FORECASTING MODELS AND INDEX DEVELOPMENT: A NON-PARAMETRIC APPROACH

Salima Taibi

MÉMOIRE
Présenté en vue de l'obtention du

Diplôme d'habilitation à diriger les recherches
UNIVERSITÉ de ROUEN

MODÈLES PRÉVISIONNELS ET ÉLABORATION D'INDICES : UNE APPROCHE NON PARAMÉTRIQUE

SALIMA TAÏBI-HASSANI
Docteure en Mathématiques Appliquées
de l'Université Paul Sabatier Toulouse
Esitpa - Unité Agri'Terr
LMRS

Soutenu le 6 JANVIER 2016 devant le Jury composé de

Dominique FOURDRINIER	Professeur Université de Rouen	Président
Christian MOUGIN	Directeur de Recherche INRA/AgroParisTech	Examinateur
Maria PARLINSKA	Professeur WULLS University Warsaw	Rapporteur
Serge PERGAMENCHTCHIKOV	Professeur Université de Rouen	Rapporteur
William E. STRAWDERMAN	Professeur RUTGERS University USA	Rapporteur

SOMMAIRE
Remerciements
Chapitre 1

1.1 Introduction générale

1.2 Production scientifique

Chapitre 2 Les problèmes d'estimation et de modélisation non paramétrique

2.1 Estimation non paramétrique de la fonction de hasard par la méthode du noyau et des k-points les plus proches pour des observations dépendantes.
2.1.1 Estimation de la fonction de hasard
2.1.2 Estimation non paramétrique du mode de la fonction de hasard
2.2 Estimation non paramétrique de la fonction de hasard dans le cas de données censurées à droite
2.3 Méthode directe pour l'estimation non paramétrique du taux de hasard sous données censurées à gauche
2.4 Prédiction modale non paramétrique : convergence uniforme de l'estimateur à noyau du mode conditionnel à partir d'observations dépendantes
2.4.1 Résultat général
2.4.2 Estimateurs à noyau
2.4.3 Résultats de convergence
2.4.4 Application à la prédiction modale
2.5 Convergence ponctuelle d'un estimateur de la régression d'une variable aléatoire réelle par rapport à une variable aléatoire dans un espace mesurable pour des observations dépendantes : cas du consentement à payer.
2.5.1 Estimation non paramétrique de la fonction de régression
2.5.2 Hypothèses générales
2.5.3 Estimation non paramétrique du consentement à payer
2.5.4 Propriétés de convergence ponctuelle de l'estimation du consentement à payer : cas des observations α et ϕ mélangeantes.

Chapitre 3 Modélisation des données socio-économiques

3.1 Modélisation des données qualitatives et application dans le cas de données socio-économiques. La méthode des Random Forests versus Analyse discriminante.
3.1.1 Étude sur les résultats d'enquête : cas du consentement à payer
3.1.2 Les forêts aléatoires
3.1.3 Résultats
3.2 Méthode des programmes : plans fractionnaires et plans de sondage
3.2.1 Introduction et contexte
3.2.2 Modélisation
3.2.3 Sélection de modèles
3.2.4 Conclusion
3.3 La méthode des programmes et le problème du statu quo pour des données longitudinales
3.4 L'évaluation du consentement à payer : méthode des programmes et hétérogénéité des préférences
3.5 Modèle du rendement en riz dans la Province de Tamatave

3.5.1 Contexte
3.5.2 Modèle explicatif du rendement en riz
Chapitre 4 Modélisation des données biologiques, agronomiques et physico-chimiques
4.1 Démarche statistique pour la sélection des indicateurs par Random Forests pour la surveillance de la qualité des sols
4.1.1 Synthèse bibliographique sur les indices d'état d'un sol
4.1.2 Construction d'un indice multiparamétré de la qualité des sols
4.2 Modèle de transfert des métaux lourds du sol vers la plante
4.3 Concentrations de polluants dans le sol
4.4 Modélisation de données agronomiques
4.4.1 Hétérogénéité intraparcellaires en agriculture de précision
Chapitre 5 Animation et responsabilités en recherche
5.1 Animation du laboratoire Lamsad.
5.2 Conférence invitée,présidence de séance, expertise
5.3 Les projets de recherche
5.3.1 Les projets de recherche régionaux
5.3.2 Les projets de recherche nationaux
5.3.3 Les projets de recherche internationaux
5.4 L'encadrement en recherche
5.4.1 L'encadrement de l'équipe du Lamsad
5.4.2 L'encadrement de post-doctorants
5.4.3 L'encadrement de doctorants
5.4.4 Encadrement d'ingénieurs d'étude et de recherche
5.4.5 L'encadrement de l'équipe BIOMATH
5.4.5 L'encadrement de stagiaires
5.5 Travaux de vulgarisation
5.5.1 Les sciences de la vie mises en équation
5.5.2 Quand mathématiques riment avec développement durable
Chapitre 6 Responsabilités et activités d'enseignement
6.1 Responsabilités d'enseignement
6.2 Activités d'enseignements
6.3 Contributions pédagogiques
6.4 Formation par apprentissage, Formation continue
6.5 Formation par la voie VAE
6.6 Mise en place d'observatoires
Chapitre 7 Conclusion et perspectives de recherche
BIBLIOGRAPHIE
ANNEXES
Curriculum Vitae
Liste des abréviations

Remerciements

Je remercie vivement le Professeur Domninique Fourdrinier qui m'a soutenue et accompagnée, d'avoir accepté de présider ce jury

Je remercie profondément les Professeurs Maria Parlinska, Bill Strawderman, Serge Pergamentchikov et Christian Mougin pour l'intérêt qu'ils ont voulu porter à ce travail et qui me font l'honneur de participer à ce jury.

Mes remerciements vont à :

Mon conjoint Jaouad, mes filles Amel et Dalel, mes parents et toute ma famille pour leur constant soutien,

Maryvonne Iacovella pour ses encouragements,

Paul Denieul pour son accueil à l'Esitpa et sa confiance,

Claude Dellacherie pour son accueil au sein du laboratoire LAMS et Gérard Grancher pour sa disponibilité,

Elie Youndjé pour sa collaboration,

Dimitri avec qui j'ai partagé des travaux pluridisciplinaires en économie et statistique ,

Manasé Bezara pour nos échanges dans les projets de développement,

Karine Laval, Saïd Koutani et Geoffroy Belhenniche pour leurs encouragements,

Jeanne Chantal Dur pour les heures passées sur la Base BIO2,

Tous les collègues de l'Esitpa et du LMRS et en particulier Nathalie, Laurence, Patrice, Jérôme, Asma, Anna, Diégo, Sylvia, Anita, Walter, Vlad Barbu,

Sorin Muntean, Ouerdia Arkoun, Iryna Petrovska, Aurore Blot, Jeanne Bodin et Saturnin Touk,

Antonio Bispo que je remercie pour la création du groupe Biomath, dans le cadre du projet " Bioindicateurs",

La Région Haute Normandie et L'Ademe pour leur soutien financier dans les projets de recherche

Enfn je remercie toutes les personnes qui de près ou de loin ont aidé ou contribué à ce travail.

RÉSUMÉ

Mes recherches ont débuté au Laboratoire de Statistique et Probabilités de l'Université Paul Sabatier-Toulouse) sur des problèmes d'estimation et de prédiction non paramétriques en septembre 1981.

Après avoir occupé un poste de maître-assistante chargée de cours au sein de l'Université d'Annaba (Algérie) de 1986 à 1992 , j'ai été recrutée en qualité d'ATER entre 1992 et 1996 à l'Université de Rouen. J'ai poursuivi mes travaux de recherche sur l'estimation de la fonction de hasard au sein du Laboratoire L.A.M.S, maintenant Rapahël Salem (UMR CNRS 6085), travaux conjointement menés avec Élie Youndjé (Hassani et Youndjé 1997, Hassani et Youndjé 2003). Les diverses fonctions assurées à l'Esitpa, liées à l'enseignement et à la recherche, m'ont amenée à coordonner ou participer à des projets de recheche pluridisciplinaires régionaux, nationaux et internationaux et sur des thématiques environnementales, agricoles et socio-économiques.

Mes travaux de recherche se décomposent en deux volets, recherche fondamentale et appliquée, et sont centrés sur l'estimation, la construction de modèles non paramétriques et l'élaboration de modèles et plus particulièrement d'indices.

À cet égard les modèles non paramétriques connaissent un véritable essor en raison que les distributions des données biologiques, économiques, sociologiques sont assez souvent méconnues.

Les résultats en recherche fondamentale portent sur l'estimation non paramétrique de fonctionnelles telles que la fonction de densité, la fonction de hasard, la fonction de régression, la densité conditionnelle, le mode conditionnel. Les problèmes sont traités pour des situations assez diverses que nous pouvons rencontrer telles que la dépendance des données, la présence d'une variable censurant l'information ou encore lorsque les valeurs prises par la variable explicative sont dans un espace mesurable. J'ai montré des résultats de convergence sur la fonction de hasard dans le cas de données censurées à droite, et construit un nouvel estimateur dans le cas de données censurées à gauche, de manière directe, alors qu'il n'existe qu'une approche utilisant l'inversion du temps mais qui est non réaliste.

Une autre thématique de recherche a porté sur l'estimation de la fonction de régression dans le cas de données dépendantes lorsque entre autres, la variable à expliquer est le consentement à payer pour un programme de préservation, problème rencontré en socio-économie. J'ai réalisé la construction d'un estimateur et montré sa convergence précisément pour des processus mélangeants et lorsque les variables explicatives sont à valeurs dans des espaces mesurables (Taïbi et *al.* 2015a).

J'ai encadré Adigaw-E-Touck Adigaw S.L en doctorat . Celui-ci a soutenu sa thèse à l'Université de Rouen en Janvier 2013. Ces travaux ont porté sur l'estimation non paramétrique de fonctionnelles dans le cas censuré.

Depuis 2012, je coencadre en thèse Jérôme Dantan enseignant-chercheur à l'Esitpa sur le thème du Big Data. Les travaux de sa thèse visent à développer des modèles de comportement de systèmes complexes et une approche systémique unifiée pour l'optimisation durable des systèmes socio- environnementaux. La soutenance est prévue courant 2016.

À l'heure actuelle je coencadre aussi à distance la thèse d'Assia Ayache, doctorante de l'Université de Constantine (Algérie), sur des méthodes de classification supervisée

et non supervisée telles que les réseaux de neurones et les forêts aléatoires (Random Forests).

J'ai encadré en modélisation statistique quatre post-doctorants, dont deux chercheurs d'universités étrangères, sur des thématiques de recherche agronomique, économique, écologique et statistique. J'ai encadré des étudiants en stage M1, M2 et Ingénieur sur des thématiques en lien avec les projets de recherche.

J'ai au total une vingtaine d' articles parus dans des revues internationales avec comité de lecture, une cinquantaine de communications orales, trois conférences invitées et expertisé plusieurs articles.

Depuis 2002, j'ai initié et/ou participé à dix projets, six projets régionaux, deux projets internationaux et deux projets nationaux. Entre 2006 et 2013, j'ai participé aux deux programmes de recherche, "Bioindicateurs de la qualité des sols" (phase 1 et phase 2) portés par l'ADEME (Agence de l'Environnement et de la Maîtrise de l'Énergie). J'ai collaboré aux traitements statistiques et élaboré des démarches de modélisation conjointement avec le laboratoire Biosol (Esitpa), l'unité PESSAC de l'Inra de Versailles et le laboratoire de microbiologie du froid de l'Université de Rouen (Laval et *al.* 2009, Taïbi et *al.* 2013, Taïbi et *al.* 2015b).

Ce projet a été reconduit par l'ADEME en 2009 pour une deuxième phase de trois années, intitulé "Bioindicateurs de la qualité des sols II". Ce programme de recherche national compte une vingtaine d'équipes de recherche. J'ai animé le groupe Biomath constitué par l'ADEME durant plus de trois années pour conduire les travaux de modélisation statistique de ce programme de recherche, et encadré Jeanne Bodin en post-doctorat de biostatistique (Février 2012-Mars 2013).

Les travaux de recherche "mono indicateur" ont permis d'établir des valeurs de références et de mettre en évidence la sensibilité à un environnement donné (pratiques agricoles et usages/stress chimique). L'approche globale utilisant des méthodes supervisées telle que la méthode des forêts aléatoires (Taïbi et *al.* 2013, Taïbi et *al.* 2015b) a permis la sélection d'une batterie d'indicateurs explicatifs et discriminants pour un stress donné et de trouver des résultats probants dans des contextes d'évaluation de l'état du sol à la fois complexe et dynamique.

Enfin une démarche pour établir l'opérationnalité d'un indicateur à partir des résultats d'enquêtes à dire d'experts a été établie. Tous ces résultats permettront de calculer un indice fonctionnel simple ou agrégé de l'état d'un sol.

J'ai participé à la coordination des projets EMIRE 1&2 (Évaluation monétaire de l'impact du ruissellement érosif) entre 2010 et 2014. Ce programme de recherche pluridisciplinaire (Crastes et *al.* 2014) sur le thème de la monétarisation regroupe des chercheurs en économie de l'environnement, en économétrie, en statistique et en géographie spatiale. Pour lutter contre le ruissellement érosif, un programme organisé autour de trois moyens d'action a été proposé : la modification des pratiques agricoles, le développement des infrastructures de protection et la communication. La mise en place de

ce programme implique bien sûr une taxe supplémentaire pour les habitants, et afin d'évaluer monétairement ce risque, nous avons eu recours à une méthode utilisée en

économie pour évaluer des biens non marchands, la méthode des programmes (choice experiment). Les principaux problèmes qui découlent de cette méthode sont l'élaboration du plan d'expérience (fractionnaire), la sélection des attributs et les niveaux de ces attributs.

Un des thèmes de mes travaux porte sur les biais liés à la mise en place d'un plan fractionnaire dans la méthode des programmes et sur la méthodologie à adopter. Sur ce sujet j'ai encadré en post-doctorat de modélisation mathématique (Octobre 2011-Mars 2012) Ouerdia Arkoun titulaire d'un doctorat en mathématiques, option statistique, de l'Université de Rouen.

D'autres travaux ont porté sur la recherche des facteurs expliquant le choix des habitants de la Valllée du Commerce en utilisant le modèle logit polytomique à paramètres aléatoires. Actuellement, mes travaux sont centrés sur les problèmes rencontrés lors d'enquêtes longitudinales (Taïbi et *al.* 2014b) et, sur ce thème, j'ai encadré Iryna Petrovska, doctorante de l'Université Wulls (Warsaw Universiy of Life Sciences), accueillie à l'Esitpa pour un séjour de cinq mois (2013-2014).

Parmi les résultats de recherche obtenus pour la modélisation de données socio-économiques, je peux citer la prédiction du consentement à payer. Á l'instar de la méthode Disqual (Saporta 1977), j'ai élaboré une méthode de prédiction utilisant les Forêts Aléatoires (Breiman 2001) sur des données qualitatives. Cette méthode innovante présente deux intérêts : la construction d'un modèle de prédiction et une solution au problème des traitements statistiques en grande dimension (Laroutis et Taïbi 2011).

La nécessité d'élaborer des indices, tels que le taux de fiabilité, la fonction de survie, le consentement à payer, l'indice de richesse ou de pauvreté, l'indice d'état d'un sol, le taux de transfert de métaux lourds du sol vers la plante, m'a amenée à construire des modèles explicatifs et prédictifs, que cela soit une construction pas à pas à partir de données expérimentales ou que cela soit une construction d'estimateurs non paramétriques. J'ai élaboré ces démarches de modélisation en tenant compte des contraintes fréquemment rencontrées que sont l'hétérogénéité des données, le problème des grandes dimensions, l'autocorrélation des variables, la présence de données incomplètes ou censurées à droite ou à gauche ou de données tronquées ou aberrantes ou manquantes.

L'animation et la participation à ces différents projets de recherche m'ont amenée à travailler dans un contexte interdisciplinaire et à avoir plusieurs perspectives de recherche fondamentale et appliquée dont les acquis seront valorisés. Notamment à travers le programme international de développement des systèmes complexes Unitwin (Unesco) qui rassemble plusieurs laboratoires de recherche en modélisation (Taïbi et *al.* 2014 a) et aussi dans le cadre de la création du Master Msc sur le thème du"'Big data" dont j'ai la responsabilité et qui devrait ouvrir à la rentrée prochaine.

Chapitre 1

1.1 Introduction Générale

Mes activités de recherche ont débuté en DEA puis en thèse de doctorat (1982-1985) sur des travaux d'estimation et de prédiction non paramétriques au sein du Laboratoire de Statistique et Probabilités de l'Université Paul Sabatier à Toulouse (Toulouse III).

J'ai effectué mes travaux de doctorat sous la direction de Gérard Collomb. J'ai soutenu ma thèse le 26 septembre 1985. J'ai construit des estimateurs non paramétriques par la méthode du noyau de fonctions telles que la fonction de densité, la fonction de régression, la fonction de hasard (appelée aussi taux de défaillance ou taux de survie) ou de paramètres tels que le mode conditionnel.

J'ai développé des estimateurs non paramétriques de la fonction de hasard par la méthode du noyau et par la méthode des k-points les plus proches dans le cas censuré à droite et non censuré pour des processus faiblement dépendants et, plus précisément, uniformément fortement mélangeants (Collomb et *al.* 1985a, Hassani 1985). J'ai étudié plus particulièrement le cas des processus φ-mélangeants. Mes principaux résultats portent sur la convergence ponctuelle et presque complète d'estimateurs non paramétriques de cette fonction.

Dans le cas multimensionnel, j'ai construit un estimateur à noyau de la fonction de densité conditionnelle et du mode conditionnel (supposé unique) dans le cas de processus mélangeants. Il s'agit d'élaborer un modèle de prédiction modale non paramétrique. J'ai montré la convergence uniforme presque complète de l'estimateur à noyau de la fonction de densité conditionnelle dans le cas d'un processus stationnaire à valeurs dans $C \times \mathbb{R}$, où C un compact de $\mathbb{R}^p$ (Collomb et *al.*, 1986, Hassani 1985).

La plupart des résultats issus de ma thèse ont fait l'objet d'articles ou de communications dans des congrès.

J'ai ensuite poursuivi ces travaux à l'Université de Rouen au sein du Laboratoire L.A.M.S , maintenant Rapahël Salem (UMR CNRS 6085). Ces travaux ont été conjointement menés avec Élie Youndjé et ont porté sur le choix de la fenêtre de lissage dans l'estimation de la fonction de hasard par validation croisée (Hassani et Youndjé 1997, Hassani et Youndjé 2003).

Recrutée en 1998 comme enseignant-chercheur à l'Esitpa, Ecole d'Ingénieurs en Agriculture, j'ai continué à être membre du laboratoire de mathématiques Raphaël Salem (UMR CNRS 6085). Étant donnée la forte demande en recherche de modèles explicatifs et/ou prédictifs dans de nombreux domaines comprenant la biologie, l'agronomie et l'économie, mes travaux de recherche ont été orientés en ce sens.

Devant l'accumulation des données expérimentales, qu'elles soient issues de la biologie, de l'industrie ou du monde socio-économique, le besoin d'élaborer des modèles de prédiction connait un réel essor. Ma fonction au sein de l'Esitpa m'a permis de collaborer avec des chercheurs agronomes, écologues, biologistes, microbiologistes, économistes, sociologues, géographes, épidémiologistes, chimistes, médecins. En 2001, j'ai créé le laboratoire de Modélisation Statistique, le Lamsad, et j'ai initié (et/ou participé) des projets, de recherche pluridisciplinaire, régionaux, nationaux et internationaux.

Ma recherche se place donc dans le cadre de construction de modèles non paramétriques. En effet les populations sur lesquelles je travaille ne sont assujetties à aucune hypothèse distributionnelle et n'obéissent pas nécessairement à des lois de probabilités connues. J'ai de plus en plus mené des travaux de recherche sur des problématiques environnementales, agricoles et socio-économiques.

Parallèlement à ces travaux appliqués à la biologie, à l'agriculture, et à l'économie, mes recherches sur la fonction de hasard ont été étendues au cas de données censurées à gauche et ont été poursuivies notamment dans le cadre de la thèse de doctorat de Adigaw-E-Touck Adigaw S.L (2013). Ces travaux de thèse ont été coencadrés avec le professeur Dominique Fourdrinier de l'Université de Rouen. Nous avons déterminé un estimateur de la fonction de hasard dans le cas de données censurées à gauche. Un tel estimateur est habituellement construit en inversant l'estimateur de la fonction de hasard dans le cas de données censurées à droite. De manière plus réaliste nous avons construit un estimateur de manière directe et la propriété de convergence de cet estimateur a été établie (Taïbi-Hassani et Touk 2011, Touk 2013). Cet estimateur est obtenu par lissage de l'estimateur de Kaplan-Meier à gauche. La convergence presque-sûre uniforme, la normalité asymptotique et la décomposition asymptotique du biais, de la variance et de l'erreur quadratique moyenne ont été mises en évidence. De nouveaux estimateurs du taux de hasard et de la densité de probabilité pour données censurées à gauche sont ainsi obtenus par approche directe et les résultats de convergence presque-sûre uniforme et de normalité asymptotique sont prouvés. Nous établissons également la décomposition asymptotique de l'erreur quadratique intégrée ainsi que l'optimalité d'un critère de validation croisée. Des résultats de convergence ponctuelle en probabilité et presque-complète d'estimateurs de type delta-suites de la fonction de régression, de la fonction de densité et de hasard conditionnelles, sont ainsi obtenus pour des processus φ-mélangeants et α-mélangeants.

J'ai aussi montré la convergene d'un estimateur de la fonction de régression dans le cas d'espaces mesurables et de variables dépendantes. En particulier pour des données socio-économiques, lorsque par exemple la variable à expliquer est le consentement à payer pour un programme de préservation . J'ai réalisé la construction d'un estimateur et montré sa convergence en probabilité et presque complète, précisément pour des processus faiblement dépendants (mélangeants) et lorsque les variables explicatives sont à valeurs dans des espaces mesurables (Taïbi et *al.* 2015b).

Depuis 2011, avec Yann Pollet, Professeur au Conservatoire National des Arts et Métiers, je coencadre la thèse de Jérôme Dantan, enseignant à l'Ésitpa, centré sur le thème du Big Data. En effet, actuellement, nous assistons au développement des moyens d'enregistrement et de stockage de données. L'un des principaux défis de la société actuelle est la gestion de données qui sont de plus en plus nombreuses et qui se

révèlent souvent imparfaites, incomplètes ou aléatoires. La gestion de telles données est présente dans de nombreux secteurs d'activité, notamment pour assister les humains dans leurs prises de décision en fusionnant différentes données issues de nombreuses sources d'informations (mesures, capteurs, observations). Les travaux de cette thèse visent à développer des modèles de comportement de systèmes complexes (Dantan *al.*, 2015a ; 2015 b).

Depuis 2013, je coencadre à distance la thèse d'Assia Ayache, doctorante de l'Université de Constantine (Algérie), sur des méthodes de classification supervisée et non supervisée telles que les réseaux de neurones et les forêts aléatoires (Random Forests).

Entre 2006 et 2008, j'ai participé au programme de recherche, " Bioindicateurs de la qualité des sols"(phase 1) porté par l'ADEME (Agence de l'Environnement et de la Maîtrise de l'Énergie). J'ai collaboré aux traitements statistiques et élaboré des démarches de modélisation conjointement avec le laboratoire Biosol (Esitpa), l'unité PESSAC de l'Inra de Versailles et le laboratoire de microbiologie du froid de l'Université de Rouen (Laval et *al.*, 2009, Taïbi et *al.* 2015). Ce projet a été reconduit par l'ADEME en 2009 pour une deuxième phase de 3 années " Bioindicateurs de la qualité des sols II ". Ce programme de recherche national compte une vingtaine d'équipes de recherche Taïbi et *al.*, 2013. L'ADEME a constitué le groupe Biomath composé de statisticiens et d'informaticiens. J'ai été donc amenée à animer ce groupe durant plus de 3 années pour conduire les traitements statistiques de ce programme de recherche et à encadrer Jeanne Bodin docteure en écologie recrutée en postdoctorat (2012-2013). Outre la conception d'un carnet de route et l'organisation des travaux inter-disciplinaires en réponse aux questions soulevées par les chercheurs, il a fallu sélectionner les indicateurs les plus discriminants pour la surveillance de l'état d'un sol et construire une démarche de modélisation pour établir un indice multiparamétré de l'état d'un sol Taïbi et *al.*,2012a, Taïbi et *al.*,2012b, Taïbi et *al.*, 2013, Taïbi et *al.*, 2014, Bodin et *al.*, 2013, Peres et *al.*, 2013.

Parallèlement, j'ai été coordinatrice des projets EMIRE 1&2 (Évaluation monétaire de l'impact du ruissellement érosif), (2009-2014). Il s'agit de programmes pluridisciplinaires sur le thème de la monétarisation regroupant des chercheurs en économie de l'environnement, en économétrie, en statistique et en géographie spatiale.

Concernant le programme EMIRE 1, l'objectif est d'estimer le consentement à payer pour un bien marchand ou non marchand et de sélectionner les facteurs amenant les personnes à révéler leur préférence pour un programme de préservation par rapport à une palette de programmes qui leur est proposée. J'ai établi une nouvelle méthode basée sur la méthode des forêts aléatoires ou Random Forests (Breiman 2001).

