008).

A large number of papers, some of which are reference works, have provided a clearer understanding of archaeological data through the use of statistics (Baxter, 1994; Orton, 1980, 2000; Orton and Tyers, 1992; VanPool and Leonard, 2011). In this article, we discuss how statistics can be used to clarify the issue of time in archaeology. The idea of using the distribution of pottery to achieve a better characterization of the time and the functional nature of archaeological contexts is not new; what is new here are the methods used to address the issue.

For each archaeological context, we used pottery to construct and analyze two different probability density curves of estimated dates, one based on the dating of events in calendar time (hereafter called “event time”), and one based on duration/time-span (hereafter called “accumulation time”). The chain of reasoning for a given archaeological context can be summarized as follows:

- Obtain two density curves representing the dating of an archaeological context.
  - Step 1: The first is a Gaussian curve derived from a linear regression model. This allows the mean date of issue of a coin to be estimated from the pottery profile. From an archaeological point of view, this estimation of a *terminus post quem*, with all the biases involved in the use of coins for dating, provides a fundamental basis for chronology-building in calendar time;
  - Step 2: The second curve is a mixture of Gaussians derived from the previous model. The date of an archaeological context is estimated using the weighted average of the estimated dates of the pottery fabrics found within it. Assuming that the fabrics are statistically independent, the associated probability density of the date can be estimated as a weighted sum of Gaussian distributions. From an archaeological point of view, this dating estimation provides a closer representation of accumulation time recorded in the soil (Olivier, 2001; Wirtz and Olivier, 2003). At best, depending on the quality of the archaeological context, it can be interpreted as a formation process reflecting the duration or succession of events on the scale of archaeological time, and at worst, as imprecise dating due to contamination of the context by residual or intrusive material.
- Compare the density curves, in order to:
  - Validate the method from a chronological perspective, exploring the duration or intensity of occupation for each context;
  - Identify the boundaries of socio-economic entities within the broader area of the centre-west region of France, using a chronological model based on the Tours reference site;
  - Obtain a clearer understanding of chrono-functional issues through a better interpretation of the type of archaeological context.

2. Archaeological material and statistical methods

2.1. The corpus of archaeological data

The choice of corpus is important on account of its strong impact on subsequent archaeological interpretations. It is essential that only the least disturbed contexts are retained, in other words those likely to reveal the greatest amount of material contemporary with the action interpreted by the archaeologist. Our corpus was composed of 278 archaeological contexts, of which 95 were reference contexts and 183 were supplementary contexts whose dates were ill-defined or unknown. These supplementary contexts, 37 in Tours and 146 from other sites in the centre-west of France, were not used in the construction of the pottery model (Appendix 1).

Apart from making a careful selection of relevant archaeological contexts, it is also important to consider their geographical location when investigating socio-economic issues in such a large area. Thirteen geographical sites were studied, from Nevers in the east to Poitiers in the west, and including Bourges, Orléans, Blois, Tours, Chinon and Châtellerault, with additional outlying contexts, for example, Limoges in the south and Jublains in the north. The 278 contexts in the corpus also cover a long calendar period, from the Middle Ages to modern periods (6th–17th centuries), albeit unevenly.

Pottery types were classified according to the workshop where they were made when this was known, or to the pottery-making tradition to which they belonged. In order to identify socio-cultural or pottery culture areas, popularity and competition between products must be taken into account. This is based on the assumption that fabrics showing technical similarities but probably made in workshops in different areas and distributed locally, could nevertheless belong to the same pottery-making tradition.

In this study, once the fabrics for each site in the study zone (the centre-west of France) had been identified, inventoried and individualized, only the 200 fabrics with features common to all the

archaeological contexts were retained in a database[2] and used to construct the corpus. In this way, we were able to work with a very large corpus of 15,044 items, calculated using a *Minimum Vessel Count* (MINVC), from a very large geographical area and covering a long period of time. This procedure, which is original and unique in the archaeological field, provides the basis for the statistical modelling of archaeological problems related to dating. Various quantification techniques (Sherd count and Estimated vessel equivalents, occurrence/non-occurrence) were used in the study (Orton, 1975, 1989, 1993). The purpose of this article is not to compare available methods, and hence we will only present the MINVC results, which provided the most satisfactory results.

This approach required extensive pottery analysis work in order to harmonize the methods developed over more than 15 years by a small group of researchers and to set up common typological tools and quantification methods in this very large study area in the west of France (Husi, 2006). To enlarge our field of investigation, a network of researchers, currently covering French-speaking areas of Europe, was created around an internet site designed as a base for spatialized data and common regional typological tools and site notices; the name of this network is ICERAMM (abbreviation of *Information sur la CERAmique Médiévale et Moderne*) (http://iceramm.univ-tours.fr/).

2.2. Statistical modelling: construction of a pottery model

Dating an archaeological context is always delicate. It is important to keep in mind what is to be dated, either a one-off event occurring at a specific moment, or a process built up over time. This study proposes datings based on the statistical modelling of pottery data, in contrast to "traditional" datings which are frequently based on an intuitive comparison of pottery assemblages. Of course, this does not in any way deny the importance of the pottery expert's knowledge in the dating process; on the contrary, it enables this knowledge to be better integrated through a systemic procedure, which is essential when there is a very large corpus of data to be analyzed.

The methodology is based on a statistical and visual approach using two estimated density curves to date each archaeological context. Two steps were required in the statistical procedure, each leading to the construction of a density curve. The first enabled us to estimate a date corresponding to the *terminus post quem* of the context, a cursor reflecting an event dated in calendar time. As further information about this step can be found in the studies referred to above and in the Appendix 2, only a brief summary is provided below. The second step, based on the results of the first, allows the chronological profile of the context to be estimated, giving a picture that is closer to archaeological time, in other words the rate of accumulation.

2.2.1. Step 1: modelling events dated in calendar time (dateEv)

This step involves estimating the date of an event recorded in the ground (an archaeological context for the archaeologist) based on the pottery assemblage of which it is comprised. This is achieved by fitting a regression model linking a known date in calendar time, in this instance the date of issue of a coin, to its pottery profile. The reference corpus used to fit the regression model comprised the contexts of the city of Tours containing coins, chosen for their chrono-stratigraphic quality. We then used the previously adjusted model to calculate a predicted value for contexts not included in the reference corpus, and sometimes stratigraphically separated or poorly documented, with a 95% prediction confidence interval for the dating.

The statistical procedure was as follows:

- A correspondence analysis (CA) was carried out to summarize the information in the reference corpus data; the data matrix consisted of the archaeological contexts and fabrics quantified using MINVC in all the Tours contexts, whether or not they had been dated with coins. This produced 95 contexts with pottery profiles composed of 200 fabrics. We then kept only the first ten factorial axes, accounting for approximately 64% of the total variance. In this way, our contingency table, crossing 95 contexts and 200 fabrics, becomes a 95*10 table, *an incomplete reconstitution of the data*. This principle is used in many factor analysis techniques, providing a way of reducing the number of explanatory variables in the linear regression model (in this case from 200 to 10. On this basis, the other 190 components may reasonably be ignored!) (Benzécri, 1973; Greenacre, 1984; Moreau *et al.*, 2000; Saporta, 2006).
- In order to estimate a date for the context, it is essential to refer to objects dated by another source, in this instance, non-residual coins. These contexts were selected on a very strict basis for their chrono-stratigraphic reliability, level of domestic occupation, or enclosures with long urban stratigraphic sequences, thereby minimizing any bias linked to disparity between the date of the coin and that of the context. The interest of this choice also lies in the fact that the dated contexts are distributed evenly between the 6th and 17th centuries. A *Gaussian multiple linear regression model* (cf Eq. (1)) is then fitted on these 25 contexts on the basis of the significant factorial components of the CA among the ten selected, in other words, only the first eight, represented as F^k ($k = 1,\ldots,8$). It can be expressed as follows:

$$dateEv_i = \beta_0 + \sum_{k=1}^{8} \beta_k \left(F^k\right)_i + \varepsilon_i \quad \forall i = 1,\ldots,25 \tag{1}$$

where ε_i are normally, identically and independently distributed random variables following an $N(0; \sigma^2)$, and $(F^k)_i$ is the ith coordinate of the kth factor component ($i = 1,\ldots,25$ and $k = 1,\ldots,8$). These explanatory variables are known constants, and the unknown β parameters have to be estimated.

After estimating the β parameters of the model using the classical results of the multiple regression analysis and checking that the model fits the data correctly (we obtained a R^2_{aj} of approximately 0.9973, thus very close to 1, and a residual standard deviation of approximately 23 years!), we can deduce the estimated date of a context and also predict that of a different context which has no coins and is thus not dated. This estimation or prediction[3] is based solely on the information contained in the pottery profile of the studied context and linked to the response variable of the regression model (date of coin), in other words on the first eight significant factor components of the CA. In this way we obtain:

- An estimated value and 95% confidence interval of the event time for the so-called "active" contexts;

[2] This is a computerized system for recording archaeological data developed by the "*Archéologie et Territoires*" Laboratory (UMR 6173 CITERES-LAT), called *ArSol* (*Ar*chives du *Sol* = Soil Archives). In addition to the part devoted to processing stratigraphic data, this system has a pottery analysis module, called BaDoC (*Ba*se de *D*onnée *C*éramique = pottery database) (Galinié et al., 2005; Husi, Rodier, due for publication in 2011).

[3] The term 'prediction' is used when no coins were found in the context.

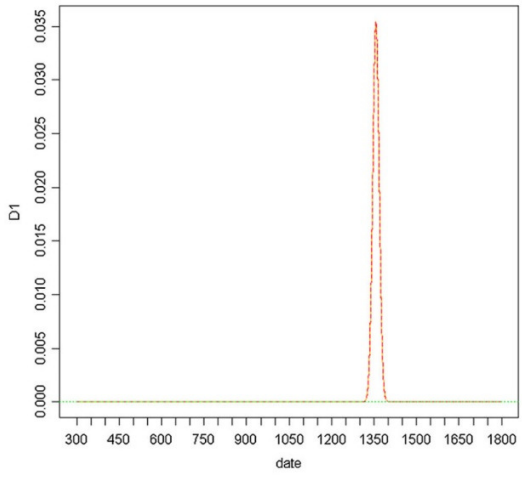


Fig. 1. Density curve *(dateEv)* of the D1 archaeological context.

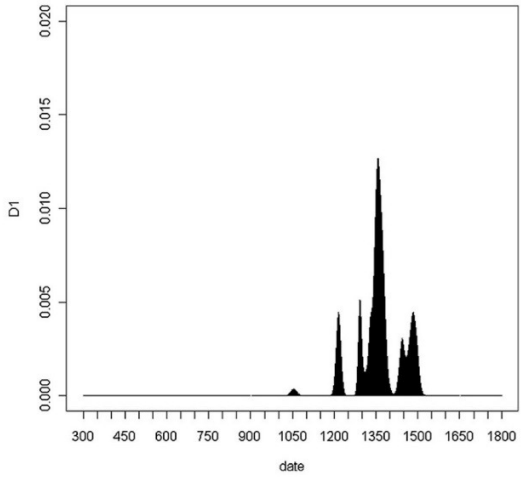


Fig. 2. Density curve *(dateAc)* of the D1 archaeological context.

- A predicted value and 95% confidence interval of the event time for the non-dated "supplementary" sets, which may or may not come from sites other than those used for constructing the model. The statistical results are then validated by external evidence such as the occurrence of coins not included in the model and stratigraphic reasoning.

In this way, whatever the nature of the context, either active or supplementary, by using the classic results of the Gaussian linear model it is possible to represent the estimated probability density corresponding to the event time (*dateEv*) of each context. It can be approached by a normal distribution of parameters varying with each context (Appendix 2 for further mathematical details). It is represented by an orange curve in the following figures. By definition, the area under the density curve has a value of 1 (i.e. 100%).

Below is an example of a curve for the D1 archaeological context (a garden used for dumping household waste) dated by coins at 1341: we obtain a 95% confidence interval for a *dateEv* of [1333; 1381] and a density curve (Fig. 1).

2.2.2. Step 2: from event time (dateEv) to accumulation time (dateAc)

We used the results of the first step and the properties of the CA to obtain an estimated value of the date of each fabric $((dateEv)^j; j = 1, \ldots, 200)$ and a 95% confidence interval. We could then define the archaeological time shown as *dateAc*, in other words the accumulation time of a context, as the weighted sum of fabric datings; the weights are the proportions of *MINVC* of each fabric in the context. Assuming that the random $dateEv^j$ variables are independent,[4] the distribution of accumulation time of every i_0 context can be approached by the Gaussian mixture.

In this way, for each context we obtained a plurimodal density curve representing the estimated law of accumulation time based on the mixture of normal densities (dating of each fabric). This density is represented by the black curve in the figures in this article. By definition, the area under the density curve has a value of 1 (hence 100%). Below is an example of the D1 context dated by coins at 1341, with a 95% interval for a *dateAc* of [1206; 1499] and its density curve (Fig. 2).

The orange and black density curves are interpreted using a graphic tool which combines them on a single graph (Fig. 3). In this example, it can be seen that the peak representing the estimated event time (orange curve) is superimposed on the main peak of the black curve. Depending on the nature of the context, the small peaks that can be seen on either side of the main peak of the black curve can be interpreted as representing residual or intrusive material, and/or as the material recording longer occupation, whose main activity corresponds to the point where the two curves overlap. If we accept this hypothesis, the occupation is likely to involve an activity occurring over a long period of time rather than

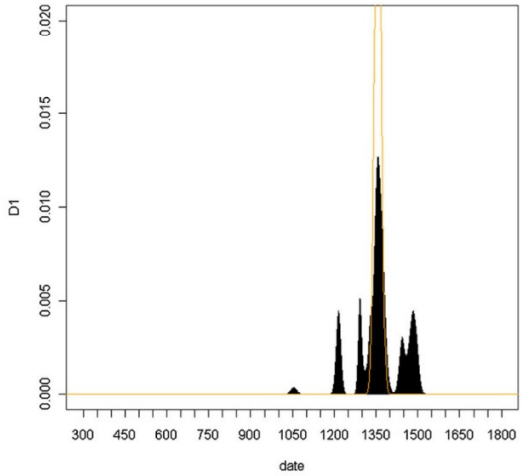


Fig. 3. Juxtaposition of two curves *(dateEv* and *dateAc)* for the D1 archaeological context.

[4] Variables are different fabrics: in this case, statistical independence appears to be a realistic archaeological hypothesis.

a one-off action, without however any evidence that the width of the main peak of the black curve represents the whole of this occupation. Herein lies the ambiguity between the notions of accumulation and time-span!

3. Results and discussion: interpretation of dating curves

The methodology presented here, illustrated by curves and hence the construction of a model, is an original archaeology dating tool. It formalizes to some extent the widely used intuitive procedure for dating archaeological contexts from pottery and the rare artefacts at our disposal that are dated in calendar time. It also provides more specific answers to archaeological issues: in the socio-economic domain, by defining a field of spatial validity of the temporal statistical model; in the functional domain, by characterizing the nature of archaeological contexts based on a typology of the chronological patterns obtained. We will now present several examples to illustrate these applications of the model, starting with the most important – dating.

3.1. Dating: relevance of the pottery model

Our purpose here is not to describe the method, but to look at its relevance through a number of examples. We will present a near perfect example in Tours and compare different dating methods, first to validate the procedure and secondly to understand better the relationship between time and object which varies according to the method used. It is up to the archaeologist to interpret these datings.

3.1.1. Tours: excellent adequacy of the model

Before trying to validate the relevance of the procedure, we will present a textbook case: a context of domestic occupation (context J1) from a site in Tours, with two curves presenting a near perfect profile. The orange curve representing the date of a coin and the black curve reflecting the chronological profile of the context are almost perfectly matched (Fig. 4). We can conclude that the residual or intrusive contaminating material is marginal in this context and that most estimated dates of the fabrics range from the end of the 15th to the beginning of the 16th centuries, as shown by the black curve. Furthermore, if any coins were found in this context, there is a 95% chance that they would come within this same period of time, as indicated by the orange curve. By comparing these results with the stratigraphy and with the actual composition of this pottery assemblage, we can narrow down the period of time, the material probably accumulating over a relatively short period.

This context (J1) is the back of a courtyard which served as a rubbish dump. Unlike the previous example (context D1, Fig. 3) which involved a process of accumulation over time, the use of this courtyard as a rubbish dump appears to have been fairly brief, linked more to a one-off action. As this near perfect example is not a typical case, it is important to use these results as an aid for the chronological interpretation of the contexts and not as an automatic dating tool. We will attempt to show this by comparing our results with those obtained using other dating methods.

3.1.2. Pottery and archaeomagnetism models: what is being dated?

The results must be interpreted critically, not only situating the dated context within the chrono-stratigraphic background of the site, but also understanding the reality of the dated object according to the source and methods used. The following example is from the site of La Vermicellerie in Fondettes, located about 10 km west of Tours on the banks of the Loire (Gaultier, 2011). It is a large rural site, composed of small, frequently isolated structures, in a small stratigraphic sequence.

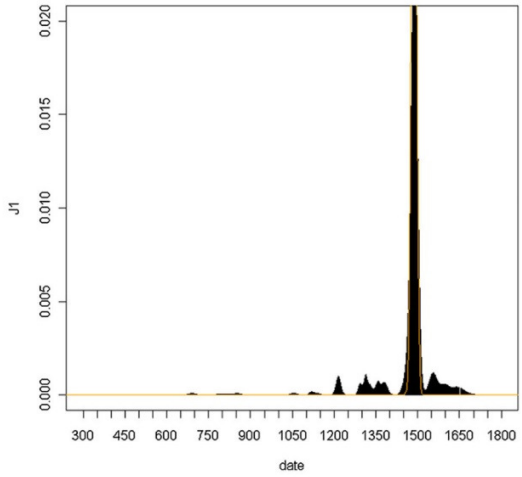


Fig. 4. Curves *(dateEv and dateAc)* for the J1 archaeological set.

We chose a context (Z028) consisting of a series of four domestic ovens, which were used as ash pits after their abandonment. These changes of use were deduced by comparing the archaeomagnetic datings of three of the ovens with those estimated by the pottery model. Before comparing the datings, it is essential to define what is actually being dated using these two distinct approaches. While archaeomagnetism dates the last use of the oven, the pottery model dates the use of the structures as ash pits after the ovens had been abandoned. At best, and logically, there may be a discrepancy between the two datings, the former being slightly earlier than, or contemporaneous with the latter. This is confirmed by our results, as the archaeomagnetic dating gives a confidence interval of (95%) between 515 and 645, while the first step of the pottery model gives a confidence interval for *dateEv* of between 631 and 690 (Fig. 5). Any coins would thus have a 95% chance of coming within this interval, and thus the *terminus post quem* of this context is very unlikely to be earlier than 631. The analysis of the results of the second step of the pottery model, with a confidence interval for *dateAc* of [530; 810], follows a similar pattern, as the main peak starts, like the former, at about 630, albeit with a slight overall shift to the right (Fig. 6). Comparison of the two curves (Fig. 6) reveals two small peaks prior to 630 on the black curve, undoubtedly corresponding to residual fabrics.

In view of the fragility of these structures, we can assume that they had a short life-span, and that as soon as their initial use ceased, they were used as ash pits before being abandoned. If we accept this hypothesis, the most likely dating for the last use of the ovens must have been between 631 and 645. Indeed, the rapid, almost contemporaneous succession of uses restricts the range of the archaeomagnetic interval. It also raises the hypothesis that, give or take a few years, the latest possible date of 645 could serve as *terminus ante quem* (TAQ) for the use of the ovens as an ash pit.

This example clearly illustrates the importance of comparing the estimates obtained by different dating methods, while keeping in mind the specific internal logic and interpretative limits of each. In this example, the comparison enabled us to validate our procedure, but in other cases, it could also reveal discrepancies between dating methods which could enhance or raise questions about the archaeological interpretation.

Pottery model *dateEv* (orange curve): CI 95%		
Earliest possible date	Estimation	Latest possible date
631	660	690
Pottery model *dateAc* (black curve) : CI 95%		
Earliest possible date	Estimation	Latest possible date
530	694	810
Archaeomagnetic dating of 3 ovens: CI 95%		
Oven 325: (470-645) most probably (515-645)		
Oven 326: (515-645) no probability		
Oven 217: (515-645) no probability		

Fig. 5. Results of the pottery model and archaeomagnetism for the domestic ovens at Fondettes (Z028).

3.2. From the pottery model to the socio-economic and cultural interpretation of archaeological contexts

- By changing the scale of application of this modelling procedure from the town (Tours) to a much larger space, in this case the centre-west of France, it indirectly becomes an aid to interpreting socio-economic mechanisms, as it allows the boundaries of socio-economic entities to be more clearly identified. This can be achieved by estimating the dates of domestic archaeological contexts further and further from the reference point (i.e. Tours). As the curves are constructed by comparing pottery assemblages including recipients of similar fabrics or with identical production traditions, the spatial range of the chronological pottery model can thus be determined. It is not surprising to observe that the further the modelled assemblages are from the reference point, the less accurate the dating estimates become. This is not due to an error in the choice of corpus or in statistical procedure, but to the distance from the reference point, revealing a chronological difference in the production and use of pottery with similar production traditions. This spatial limit of the model's field of application indicates that identical traditions existed at different times in areas that had no real economic contact with each other, apart from phenomena of know-how, popularity or competition (Husi dir. for publication in 2012). Hence, using chronology to explain socio-economic facts provides a wealth of information, as illustrated by the following examples.