J'ai construit une démarche de classement pour des variables qualitatives. En effet les variables socio-économiques étant souvent qualitatives ou ordinales, il s'agit de construire un modèle prédictif intégrant cette spécificité. La méthode d'évaluation contingente constitue une méthode économique quantifiant monétairement l'ensemble des valeurs que les individus attribuent à un bien environnemental donné. Cette méthode nécessite un questionnaire visant à révéler le consentement à payer (CAP) des individus pour la préservation des zones humides de l'estuaire de la Seine. Nous nous plaçons dans le cadre assez général de traitements de données d'enquêtes avec pour

objectif, d'une part, de prédire le consentement à payer, et d'autre part, de pouvoir reproduire cette méthodologie pour des vagues d'enquêtes successives. Nous nous sommes limités au cas où la variable dépendante est binaire, la procédure pouvant être étendue au cas de variables qualitatives polytomiques. L'objectif est de construire un modèle permettant de prédire le consentement à payer. Les prédicteurs étant pour la plupart des variables qualitatives (nominales ou ordinales), nous avons utilisé une procédure permettant de les transformer en variables quantitatives. Nous effectuons l'analyse des correspondances multiples des prédicteurs c'est-à-dire l'analyse des correspondances du tableau disjonctif. Les p variables explicatives sélectionnées $X_1, X_2, \ldots, X_p$ sont remplacées par les coordonnées des n individus sur les q axes factoriels ($q < p$), en opérant une pondération permettant de prendre en compte l'importance des composantes. C'est une méthode basée sur la méthode Disqual (Saporta 1977).

Deux méthodes de classement ont été mises en œuvre sur une base de données d'enquête administrée auprès d'un échantillon représentatif de 300 individus : l'analyse discriminante et la méthode des forêts aléatoires. Nous avons comparé leurs performances en termes de classement (Laroutis et Taïbi 2011).

Pour la deuxième phase du projet EMIRE (EMIRE II), afin de lutter contre le ruissellement érosif, un programme organisé autour de trois moyens d'action a été proposé : la modification des pratiques agricoles, le développement des infrastructures de protection et la communication. La mise en place de ce programme implique bien sûr une taxe supplémentaire pour les habitants. Il est alors nécessaire d'établir donc le montant de cette taxe et les moyens à allouer à chacun des trois moyens d'action. Ainsi une enquête a été réalisée dans l'objectif de définir la somme que les habitants seraient prêts à verser pour bénéficier d'un programme de lutte contre le ruissellement.

Afin d'évaluer monétairement ce risque, nous avons eu recours à une méthode utilisée en économie pour évaluer des biens non marchands, la méthode des programmes (Choice Experiment) Louviere et *al.* (2000). Cette méthode consiste en un sondage portant sur les différents programmes proposés. Il est présenté à chaque individu enquêté 6 choix. Chaque choix porte sur différentes combinaisons des 3 actions envisagées, associées à un coût, et ce afin d'estimer le consentement à payer des habitants de la Vallée du Commerce. Les principaux problèmes qui découlent de cette méthode sont l'élaboration du plan d'expérience (fractionnaire), la sélection des attributs et les niveaux de ces attributs (Hanley et *al.*, 1998). En effet les choix selon le genre, l'âge ou la catégorie socio-professionnelle, peuvent être différents (Ladenburg et *al.*, 2008).

Notre travail porte sur les biais liés à la mise en place d'un plan fractionnaire dans la méthode des programmes et sur la méthodologie à adopter. L'objectif est de développer une démarche prenant en compte les biais inhérents à cette méthode tels que les niveaux des différents attributs, le plan d'expériences utilisé et le problème du statu quo (non pris en compte dans cette méthode). Le but de ce travail est de développer une démarche limitant les biais. Sur cette thématique, j'ai encadré Ouerdia Arkoun docteure de l'Université de Rouen en post-doctorat au Lamsad (Arkoun et *al.*, 2012). D'autres travaux portent sur la recherche des facteurs expliquant le choix des habitants de la Vallée du Commerce en utilisant le modèle logit polytomique à paramètres aléatoires (Crastes et *al.* 2014). Actuellement, mes travaux sont centrés sur les problèmes rencontrés lors d'enquêtes longitudinales (Taïbi et *al.*, 2014c) et, sur ce thème, j'ai encadré

Iryna Petrovska doctorante de l'Université Wulls (Warsaw University of Life Sciences) pour un séjour de cinq mois.

Nous avons conçu, en partenariat avec l'Université de Tamatave et la Région Haute Normandie, un modèle de développement durable, le Campus Paysan. Le Campus Paysan, espace de promotion de la ruralité, est un projet de développement agricole initié en 2005 dans la Province de Tamatave, par l'Université de Tamatave, l'Esitpa et la Région Haute-Normandie. Toujours dans l'optique de construction d'indices, j'ai établi un modèle du rendement en riz ou encore un indice de richesse conçu à partir de données recueillies auprès de paysans malgaches dans la province de Tamatave (Taïbi et *al.* 2006, Taïbi et *al.* 2010 ; Taïbi et Bezara 2011).

La pression démographique galopante pousse les agriculteurs à diminuer les temps de jachère, dans un système de culture sur brûlis qui provoque des pertes de rendements non négligeables. Celles-ci peuvent être évitées au moyen de pratiques culturales adaptées comme, par exemple, le SRI (Système de Riziculture Intensive).

Afin d'expliquer les raisons de la dégradation subie par la province de Tamatave, des vagues d'enquêtes ont été lancées afin de créer un Observatoire de la Ruralité dans la Région d'Atsinanana. Les objectifs de cet observatoire sont de décrire les spécificités de la population rurale, d'identifier des indicateurs technico-économiques et de proposer un modèle de développement durable dans une des 22 régions malgaches, la Région d'Atsinanana. Au total, 1100 questionnaires ont pu être exploités, les paysans ayant été sélectionnés en utilisant la méthode de sondage stratifié. Les thèmes retenus sont la situation globale de l'exploitation, les critères de productivité rizicole, les modes de commercialisation, la trajectoire de l'exploitation, les attentes en terme de développement agricole et le niveau de richesse. Pour établir la typologie des exploitations, les méthodes de classement ont été mises en œuvre et des outils d'aide à la décision ont été créés. Cet observatoire a pour objectif à terme de mesurer l'impact du projet Campus Paysan sur le dévelopement de la province de Tamatave.

Dans le cadre d'un autre projet régional de recherche, le projet ET LIN(Éléments Trace Métalliques dans le lin) en partenariat avec l'institut technique du Lin en Normandie et l'Esitpa, j'ai accueilli et encadré Sorin Muntean de l'Université de Cluj Napoca (Roumanie), en Post-Doc au Lamsad. Nous avons établi un modèle de transfert des éléments traces métalliques (ETM) du sol vers la plante, en l'occurrence sur deux variétés de lin, le lin d'hiver et le lin de printemps (Muntean et *al.*, 2007, Taïbi et Lemaire *al.*, 2005). La démarche que nous avons adoptée peut être reproduite pour d'autre plantes et d'autres contextes de pollution, qui peut se transmettre à travers les parois de la racine vers la tige, la feuille, la fleur ou la graine.

Depuis 2005, je collabore avec des chercheurs de l'INRA de Versailles (Unité PESSAC). Une partie de ces travaux de collaboration ont porté sur le suivi d'une année d'herbicides (glyphosate, diuron et atrazine) dégradées et présentes dans les boues d'épuration, avec un échantillonnage pour trois traitements d'eaux usées. Le fait est que ces boues d'épuration peuvent produire indirectement une pollution des sols. Il est donc essentiel de définir la nature de ces polluants et de les quantifier (Ghanem et *al.* 2007).

Dans le domaine médical, l'équipe de recherche en urologie du Centre Hospitalier Universitaire de Rouen a mené une enquête prospective, en vue d'évaluer l'impact d'un appareil permettant de diminuer les problèmes d'incontinence. J'ai participé aux traitements statistiques et à l'analyse des résultats de cette enquête prospective (Berthier et *al.* 2008).

J'ai mené des travaux relevant d'un domaine spécialisé de l'agriculture, l'agriculture de précision qui consiste à utiliser les systèmes embarqués en vue d'optimiser les rendements agricoles en tenant compte de l'hétérogénéité intra-parcellaires (Duval et *al.* 2007a ; Duval et *al.*, 2007b).

En résumé, la nécessité d'élaborer des indices, tels que le taux de fiabilité, la fonction de survie, le consentement à payer, l'indice de richesse ou de pauvreté, l'état d'un sol, le taux de transfert de métaux lourds, m'a amenée à construire des modèles explicatifs et prédictifs, que cela soit une construction pas à pas à partir de données expérimentales ou que cela soit une construction d'estimateurs non paramétriques. J'ai élaboré ces démarches de modélisation en en tenant compte des contraintes fréquemment rencontrées que sont l'hétérogénéité des données, le problème des grandes dimensions, l'autocorrélation des variables, la présence de données incomplètes ou censurées à droite ou à gauche ou de données tronquées, de données aberrantes ou de données manquantes. Les perspectives de recherche sont encore nombreuses notamment à travers le programme de développement des systèmes complexes : Unitwin (Unesco) qui vient de démarrer et qui rassemble les chercheurs de très nombreux laboratoires internationaux (Taïbi et *al.*, 2014a).

1.2 Production scientifique

Articles dans des revues internationales et nationales avec comité de lecture

1. Taïbi-Hassani S., Laroutis D., Adigaw-E-Touck S., 2015. Pointwise Convergence of a nonparametric estimator of regression in a measurable space used in Contingent Valuation Method. *Journal of Mathematics and System Science*, **5**, 188-195.
2. Taïbi-Hassani S., Lepelletier P., Blot A., Thoisy-Dur J-C., 2015. A statistical approach to the evaluation and modelling of contamination in an agro-ecosystem. *International Journal of Ecology Economics and Statistics (IJEES)*, **36**, (1),83-97.
3. Dantan J., Pollet Y., Taîbi S., 2015. Combination of Imperfect Data in Fuzzy and Probabilistic Extension Classes, *Journal of Environmental Accounting and Management*,**3**, (2), 123-150.
4. Taïbi-Hassani S., Adigaw E-Touck S.,2015. A direct approach of nonparametric estimation of the hazard rate with left censored data. Soumis
5. Taïbi S., Petrovska I., Laroutis D., 2014. Status Quo and willingness to pay for reduction of risk of erosive runoff. *Scientific Journal Warsaw University of Life Sciences : Problems of World Agriculture*, **14** (XXIX) n°4,173-177.
6. Crastes R., Beaumais O., Arkoun O., Laroutis D., Mahieu P.A, Rulleau B., Hassani-Taïbi S. Barbu V., Gaillard D., 2014. Erosive runoff events in the European Union : using discrete choice experiment to assess the benefits of integrated management policies when preferences are heterogeneous.*Ecological Economics*,**102**, 105-112.
7. Taïbi-Hassani S., Thoisy-Dur J-C, Lepelletier P., Bodin J., Bennegadi-Laurent N., Bessoule J-J., Bispo A., Bodilis J., Chaussod R., Cheviron N., Cortet J., Criquet S., Dantan J., Dequiedt S., Faure O., Gangneux C., Harris-Hellal J., Hedde M., Hitmi A., Le Guedard M,, Legras M., Pérès G., Repinçay C., Rougé L. , Ruiz N., Trinsoutrot-Gattin I. ,Villenave C., 2013. Démarche statistique pour la sélection des indicateurs par Random Forests pour la surveillance de la qualité des sols, *Etude et Gestion des Sols*, **20**(2), 127-135.
8. Laroutis D., Taïbi-Hassani S., 2011. Discriminant Analysis Versus Random Forests on Qualitative Data : Contingent Valuation Method Applied to the Seine Estuary Wetlands. *International Journal of Ecological Economics & Statistics* ,**20** (11), 1-13.
9. Berthier A., Sentilhes L., Taïbi S. , Loisel C. ,Philippe Grise , Marpeau L., 2008. Sexual function in women following the transvaginal tension-free tape procedure for incontinence. *International Journal of Gynecology and Obstetrics*, **102**(2),105-109 .
10. Laval K., Mougin C., Akpa-Vinceslas M., Barray S., Dur J.C., Gangneux C., Lebrun J., Legras M., Lepelletier P., Plassart P., Taïbi S., Trinsoutrot-Gattin I.,2008. Nouvelles avancées vers la compréhension des données biologiques,*Etude et Gestion des Sols* ,**16**, 275-287.

11. Ghanem A., Bados P., Estaun R.A, Felippe L.de Alencastro, Taïbi S., Einhorn J., Mougin C.,2007. Concentrations and specific loads of glyphosate, diuron, atrazine, nonylphenol and metabolites thereof in French urban sewage sludge. *Chemosphere*, **69** , 1368-1373.
12. Muntean S., Legras M., Llorens J.M, GIRO F., Allaoui J., Taïbi S., 2007. Estimation of rates of uptake of trace elements from the soil to seeds of oilseed flax. *USAMV-CN,Bulletin of the University of Agricultural Sciences and Veterinary Medicine Cluj-Napoca*, **63** 337.
13. Taïbi-Hassani S.Youndjé E., 2003. Validation croisée pour l'estimateur lissé de la fonction de hasard : cas des données censurées. *Revue de Statistique Appliquée*, **LI**(I),73-86.
14. Taïbi-Hassani S., Youndjé E., 1997. Estimation lisse d'une fonction de hasard : Choix optimal de la fenêtre pour des observations censurées. *Comptes Rendus de l'Académie des Sciences de Paris*. Tome 324, Série I, 481-484.
15. Collomb G. Härdle W., Hassani S., 1987. A note on prediction via estimation of the conditionnal mode function, *Journal of Statistical Planning and Inference*, **15** ; 227-236.
16. Hassani S., Collomb G.,Sarda P. Vieu. P., 1986. Approche non paramétrique en théorie de la fiabilité : revue bibliographique, Revue de Statistique Appliquée, **35** (4) 27-41.
17. Collomb G., Hassani S., Sarda P., Vieu P.,1985. Estimation non paramétrique de la fonction de hasard pour des observations dépendantes., *Statistique et Analyse des Données*, **10** (13) ; 42-49. .
18. Collomb.G., Hassani S.,Vieu P.,Sarda P., 1985. Convergence uniforme d'estimateurs de la fonction de hasard pour des observations dépendantes : méthodes du noyau et des k-points les plus proches. *Comptes-Rendus de l'Académie des Sciences de Paris tome 301*, **série 1** (12), 653-656..
19. Antoch J., Collomb G., Hassani S., 1984. Robustness in parametric and non parametric regression estimation : An investigation by computer simulations. *COMPSTAT, Physica Verlag,Vienna* , 49-54.

Chapitre d'ouvrage

1. Taïbi S., Bezara M., 2011. La méthode des Forêts aléatoires appliquée à l'Observatoire de la ruralité à Tamatave. Pratiques et méthodes de sondage. Dunod, Collection Cours et Cas Pratiques, 382 p.

Conférences données à l'invitation du comité d'organisation dans des congrès ou colloques

1. Taïbi S., Petrovska I., Laroutis D., 2014. Status Quo and willingness to pay for reduction of risk of erosive runoff. 11th International Science Conference on Global Problems of Agriculture, forestry and food economy Warsaw.

2. Taïbi, S., Lepelletier, P., Dantan J., Thoisy-Dur, J.-C.,Bodin, J. A., 2014. Statistical approach for soil monitoring, risk assessment and soil characterization, e-Kickoff ICCSA'14 , Complex Systems Digital Campus, UNITWIN-UNESCO.
3. Taïbi, S., Rougé, L., Thoisy-Dur, J.-C., Bodin, J., Lepelletier, P., Dantan, J., Pérès, G., Grand, C. and Bispo, A., 2012. « Gestion et traitement des données du programme. Approche statistique de sélection d'Indicateurs et de biomarqueurs dans la surveillance de la qualité des sols et l'évaluation des risques. Journées Techniques Nationales, Bioindicateurs pour la caractérisation des sols, ADEME, Paris, 10 p.
4. Taïbi S., 2007. Statistical modelling and sustainable development. Ells University SGGW Varsow.Conference Erasmus Mundus Warsaw Poland.
5. Taïbi S., Problèmes d'estimation et de prédiction non paramétriques sous censure aléatoire à droite, Université de Dijon Juin 2004.

Communications avec comité de lecture dans des congrès internationaux

1. Dantan, J., Pollet, Y., Taïbi, S. 2015. A formal model to compute uncertain continuous data. In proceedings of CCS 2015 (international Conference on Complex Systems) - CS-DC'15 World e-conference (Complex Systems Digital Campus) UNITWIN/UNESCO. September 28 - October 2, 2015.
2. Dantan, J., Pollet, Y., Taïbi, S. 2015. A systemic meta-model for socio-environmental systems. In proceedings of the Sixth International Conference on Complex Systems Design Management, CSDM 2015. Editors : Auvray, G., Bocquet, J.-C., Bonjour, E., Krob, D. (Eds.). November 23-25, 2015. P. 307. Paris, France
3. Pauget B., Rougé L., Bispo A., Grand C., Beguiristain T., Bessoule J.-J., Bodilis J., Chaussod R., Cheviron N., Coeurdassier M., CortetJ., Criquet S., Dequiedt S., Faure O., Gangneux C., Gattin I., le Guedard M., Hitmi A., Laurent N., Legras M., Nélieu S., Ruiz N., Taïbi S., Vandenbulcke, F., de Vaufleury, A., Villenave C. and Pérès G. 2015. « Soil bioindicators : how soil properties influence their responses and how to select them in function of the site issues ? ». SETAC Europe 25th Annual Meeting. 3-7 May, Barcelona, Spain.
4. Dantan J., Pollet Y., Taïbi S. 2014. Taking account of uncertain, imprecise and incomplete data in sustainability assessments in agriculture. In proceedings of Computational Science and Its Applications - ICCSA 2014 - 14th International Conference, Part III, Lecture Notes in Computer Science LNCS 8581, ISBN 978-3-319-09149-5, pp. 625639. Springer International Publishing Switzerland. Guimarães, Portugal, June 30 - July 3, 2014. Communications orales*et al.* ICCSA-CLASS 2014.pdf
5. Dantan J., Pollet Y., Taïbi S. 2014. A goal-oriented meta-model for scientific research. In proceedings of Computational Science and Its Applications - ICCSA 2014 - 14th International Conference, Part V, Lecture Notes in Computer Science LNCS 8583, ISBN 978-3-319-09155-6, pp. 762774. Springer International Publishing Switzerland. Guimarães, Portugal, June 30 - July 3, 2014. ICCSA-AEIDSS 2014.pdf

6. Pauget B., Rougé L., Bispo A., Grand C., Beguiristain T., Bessoule J.-J., Bodilis J., Chaussod R., Cheviron N., Coeurdassier M., Cortet J., Criquet S., Dequiedt S., Faure O., Gangneux C., Gattin I., le Guedard M., Hitmi A., Laurent N., Legras M., Nélieu S., Ruiz N., Taïbi S., Vandenbulcke F., de Vaufleury A., Villenave C., Cluzeau D., Pérès G., 2014. Soil bioindicators : how soil properties influence their responses and how to select them in function of the site issues ? 1er GSBI. 3-5 Décembre 2014, Dijon, France.
7. Pérès G., Pauget B., De Vaufleury A., Coeurdassier M., Leguedard M., Bessoule J.J., Dequiedt S., Chaussod R., Ranjard L., Cluzeau D., Guernion M., Rougé L., Hedde M., Cheviron N., Dur J.C., Nelieu S., Mougin C., Gattin I., Gangneux C., Laurent N., Legras M., Laval K., Lepelletier P., Taïbi S., Villenave C., Faure O., Hellal J., Cortet J., Beguiristain T., Leyval C., Bodilis J., Criquet S., Hitm A., Ruiz N., Vandenbulcke F., Grand C., Galsomies L., Bispo A. 2014. Which bioindicators are suitable for soil quality monitoring and risk assessment ? From relevance study to transfer tool development. 1er GSBI. 3-5 Décembre 2014, Dijon, France.
8. Taïbi S., Thoisy-Dur J.C., Bodin J., Rougé L., Dantan J., Lepelletier P., Michaud A., Houot S., Pérès G., Bispo A., 2013.A statistical approach to assess soil biodiversity and biological activity responses to repeated organic amendment applications in cultivated soils - Relationships with soil functions. *RAMIRAN , 15th international conference*, Versailles, France.
9. Dantan J., Pollet Y., Taïbi S. 2013. The G.O.A.L. Approach. In proceedings of ENASE International Conference on Evaluation of Novel Approaches to Software Engineering. ,173-180. Angers, France, July 4-6, 2013.
10. Pérès G., Bispo, A., Grand C., Cluzeau D., Gattin I., Hedde M., Cheviron N., Harris-Hellal J., LeGuedard M., Bessoule J.J., RuizN., Pauget B., de Vaufleury A., Beguiristain T., Dequiedt S., Chaussod R., Faure. O., Hitmi A., Criquet S., Legras, M., Laurent N., Vandenbulcke F., Coeurdassier M., Ponton S., Cortet J., Villenave C., Bodillis J., Lepelletier P., Taïbi S., Dur J.-C., Bodin J. 2013. Application of soil bioindicators for risk assessment, monitoring and soil characterization in contaminated soils. Results from the French national "Bioindicators Programme". *12th International UFZ-Deltares Conference on Groundwater-Soil-Systems and Water Resource Management (AquaConSoil)*. Barcelona, Spain.
11. Crastes R., Beaumais, O., Arkoun, O., Laroutis, D., Mahieu, P.A., Rulleau, B., Hassani-Taïbi, S., Barbu, V.S., Gaillard, D., 2013. Erosive Runoff Events in the European Union : Using Discrete Choice Experiment to Assess the Benefits of Integrated Management Policies when Preferences are Heterogeneous. Workshop on non-market valuation. , Nantes, France..
12. Arkoun O., Barbu V., Crastes R, Laroutis D., Jia F., Taïbi-Hassani S., 2012. Sondage et plans fractionnaires appliqués à la méthode des programmes. 7ème Colloque Francophone sur les sondages. Rennes.
13. Thoisy-Dur J.-C., Lepelletier P., Taïbi S., Rougé L., Dantan J., Pérès G., Grand C. and Bispo A., 2012. Statistical approach to select soil bioindicators for soil monitoring, risk assessment and soil characterization ». Results from the French national Programme Bioindicators. 6th SETAC World Congress/SETAC Europe 22nd Annual Meeting, Berlin.

14. Bodin J., Dur J.-C., Rougé L., Dantan J., Lepelletier P., Grand C., Pérès G., Bispo A. Taïbi S., 2013. Soil bioindicators to assess soil biodiversity and activity responses to land-use practices. Final results of the research project Bioindicators ». International Interdisciplinary Conference on Land Use and Water Quality : Reducing Effects of Agriculture, The Hague, Netherland.

15. Gaillard D, Bonnet E., Bensaïd A., Arkoun O., Barbu V., Beaumais O., Crastes R., Laroutis D., Mahieu P.A., Rulleau B., Taïbi S., 2012. Analyse spatialisée de la perception du risque de ruissellement érosif. Modélisation et consentement à payer. Intérêt et apports d'une approche pluridisciplinaire », 2ème séminaire international euro-méditerranée sur l'Aménagement du Territoire la Gestion des risques et la Sécurité civile, Algérie.

16. Pérès G., Bispo A., Grand C., Gattin I., Hedde M., Harris-Hellal J., Leguedard M., Ruiz N., Alaphilippe A., Beguiristain T., Douay F., Faure O., Hitmi A., Houot S., Legras M., Guernion M., Vian J.F., Conil S., Rougé L., Lepelletier P., Taïbi S., Dur J.C., Cluseau D., 2012. Soil bioindicators for soil monitoring, risk assessment and soil characterization. Results from the French national "Bioindicators Programme". , 4th EUROSOIL , Bari, Italy.

17. Dantan J., Pollet Y., Taïbi S., 2012. Semantic indexation of Web services for collaborative expert activities. In proceedings of IADIS International Conference Information Systems. March 10-12, , 57-64, Berlin, Germany.

18. Dantan J., Pollet Y., Taïbi S., 2012. A KDD Process to retrieve and aggregate data from relational databases. *In proceedings of IADIS International Conference Information Systems.*, 443-445, Berlin, Germany.

19. Pérès G., Grand C., GattinI., Hedde M., Harris-Hellal J., Leguedard M., Ruiz N., Alalaphilippe A., Beguiristain T., Pruvot,C., FaureO., Hitmi A., Houot S., Legras M., Guernion M., Vian J.F., Conil S., Rougé L., Taïbi, S., Cluzeau , D., 2011. A national research programme to validate a battery of soil bioindicators for impact and risk assessment in urban soils. SUITMA 6, Marrakech, Morocco.

20. Taïbi S., Adigaw-E-Touck S.,2011. Validation croisée pour un estimateur lisse de la fonction de hasard sous données censurées à gauche 44èmes Journées de Statistique de la SFDS. Gammarth, Tunisie.

21. Taïbi S., Lepelletier P., G Perez, Rougé L., Dur J-C, Bispo A., 2011. Démarche en vue d'élaborer un indice d'état du sol. 44èmes Journées de Statistique de la SFDS. Gammarth-Tunisie.

22. Pérès G., Ruiz N., Hedde M. , Le Guedard M., Gattin I., d'Hugues P., Beguiristain T., Douay F. , Houot S. , Vian J.F., Faure O., Hitmi A., Alalaphilippe A., Dubs F., Rougé L., Taïbi S., Bispo A. , Grand C., Galsomies L. , Cluzeau D., 2011. Development and relevance assessment of bioindicators for soil monitoring, characterization and risk assessment. Example of a Bioindicator Program developed at National scale . *EJSB* , France.

23. Laroutis D., Taïbi S., 2010. Discriminant Analysis versus random forests on qualitative data : Contingent Valuation Method applied to the Seine estuary wetlands. 44th Annual Conference of the Canadian Economics Association, Quebec Canada.

24. Taïbi S., Bezara M., Lepelletier P., Nodjirim D., 2010. Mise en place de l'Observatoire de la ruralité dans le cadre du projet Campus Paysan à Madagascar, 6ème Colloque Francophone sur les Sondages, Tanger, Maroc.
25. Taïbi S, Lepelletier P., 2010. L'Observatoire des Jeunes Diplômés en Agriculture. 6ème Colloque Francophone sur les Sondages, Tanger, Maroc.
26. Taïbi S., Laroutis, D., 2009. Discriminant analysis versus random forests on qualitative data : Contingent Valuation Method applied to the Seine estuary wetlands. Applied Statistics International Conference, Ribno Slovenia.
27. Adigaw Touk S., Laroutis D, Taïbi S., 2009. Estimation non paramétrique de la régression : cas du consentement à payer. Journées d'études en statistique , Bordeaux.
28. Laroutis D. Taïbi S.,2009. Analyse discriminante versus forêts aléatoires : méthode d'évaluation contingente appliquée à l'estuaire de la Seine. Journées d'études en statistique , Bordeaux.
29. Taïbi S., Rouen C., Bezara M., 2008 Modélisation du rendement du riz à partir de données longitudinales. Congrès SFDS SSC, Montréal, Canada.
30. Taïbi S., Lambert A., Lepelletier P., Laval K., Mougin C., 2008. Elaboration d'un indice de la qualité des sols. Congrès Statistical Society of Canada (SSC) et Société Française de Statistique Montréal, Canada.
31. Mougin C. Dur J-C , Ridreau C., Huard E., Taïbi S., Tessier D., 2008. High levels of enzymatic activities are measured in soils managed under no-tillage leading to acidification and increased bioavailability of toxic metals. SETAC Europe Varsaw .
32. Dur J.C., Legras M., Gangneux C., Gattin I., Bailleul C., Akpa M., Plassart P., Barray S., Taïbi S., Massignam A., Pandolfo C., Lebrun J., Hedde M., Mougin C., Laval K., 2007. Interest in the Development on one Risk Indicators in Soil Ecotoxicology. Soil and Wetland Ecotoxicology, SOWETOX Barcelona.
33. Duval C., Debandt V, Eveillé J-P, Mahieu D., Lepelletier P., Taïbi S., LLorens J.M., 2007a.Influence de l'hétérogénéité pédoclimatique en Haute-Normandie sur la variabilité intraparcellaire des rendements en blé et en colza,Journées d'études en statistique, Angers.
34. Duval C., Debandt V., Eveillé J-P., Mahieu D., Taïbi S., LLorens J-M., 2007b. Influence of the pedo-climatic variability in Haute Normandie (NW France) on the intra field spatial variability on yields of wheat and oilseed rape.6th European Conference on Precision Agriculture and the 3rd Euro pean Conference on Precision Livestock Farming Skiathos, Greece, 87-94 .
35. Mougin C., Laval K., Legras M., Barray S., Taïbi S., Tessier, D., 2007. Chemical contamination versus non chemical stressors : the case study of agricultural soils. SETAC Europe 17th Annual Meeting, Porto, Portugal.
36. Muntean S., Legras M., Llorens J.M., Giro F., Allaoui J., Taïbi S., 2007. Estimation of rates of uptake of trace elements from the soil to seeds of oilseed flax, 6th Iinternational Symposium "Prospects for the 3rd millennium agriculture", University of Agricultural Sciences and Veterinary Medicine, Cluj-Napoca, Romania, 4-6 October 2007.

37. Legras M., Bailleul C., Tessier D.,Dur J.C,, Gangneux C., Taïbi S., Laval K., 2006. Effect of physicochemical characteristics of agricultural soils on fungal biomass. Impact of copper, EMEC 7, European Meeting on Environmental Chemistry, Brno, République Tchèque, 6-10 december 2006.
38. Legras M., Gangneux C., Tessier D., Dur J.C, Bailleul C., Taïbi S., Laval K. Effect of physicochemical characteristics of agricultural soils on fungal biomass, ISME-11, 11th International Symposium on Microbial Ecology, Vienna (Autriche), 20-25 august 2006.
39. Mougin C ; Laval, K. Taïbi S., Tessier, D. Lemaire A-S.and Barray S.,2005. Towards an index of biological state of the soil as a new tool for ecotoxicological studies, SETAC Europe 15th Annual Meeting, Lille.
40. Taïbi-Hassani S., Youndjé E., 1996. Estimation lisse d'une fonction de hasard : Choix optimal de la fenêtre pour des observations censurées. XVIIèmes Rencontres Franco-Belges de Statisticiens. Marne-La-Vallée, France.
41. Hassani S., 1984. Régression non paramétrique pour des variables aléatoires à valeurs dans un espace mesurable. Journées de Statistique. A.S.U. Montpellier.

Communications avec comité de lecture dans des congrès nationaux

1. Hedde M., Peres G., Villenave C., Gattin I,. Leguedard M., Harris-Hellal J., Dequiedt S., de Vaufleury A., Taïbi S., Grand C. Bispo A., 2014. Comment calculer les services écosystémiques rendus par les sols : un essai sur la base des données du programme « Bioindicateurs de qualité des sols » de l'ADEME. *Les 12èmes Journées d'Etudes sur les sols.*
2. Crastes R., BeaumaisO., LaroutisD., ArkounO., MahieuP.A., Rulleau B., Taïbi S., 2012. Valuing the reduction of risks provoked by erosive runoffs using choice experiment. *Journée d'étude en Econométrie Appliquée.* Le Havre, France.
3. Taïbi S., Dur J.C Lepelletier P., Rougé L., Dantan J., Bispo A, Grand, C., G Perez., Approche statistique de sélection d'Indicateurs et de Biomarqueurs dans la surveillance de la qualité des sols et l'évaluation des risques. Résultats du programme national ADEME, Bioindicateurs II, JES Versailles, Mars 2012.
4. Taïbi S., Lemaire A.S., 2005. Estimation du taux de transfert des éléments trace du sol vers les graines de lin oléagineux, Statistique des Processus. Angers, France.

Communications sans actes dans des congrès ou des colloques nationaux

1. Laroutis D., Taïbi S.,2009. Analyse discriminante versus Forêts Aléatoires sur des données qualitatives : Méthode d'évaluation contingente appliquée aux zones humides de l'estuaire de la Seine, *XLVIème Colloque de l'Association de Science Régionale De Langue Française (ASRDLF)*, Entre enjeux locaux de développement et globalisation de l'économie : quels équilibres pour les espaces régionaux ?, Clermont-Ferrand, France 6-8 juillet .
2. Barray S., Laval K., Legras M., Mougin C., Taïbi S., Tessier D.,2006. Elaboration et validation d'un indice d'état biologique des sols, *3ème Séminaire d'Ecotoxicologie de l'INRA*, Dinard, France .

3. Taïbi S., 2004. Statistique et analyse sensorielle. Évaluation de la performance des juges dans le cadre d'une épreuve de profils sensoriels. Université de Rouen Janvier .
4. Taïbi-Hassani S.,2004. Statistique et analyse sensorielle. Évaluation de la performance des juges dans le cadre d'une épreuve de profils sensoriels. Séminaire du laboratoire LMRS, Université de Rouen.

Ouvrages de vulgarisation

1. Taïbi S., Bezara M., « Quand mathématiques riment avec développement durable ». Conférence "30 minutes pour comprendre", Université de Rouen. Octobre 2010.
2. Taïbi, S., Les sciences de la vie mises en équation. Conférence Fête de la Science, 17-23 novembre 2008, Rouen, France.

Projets de recherche

1. Projet UNITWIN UNESCO depuis 2013. Membre référente pour l'Esitpa (partenaire).
2. Programme National ADEME - Bioindicateurs II - Bioindicateurs de la qualité des Sols-Coordination et animation du groupe Biomath «Gestion et traitement des données du programme. Approche statistique de sélection d'indicateurs et de biomarqueurs dans la surveillance de la qualité des sols et l'évaluation des risques» 2009-2013 (coordination).
3. Projet « Identification des freins et leviers de l'agriculture intégrée » - Chambre d'agriculture Régionale de Normandie, Oct 2011/ Octobre 2013(partenariat).
4. Projet EMIRE I et II - Impacts sur le ruissellement érosif dans la Vallée du Commerce 2010-2013 (coordination).
5. Programme Ademe Bioindicateurs I 2005-2008 (collaboration).
6. Projet Campus Paysan 2005-2010 (coordination).
7. Projet Agriculture de Précision 2006-2007 (participation).
8. Projet ET LIN 2003-2005, (collaboration).

Responsabilités et animations

1. Responsabilité de l'équipe du Lamsad Laboratoire de modélisation statistique, 2002-2013 Esitpa
2. Animation du groupe Biomath,2010-2013 Projet Bioindicatieurs II

Encadrement

-Encadrement de post-doctorants

1. Jeanne Bodin (Février 2012-Mars 2013), Docteur en Ecologie de l'Université de Nancy et Berlin .Projet Bioindicateurs de la qualité des sols(Bodin et *al.* 2013, Taïbi et *al.* 2012).
2. Ouerdia Arkoun (2011-2012) docteure en mathématiques de l'Université de Rouen . Projet EMIRE (Arkoun et *al.* 2012, Crastes et *al.* 2014).
3. Manassé Bezara(2005-2006, 2009) Enseignant-Chercheur , Université de Tamatave. Projet Campus Paysan (Taïbi et *al.* 2012, Bezara et *al.* 2010,Taïbi et 2009, Taïbi et *al.* 2010, Taïbi et *al.* 2007).
4. Sorin Muntean (2004-2005), Docteur 3ème cycle, Université des Sciences Agricoles et Médecine Vétérinaire de Cluj-Napoca (Roumanie) et Assistant-Professeur à la Chaire de Phytotechnie. Projet ETlin (Muntean et *al.* 2006).

-Encadrement de doctorants

1. Saturnin Adigaw-E-Touck, Thèse soutenue le 11 Janvier 2013 intitulée "Modèles non paramétriques de survie pour données incomplètes" , Université de Rouen.
2. Iryna Petrovska, doctorante de l'Université SGGW (Warsaw University of life Sciences), a effectué un séjour de recherche à l'Esitpa (Septembre 2013-Février 2014). Ses travaux ont porté sur les facteurs du statu quo (Taïbi et *al.* 2014).
3. Jérôme Dantan, doctorant CNAM-Esitpa, directeur de thèse Yann Pollet. Le travail a pour thème "une approche systémique unifiée pour
   l'optimisation durable des systèmes socio-environnementaux". Soutenance en 2016.
4. Assia Ayache, doctorante de l'Université de Constantine, bénéficie de séjours de recherche financés par son université. En collaboration avec son directeur de thèse, Fouad Rahmani, responsable de l'ecole doctorale de mathématiques de l'Université de Constatine, j'assure un suivi à distance. Sa thématique de recherche porte sur les réseaux de neurones, et plus précisément sur les méthodes supervisées et non supervisées dans le cas de données bruitées (2013).

-Encadrement d'ingénieurs d'étude et de recherche

Emmanuelle Nieullet, 2006-2007 Ingénieur en Agriculture et titulaire d'un Master de l'Université de Montpellier. Projet Campus Paysan à Madagascar (Taïbi et *al.* 2007).

- Encadrement de stagiaires

1. 2012 Jia Fan, INSA de Rouen / LMRS Projet de Fin d'études 5ème Année Génie Mathématiques/AIMAF2
2. 2011 Christophe Marborough,Stage Technicien Insa de Rouen,

3. 2010 Pierre Parent, P.,INSA-LAMSAD.
4. 2010 Théophile Chaumont-Frelet, Insa de Rouen, 3ème Année Génie Mathématiques,
5. 2009 Halima Chtioui, Université de Bourgogne Master 2 MIGS,
6. 2008 Sébastien Bellet,3ème Année Génie Mathématiques Insa de Rouen,
7. 2007-2008 Aurore Lambert, Insa de Rouen Projet de fin d'études Génie Mathématiques,
8. 2007 Arles Fanampindrainy, LAMSAD- Université de Tamatave,Maîtrise de gestion,
9. 2007 Valentin Vlaaz, Université Université de Galati (Roumanie) Master 2,
10. 2007 Valentina Contantinescu, Université de Galati (Roumanie) Master 2,
11. 2006 Candice Rouen, Insa de Rouen, Génie Mathématiques, Stage Ingénieur,
12. 2005 Jawad Alaoui, Insa de Rouen Projet de Fin d'études, Stage Ingénieur,
13. 2004 Mounir Lafkahi, Université de Caen Master 2,
14. 2004 Youssef Kacimi, Université de Caen Master 2,
15. 2003 Mélanie Frémont, Insa de Rouen Projet de Fin d'études Ingénieur,
16. 2002 Valérie Chauvenssy, Université de Rouen. Master 1
17. 2002 Delphine Grancher, Université de Rouen. Master 1

Rapporteur pour des revues

1. Journal of Applied Statistics.
2. Scientific Journal of the Warsaw University of Life Sciences.
3. Environmental Chemistry Letters.

Représentations

Membre du Conseil Scientifique - Esitpa, 2010-2013.

Rapports de recherche

1. Bodin J., Taïbi, S., Thoisy-Dur J.C.,Dantan J., Lepelletier, P., Rougé L., Bioindicateurs de la qualité des sols. Démarche d'analyse globale. Rapport d'activités. Ademe- Esitpa Fév. 2013.
2. Taïbi S., Lepelletier P., Dantan J., Bodin J., Thoisy-Dur J.C., Rougé L., 2012. Gestion et traitement des données du programme. Approche statistique de sélection d'Indicateurs et de biomarqueurs dans la surveillance de la qualité des sols et l'évaluation des risques. Rapport Final Ademe -Esitpa, 101 pp.
3. Taïbi, S., Rougé, L., Thoisy-Dur, J.C., Bodin, J., Lepelletier, P., Dantan J.2011. Gestion et traitement des données du programme. Approche statistique de sélection d'Indicateurs et de biomarqueurs dans la surveillance de la qualité des sols et l'évaluation des risques, Rapport Intermédiaire Ademe- Esitpa .

4. Taïbi S. Roche D. Rapport d'évaluation du projet « Campus Paysan ».- Région Haute Normandie - Université de Tamatave (Madagascar), Déc. Esitpa, 2009.
5. , Taïbi, S., Lepelletier P., Barray, S., Plassart, P., Brault, A., Dur, J.C., Huard, E., Lebrun, J., Mougin, C., Tessier, D., 2008. "Elaboration et validation d'un indice d'état biologique des sols", Rapport Final ADEME.
6. Barray S., Laval K., Lemaire A-S., Mougin C., Taïbi S.,2008. Rapport d'activité Programme Bioindicateurs I - Ademe.
7. Taïbi, S., Bezara M., Nieullet E., Nodjirim D., 2007. Mise en place du Projet Campus dans la province de Tamatave.
8. Taïbi, S., Bezara M., Nodjirim D.,2006 Mise en place d'un modèle de développement durable dans la province de Tamatave.

Rapports de stagiaires

1. Marborough, C., Bioindicateurs et indices de la qualité des sols INSA Génie Mathématiques- LAMSAD,2011.
2. Chaumont-Frelet T., Analyse statistique d'une base de données concernant l'efficience énergétique et économique des exploitations agricoles de polyculture-élevage de Haute-Normandie. INSA- LAMSAD, 2010.
3. Parent, P., Travaux préliminaires à l'établissement d'un indice de qualité des sols. Génie Mathématiques INSA -LAMSAD, 2010.
4. Chtioui H., Etude de la qualité d'indicateurs socio-économiques et biologiques à partir de méthodes de classement, Mémoire de Master 2, MIGS Uiversité de Dijon LAMSAD, 50pp., 2009.
5. Lambert, A., Mise en place d'un bio-indicateur de la qualité des sols. Projet de fin d'études,5ème année Génie Mathématiques INSA, LAMSAD, 2008.
6. Bellet S. 3ème annnée Génie Mathématiques INSA- LAMSAD, 2008.
7. Constantinescu, V.,. La mise en place d'un site web illustrant le projet -Campus Paysan de Madagascar,Lamsad Esitpa 2007.
8. Fanampindrainy, A., Constitution de la base de données en vue de la mise en place du Campus Paysan à Madagascar. LAMSAD- Université de Tamatave, 2007.
9. Vlaaz, V., Analyse des données issues de l'enquête Observatoire pour le projet Campus Paysan à Madagascar. Lamsad- 2007.
10. Rouen, C. Etude du rendement des exploitations, dans la province de Tamatave, dans le cadre du projet Campus Paysan. Projet de fin d'études Génie Mathématiques, 5ème année INSA-LAMSAD, 2006.
11. Alaoui J. , Transfert des métaux lourds dans le lin. Travaux de simulation. Projet de fin d'études Génie Mathématiques, 5ème année INSA-LAMSAD, 2005.
12. Kacimi Y.,Méthodes de discrimination dans le cadre d'une épreuves sensorielles, . Master 2 Université de Caen,2004.
13. Lafkahi M., Les séries chronologiques et données climatiques. Master 2 Université de Caen, 2004.

14. Frémont M. Statistique des procédés. Cartes de contrôles non paramétriques. Projet de fin d'études Génie Mathématiques, 5ème année INSA-LAMSAD, 2003.
15. Chauvenssy V., Analyse des données pluriannuelles recueillies auprès de l'observatoire de la qualité de l'air en Seine Maritime. Problèmes d'estimation et de prédiction. Mémoire Maîtrise d'Ingénierie Mathématique, Université de Rouen, 58pp., 2002.
16. Grancher D., Simulations en modélisation paramétrique et non paramétrique- Mémoire Maîtrise d'Ingénierie Mathématique Lamsad-Université de Rouen,2002.

Chapitre 2

Les problèmes d'estimation et de modélisation non paramétriques

2.1 Estimation non paramétrique de la fonction de hasard par la méthode du noyau et des K-points les plus proches ou KNN (K-Nearest Neighbors) pour des observations dépendantes

2.1.1 Estimation de la fonction de hasard

On s'intéresse à des modèles de durée de vie. Considérons l'exemple de la durée de vie d'une plante. En prenant l'origine des temps au moment de la germination, on cherche à évaluer la probabilité que la plante meure entre les temps t et $t+\Delta t$ sachant qu'elle était encore en vie au temps t (Δt est pris au sens physique du terme). On note X la variable aléatoire réelle positive "durée de vie d'une plante". La variable aléatoire X est à valeurs dans une partie E non vide et mesurable de $\mathbb{R}^p$ dont la loi de probabilité dans $\mathbb{R}^p$ admet F (resp. f) pour fonction de répartition (resp. densité). On suppose que la densité f est uniformément continue sur E. La fonction de hasard est notée λ et est définie par $\lambda = \frac{f}{1-F}$, pour tout $x \in E$ tel que $F(x) < 1$.

Le cas où $E = \mathbb{R}^+$ constitue un important champ d'applications (Hassani et *al.*, 1986). La fonction de hasard de X est aussi appelée taux de hasard, taux de défaillance, taux de mortalité, fonction d'intensité selon les applications envisagées.

On estime la fonction de hasard λ à partir d'une suite $X_i, i \in \mathbb{N}^*$, de variables aléatoires ayant même loi que X. On suppose que le processus $(X_n)_{n\in\ N^*}$ est uniformément fortement mélangeant ou φ-mixing en ce sens que (Billingsley 1968), pour tous entiers positifs i et j et tout événement A (resp. B) appartenant à la tribu engendrée par $(X_1, \ldots, X_i)$ (resp. $(X_{i+j}, X_{i+j+1}, \ldots)$) on a :

$$|P(A \cap B) - P(A)P(B)| \leq \phi_j P(A),$$

où $(\phi_n)_{\mathbb{N}}$ est une suite réelle positive de limite nulle. Dans toute la suite, on supposera sans restriction de généralité, que la suite $(\phi_n)_{\mathbb{N}}$ est décroissante.

On appelle i la fonction de $\mathbb{N}$ dans $\mathbb{N}$ associée à $\phi = (\phi_n)_{n\in\mathbb{N}}$ définie par

$$i_\phi(n) = \inf\left\{ j \leqslant n \ / \ \frac{\phi_j}{j} \leqslant \frac{1}{n} \right\}.$$

Plusieurs estimateurs de la fonction de hasard existent, voir à ce sujet la revue bibiographique (Hassani et al 1986). Nous retenons pour la suite l'estimateur à noyau et l'estimateur par la méthode des k-points les plus proches. En effet ces estimateurs présentent deux intérêts. Le premier est celui de pouvoir établir des résultats de convergence uniforme sur un compact, ce qui permet d'estimer le mode (si celui-ci est supposé unique). Le second pôle d'intérêt est que ces résultats peuvent être étendus au cas de processus mélangeants.

L'estimateur non paramétrique de λ par la méthode du noyau, noté λ_n, est défini par

$$\lambda_n(x) = \frac{f_n(x)}{1 - F_n(x)}, \forall x \in E,$$

et le second estimateur non paramétrique de λ par la méthode des K-NN, noté $\tilde{h}_n$, est donné par

$$\tilde{\lambda}_n(x) = \frac{\tilde{f}_n(x)}{1 - F_n(x)}, \ \forall x \in E$$

où

$$F_n(x) = n^{-1} \sum \mathbb{1}_{[X_i \leq x]}, \ \forall x \in E.$$

Soit K définissant un noyau de $\mathbb{R}^p$, c'est-à-dire une fonction réelle bornée de $L^1(\mathbb{R}^p)$ telle que

$$\int K(z)dz = 1 \quad \text{et} \quad |\mathrm{z}|\,\mathrm{K(z)} \underset{|\mathrm{z}|\to+\infty}{\to} 0.$$

Les deux estimateurs $f_n(x)$ et $\tilde{f}_n$ (Parzen (1962), Rosenblatt (1956) et Loftsgarden et Quesenberry (1965)), sont définis par

$$f_n(x) = (nh_p^n)^{-1} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right), \forall x \in E$$

où $(h_n)_{\mathbb{N}}$ est une suite réelle strictement positve de limite nulle, et

$$\tilde{f}_n(x) = \frac{1}{(nR(k_n, x)^p)} \sum_{i=1}^{n} K\left(\frac{x - X_i}{R(k_n, x)}\right), \ \forall x \in E$$

où $R(k_n, x)$ est la distance entre x et la k_n-ième observation la plus proche de x, soit

$$R(k_n, x) = \inf\{\mu \in \mathbb{R}^+ : \mathrm{card}\{X_i, i = 1, \ldots, n, \ |x - X_i| \leqslant \mu\} \geqslant k_n\},$$

la suite k_n étant entière, strictement positive et vérifiant $\frac{k_n}{n} \underset{n\to\infty}{\longrightarrow} 0$.

Hypothèses

C désigne un compact de E et $\tilde{C}$ un ϵ-voisinage de compact de C dans E avec $\tilde{C} = E$ si $C = E$. On suppose que f vérifie

$$\exists\, \Gamma < +\infty, f(x) \leqslant \Gamma, \forall x \in E,$$

$$\exists\, \gamma > 0, f(x) \geqslant \gamma, \forall x \in \tilde{C},$$

et

$$\exists\, \tau > 0, F(x) \leqslant 1 - \tau, \forall x \in C.$$

Nous énonçons ci-après les principaux résultats de convergence concernant ces deux estimateurs.

On supposera dans toute la suite que le noyau K est lipschitzien.

Proposition 1.

Si la suite $(h_n)_{\mathbb{N}}$ vérifie, conjointement avec la suite $i_\phi(n)$ associée à ϕ_n, la condition

$$\frac{nh_n^p}{i_\phi(n)\ln(n)} \to \infty$$

alors $\sup_{x\in C} |\lambda_n(x) - \lambda(x)|$ converge presque complètement vers 0 quand $n \to +\infty$.

Nous donnons un résultat analogue concernant l'estimateur par la méthode des k-points les plus proches $\tilde{\lambda}_n$.

Proposition 2.

On suppose que le noyau K vérifie les deux hypothèses suivantes :

$$K(cz) \geqslant K(z),\ \forall c \in [0,1], \forall z \in \mathbb{R}^p$$

et

$$K(z) = 0,\ \forall z \in \mathbb{R}^p,\ |z| > 1.$$

Si la suite k_n vérifie, conjointement avec la suite $i_\phi(n)$, la condition

$$\frac{k_n}{i_\phi(n)\ln(n)} \underset{n\to\infty}{\longrightarrow} +\infty$$

alors $\sup_{x\in C} \left|\tilde{\lambda}_n(x) - \lambda(x)\right|$ converge presque complètement vers 0 quand $n \to +\infty$.

Les démonstrations de ces deux propositions utilisent une extension de l'inégalité de Bernstein au cas de variables ϕ-mélangeantes (Collomb, 1984), des résultats classiques de Bochner relatifs aux propriétés des noyaux, une conséquence de Trèves du lemme d'Urysohn et un résultat général de Moore et Yackell (1977) sur la méthode des k-points les plus proches.

2.1.2 Estimation non paramétrique du mode de la fonction de hasard

Supposons que le mode θ de la fonction λ défini par

$$\theta = \arg\max_{x \in C} \lambda(x)$$

soit unique.

On construit à partir des deux estimateurs λ_n et $\tilde{\lambda}_n$ deux estimateurs de θ, notés θ_n et $\tilde{\theta}_n$, avec

$$\theta_n = \arg\max \lambda_{x \in Cn}$$

et

$$\tilde{\theta}_n = \arg\max_{x \in C} \tilde{\lambda}_n.$$

Proposition 3. Sous les hypothèses de la proposition 1, l'estimateur θ_n converge presque complètement vers θ quand n tend vers l'infini.

Proposition 4. Sous les hypothèses de la proposition 2, l'estimateur $\tilde{\theta}_n$ converge presque complètement vers θ quand n tend vers l'infini.

En effet quand $\check{\lambda}_n - \lambda(x)$ converge vers 0 presque complètement, la continuité de la fonction h et l'unicité de θ sur C impliquent que, si

$$sup_{x \in C} \left| \check{\lambda}_n(x) - \lambda(x) \right|$$

converge presque complètement vers 0 quand $n \to +\infty$, alors $\check{\theta}_n$ converge presque complètement vers $\check{\theta}$ quand n tend vers l'infini, avec $\check{\theta}_n = \theta_n$ si $\check{h}_n = \lambda_n$ et $\check{\theta}_n = \tilde{\theta}_n$ si $\check{\lambda}_n = \tilde{\lambda}_n$.