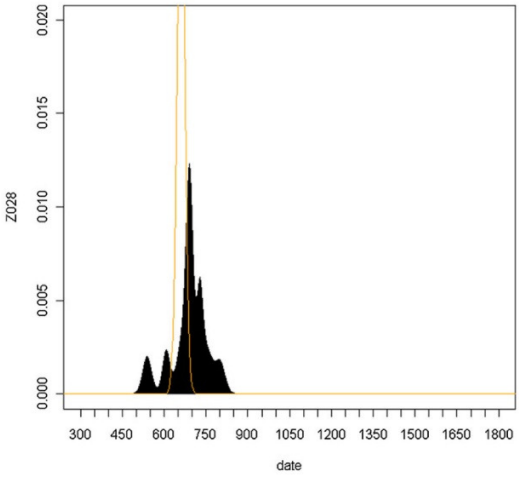


Fig. 6. Curves *(dateEv* and *dateAc)* for the domestic ovens.

By applying our model to contexts that are at a greater or lesser distance from the reference point, we can define a homogeneous socio-economic area around that point, and identify the sites which are further away on the basis of their pottery profile (Appendix 1).

3.2.1. Application of the model to the Touraine area

Our purpose here is not to look further at examples from the reference site of Tours or neighbouring sites such as Fondettes which belong to the same socio-economic area, but to move gradually further afield in order to analyze cases where our model is either no longer valid or continues to be valid despite the distance. The rural site of Truyes, 30 km south-east of Tours, comprises a series of five agricultural units with successive levels of settlement, occupation and abandonment (Tourneur, 2005). The units are non-contiguous and thus have no stratigraphic link. To illustrate our point, we selected the occupation contexts of two units only. Generally speaking, all the agricultural units of this site have identical curve profiles. Analysis of the two selected contexts reveals that the main peaks of the two curves are very close, even if they do not exactly match (Fig. 7). The slight shift of the orange curve to the right can be explained by the fact that a *terminus post quem* was estimated from the date of issue of coins, an artefact that is often hoarded and thus at best contemporaneous with but often slightly earlier than the pottery. In fact, the issue date rarely reflects the period during which the coins were used. We can also observe small peaks on either side of the main peak of the black curve, indicating residual and intrusive material.

The quality of the curves clearly shows that this site comes within the Touraine economic area. Support for this hypothesis can be found by referring to the last paragraph of this article, devoted to functional analysis and which also includes curves for another site, Neuvy-le-Roi, 30 km north of Tours (*infra* Fig. 13) (Tourneur, 2004). While these curves are used to demonstrate a different point, they show similar profiles to those of Truyes and reinforce the idea that pottery areas do not extend beyond a radius of 30 km around a main consumption centre, such as Tours. All the estimated dates match the stratigraphic analysis of the sites perfectly. The only slight short-coming is that there are currently no elements for comparison with laboratory methods, as in the case of the Fondettes site described above.

Z090: Pottery model *dateEv* (orange curve): CI 95%		
Earliest possible date	Estimation	Latest possible date
635	666	697
Z090: Pottery model *dateAc* (black curve): CI 95%		
Earliest possible date	Estimation	Latest possible date
597	687	814
Z091: Pottery dating *dateEv* (orange curve): CI 95%		
Earliest possible date	Estimation	Latest possible date
611	641	671
Z091: Pottery dating *dateAc* (black curve): CI 95%		
Earliest possible date	Estimation	Latest possible date
594	673	807

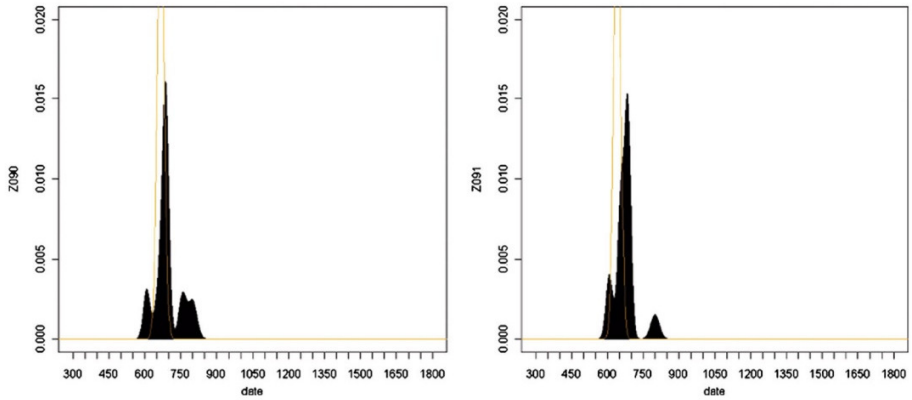


Fig. 7. Results of the pottery model and curves *(dateEv* and *dateAc)* for the agricultural units of Truyes.

3.2.2. Limits of the model linked to the geographic distance from the reference site

The example chosen to illustrate this point is the historic city of Blois, located on the River Loire approximately 65 km east of Tours. The selected archaeological contexts, unearthed on the castle site, are a series of middens, represented here by two rubbish pits (Aubourg et al., 1993). The almost perfectly matched, overlapping curves, together with the very narrow confidence intervals, demonstrate that the model is not in doubt (Fig. 8). The first midden (Z002) includes a coin (864–875) which validates the results of the model perfectly (this external date was not involved in the construction of the model, as all the Blois contexts are supplementary). By contrast, the results obtained for the second midden (Z004) show a discrepancy between the estimate based on the model and the dates provided by the *in situ* coins. In fact, the latest possible date proposed by the model (steps 1 and 2) does not extend beyond the end of the 9th century (888 and 881 respectively), while the coins give an earliest possible date (*terminus post quem)* of 887. The same observation can be made for a large number of other contexts in Blois.

A dual explanation is possible: a small number of fabrics common to the two towns, combined with a continuity of fabric in common use in Blois during these periods. These two elements make it difficult to compare the pottery assemblages and thus to use the model constructed from contexts from the Tours reference site to estimate the date of contexts unearthed in Blois. This result provides valuable elements for investigating socio-economic issues regarding the circulation of products and trade networks within short distances. While the orange and black curves are almost perfectly matched, the model no longer appears to be as relevant; this reinforces the hypothesis that the date provided by the model will become increasingly inaccurate as the distance between the reference site and the context site to be dated increases. Thus, while the sites at Fondettes, Truyes and Neuvy-Le-Roi (to which could be added Chinon and Rigny which are about 40 km south-west of Tours but not discussed here) appear to belong to the same socio-economic entity as Tours, Blois does not, which in itself is an invaluable finding for the archaeologist!

In this way, the limit of the model's application in relation to the distance from the reference site provides a means of identifying socio-economic and cultural spaces. Is distance the only factor involved, or is it just one among several?

The "Pouthumé" site in Châtellerault has provided a certain number of archaeological contexts, including a level of occupation of a building (Z019), which offers a good example of the boundaries of the pottery model. The choice of this site is justified largely by a calibrated 14C dating (892–1022). It contradicts neither the pottery chrono-typology nor the chrono-stratigraphic coherence of the site (Cornec et al., 2006). However, the pottery model gives a *dateEv* confidence interval of between 1045 and 1084 and

Z002: Pottery dating *dateEv* (orange curve): CI 95%		
Earliest possible date	Estimation	Latest possible date
811	858	894
Z002: Pottery dating *dateAc* (black curve): CI 95%		
Earliest possible date	Estimation	Latest possible date
675	853	1126
Z004: Pottery dating *dateEv* (orange curve): CI 95%		
Earliest possible date	Estimation	Latest possible date
811	849	888
Z004: Pottery dating *dateAc* (black curve): CI 95%		
Earliest possible date	Estimation	Latest possible date
735	853	881

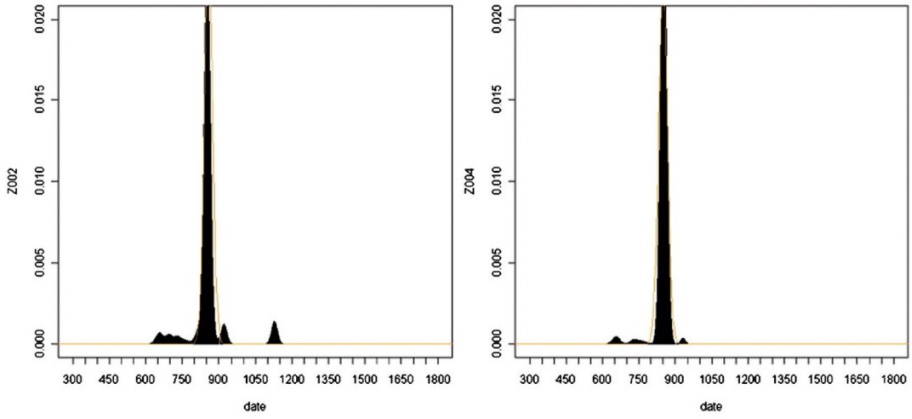


Fig. 8. Results of the pottery model and curves *(dateEv* and *dateAc)* for the enclosed contexts of Blois.

a chaotic estimated *dateAc* density (black curve) with numerous peaks (Fig. 9). How can we explain this complete mismatch between the results of the pottery model and the 14C dating?

As in the case of Blois, the main reason appears to be the distance between the reference site (Tours) and the site with contexts to be dated. However, while this explanation was acceptable for Blois, it is not sufficient in the case of "Pouthumé" because of the very chaotic shape of the *dateAc* density (black curve), although Châtellerault and Blois are equidistant from Tours. A second explanation is to be found by comparing the fabrics from Tours and Châtellerault, which show technical similarities but are more than a century apart! This chrono-typological difference arises from different preferences, potentials of the clay deposits used, or know-how of the potter. Are we looking at two socio-economic areas without strong economic ties, or indeed at two distinct cultural entities?

This seems to be confirmed by applying the pottery model to the "Aubaret" site in Poitiers, a city approximately 120 km south of Tours. A rubbish pit (Z094) was chosen here both for the quality of its pottery assemblage and because it contained coins which can be dated to the end of the 10th century (Bocquet, dir. 1997). The curves clearly show the inconsistency of the results, the orange curve giving a *dateEv* estimation of the beginning of the 11th century, while the main peak of the black curve is situated at the beginning of the 13th century (Fig. 10). The distance from the reference site and the fact that this city belongs to a different cultural entity are the main causes of these chronological disparities. Some fabrics in the two economic spaces have similar technical and aesthetic features, but come from different periods, indicating (as in Châtellerault) technical constraints or preferences, but without evidence of any real contact between these areas.

- These examples, used to illustrate our argument, were chosen from a corpus of approximately 150 modelled archaeological contexts in the centre-west of France. The trend is identical for all the contexts, with the pottery model becoming progressively less reliable the further the application sites are from the reference site. We can conclude that our model, based on a single archaeological source, allowed us to define the boundaries of the economic environment of Tours, namely a radius of approximately 30–40 km around the town (Appendix 1 : limits of the model presented by grey parallel lines). Beyond that, there are two possible situations: areas such as the Blois district which are economically distinct but show evidence of strong links with Touraine, and those such as upper Poitou, which are not necessarily further away, but which belong to distinct cultural entities, as shown by the pottery.

With regard to future objectives, changing the geographic scale but using the same procedure with other reference points in the

Z019: Pottery dating *dateEv* (orange curve): CI 95%		
Earliest possible date	Estimation	Latest possible date
1045	1065	1084
Z019: Pottery dating *dateAc* (black curve): CI 95%		
Earliest possible date	Estimation	Latest possible date
580	1147	1230

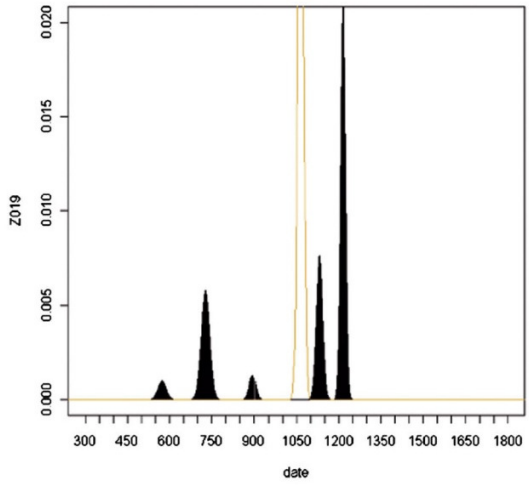


Fig. 9. Results of the pottery model and curves *(dateEv* and *dateAc)* for the Châtellerault context.

centre-west of France and/or other regions would help identify new pottery cultural areas. This would allow us to draw up a dynamic image of the socio-economic and cultural organisation of the region over long time spans, albeit imperfectly because it would be based on a single source.

3.3. The pottery model: an aid for the functional interpretation of archaeological contexts

The distribution of pottery over time is used here to achieve a better characterization of the functional nature of archaeological contexts. The procedure involves analyzing, comparing and classifying the shapes of the density curves by function in order to establish groups of curve shapes that would ultimately constitute a reference base constructed from contexts with a known function. Once this classification has been made, the objective would be to use this typo-functional reference of the curves as a tool for interpreting archaeological contexts whose functions are not completely understood.

The purpose of this article is to show the interest of this procedure through a few examples and to provide succinct ideas for looking more generally at the constitution of archaeological deposits and the biases imposed on the archaeologist when trying to interpret findings.

The "Marmaudière" site in the village of Neuvy-Le-Roi, about 30 km north of Tours, is a good example of our procedure (Tourneur, 2004). Like Truyes, this site comprises a series of agricultural units from the 7th to 9th centuries, where it is possible to trace the course of events from settlement, domestic occupation and then abandonment of the space.

To understand the reasoning, it should be noted that the units are spatially isolated from one another on the site and therefore have no stratigraphic connections. Sometimes we know only the details of the evolution of each unit (phases of settling, domestic occupation and abandonment), but at other times we do not have this functional information and have to find it. To better understand the general organization of agricultural units on the site, it is essential to know the chronology between units, and consequently the date and functional nature of each stage of their formation. Furthermore, the absence of stratigraphic relationships between, and sometimes even within, agricultural units, raises problems of interpretation linked to land use, and of course to the chronology of the site. We will now describe in detail how a site can be interpreted functionally using the curves produced by the model, a decision-making tool.

The first choice of contexts for modelling was an agricultural unit where all the stages of its evolution were known, allowing the curves obtained to be analyzed in detail. This unit then served as a reference for analyzing the next one which was more difficult to interpret (Figs. 11 and 12).

The three orange curves (Fig. 13) for agricultural unit 1, the reference unit, clearly show how the site evolved: settled during

Z094 : Pottery dating *dateEv* (orange curve): CI 95%		
Earliest possible date	Estimation	Latest possible date
1044	1062	1081
Z094 : Pottery dating *dateAc* (black curve): CI 95%		
Earliest possible date	Estimation	Latest possible date
680	1203	1230

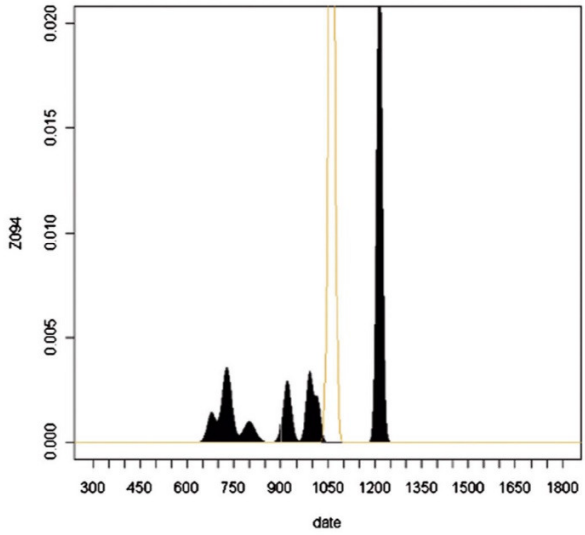


Fig. 10. Results of the pottery model and curves *(dateEv* and *dateAc)* for the Poitiers context.

Z069 : (Settlement) Pottery dating *dateEv* (orange curve): CI 95%		
Earliest possible date	Estimation	Latest possible date
642	669	695
Z069 : (Settlement) Pottery dating *dateAc* (black curve): CI 95%		
Earliest possible date	Estimation	Latest possible date
506	685	1119
Z072 : (Domestic occupation) Pottery dating *dateEv* (orange curve): CI 95%		
Earliest possible date	Estimation	Latest possible date
720	757	794
Z072 (Domestic occupation) Pottery dating *dateAc* (black curve): CI 95%		
Earliest possible date	Estimation	Latest possible date
603	754	1119
Z079 (Abandonment) Pottery dating *dateEv* (orange curve): CI 95%		
Earliest possible date	Estimation	Latest possible date
861	890	918
Z079 (Abandonment) Pottery dating *dateAc* (black curve): CI 95%		
Earliest possible date	Estimation	Latest possible date
514	857	1497

Fig. 11. Results of the pottery model for the reference agricultural unit 1.

Z067 (Settlement?) Pottery dating *dateEv* (orange curve): CI 95%		
Earliest possible date	Estimation	Latest possible date
609	640	671
Z067 (Settlement?) Pottery dating *dateAc* (black curve): CI 95%		
Earliest possible date	Estimation	Latest possible date
634	680	713
Z070 (Domestic occupation?) Pottery dating *dateEv* (orange curve): CI 95%		
Earliest possible date	Estimation	Latest possible date
737	785	834
Z070 (Domestic occupation?) Pottery dating *dateAc* (black curve): CI 95%		
Earliest possible date	Estimation	Latest possible date
682	796	879
Z074 (Abandonment?) Pottery dating *dateEv* (orange curve): CI 95%		
Earliest possible date	Estimation	Latest possible date
795	831	867
Z074 (Abandonment?) Pottery dating *dateAc* (black curve): CI 95%		
Earliest possible date	Estimation	Latest possible date
605	800	1139

Fig. 12. Results of the pottery model for agricultural unit 2 whose functional interpretation requires clarification.

the second half of the 7th century (Z069), occupied during the first half of the 8th century (Z072), and abandoned during the 9th century (Z079). If we compare these curves with the black curves for each stage, ignoring the small marginal fluctuations indicating residual or intrusive material, we can see that the main peaks match almost perfectly.

Apart from the chronological input, which is not of particular concern here, the profiles of the black curves also provide an indication of the nature of the archaeological contexts. Comparing the two agricultural units, the similarities between the orange and black curves corresponding to settlement and abandonment are not perfect, but are generally fairly close in appearance. By

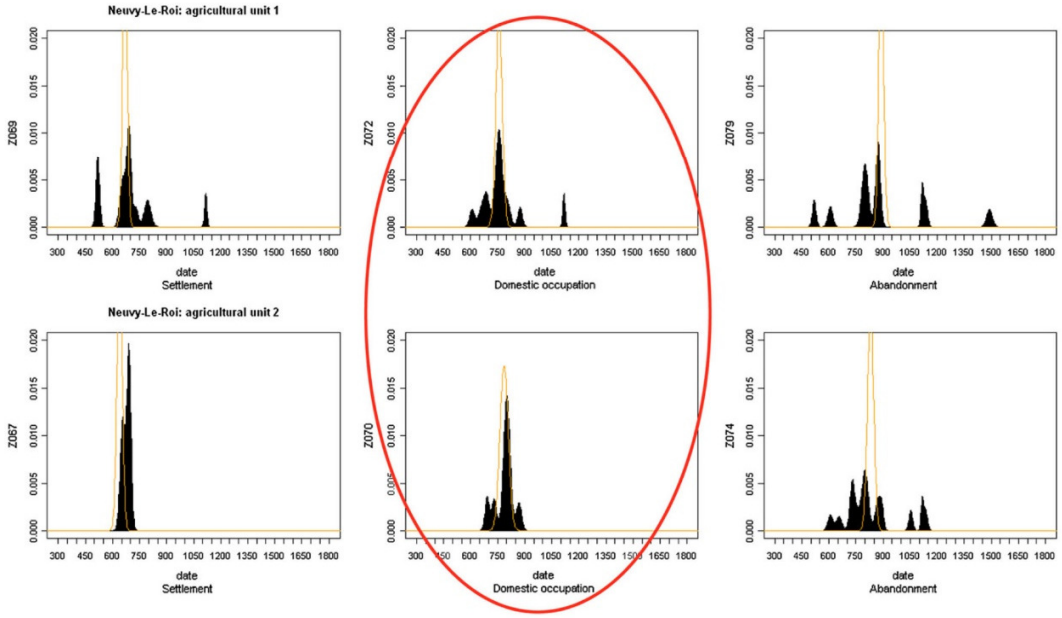


Fig. 13. Chrono-functional comparison of the density curves of agricultural units 1 and 2.

contrast, the curve profiles for the occupation period are identical, suggesting that the Z070 archaeological context is indeed the domestic occupation level of the second agricultural unit. This hypothesis is further reinforced by the same reading of the occupation curves of the agricultural units of the Truyes site presented above (*supra* Fig. 7).