2.2 Estimation non paramétrique de la fonction de hasard dans le cas des données censurées à droite

Pour une variable aléatoire X à valeurs dans $\mathbb{R}^p$ dont la loi de probabilité admet F pour fonction de répartition et f pour densité, on estime la fonction de hasard $h = \dfrac{f}{1-F}$ à partir d'une suite $X_i, i \in \mathbb{N}^*$, de variables aléatoires ayant même loi que X, lorsque le processus$(X_n), n \in N$ est uniformément fortement mélangeant. L'estimation de la fonction de hasard est un problème qui apparaît dans de nombreux modèles statistiques en analyse des durées de vie. On peut citer l'étude de la survie en médecine, les problèmes de fiabilité en industrie, le taux de réemploi en socio-économie, l'analyse de cycles de vie en agroalimentaire.

Plusieurs estimateurs de la fonction de hasard ont été établis. Nous renvoyons aux revues bibliographiques Eryer (1977), Tapia et Thompson (1973), Wertz (1978), Bean et Tsakas (1980), Deheuvels(1977), Hassani et *al.* (1986). La majorité des estimateurs de la fonction de hasard sont construits à partir d'estimateurs de la fonction de densité et de la fonction de survie. En analyse de survie, comme dans plusieurs autres domaines d'application en statistique, il est fréquent que la variable d'intérêt soit observée de façon incomplète : les données dont on dispose alors sont censuréees ou tronquées, et l'enjeu est de trouver les moyens de les traiter. Dans de nombreux domaines, notamment

en médecine, agriculture, chimie, environnement, sociologie, etc. la littérature traitant de l'estimation de la fonction de hasard dans le cas de données complètes, tronquées et censurées à droite est assez abondante. Citons les articles de Watson et Leadbetter (1964a), Blum et Susarla (1980), Singpurwalla et Wong (1983) pour les données complètes. Dans le cas de données censurées à droite, citons Tanner et Wong (1983), Ramlau-Hansen (1983) et Yandell (1983), Hassani et *al.* (1986), Collomb et *al.* (1986), Taïbi-Hassani et Youndjé (2003).

2.2.1 L'estimateur à noyau sous censure à droite de la fonction de hasard

1.- Introduction

Dans le cadre des données complètes, la plupart des estimateurs non-paramétriques de la fonction de hasard sont définis comme rapport d'un estimateur de la densité et d'un estimateur de la fonction de survie. La survie étant généralement estimée à l'aide de la fonction de répartion empirique, estimer la survie par le biais d'un estimateur à noyau de la fonction de répartition apporte au moins deux avantages (Youndjé et *al.* (1994)). En effet on obtient un estimateur continu, et l'estimation de la survie en termes d'erreur quadratique est améliorée. Dans le cadre des données censurées, les estimateurs de la fonction de hasard sont aussi en général des rapports d'estimateurs dont le dénominateur est un estimateur de type survie (Hassani et *al.* 1986). Comme dans le cas des données complètes, il est avantageux d'utiliser un estimateur à noyau au dénominateur ; nous allons considérer ici un tel estimateur (Taïbi-Hassani et Youndjé 2003).

Comme dans la plupart des problèmes d'estimation utilisant des estimateurs à noyau, la performance d'un estimateur à noyau de la fonction de hasard va dépendre fortement du choix de la largeur de fenêtre. Dans le cadre des données complètes, une procédure de sélection de la largeur de fenêtre a été proposée par Sarda et Vieu (1991) lorsque la fonction de survie est estimée par la fonction de répartion empirique. Ce résultat a été étendu au cas des données censurées par Patil (1993a) et étudié davantage dans Patil (1993b). Une règle de sélection de la fenêtre a été introduite par Youndjé et *al.* (1994) dans le cadre des données complètes lorsque la survie est estimée par un estimateur à noyau.

Nous proposons une méthode basée sur les idées de validation croisée pour calculer la fenêtre minimisant asymptotiquement l'erreur quadratique intégrée.

2.- Estimateur à noyau sous censure

Nous considérons un estimateur à noyau de la fonction de hasard dans le cas des données censurées à droite. Nous donnons la décomposition asymptotique de l'erreur quadratique intégrée et proposons une méthode pour sélectionner la largeur de fenêtre asymptotiquement optimale.

Soit X^0 une variable aléatoire réelle représentant une durée de vie de densité f_0 et de fonction de répartion F_0. Soit C une variable aléatoire réelle de densité g et de fonction de répartition G représentant la censure ; on suppose que X^0 et C sont indépendantes. Soit $(X_i^0,\ C_i),\ i = 1, \ldots, n$ un n-échantillon i.i.d de $(X^0,\ C)$. Dans le modèle de durées

censurées à droite, les variables observables sont $(X_i,\ \Delta_i),\ i = 1, \ldots, n$ où

$$X_i = \min(X_i^0,\ C_i) \qquad \text{et} \qquad \Delta_i = I_{[X_i \leqslant C_i]}.$$

Les $(X_i, \Delta_i),\ i = 1, \ldots, n$ ainsi définis forment un n-échantillon i.i.d d'un couple aléatoire (X, Δ). X étant de densité f et de fonction de répartition F avec $1 - F = (1 - F_0)(1 - G)$. D'après Tanner et Wong (1983) on a la relation

$$E[\Delta \ / \ X = x]f(x) = f_0(x)(1 - G(x)). \tag{2.1}$$

On pose

$$\phi(x) = f_0(x)(1 - G(x)). \tag{2.2}$$

Nous nous proposons ici d'estimer le taux de hasard $\lambda = f_0/(1 - F_0)$, utilisant $(X_i, \Delta_i), i = 1, \ldots, n$ et, comme estimateur de cette fonction, nous retenons l'estimateur à noyau défini par

$$\widehat{\lambda}(x) = \frac{\phi_h(x)}{1 - F_h(x)} \tag{2.3}$$

où, compte tenu de (2.1) et (2.2) (voir Collomb (1976))

$$\phi_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)\Delta_i \tag{2.4}$$

est l'estimateur à noyau de ϕ et

$$F_h(x) = \frac{1}{n}\sum_{i=1}^{n} H\left(\frac{x - X_i}{h}\right) \tag{2.5}$$

est l'estimateur de F (voir par exemple Lejeune et Sarda (1992)). K un noyau et $H(x) = \int_{-\infty}^{x} K(t)dt$ et $h = h(n)$ est la largeur de fenêtre, ou paramètre de lissage ou de régularisation.

De nombreux estimateurs de la fonction de hasard ont été proposés lorsque les durées sont censurées ou non, voir entre autres Singpurwalla et Wong (1983), Hassani (1985), Hassani et *al.* (1986) et Droesbeke et *al.*(1989) pour des revues bibliographiques. L'estimateur $\widehat{\lambda}$ peut être vu comme une extension au cas des données censurées de l'estimateur introduit par Watson et Leadbetter (1964a et 1964b) et étudié davantage par Murthy (1965), Ahmad (1976), Youndjé et *al.* (1994).

3.- Validation croisée pour l'estimateur lissé de la fonction de hasard : cas des données censurées à droite

En estimation fonctionnelle, l'indicateur permettant d'analyser la performance d'un estimateur est l'erreur quadratique intégrée. Pour tout estimateur $\widehat{\lambda}$ cette quantité est définie par

$$ISE(\widehat{\lambda}, h) = \int (\widehat{\lambda}(x) - \lambda(x))^2 W(x)dx,$$

W étant une fonction de poids positive. Le lecteur trouvera dans Marron et Padgett (1987) les raisons pour lesquelles il est plus intéressant de considérer l'ISE plutôt que son espérance mathématique, appelée MISE, pour analyser la performance d'un estimateur.

Hypothèses

Dans toute la suite nous supposerons que :

– la fonction W est à support compact de support S_W et $\overset{\circ}{S_W} \neq \emptyset$; $(H.1)$

– la largeur de fenêtre h appartient à l'intervalle d'intérêt

$$H_n = \left[An^{-\frac{1}{5}-\varepsilon}, \quad Bn^{-\frac{1}{5}+\varepsilon}\right], \quad 0 < \varepsilon < \frac{1}{5}, \quad 0 < A < B < \infty \qquad (H.2)$$

– K est à support compact, lipschitzien, symétrique et $\int K = 1$; $(H.3)$

– Les fonctions G et f_0 sont lipschitziennes et de classe $\mathcal{C}^2$; $(H.4)$

– il existe $\gamma > 0$ tel que

$$\forall x \in S_W, \qquad f(x) \geq \gamma, \qquad F_0(x) < 1-\gamma \qquad \text{et} \qquad G_0(x) < 1-\gamma. \quad (H.5)$$

– La fonction f est bornée. $(H.6)$

Théorème 3.1 *Sous les hypothèses* $(H.1) - (H.6)$ *on a :*

$$\sup_{h \in H_n} \left| \frac{ISE(\widehat{\lambda}, h) - C_1(nh)^{-1} - C_2 h^4}{C_1(nh)^{-1} + C_2 h^4} \right| \longrightarrow 0 \qquad \text{p.s.} \qquad (3.1)$$

où

$$C_1 = \left[\int K^2\right] \int \frac{\phi(x)}{(1-F(x))^2} W(x)dx \quad \text{et} \quad C_2 = \int (B_\phi(x) + \lambda(x) B_F(x))^2 \frac{W(x)}{(1-F(x))^2} dx,$$

avec

$$B_\phi(x) = \frac{1}{2}\phi''(x)\left[\int u^2 K(u)du\right] \qquad \text{et} \qquad B_F(x) = \frac{1}{2}f'(x)\left[\int u^2 K(u)du\right]$$

Remarque 3.1 Le théorème 3.1 est une extension au cas des données censurées d'un résultat de Vieu (1991), mais ce résutat est énoncé sous α-mélangeance. Ce théorème est à rapprocher du théorème 3.1 de Marron et Padgett (1987) et du Théorème 3.1 de Patil (1993a). Il faut également remarquer que (2.2) et $(H.4)$ assurent l'existence de ϕ'' et de $F'' = f'$.

4.- Le critère de validation croisée

Le théorème 3.1 montre l'influence de la largeur de fenêtre h sur l'estimateur $\widehat{\lambda}$. Il découle de ce résultat qu'une largeur trop petite augmentera la composante proportionnelle à $(nh)^{-1}$ de l'ISE asymptotique et qu'une largeur trop grande augmentera la composante proportionnelle à h^4. Une méthode usuellement considérée pour choisir la largeur de fenêtre minimisant asymptotique l'ISE est la validation croisée. Le critère de validation considéré ici est motivé comme suit :

– L'ISE peut se décomposer comme

$$ISE(\widehat{\lambda}, h) = \int \widehat{\lambda}^2(x) W(x)dx - 2\int \left[\frac{\widehat{\lambda}(x)W(x)}{1-F_0(x)}\right] f_0(x)dx + \int \lambda^2(x)W(x)dx.$$

– Le troisième terme est indépendant de h, donc il suffit de choisir h minimisant les deux premiers termes.
– Un estimateur "presque" sans biais du second terme est

$$\frac{2}{n}\sum_{i=1}^{n}\frac{\widehat{\lambda}^{-i}(X_i)W(X_i)\Delta_i}{(1-F_n(X_i)} \qquad \text{avec} \qquad \widehat{\lambda}^{-i}(x)=\frac{\phi_h^{-i}(x)}{1-F_h^{-i}(x)}, \tag{4.1}$$

$$\phi_h^{-i}(x)=\frac{1}{(n-1)h}\sum_{\substack{j=1\\ j\neq i}}^{n}K\left(\frac{x-X_j}{h}\right)\Delta_j, \qquad F_h^{-i}(x)=\frac{1}{(n-1)}\sum_{\substack{j=1\\ j\neq i}}^{n}H\left(\frac{x-X_j}{h}\right) \tag{4.2}$$

et F_n représente la fonction de répartition empirique définie par

$$F_n(x)=\frac{1}{n}\sum_{i=1}^{n}I_{[X_i\leq x]}. \tag{4.3}$$

Soit $\widehat{h}$ le minimisateur sur H_n du critère de validation

$$CV(h)=\int\widehat{\lambda}^2(x)W(x)dx-\frac{2}{n}\sum_{i=1}^{n}\frac{\phi_h^{-i}(X_i)\Delta_i W(X_i)}{(1-F_h^{-i}(X_i)(1-F_n(X_i))}. \tag{4.4}$$

Le résultat immédiatement ci-dessous montre que $\widehat{h}$ minimise asymptotiquement l'ISE.

Théorème 4.1 *Sous les hypothèses* $(H.1)-(H.6)$ *on a :*

$$\frac{ISE(\widehat{\lambda},\widehat{h})}{\inf\limits_{h\in H_n}ISE(\widehat{\lambda},h)}\longrightarrow 1 \qquad \text{p.s.} \tag{4.5}$$

Remarque 4.1 Ce résultat est une extension au cas des données censurées du théorème 2 de Youndjé et *al.* (1994) et est à rapprocher du théorème 4.1 (resp. 3.2) de Marron et Padgett (1987)(resp. Patil (1993a)).

2.3 Méthode directe pour l'estimation non paramétrique du taux de hasard sous données censurées à gauche

La censure aléatoire à gauche survient lorsque le temps d'origine de la durée de vie précède celui de l'étude. En analyse de survie, comme dans plusieurs autres domaines notamment en médecine, agriculture, chimie, environnement, sociologie, etc... il est fréquent que la variable d'intérêt soit observée de façon incomplète et plus particulièrement censurées à gauche. Peu de références existent sur ce sujet. En effet ce schéma de censure est vu comme une inversion du cas de données censurées à droite, appelée la méthode d'inversion du temps.

Soit X_0 une variable aléatoire d'intêret représentant la durée de vie, de densité f_0 et de fonction de répartition F_0. Soit C une variable aléatoire réelle de densité g et de fonction de répartition G modélisant la censure.

Nous supposons que X_0 et C sont indépendantes. Soient $X = \max(X_0; C)$ et $\Delta = I_{[C \leqslant X_0]}$ où $I_{[.]}$ est la fonction indicatrice. Les fonctions f et F désignent respectivement la densité et la fonction de répartition de X. Dans les modèles des durées de vie censurées à gauche, c'est en réalité un échantillon du couple $(X; \Delta)$ qui est observé. Soient donc $(X_i; \Delta_i)$ un n-échantillon du couple $(X; \Delta)$. Nous nous proposons ici, sur la base de cet échantillon, d'estimer le taux de fiabilité ou taux de survie. Citons les papiers de Gurler et Wang(1993), où le taux de hasard est estimé pour des données tronquées à gauche, Uzunogullari et Wang (1992), Sun (1997) pour le cas des données doublement incomplètes, en l'occurence des données à la fois censurées à droite et tronquées à gauche.

Le nombre de travaux consacrés à la problématique de la censure à gauche est beaucoup moins abondant que pour la censure à droite. Pour traiter cette question qui n'est simplement que le symétrique de celui de la censure à droite, un certain nombre d'approches a été proposé. La technique la plus employée, assez intuitive consiste en une transformation des données censurées à gauche en données censurées à droite en les multipliant par -1.

Dans cet ordre d'idées, Ware et Demets (1976) et Csörgo et Horvath (1980) proposent d'estimer la fonction de survie $S = 1 - F$ par l'estimateur $S(x) = 1 - \hat{S}_{KM}(-x)$ où $\hat{S}_{KM}$ désigne l'estimateur de Kaplan-Meier qui, rappelons-le, estime la fonction de survie sous données censurées à droite. Cette approche indirecte présente quelques inconvénients. En effet elle exclut les variables à support positif, ce qui dans le cadre des modèles de durées de vie pose problème. De plus, très souvent des situations dans lesquelles les données sont à la fois censurées à gauche et à droite ont pour conséquence de rendre la méthode d'inversion des données inappropriée.

Afin de contourner cette diffculté, nous proposons une méthode directe d'estimation de la fonction de hasard λ ne nécessitant pas le renversement de l'ordre des observations. Nous présentons donc un estimateur à noyaux de λ qui peut être vu comme la version gauche de celui étudié dans Taïbi-Hassani et Youndjé(2003).

Ce nouveau champ d'étude est introduit par Gomez et *al.*(1994) pour l'étude de la fonction de survie S. Ces auteurs proposent pour ce paramètre, une approche directe basée sur l'équation intégrale de Doléans à partir de laquelle ils construisent ce qui est communément appelée l'estimateur LKM (Left Kaplan Meier), version censurée à gauche de l'estimateur de Kaplan-Meier(1958).

À notre connaissance, il n'existe aucun estimateur du taux de hasard dans le cas de la censure gauche, en approche directe. Deux nouveaux estimateurs de type noyau de S sont construits comme un rapport d'un estimateur à noyau d'une sous-densité et d'une combinaison d'estimateurs de fonctions de survie.

Nous avons élaboré la construction de nouveaux estimateurs par approche directe de la fonction de survie et de la fonction de hasard pour des données censurées à gauche, utilisant l'estimateur de Kaplan-Meier lissé par noyaux. Les résultats de convergence uniforme presque sûre, de normalité asymptotique ainsi que les expressions asymptotiques de l'erreur quadratique intégrée sont établis (Taïbi et Adigaw-E-Touck (2011), Adigaw-E-Touck (2013), Taïbi et Adigaw-E-Touck (2014e)).

2.4 Prédiction modale non paramétrique : convergence uniforme de l'estimateur à noyau du mode conditionnel à partir d'observations dépendantes

Soit (X, Y) un couple aléatoire à valeurs dans un espace probabilisé $\left(\mathbb{R}^{p+1}, \mathcal{B}_{\mathbb{R}}^{p+1}, P\right)$ où P appartient à l'ensemble des probabilités possédant une densité $f = f_P(.,.)$.

Soit $f(.,.)$ la densité conditionnelle de Y par rapport à X définie par

$$f(y/x) = f(x,y) \Big/ \int_{\mathbb{R}} f(x,y)\, dy$$

pour tout $y \in \mathbb{R}$ et $x \in \mathbb{R}^p$ lorsque la marginale $\mathrm{f}_0(.) = \int_{\mathbb{R}} f(.,y)\, dy$ est strictement positive.

On désigne par C un compact de R^p sur lequel cette dernière condition est satisfaite et on s'intéresse à l'estimation du mode conditionnel, fonction notée θ et définie sur C par

$$\theta(x) = \arg\max_{y \in \mathbb{R}} f(y/x).$$

On suppose que cette fonction θ existe et satisfait la condition d'unicité uniforme suivante :

$$\forall \epsilon > 0, \exists \alpha > 0 \ : \ \forall t : C \to \mathbb{R}, \ \sup_{x \in C} |\theta(x) - t(x)| \geqslant \epsilon \Rightarrow \sup_{x \in C} |f(\theta(x)/x) - f(t(x)/x)| \geqslant \alpha \tag{1.1}$$

On se propose d'estimer le mode conditionnel θ par θ_n :

$$\theta_n(x) = \arg\max_{y \in \mathbb{R}} f_n(y/x), \forall x \in C$$

où $f_n(./.)$ est un estimateur de la densité conditionnelle, défini à partir de n couples aléatoires (X_i, Y_i), $i = 1, \ldots, n$, de même distribution que le couple (X, Y).

Nous tenons à préciser qu'il faut donc établir la convergence de l'estimateur de la densité conditionnelle sur R, et que seule la convergence sur un compact a été établie.

Nous donnons d'abord un résultat général permettant de déduire les propriétés de convergence stochastique de θ_n de celles de $f_n(./.)$ sans aucune autre hypothèse que celles introduites jusqu'ici. Ensuite nous énonçons les propriétés de convergence uniforme presque complète de l'estimateur θ_n défini classiquement à l'aide de noyaux sur la base d'un processus fortement mélangeant.

2.4.1 Résultat général

Remarque

Pour $C = \{x\}$ la condition d'unicité uniforme est satisfaite dès que $\theta(x)$ est unique. Avec $f(y/x) \to 0$ quand $y \to +\infty$, cette dernière condition est vérifiée dès que la marginale $f(./x)$ est uniformément continue.

On désigne par M l'un des deux modes de convergence"en probabilité" ou "presque complète".

Lemme

Si

$$\sup_{x\in C}\ \sup_{y\in\mathbb{R}}\ |f_n(y/x)-f(y/x)| \xrightarrow[n\to\infty]{M} 0$$

alors

$$\sup_{x\in C}|\theta_n(x)-\theta(x)| \xrightarrow[n\to\infty]{M} 0$$

2.4.2 Estimateurs à noyau

L'estimateur de la densité conditionnelle de $f(./.)$, f_n est défini comme suit (on convient que $\frac{0}{0}=0$) :

$$\forall (x,y)\in\mathbb{R}^{p+1},\ \forall n\in\mathbb{N},\ f_n(y/x)=\frac{\sum_{i=1}^n h_n^{-1}K_1\left(\frac{y-Y_i}{h_n}\right)K_0\left(\frac{x-X_i}{h_n}\right)}{\sum_{i=1}^n K_0\left(\frac{x-X_i}{h_n}\right)}$$

où $(h_n)_{n\in\mathbb{N}}$ est une suite réelle strictement positive telle que $\lim\limits_{n\to+\infty} h_n=0$ et où K_0 (resp. K_1) est un noyau de $\mathbb{R}^p$ (resp. $\mathbb{R}$).

2.4.3 Résultats de convergence

Proposition On suppose que la densité conjointe $f(.,.)$ est uniformément continue sur $C\times\mathbb{R}$, que les noyaux K_0 et K_1 sont lipschitziens, que le support de K_1 est borné et que l'une des deux hypothèses ci-dessous

– $\exists\, t>0:\ \mathbb{E}\,|Y|^t<+\infty$ et $\exists\,\nu>0:\ \sum\limits_{n=1}^{+\infty} h_n^{-\nu}<+\infty$

ou

– la v.a.r. Y est bornée

est satisfaite.

Si la suite $(h_n)_{n\in\mathbb{N}}$ vérifie conjointement avec la suite $i(n)$ associée à $(\phi_n)_{n\in\mathbb{N}}$ l'hypothèse

$$\frac{nh_n^{p+1}}{i(n)\ln(n)} \xrightarrow[n\to+\infty]{} \infty$$

alors on a

$$\sup_{x\in C}|\theta_n(x)-\theta(x)| \xrightarrow[n\to\infty]{p.co.} 0.$$

2.4.4 Applications à la prédiction modale

Soit $(Z_n)_{n\in\mathbb{N}}$ un processus strictement stationnaire à valeurs dans $\mathbb{R}^m$. Le mode autorégressif d'ordre q, noté θ, soit

$$\theta(x)=\arg\max_y f(y/x),\ \forall x\in\mathbb{R}^{mq}$$

où f représente la densité conditionnelle de Z_{q+1} sachant $Z_1,\ldots,Z_q$ est estimé par

$$\theta_n(x)=\arg\max_y f_n(y/x),\ \forall x\in\mathbb{R}^{mq}$$

avec f_n défini par :

$$f_n(y/x) = \frac{\sum_{i=1}^{n} h_n^{-1} K_1\left(\frac{y-Y_i}{h_n}\right) K_0\left(\frac{x-X_i}{h_n}\right)}{\sum_{i=1}^{n} K_0\left(\frac{x-X_i}{h_n}\right)}$$

avec

$$X_i = [Z_i, \ldots, Z_{i+q-1}], \quad Y_i = Z_{i+q}, i = 1, \ldots, n, \quad n = N - (q+1) - 1, \quad p = mq.$$

On suppose que le processus $(Z_n)_{n\in\mathbb{N}}$ est ϕ-mélangeant (voir Collomb (1984, 1985b) ou Collomb et *al.*1986) et que les hypothèses des paragraphes 1 et 3 sont satisfaites par X_i, Y_i. Nous appliquons ce dernier résultat à la prédiction modale de Z_{n+1} à partir de $Z_1, \ldots, Z_n$.

Proposition Sous les mêmes conditions que la proposition précédente, on a

$$\|\theta_n(Z_{n-q+1}, \ldots, Z_n) - \theta(Z_{n-q+1}, \ldots, Z_n)\| \mathbb{1}_{\{(Z_{n-q+1}, \ldots, Z_n)\in C\}} \xrightarrow[n\to\infty]{p.co.} 0.$$

2.5 Convergence ponctuelle d'un estimateur de la régression d'une variable aléatoire réelle par rapport à une variable aléatoire dans un espace mesurable pour des observations dépendantes : cas du consentement à payer

La méthode d'évaluation contingente permet d'évaluer les préférences individuelles pour un changement concernant un bien public. Au centre de cette méthode, nous trouvons un questionnaire qui permet de révéler le consentement à payer (CAP), c'est-à-dire le niveau de revenu que des individus seraient prêts à donner pour des biens hors-marchés. C'est l'une des techniques les plus communément appliquées pour définir une valeur monétaire de biens non-marchands, et notamment de zones humides, d'espèces en danger et de biodiversité etc. Récemment, de nombreuses études ont souligné l'impact significatif que peut jouer la localisation des enquêtés sur leur consentement à payer. Des travaux ont par exemple souligné, concernant l'évaluation monétaire des zones humides de l'estuaire de la Seine, que les habitants des grandes villes consentent à payer plus que les ruraux pour la protection de ces espaces naturels. Ainsi, la nécessité d'analyser plus précisément l'impact de la localisation des individus sur leur consentement à payer apparaît naturelle.

Notre objectif consiste à construire un modèle de prévision non paramétrique du consentement à payer (CAP) d'un individu en fonction de sa localisation géographique. En considérant le consentement à payer comme étant une variable aléatoire réelle $Y = CAP$, nous nous proposons donc d'estimer la régression r du CAP par rapport à sa localisation, soit la variable aléatoire X à valeurs dans un espace mesurable quelconque E. Nous pouvons voir cette méthode comme une extension du cas classique et largement étudié où $E = \mathbb{R}^p$. Ainsi, après avoir défini un estimateur de type Nadaraya-Watson (r_n) de r, nous présentons des résultats de convergence ponctuelle de cet estimateur. Plusieurs modèles de prédiction du CAP (notamment paramétriques) sont généralement utilisés dans le cadre de la méthode d'évaluation contingente : régressions logistiques

[simple ou multiple (Calkins et *al.*., 2002 ; Bekele et Drake, 2003 ; Soliño et *al.*, 2009)], modèle Probit (Maruyama et Takimoto, 2008) ou encore modèle Tobit (Beaumais et *al.*, 2008). Des comparaisons entre estimations paramétriques et semi-non paramétriques ont également été réalisées. Par exemple, Crooker et Herriges (2004) cherchent, à travers ce type de comparaison, à évaluer la sensibilité des résultats obtenus pour l'évaluation du CAP en fonction de la distribution des préférences et de la procédure d'estimation employée.

2.5.1 Estimation non paramétrique de la fonction de régression, Taïbi-Hassani *al.* 2015a

$(\Omega, \mathcal{A}, \mathbb{P})$ est un espace de probabilité. Soit $(E, \mathcal{E}, \mu)$ un espace mesuré où μ est une mesure positive et bornée et $(X_i, Y_i)_{i\in\mathbb{N}}$ une suite de couples de variables aléatoires définies sur $(\Omega, \mathcal{A}, \mathbb{P})$ et à valeurs dans $(E \times \mathbb{R}, \mathcal{E} \times \mathcal{B}_{\mathbb{R}})$. On suppose que pour tout i de $\mathbb{N}$, Y_i est intégrable et nous admettons l'existence d'une fonction réelle r définie sur E vérifiant

$$\mathbb{E}(Y_i/X_i) = r(X_i) \qquad \forall i \in \mathbb{N} \tag{2.1}$$

et considérons le problème de l'estimation de cette régression r. Nous étudions la convergence ponctuelle de la suite d'estimateurs $(r_n)_{n\in\mathbb{N}}$, définie par

$$r_n(x) = \begin{cases} \sum_{i=1}^n Y_i\, K_n(x, X_i) \,/\, \sum_{i=1}^n K_n(x, X_i) & \text{si} \quad \sum_{i=1}^n K_n(x, X_i) \neq 0 \\ 0 & \text{sinon} \end{cases} \tag{2.2}$$

où x est un élément fixé de E et $(K_n)_{n\in\mathbb{N}}$ une suite de fonctions réelles mesurables définies sur $E \times E$. Plus particulièrement, lorsque $E = \mathbb{R}^p$ et les couples $(X_i, Y_i)_{i\in\mathbb{N}}$ sont équidistribués, il s'agit d'estimer la régression r de Y_1 par rapport à X_1 : ce problème a été abondamment étudié dans le cas où les couples $(X_i, Y_i)_{i\in\mathbb{N}}$ sont indépendants [voir la revue bibliographique de Collomb, (1981), (1985a)]. L'estimateur non paramétrique de r le plus usuel est alors l'estimateur de Nadaraya-Watson avec

$$K_n(x, z) = h_n^{-p} K\left(\frac{x - z}{h_n}\right) \tag{2.3}$$

où K est un noyau de $\mathbb{R}^p$.

Les propriétés de convergence presque sûre de tels estimateurs ont été étendues au cas de couples $(X_i, Y_i)_{i\in\mathbb{N}}$ non nécessairement indépendants mais formant un processus uniformément fortement mélangeant (Collomb 1984,1985b). Certains de ces derniers résultats sont maintenant étendus à des estimateurs de $r(x)$ définis à l'aide de suites $(K_n(x, .))_{n\in\mathbb{N}}$ plus générales et pour des variables aléatoires $(X_i)_{i\in\mathbb{N}}$ à valeurs dans un espace E qui n'est pas nécessairement $\mathbb{R}^p$.

Soit $(X_n)_{n\in\mathbb{N}}$ à valeurs dans $E = \{1, 2, \ldots, m\}$, $m \in \mathbb{N}$. Nous supposons que

$$P\left(X_{n+1} = j / X_n = i\right) = \pi(i, j), \quad \forall (i, j) \in E, \quad \forall n \in \mathbb{N} \qquad ,(*1)$$

ce qui est satisfait dès que $(X_n)_{n\in\mathbb{N}}$ est une chaîne de Markov stationnaire.
Nous estimons la probabilité de transition

$$\pi(i, j) = \mathbb{E}\left[\mathbb{1}_{\{X_2=j\}} / X_1 = i\right] \tag{*2}$$

par

$$\pi_n(i,j) = \frac{\displaystyle\sum_{k=1}^{n-1} \mathbb{1}_{\{X_{k+1}=i\}} \mathbb{1}_{\{X_k=j\}}}{\displaystyle\sum_{k=1}^{n-1} \mathbb{1}_{\{X_k=j\}}}$$

et, dans ce cas,

$$K_n(x,z) = \mathbb{1}_{\{x=z\}}.$$

La généralisation de l'estimateur de Nadaraya-Watson permet de traiter le cas du régressogramme que nous retrouvons en prenant :

$$K_n(x,z) = \begin{cases} \dfrac{1}{h_n^p} & \text{si} \quad z \in \displaystyle\prod_{i=1}^{p} \left[h_n \left[\frac{x_i}{h_n} \right], h_n \left[\frac{x_i}{h_n} \right] + 1 \right] \\ 0 & \text{sinon} \end{cases}$$

$\forall z \in \mathbb{R}^p$, $\forall x = (x_1, \ldots, x_p) \in \mathbb{R}^p$ et $\forall n \in \mathbb{N}$. $[x]$ désigne la partie entière de x.

2.5.2 Hypothèses générales

Nous supposons que

(H_1) la loi de probabilité de X_i admet une densité f_i par rapport à μ telle que

$$\exists \quad \gamma > 0, \quad \Gamma > 0 \quad : \quad \gamma < f_i < \Gamma, \quad \forall i \in \mathbb{N}$$
$$\exists M > 0, \quad |Y_i| < M, \quad \forall i \in \mathbb{N};$$

(H_2) le processus $(X_i, Y_i)_{i\in\mathbb{N}}$ est uniformément fortement mélangeant (ou ϕ-mélangeant),

(H_3) les fonctions K_n satisfont les hypothèses suivantes, pour x fixé dans E,

a) $\forall x \in E$ et $\forall n \in \mathbb{N}$, $\displaystyle\int K_n(x,z)\, \mu(dz) = 1$,

b) $\forall n \in \mathbb{N}$, K_n est strictement positive sur $E \times E$.

c) il existe une constante C telle que pour tout $n \in \mathbb{N}$ et tout x, y dans E,

$$K_n(x,y) \leqslant C \int K_n^2(x,.)\, d\mu$$

Soit ψ une fonction réelle mesurable définie sur E. Nous dirons que la suite $(K_n)_{n\in\mathbb{N}}$ et la fonction ψ vérifient "la condition $\mathcal{C}$", si pour tout x de E,

$$\lim_{n\to+\infty} \int_E \psi(.) K_n(x,.)\, d\mu = \psi(x).$$

Remarque

Lorsque $E = \mathbb{R}^p$ et $(K_n)_{n\in\mathbb{N}}$ est une suite de noyaux définie de la façon suivante à partir d'un noyau K à support compact par

$$\forall n \in \mathbb{N}, \quad K_n(x,.) = \frac{1}{h_n} K\left(\frac{x - .}{h_n} \right)$$

la condition $\mathcal{C}$ est vérifiée pour toute fonction ψ continue.

2.5.3 Estimation non paramétrique du consentement à payer

Supposons que nous cherchons à prévoir le consentement à payer y à partir de la géolocalisation x exprimée par sa latitude et sa longitude, sachant que nous disposons d'un échantillon de couples (X_i, Y_i), $i = 1, \ldots, n$ où

$$\begin{cases} X_i : & \text{est le couple (latitude, longitude) } i \\ Y_i : & \text{est le prix que l'individu consente à payer.} \end{cases}$$

Les personnes $i = 1, \ldots, n$ étant en général assez proches, il n'y a pas indépendance entre les couples (X_i, Y_i), $i = 1, \ldots, n$. Nous pouvons toutefois supposer que pour des habitants suffisamment éloignés il y a "presque indépendance", ce qui est exprimé par la condition de ϕ-mélange. Nous supposons par ailleurs l'équidistribution des couples (X_i, Y_i), $i = 1, \ldots, n$ et nous désignons par (X, Y) un couple de même loi. Il est donc naturel de prévoir y à partir de x par

$$y = r(x) = \mathbb{E}(Y/X = x).$$

Il s'agit donc de l'estimation de la régression d'une variable aléatoire réelle Y par rapport à une variable aléatoire X à valeurs dans l'espace E qui, par commodité, peut être considéré comme espace métrique. E est ici la sphère de rayon unité muni de la distance

$$d(x, z) = \inf \{\text{arc}(x, z), 2\pi - \text{arc}(x, z)\}$$

où $\text{arc}(x, z)$ désigne l'arc du plus grand cercle passant par x et z.

2.5.4 Propriétés de convergence ponctuelle de l'estimateur du consentement à payer : cas des observations α et ϕ mélangeantes

Nous donnons ci-dessous des propriétés de convergence ponctuelle et presque coplète de l'estimateur de régressionr_n. Soit i la fonction associée à $\phi = (\phi_n)_{n\in\mathbb{N}}$:

$$i_\phi(n) = i(n). \tag{5.1}$$

Proposition 1.

Si r et K_n vérifient la condition $\mathcal{C}$ et si

$$\lim_{n\to+\infty} n^{-1} i(n) \int K_n^2(x, .)\, d\mu = 0 \tag{5.2}$$

alors, pour tout point x fixé dans E, lorsque n tend vers $+\infty$:

$$r_n(x) \xrightarrow{\text{P}} r(x) \tag{5.3}$$

Proposition 2.

Si r et K_n vérifient la condition $\mathcal{C}$ et si

$$\lim_{n\to+\infty} n^{-1} i(n) \ln n \int K_n^2(x, .)\, d\mu = 0 \tag{5.4}$$

alors, pour tout point x fixé dans E, lorsque n tend vers $+\infty$:

$$r_n(x) \xrightarrow{\text{P.Co}} r(x) \tag{5.5}$$

Pour une géolocalisation connue x, le consentement à payer d'un habitant peut alors être estimé par $r_n(x)$ selon les deux modes de convergences, en probabilité et presque sûrement.

Remarques

- Lorsque le processus $(X_i, Y_i)_{i\in\mathbb{N}}$ est m-dépendant, on peut choisir la suite i_ϕ constante et les conditions (5.2) et (5.4) deviennent

$$\lim_{n\to+\infty} n^{-1} \int K_n^2(x,.)\, d\mu = 0 \text{ et } \lim_{n\to+\infty} n^{-1} \ln(n) \int K_n^2(x,.)\, d\mu = 0.$$

- Si $(X_i, Y_i)_{i\in\mathbb{N}}$ est markovien et stationnaire, Rosenblatt(1971) montre que si, de plus, il vérifie la condition de Doeblin (Doob 1953) ou la condition de norme L^p pour $p = 1$ ou $p = \infty$, alors $i(n) \geqslant \ln(n)$.

Chapitre 3

Modélisation de données socio-économiques

3.1 Modélisation de données qualitatives et application dans le cas de données socio-économiques. La méthode des Random Forest versus Analyse discriminante.

Les travaux de recherche que je vais décrire dans cette partie (Laroutis et Taïbi 2011) sont basés sur la méthode d'évaluation contingente (MEC). Cette méthode constitue une méthode économique quantifiant monétairement l'ensemble des valeurs que les individus attribuent à un bien environnemental donné. La démarche consiste à administrer un questionnaire qui vise à révéler le consentement à payer (CAP) des individus pour la préservation des zones humides de l'estuaire de la Seine. L'objectif est de prédire une variable quantitative (CAP) à l'aide de variables qualitatives (socio-économiques).

Nous nous limitons au cas où la variable dépendante est binaire, la procédure pouvant être étendue au cas de variables qualitatives polytomiques. Notre objectif est donc de construire un modèle permettant de prédire la variable à expliquer. Les prédicteurs étant pour la plupart des variables qualitatives (nominales ou ordinales), nous avons utilisé une procédure permettant de les transformer en variables quantitatives en s'inspirant de la Méthode Disqual (Saporta, 1977). Nous effectuons l'analyse des correspondances multiples des prédicteurs c'est-à-dire l'analyse des correspondances du tableau disjonctif. Les p variables explicatives sélectionnées $X_1, X_2, \ldots, X_p$ sont remplacées par les coordonnées des n individus sur les q axes factoriels ($q < p$) en opérant une pondération permettant de conserver l'importance des composantes. Les méthodes prédictives dans le cas où la variable endogène ou à prédire esr qualitative sont dites méthodes de classement. Parmi les méthodes de classement classiques on peut citer l'analyse discriminante, les réseaux de neurones de Kohonen , la régression logistique, la méthode CART, la méthode CHAID, Support Vector Machine (SVM) et la régression logistique PLS. Dans le cadre de la méthode d'évaluation contingente, le traitement des variables explicatives du CAP des individus se réalise généralement par un traitement économétrique de type Logit, Probit ou encore Tobit. La méthode des Forêts Aléatoires (Random Forests) offre des résultats encourageants (Breiman, 2001). Les travaux récents sur le sujet mettent en évidence la supériorité prédictive de ce type de

méthodes par rapport aux modèles de régression généralement utilisés (Iverson et *al.*, 2004 ; Prasad et *al.*, 2006 ; Peters et *al.*, 2007).

L'extension de la méthode des Random Forest au cas de données qualitatives est une innovation et peut offrir de nombreuses perspectives pour des applications en sociologie, médecine, sciences agronomiques, sciences animales et vétérinaires,...

Deux méthodes de prédictives ont été mises en œuvre : l'analyse discriminante et la méthode des forêts aléatoires. Le but est aussi de comparer leurs performances en termes de classement.

Peters et *al.* (2007) soulignent l'absence, dans certains domaines, de travaux utilisant la méthode des forêts aléatoires ce qui conduit ainsi à limiter les comparaisons et les analyses. Par exemple, très peu d'études se sont intéressées à la modélisation de la distribution écologique. Il apparaît également qu'aucune recherche dans le cadre de la méthode d'évaluation contingente n'a été jusqu'alors réalisée. Notons que, même si les forêts aléatoires ont fait l'objet de travaux multiples consistant à les comparer à d'autres modèles prédictifs (principalement la régression logistique), aucune comparaison n'a encore été mise en œuvre avec l'analyse discriminante. L'extension de la méthode des forêts aléatoires au cas de variables qualitatives est une méthode statistique innovante.

3.1.1 Étude sur des résultats d'enquête : cas du consentement à payer

L'enquête a été administrée auprès d'un échantillon représentatif de 300 individus. Les prédicteurs (sexe, zone géographique, situation familiale, niveau d'études, avis sur le programme, etc.) sont pour la plupart des variables qualitatives (nominales ou ordinales). Afin d'expliquer la participation des individus au programme de préservation des zones humides de l'estuaire de la Seine, nous avons utilisé une procédure permettant de transformer ces variables qualitatives en variables quantitatives. Nous effectuons donc une analyse des correspondances multiples (ACM) des prédicteurs pour transformer les variables qualitatives en variables quantitatives. L'intérêt d'une telle démarche est de s'affranchir du problème de multicolinéarité tout en gardant la structure d'origine du tableau de données.

L'ACM est considérée comme une analyse factorielle des correspondances du tableau disjonctif complet. L'ACM permet de mettre en évidence les proximités entre les observations et de synthétiser l'information. L'ACM fournit des axes factoriels qui sont définis comme combinaisons linéaires des variables indicatrices des modalités. Nous notons k_{ij} le terme général du tableau disjonctif complet. Les p variables explicatives sélectionnées $X_1, X_2, \ldots, X_p$ sont remplacées par les coordonnées sur les q axes factoriels ($q \leq p$). Dans notre cas, nous supposerons que $q = p$. La loi des valeurs propres issues d'une ACM dépend de trop de paramètres pour que nous puissions la tabuler. L'extraction d'un nombre de facteurs significatifs n'a pas d'intérêt dans le cas présent, nous gardons dans toute la suite l'ensemble des facteurs issus de cette analyse. En effet nous voulons conserver la totalité de l'information de la base de données et rendre cette procédure automatique.

Au total, les 34 facteurs issus de l'ACM ont été exploités afin de conserver toute l'information de la base de données initiale. Cette procédure permet aussi de pouvoir reconduire l'analyse pour d'autres vagues d'enquêtes sans sélection des facteurs. De plus,

en conservant tous les facteurs obtenus par l'analyse des correspondances multiples la quantification des variables X_j est celle qui donne la distance de Mahalanobis la plus grande entre les deux groupes.

Nous avons transformé les coordonnées des individus en les pondérant par les pourcentages d'inertie. Cette opération permet en effet de conserver l'inertie de chaque facteur.

La procédure suivante est de trouver un modèle prévisionnel du consentement à payer. La méthode des forêts aléatoires ou RF (Breiman 2001) et l'analyse discriminante linéaire de Fisher (LDA) ont été choisies. En effet l'analyse discriminante linéaire est une méthode destinée à décrire et classer des individus caractérisés par un nombre important de variables quantitatives (Saporta, 1977). Parmi les groupes connus, les principales différences sont caractérisées à l'aide des variables mesurées et le groupe d'appartenance d'une nouvelle observation est déterminé uniquement à partir des variables mesurées. Des transformations linéaires sont faites sur les variables initiales pour créer de nouvelles variables composites : les fonctions canoniques discriminantes. La pondération des variables définissant les fonctions discriminantes est là pour maximiser la séparation des groupes.

Cette technique consiste à chercher des axes sur lesquels nous projetons les observations de telle sorte que :

- les centres des k groupes soient projetés avec la dispersion maximale ;
- les projections des observations de chaque groupe soient en moyenne peu dispersées.

Soient g_j les centres de gravité affectés des poids q_j des k nuages. La matrice de variance inter-groupes B est donc définie par : $B = q_j(g_j - g)(g_j - g)'$, et la matrice de variance intra-groupes W est donnée par $W = \sum_{j=1}^{=n} q_j V_j$ où V_j étant la variance du sous-nuage j.

La matrice de variance-covariance de l'ensemble des observations E est notée V. Nous avons $V = B + W$. On cherche donc une combinaison linéaire u telle que $u'Bu$ soit maximal et $u'Wu$ soit minimal. Nous choisissons comme critère de maximiser $u'Bu/u'Wu$ (ou $u'Bu/u'Vu$) pour séparer au mieux les groupes (inter) tout en gardant l'homogénéité intra-groupes.

La variable " CAP " étant binaire, $k = 2$, il n'existe donc qu'un axe discriminant, la droite des centres, de vecteur directeur $(g_1 - g_2)$ sur laquelle nous projetons avec la métrique W^{-1} (métrique de Mahalanobis). La combinaison linéaire $W^{-1}(g_1 - g_2)$ est appelée fonction de Fisher.

Nous exprimons alors directement d comme combinaison linéaire des indicatrices des X_j ce qui revient à attribuer à chaque catégorie de chaque variable une valeur numérique ou score. La fonction discriminante d est alors égale à l'addition des scores obtenus dans les catégories des p variables (Thiria et *al.*,1997).

3.1.2 Les forêts aléatoires

Soient Y la variable réponse à prédire,$X_1, X_2, \ldots, X_p$ les p variables d'entrées et n le nombre d'observations.

Breiman (2001) a proposé une famille de méthodes de classification appelée Random Forests basées sur le concept de randomisation. Une forêt aléatoire (RF) consiste

en un nombre arbitraire d'arbres simples, utilisés pour calculer un vote pour la classe la plus populaire (classification), ou dont les réponses sont combinées (moyennées) pour obtenir une estimation de la variable dépendante (régression). En utilisant les RF, nous obtenons une amélioration significative de la prévision par rapport aux techniques classiques (CART,...).

Une forêt aléatoire est un ensemble de m arbres de classification ou de régressions construits à partir des données disponibles, ainsi que de n échantillons bootstrap. Pour chaque échantillon i, nous construisons le i-ème arbre tel qu'à chaque nœud, nous choisissons la meilleure partition obtenue à partir de k variables prises aléatoirement dans $X_1, X_2, \ldots, X_p$. Le résultat pour le vecteur d'entrée $(y, x_{i_1}, x_{i_2}, \ldots, x_{i_p})$ est alors la classe la plus populaire parmi les m arbres lorsqu'il s'agit d'une classification, ou la moyenne obtenue à partir des m arbres lorsqu'il s'agit d'une régression.

La réponse de chaque arbre dépend du sous-ensemble de prédicteurs choisis indépendamment (avec remplacement) et avec la même distribution pour tous les arbres de la forêt.

Les RF peuvent être définies comme un principe générique de combinaisons de classifieurs composé de L classifieurs élémentaires de type arbres de décision et notés $h(x, \theta_k)$, où θ_k est une suite de famille de vecteurs aléatoires indépendants et identiquement distribués et où x représente une donnée d'entrée (Bernard et *al.* 2009)

Soit n le nombre de données d'apprentissage. Un arbre de décision est construit selon l'algorithme ci-après

1. Tirer aléatoirement n individus avec remise. L'ensemble résultant est utilisé pour induire l'arbre.
2. Pour p variables, un sous-ensemble de k variables est tiré aléatoirement. Le meilleur sous-ensemble est sélectionné pour le partitionnement.
3. Pas à pas on construit un arbre jusqu'à atteinte d'une taille maximale.

Ce processus implique le choix de deux paramètres : k et le nombre d'arbres m. Breiman (2001) a choisi de fixer l'hyperparamètre k, avec $k = 1$ ou $k = Pe[1+\log_2(p+1)]$ mais sans justification pour le choix de k. À ce sujet , d'ailleurs, très peu de travaux existent dans la littérature. Dans Bernard *et al.* 2007, un intervalle de k optimal a été identifié, celui-ci contient $Pe[1+\log_2(p+1)]$ et $\sqrt{p}$. Les conclusions données à partir de travaux de Heutte et *al.* 2008 montrent que $Pe[1+\log_2(p+1)]$ et $\sqrt{p}$ sont comparables et ne donnent pas de différence significative.

Les résultats donnés ci-après sont issus de l'implémentation de l'algorithme Random Forests du module STATISTICA Data Miner. L'hyperparamètre k a été fixé à $Pe[1 + \log_2(p + 1)]$, où p est le nombre de variables retenues pour l'échantillon d'apprentisage.

Le nombre m d'arbres doit aussi être fixé a priori, en général ce nombre est compris entre 100 et 500. Breiman (2001) montre que lorsque m est grand, nous ne rencontrons pas de problèmes liés au surdimensionnement mais, au contraire, l'erreur de généralisation converge presque sûrement selon la loi forte des grands nombres. Cette erreur de généralisation est estimée par l'intermédiaire de l'erreur out-of-bag (OOB), calculée au fur et à mesure des itérations de l'algorithme. L'erreur OOB correspond à l'erreur faite lors de la prédiction des données n'appartenant pas à l'échantillon bootstrap utilisé pour construire l'arbre. Elle est également utile à la sélection des variables

d'importance et à comprendre l'interaction entre les variables d'entrée. En effet, si deux variables contiennent la même information, une seule d'entre elles est utile et l'ajout de la seconde n'aura pas pour effet de réduire l'erreur.

L'importance d'une variable X_j se calcule grâce à deux estimateurs correspondant tous deux à l'augmentation de l'erreur de régression lorsque nous permutons les valeurs de X_j, les autres variables restant inchangées. De ce fait, lorsque la valeur de l'estimateur d'importance est élevée pour la variable X_j, cela signifie qu'une variation de X_j induit une erreur importante, et donc que cette variable influe fortement sur la variable réponse. La différence entre les deux estimateurs utilisés, hormis le fait que le premier est normalisé, vient uniquement de l'ensemble de données utilisé.

Le premier estimateur, appelé pourcentage de gain de l'erreur quadratique moyenne (%IncMSE, Increased Mean Square Error), est calculé à partir des données out-of-bag :

$$\%IncMSE(X_j) = \sigma^{-1} \times [MSE_{OOB}(Xj) - MSE'_{OOB}(X_j)]$$

avec
- σ l'erreur-type, i.e. l'écart-type de l'erreur de régression ;
- $MSE_{OOB}(X_j) = E[(Y_{OOB}(X_j))^2]$, l'erreur quadratique moyenne pour la variable d'entrée X_j ;
- $MSE'_{OOB}(X_j)$ la nouvelle erreur quadratique moyenne pour la variable d'entrée X_j après permutation des valeurs de X_j.

Le second estimateur, le gain de pureté du nœud (GPN), est en fait calculé à partir de la somme des carrés résiduelle, et utilise les données in-bag servant à la construction des arbres avec :
- y_i la variable réponse du i-ème échantillon bootstrap ;
- $\hat{y}_i(X_j)$ son estimation avec les valeurs de $(X_1, \ldots, X_K)$;
- $\hat{y}'i(X_i)$ sa nouvelle estimation après permutation des valeurs de la variable X_j.

3.1.3 Résultats

Afin de comparer les deux méthodes, la LDA et les RF, nous nous intéressons aux matrices de confusion. Afin d'évaluer les modèles, l'échantillon total a été subdivisé en deux échantillons aléatoires : l'échantillon d'apprentissage et l'échantillon test. La base de données comporte 299 observations, un seul individu (atypique) a été retiré de la base. La taille de l'échantillon d'apprentissage correspond au 2/3 de la base (soit 200 individus) et celle de l'échantillon test à 1/3 (soit 99 individus) (cf. Garzón et *al.*, 2006).

Les résultats de l'analyse factorielle discriminante de Fisher sur un échantillon (randomisé) de taille 200 montrent que l'axe a un pouvoir discriminant significatif (Pvalue $<$ 0,0005). La fonction score aboutit à un bon taux de classement puisque le pourcentage d'observations bien classées est de 75,5% pour les deux groupes .

À notre connaissance, aucune étude n'a comparé les résultats des forêts aléatoires à ceux de l'analyse discriminante (Laroutis et Taïbi 2011). Cependant, l'apparente supériorité des forêts aléatoires par rapport aux autres méthodes (et notamment la régression logistique) pourrait conduire à considérer également une supériorité par rapport à l'analyse discriminante. Peters et *al.* (2008) par exemple, lors de leur comparaison des forêts aléatoires à la régression logistique ont conclu que la méthode des RF serait plus précise que la régression logistique polytomique.

Dans le cadre de notre étude, la statistique de McNemar est inférieure à la valeur théorique (3,84) conduisant ainsi à considérer qu'il n'y a aucune différence significative de performance entre les forêts aléatoires et l'analyse discriminante.