However, there are several causes of variability in the formation of archaeological contexts. The first is linked to the actual constitution of archaeological deposits through a process of accumulation reflecting successive human or natural actions disturbed by time (Schieffer, 1987; Macphail et al., 2003). The second is due to the inherent bias in the way the archaeologist excavates successive anthropogenic levels where the interface is sometimes blurred. The final cause of variability comes from the chrono-typological construction of the artefacts, in this case pottery, which has its own specific pattern. This consistency in the appearance of the curves for occupation levels can therefore only be explained by the importance of the functional factor.

One way of observing this is to look at the curves from a different perspective, in this instance those obtained for the second agricultural unit (Fig. 14). The evolution of this part of the site, divided up on the basis of the succession of archaeological contexts identified and interpreted by the archaeologist, follows a logic that differs to some extent from that based on the on-going accumulation of archaeological levels over time. In fact, some installation deposits can be found during domestic occupation, and

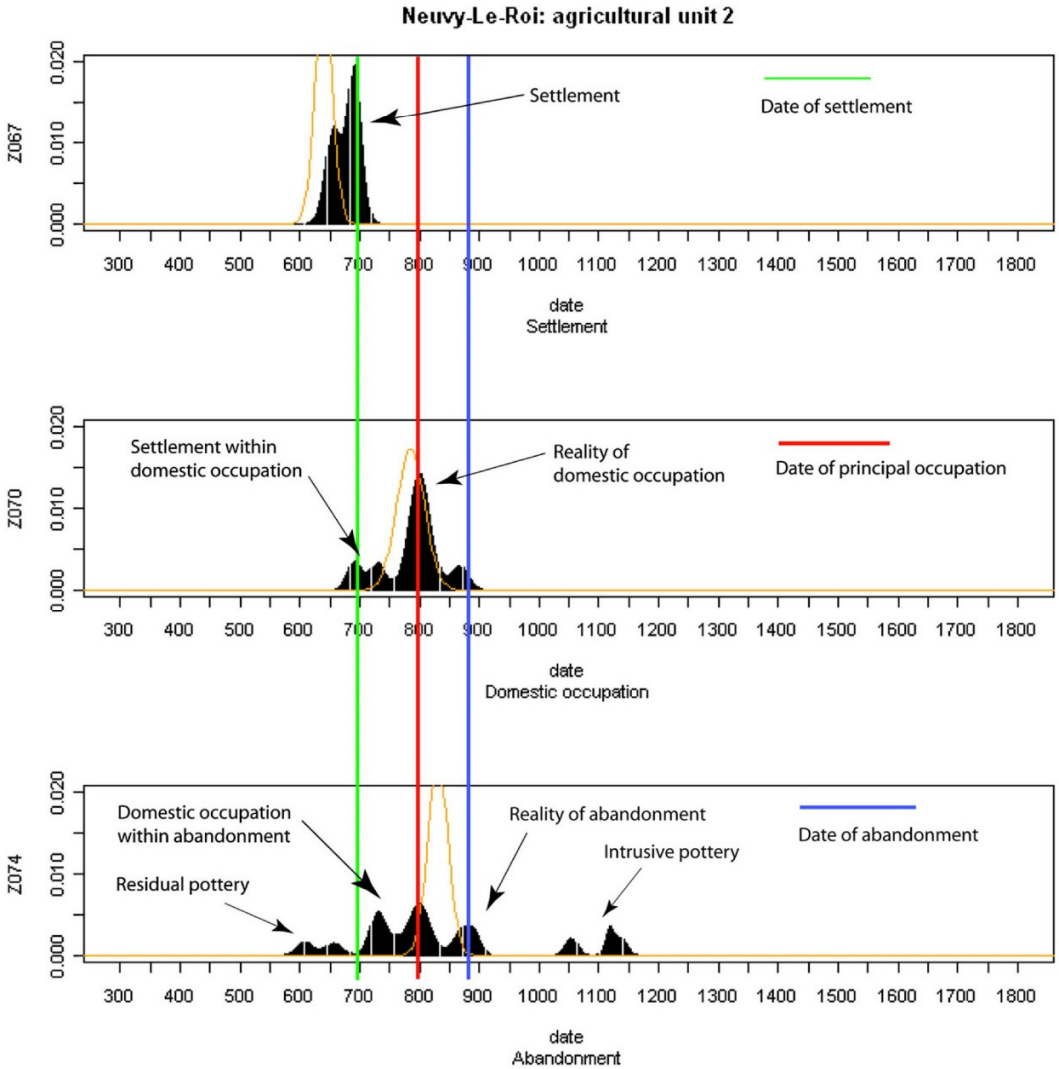


Fig. 14. Succession of archaeological contexts (settlement, occupation and abandonment) based on stratigraphic divisions proposed by the archaeologist.

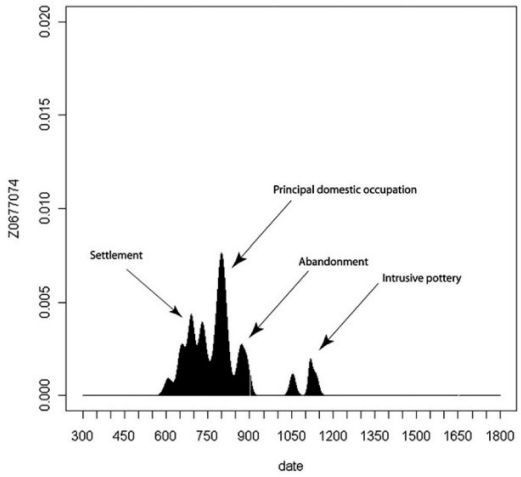


Fig. 15. Black curve produced by combining the three previous contexts within a single pottery corpus.

some deposits of this occupation can be observed during abandonment. This is perfectly logical in the process of the constitution of archaeological deposits over time; the continuity can be observed in material traces or stratigraphic gaps that constitute not only dated events but also chronological hiatuses (Galinié et al., 2004; Husi, 2006).

Extending this argument, the agricultural unit was modelled on the entire corpus without differentiating between the three archaeological contexts of which it is composed, (settlement, domestic occupation and abandonment). The black curve (Fig. 15) clearly shows the three previously perceived peaks, while the remainder of the preceding states are of course not shown. We have moved from a sequential image integrating stratigraphy (Fig. 14) to an image based on the estimation of pottery datings (Fig. 15). It is however impossible to interpret Fig. 15 without Fig. 14. By contrast, for a long stratigraphic urban sequence, it is possible to imagine moving from a series of curves constructed from interpreted chrono-functional contexts (Fig. 14) to a continuous re-reading of the site from a single fluctuating curve, representing actions and gaps visible between peaks (Fig. 15).

4. Conclusions and further prospects

The philosophy of this paper is to place archaeological issues at the heart of scientific procedures, and consequently to attempt to address these issues using statistical methods, recent or otherwise, adapted to the archaeological data to be processed. Moreover, in any modelling attempt, it is important to take into account the dialectical relationships between chronological, socio-economic and functional factors, which must be analyzed and reviewed whenever new archaeological data are introduced into the model. There is thus a dual objective here, on the one hand to present a dating method for contexts based on archaeological and statistical reasoning, and on the other to apply it and validate its results on contexts that are representative of the archaeological interest of the procedure.

From a chronological standpoint, the comparison of our results to those obtained with other dating methods, such as archaeomagnetism, illustrates the quality of the pottery model, and also its simplicity which in no way impairs the archaeological interpretation. It allows to scrutinize model errors and to take them into account in the readjustment of model parameters. As suggested by M. J. Baxter: "The view that developments in the analysis of radiocarbon data, associated with an explicitly Bayesian approach, is one of the most important contributions that mathematics/statistics has made to archaeology is based on my perception that it had altered *archaeological* interpretations in a fundamental way…" (Baxter, 2008: 980).

In addition to the purely chronological aspects, the approach that we have used to analyze the archaeological material of several geographically separated sites also addresses socio-economic and functional questions. For the little known period of the 8th to 10th centuries, the spatial interpretation of the pottery model shows that the pottery area of Touraine did not extend beyond a 40 km radius around Tours, its main consumption centre and the reference site for constructing the model. Our model thus allows socio-economic spaces to be defined using an archaeological source, namely pottery. In the future, it should incorporate other carefully chosen reference sites in order to draw up a map of socio-economic mechanisms on a much broader geographic scale, such as the middle Loire valley.

Many further developments of the model can then be envisaged. For example, it could be worth including more information from the physical relationships between archaeological contexts, in other words the relative chronology, using statistical methods such as Asymmetric Eigenvector Maps (AEM) (Blanchet et al., 2008). AEM is a spatial eigenfunction method used to model the spatial distribution of species, generated by an asymmetric, directional physical process which could be interesting to adapt to our archaeological situation. The results of the model analyzed from a functional viewpoint also open up interesting research possibilities. In the medium term, the prospects include the construction of a functional reference base of the curves using archaeological contexts that can be interpreted without any doubt. Establishing this reference base would then make it possible to determine the nature of contexts whose interpretation is more hypothetical. There are several possible statistical methods that could be used: curve classification by functional analysis (Ramsay, Silverman, 2002; Ferraty, Vieu, 2006) or by generalized Procrustes analysis (Gower, Dijksterhuis, 2004).

Appendix. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jas.2011.06.031.

References

Aubourg, V., Josset, D., Joyeux, P., 1993. Cour du château de Blois (Loir-et-Cher). Rapport d'opération S.R.A. Centre. 97 p., 71 fig. Archives du service régional de l'Archéologie du Centre.
Baxter, M.J., 1994. Exploratory Multivariate Analysis in Archaeology. Edinburgh University Press, Edinburgh.
Baxter, M.J., 2008. Mathematics, statistics and archaeometry: the past 50 years or so. Archaeometry 50, 968–982.
Bellanger, L., Husi, P., Tomassone, R., 2006a. Statistical aspects of pottery quantification for dating some archaeological contexts. Archaeometry 48, 169–183.
Bellanger, L., Husi, P., Tomassone, R., 2006b. Une approche statistique pour la datation de contextes archéologiques. Revue de Statistique Appliquée LIV (2), 65–81.
Bellanger, L., Tomassone, R., Husi, P., 2008. A statistical approach for dating archaeological contexts. Journal of Data Science (JDS) 6 (2) revue en ligne. http://www.sinica.edu.tw/~jds/Content-v-6-2.html.
Benzécri, J.-P., 1973. Analyse des données. Tome II: Analyse des correspondances. Dunod, Paris.
Blanchet, F.G., Legendre, P., Borcard, D., 2008. Modelling directional spatial processes in ecological data. Ecological Modelling 215, 325–336.
Bocquet, A. dir. 1997, Poitiers « Hôtel Aubaret », DFS de sauvetage programmé, Poitiers, 97 pp.

Cornec, T., Farago-Szekeres, B., Brisach, B., 2006. D'une résidence mérovingienne vers un cimetière carolingien, Châtellerault-Pouthumé (86), RFO, Poitiers, 3 vol.
Ferraty, F., Vieu, P., 2006. Nonparametric Functional Data Analysis. Edition Springer. http://www.lsp.ups-tlse.fr/staph/npfda.
Ferdière, A., 2007. Le temps des archéologues, le temps des céramologues, SFECAG, Actes du congrès de Langres, 15–24.
Galinié, H., Rodier, X., Saligny, L., 2004. Entités fonctionnelles et dynamique urbaine dans la longue durée. Histoire et Mesure XIX (3/4), 223–242.
Galinié, H., Husi, P., Rodier, X., Theureau, C., Zadora-Rio, E., 2005. ARSOL. La chaîne de gestion des données de fouilles du Laboratoire Archéologie et Territoires. Les petits cahiers d'Anatole, n° 17, 27/05/2005, 36.772 signes. http://www.univ-tours.fr/lat/pdf/F2_17.pdf.
Gaultier, 2011. Gaultier M. – Boulevard Périphérique nord-ouest de Tours: La Vermicellerie, un site du haut Moyen Âge (Fondettes, 37), Rapport de Fouille Archéologique, Conseil Général d'Indre-et-Loire. SRA Centre, Orléans.
Gower, J.C., Dijksterhuis, G.B., 2004. Procrutes Problems. Oxford University Press, New York.
Greenacre, M.J., 1984. Theory and Applications of Correspondence Analysis. Academic Press, New York.
Husi, P., 2006. Echelles de temps et chronologie du site jusqu'à la construction de l'église. Archéologie et Histoire de l'Art, n°22. In: Lorans, E. (Ed.), dir. – Saint-Mexme de Chinon, Ve – XXe siècle. Comité des travaux historiques et scientifiques, Paris.
Husi, P., Rodier, X., (due for publication in 2011), ArSol: an Archaeological Data Processing System, Actes du 36th Annual Conference on Computer Applications and Quantitative Methods in Archaeology (CAA-2008), Budapest (Hongrie), 2–6 Avril 2008.
Kulpa, Z., 1997. Diagrammatic representation of interval space in proving theorems about interval relations. Reliable Computing 3, 209–217.
Macphail, R.I., Galinié, H., Verhaeghe, F., 2003. A future for dark earth ? American Antiquity 77 (296), 349–358.
Moreau, J., Doudin, P.A., Cazes, P., 2000. L'Analyse des correspondances et les techniques connexes. Springer-Verlag, Berlin.
Olivier, L., 2001. Temps de l'histoire et temporalités des matériaux archéologiques: à propos de la nature chronologique des vestiges matériels. Antiquités Nationales 33, 189–201.
Orton, C.R., 1975. Quantitative pottery studies: some progress, problems and prospects. Science and Archaeology 16, 30–35.
Orton, C.R., 1980. Mathematics in Archaeology. Collins, London.
Orton, C.R., 1989. An introduction to quantification of assemblages of pottery. Journal of Roman Pottery Studies 2, 94–97.
Orton, C.R., 1993. How many pots make five?—An historical review of pottery quantification. Archaeometry 35, 169–184.
Orton, C.R., 2000. Sampling in Archaeology. Cambridge University Press, Cambridge.
Orton, C.R., Tyers, P.A., 1992. Counting broken objects: the statistics of ceramic assemblages. Proceedings of the British Academy. In: Pollard, A.M. (Ed.), New Developments in Archaeological Science; a Joint Symposium of the Royal Society and the British Academy, vol. 77. published for The British Academy by Oxford University Press, pp. 163–184.
Ramsay, J.O., Silverman, B.W., 2002. Applied Functional Data Analysis. Edition Springer.
Saporta, G., 2006. Probabilités, Analyse des Données et Statistique, 2ème édition révisée et augmentée. Editions Technip, Paris.
Seigne, J., 2007. Dendrochronologie et datations archéologiques pour la période antique. Compte-rendu de la table-ronde du 23-01-2006 à Tours. Les petits cahiers d'Anatole. http://citeres.univ-tours.fr/doc/lat/pecada/pecada_20.pdf n° 20, 32/01/07, 18724 signes.
Schieffer, M.B., 1987. Formation Processes of the Archaeological Record. University of New Mexico Press, Albuquerque, 428 pp.
Tourneur, J., 2004. A28-section Montabon-Tours, Un habitat rural du haut Moyen Age: le site de Neuvy-le-Roi « La Marmaudière » (Indre-et-Loire): 37.170.172 AH, Document Final de Synthèse. INRAP/SRA Centre.
Tourneur, J., 2005. A85, L'habitat mérovingien de Truyes: Les Grandes Maisons » (Indre-et-Loire): 37.263.020 AH, Document Final de Synthèse. INRAP/SRA Centre.
Van de Weghe, N., Docter, R., De Maeyer, P., Bechtold, B., Ryckbosch, K., 2007. The triangular model as an instrument for visualising and analysing residuality. Journal of Archaeological Science 34, 649–655.
VanPool, T.L., Leonard, R., 2011. Quantitative Analysis in Archaeology, Editions Wiley.
Wirtz, B., Olivier, L., 2003. Recherches sur le temps archéologique: l'apport de l'archéologie du présent. Antiquités Nationales 35, 255–266.

Further Reading

Galinié, H., 2000. Ville, espace urbain et archéologie, col. Sciences de la ville, n 16, Maison des Sciences de la Ville, de l'Urbanisme et des Paysages, CNRS-UMS 1835. Université de Tours.
Husi, P., Tomassone, R., Chareille, P., Chareille, P., 2000. Céramique et chronologie: de l'analyse factorielle au modèle linéaire, application aux sites d'habitats de Tours. Histoire & Mesure XV (1/2), 3–32.
Laxton, R.R., 1990. Methods of chronological ordering. In: Voorrips, A., Ottaway, B. (Eds.), New Tools From Mathematical Archaeology. Scientific Information Centre of the Polish Academy of Sciences, Warsaw, pp. 37–44.
Lebart, L., Morineau, A., Piron, M., 1995. Statistique exploratoire multidimensionnelle. Dunod, Paris.
Madsen, T., 1988. Multivariate statistics and archaeology. In: Madsen, T. (Ed.), Multivariate Archaeology. Aarhus University Press, Aarhus, pp. 7–27.
Tyers, P.A., Orton, C.R., 1991. Statistical analysis of ceramic assemblages. In: Lockyear, K., Rahtz, S. (Eds.), Computer Applications and Quantitative Methods in Archaeology. British Archaeological Reports, International Series 565, Oxford, pp. 117–120.

4. DISCRIMINATION OF PSYCHOTROPIC DRUGS OVER-CONSUMERS USING A THRESHOLD EXCEEDANCE BASED APPROACH.

Bellanger L., Vigneau C., Pivette J., Jolliet P. and Sébille V. (2013). *Statistical Analysis and Data-Mining*, **6**(2): 91-101

Discrimination of Psychotropic Drugs Over-Consumers Using a Threshold Exceedance Based Approach

Lise Bellanger[1]*, Caroline Vigneau[2,3], Jacques Pivette[4], Pascale Jolliet[2,3] and Véronique Sébille[3,5]

[1]*Laboratoire de Mathématiques Jean Leray, Université de Nantes, France*

[2]*Centre d'Evaluation et d'Information sur la Pharmacodépendance, Service de Pharmacologie Clinique, CHU, Nantes, France*

[3]*Equipe d'Accueil EA 4275 "Biostatistique, Recherche Clinique et Mesures Subjectives en Santé", Faculté de Médecine-Pharmacie, Université de Nantes, France*

[4]*Service Médical de l'assurance maladie, Nantes, France*

[5]*Cellule de Promotion de la Recherche Clinique—Plateforme de Biostatistique, CHU de Nantes, France*

Received 26 July 2011; revised 28 June 2012; accepted 16 September 2012
DOI:10.1002/sam.11165
Published online 5 November 2012 in Wiley Online Library (wileyonlinelibrary.com).

Abstract: Use of some medication, such as tranquilizers or hypnotics may carry important risks for patients including the emergence of abuse and/or dependence. The problem we tackle consists of identifying and discriminating the group most "at risk" of abuse and/or dependence for a given drug, in order to provide an estimation of its prevalence and to develop preventive measures targeted toward the corresponding drug prescription. A criterion, currently employed to characterize patients' consumption of a drug, is the ratio between their daily average consumption and the maximum recommended daily dose as specified in the drug monograph, called the F factor. In theory, any patient having an F factor greater than 1 should be classified as an over-consumer for the corresponding drug, but in practice this threshold might not be very relevant for all drugs. The proposed approach, combining different statistical methods (extreme value theory with the Peaks Over Threshold Model, logistic regression, ROC curve), is an innovative way to study consumption behaviors of psychotropic drugs. Two drugs are studied: an antidepressant, tianeptine and a hypnotic, zolpidem. From one drug to another, different thresholds for the F factor and patient's characteristics associated with the risk of extreme consumption are found, revealing different consumption behaviors. © 2012 Wiley Periodicals, Inc. Statistical Analysis and Data Mining 6: 91–101, 2013

Keywords: extreme value theory; cluster analysis; discrimination; logistic regression; ROC curve; misuse; evaluation

1. INTRODUCTION

Drug abuse consists in the use of a drug in an excessive amount or for purposes for which it was not medically intended, and this may lead to drug dependence or addiction. Drug approval relies on clinical trial data that often have limited follow-up periods and sample sizes. Moreover, patients displaying risk factors for abuse or addiction are usually excluded from such trials, which make the evaluation of the drug's over-consumption difficult in these settings. Post-approval surveillance data, aiming at collecting spontaneous notification from health professionals, could be an alternative. Unfortunately, this data mostly concerns the most severely addicted patients and hence, might also suffer from sample selection bias. By contrast, health insurance databases, containing public health information related to patient's consumption of drugs in 'real life' situations, could provide a valuable tool for identifying the dependence potential of a given drug and over-consumption behaviors. Using these databases, a patient's drug consumption can be characterized as the

Correspondence to: Lise Bellanger (lise.bellanger@univ-nantes.fr)

Additional Supporting Information may be found in the online version of this article.

ratio between his/her daily average consumption and the maximum recommended daily dose, as specified in the drug monograph [1], sometimes designated as the F factor. Even if the recommended daily dose can vary according to different criteria (age, symptoms, etc.), its maximum dose does not change and remains drug specific. One could assume that a patient having an F factor greater than 1 should be classified as an over-consumer of the corresponding drug, but in practice this threshold might not be appropriate. Indeed, overconsumption possibly reflects (i) prescribing doctor's decision to seek better efficiency, (ii) abuse and/or dependence behavior characterized by the patient's compulsion to demand the drug be prescribed. Indeed, the threshold should probably depend on the specific drug since it can be hypothesized that the dependence potential and the possible misuse of drugs might differ importantly from one drug to the other [2]. Thus, it is very likely that the threshold value is drug dependent and is not necessarily equal to one, whatever the drug considered.

The methods used in most pharmacoepidemiological studies are based on standard descriptive approaches for estimating the prevalence of drug prescriptions and of drug misuse for instance. Lorenz curves are also used to investigate drug use and prescription patterns [3,4]. Latent class analysis has also been proposed to better understand and identify patterns of psychotropic or illicit drug consumption [5,6]. Finally, cluster analysis has also been used to assess prescription drug abuse and potential misuse from reimbursement databases in ref. 7. However, there is not, to our knowledge, any research aimed at identifying a threshold for overconsumption. Moreover, it seems that there is no consensus on the threshold values that should be used to define possibly 'deviant behavior' and thus indicate possible abuse.

Our main objective is to propose a novel statistical approach in this context for identifying a threshold, possibly drug-dependent, based on the F factor in order to produce a division into two groups of patients (over and normal consumption groups). Secondary objectives are to identify the main demographic and clinical characteristics associated with the risk of over-consumption defined by this threshold. This article is organized as follows. In Section 2, we describe the proposed methodology, based on the combination of two different approaches: the Peaks Over Threshold (POT) Model coming from extreme value theory, used as a clustering method and logistic regression used as a discrimination method. In Section 3, we apply our approach to post-marketing surveillance data on two prescription drugs: an antidepressant drug (tianeptine) and a hypnotic drug (zolpidem) for which abuse and dependence potential are known for the former and very likely for the latter. In section 4, we discuss the results obtained on our data,

Statistical Analysis and Data Mining DOI:10.1002/sam

2. METHODOLOGY

2.1. POT model

The goal of extreme value analysis is to quantify the stochastic behavior of a process at its high (or low) levels. In particular, it provides estimates of the probability of events (called extreme events) that are higher (or lower) than what has been already observed. In the POT model, an extreme event occurs when a high (low) threshold is exceeded by the process. In our context, it consists in choosing a high threshold and in modeling the process of exceeding it, conditionally upon the chosen level. Many monographs have been published; for example, refs 8, 9, or 10 detailing also the role of extreme value theory for insurance and financial applications, or more recently refs 11 and 12. It has also been used in climate and hydrology [13], air pollution [14], insurance or finance [15], and biomedical data processing [16]. However, to our knowledge, the POT model has never been used in the specific domain of pharmacoepidemiology. We use it as a particular method of cluster analysis to produce a partition in two groups of patients (over and normal consumption groups), depending on a threshold for the F factor named u. This threshold u will be fixed with the POT model. This approach provides a valuable tool to characterize extreme consumptions behaviors of drugs.

We shall only review the necessary foundations of classical threshold models.

Let Y be a random variable, with realization y, having distribution function (df) F_Y. Define the right endpoint of F_Y as $y^* = \sup\{y \in R : F_Y(y) < 1\}$. Given a real number $u < y^*$, the event $Y = y$ is called *an exceedance* over the threshold u if $y > u$. Given $Y > u$, the random variable $Y - u$ is called the *excess* over the threshold u. F_u, the *excess distribution* of Y over u, has a df given by:

$$F_u(x) = P[Y - u \leq x | Y > u] = \frac{F_Y(x+u) - F_Y(u)}{1 - F_Y(u)}, 0 \leq x < y^* - u. \quad (1)$$

In our work, Y_i will be the ratio between the daily average consumption and the maximum recommended daily dose for patient i, designated as the F factor.

If the common distribution function F_Y was known, the excess distribution in Eq. (1) would also be known. But in practice, this is generally not the case and we have to use a model to approximate F_u. Extreme value theory provides a parametric model, derived using asymptotic arguments demonstrated in ref. 17, allowing to approximate this distribution for high values of the threshold u. This result says that for sufficiently large u, the distribution of

the excesses over high thresholds $\{Y - u\}$ given that $\{Y > u\}$ is approximately a Generalized Pareto Distribution (GPD) $G_{\xi,\sigma}$:

$$F_u(x) \approx G_{\xi,\sigma}(x), \tag{2}$$

where $G_{\xi,\sigma}$ is defined by:

$$G_{\xi,\sigma}(x) = \begin{cases} 1 - \left(1 + \xi \frac{x}{\sigma}\right)^{-1/\xi}, & \xi \neq 0, \sigma > 0, \\ 1 - \exp\left(-\frac{x}{\sigma}\right), & \xi = 0, \sigma > 0, \end{cases} \tag{3}$$

where, ξ is the shape parameter and σ the scale parameter, for $\xi \neq 0$, the range of x is: $x \geq 0$ if $\xi \geq 0$, and $0 \leq x \leq -\sigma/\xi$ if $\xi < 0$. Many techniques exist to estimate the GPD parameters [18], traditional methods being the Maximum Likelihood (ML) and the moments-based methods. We use the ML method for which standard properties of ML estimators apply if $\xi > -0.5$. In fact, the case where $\xi \leq -0.5$, rarely arises in practice.

To implement the POT method, we must choose a suitable threshold u, providing a balance between bias and variance. In practice, a threshold as low as possible, subject to the asymptotic model and providing a reasonable approximation, is generally adopted. Different exploratory methods are available: the first one we use is based on the mean of the GPD, and uses the well-known *Mean excess plot*, also termed as the *mean residual life plot*; the second one (*threshold Choice plot*) is an assessment of the stability of parameters estimates, called *threshold invariance property* (if the distribution function of $\{Y - u\}$ given that $\{Y > u\}$ is a GPD, then for any threshold $v \geq u$ the distribution function of $\{Y - v\}$ given that $\{Y > v\}$ is also a GPD) (see, for more details, refs 10, 12, 19).

For assessing the quality of the fitted generalized Pareto models, classical diagnostic plots are used such as probability plots or quantile plots. Once the threshold exceedance u is determined, the POT model allows producing a partition in two groups: over (F factor $> u$) and normal consumption (F factor $\leq u$) groups.

2.2. Logistic Regression

In order to depict and identify predictors of over-consumption, a logistic regression model (see, for example, ref. 20 for more mathematical and practical details) for threshold exceedance is constructed. We consider a binary outcome z_i which represents for the ith patient ($i = 1, \ldots, N$) exceedance or nonexceedance of the threshold u previously obtained using the POT method (1 = yes, 0 = no). Classically, the logistic regression model can be written as:

$$\pi_i = P[z_i = 1] = \text{logit}^{-1}\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}\right),$$

where $\text{logit}(\pi) = \log(\pi/1 - \pi)$, β_0, $\beta_1, \ldots, \beta_p$ are the regression parameters and x_{ij} ($i = 1, \ldots, N$; $j = 1, \ldots, p$) are the covariates.

Maximum likelihood approach for inference is used as well as the bootstrap to obtain nonparametric bootstrap confidence limits and the plotting of the distributions of the coefficient estimates (using histograms and kernel smoothing estimates, e.g., ref. 21). Besides providing more accurate point estimates for prediction error, bootstrap is also used for the construction of confidence intervals for odd-ratios derived from fitted logistic models [22,23]. An iterative backward selection procedure was used to select the variables that were significantly associated with over-consumption (variable candidates for the model were those associated with over-consumption in bivariate analyses with $p < 0.20$ criterion and subsequently retained in the multivariate model using $p < 0.05$ criterion).

In order to choose the model, p-values, Hosmer test, and area under the receiver operating characteristic (ROC) curve are also taken into account. ROC curves plot the true positives (sensitivity) against false positives (1 − specificity). In our case, sensitivity is the proportion of patients exceeding the fixed threshold (identified using the POT model) who are correctly identified by the logistic model and specificity is the proportion of patients not exceeding this threshold who are correctly identified by the logistic model. The ROC curve shows the relative tradeoffs between true positives and false positives and the diagonal line represents the strategy of random guessing.

The area under the ROC curve is a single index for measuring the performance of a model. The larger the AUC, the better is the overall performance of the logistic model, that is its ability to distinguish between exceedance or nonexceedance of the threshold.

2.3. Optimal Cut-Off Probability for the Purposes of Patients' Classification

The results of the fitted logistic regression model were also summarized using a classification table (also named confusion matrix in the case of the Presence Absence model). The table is the result of a cross-classification of the outcome variable (F factor over u or not) with a dichotomous variable whose value is derived from the estimated logistic probability. To obtain this probability value, we must define a cut point c and compare each estimated probability to c: if the estimated probability

exceeds c, the derived variable is 1 (over consumer); otherwise, it is equal to 0 (normal consumer).

The cut-off value is chosen as the threshold where sensitivity (Se: proportion of observed over consumers correctly predicted by the model) and specificity (Sp: proportion of observed normal consumers correctly predicted by the model) are equal, which is approximately the point where Se and Sp curves cross, and that maximizes both Se and Sp.

All statistical analysis were performed using R (http://www.r-project.org) with packages ismev; POT and evir for POT model and packages Design and PresenceAbsence for logistic regression, ROC curve, and confusion matrix.

3. DATA RESULTS

Data on the dispensing of reimbursed drugs to people affiliated to the General Health Insurance Scheme (GHIS), an obligatory public health insurance system in France, are provided by pharmacists for reimbursement purposes, thus constituting a quasi-exhaustive data base of drug prescriptions. The data motivating our research arise from pharmacoepidemiological phase IV (post-marketing surveillance) files generated by the GHIS in the Pays de la Loire region in north-western France.

Two drugs are studied:

- tianeptine, an antidepressant drug flagged for its abuse and dependence potential for which case reports are published in the literature [24,25];
- zolpidem, an hypnotic drug for which abuse and dependence potential is suspected [26].

The data consist of patients that had at least two deliveries of the drug studied between July 1, 2005 and December 31, 2005. For tianeptine, $N = 7263$ patients were evaluated and $N = 33,584$ for zolpidem. Information on each delivery of the selected drugs during the studied period is provided as well as data regarding several covariates. For each patient, we have variables related to:

- patient's consumption of the studied drug (tianeptine or zolpidem) provided by the F factor which is the ratio between the daily average consumption of the patient and the maximum recommended daily dose as specified in the drug monograph
- socio-demographics: age in years and gender
- drug's prescription:
  - doctor shopping behavior: visiting more than three doctors to obtain prescriptions during the follow-up period
  - need for treatment by a psychiatrist
- drug's delivery:
  - pharmacy shopping behavior: having more than three pharmacies that delivered the studied drug during the follow-up period
  - R ratio: number of dispensations/28 days
- associated treatments delivered during the period of treatment:
  - number of psychotropic drugs
  - number of benzodiazepines
    - more than one anxiolytic benzodiazepine (BZD) drug
    - more than one other anxiolytic drug
    - more than one hypnotic BZD drug
    - more than one other hypnotic drug (in addition to the studied drug)
    - rivotril drug
  - other antidepressant drugs, neuroleptic or morphine drugs
    - more than one other antidepressant (ATD) drugs (in addition to the studied drug)
    - more than one neuroleptic drug
    - at least one morphine drug

In Table 1, we report the general characteristics of the patients for tianeptine and zolpidem. As expected for this type of medication, a majority of women was observed in our sample and the median ages were above 60 years old for both drugs. The maximum number of psychotropic drugs delivered during the study period was large for both drugs (14 for tianeptine and 16 for zolpidem); this might suggest that some patients are more seriously ill than others among the consumers of these drugs. This hypothesis is reinforced by the fact that about 20% of the patients were also

Table 1. Characteristics of the patients for tianeptine and zolpidem.

Variable (name of variable)	Tianeptine $N = 7263$	Zolpidem $N = 33{,}584$
Age	66; 29 [15–102]	63; 24 [6–106]
Gender (female)	4962 (68.3%)	23240 (69.2%)
Doctor shopping behavior (yes)	32 (0.4%)	300 (0.9%)
Specialist follow-up (yes)	1529 (21.1%)	6098 (18.2%)
Pharmacy shopping behavior (yes)	78 (1.1%)	438 (1.3%)
Number of psychotropic drugs delivered during the study period	1; 1 [0–14]	1; 2 [0–16]
Number of benzodiazepine drugs	1; 1 [0–10]	1; 1 [0–10]
• Number of anxiolytic BZD >1 (yes)	4121 (56.7%)	14578 (43.4%)
• Number of other anxiolytic drug >1 (yes)	412 (5.7%)	930 (2.8%)
• Number of hypnotic BZD drug >1 (yes)	1074 (14.8%)	3565 (10.6%)
• Number of other hypnotic drug >1 (yes)	2183 (30.1%)	2846 (8.5%)
• Rivotril drug (yes)	354 (4.9%)	1596 (4.8%)
Number of other antidepressant drug >1 (yes)	1518 (20.9%)	13359 (39.8%)
Number of neuroleptic drug >1 (yes)	1043 (14.4%)	3388 (10.1%)
Number of morphine drug >1 (yes)	126 (1.7%)	839 (2.5%)
R ratio >1 (yes)	2041 (28.1%)	7120 (21.2%)
Consumption factor, F	0.71; 0.39 [0.06–10.95]	0.66; 0.52 [0.03–30.07]

Notes: BZD: benzodiazepine, ATD: antidepressant, R ratio: number of dispensations/28 days.
Data summaries are median; interquartile range (IQR) [minimum–maximum] for continuous variables or numbers of patients (percentages) for categorical variables.

seen by a psychiatrist. Similarly, the maximum F factor could reach high values for both drugs. These extreme levels of consumption, often corresponding to misuse of the drug, usually illustrate a situation of abuse and/or major pharmacodependence.

3.1. POT Model for Tianeptine and Zolpidem

We use the POT method to obtain the generalized Pareto distribution (GPD) as an approximation to the distribution of excess amounts over high thresholds. To use the POT model approach, we are first required to choose a threshold value u. The mean residual life plot for tianeptine data (Fig. 1) is reasonably linear for $u > 1$, suggesting a reasonable choice of $u = 1.1$. This choice leads to 524 exceedances (7.22% of the patients).

Figure 2 represents the threshold choice plot for tianeptine. It allows looking for the stability of parameter estimates for a range of thresholds. We select the lowest threshold u_0 for which the estimates remain near-stable. The threshold u (upper horizontal axis) and the number k of exceedances of u (lower horizontal axis) are plotted versus their corresponding estimates of ξ along with confidence intervals. The resulting plot is relatively stable with estimated values ranging between (0.3; 0.5), with an increase in statistical uncertainty for very high threshold, that was already observed in the mean residual life plot Fig. 1. Hence, the selected threshold of $u = 1.1$ appears reasonable.

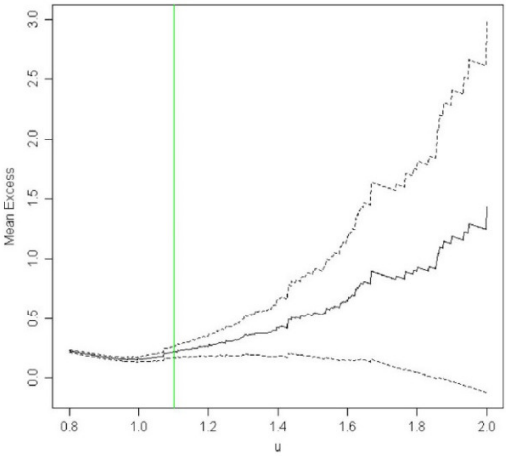


Fig. 1 Mean residual life plot for tianeptine data. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Same graphical methods are used for zolpidem, leading to $u = 2.0$, resulting in 318 exceedances (0.95% of the patients) (cf. Figs S1 and S2 in supporting information).

Having determined a threshold, the parameters of the generalized Pareto distribution can be estimated by maximum likelihood if $\xi > -0.5$. because the classical properties of the ML estimator hold. Table 2 shows that ML method can be applied in our context. It gives, for

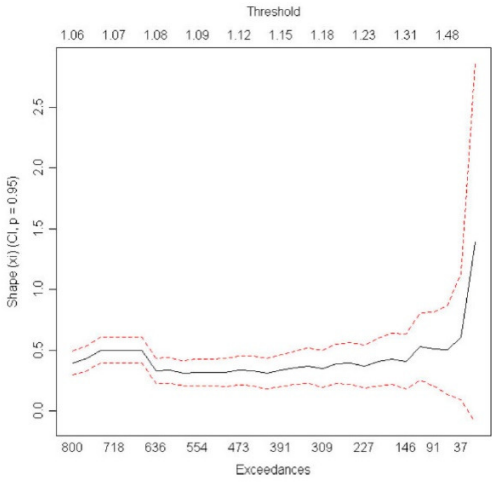


Fig. 2 Shape parameter estimate against threshold for tianeptine data. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Table 2. Maximum likelihood estimates of the GPD parameters for tianeptine and zolpidem.

	Tianeptine	Zolpidem
F Threshold	1.1	2.0
nexc	524	318
nllh	−339.135	176.665
Shape ξ (SE)	0.307 (0.052)	0.607 (0.092)
Scale σ (SE)	0.142 (0.009)	0.350 (0.036)

Notes: nexc = the number of data points above the threshold, nllh = the negative logarithm of the likelihood evaluated at the maximum likelihood estimates, SE = standard error.

both studied drugs and chosen u, maximum likelihood estimates of each GPD parameters and also standard errors obtained in the usual way from standard likelihood theory. As the distributions encountered in both cases have support unbounded to the right, we find, as expected, that the 95% confidence interval for ξ is in the positive domain (e.g. $0.307 \pm 0.052 = [0.205, 0.409]$ for tianeptine): data exhibit heavy tail behavior.

Diagnostics plots for the tianeptine's (zolpidem, respectively) fitted GPD are shown in Fig. 3 (Fig. S3, respectively). The fitted model seems to be appropriate and we can therefore model the excesses as GPD.

3.2. Multivariate Logistic Regression

Over-consumption is defined as exceedance of the threshold identified using the POT model for both drugs: tianeptine's threshold being equal to 1.1 and zolpidem's threshold

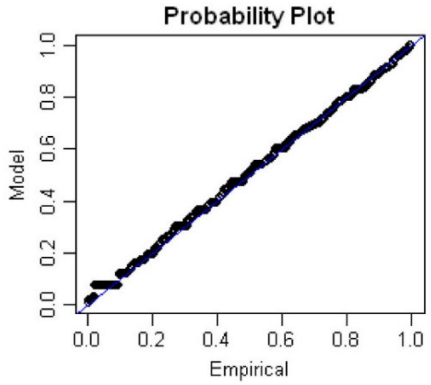


Fig. 3 $P-P$ plot for threshold excess model fitted to tianeptine data ($u = 1.1$). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

to 2.0. Tables S1 and S2 summarize the bivariate relationships between the independent and dependent variables. Backward elimination procedure was used as previously mentioned. Among the variables that are not subsequently retained, we may notice that no statistically significant differences in consumption behavior are observed between men and women, whatever the drug.

Results of multivariate logistic regression appear in Table 3 for both drugs.

To validate our model, we performed bootstrap replicates (see ref. 27 for an overview) of the analysis to determine how sensitive estimated parameters are to small changes in the data. We estimated all of the regression coefficients, computed nonparametric bootstrap confidence limits, and finally plotted the distributions of the coefficient estimates (using histograms and kernel smoothing estimates). The estimated densities appear approximately normal (Figs 4 and S4a–c). It confirms that the results obtained using maximum likelihood methods are robust and that the detected relationship is therefore likely to be true.

For both drugs, younger age, pharmacy shopping behavior, consumption of at least one anxiolytic drug including BZD or an hypnotic BZD, and displaying an R ratio >1 were identified as risk factors for over-consumption. For instance, pharmacy shopping behavior increases considerably the odds of being in the over-consumption group by 168.5% for tianeptine (resp. 518% for zolpidem).

Conversely, the need for treatment by a psychiatrist has an opposite effect on the risk of over-consumption, according to the considered drug. More precisely, for tianeptine, the need for treatment by a psychiatrist increases the odds of being in the over-consumption group by 63% while for zolpidem, it decreases the odds of being in the over-consumption group by 35.6%. Other variables are

Table 3. Multivariate logistic regression analysis of over-consumption risk for tianeptine and zolpidem; results of stepwise selection procedure. p-Value <5%.