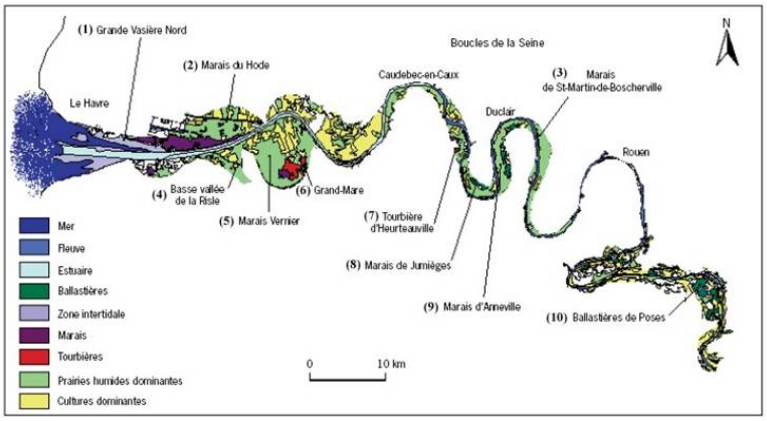
Afin de valider ces conclusions, nous avons également calculé le Percentage Correctly Classified (Burez et Van den Poel, 2007 ; Coussement et Van den Poel, 2008) qui confirme nos résultats. Comme le mettent en avant Coussement et Van den Poel (2008), cet indicateur est sans doute l'un des plus communément utilisé afin d'évaluer les performances d'un classificateur.

Ce champ de travaux de modélisation et prédiction de données qualitatives étudiées par Bouroche et *al.* (1977), ouvre de nouvelles perspectives. Par exemple, dans les traitements de données tels les scores de notations dans le diagnostic de maladie des plantes. D'autres types de modélisation peuvent être traités avec la même démarche. Les résultats décrits ci-dessus sont issus de modélisation sur des données réelles. Ces recherches doivent être poursuivies sur d'autres bases de données.

3.2 Méthode des programmes : plans fractionnaires et plans de sondage

Introduction

Le contexte du projet pluridisciplinaire EMIRE est le suivant(Arkoun *al.* 2012). Le phénomène du ruissellement érosif engendre de fortes dégradations du sol, inondations avec coulées de boues, destruction des infrastructures, pollution de l'eau via les nappes phréatiques. Ce phénomène est assez important dans la Vallée du Commerce.



Un programme de recherche financé par les Grands Réseaux de recherche Régionaux (GRR VATA) a donc été lancé en 2009 afin de lutter contre le ruissellement érosif.

La méthode d'évaluation contingente a connu une progression importante ces 20 dernières années en raison de sa relative simplicité et son large éventail d'applications (Hanley et *al.*, 1998). Toutefois, elle n'est pas non plus adaptée à l'étude de changements multidimensionnels (Hanley et *al.*, 2005). D'autres sont apparues qui visent à modéliser les préférences individuelles pour des biens décrits par un ensemble d'attributs prenant différentes valeurs. Ces méthodes sont fondées sur des enquêtes où les personnes interrogées sont confrontées à plusieurs scénarii. Ainsi, chaque expérience présentée aux enquêtés correspond généralement à un ensemble de trois scenarii. Chacun des deux

premiers scénarii représente la situation (hypothétique) qui prévaudrait suite à la mise en place d'une mesure à appliquer. Ces scenarii associent donc pour chaque attribut considéré un certain niveau (ou état). Le troisième scénario est qualifié de statu quo (non-intervention). Le répondant devra choisir le scénario qu'il préfère parmi ceux qui lui sont proposés. Cette expérience se renouvelle plusieurs fois consécutives face à des scenarii alternatifs différents.

Cette démarche itérative permet à l'évaluateur de disposer d'un éventail de préférences révélées. Cette méthode est connue aussi sous le nom de la méthode des programmes (Choice experiment).

Globalement, à chaque itération, l'enquêté arbitre entre deux alternatives et le statu quo. Selon la méthode utilisée, les personnes interrogées sont invitées à classer les alternatives/actions, à les noter ou à indiquer celle qu'elles préfèrent. Un autre attribut est ajouté : il s'agit d'un attribut monétaire pour lequel plusieurs niveaux sont également définis. Cet exercice est répété par chaque enquêté un certain nombre de fois pour des choix différents d'alternatives, la sélection des choix obéissant aux modalités d'un plan d'expériences. Cet attribut permet de prendre en compte la contrainte budgétaire du consommateur dans la démarche de choix et présente l'avantage, par rapport à d'autres techniques d'évaluation directe, de ne pas attendre de l'individu une construction de la valeur qu'il associe au bien. On adopte un processus de génération fractionnaire des expériences qui consiste, en respectant des critères statistiques bien définis (dont le critère d'orthogonalité), à combiner les différents attributs, et leurs différents niveaux pour générer plusieurs scénarii. Chaque scénario traduit un état spécifique du bien étudié au regard de chacun des attributs et de l'attribut monétaire qui prévaudrait sous l'effet d'une mesure publique particulière. Á ce stade, il est important de noter que l'attribut monétaire est considéré au même titre que les autres attributs du bien pour générer les scenarii, ce qui implique que les scénarii qui résultent de ce processus de génération peuvent très bien illustrer des mesures politiques coûteuses, mais pour lesquelles l'attribut monétaire associé est fondamental dans la construction des expériences et permet justement de tester le rôle de contrainte budgétaire des agents lors de l'enquête.

Afin d'estimer au mieux le choix préféré des habitants de la Vallée du Commerce, un échantillon de 619 habitants a été sondé.

La méthode des programmes, comme définie plus haut, est inhérente à un questionnaire structuré de la manière suivante :

- une partie portant sur les critères sociodémographiques,
- une partie présentant la problématique et la description des programmes,
- une partie invitant l'enquêté à choisir entre différents scénarios.

La dernière partie du questionnaire implique de faire des choix de blocs de scénarii. En effet il est clair que chaque répondant ne peut être soumis à choisir parmi toutes les batteries de combinaisons de programmes ou d'actions. Dans notre cas les blocs au nombre de quatre sont formés de six scénarios. En effet, hormis donc un noyau de questions commun à tous les répondants, quatre blocs différents sont présentés. Tous les enquêtés n'ont donc pas les mêmes choix. Le dispositif est par conséquent incomplet. Ce point nous a conduits à développer une approche méthodologique spécifique pour le choix du plan d'échantillonnage.

Nous tenons à préciser que notre démarche peut être aisément reproduite, et

X = Champ d'actions réalisé

	Situation 1	Situation 2	Situation 3 (SQ)
Modif. des prat. agricoles		X	
Amél. des infra. de protec.	X		
Dévelop. de la commu.		X	
Coût	25€ /an	37,5€ /an	0
La meilleure situation est	1	2	3

TABLE 3.1 – **Un choix de programme parmi les 6 proposés**

l'exemple que nous prenons ne sert qu'à illustrer notre travail. Pour le choix de la méthode d'échantillonnage, nous avons donc opté pour la méthode des quotas. L'un des avantages de cette méthode est qu'elle ne nécessite pas de connaître la base de sondage. Nous avons accès à la répartition des variables telles que le sexe, la classe d'âge, la profession et la localisation sur toute la vallée du Commerce (données Insee). La méthode de sondage par quotas est reprise dans la plupart des études faisant référence à la méthode des programmes Czajkowski et *al.* (2009). Ces articles ne font référence à la méthode d'échantillonnage que pour constituer l'échantillon global. Il n'est nullement mentionné que les quotas utilisés pour la stratification de l'échantillon total sont également conservés pour tous les sous-échantillons relatifs à chaque bloc. Il peut donc arriver que nous puissions proposer le bloc X à une certaines frange de la population et qu'au final les caractéristiques des enquêtés des différents blocs ne répondent pas aux quotas considérés pour constituer l'échantillon mère. D'autre part Ladenburg et *al.* (2008) montrent à partir d'une étude empirique que le biais est engendré par la variable "genre". Nous nous proposons dans ce travail d'administrer le questionnaire à quatre échantillons de taille égale à 155 et répondant aux mêmes quotas que ceux appliqués à l'échantillon total. Les hypothèses que nous émettons sont de confirmer ou d'infirmer que les modèles issus des quatre échantillons sont similaires. S'il y a un lien significatif entre les caractéristiques des enquêtés et le choix des programmes, est-il le même quelque soit le modèle ?

L'échantillon étudié est bien représentatif dans la vallée du commerce selon les trois critères genre, âge et localisation conformément aux données fournies par l'INSEE. Trois actions (ou programmes) susceptibles d'être mises en place pour lutter contre le ruissellement ont été proposées :

– la modification des pratiques agricoles,
– l'amélioration des infrastructures de protection,
– le développement de la communication.

Il est proposé à chaque enquêté un bloc avec six choix, et pour chaque choix, il est demandé à l'enquêté de choisir une situation parmi les trois proposées. Les situations font référence aux actions menées associées à un coût annuel par foyer. La situation 3 est le statu quo, c.à.d, aucune amélioration et donc aucun coût supplémentaire n'est proposé.

3.2.1 Problématique

Dans le questionnaire, nous considérons les trois critères comme trois facteurs à deux niveaux, avec -1 (niveau bas) et $+1$ (niveau haut). Le montant à payer est considéré comme un facteur à trois niveaux 12.5 €, 25 € et 37.5€, que nous codons par -1, 0 et $+1$.

Le nombre possible de combinaisons est égal à $2 \times 2 \times 2 \times 3 = 24$. Pour deux alternatives, $\binom{24}{2} = 276$ possibilités. Ce nombre est trop élevé, ce qui nous amène à utiliser un plan fractionnaire.

On a proposé aux enquêtés des blocs différents de choix de programmes, la question que l'on se pose est la suivante : le bloc a t-il une influence sur le montant que l'enquêté est prêt à payer ?

L'objectif est de développer une démarche innovante prenant en compte les biais inhérents à cette méthode tel que le plan d'expériences utilisé.

Nous rappelons brièvement la définition d'un plan d'expériences fractionnaire (Goupy 2006).

Definition Un plan d'expériences complet est un plan dans lequel toutes les combinaisons distinctes de niveaux des facteurs sont présentes, pour un tel plan avec k facteurs à 2 niveaux, il faut effectuer au minimum 2^k expériences.

Lorsque le nombre de facteurs est élevé, augmente ainsi le nombre d'expériences requis, d'où l'idée de diminuer la taille des plans et donc d'utiliser pour l'étude de k facteurs des matrices d'expériences issues de plans $2^{k-1}, 2^{k-2}, \cdots, 2^{k-p}$.

Definition Un plan fractionnaire est une partie d'un plan complet, on parle de plan 2^{k-p}, ce qui veut dire k facteurs mais 2^{k-p} essais.

Theorem 3.2.1 *Un plan d'expériences est orthogonal et équilibré vis-a-vis d'un modèle M donné, si et seulement si, pour chaque couple d'actions disjointes du modèle M, tous ces couples de niveaux possibles sont présents un même nombre de fois dans le plan d'expériences.*

Autrement dit, le plan d'expériences orthogonal est un sous ensemble d'un plan d'expériences complet qui satisfait les conditions suivantes :

- pour chaque facteur, les niveaux pris sont équilibrés (chaque niveau du facteur apparaît le même nombre de fois dans le plan d'expériences),
- le produit scalaire de deux colonnes quelconques de la matrice d'expériences est nul,
- les niveaux pris par chacun des attributs ne sont pas corrélés entre eux,
- la distance entre les différents niveaux pris par chacun des attributs doit être la même quelque soit le niveau (règle d'équidistance). La règle d'équidistance vise à minimiser la variance des paramètres du modèle estimé a posteriori et donc à maximiser la significativité.

Bloc1		Prog1	Prog2	Prog3	Coût
Choix 1	1	−1	−1	−1	−1
	2	1	1	1	0
Choix 2	1	1	1	1	−1
	2	−1	−1	−1	0
Choix 3	1	−1	1	−1	0
	2	1	−1	1	1
Choix 4	1	1	−1	1	0
	2	−1	1	−1	1
Choix 5	1	−1	−1	−1	1
	2	1	1	1	−1
Choix 6	1	1	1	1	1
	2	−1	−1	−1	−1

Bloc2		Prog1	Prog2	Prog3	Coût
Choix 1	1	−1	−1	1	−1
	2	1	1	−1	0
Choix 2	1	1	−1	−1	−1
	2	−1	1	1	0
Choix 3	1	−1	1	1	0
	2	1	−1	−1	1
Choix 4	1	1	1	−1	0
	2	−1	−1	1	1
Choix 5	1	−1	−1	1	1
	2	1	1	−1	−1
Choix 6	1	1	−1	−1	1
	2	−1	1	1	−1

Bloc3		Prog1	Prog2	Prog3	Coût
Choix 1	1	−1	1	−1	−1
	2	1	−1	1	0
Choix 2	1	1	−1	1	−1
	2	−1	1	−1	0
Choix 3	1	−1	−1	−1	0
	2	1	1	1	1
Choix 4	1	1	1	1	0
	2	−1	−1	−1	1
Choix 5	1	−1	1	−1	1
	2	1	−1	1	−1
Choix 6	1	1	−1	1	1
	2	−1	1	−1	−1

Bloc4		Prog1	Prog2	Prog3	Coût
Choix 1	1	−1	1	1	−1
	2	1	−1	−1	0
Choix 2	1	1	1	−1	−1
	2	−1	−1	1	0
Choix 3	1	−1	−1	1	0
	2	1	1	−1	1
Choix 4	1	1	−1	−1	0
	2	−1	1	1	1
Choix 5	1	−1	1	1	1
	2	1	−1	−1	−1
Choix 6	1	1	1	−1	1
	2	−1	−1	1	−1
Somme		$\sum = 0$	$\sum = 0$	$\sum = 0$	$\sum = 0$

TABLE 3.2 – **Le plan d'expériences orthogonal appliqué lors de l'enquête**

– Prog1 : Modification des pratiques agricoles.
– Prog2 : Amélioration des infrastructures de protection.
– Prog3 : Développement de la communication.

Remarque L'orthogonalité du plan d'expériences est bien satisfaite. Cependant, demander à un individu d'effectuer 24 choix consécutifs ne semble pas envisageable (biais cognitif, contrainte temporelle, etc.). Il s'agit alors de diviser l'ensemble des choix obtenus en plusieurs blocs. Dans notre cas, on divise ces 24 ensembles en 4 blocs de 6 ensembles chacun. L'orthogonalité n'est pas respectée dans chaque bloc, mais l'orthogonalité globale est préservée.

3.2.2 Modélisation

Nous cherchons à expliquer Y à partir des variables socio-économiques. Nous avons choisi le modèle logistique polytomique. Dans cette partie, on présente les modèles obtenus pour chaque bloc et l'interprétation des résultats obtenus. La modélisation est d'abord faite par bloc, puis sur la base de données complète.

X est un vecteur de p variables explicatives, $X = (X_1, \cdots, X_p)$. Nous allons

	Bloc 1		Bloc 2		Bloc 3		Bloc 4		Echantillon total	
Attribut	**Value** (β_j/α_i)	**t**	**Value**	**t**	**Value**	**t**	**Value**	**t**	**Value**	**t**
Chef de famille (oui)	-0.900	-1.060	1.110	2.024	0.509	1.197	0.09525	0.2553	0.194	0.972
Touché (oui)	-0.365	-0.293	9.564	0.263	0.494	0.741	-1.00595	-1.8724	-0.139	-0.426
Maison	2.658	1.770	0.121	0.125	0.525	0.765	0.88502	1.4006	-0.430	-1.345
Locataire	0.169	0.037	0.381	0.552	0.385	0.711	0.85562	0.6120	0.465	1.979
Rurale	1.278	1.006	0.387	0.658	0.148	0.321	0.20434	0.5062	0.128	0.569
Tranche de revenu 2	3.006	1.644			14.507	0.836	2.16147	3.1657		
Tranche de revenu 3	3.065	1.703			0.710	1.007	2.35578	3.1657		
Tranche de revenu 4	2.196	1.128			0.062	0.077	2.58377	3.1128		
Tranche de revenu 5	1.883	0.637			0.202	0.179	1.23971	1.0216		
Tranche de revenu 6	0.642	0.291			0.439	0.576	1.74011	2.3378		
Programme 1	15.946	5.383					3.06383	6.3898		
Programme 2	15.356	5.895	24.024	0.949	13.253	0.520			8.455	8.688
Programme 3			14.507	0.836						
Interaction	20.941	5.544	16.167	0.931			11.634	10.531		
$\chi_{\{Y\in[0,12.5]\}}$	8.805	3.147	13.603	0.374	2.216	2.531	1.6996	2.1482	1.487	3.975
$\chi_{\{Y\in[12.5,25]\}}$	16.147	4.570	34.031	0.815	12.393	0.486	4.9655	5.3564	5.550	11.282
$\chi_{\{Y\in[25,37.5]\}}$	19.510	5.146	70.432	1.687	16.407	0.644	5.3317	5.7098	8.688	14.903
AIC	99.286		126.830		207.402		265.2817		802.790	

TABLE 3.3 – **Modèles complets issus des blocs et de l'échantillon total**

maintenant introduire plusieurs seuils α_1, α_2 et α_3 tels que

$$Y|_{\{X=x\}} = \begin{cases} 0 & si\ Y^* < \alpha_1 \\ 12.5 & si\ \alpha_1 \leqslant Y^* < \alpha_2 \\ 25 & si\ \alpha_2 \leqslant Y^* < \alpha_3 \\ 37.5 & si\ Y^* \geqslant \alpha_3 \end{cases}$$

où $Y^* = \beta x + \epsilon$. On a

$$\text{logit}\mathbb{P}(Y \leqslant j/X = x) = \alpha_j - \beta_1 x_1 - \beta_2 x_2 - \cdots - \beta_p x_p,\ \forall j = 1, \cdots, 4,$$

ou encore

$$\mathbb{P}(Y \leqslant j/X = x) = \frac{\exp\{\alpha_j - \beta_1 x_1 - \beta_2 x_2 - \cdots - \beta_p x_p\}}{1 + \exp\{\alpha_j - \beta_1 x_1 - \beta_2 x_2 - \cdots - \beta_p x_p\}}.$$

Les résultats obtenus sont consignés au sein du tableau 3.3.

Nous remarquons que les modèles issus des quatres blocs et de l'échantillon-mère sont significativement différents. La sélection des modèles est basée sur le critère AIC

3.2.3 Sélection de modèle

Les sous-modèles sont sélectionnés selon le critère AIC aussi bien pour les quatre blocs que pour l'échantillon total. On utilise ici la sélection descendante basée sur le critère AIC. Après la sélection de modèle, nous obtenons des sous-modèles différents par rapport aux variables explicatives. Les informations sont consignées dans le tableau suivant

	Bloc 1		Bloc 2		Échantillon total	
Attribut	**Value**	**t-value**	**Value**	**t-value**	**Value**	**t-value**
Chef de famille (oui)			1.058	2.016		
Touché (oui)			8.794	0.343		
Maison	1.386	1.848				
Locataire					0.3431	1.661
Programme 1	2.112	3.584			7.5059	16.443
Programme 2	2.450	4.295	22.885	1.987	8.4251	8.769
Programme 3			11.096	0.517		
Interaction	4.491	6.179	−12.674	−0.590	11.5908	10.601
$\chi_{\{Y\in[0,12.5]\}}$	−3.978	3.994	12.419	0.485	1.6524	10.3687
$\chi_{\{Y\in[12.5,25]\}}$	2.562	3.402	31.763	2.230	5.7045	15.9125
$\chi_{\{Y\in[25,37.5]\}}$	5.744	5.585	655.596	46.033	8.8312	18.4550
AIC	89.217		121.496		798.2087	

TABLE 3.4 – **Sous modèles selon les blocs et l'échantillon total**

Échantillon	**Modèle**
Bloc 1	CAP → Type de logement + Prog 1 + Prog 2 + Interaction
Bloc 2	CAP → Chef de famille + Être touché + Prog2 + Prog 3 + Interaction
Bloc 3	CAP → Prog 2
Bloc 4	CAP → Revenu + Être touché + Type de logement + Prog 1
Échantillon total	CAP → Locataire + Prog 1 + Prog 2 + Interaction

TABLE 3.5 – **Modèles obtenus selon les blocs et de l'échantillon total**

Les modèles finaux pour chaque bloc et celui de l'échantillon total sont présentés dans le tableau (3.5).

3.2.4 Conclusion

Les quatre blocs ont été constitués de manière à ce que les répondants aient les mêmes caractéristiques (genre, localisation, CSP). Or nous observons que les modèles issus des quatre blocs sont corrects mais que les paramètres estimés de chaque variable entrée dans le modèle sont différents. On aurait pu s'attendre à des modèles quasi-semblables.

On met en évidence le fait que l'échantillonnage ou la méthode de sondage a bien eu un impact sur les résultats. On prend un risque à ne pas contrôler la méthode de sondage inter-blocs. Notons que la méthode des programmes ne propose pas dans son protocole de vérifier que les sous-échantillons ou blocs soient semblables. Cette étude doit être poussée en la menant sur des enquêtes longitudinales et et en tenant compte du plan de sondage.

3.3 La méthode des programmes et le problème du statu quo pour des données longitudinales

Un autre problème soulevé en utilisant la méthode des programmes est le statu quo qui peut être vu comme une réponse de protestation (protest bids). Nous avons essayé d'identifier les enquêtés et les raisons qui les ont amenés à choisir le statu quo. Comme la même enquête a été reproduite, nous avons voulu aussi mettre en évidence si les répondants avaient changé d'avis. Nous avons mis en œuvre le modèle mixte pour des données longitudinales (Vebeke et Molenlebergh 2000), et comparé les modèles issus des données recueillis pour les années 2012 et 2013. Nous concluons qu'il n'y a pas de différence significative entre les deux modèles (Taïbi et *al.* 2014).

3.4 L'évaluation du consentement à payer : méthode des programmes et hétérogénéité des préférences

Afin de tenir compte de l'hétérogénéité non observée dans les préférences des répondants, le modèle Logit à coefficients aléatoire (mixed logit) a été mis en œuvre. En effet ce modèle est plus adapté ici que le modèle logit multinomial classique dans le sens où il permet de mieux comprendre les déterminants des choix des répondants à un questionnaire d'enquête. Pour estimer les paramètres du modèle on utilise des techniques de simulations. McFadden et Train (2000) ont montré que tout modèle de choix discrets découlant de la maximisation d'une utilité stochastique peut être approché, avec toute la précision voulue, par un modèle logit à paramètres aléatoires, sous réserve d'un choix approprié des variables et d'une spécification adéquate de la loi de distribution des paramètres.

La modélisation des données d'enquête du projet Emire 2 (Crastes et *al.* 2014), en appliquant le modèle logit à coefficients aléatoires, nous a permis de mettre en évidence que c'est le développement des infrastructures qui a été plébiscité avec un coût moyen de $16,09e$. Ce coût moyen est de $12,92e$ pour l'amélioration des pratiques agricoles et enfin de $5,40e$ pour le développement de la communication. Cependant, nous constatons une forte hétérogénéité dans les préférences en fonction de la zone où les répondants vivent. Nous retrouvons encore ici le problème exposé dans la partie sur l'estimation du consentement à payer dans le cas d'un espace mesurable et pour un processus faiblement dépendant.

3.5 Modèle du rendement en riz dans la province de Tamatave

La méthode des forêts aléatoires a été appliquée à l'Observatoire de la ruralité dans le cadre du projet Campus Paysan à Madagascar (Taïbi-Hassani et Bezara, 2011).

Dans le projet de développement « Campus Paysan » à Tamatave-Madagascar (2005-2009), il a été reconnu que le volet observatoire de la ruralité représente un des éléments clés du programme (Taïbi et *al.* 2006). La majorité de la population de Tamatave vit principalement de l'agriculture, mais sous forme traditionnelle. Il s'ensuit une faible productivité et un niveau de rendement très bas.

Face à cette situation, l'Université de Tamatave a fait appel à différents partenaires pour l'accompagner à s'engager dans une voie de recherche de solutions et devenir un outil développement local. Dans cette perspective, le projet Campus Paysan a été mis en place et a pour mission de former des paysans (campusards) et de détecter des paysans leaders. Le campus paysan est un espace universitaire délocalisé, pour la formation et la promotion de la ruralité. Il a été initié en 2005 dans le cadre d'une coopération décentralisée entre les Régions Haute-Normandie (France) et Tamatave (Madagascar) en partenariat avec l'Ésitpa. Il s'inscrit résolument dans le cadre des orientations stratégiques du développement rural.

C'est dans ce contexte qu'un dispositif de collecte de données a été mis en place, l'observatoire de la ruralité de Tamatave (ORDT). En effet, il est important de disposer de données exactes et actualisées, tant pour évaluer les progrès et planifier les investissements, que pour assurer l'efficacité des analyses et de la mise en œuvre du projet

(Reyes et Due, 2009). L'analyse des résultats a pour premier objectif de mieux apprécier la situation des paysans et d'avoir une cartographie des productions de la Région et, pour second objectif, de mesurer l'impact du projet. Plus précisément, cet observatoire permettra de mieux connaître les spécificités de la population rurale, d'identifier les déterminants de la productivité rizicole des exploitations agricoles locales, d'élaborer des indicateurs technico-économiques. Enfin, il permettra aussi de tester le modèle développement «Campus paysan», d'adapter le contenu des formations à l'évolution des besoins du milieu rural et d'orienter les activités de production aux demandes des marchés.

Le questionnaire porte sur les critères de différenciation des structures d'exploitation et des niveaux de productivité rizicole ainsi que des degrés de diversification et de valorisation des productions et des conditions socio-économiques d'amélioration des revenus.

Le questionnaire est structuré en plusieurs parties décrites ci-dessous.