	Tianeptine	Zolpidem
	1.1	2.0
F Threshold Logistic variable	Estimate (SE) OR [95% CI]	
Age , 10 years increase	−0.011 (0.003) 0.896 [0.849–0.946]	−0.010 (0.004) 0.907 [0.843–0.975]
Doctor shopping behavior		1.317 (0.215) 3.733 [2.451–5.688]
Specialist follow-up	0.489 (0.108) 1.630 [1.320–2.014]	−0.440 (0.142) 0.644 [0.488–0.850]
Pharmacy shopping behavior	0.988 (0.323) 2.685 [1.427–5.052]	1.821 (0.180) 6.180 [4.344–8.792]
Number of anxiolytic BZD > 1	0.250 (0.109) 1.284 [1.036–1.591]	0.435 (0.138) 1.545 [1.178–2.026]
Number of other anxiolytic drug > 1	0.436 (0.183) 1.546 [1.080–2.213]	0.797 (0.198) 2.218 [1.505–3.270]
Number of hypnotic BZD drug > 1		0.521 (0.139) 1.683 [1.281–2.212]
Number of other hypnotic drug > 1	0.286 (0.104) 1.332 [1.086–1.633]	
Number of other antidepressant drug > 1	0.228 (0.110) 1.255 [1.012–1.558]	
Number of neuroleptic drug > 1		0.400 (0.147) 1.492 [1.118–1.992]
Number of morphine drug > 1		0.574 (0.263) 1.776 [1.061–2.971]
R ratio > 1	3.289 (0.150) 26.807 [19.971–35.983]	4.439 (0.342) 84.656 [43.348–165.331]
ROC area	0.87	0.93

Notes: BZD = benzodiazepine, ATD = antidepressant, SE = standard error, OR = odds ratio, CI = confidence interval.

selected as having a specific effect on one of the drugs but not on the other: having at least one other hypnotic drug or an antidepressant is only associated with the risk of over-consuming tianeptine while doctor shopping behavior, having at least one hypnotic benzodiazepine, neuroleptic or morphine drug is merely associated with the risk of over-consuming zolpidem.

3.3. ROC Curves—Classification

We obtain an area under the ROC curve equal to 0.87 (with associated standard deviation 0.01) for the tianeptine model and 0.93 (with associated standard deviation 0.01) for the zolpidem model; which is considered in both cases as very good discriminations (Figs 5 and S5).

A cut-off probability value is then calculated for translating predicted probabilities from the logistic regression into 0/1 values (i.e., non-exceedance/exceedance of the fixed threshold u for F) for each patient. In our context of pharmacodependence, without any *a priori* guidelines and prevalence assumptions, the optimized cut-off giving equal weights to Se and Sp, is calculated and found to be 0.15 for tianeptine and 0.02 for zolpidem. Indeed, for tianeptine if a patient has a probability of exceedance estimated by the logistic model lower than 0.15, he/she will be classified as having an F factor <1.1 otherwise he/she will be considered at risk of over-consumption. The statistics of exceedances discovery are best seen in a simple two-by-two confusion table (Tables 4 and S4, respectively), where 7243 (respectively 33,584) patients are classified according to their status and the logistic result. In conventional terms, we have a classification's rule with 83% (90%, respectively) Se and 81% (84%, respectively) Sp, along with a percent correctly classified (PCC) of 81% (85%, respectively) for tianeptine (zolpidem, respectively) data.

4. DISCUSSION

We proposed a statistical procedure to determine, for a given drug, a threshold for the F factor allowing to define and discriminate two groups of patients (over and normal consumption groups). We then analyzed the demographic and clinical characteristics associated with the risk of over-consumption of drugs.

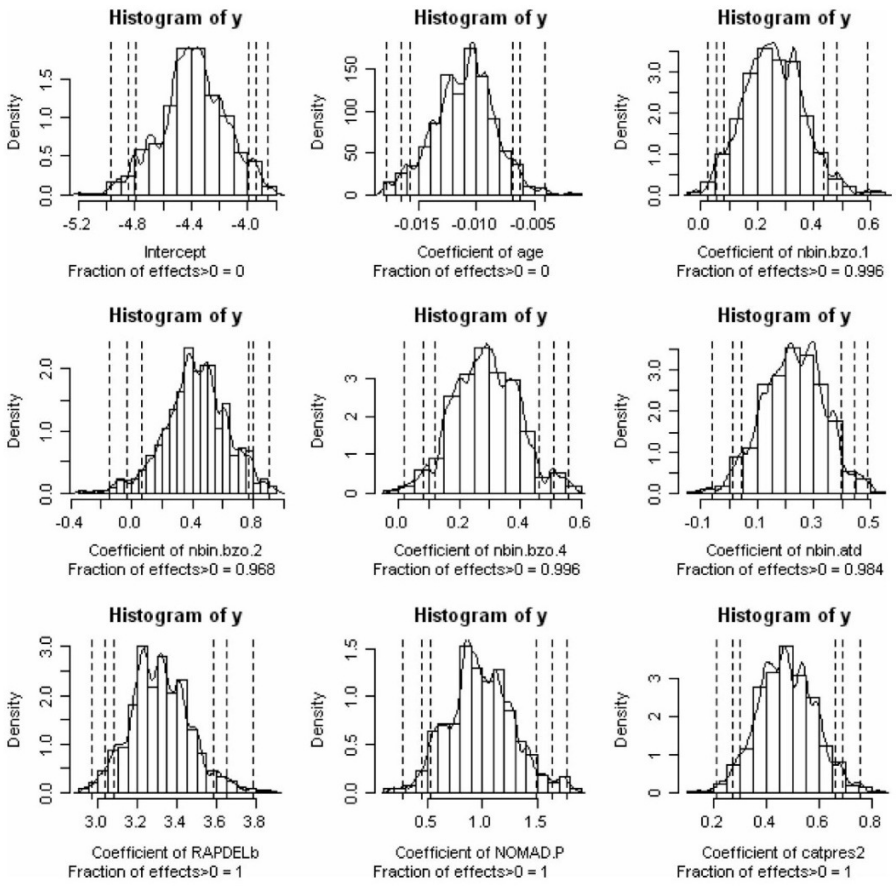


Fig. 4 Bootstrap distribution for logistic regression coefficients for tianeptine.

In the context of our pharmacoepidemiological data, constituting a quasi-exhaustive data base of drug prescriptions in the Pays de la Loire region of France, the POT model allowed identifying two different threshold values that seemed to be clinically relevant for the two studied drugs. For tianeptine, the threshold of 1.1 means that most of the times, when the treatment is appropriately used, no exceedance is expected [24,28]. The group of patients exceeding this threshold represents a small portion (7.22%) of the studied population (see Table S1). This group seems to display either i) addictive behavior associated with elevated doses of the drug or ii) a seriousness of their depressive pathology associated with a rise in their dosage determined by their practitioner for therapeutic purposes (eventually combined with other antidepressants). Indeed, for tianeptine, the factors associated with the risk of extreme consumption are: treatment by a psychiatrist, pharmacy shopping behavior, having other hypnotic drugs or antidepressants during the same period. For zolpidem, the threshold of 2.0 seemed to be also quite relevant [26]. Indeed, guidelines suggest that the maximum recommended daily dose should not exceed one pill (10 mg). The threshold of 2.0 indicates that patients belonging to the extreme consumption group take at least twice this maximum recommended dose. Besides, this value of 2.0 might be explained by the nature and the well-known pharmacological characteristics of the drug itself. Indeed, zolpidem is a hypnotic drug characterized by a short half-life that may explain the fact that patients may take an additional pill during the night. Moreover, a pharmacological tolerance has been demonstrated for this drug and may therefore explain why some patients tend to increase their dosage in order to experience or maintain the hypnotic effect of the drug. Furthermore, this also underlines

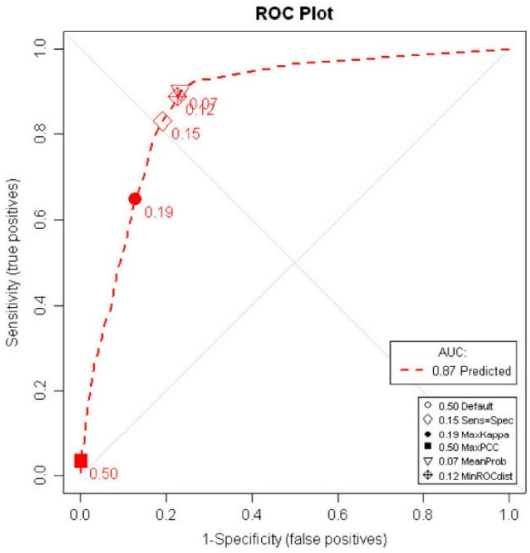


Fig. 5 ROC curve for tianeptine. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

the likely lack of efficacy of the drug at the recommended dosage and might explain the association of other hypnotics (such as hypnotic benzodiazepine). The small group of patients (0.95% of the sample) exceeding the threshold of 2.0 (see Table S2) could correspond to addictive patients often using very large amounts of the drug and displaying fraudulent behavior to get zolpidem. Indeed, the factors associated with the risk of extreme consumption are: (see Table 3): pharmacy shopping behavior, and doctor shopping behavior mostly concerning patients being treated by a general practitioner.

In order to describe and discriminate the two subpopulations (over- and normal consumption groups), and to predict the status of a new patient, we applied a logistic model. As this analysis consists in a single study, we used bootstrap to validate our model. Of course bootstrapping cannot replace validation in another sample of patients, but it increases the confidence that the detected relationships reflect true associations. The distinction between high-risk and low-risk patients regarding overconsumption is also important in a predictive perspective; ROC curves provide a measure of discrimination accuracy of our regression logistic models (see Figs 5 and S5); Both areas under the ROC curves were quite high (0.87 for tianeptine and 0.93 for zolpidem).

Our cluster problem is very specific. An approach by mixture models and the expectation-maximization (EM) algorithm or by other clustering techniques like Hierarchical

Table 4. Classification table based on the logistic regression model in Table 3 using a Cutpoint of 0.15 (sens = spec) for tianeptine.

	Observed		
Classified	1	0	Total
1	435	1287	1722
0	89	5432	5521
Total	524	6719	7243

Notes: Se = 435/524 = 83%; Sp = 5432/6719 = 80.8%; PCC = 81.1%.

or partitioning ones would not allow identifying and interpreting the threshold as does the POT model. Indeed, by construction, the threshold obtained by the POT model is related to the extreme consumption of the studied drug. This constitutes pharmacologists' main interest since they are trying to evaluate the drugs' potential for overconsumption. The strengths of our proposed statistical procedure, in the context of pharmacoepidemiological data, are to confirm the important variability, from one drug to the other [2] and among patients, of the dependence potential and the possible misuse of drugs. This methodology can be generalized to any drug to detect patients with extreme-consumption behavior in the huge mass of data gathered in health insurance databases.

However, there are some limitations and also some necessary further developments to our findings, associated both with the available data as with some methodological issues related to the statistical approach. Indeed, from a pharmacoepidemiological viewpoint, the interpretation of the results must also take into account the specificity of these data. Indeed, the data focuses on patients having at least two deliveries of the drugs during the study period (6 months) and only reflects pharmacy delivery but not necessarily patients' true consumption of the drugs [29]. It might therefore be hypothesized that these data can highlight the drugs most 'at risk' of abuse and/or dependence and that longer follow-up periods would be necessary to allow for the possibility of studying consumptions trends more precisely. In addition, the available data does not provide information regarding other important patients' characteristics such as: general consumption behavior (tobacco, alcohol...), underlying disease, individual characteristics (socio-professional status, etc.), consumption of other potential illicit drugs (heroin, cocaine, amphetamines, other psycho-active substances). Such additional information would be very valuable to provide a better understanding and insight into the population of over-consuming patients. For example, a drug can have multiple recommended doses for different symptoms, but this database does not provide information on the symptoms nor the pathology of the patient.

Nevertheless, as we use the maximal recommended dose, we should not over estimate over-consumption.

Another issue concerns the POT method itself. It is well known that in practical applications, choosing the threshold u is perhaps the greatest practical problem of this method. Yet, this selection is critical to estimation accuracy. Our findings need to be confirmed by other studies. Another method, more marginally used to choose a suitable threshold u, is the L-Moments plot: L-moments are summary statistics for probability distributions and data samples. They are analogous to ordinary moments (they provide measures of location, dispersion, skewness, kurtosis, and other aspects of the shape of probability distributions or data samples; they are computed from linear combinations of the ordered data values). Indeed, it would be interesting to try to use the theoretical relation between the L-Kurtosis and the L-Skewness existing for the GPD to detect an optimal threshold and then to use a resampling method for estimating the bias and standard error of such a threshold. In our case, it is not possible to use the bootstrap because we would like to study extreme structure in our data set. The jackknife is also not adequate because (i) it can fail if the studied statistic is not 'smooth' (like the median), and (ii) the jackknife data sets differ from the original data set by only one data point (in our case, $N = 7263$ or 33,584 patients). A possibility that can be used to fix up the inconsistency of the jackknife for non-smooth statistics is to leave out d observations instead of one at a time. Indeed, it could be interesting to try to use the jackknife by groups or delete-d jackknife [27] to estimate bias and standard error of the optimal threshold detected using the theoretical relation between the L-Kurtosis and the L-Skewness existing for the GPD.

Choosing logistic regression as a discrimination method could also be discussed. Other possible methods, such as discriminant analysis or CART (classification and regression tree), exist. It would be interesting to apply the three techniques to our data set and to compare their performances. Whatever the discrimination method, cut-off criteria for translating predicted probabilities into 0/1 values could also be discussed. Optimal cut-off depends on the relative costs of making false positive vs. negative errors. Different cut-offs result in different values of misclassification rate. The classical default criterion is to set the cut-off probability at 0.5, but it might not be relevant in many cases. The standard concept of sensitivity or equivalently the false negative rate (FNR $=$ $1 - \text{Se}$) is useful in public health. In this study, without any *a priori* guidelines, we retained the cut-off equaling high sensitivity and specificity, giving same cost to false positive and negative errors. The resulting FNR is $89/524 = 1 - 83\% = 17\%$ (10.1%, resp.), which is directly informative on the proportion of truly exceedences missed. Even if the FNR are quite satisfying, these classification results could perhaps be improved after discussion with pharmacologists since from a public health point of view, the cost of not detecting an overconsumption behavior might overcome the cost of falsely detecting a normal consumption behavior (hence promoting high sensitivity in the first place).

Finally, the POT model was used as a clustering method to split the patient population into two subgroups (normal and over-consumers) according to the F factor's value. Since the over-consumer group might include two different subtypes of patients, such as seriously ill patients and heavily addictive patients, three or more groups could be more appropriate to describe and characterise drug consumption. The use of Latent Class analysis, LCA (see, e.g., ref. 30 for a short overview and ref. 6 for a possible use in the context of pharmacoepidemiological data) offers such possibilities with the identification of homogeneous subgroups or 'classes' based on observed patients characteristics and consumption patterns. Another advantage of this approach is that observed characteristics are treated as fallible indicators of unseen states which seem to be well suited to study pharmacodependence since it cannot be directly observed and hence could be considered as a latent variable [31]. However, the use of LCA with dichotomous indicators requires that they have been well-defined in a preliminary step. For instance, in ref. 6, over-consumption is defined by an F factor above 1 which might not be appropriate. It could be of value to combine the POT model and LCA. Indeed, POT model could be used as a preliminary step for identifying the most appropriate threshold for the F factor. This value could then be subsequently used for LCA.

In summary, in this study, we identified with the POT model, a threshold value for overconsumption that should be used to define possibly 'deviant behavior' and thus indicate possible abuse. We demonstrated that this threshold value is drug dependent. We quantified the dependence potential of psychotropic drugs using observed patients' consumption of such medications. In the future, the identification of the most 'at risk' group of abuse and/or dependence for a given drug using the threshold identified with the POT model could help to develop target preventive measures toward the corresponding drug prescriptions. Such a topic is relevant at the same time to patients, clinicians, researchers and policy makers. The association of methods coming from extreme value theory (POT model) and methods often used in epidemiology (logistic regression) could provide a valuable approach in pharmacoepidemiology to study and predict extreme consumption behaviors of psychotropic drugs.

Statistical Analysis and Data Mining DOI:10.1002/sam

REFERENCES

[1] Dictionnaire Vidal, 77 ème ed., Paris, Ed. du Vidal, 2001.

[2] C. Victorri-Vigneau, G. Basset, and P. Jolliet, How a novel programme for increasing awareness of health professionals resulted in a 14% decrease in patients using excessive doses of psychotropic drugs in western France, Eur J Clin Pharmacol 62 (2006), 311–316.

[3] G. J. Bramness, K. Furu, and A. Engeland, Carisoprodol use and abuse in Norway. A pharmacoepidemiological study, Br J Clin Pharmacol 64 (2007), 210–218.

[4] P. Gierden, G. J. Bramness, and L. Slodal, The use and potential abuse of anticholinergic antiparkinson drugs in Norway: a pharmacoepidemiological study, Br J Clin Pharmacol 67 (2008), 228–233.

[5] P. T. Harrell, B. Mancha, H. Petras, R. C. Trenz, and W. W. Latimer, Latent classes of heroin and cocaine users predict unique HIV/HCV risk factors, Drug Alcohol Depend 122 (2012), 220–227.

[6] L. Wainstein, C. Victorri-Vigneau, V. Sébille, J. B. Hardouin, F. Feuillet, J. Pivette, A. Chaslerie, and P. Jolliet, Pharmacoepidemiological characterization of psychotropic drugs consumption using a latent class analysis, Int Clin Psychopharmacol 26 (2011), 54–62.

[7] E. Frauger, V. Pauly, F. Natali, V. Pradel, P. Reggio, H. Coudert, X. Thirion, and J. Micallef, Patterns of methylphenidate use and assessment of its abuse and diversion in two french administrative areas using a proxy of deviant behaviour determined from a reimbursement database. Main trends from 2005 to 2008, CNS Drugs 25 (2011), 415–424.

[8] M. R. Leadbetter, G. Lindgren, and H. Rootzen, Extremes and related properties of random sequences and processes., New York, Heidelberg, Berlin, Springer-Verlag, 1983.

[9] M. Falk, Laws of Small Numbers: Extremes and Rare Events (3rd ed), Basel, Birkhäuser Springer Basel, 2011.

[10] P. Embrechts, C. Klüppelberg, and T. Mikosch, Modelling Extremal Events for Insurance and Finance, Berlin, Springer, 1997.

[11] R. Reiss and M. Thomas, Statistical Analysis of Extreme Values: from Insurance, Finance, Hydrology, and Other Fields (2nd ed), Basel, Boston, Birkhäuser Verlag, 2001.

[12] S. Coles, An Introduction to Statistical Modeling of Extreme Values: With 77 Illustrations, London, Springer, 2001.

[13] F. Ashkar, N. El-Jabi, and S. Sarraf, Study of hydrological phenomena by extreme value theory, Nat Hazard 4 (1991), 373–388.

[14] L. Bellanger and R. Tomassone, Trend in high tropospheric ozone levels. Application to Paris monitoring sites, Stat J Theoret Appl Stat 38 (2004), 217–241.

[15] M. Gilli and E. Këllezi, An application of extreme value theory for measuring financial risk, Comput Econ 27 (2006), 207–228.

[16] S. J. Roberts, Extreme value statistics for novelty detection in biomedical data processing, IEEE Proc Sci Measure Technol 147 (2000), 363–367.

[17] J. Pickands, Statistical inference using extreme order statistics, Ann Statist 3 (1975), 119–131.

[18] J. R. M. Hosking, J. R. Wallis, and E. F. Wood, Estimation of the generalized extreme-value distribution by the method of probability-weighted moments, Technometrics 27 (1985), 251–261.

[19] A. C. Davison and R. L. Smith, Models for exceedances over high thresholds (with discussion), J Roy Stat Soc B52 (1990), 393–442.

[20] D. Hosmer, Applied Logistic Regression (2nd ed), New York, Wiley, 2000.

[21] T. J. Hastie, R. J. Tibshirani, and J. H. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, New York, NY, Springer, 2009.

[22] B. Efron and R. Tibshirani, Improvements on cross-validation: The .632+ bootstrap method, J Am Stat Assoc 92 (1997), 548–560.

[23] G. Gong, Cross-validation, the jackknife, and the bootstrap: excess error estimation in forward logistic regression, J Am Stat Assoc 81 (1986), 108–113.

[24] P. Vandel, W. Regina, W. Bonin, D. Sechter and P. Bizouard. Abuse of tianeptine. A case report, Encephale 25 (1999), 672–673.

[25] E. Guillem and J. P. Lépine, Does addiction to antidepressants exists? About a case of one addiction to tianeptine, Encephale 29 (2003), 456–459.

[26] C. Victorri-Vigneau, E. Dailly, G. Veyrac, and P. Jolliet. Evidence of zolpidem abuse and dependence: results of the French Centre for Evaluation and Information on Pharmacodependence (CEIP) network survey, Br J Clin Pharmacol 64 (2007), 198–209.

[27] B. Efron and R. J. Tibshirani, An Introduction to the Bootstrap, Chapman & Hall/CRC Monographs on Statistics & Applied Probability (1st ed), Chapman and Hall/CRC, New York, 1994.

[28] L. Leterme, Y. Singlan, V. Auclair, R. Le Boisselier, and V. Frimas. Usage détourné de tianeptine. A propos de cinq cas de surconsommation, Annales de médecine interne 154 (2003), 2858–2863.

[29] J. Micallef, V. Pradel, X. Thirion, P. Jolliet, and M. Lapeyre-Mestre. Use of the health insurance database by the centres for evaluation and information on pharmacodependance: examples, interests and future prospects, Therapie 59 (2004), 581–588.

[30] P. F. Lazarsfeld, Latent Structure Analysis, Mifflin, Houghton, 1968.

[31] L. M. Scheier, A. Ben Abdallah, J. A. Inciardi, J. Copeland, and L. B. Cottler. Tri-city study of Ecstasy use problems: a latent class analysis, Drug Alcohol Depend 98 (2008), 249–263.