1. Situation globale de l'exploitation : main d'œuvre familiale, niveau d'instruction de l'individu enquêté, surface exploitée, assolements, types d'élevages et effectifs, statut des terres... .
2. Critères de productivité rizicole : type de rizière, quantité produite sur une année et niveau d'autosuffisance, capacité d'achat en riz, densité de semis, modes de fertilisation, accès à l'irrigation.
3. Modes de commercialisation : lieux de vente, fréquence des ventes, temps de déplacement sur les lieux de vente.
4. Trajectoire de l'exploitation : destination des productions, motivation au changement, transmission de l'exploitation.
5. Attentes en terme de développement agricole : types de services, proximité et accès, facteurs limitants de la productivité agricole, intérêts des regroupement de paysans.
6. Niveau de richesse : sources de revenus, revenus et dépenses.

Résultats pour la Région d'Atsinanana

Une pré-enquête a été réalisée pour évaluer le dispositif, 12 villages ont été sélectionnés, 180 questionnaires ont été exploités dans la Région d'Analanjirofo et ont permis d'améliorer le questionnaire. Le sondage stratifié à deux degrés a été privilégié. La Région d'Atsinanana compte 84 communes rurales que nous avons classées par taille. Trois classes ont été retenues. Nous avons procédé au tirage au sort de 10 communes dans chaque classe et ensuite tiré au hasard 50 habitants par commune. Au total 1500 exploitations ont été sélectionnées. Seuls 1400 questionnaires ont été totalement exploités pour la suite de l'étude.

Résultats

La méthode des forêts aléatoires a été mise en œuvre sur le logiciel Statistica Data Miner en utilisant un échantillon d'apprentissage pour estimer la règle d'affectation et un échantillon test pour tester le modèle respectivement de taille 1000 et 400. Au total

200 arbres ont été utilisés dans les deux analyses qui suivent. Une centaine de facteurs a été prise en compte.

- Étude sur l'autosuffisance alimentaire

Nous avons essayé d'expliquer les raisons de l'autosuffisance à partir de l'ensemble des données d'enquêtes. La variable « êtes vous autosuffisant en riz ? » est à réponse binaire (oui ou non). Les matrices de confusion sur les échantillons d'apprentissage et test nous renvoient d'excellents taux de classement (86% et 83%). La méthode des forêts aléatoires a permis de dégager les critères qui expliquent le mieux l'autosuffisance alimentaire. En effet, un agriculteur est autosuffisant s'il

- a fait des études et a des possibilités d'échanger avec le monde extérieur,
- entrevoit d'autres possibilités que l'agriculture comme métier pour ses enfants ou pour lui-même, anticipe sur l'avenir,
- vit plutôt en famille et dans un lieu accessible par la route ou au moins proche d'un lieu de vente,
- s'oriente vers une culture de rente comme le poivre,
- s'oriente vers la production vivrière en priorité (riz et maïs),
- utilise une densité de semis élevée,
- évite de mélanger les types différents d'agriculture.

- Étude sur le rendement en riz

En appliquant la méthode des forêts aléatoires pour expliquer le rendement en riz, il en résulte les constats suivants. Le rendement en riz est plus élevé si, par ordre d'importance, le paysan malgache

- utilise une densité de semis élevé,
- vend principalement dans une grande ville (et non pas au village ou au marché),
- a accès aux produits vétérinaires directement au sein de son village,
- ne vit pas loin des lieux de vente,
- maîtrise les techniques d'irrigation.

La matrice de confusion nous donne un taux d'individus bien classés excellent, aussi bien sur l'échantillon d'apprentissage (95%) que sur l'échantillon test (90%).

Conclusion

Cette étude permet de nous éclairer sur les critères à prendre en compte pour améliorer le rendement d'une exploitation. Les résultats de cette enquête ont permis à l'équipe technique du Campus Paysan de mieux adapter les plans de formation et de répondre aux besoins des campusards (rizipisciculture, gestion, irrigation, soins vétérinaires, micro crédits...). L'observatoire de la Ruralité de Tamatave constitue un outil d'analyse de la situation agricole dans la Région et permet de collecter des données pour dégager des indicateurs fiables de développement local et mesurer l'impact du Campus Paysan.

Chapitre 4

Modélisation de données biologiques, agronomiques et physico-chimiques

4.1 Démarche statistique pour la sélection des indicateurs par Random Forests pour la surveillance de la qualité des sols

Nous avons initié et élaboré une démarche d'élaboration d'un indice d'état d'un sol lors de l'étude du projet Bioindicateurs Phase I (Laval et *al.* 2008). Deux sites St Georges et Yvetot ont fait l'objets d'expérimentations. Notons qu'il s'agit de mettre en évidence les indicateurs biologiques les plus discriminants pour l'évaluation d'un état donné. Les prélévements ont été reproduits sur plusieurs saisons Taïbi et *al.* 2014.

Dans le cadre du programme BIO2, le groupe Biomath formé de statisticiens et d'informaticiens partenaires du projet a eu pour objectifs de concevoir la base de données, d'en vérifier la qualité, de proposer des démarches d'analyse des données et d'harmoniser les traitements statistiques pour les groupes biologiques. Cette seconde phase du programme Bioindicateurs utilise les données acquises par plus de 20 laboratoires sur 47 placettes de prélèvement (Pérès *et al.*, 2012).

Compte tenu du nombre très important de données (>200 000), il s'est très vite avéré nécessaire d'harmoniser et de centraliser les résultats obtenus à travers la création d'une base de données accessible de tous. Afin de faciliter l'accès aux données un programme d'automatisation sous Java a été conçu. Ceci a ainsi permis la mise à disposition des résultats pour l'ensemble des équipes en vue de croiser les données obtenues par des laboratoires différents et de faciliter la manipulation et l'analyse des données.

L'objectif des traitements de données est de hiérarchiser les bioindicateurs en fonction de leur sensibilité aux facteurs environnementaux et aux perturbations (contaminations et usage du sol) d'une part, et de proposer un indicateur agrégé de la réponse à ces facteurs d'autre part. Finalement, une démarche de sélection des bioindicateurs intégrant des critères de faisabilité pour la mise en pratique et le développement de ces outils est proposée afin de moduler les résultats précédents par la prise en compte de leur aspect technique et socio-économique.

Nous avons mis en place une méthodologie pour sélectionner une batterie d'indicateurs sensibles à des facteurs tels les pollutions organiques, métalliques mais aussi à des facteurs d'usage, de concentration en MO, en CO. La taille de la base de données, même après un travail de réduction nécessite de mettre en place un data mining, basé sur les Random Forests.

Une enquête utilisateurs a ensuite été lancée pour évaluer l'opérationnalité des bioindicateurs. Une proposition de démarche a été conçue par Biomath en vue d'établir une échelle de transférabilité vers les utilisateurs de sites agricoles ou pollués.

Enfin nous illustrons l'algorithme permettant de construire un outil d'aide à la décision " ModelBio ". La définition de la qualité des sols n'étant pas fixée, la construction d'indices d'état du sol est une alternative. En effet la notion de qualité d'un sol ne fait pas l'unanimité et peut revêtir plusieurs définitions. Elle dépend étroitement de la notion de service fourni. Quatre types de services fournis par les écosystèmes ont été définis dans le MEA (Millenium Ecosystem Assessment) : services d'approvisionnement, services de régulation, services culturels et services de soutien des écosystèmes. Les organismes du sol contribuent à l'ensemble de ces services (Hedde et *al.* 2014).

4.1.1 Synthèse bibliographique sur les indices

Les indices de qualité de l'air ou de l'eau impliquent l'analyse des contaminants spécifiques notamment la connaissance des seuils de toxicité (épidémiologie). Les usages sont moins variés et ne nécessitent pas de prendre en compte des scénarii différents comme pour le sol. Différentes définitions de la qualité du sol ont été établies par Parr et *al.* 1992, Harris et *al.* 1996, Karlen et *al.* 1997 du point de vue de l'agriculture (rendement, ou écologique ou environnemental). Mais on ne retrouve pas dans la littérature une définition universelle vue la complexité du sujet. Ainsi on retrouve dans Sojka et Upchurch (1999), la conclusion suivante " si un terme ne peut être défini indéfiniment c'est qu'il est réellement indéfinissable ". Cependant maintenir la qualité d'un sol est nécessaire pour assurer un environnement viable, durable (Smith et *al.* 1993) malgré une thématique complexe due au climat, au sol, aux plantes, aux facteurs anthropiques et à leurs interactions. Mais il est difficile de quantifier la qualité d'un sol ainsi lorsqu'on effectue des recherches :

1. les mots clefs qualité + sol renvoient à pas moins de 162 618 articles,

2. les mots clefs qualité + sol + indice renvoient à pas moins de 56987 articles,
3. les mots clefs qualité + sol + indice + modèle renvoient à pas moins de 243 articles.

La qualité ou l'état d'un sol peuvent être définis comme le plus petit ensemble de paramètres qui, mis en relation, permet de renseigner sur la capacité d'un sol à avoir une fonction donnée. Un indicateur est une variable mesurable qui influence la capacité d'un sol à avoir une fonction donnée (Acton et Padbury 1993). Toutes les études montent la difficulté d'établir un indice de qualité en raison de la diversité des propriétés physicochimiques, microbiologiques, biologiques et la nécessité de les intégrer pour établir cette qualité (Parr et Papendick 1992), Garcia et *al.* 1994, Halvorson et *al.* 1996)

La connaissance d'un sol de qualité à travers le monde est liée à l'expérience des agriculteurs (expertise) : couleur du sol, texture, rendement, performance, profondeur du sol. Ces indicateurs sont utilisés au Kenya, en Amérique Latine, en Inde... On peut classer les indices en n catégories.

Les indices simples de la qualité des sols

L'indice le plus abondamment utilisé est le quotient qCO^2 (respiration/biomasse microbienne). Cet indice a été acté comme un indicateur de la maturité d'un écosystème (Insam et Hasewander (1989), Anderson et Domsh 1985, Insam et Domsh 1988, Anderson et Domsh 1990). Ce rapport décroît dans les sols où l'on pratique la monoculture par comparaison à des sols ayant subi des rotations de culture. Ce quotient a aussi été utilisé comme un indicateur de l'altération d'un sol dans le cas de contaminations par des métaux, ou suite à une déforestation, ou une variation de température, ou en cas de changement de pratique (Bastida et *al.* 2008, Brookes 1995, Dilly et *al.* 2003, Joergensen et *al.* 1990, Liao et Xiao 2000). Cependant il est non sensible à certaines perturbations (Wardle et Gham 1995). Le rapport Cmicr/Corg total (Anderson et Domsh 1990), est un indice plus sensible que le qCO^2, car la biomasse organique répond plus rapidement au changement que la MO (Powlson et Jenkinson 1981). Jenkinson et Ladd (1981) ont proposé pour des sols cultivés qu'une valeur de 2,2 reflète un bon équilibre entre les 2 fractions de carbone. Le point faible de ces ratios : les changements dus aux modifications de pratiques culturales peuvent être masqués par des facteurs climatiques (Insam et *al.* 1989). On peut citer d'autres indices simples de la qualité d'un sol :

SQI = Soil Quality Index = Somme de chaque composant en liaison avec la biomasse microbienne

IR = Immobilisation Ratio= Somme de tous les composants extraits de la biomasse microbienne sur le total.

Hydrolysing Coefficient (testé en Italie) Perucci (1992)

Metabolic Ratio = Dehydrogenase/ C soluble dans l'eau (Masciandaro et *al.* 1998) donne une information qualitative de la dégradation causée par un usage intensif. Cet indexe a été utilisé pour établir l'effet des pratiques ou des types de cultures (Caravaca et *al.* 2002, Saviozzi et *al.* 2001, Riffaldi et *al.* 2002).

Un indice basé sur les microarthropodes

Le BSQ Biological Soil Quality a été testé dans la vallée de PO (Italie du Nord) sur une parcelle de 200 km^2. Un score est attribué à chaque groupe de microarthropodes,

le BSQ = Somme Scores de chaque groupe.

Le dispositif mis en place consiste en parcelles homogènes avec six répétitions (aléatoires). Les facteurs pente et végétation ont été pris en compte. Le BSQ a été testé afin d'évaluer sa reproductibilité et l'effet spatio-temporel dans un objectif de standardisation et pour fournir des valeurs de référence pour l'indice des sols cultivés d'Italie du Nord.

Pour les neuf exploitations étudiées en utilisant le modèle mixte, le BSQ donne des résultats homogènes. Pas de fluctuations annuelles mais des fluctuations saisonnières (apports d'intrants, températures). Une des difficultés réside en la subjectivité dans l'attribution des scores et la non prise en compte de la grande mobilité des arthropodes (Aspetti et *al.* (2010), Parisi et *al.* (2005), Knoepp et *al.* (2000).

En résumé, les indices simples de la qualité des sols ont pour avantage leur simplicité d'utilisation mais présentent un point faible car on a un manque d'information globale sur l'état du sol, d'où l'intérêt de développer des indices agrégés ou " multiparamétrés ".

Les indices multiparamétrés de la qualité des sols agricoles

Le premier indice a été établi par Karlen et *al.* (1994) et est basé sur l'utilisation de la fonction score. Soil quality = $qwe(wt) + qwma(wt) + qrd(wt) + qfqp(wt)$ avec

- we : Capacité du sol à s'accommoder à l'eau,
- wma : Capacité du sol à faciliter le transfert d'eau et son absorption,
- rd : Capacité du sol à résister à la dégradation,
- fqp : Capacité du sol à assurer la croissance des végétaux,
- wt : Poids associé à chaque fonction du sol.

Son point faible : les poids sont subjectifs et ne sont associés à aucun fondement mathématique. Karlen et *al.* (1994) ont évalué la qualité du sol (maïs) mais on ne trouve que de brèves explications sur la sélection des paramètres. Karlen et *al.* (1994) ont identifié le rôle de l'eau comme important, en utilisant des paramètres physiques, climatiques et biologiques. Ils ont étudié l'effet du labour sur la qualité des sols, soulignant que l'absence de labour affectait la qualité des sols.

L'indice SQI a été ainsi défini (Andrews et *al.* 2002a 2002b)

$SQI = \sum 0,61S_{(OMi)} + 0,61S_{ECi} + 0,16\S_{pHi} + 0,16S_{WSAi} + 0,15S_{Zni} + 0,09S_{BDi}$

- OM : organic matter
- EC : Electrical conductivity
- WSA : water-stable aggregates
- Zn : Zinc
- BD : bulk density
- S : score de la variable

Les coefficients du modèle sont basés sur des analyses multivariées, obtenus en comparant les effets de plusieurs systèmes de culture sur un sol de qualité (dryland). On peut remarquer que ce sont la matière organique et la conductivité électrique qui sont les indicateurs les plus importants dans le calcul de ce score.

On peut citer d'autres indices. L'indice « Sustainable Index » est obtenu à l'aide d'une approche trigonométrique basée sur trois sous indices, nutritionnel, microbiologique et relation sol-plante. Cet indice a été établi à partir d'expérimentations dans le Punjab

(Kang et *al.* (2005)).
Un indice physique de la qualité des sols (riz et blé) basé sur quatre paramètres, la densité réelle, la matière organique, la résistance du sol (à la pénétration racinaire) et les agrégats (densité), en utilisant des régressions multiples établi par Mohanty et *al.* 2007. On peut remarquer qu'il n'a pas été effectué de validation spatiale.

Indices utilisant des activités enzymatiques

Même s'il a été montré que les activités enzymatiques sont sensibles au changement dans le sol (Nannipieri et *al.* 1990), peu d'indices utilisant les activités enzymatiques ont été développés. Puglisi et *al.* (2006) ont proposé trois indices évaluant la dégradation d'un sol en utilisant différentes activités enzymatique (dues aux pratiques culturales) incluant la densité des semis et l'application de la MO (fertilisation) dans différentes localités en Italie. Le premier indice (AI1) est défini à partir de sept activités enzymatiques (Arylsulphatase, gluco, phosphatase, urease, invertase, dehydrogénase, phenoloxidase). Le deuxième indice (AI2) utilise quatre activités (phosphatase, urease, invertase, glu). Le troisième indice (AI3) utilise trois enzymes (glu, phosphatase, uréase).

D'autres indices existent

- L'indice EAN (Enzymatic Activity Number) a été proposé par Beck (1984)
  $EAN = 0,2x(0,15xDH + CA1,25x10/5xP + 4x10/2xPR + 6x10/4Xam$,
- L'indice BIF (Biological Index of Fertility), Stefanic et *al.* (1984)
  $BIF = (1,5xDH + kx100xCA)/2$,
- Le $BISF$ (Biological Index of soil Fertility) proposé par (Koper et Piotrowka 2003),
  $BISF = C(Carbone)+N(Azote)+DH(Dehydro)+P(Phosphatase)+PR(Protease)+AM(Amylase)$

Indices basés sur la faune du sol

L'Indice Biotique de Qualité des Sols, l'IBQS reprend la procédure de calcul de l' IBGN (indice la qualité de l'eau et se calcule en fonction de l'abondance moyenne des taxons indicateurs présents dans les échantillons (Di) et du pouvoir indicateur (Si) des taxons indicateurs. Les taxons indicateurs sont au nombre de 22, ont été choisis en fonction de leur sensibilité aux perturbations. En fonction de celle-ci, un pouvoir indicateur leur a été attribué. Celui-ci est d'autant plus petit que le taxon est sensible aux perturbations

Indices de la qualité de l'eau

L'indice IBCH

L'IBCH (OFEV 2010) reprend in extenso les directives de la norme française de l'indice biologique global normalisé ou IBGN pour le calcul de l'indice. L'IBGN permet d'évaluer la qualité d'un cours d'eau à partir d'analyses sur les macro-invertébrés benthiques. L'IBCH est déterminé à partir du tableau de détermination comprenant les 9 groupes faunistiques indicateurs (GI) en ordonnées et en abscisses les 14 classes de variétés taxonomiques (VT) : $IBCH = GI + VT - 1$ avec $IBCH < 21$. L'IBCH

atteint une valeur maximale de 20. En l'absence significative de taxons indicateurs (3 ou 10 individus), la note IBCH est égale à zéro.

Affectation d'un tronçon de cours d'eau à une des cinq classes de qualité en fonction du score obtenu pour l'IBCH.

L'indice Seq-eau

Le Système d'Évaluation de la Qualité des Eaux (SEQ-Eau) est l'outil national d'évaluation de la qualité des eaux des cours d'eaux.Il est utilisé depuis les années 2000. Les outils d'évaluation ont été construits de façon modulaire et adaptable aux évolutions scientifiques et techniques ainsi qu'aux spécificités régionales. Les altérations de la qualité sont traduites en indices de qualité exprimées sur une échelle de 0 à 100. Cette échelle est subdivisée en cinq classes aux fins de représentation cartographique de la qualité. Pour chaque variable des classes de qualité sont définies en fonction de sa valeur.

Indices de la qualité de l'air

- L'indice ATMO

L'indice ATMO est basé sur les niveaux de dioxyde de soufre, dioxyde d'azote, ozone et particules fines. L'indice ATMO est le plus grand des quatre sous indices : un sous-indice est donné à chacun de ces quatre polluants. Les sous indices sont calculés à partir de recommandations sur les concentrations des différentes substances dans l'air (Air-Normand).

Les lichens comme bioindicateur

L'utilisation de ce bio-indicateur pour mesurer la qualité de l'air est plus facile et moins couteuse à mettre en œuvre. La méthode la plus utilisée est celle mise au point en 1986 par Van Haluwyn et Lerond. À partir de lichens très représentatifs et faciles à identifier, ces chercheurs ont établi une échelle de correspondance entre groupements épiphytes et degré de pollution. C'est l'échelle de toxitolérance de Van Haluwyn-Lerond, graduée de l'indice A (la plus forte pollution) à l'indice G (l'air le plus sain). Cette échelle a été établie à la suite d'une étude poussée sur les lichens et ses différentes espèces.

Conclusion

On retrouve une abondance d'indices simples et agrégés mais les phases d'apprentissage sont nombreuses et les étapes de tests sont rares. Les paramètres calculés sont assez souvent liés aux résultats de l'expérimentation même, du protocole, du contexte pédoclimatique. Les indicateurs sensibles au climat ne devraient être utilisés qu'à la même saison. Dans les divers articles, la reproductibilité des expériences est citée mais le nombre d'unités est assez limité même si dans certains cas les étapes de validation pour les indices agrégés standardisés sont en cours. Les facteurs devant être considérés comme aléatoires sont considérés comme fixes et l'on ne tient pas compte de la hiérarchie des différents facteurs exogènes. Les indices de la qualité du sol sont assez souvent inspirés des indices de la qualité de l'air ou de l'eau comme par exemple l'IBCH.

4.1.2 Constructon d'un indice multiparamétré de la qualité des sols

Au cours des deux phases du projet Bioindicateurs (BIO 1 et BIO 2) de la qualité des sols soutenu par l'Ademe nous avons, d'une part, élaboré une démarche de sélection des indicateurs et de biomarqueurs les plus discriminants pour la surveillance de la qualité des sols et l'évaluation des risques et, d'autre part, proposé une construction d'un indice d'état d'un sol.

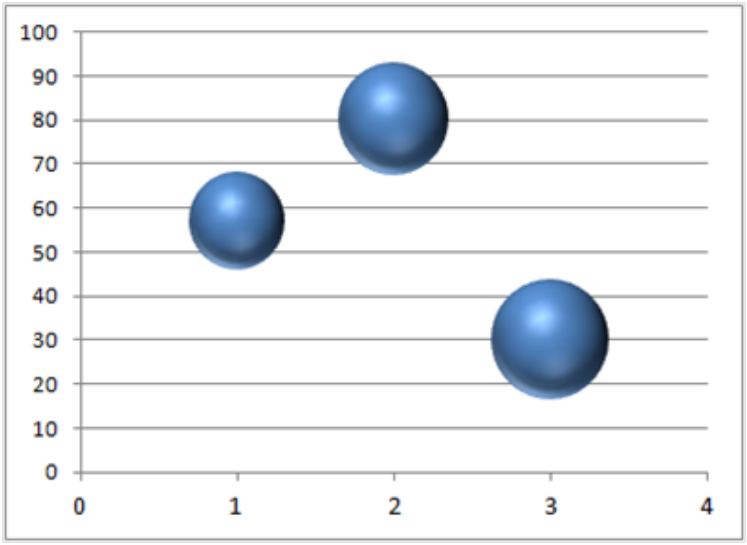
Le volume des données définies dans le programme Bioindicateurs 2 (Ademe) et la grande batterie de variables biologiques à tester (une centaine) nécessitent des techniques d'analyse telles que les Random Forests. En effet ces méthodes peuvent s'affranchir du problème de multi-colinéarité pour la sélection d'indicateurs sensibles aux différents facteurs étudiés. La sélection des variables les plus discriminantes a été éffectuée par la méthode des Random Forests. Ainsi nous avons recherché la meilleure sélection en étudiant l'ensemble des variables biologiques appartenant aux compartiments Microflore et Faune. Cette démarche a porté sur l'ensemble de ces indicateurs. Ces travaux ont été mis en œuvre sur les trois facteurs de discrimination : l'usage des sols, les niveaux de contamination en ETM, et les niveaux de contamination en polluants organiques.

Nous avons ensuite regroupé les variables les plus discriminantes issues de chaque analyse par RF. Une analyse discriminante linéaire a ensuite été mise en œuvre pour chaque facteur en vue d'élaborer un modèle prédictif.

L'objectif des traitements de données est de hiérarchiser les bioindicateurs en fonction de leur sensibilité aux facteurs environnementaux et aux perturbations (contaminations et usage du sol) d'une part, et de proposer un indicateur agrégé de la réponse à ces facteurs d'autre part. La méthode des Random Forest (Breiman 2001) trouve déjà de nombreuses applications dans différents domaines telles que l'écologie (Brosteaux, 2005) ou l'agriculture (Cutler et *al.*, 2007).

Vers un indice global

La synthèse bibliographique nous montre bien qu'un indice universel unique et simple de la qualité des sols n'existe pas. Au-delà de la capacité d'un indicateur à distinguer des différences entre sites contrastées, de se comporter différemment selon des usages du sol ou encore de donner des informations sur l'état d'un sol, il est aussi important que celui-ci soit opérationnel et réponde aux critères des décideurs (Ritz et al , 2008). Un bioindicateur peut donc être caractérisé par deux scores : l'un de sensibilité à un facteur et l'autre d'opérationnalité. Deux dimensions permettent donc de mieux classer les différents bioindicateurs. Le graphique suivant permet de visualiser les scores de sensibilité et de faisabilité. La taille de la bulle permet d'identifier le score de sensibilité.



Nous avons décrit une manière de calculer les scores de sensibilité et d'opérationnalité d'un indicateur (Taïbi et *al.* (2012), Taïbi et *al.* 2013). Nous avons aussi obtenu une sélection d'indicateurs permettant d'établir un modèle explicatif et prédictif de l'état d'un sol (Taïbi et *al.* 2014). Nos travaux ont porté sur deux échelles de contamination : organique et métallique.

- Le score de sensibilité à une source de contamination : hiérarchie des indicateurs

Afin d'évaluer la réponse des 74 variables biologiques aux facteurs, contamination métallique, contamination organique nous avons appliqué le test de Kruskal-Wallis ou Mann-Whitney. Notons que les seuils des facteurs d'explication "'Métalliques" et "Organiques" sont déduits en combinant les valeurs des vibrisses internes et externes des différents éléments métalliques d'après les données du RMQS (Villaneau et *al.*, 2008) et les données INRA-ASPITET pour l'arsenic (Baize, 2000). Les analyses sur les vibrisses et travaux de classification en utilisant l'ultramétrique de Ward ont permis finalement d'assigner les 47 modalités à une classe de contamination qu'elle soit organique ou métallique. Les tests sont réalisés sur le jeu de données propre à la problématique abordée, les sites pollués et les sites agricoles sont donc étudiés indépendamment. Les variables dites de bioaccumulation et d'exposition aux métaux sont testées sur sites pollués seulement.