5. MULTILEVEL MODELLING ON SURVEY DATA: IMPACT OF THE 2-LEVEL WEIGHTS USED IN THE PSEUDOLIKELIHOOD.

Lucas J.-P; Sébille V., Le Tertre A., Le Strat Y.; Bellanger L. (2014). *Journal of Applied Statistics* (Under Press)

Journal of Applied Statistics, 2013
http://dx.doi.org/10.1080/02664763.2013.847404

Multilevel modelling of survey data: impact of the two-level weights used in the pseudolikelihood

Jean-Paul Lucas[a,b*], Véronique Sébille[b], Alain Le Tertre[c], Yann Le Strat[c] and Lise Bellanger[d]

[a]*Scientific and Technical Building Centre (CSTB), Paris Est University, Marne-la-Vallée, France;* [b]*EA4275-Sphere, University of Nantes, Nantes, France;* [c]*French Institute for Public Health Surveillance (InVS), Saint-Maurice, France;* [d]*UMR CNRS 6629 Laboratory of Mathematics Jean Leray, University of Nantes, Nantes, France*

(*Received 25 March 2013; accepted 18 September 2013*)

Approaches that use the pseudolikelihood to perform multilevel modelling on survey data have been presented in the literature. To avoid biased estimates due to unequal selection probabilities, conditional weights can be introduced at each level. Less-biased estimators can also be obtained in a two-level linear model if the level-1 weights are scaled. In this paper, we studied several level-2 weights that can be introduced into the pseudolikelihood when the sampling design and the hierarchical structure of the multilevel model do not match. Two-level and three-level models were studied. The present work was motivated by a study that aims to estimate the contributions of lead sources to polluting the interior floor dust of the rooms within dwellings. We performed a simulation study using the real data collected from a French survey to achieve our objective. We conclude that it is preferable to use unweighted analyses or, at the most, to use conditional level-2 weights in a two-level or a three-level model. We state some warnings and make some recommendations.

Keywords: lead exposure data; level-2 weights; multilevel model; pseudolikelihood; public database; survey data

1. Introduction

Analysis of survey data traditionally aims to estimate (census) the parameters of interest, θ, of a finite population, U, of size N : $\theta = f(y_1, y_2, \ldots, y_N)$. θ can be a total, a percentile or a mean, for instance. Estimating such quantities belongs to *descriptive inference*. θ is estimated from a sample s of size n, i.e. from the values of a subset of the units of U : $\hat{\theta} = f(y_1, y_2, \ldots, y_n)$. s is selected at random by a sampling design, $p(\cdot)$, and all the *selection probabilities* of each unit i of U, π_i, to be selected in s are known.

*Corresponding author. Email: jean.paul.lucas@free.fr

Design-based inference concerns the sampling distribution of θ over all possible sample selections from the same $p(\cdot)$; the randomness is only located in $p(\cdot)$, and the population values $y_1, y_2, \ldots, y_N$ are considered fixed. Design-based inference is traditionally used for descriptive inference.

Another approach is *model-based* inference [14]. The population values $y_1, y_2, \ldots, y_N$ are considered as realisations of a *superpopulation* model: U is thought of as a sample drawn with replacement from a superpopulation. A unique sample s is then drawn from a sampling design. Such superpopulation models depend on unknown parameters of interest. Analysis of such parameters is called *analytical inference*. Model-based inference is often used for analytical inference.

In descriptive inference, the results obtained from s must be valid for the entire U. For this condition to hold, the distribution of the Y-values in s, $f_s(y_i)$, and the distribution of Y-values in U, $f_p(y_i)$, must be the same. If not, the sampling is *informative* for Y. The sample distribution is expressed as follows:

$$f_s(y_i) = \frac{\Pr(i \in s|y_i)}{\Pr(i \in s)} f_p(y_i) \quad [7]$$

For instance, if the outcome values, y_i, are related to the selection probabilities, $\Pr(i \in s)$, $\Pr(i \in s|y_i)$, and $\Pr(i \in s)$ are not equal and the sampling is thus informative for Y. Informative sampling typically leads to biased inference. In design-based analysis, bias under randomisation is treated with weighted estimators, for which the value of each unit i of the sample is weighted by the inverse of its selection probability, $w_i = 1/\pi_i$, as in the Horvitz–Thompson estimator $\hat{\theta} = \sum_{i \in s} w_i y_i$ of the Y-values population total, $\theta = \sum_{i=1}^{N} y_i$, for example. In practice, the weights w_i are often adjusted with methods such as *post-stratification* to provide new weights, $\tilde{w}_i$, thereby improving the estimates [14].

In analytical inference, all the units of U are also of interest. If we intend to model the outcome values, y_i, as a function of covariates values, $\boldsymbol{x}_i$, the probability density function $f_s(y_i|\boldsymbol{x}_i)$ in s is expressed as follows:

$$f_s(y_i|\boldsymbol{x}_i) = \frac{\Pr(i \in s|\boldsymbol{x}_i, y_i)}{\Pr(i \in s|\boldsymbol{x}_i)} f_p(y_i|\boldsymbol{x}_i) \quad [6]$$

Similarly, the sample model differs from the population model, except if $f_s(y_i|\boldsymbol{x}_i)$ equals $f_p(y_i|\boldsymbol{x}_i)$ for all y_i; this possibility cannot be ignored in the inference process. To mitigate this problem, we may introduce the design variables as covariates in the model. However, the analyst may not have all of the design covariates, in particular if he/she uses public data. The model incorporating the design variables may also not be of scientific interest if it does not reflect the desired original model. Another solution is to introduce weights into the estimator expressions; this solution leads to a combination of the randomisation distribution and the hypothetical distribution underlying the population model. For instance, in the case of a (single-level) regression model, regression coefficients may be obtained using maximum likelihood (ML) estimation. The ML estimators solve the census estimating equations $\sum_{i=1}^{N} (y_i - \boldsymbol{x}_i\boldsymbol{\beta})\boldsymbol{x}_i = 0$. The use of the available data from $s \subset U$ and sampling weights leads to a *pseudomaximum likelihood* (PML) estimation (PMLE): $\sum_{i \in s} w_i(y_i - \boldsymbol{x}_i\hat{\boldsymbol{\beta}}_{PMLE})\boldsymbol{x}_i = 0$. PML estimators are consistent in that they approach the finite population parameters when n and N both tend to ∞ (see section 2.2 of reference [11] and the references therein for more details).

Instead of single-level models, we may want to fit *multilevel* models. This situation typically occurs when the units of s are sampled with a multistage sampling. For instance, in a two-stage sampling, *primary sampling units* (PSUs) are selected at the first stage; then, units are selected at stage 2 from within the units selected at stage 1. For example, PSUs may be regions, and units at stage 2 may be schools. If we then sample pupils within the schools, a three-stage sampling is

thus defined. In the case of multilevel modelling, some specific weights must be introduced for each level.

In this paper, we aim to study the impact of the level-2 weights incorporated into the pseudolikelihood on the parameter estimates in a two-level model and a three-level model, when the sampling design and the hierarchical structure of the multilevel model do not match. To achieve this, we generated populations based on real data collected from a French survey about lead in housing performed in 2008–2009.

In Section 2, we introduce multilevel modelling, the issue of weighting and the pseudolikelihood estimation for a multilevel model. In Section 3, we describe the study based on lead contamination data motivating this work, its sampling design, the different level-2 weight candidates that could be introduced into the pseudolikelihood to fit the model and the different model results. In Section 4, we perform a simulation study based on our real data set to investigate the performance obtained with each candidate's weights. In Section 5, we present our results, and we discuss their implications in Section 6. Finally, in Section 7, we conclude by making some recommendations.

2. Multilevel modelling of survey data

2.1 *Multilevel models*

We restrict the presentation to two-level and three-level models with a random intercept. For a more general context, see reference [11].

In three-stage sampling, PSUs are selected with probabilities denoted π_k, where $k = 1, \ldots, n^{(3)}$. Secondary sampling units (SSUs) are chosen within each selected PSU with probabilities $\pi_{j|k}$, where $j = 1, \ldots, n_k^{(2)}$. Tertiary sampling units (TSUs) are then chosen within each selected SSU with probabilities $\pi_{i|j}$, where $i = 1, \ldots, n_j^{(1)}$. TSUs are called level-1 units, SSUs are level-2 units, and PSUs are level-3 units. We denote the outcome with Y. We denote by the outcome value of the ith level-1 unit within the jth level-2 unit within the kth level-3 unit as y_{ijk}. Similarly, level-1 information is stored in level-1 covariates denoted by $X_{ijk}^{(m)}$ for the mth covariate. The rth level-2 covariate is denoted $X_{jk}^{(r)}$, and the pth level-3 covariate is denoted $X_k^{(p)}$.

A three-level model with a random intercept is described by the following equations and assumptions:

Level 1:

$$y_{ijk} = \beta_{0jk} + \sum_m \varphi_m x_{ijk}^{(m)} + \epsilon_{ijk}, \tag{1}$$

Level 2:

$$\beta_{0jk} = \beta_{0k} + \sum_r \psi_r x_{jk}^{(r)} + \xi_{jk}, \tag{2}$$

Level 3:

$$\beta_{0k} = \beta_0 + \sum_p \theta_p x_k^{(p)} + \zeta_k. \tag{3}$$

with $\epsilon_{ijk} \sim N(0, \sigma_1^2)$, $\xi_{jk} \sim N(0, \sigma_2^2)$, and $\zeta_k \sim N(0, \sigma_3^2)$. β_0 is the overall average, and ζ_k are the random level-3 unit effects, which have mean 0 and variance σ_3^2 and represent the dispersion around the average quantity β_0. ζ_k are uncorrelated across the level-3 units. ξ_{jk} are the random level-2 unit effects, which have mean 0 and variance σ_2^2 and represent the dispersion within a level-3 unit around the random average quantity β_{0k}. The variance is assumed to be constant among the level-3 units; ξ_{jk} are uncorrelated across the level-2 and level-3 units and uncorrelated with the covariates. ϵ_{ijk} are the disturbances, which have mean 0 and variance σ_1^2 and represent the outcome dispersion within a level-2 unit, which is assumed to be constant among the level-2 units.

ϵ_{ijk} are uncorrelated across level-1, level-2, and level-3 units and uncorrelated with covariates. The random effects ζ_k, ξ_{jk}, and ϵ_{ijk} are classically supposed to be uncorrelated. φ_m, ψ_r, and θ_p are the coefficients associated with covariates that represent the fixed effects.

A two-level model with a random intercept is defined by the following two equations under analogous assumptions:

Level 1:

$$y_{ij} = \beta_{0j} + \sum_m \varphi_m x_{ij}^{(m)} + \epsilon_{ij}, \tag{4}$$

Level 2:

$$\beta_{0j} = \beta_0 + \sum_r \psi_r x_j^{(r)} + \zeta_j, \tag{5}$$

for $j = 1, \ldots, n^{(2)}$ and $i = 1, \ldots, n_j^{(1)}$.

2.2 *Weighting issue and pseudolikelihood*

Very often, observations are collected from multistage sampling in surveys with potentially unequal selection probabilities at some stages that induce informative sampling. Biases in parameter estimates may thus arise when these unequal probabilities are not considered in multilevel modelling [8]. To fit a multilevel model to such data, some estimation methods – in particular, methods based on pseudolikelihood [8,11,15] – that incorporate weights into the classical likelihood expression have been proposed. Such weighting methods only attempt to adjust the effects due to the sampling that are not accounted for by the covariates of the studied model [8]. This adjustment should make the sampling model identical to the population model. The weighting does not protect against misspecification of the population model, which is assumed to be correctly specified to explain the process that generated the data [6]. Whereas in a single-level linear regression, an overall sampling weight can be associated with each independent observation; in multilevel modelling, observations are not independent and such a weight does not carry sufficient information for bias correction [8].

Following the notation used in reference [10], the log-pseudolikelihood for a two-level model is expressed as

$$\sum_{j=1}^{n^{(2)}} w_j^{(2)} \log \int \exp \left\{ \sum_{i=1}^{n_j^{(1)}} w_{i|j} \log f(y_{ij}|\zeta_j) \right\} g(\zeta_j) \mathrm{d}\zeta_j, \tag{6}$$

where $n_j^{(1)}$ is the number of level-1 units within the level-2 unit j, $n^{(2)}$ is the number of level-2 units, and $\sum_{i=1}^{n_j^{(1)}} w_{i|j} \log f(y_{ij}|\zeta_j)$ is the log-likelihood contribution of level-1 units, which are conditional on the random effect ζ_j at level-2. $g(\zeta_j)$ is the normal density of the random effect ζ_j. $w_j^{(2)}$ is the level-2 weight of the level-2 unit j. $w_{i|j}$ is the level-1 conditional weight of the one-level unit i within the level-2 unit j. Note that $f(y_{ij}|\zeta_j)$ and $g(\zeta_j)$ are simplified notations for $f(y_{ij}|\zeta_j, \boldsymbol{\beta}, \sigma_1^2)$ and $g(\zeta_j|\sigma_2^2)$, respectively, where $\boldsymbol{\beta}$ is the vector of the fixed effects. σ_1^2 and σ_2^2 are the variance parameters. For a general expression for the pseudolikelihood, see section 4.2 in reference [11]. The standard errors are derived from a Taylor linearisation variance estimator [11].

In the case of a two-level model, for instance, the weights of each level unit, $w_j^{(2)}$ and $w_{i|j}$ in Equation (6), must be the reciprocals of π_j and $\pi_{i|j}$, respectively, i.e. the *conditional probabilities* of selection (we also use the term 'conditional' for π_j for simplicity). We call these weights *conditional weights* in this paper. However, such weights when associated with level-1 units may lead to bias in variance components estimators if they are large. Thus, a crucial question is

how these level-1 weights should be corrected to reduce bias. Several correction methods, called scaling methods, have been proposed (see reference [11] for a summary of the scaling methods), and their efficiency was assessed in some simulation-based studies (in most cases, for two-level models).

In practice, sampling with at least two stages is common in health surveys, and we may want to fit a two-level model using data collected from a sampling design with more than two stages. This situation may arise when little information is available about the highest level of the sampling design. For instance, in the PISA (Programme for International Student Assessment) files, the identifiers of the PSUs are not provided. The authors of reference [11] obtained these identifiers by requesting them from the data provider, but they did not consider the PSUs as level-3 units because the variance among the PSUs was not regarded as interesting by the authors. Instead, they fitted a two-level model. In any case, a three-level model could not be fitted with the requested conditional weights because the level-3 unit selection probabilities were not provided.

3. Context and problematic issues

3.1 *Motivation for the present study*

The present study was motivated by a study that aimed to estimate the contribution of lead sources to polluting the interior floor dust in French housing. Data were collected from a survey called 'Plomb-Habitat' (PH) performed from 2008–2009 [5]. The population of interest, U, was the set of primary residences (as opposed to second homes) in which at least one child aged six months to six years lived; the sample corresponds to 3,581,991 residences in mainland France. The sampling design is illustrated in Figure 1. This survey was nested in a two-stage survey called 'Saturn-Inf' (SI) that was conducted to estimate the prevalence of lead poisoning among children in France [2].

At the first stage of the SI survey, hospitals (PSUs) indexed by the letter k were selected. Then, at the second stage, children, indexed by the letter j, were surveyed. Housing units in the PH survey cannot be considered as units that belong to a third stage because the PH survey was *stricto sensu* a second phase (indicated by the superscript^b in Figure 1; the first phase is indicated by the superscript^a) from the second stage of the SI survey. Several rooms were investigated within the homes.

The outcome, Y, was the lead loading ($\mu g\, m^{-2}$) of the interior floor dust of each investigated room. Between two and five rooms within a housing unit were investigated. To account for the correlation among the lead loadings within a housing unit, a two-level model was considered, as described by Equations (4) and (5), where the level-1 units were rooms within housing units, which were the level-2 units. Such modelling enables estimates of the intra-class correlation coefficient between two lead loadings, which are also of interest. Only the covariates relative to the rooms and housing units were used in the model (see the covariates list in Appendix 1).

The PSUs were oversampled in the administrative regions with a lead hazard and in the at-risk areas. Moreover, during the second phase, children with high blood lead levels were oversampled. These oversamplings indicate that the SI/PH sampling was informative for Y.

Because the continuous covariates were very right-skewed, they were log-transformed before they were included in the multilevel model. Such a transformation for covariates was judged useful for environmental data about lead exposure [4,13]. The outcome was also log-transformed to approach the normality assumption of the disturbances.

In our survey, the child's bedroom, the living room, the main entrance, the kitchen, the playroom, and the bedroom of another child were exhaustively investigated. These rooms are called the 'PH rooms'. The level-1 conditional inclusion probabilities, $\pi_{i|j}$, were equal to 1 (as well as the corresponding conditional weights, $w_{i|j}$, obviously). Such weights do not induce bias for parameter

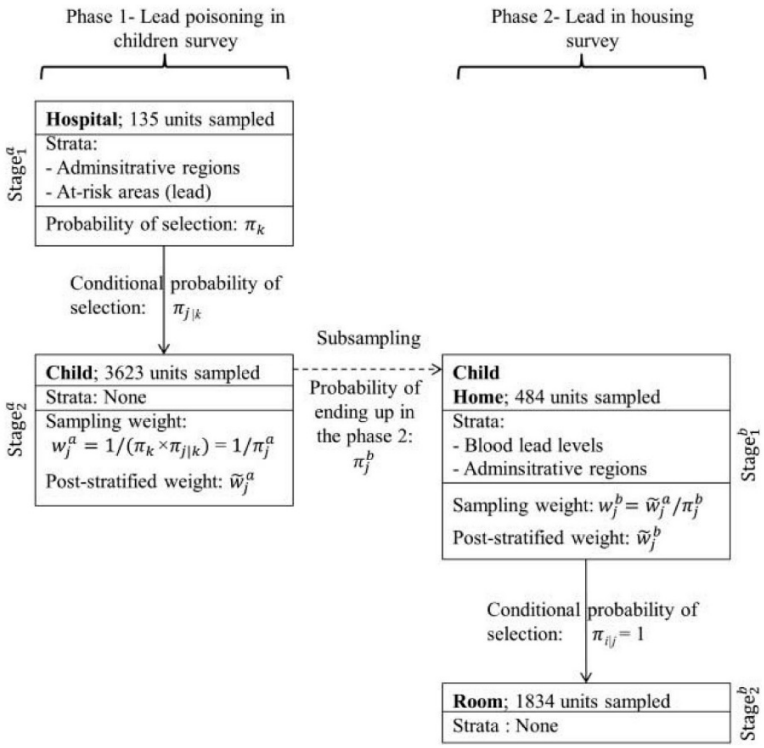


Figure 1. Sampling design of the lead poisoning in the children survey 'SI' and sampling design of the nested lead in the housing survey 'PH'. France 2008–2009.

estimates because they are not based on unequal selection probabilities. Thus, we were able to study the overall impact on estimations of the weights at higher levels than level-1.

3.2 *The different level-2 candidate weights*

In Equation (6), $w_j^{(2)}$ is the inverse of the inclusion probability of unit j at level-2, when level-2 is the highest level (i.e. the first stage of a two-stage design). Otherwise, the choice of level-2 weight introduced into Equation (6) can be questioned.

For a two-level model described by Equations (4) and (5), we identified the six following candidate level-2 weights, $w_j^{(2)}$, to be incorporated into the pseudolikelihood described in Equation (6). Following the notations in Figure 1, we obtained the following:

- $\boldsymbol{w}_1 : 1/\pi_j^b$
- $\boldsymbol{w}_2 : w_j^b$
- $\boldsymbol{w}_3 : \tilde{w}_j^b$
- $\boldsymbol{w}_4 : 1/(\pi_j^a \times \pi_j^b)$
- $\boldsymbol{w}_5 : 1$
- $\boldsymbol{w}_6 : 1/(\pi_{j|k} \times \pi_j^b)$.

$\boldsymbol{w}_1$-weights were considered conditional weights because π_j^b was not an overall inclusion probability for a housing unit (although π_j^b was not really a conditional probability because it was

located between two phases and not between two stages). $\boldsymbol{w}_2$-weights were the 'pure' overall sampling weights for a housing unit, i.e. not post-stratified based on housing criteria. $\boldsymbol{w}_3$-weights were the post-stratified overall sampling weight for a housing unit, i.e. also post-stratified based on housing criteria. $\boldsymbol{w}_4$-weights were design weights, i.e. not post-stratified at the child-level and not post-stratified at the home-level. $\boldsymbol{w}_5$-weights involved unweighted analysis. $\boldsymbol{w}_6$-weights can be considered as intermediate weights between $\boldsymbol{w}_1$ and $\boldsymbol{w}_4$, and their presence is justified by the subsampling.

We also tested a three-level model described by Equations (1–3), despite the presence of subsampling, instead of a stage between the SI and PH surveys. For the level-2 weights, we identified the three following candidates:

- $\boldsymbol{w}_7 : 1/\pi_j^b$
- $\boldsymbol{w}_8 : 1$
- $\boldsymbol{w}_9 : 1/(\pi_{j|k} \times \pi_j^b)$.

In a three-level model, the only possible weighting for the PSUs (here, hospitals) was $1/\pi_k$. Thus, in the two 3-level models of cases $\boldsymbol{w}_7$ and $\boldsymbol{w}_9$, the three-level weights were $1/\pi_k$. For case $\boldsymbol{w}_8$, the three-level weights were equal to 1 to produce an unweighted three-level model. The other

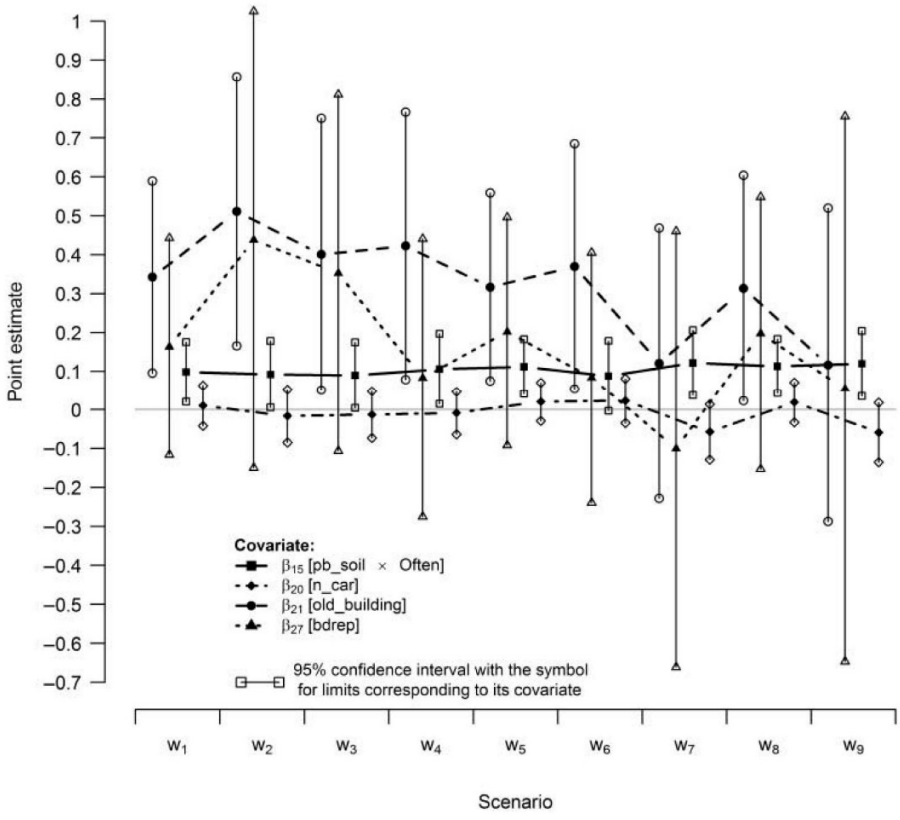


Figure 2. Behaviour of estimates obtained with a random-intercept two-level model ($\boldsymbol{w}_1$–$\boldsymbol{w}_6$) and a random-intercept three-level model ($\boldsymbol{w}_7$–$\boldsymbol{w}_9$) depending on the different scenarios for the two-level weights for four of the covariates used in the model. Model fitted on complete cases (1605 observations).

two-level weights that were candidates for the two-level model ($\boldsymbol{w}_2$–$\boldsymbol{w}_4$) were irrelevant because they were not based on the conditional probabilities of selection.

3.3 *Application*

The nine scenarios $\boldsymbol{w}_1$–$\boldsymbol{w}_9$ were tested and compared on complete cases (1605 rooms/429 housing units). For some covariates, their parameter estimate varied substantially depending on the scenarios. Indeed, Figure 2 displays the point estimates of four covariates and illustrates what we observed among the 33 parameters estimated (30 covariates + 1 intercept + 2 variance parameters). The point estimates of the coefficients associated with some covariates were relatively stable, such as for the lead concentration of the soil of the outdoor play area of the child when he/she often uses it (labelled [pb_soil × often]) and the car annual flow ([n_car]). For the two other covariates, we observed some substantial differences between the estimates, sometimes with a change in the sign. Thus, the use of some different level-2 weights seemed to have an impact on estimates. Further investigation, described in the Section 4, was performed to identify the best candidate weights.

4. Population generation and sampling

4.1 *General strategy*

First, to choose the appropriate level-2 weights to use in the study of the lead source contribution to pollution in the interior floor dust, we generated populations of dwellings/rooms. The covariates selected in our model were generated from distributions estimated from the data of the PH survey. The response variable was thus generated with a multilevel model equation. Second, we sampled housing units using a sampling design similar to that described in Figure 1. Third, we applied the nine scenarios to the samples drawn. Fourth, we compared the coefficient estimates with their true values, which we previously set. Because multilevel modelling belongs to model-based inference, generation replication was applied to the population U and not to the sample, i.e. we focused on estimates of the population model. Populations were generated, and the same sample was drawn from each population (i.e. a unique list of home identifiers).

4.2 *Generation of populations*

In practice, we generated several files where the rows represented housing units. The columns represented both the covariates and the response variable used in the multilevel model. A file that represented all of the housing units in France was extracted from the 2006 French population census [3] carried out by the French Institute INSEE (Institut National de la Statistique et des Études Économiques). From this file, only the housing units of our population of interest U were kept. For each housing unit, we sampled a number L(=2–5) of PH rooms (see Section 3.1 for the definition). Then, we sampled the type of the L rooms from among the child's bedroom, the living room, the main entrance, the kitchen, the playroom, and the bedroom of another child. The child's bedroom was always selected. The number L of PH rooms was chosen depending on the design-based distribution in the cross-table 'No. of PH rooms investigated × No. of rooms for living'. We refer to the bedrooms, the living rooms, the kitchens of area $>12\,\mathrm{m}^2$ and the rooms used as offices as *rooms for living*. The type of room was selected depending on the design-based distribution in the cross-table 'Type of PH rooms × No. of investigated rooms'. The sampled rooms constituted the final file, equivalent to our PH-table, which contained 1834 rows.

Some factor variables (which we called *auxiliary information*) collected from the PH survey matched those presented in the INSEE file. Let X be a covariate used in our multilevel model.

We generated X in the population by drawing a quantile value from its log-normal distribution estimated from the PH survey data (weighted estimation by the method of moments). Information from our real data was used in this manner. Assuming that X was linked to auxiliary information Z, its quantile values were selected by the levels of Z. For instance, we generated the values of the annual traffic flow of vehicles on the road closest to the home using the administrative region (which is a stratification variable in the sampling design of the SI and PH surveys) as primary auxiliary information and also using the urbanisation (rural/urban) as secondary auxiliary information. Another example of auxiliary information was the number of rooms for living used in the paragraph above. For categorical covariates, instead of using quantiles, we proceeded with modalities. The details of simulation for each covariate are given in the supplementary information.[1]

When all of the covariates were generated, the outcome was generated from Equations (4) to (5). We set the true parameter values at a similar order of magnitude as the point estimates obtained in Section 3.3 for the different scenarios of our PH survey sample in the complete cases (results not shown). Similarly, we set $\epsilon_{ij} \sim \mathcal{N}(\mu = 0; \sigma_1^2 = 0.80)$ and $\zeta_j \sim \mathcal{N}(\mu = 0; \sigma_2^2 = 0.45)$. Table A1 in Appendix 3 contains the true value of the regression coefficients. We generated 500 data sets.

4.3 *Sampling design*

We performed a sampling design that was as similar to the SI/PH sampling design as possible. The number of sampled units by stage and by stratum was the same as that of SI/PH survey, and the post-stratification was reproduced. However, some information about French hospitals and children used in SI/PH survey was not included in the generated population files because this information was very complex to gather and to generate, not to mention the fact that it was restricted. We replaced hospitals at the first stage with groups of cities called 'EPCIs' (Établissement Public de Coopération Intercommunale) defined by INSEE. The stratum 'region at risk (lead)' was not reproduced because it was defined according to hospitals; thus, it could not be reconstructed for EPCIs. We replaced children at the second stage of phase 1 with housing units because we did not have 'child' as a statistical unit in our generated population.

We use the same notations as Figure 1. The sampling frame contained 2594 EPCIs (PSUs). At the first stage of phase 1, between 2 EPCIs and 19 EPCIs were sampled by stratum. At the second stage, 3623 housing units (SSUs) were sampled from the 135 EPCIs drawn during the first stage. A minimum of 3 SSUs and a maximum of 45 SSUs by PSU were drawn. Post-stratification was applied on the overall sampling weights of the SSUs, w_j^a, to ensure that (i) the sum of the weights of the units in a given EPCI was equal to the real number of housing units within this EPCI and (ii) the sum of the weights of the units in a given administrative region was equal to the real number of housing units within this region. This post-stratification mimicked the post-stratification performed on children's weights in the SI survey to obtain the weights $\tilde{w}_j^a$ of Figure 1.

A subsample of 1032 housing units, stratified by region, was drawn to represent the 1032 addresses of volunteers (children's parents) who agreed to participate in the PH survey. Then, 484 housing units were randomly selected, with a minimum of 10 units and a maximum of 173 units in a stratum. All the PH rooms of a selected housing unit were included. Post-stratification was applied to the overall sampling weights, w_j^b, to ensure that the sum of the weights of the units in a given administrative region was equal to the real number of housing units within this region. We also applied post-stratification to ensure that the sum of the weights was equal to the real number of units of the post-strata. The post-strata were defined by crossing the 22 French administrative regions, the year of construction of the housing unit ($<1949/\geq 1949$) and the type of dwelling (individual/apartment building). This post-stratification mimicked the post-stratification performed on the housing units' design weights in the PH survey to obtain the weights $\tilde{w}_j^b$ of Figure 1.

The housing units selected had 1873 observations (rooms), with 2–5 rooms per housing unit. Then, for each of the 500 samples, we computed the weights of the 9 scenarios described in Section 3.2.

The oversampling of PSUs in the SI/PH sampling in the at-risk areas could not be reproduced, nor could the oversampling during the second phase based on the blood lead levels. However, the oversampling of PSUs in administrative regions with lead hazard was reproduced, thereby ensuring that the sampling design of the simulation study remained informative for the interior floor dust lead loading (Y).

4.4 *Statistical software*

The multilevel model of each scenario was fitted using Stata SE (StataCorp. 2011. Stata Statistical Software: Release 12. College Station, TX: StataCorp LP) by applying the 'xtmixed' command, which performs PMLE based on the method presented in reference [11]. See Appendix 2 for the Stata commands used.

4.5 *Criteria of comparison*

For each parameter, we calculated the (model) relative bias. Our strategy consisted of using the relative root mean square error (RMSE) to decide which scenario provided the best estimators if the relative bias was not able to provide such a decision. Let $\hat{\beta}$ be the parameter estimator, where β represents φ, ψ, β_0, σ_2^2, or σ_1^2 in the model described by Equations (4) and (5) (or by Equations (1–3)). The bias of $\hat{\beta}$, denoted $B(\hat{\beta})$, was estimated as $\hat{B}(\hat{\beta}) = \bar{\hat{\beta}} - \beta_{\text{true}}$ based on the 500 replications, where β_{true} is the true value of the regression coefficient. The variance of $\hat{\beta}$ was estimated as $\hat{V}(\hat{\beta})$, and the RMSE ($\hat{\beta}$) was estimated as $\sqrt{\hat{B}(\hat{\beta})^2 + \hat{V}(\hat{\beta})}$. The relative bias, $B_R(\hat{\beta})$, the relative variance, $V_R(\hat{\beta})$, and the relative RMSE were computed as $[\hat{B}(\hat{\beta})/\beta_{\text{true}}]$, $\hat{V}(\hat{\beta})/\beta_{\text{true}}^2$, and $\sqrt{\hat{B}_R(\hat{\beta})^2 + \hat{V}_R(\hat{\beta})}$, respectively.

5. Results

Figure 3 shows the relative bias distribution of the overall 33 estimated parameters $\beta_0, \beta_1, \ldots, \beta_{30}, \sigma_1^2, \sigma_2^2$, for each scenario. The relative bias was approximately 0 for all of the scenarios. However, the $\boldsymbol{w}_1$-estimator had the lowest dispersion for the relative bias values, as opposed to estimators based on $\boldsymbol{w}_2$, $\boldsymbol{w}_3$, $\boldsymbol{w}_4$, $\boldsymbol{w}_6$, and $\boldsymbol{w}_9$. Each scenario provided the least-biased estimator in absolute value for at least one individual parameter (see Table A1 in Appendix 3).

Figure 4 shows that scenarios $\boldsymbol{w}_1$, $\boldsymbol{w}_5$, $\boldsymbol{w}_7$, and $\boldsymbol{w}_8$ provided estimators with lowest variance. By considering each parameter individually (see Table A1 in Appendix 3), the $\boldsymbol{w}_5$ scenario provided estimators with the lowest variance most of the time, followed by $\boldsymbol{w}_8$ and $\boldsymbol{w}_1$.

Figure 5 shows results for the RMSE that are similar to those of the variance. The unweighted two-level model ($\boldsymbol{w}_5$) provided the best RMSE values, followed by the unweighted three-level model ($\boldsymbol{w}_8$) and the two-level model based on $\boldsymbol{w}_1$ (see Table A1 in Appendix 3). Small differences are observed among these three scenarios. For 1 of the 500 generated samples, the optimisation algorithm of multilevel model failed for $\boldsymbol{w}_8$, and it failed for 88 generated data sets for scenario $\boldsymbol{w}_9$. The failure is possibly due to the subsampling leading to a non-concave log-pseudolikelihood function, in particular with the weights associated with scenario $\boldsymbol{w}_9$. Figure 6 displays the RMSE of some covariates and illustrates what we observed among the 33 individually estimated parameters. Estimators based on $\boldsymbol{w}_1$, $\boldsymbol{w}_5$, $\boldsymbol{w}_7$, and $\boldsymbol{w}_8$ have similar RMSE results. From all of these results, it seems that the best scenarios are those based on no weighting ($\boldsymbol{w}_5$ or $\boldsymbol{w}_8$) or based on conditional

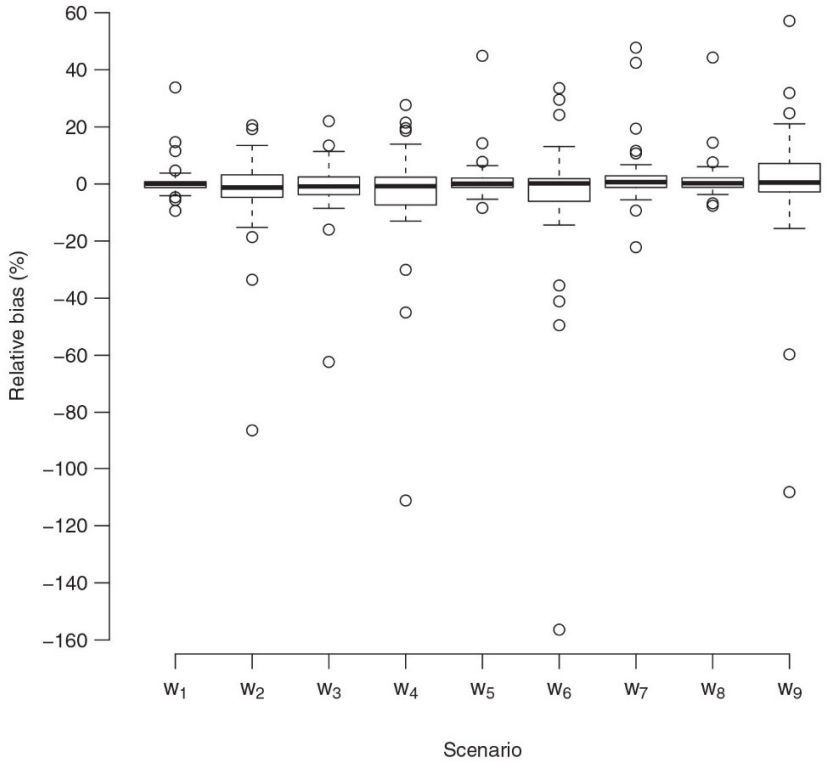


Figure 3. Distributions of the relative bias estimated values of 33 pseudolikelihood estimators (30 covariates, 1 intercept, and 2 variance parameters) obtained from a two-level random-intercept model ($\boldsymbol{w}_1$–$\boldsymbol{w}_6$) and a three-level random-intercept model ($\boldsymbol{w}_7$–$\boldsymbol{w}_9$), depending on the type of level-2 weights used. Relative bias estimated from 500 replications.

weights as level-2 weights ($\boldsymbol{w}_1$ or $\boldsymbol{w}_7$). We obtained similar results when performing the comparison based only on the sample set without a failure of the optimisation algorithm (411 samples); the ranking of the nine scenarios depending on their bias or their efficiency was unchanged.

6. Discussion

We studied the impact of the weights for level-2 units on the parameter estimates of a multilevel model on survey data. Our objective was to choose the type of level-2 weights to fit a multilevel model for estimating the contribution of lead sources contaminating residential interior floor dust. The results of our simulation study demonstrate that the best strategy for us is to use no level-2 weights (i.e. to use scenarios $\boldsymbol{w}_5$ or $\boldsymbol{w}_8$). If level-2 weights must be used, we should use the inverse of the conditional inclusion probabilities for the level-2 weights (scenarios $\boldsymbol{w}_1$ or $\boldsymbol{w}_7$).

We studied a situation in which the hierarchical structure of the multilevel model and the hierarchical structure of the sampling design do not match, in particular when the highest level of the sampling design is not used as the top level in the multilevel model. This situation may often occur when we use public data. In this situation, the use of overall sampling weights (whether post-stratified or not) for the top-level units of a multilevel model seems to not be a good practice. Overall sampling weights were used in the scenarios $\boldsymbol{w}_2$, $\boldsymbol{w}_3$, $\boldsymbol{w}_4$, and $\boldsymbol{w}_6$, which provided the worst results.

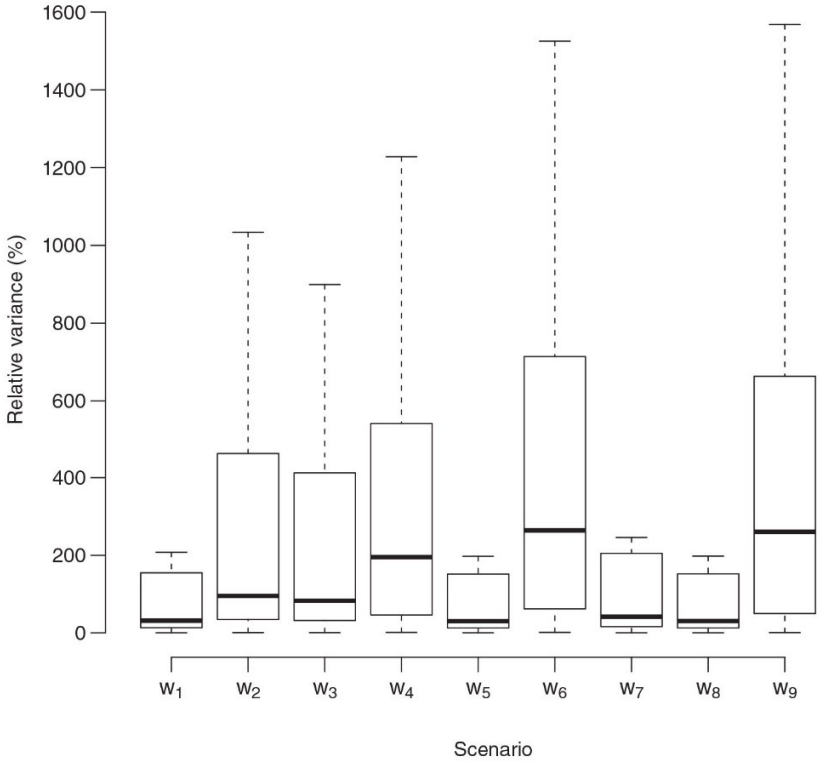


Figure 4. Distributions of the relative variance estimated values of 33 pseudolikelihood estimators (30 covariates, 1 intercept, and 2 variance parameters) obtained from a two-level random-intercept model ($\boldsymbol{w}_1$–$\boldsymbol{w}_6$) and a three-level random-intercept model ($\boldsymbol{w}_7$–$\boldsymbol{w}_9$), depending on the type of level-2 weights used. Relative variance estimated from 500 replications.

Even so, it can be attractive for the analyst to recover the weighting that was lost by not using units of a higher level of clustering than the top-level units of the multilevel model; one way to recover this weighting is through the introduction of overall sampling weights for these top-level units. However, the log-pseudolikelihood is a sum of lower-level weighted log-pseudolikelihood expressions as described in Equation (6) and more generally described by Equation (5) in reference [11]. Thus, the form of the log-pseudolikelihood involves the use of separate weights at each level and not weights based on the overall inclusion probabilities. Therefore, if we must use some weights, it is illogical and not rigorous to use overall sampling weights for units of any levels and, in particular, for the top-level units of a multilevel model. If there are in fact higher levels than the top level of the model, there is no justification in the log-pseudolikelihood expression for allowing us to use weights other than the conditional weights for the top-level units of the model.

The sampling design may produce different estimates based on the extent to which it is informative for Y [11]. Different results for the unweighted scenarios ($\boldsymbol{w}_5$ and $\boldsymbol{w}_8$) and the other (weighted) scenarios may indicate informative probabilities in general. In our study, unweighted models appeared to estimate the contribution of lead sources polluting the interior floor dust more accurately and more precisely. This finding implies that weighted analyses do not systematically outperform unweighted analyses to obtain efficient estimators when the sampling design is informative. This observation applies for a situation in which no weights were needed for level-1 and

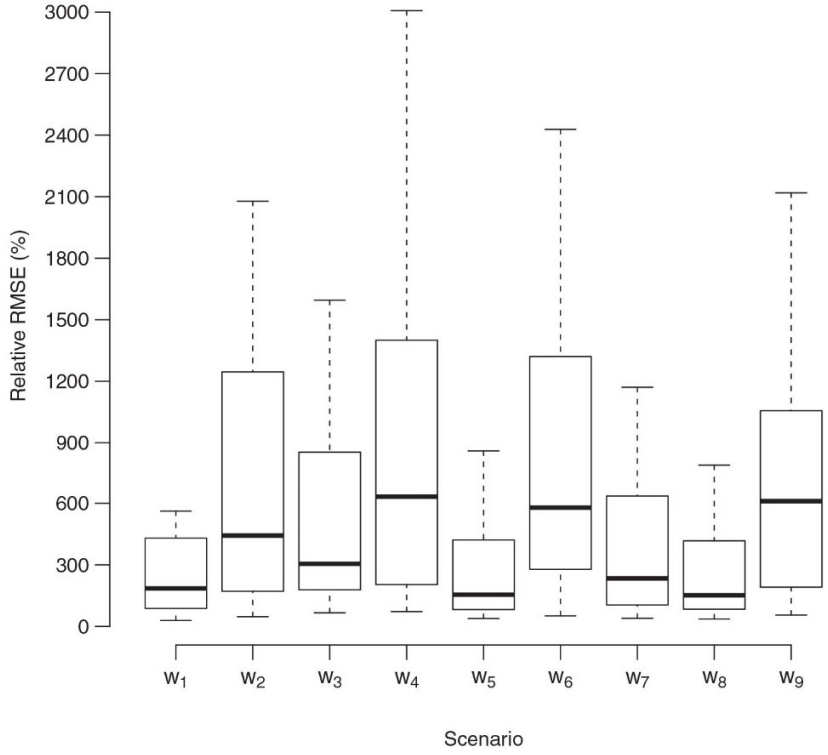


Figure 5. Distributions of the relative RMSE estimated values of 33 pseudolikelihood estimators (30 covariates, 1 intercept, and 2 variance parameters) obtained from a two-level random-intercept model ($\boldsymbol{w}_1$–$\boldsymbol{w}_6$) and a three-level random-intercept model ($\boldsymbol{w}_7$–$\boldsymbol{w}_9$), depending on the type of level-2 weights used. Relative RMSE estimated from 500 replications.

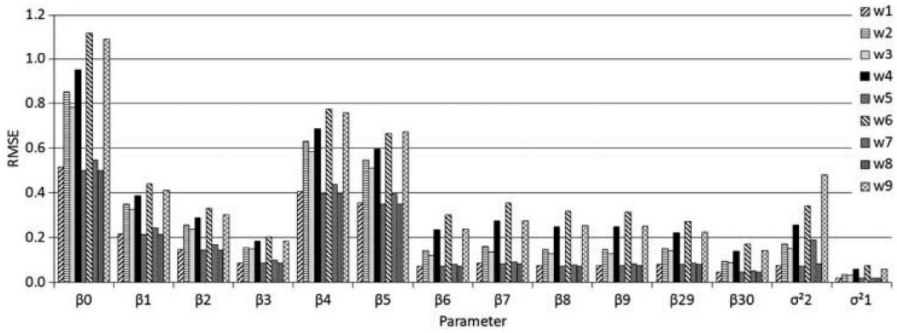


Figure 6. RMSE of some estimators obtained from a two-level random-intercept model ($\boldsymbol{w}_1$–$\boldsymbol{w}_6$) and a three-level random-intercept model ($\boldsymbol{w}_7$–$\boldsymbol{w}_9$), depending on the type of level-2 weights used. RMSE estimated from 500 replications.

is based on a particular sampling design. Further results are thus needed to describe the situations in a more general context for which unweighted analyses can be used when the sampling is informative.

Higher units than the top-level units of our two-level models ($\boldsymbol{w}_1$ through $\boldsymbol{w}_6$) did not really constitute a level because of the subsampling. The case of three levels could be easily studied by adapting our sampling programme to generalise our results. Additional parameters, such as the cluster size (for us, the number of rooms within each housing unit) or the sampling fraction (at a stage, the ratio between the number of sampled units and the population size), which most likely have an impact on estimates, should also be considered to study a more general case. In our study, the cluster size was small; in this situation, the pseudolikelihood estimators of the regression coefficients were noticeably biased when level-1 weights are used [8]. We highlight that when no level-1 weights are used, bias may also appear if unsuitable level-2 weights are used. Furthermore, it would be useful to perform an analogous study with unequal inclusion probabilities for level-1 units. In reference [1], using scaled level-1 weights is recommended when there are unequal inclusion probabilities for the level-1 units. It would be important to study the effect of level-2 weights on estimation in this situation. In this case, we believe that it would be more cautious to use conditional level-2 weights as usual (corresponding to our scenario $\boldsymbol{w}_1$). Indeed, we demonstrated that the results of scenario $\boldsymbol{w}_1$ were close to those of the best scenarios, i.e. scenarios with no level-2 weights. Further simulation studies with complementary findings are needed to generalise these first results about two-level weights in the context of multilevel modelling.

In our multilevel models, we did not have the option to declare the stratification of the sampling design in the pseudolikelihood calculated by the software. Its use may have reduced the variance of estimators and might have changed the ranking of the nine scenarios depending on their efficiency (RMSE). Although the pseudolikelihood approach seems to be able to account for stratification (at stage 1), as explained in section 5 of reference [11], the 'xtmixed' command of Stata does not. Moreover, it is not easy to treat the stratification in multilevel modelling on survey data for analysts, even if reference [11] described this point theoretically. In the manual that describes the Stata programme 'gllamm', which performs multilevel modelling on survey data and was developed by the authors of reference [12], stratification is not discussed.

Sensitivity analyses are advised to compare different weightings [1,11]. In reference [1], analysts are advised to conduct weighted and unweighted analyses. In our study, we applied the advice of reference [11] by simulating a finite population from an estimated model. Then, we selected a sample using a sampling design analogous to the actual design, and we performed an investigation to discover whether the model parameters were properly recovered depending on the different weighting methods. This advice was about level-1 weights. We applied it to level-2 weights, and in regard to our survey data, such an analysis was useful. Indeed, the findings of the literature about sampling design with two stages do not seem to be directly applicable to other designs (e.g. designs with more than two stages, which are frequently used in health surveys). Weighting is still debated even in single-level modelling [9]. Thus, in the more complex context of multilevel modelling, the use of weighting must be thorough and should currently be applied cautiously.

Researchers working on other environmental data may use our findings to build their multilevel models. For instance, data about indoor air quality usually have log-normal distributions such as those for lead levels. Moreover, the factors we used in our simulation protocol, such as the urbanisation and the period of construction of dwellings, are often related to air pollution levels.

7. Conclusion and recommendations

In this paper, we sought level-2 weights to use in a multilevel model for estimating the contribution of lead sources polluting the interior floor dust in French housing. Rooms were level-1 units, and dwellings were level-2 units. We used data collected in a French survey, with units hierarchically higher than dwellings. Based on this particular data set, we demonstrated that level-2 weights can

induce differences in point estimates and, thus, must be used cautiously. Although our results are not generalisable to all situations, we were able to suggest some warnings and recommendations.

When the highest level of the sampling design is not used as the top level in a multilevel model, attempting to recover the lost information by introducing overall sampling weights for the top-level units of the model is not a good idea. In this situation, using only the reciprocal of the conditional selection probability of the units of the top level of the model may be the best method when top-level weights will be used. Users of public databases for which only one type of weight is provided should not use the weights for the top level of their model if the weights are not conditional weights. Similarly, when some weights are provided but not fully described, it appears better to not use weights and to thus perform an unweighted analysis. Unweighted models should be fitted regardless. If weighted and unweighted results differ too greatly in that they lead to different inferential decisions, further investigation must be made. The user should perform a simulation sensitivity analysis based on her/his survey data to ensure that weights do not lead to biased and inefficient estimators. Further research regarding multilevel modelling on survey data must be conducted to extend the results of the present study because such modelling provides the potential to study some issues about survey data that are difficult to treat with other statistical tools from the applied statistics literature.

Note

1. Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1080/09593330.2013.847404.

References

[1] A. Carle, *Fitting multilevel models in complex survey data with design weights: Recommendations*, BMC Med. Res. Methodol. 9 (2009), p. 49.

[2] A. Etchevers, C. Lecoffre, A. Le Tertre, Y. Le Strat, S.-I. Groupe Investigateurs, C. De Launay, B. Bérat, M.-L. Bidondo, M. Pascal, N. Fréry, P. De Crouy-Chanel, M. Stempfelet, J.-L Salomez, and P. Bretin, *Blood lead level in children in France, 2008–2009*, BEHweb (2010). Available at www.invs.sante.fr/behweb/2010/02/index.htm.

[3] INSEE, *French Population Census 2006*, French National Institute of Statistics and Economic Studies (2008). Available at http://www.recensement.insee.fr/accesDonneesTelechargeables.action

[4] Q. Jiang and P.A. Succop, *A study of the specification of the log-log and log-additive models for the relationship between blood lead and environmental lead*, J. Agric. Biol. Environ. Stat. 1 (1996), pp. 426–434.

[5] J.-P. Lucas, B. Le Bot, P. Glorennec, A. Etchevers, P. Bretin, F. Douay, V. Sébille, L. Bellanger, and C. Mandin, *Lead contamination in French children's homes and environment*, Environ. Res. 116 (2012), pp. 58–65.

[6] D. Pfeffermann, *Modelling of complex survey data: Why model? Why is it a problem? How can we approach it?* Survey Methodol. 37 (2011), pp. 115–136.

[7] D. Pfeffermann, A.M. Krieger, and Y. Rinott, *Parametric distributions of complex survey data under informative probability sampling*, Stat. Sin. 8 (1998), pp. 1087–1114.

[8] D. Pfeffermann, C.J. Skinner, D.J. Holmes, H. Goldstein, and J. Rasbash, *Weighting for unequal selection probabilities in multilevel models*, J. R. Stat. Soc. Ser. B Stat. Methodol. 60 (1998), pp. 23–40.

[9] R.W. Platt and S.B. Harper, *Survey data with sampling weights: Is there a 'best' approach?* Environ. Res. 120 (2013), pp. 143–144.

[10] S. Rabe-Hesketh, *Multilevel modeling of complex survey data*, West Coast Stata Users' Group Meetings 2007, Marina del Rey, CA, 2007.

[11] S. Rabe-Hesketh and A. Skrondal, *Multilevel modelling of complex survey data*, J. R. Stat. Soc. Ser. A Stat. Soc. 169 (2006), pp. 805–827.

[12] S. Rabe-Hesketh, A. Skrondal, and A. Pickels, *GLLAMM Manual, Tech. Rep. 160, Division of Biostatistics*, University of California, Berkeley, 2004.

[13] S.W. Rust, D.A. Burgoon, B.P. Lanphear, and S. Eberly, *Log-additive versus log-linear analysis of lead-contaminated house dust and children's blood-lead levels*, Environ. Res. 72 (1997), pp. 173–184.

[14] C.-E. Särndal, B. Swensson, and J. Wretman, *Model Assisted Survey Sampling*, Springer-Verlag New York, Inc., New York, NY, 1992.

[15] C.J. Skinner, *Domain means, regression and multivariate analysis*, in *Analysis of Complex Surveys*, C.J. Skinner, D. Holt and T.M. F. Smith, eds., Wiley, Chichester, 1989, pp. 59–88.

Appendix 1. Covariates used in the model

Notation: 'β_i denotes the regression coefficient: [The covariate label] – The covariate description. The information level. Details about categorical covariates if applicable.' Numeric covariates are used as log-transformed in the model.

- β_1, β_2 : [fl_entry] – Floor of the home entrance. Level-2 (home). Three modalities.
- β_3 : [season] – Period of the year in which the home was investigated. Level-2 (home). Dummy.
- β_4, β_5 : [wet_land] – Whether wet cleaning is applicable to the landing of the apartment. Yes if mop or sponge can be used; no if only vacuum cleaner or broom can be used. Level-2 (home). Three modalities.
- $\beta_6, \beta_7, \beta_8, \beta_9$: [room] – Type of investigated room. Level-1 (room). Five modalities.
- β_{10} : [freq_wet] – Weekly frequency of wet cleaning of the floor (number of times per week). Level-1 (room).
- β_{11} : [freq_dry] – Weekly frequency of dry cleaning of the floor (number of times per week). Level-1 (room).
- β_{12} : [pl_sample] – Location in the room where the dust sample was collected. Level-1 (room). Dummy.
- β_{13} : [occ_risk] – Number of occupational activities related to lead practiced by the household members (also possibly practiced as leisure). Level-2 (home).
- β_{14} : [xrf_bal] – X-ray fluorescence (XRF) measurement (mg cm^{-2}) of the lead loading of the coating of the exterior railing of the terrace, loggia, or balcony; set to 0 if no exterior railings. Level-2 (home).
- β_{15} : [pb_soil × often] – Lead concentration (mg kg^{-1}) of the soil of the outdoor play area of the child when the child *often* plays in this play area. Level-2 (home).
- β_{16} : [pb_soil × all the time] – Lead concentration (mg kg^{-1}) of the soil of the outdoor play area of the child when the child plays in this play area *all the time*. Level-2 (home).
- β_{17} : [pb_hard × often] – Lead loading (µg m^{-2}) of the outdoor play area of the child, when it is on hard surface and when the child *often* plays in this play area. Level-2 (home).
- β_{18} : [pb_hard × all the time] – Lead loading (µg m^{-2}) of the outdoor play area of the child, when it is on hard surface and when the child plays in this play area *all the time*. Level-2 (home).
- β_{19} : [pb_land] – Floor dust lead loading (µg m^{-2}) of the landing of the apartment measured using wipe sampling; set to 0 if no landing. Level-2 (home).
- β_{20} : [n_car] – Annual flow of vehicles of the closest road to the home divided by the distance (km) between the road and the home. Level-2 (home).
- β_{21} : [old_build] – Old buildings within a radius of 50 m have been demolished or renovated in the past. Level-2 (home). Dummy.
- β_{22} : [leisure] – How often a hobby related to lead is practiced inside the home? (number of times per year). Level-2 (home).
- β_{23} : [extwork] – Whether work outside the home was performed in the past six months before the survey. Level-2 (home). Dummy.
- β_{24} : [inwork] – Whether work inside the home was performed in the past six months before the survey. Level-2 (home). Dummy.
- β_{25} : [basias] – Score assigned to industrial sites or service activities around the home, current or former, having a potentially polluting activity (lead). Level-2 (home).
- β_{26} : [basol] – Score about polluted sites and soils (potentially lead contaminated) around the home, involving a government action, preventive or curative. Level-2 (home).
- β_{27} : [bdrep] – Score about the lead emission in air of plants subject to authorisation (industrial plant and stockbreeding). Level-2 (home).
- β_{28} : [smoking] – Average daily time when someone smokes inside the home. Level-2 (home).
- β_{29} : [det_xrf] – Sum of the maximal XRF measurements of each diagnosis unit of the room. Only diagnosis units with a deteriorated coating. Level-1 (room).
- β_{30} : [use_xrf] – Sum of the maximal XRF measurements of each diagnosis unit of the room. Only diagnosis units with a coating in usual condition. Level-1 (room).

Appendix 2. Stata commands for multilevel fitting

The two-level model estimates were obtained using the following command:

```
xtmixed Y list_of_covariates [pw = Pi_i_sachant_j] || ID_LGT2:, variance ml
pweight(w_1) vce(cluster EPCI3),
```

where `Y` is the outcome; `list_of_covariates` is the list of our 30 covariates separated by a blank, as described in Appendix 1; `Pi_i_sachant_j` is the column name of level-1 weights (here, they are all equal to 1); `ID_LGT2` is the column name of the level-2 identifiers; `variance` enables us to obtain covariance parameter estimates returned in terms of variance rather than standard deviation; `ml` indicates that the estimation is performed using the ML (it could be removed because restricted ML is not supported in weighted analyses); and `pweight(w_1)` indicates the weights to use for level-2 (here, w_1, for instance). The option `vce(cluster EPCI3)` makes it possible to account for the clustering of the level-2 units in `EPCI3` in the pseudolikelihood estimation without declaring `EPCI3` as a higher level. Here, this option is useless because it allows us to adjust the computation of the standard error of each parameter estimate, which is not actually helpful in this study. However, it is a useful pseudolikelihood option in general, as described by [11].

The three-level model estimates are obtained with a very similar command:

```
xtmixed Y list_of_covariates [pw = Pi_i_sachant_j] || EPCI3:,
pweight(SamplingWeight_deg1) || ID_LGT2:, variance ml pweight(w_1),
```

where we added the identifiers of the highest level unit with `EPCI3` and their weight, `pweight(SamplingWeight_deg1)`; here `SamplingWeight_deg1` are $1/\pi_k$.

Appendix 3. Estimated bias, variance, and RMSE

Table A1. Best individual estimated bias, variance, and RMSE of each of the 33 parameters estimated with a two-level random-intercept model (w_1–w_6) and a three-level random-intercept model (w_7–w_9) depending on the type of level-2 weights used.

Parameter	True value	Best |bias|; top 3	Best variance; top 3	Best RMSE; top 3
β_0	−0.451	w_7, w_3, w_2	w_5, w_8, w_1	w_5, w_8, w_1
β_1	−0.185	w_7, w_5, w_1	w_5, w_8, w_1	w_5, w_8, w_1
β_2	−0.262	w_8, w_5, w_1	w_5, w_8, w_1	w_5, w_8, w_1
β_3	0.435	w_6, w_1, w_5	w_5, w_8, w_1	w_5, w_8, w_1
β_4	1.572	w_9, w_7, w_8	w_5, w_8, w_1	w_5, w_8, w_1
β_5	0.170	w_6, w_2, w_9	w_5, w_8, w_1	w_5, w_8, w_1
β_6	−0.247	w_1, w_5, w_8	w_5, w_8, w_1	w_5, w_8, w_1
β_7	0.170	w_8, w_5, w_1	w_5, w_8, w_1	w_5, w_8, w_1
β_8	−0.007	w_1, w_5, w_8	w_5, w_8, w_1	w_5, w_8, w_1
β_9	0.211	w_8, w_5, w_1	w_5, w_8, w_1	w_5, w_8, w_1
β_{10}	0.037	w_7, w_8, w_5	w_5, w_8, w_1	w_5, w_8, w_1
β_{11}	−0.072	w_8, w_5, w_1	w_5, w_8, w_1	w_5, w_8, w_1
β_{12}	0.030	w_1, w_6, w_5	w_5, w_8, w_1	w_5, w_8, w_1
β_{13}	0.084	w_7, w_1, w_8	w_5, w_8, w_1	w_5, w_8, w_1
β_{14}	0.473	w_7, w_5, w_8	w_5, w_8, w_1	w_5, w_8, w_1
β_{15}	0.105	w_5, w_8, w_7	w_5, w_8, w_1	w_5, w_8, w_1
β_{16}	0.052	w_1, w_8, w_5	w_5, w_8, w_1	w_5, w_8, w_1
β_{17}	0.099	w_9, w_4, w_3	w_5, w_8, w_1	w_5, w_8, w_1
β_{18}	0.114	w_8, w_5, w_7	w_5, w_8, w_1	w_5, w_8, w_1
β_{19}	0.297	w_1, w_9, w_8	w_5, w_8, w_1	w_5, w_8, w_1
β_{20}	0.020	w_6, w_3, w_8	w_5, w_8, w_1	w_5, w_8, w_1
β_{21}	0.399	w_1, w_9, w_8	w_5, w_8, w_1	w_5, w_8, w_1
β_{22}	0.134	w_4, w_3, w_2	w_5, w_8, w_1	w_5, w_8, w_1
β_{23}	−0.069	w_1, w_5, w_8	w_5, w_8, w_1	w_5, w_8, w_1
β_{24}	0.261	w_5, w_1, w_7	w_8, w_5, w_1	w_8, w_5, w_1
β_{25}	0.125	w_2, w_6, w_3	w_8, w_5, w_1	w_8, w_5, w_1
β_{26}	0.022	w_3, w_2, w_1	w_5, w_8, w_1	w_5, w_8, w_1
β_{27}	0.248	w_6, w_5, w_8	w_5, w_8, w_1	w_5, w_8, w_1
β_{28}	0.263	w_5, w_8, w_7	w_5, w_8, w_1	w_5, w_8, w_1
β_{29}	0.133	w_1, w_4, w_5	w_5, w_8, w_1	w_5, w_8, w_1
β_{30}	0.015	w_1, w_5, w_8	w_8, w_5, w_1	w_8, w_5, w_1
σ_2^2	0.800	w_5, w_1, w_8	w_5, w_1, w_8	w_5, w_1, w_8
σ_1^2	0.450	w_1, w_5, w_8	w_5, w_8, w_1	w_5, w_8, w_1

Notes: The detailed bias values, the detailed variance values, and the detailed RMSE values are available in the supplementary information. A total of 500 replications.