Le score d'opérationnalité

Dans une optique de développement, les résultats précédents de sensibilité des indicateurs biologiques nécessitent d'être replacés dans un contexte de faisabilité technique d'une part, mais aussi de compréhension par le public des utilisateurs. Il a donc été décidé d'établir des scores permettant d'apprécier l'importance de ces critères dans le choix de l'utilisation de l'outil. Une première enquête a été menée auprès des responsables d'indicateurs. Le questionnaire est structuré en deux grandes parties rassemblant des critères techniques et socio-économiques

– critères Techniques : méthode normalisée, protocole publié, protocole globalement simple et transférable, outils mis en œuvre entièrement in situ, restrictions d'utilisation selon les périodes de l'année,...
– critères économiques et sociétaux : durée de l'échantillonnage, coût des consommables pour une analyse, perception simple par un public non spécialisé, durée de la phase d'interprétation et de remise des résultats,...

Score_{tech} = Somme des notes C_{tech} x W_{tech}
$\text{Score}_{eco-socio}$ = Somme des notes $\text{C}_{eco-socio}$ x $\text{W}_{eco-socio}$

Le dépouillement des questionnaires retournés par chaque responsable d'indicateur nous a permis de calculer ces scores.

Le score global

Le score attribué à la sensibilité étant primordial relativement aux deux autres scores, peut être déterminé ainsi :

Score global = aScore de sensibilité x (bS_{tech} + $\text{cScore}_{eco-socio}$)

a, b, c sont des poids attribués aux différentes composantes.

En conclusion, par une approche pluridisciplinaire, les travaux de recherche " mono indicateur " permettent d'établir les valeurs de références, de mettre en évidence la sensibilité à un environnement donné (pratiques agricoles et usages/stress chimique).

L'approche globale ainsi que la démarche pour la sélection d'une batterie d'indicateurs explicatifs et discriminants pour un stress donné à l'aide des méthodes supervisées telle que la méthode des forêts aléatoires (Laroutis et Taïbi, 2011) donnent des résultats probants dans des contextes d'évaluation de l'état du sol à la fois complexe et dynamique.

Le sondage des utilisateurs versus experts apporte une plus-value pour le choix d'un indicateur dans le sens où le critère de sensibilité est certes nécessaire mais non suffisant, et ceci doit être complété par des études confirmant son opérationnalité. Finalement la complémentarité entre réponses à dire d'experts, celles des utilisateurs et les résultats fondés sur les analyses statistiques permettent de calculer un score final pour chaque indicateur.

Nous projetons de construire un indice fonctionnel d'état d'un sol à l'aide d'une démarche établie qui tienne compte de l'intégration des propriétés physicochimiques, microbiologiques, biologiques nécessaire pour établir cet indice (Taïbi et *al.*, 2011 & 2012). Cet indice peut être défini comme le plus petit ensemble de paramètres, qui, mis en relation, permet de renseigner sur la capacité d'un sol à avoir une fonction donnée.

4.2 Modèle de transfert des métaux lourds du sol vers la plante

Les travaux et résultats de cette partie découlent du projet ETlin (Bury et *al.* 2005, Muntean et *al.* 2007). Plusieurs partenaires ont contribué à ce projet. Les protocoles expérimentaux et les analyses physico-chimiques ont été élaborés par Marc Legras (Biosol-Esitpa) et Frédéric Giro (Institut Technique du Lin).

Deux sites d'essais ont été étudiés pour leurs caractéristiques pédologiques et pour leurs conditions d'apport de matières organiques. Le site de Gamaches dans l'Eure et le site d'Herblay dans le Val d'Oise. La partie polluée de ce dernier site a reçu par irrigation via une buse des eaux usées du début du XXème siècle jusqu'en 1999. La

partie témoin correspond à une pratique culturale raisonnée. Le site de Gamaches a reçu des boues d'épuration jusqu'en 1998. Ces sites vont permettre de cultiver deux variétés de lin oléagineux : A et B.

Chaque site est divisé en deux : une partie permet la culture de la variété A et une autre la variété B. On effectue des prélèvements du sol sur trois horizons ($0-25$ cm, $25-50$ cm et $50-75$ cm de profondeur), et cela durant les quatre stades de la croissance du lin : semis, croissance, floraison, maturation. Des prélèvements ont aussi été pratiqués sur les végétaux. Les teneurs des 5 métaux : Cd, Cu, Ni , Pb et Zn ont été mesurés. En premier lieu une étude des sols s'impose pour une meilleure compréhension de la dynamique des éléments traces dans les sols. Nous nous sommes contentés ici d'étudier les transferts pour le cadmium et le nickel.

Nous définissons le coefficient de transfert pour tous les métaux comme la teneur d'Eléments Traces (ET) absorbé par l'organe (la graine) par rapport à la teneur totale en ET dans le sol au moment du semis. Pour établir le modèle on suppose que les métaux lourds se diffusent à travers la plante dans un sens unique en partant du sol. Nous supposerons de plus que tous les coefficients de transfert des métaux lourds restent constants durant le temps. Le cas de la canne à sucre a été étudié par Da Silva et *al.*(2000).

$$
\begin{aligned}
\frac{dS1}{dt} &= -\lambda_1 S1 - \alpha_1 S1 \\
\frac{dS2}{dt} &= -\lambda_2 S2 - \alpha_2 S2 \\
\frac{dS3}{dt} &= -\lambda_3 S3 - \alpha_3 S3 \\
\frac{dR}{dt} &= \sum_{i=1}^{3} \alpha_i S3 - \beta R \\
\frac{dT}{dt} &= \beta R - \gamma T \\
\frac{dG}{dt} &= \gamma T
\end{aligned}
$$

Avec :

S_1 = Concentration de métal lourd à l'horizon (0cm -25cm) à l'instant t,
S_2 = Concentration de métal lourd à l'horizon (25cm-50cm) à l'instant t,
S_3 = Concentration de métal lourd à l'horizon (50cm-75cm) à l'instant t,
R = Concentration de métal lourd dans la racine à l'instant t,
T = Concentration de métal lourd dans tige à l'instant t,
G = Concentration de métal lourd dans la graine à l'instant t,
λ_1 = Vitesse de lessivage du sol pour la première profondeur,
λ_2 = Vitesse de lessivage du sol pour la deuxième profondeur,
λ_3 = Vitesse de lessivage du sol pour la troisième profondeur,
α_1 = Coefficient de transfert du métal de la première profondeur à la racine,
α_2 = Coefficient de transfert du métal de la deuxième profondeur à la racine,
α_3 = Coefficient de transfert du métal la troisième profondeur à la racine,
β = Coefficient de transfert du métal de la racine à la tige,
γ = Coefficient de transfert du métal de la tige à la graine,
t_g = Début du stade maturation.

et les conditions initiales :

$C_{1,0}$ =Concentration initiale du métal dans la première profondeur.
$C_{2,0}$ =Concentration initiale du métal dans la deuxième profondeur.
$C_{3,0}$ = Concentration initiale du métal dans la troisième profondeur.

Les trois premières équations différentielles ont été réduites à une seule et, de plus, on a observé que la teneur en métal dans le sol est une fonction affine du temps.
La résolution du système donne la solution suivante, en considérant les conditons initiales

$S_0 = C_0$, $R(0) = 0$, $T(0) = 0$ et $G_{t_g} = 0$ avec
C_0 =concentration initiale du métal dans le sol.

La résolution du système d' équations différentielles et l'utilisation de la régression linéaire nous donne dans le cas du Cadmium par exemple :

$$S(t) = \exp^{-7.6*10^{-3}x^2 - 6.3*10^{-5}x}$$

Pour le Cuivre et le Zinc et dans le cas de pratiques culturales raisonnées, la quantité de métaux lourds dans la graine peut être estimée par un modèle linéaire simple : $G = \alpha Sol$.
Nous avons classifié les terrains en deux catégories, les terrains de pratiques culturales raisonnées et les terrains pollués. Les résultats montrent que la plante réagit différemment suivant les trois familles de métaux : Cu-Zn, Ni-Pb et Cd., et ce quelle que soit la variété.
Enfin, pour les seuils de risque, les répercussions des quantités apportées par les graines de lin sont tout à fait négligeables au regard des normes établies (Taïbi et Lemaire 2005). Des recherches doivent continuer dans ce sens afin d'expliquer au mieux la mobilité de certains métaux tels que le Nickel en prenant en compte les interactions entre métaux, texture du sol et facteurs physico-chimiques.

4.3 Concentrations de polluants dans le sol

L'intensification de l'agriculture a soulevé la question de la durabilité à long terme des agroécosytèmes. Les pratiques en matière de non-labour sont présentées comme solutions de rechange. Elles apparaissent comme de meilleures pratiques pour la conservation du sol en permettant la séquestration de carbone et de réduire les gaz à effets de serre (Balesdent et *al.*, 2000). L'étude, effectuée en collaboration avec l'Inra, porte sur l'effet du non-labour sur le sol à partir de cinq activités enzymatiques (Ghanem et *al.* 2007). La pollution des sols par les métaux lourds et les composés organiques peut se produire lorsque les boues d'épuration sont utilisées comme engrais. Il est alors essentiel de définir la nature et les teneurs en polluants contenus dans les boues d'épuration afin d'évaluer les risques environnementaux.

Les boues d'épuration ont été échantillonnées à partir de trois eaux usées urbaines (usines de traitement) et une unité de compostage dans la région de Versailles (France). Les résultats d'un suivi d'un an d'herbicides (glyphosate, diuron et atrazine) et leurs principaux produits dégradés ont été analysés. Les concentrations de ces composés ont

été déterminées. Nous avons démontré la présence de glyphosate et de l'acide aminométhylphosphonique au kg^{-1} (matière sèche) pour tous les échantillons. Le diuron a été détecté à un niveau de g^{-1} , alors que ses composés et les triazines sont inférieures aux limites de quantification. Les quantités de nonylphénol étaient supérieures à la valeur seuil européenne prévue qui est de $50kg^{-1}$.

4.4 Modélisation de données agronomiques

4.4.1 Hétérogénéité intraparcellaires en Agriculture de précision

Le contexte pédoclimatique de Haute-Normandie et l'hétérogénéité de certaines parcelles ont amené un groupe d'agriculteurs à mettre en place des essais d'agriculture de précision pour tester les gains économiques et environnementaux de cette technique. Des modèles ont permis de déterminer les optimums techniques d'apports d'azote en sol superficiel et en sol profond qui ont été comparés à des optimums économiques. L'hypothèse économique choisie ici est que 50 unités d'azote doivent générer un gain de rendement de trois quintaux. Les résultats montrent que dans ces conditions on peut potentiellement diminuer les conseils de fertilisation de 20 unités pour le blé et de plus de 35 pour le colza. Les premiers résultats de la comparaison des optimums techniques et économiques obtenus permettent de diminuer les conseils en fertilisation azotée d'environ 20 unités en blé (12 €/ha) et 40 unités en colza (24 €/ha). Enfin, nous avons évalué les conséquences technico-économiques de l'utilisation de l'agriculture de précision dans cette région, en faisant varier des hypothèses économiques sur le prix de l'azote et du blé : l'augmentation du coût de l'azote ou une diminution du prix de vente du blé iront dans le sens d'une réduction plus grande des doses à apporter. Les modèles utilisés dans cette étude ne permettent pas de simuler des scénarios climatiques en fonction des types de sols, comme seraient susceptibles de le faire des modèles mécanistes (Brisson et *al.*, 1998). Le prolongement de ces travaux est de tester sur ce jeu de données le modèle STICS (Simulateur multidisciplinaire pour les cultures standard) qui modélise, à l'échelle de la parcelle, le développement d'une culture, en fonction de paramètres agronomiques ou d'autres intrants (Brisson *al.* 1998).

Les résultats montrent que l'hétérogénéité de certaines parcelles prouve l'intérêt économique et environnemental de moduler les apports d'azote dans les rotations culturales incluant du blé et du colza en Haute-Normandie. Ainsi, la modélisation a permis d'évaluer le potentiel de réduction de fertilisation azotée, ainsi que les gains économiques et environnementaux de ces techniques dans les conditions pédoclimatiques de production des quatre années étudiées. Cependant la variabilité des résultats notamment en blé nécessite de continuer cette étude pour valider les premiers résultats obtenus. Pour simuler d'autres conditions pédoclimatiques et étendre les références obtenues à l'ensemble de la région, il serait intéressant d'utiliser des modèles mécanistes pour déterminer les optimums techniques.

Chapitre 5

Animation et responsabilités en recherche

5.1 Animation du laboratoire Lamsad

J'ai assuré la responsabilité du laboratoire Lamsad de 2001 à 2013. J'ai participé à la conception du projet Campus Paysan, coordonné les projets de recherche Bioindicateurs de la qualité des sols, assuré la responsabilité scientifique des programmes Emire 1 et 2 et participé à d'autres projets tels que le Projet et Freins de l'Agriculture intégrée et le projet ETlin. Je fais partie du Projet Unitwin Unesco. Ces projets seront présentés en détail ci-après. J'ai recruté et animé une équipe de chercheurs, de doctorants, de post-doctorants, de stagiaires de master recherche ou en projets de fin d'études d'ingénieurs.

5.2 Conference invitée,présidence de séance, expertise

J'ai été invitée à deux congrès en qualité de Chairman. Le premier congrès, ACE (Association Canadienne d'Économie) s'est tenu à Québec en mai 2009 et le second à Varsovie en Juin 2014, 11th International Science Conference on Global Problems of Agriculture, forestry and food Economy.

J'ai été experte pour les journaux suivants : Applied Statistics, La majorité des articles que j'ai expertisés ont pour thèmes les méthodes de prédiction, les modèles non paramétriques tels que les Forêts aléatoires, l'analyse discriminante, ou l'estimation non paramétrique.

5.3 Les projets de recherche

5.3.1 Les projets de recherche régionaux

Les programmes EMIRE I et EMIRE II

J'ai assuré conjointement avec D.Laroutis la responsabilité scientifique du projet EMIRE, projet d'un des Grands Réseaux de Recheche Régionaux, VASI. Ce programme

de recherche a pour objectif de structurer des actions scientifiques régionales pluridisciplinaires pour répondre à une problématique agronomique d'intérêt régional et permettant une valorisation académique de premier plan. Cette problématique agronomique fait référence aux impacts du ruissellement érosif sur la société.

Pour Jayasuriya (2003), l'érosion des sols est peut-être l'une des plus sérieuses formes de dégradation des territoires à travers le monde. Les problèmes d'érosion et de dégradation des sols sont aujourd'hui de plus en plus au centre des discussions en raison du déclin induit de la production agricole (Dale et Polasky, 2007). Face au problème d'érosion, les agriculteurs et les fermiers sont de plus en plus mis à contribution afin de réduire le taux d'érosion. Récemment les pratiques culturales orientées vers la préservation de l'environnement se développent fortement (Dale et Polasky, 2007 et font l'objet pour certaines d'un label « Agriculture Conservation » par la Food and Organization of the United State Nation (NOAA) et la European Conservation Agriculture Federation (ECAF). Deux principales causes de l'érosion peuvent être mises en avant (Jayasuriya, 2003) : des causes abiotiques (conséquences dues par exemple aux vents et à l'eau) et des causes biotiques (conséquences dues entre autres à l'activité anthropique). Le développement anthropique à travers l'urbanisation des territoires et la croissance des terres agricoles favorise aujourd'hui grandement les phénomènes de ruissellement érosif. Ce phénomène particulier, qui se caractérise par une érosion due à l'action de l'eau, génère de nombreuses externalités négatives pour la société à travers les inondations ou les éboulements. Le ruissellement érosif est particulièrement important dans le département de la Seine-Maritime et notamment dans le bassin versant de la Vallée du Commerce. Ces territoires sont reconnus comme sensibles au ruissellement et à l'érosion en raison de la topographie, de la nature des sols et de la pluviométrie. De plus, l'évolution de l'agriculture a modifié la structure paysagère, entraînant le développement des phénomènes de ruissellement érosifs (Ouvry, 1992). En effet, chaque année, sur une parcelle cultivée, 1 à 10 % des précipitations hivernales ruissellent, ce qui conduit à de forts impacts sur la société à travers les coulées de boues et les inondations. Les événements de type inondation se sont multipliés depuis le milieu du XXème siècle, et ont toujours des conséquences importantes pour la société, que ce soit par le coût des ouvrages à réaliser pour réduire les externalités négatives ou par le fait que des vies humaines soient également en jeu. Le coût des ouvrages de protection est d'autant plus élevé que le phénomène est généralisé sur l'ensemble du territoire et concerne un grand nombre d'acteurs (agriculteurs, industriels et habitants). Face aux désagréments causés, les politiques évoluent et les collectivités sont à la recherche de solutions préventives et curatives.

L'objectif du projet EMIRE est précisément d'apporter une réponse en combinant les apports économiques (méthode des programmes), géographiques (géolocalisation) et statistiques (traitement de l'information). Par la méthode des programmes et en associant les apports en statistique et en géographie, nous souhaitons identifier les actions prioritaires pour les habitants afin de réduire le risque de ruissellement et d'évaluer les effets marginaux monétaires de ces différentes actions.

Plus globalement, dans le cadre du projet structurant EMIRE, l'objectif est

- d'analyser le comportement des individus habitant la Vallée du Commerce par rapport aux risques de ruissellement et d'inondation ;
- d'identifier le niveau de perception du risque par les habitants et leur prise de

conscience des efforts réalisés pour réduire ces risques ;
– de quantifier monétairement l'impact de ces phénomènes sur la société dans son ensemble.

Le programme Freins et Leviers de l'Agriculture Intégrée

Ce projet, porté par la Chambre Régionale d'Agriculture de Normandie et en partenariat avec l'Esitpa, les Chambres d'Agriculture du Calvados, de la Manche, de l'Orne et de l'Eure s'est déroulé entre Juin 2011 et Décembre 2012. L'objectif de ce projet est d'identifier les freins liés au conseil et à l'évolution des pratiques sur le thème de l'agriculture intégrée, auprès des conseillers agricoles, des agriculteurs et des responsables des filières bas-normands. Au travers de ce projet, la thématique de la réduction des intrants et de l'adaptation des systèmes de production sont mis en avant. J'ai participé aux conceptions des deux enquêtes par entretiens et quantitative ainsi qu'à la démarche d'analyse des résultats et du recueil d'informations pour la mise en évidence des leviers.

5.3.2 Les projets de recherche nationaux

Les programmes Biondicateurs Phase I et II

L'ADEME étant notamment missionnée sur la prévention de la pollution des sols, la gestion des sites et sols pollués et l'évaluation des impacts environnementaux liés aux retombées atmosphériques et à la gestion biologique des déchets, elle souhaite promouvoir le développement de bioindicateurs destinés à

(a)la surveillance de la qualité des sols pour laquelle des bioindicateurs simples à mesurer, peu chers et répétables sont nécessaires.

(b)la caractérisation approfondie de l'état biologique des sols pour laquelle des techniques plus complexes pourront être mises en oeuvre.

(c)l'évaluation détaillée des risques pour les écosystèmes sur les sites pollués pour laquelle une méthodologie plus complète est nécessaire notamment afin de relier les effets observés à des niveaux d'exposition.

J'ai participé à la phase 1 du projet Bioindicateurs de la qualité des sols, puis j'ai assuré la responsabilité du groupe du groupe Biomath lors de la seconde phase du programme BIO2. La seconde phase du programme Bioindicateurs utilise les résultats acquis par plus de 20 laboratoires sur 47 placettes de prélèvement (Pérès *et al.*, 2012). Compte tenu du nombre très important de données (>200 000), il s'est très vite avéré nécessaire d'harmoniser et de centraliser les résultats obtenus à travers la création d'une base de données accessible de tous. Celle-ci a ainsi permis la mise à disposition des résultats pour l'ensemble des équipes en vue de croiser les données obtenues par des laboratoires différents et de faciliter la manipulation et l'analyse des données. L'objectif des traitements de données exposés ici est de hiérarchiser les bioindicateurs en fonction de leur sensibilité aux facteurs environnementaux et aux perturbations (contaminations et usage du sol) d'une part, et de proposer un indicateur agrégé de la réponse à ces facteurs d'autre part. Finalement, une démarche de sélection des bioindicateurs intégrant

des critères de faisabilité pour la mise en pratique et le développement de ces outils est proposée afin de moduler les résultats précédents par la prise en compte de leur aspect technique et socio-économique.

Les objectifs généraux de ce programme sont de fournir aux secteurs économiques et aux acteurs publics de nouveaux outils de surveillance, de caractérisation et d'évaluation des risques basés sur les propriétés biologiques du sol.Le groupe Biomath a proposé une méthodologie statistique pour atteindre ces objectifs.

5.3.3 Les projets de recherche internationaux

Le projet Campus Paysan

Conscient de la dégradation continue de la ruralité ainsi que de l'appauvrissement croissant de la population, le système de l'Enseignement Supérieur et de la Recherche Malgache, se veut être attentif pour développer et mettre en œuvre des actions de lutte contre l'insuffisance alimentaire. Le mécanisme universitaire cherche à être apte à répondre aux besoins et attentes des agents de développement local. L'Université de Tamatave, principal acteur malgache avec l'Esitpa et soutenus par la Région Haute-Normandie ont mis en place un projet de développement durable " Campus Paysan ".
Ainsi le Campus Paysan est un espace de promotion de la ruralité, terrain d'échanges, de communication et de formation. Les acteurs concernés sont l'Université de Toamasina (Madagascar), le Lamsad-Esitpa-Rouen, le Conseil Régional Haute Normandie, le Ministère de l'éducation Nationale et de la Recherche Scientifique Malgache, MadSup (Coopération Française), FOFIFA à Madagascar, avec la

1. mise en place d'un laboratoire pluridisciplinaire à l'Université de Tamatave,
2. réalisation d'une base de données visant à la mise en place d'un observatoire de la vie paysanne,
3. participation à la conception de cycles de formation professionnelle adaptés en agriculture.

Le projet Unitwin

UNITWIN est l'abréviation de « University Twinning and Networking » (Système de jumelage et de mise en Réseaux des universités). Ce programme de l'UNESCO a été mis en place en 1992, conformément à une résolution adoptée par la Conférence générale de l'UNESCO à sa 26ème session. Le Programme UNITWIN/Chaires UNESCO se caractérise par la création de Chaires UNESCO et de Réseaux UNITWIN dans les institutions d'enseignement supérieur.

Ce programme de l'UNESCO est l'un des instruments privilégiés du renforcement des capacités des institutions d'enseignement supérieur et de recherche par la mise en commun et le transfert des connaissances dans un esprit de solidarité internationale. Il promeut ainsi la coopération Nord-Sud, Sud-Sud et la coopération triangulaire comme stratégie de développement des institutions. Ces institutions agissent en partenariat avec les ONG, fondations et organisations des secteurs public et privé, jouant un rôle important dans le domaine de l'enseignement supérieur. Le Programme

UNITWIN/Chaires UNESCO offre à la communauté de l'enseignement supérieur et de la recherche la possibilité de s'associer à l'action de l'UNESCO, de contribuer à la mise en œuvre de son programme et de ceux du Millénaire pour le Développement (OMD).

5.3.4 L'encadrement en recherche

L'encadrement de l'équipe du Lamsad

L'Esitpa ayant souhaité structurer sa recherche en 2001, le laboratoire de modélisation statistique a été créé en 2002. J'ai assuré la reponsabilité de ce laboratoire jusqu'à Avril 2013. J'ai animé et encadré l'équipe constituée de Patrice Lepelletier (Statistique), Jérôme Dantan(Informatique) enseignants-chercheurs, de chercheurs en post-doctorat, de doctorants et de stagiaires en master ou en stage ingénieur.

L'encadrement de post-doctorants

1. Sorin Muntean, Docteur 3 ème cycle de l'Université des Sciences Agricoles et Médecine Vétérinaire de Cluj-Napoca (Roumanie) et Assistant-Professeur à la Chaire de Phytotechnie, a bénéficié d'un post-doctorat dans le cadre des bourses affectées par la Région Haute Normandie aux chercheurs étrangers. Il a effectué un séjour post-doctoral de 9 mois au sein du Lamsad de septembre 2004 à Juin 2005. Je l'ai encadré pour des travaux de recherche relatifs au projet ETlin et pour établir les bases d'un modèle prévisionnel de phyto-disponibilité des métaux lourds dans le lin(Muntean et *al.* 2006).
2. Manassé Bezara, enseignant-chercheur de l'Université de Tamatave, avait effectué un séjour post-doctoral entre Janvier 2005 et Mai 2005. Les résulats ont débouché sur la conception d'un modèle de développement pour la Province de Tamatave ,le projet Campus Paysan (Taïbi et *al.* 2012, Taïbi et Bezara 2011, Taïbi et *al.* 2010, Taïbi et *al.* 2007).
3. Dans le cadre du projet EMIRE, j'ai encadré Ouerdia Arkoun, Docteure en mathématiques de l'Université de Rouen durant 6 mois en 2012. Ses travaux (Arkoun et al 2012, Crastes et *al.* 2014) portent sur la problématique du statu quo et les biais liés à la mise en place des expériences.
4. Pour la seconde phase du projet Bioindicateurs, j'ai encadré Jeanne Bodin, Docteure en Écologie Forestière, et qui a effectué une thèse en cotutelle avec l'Université Henri Poincaré, Nancy I et l'Université Leibniz de Hanovre (Allemagne). Son post-doc d'une durée d'une année a été cofinancé par l'ADEME (Février 2012-Mars 2013). Elle a assuré des travaux de recherche sous ma reponsabilité et en collaboration avec un groupe de travail composé de partenaires du projet (Bodin et *al.* 2013, Taïbi et *al.* 2012).

L'encadrement de doctorants

- Encadrement en thèse de doctorat de Saturnin Adigaw-E-Touck

J'ai coencadré Saturnin Adigaw-E-Touck en thèse. Il a soutenu sa thèse, intitulée "Modèles non paramétriques de survie pour données incomplètes", le 11 Janvier 2013.

Les travaux de cette thèse sont consacrés à la construction d'estimateurs non paramétriques et à l'étude de leurs propriétés de convergence. Deux thèmes sont traités. Le premier thème traite du problème des durées de vie pour des données censurées à gauche. La censure aléatoire à gauche survient lorsque le temps d'origine de la durée de vie précède celui de l'étude. Dans de nombreux domaines notamment en médecine, agriculture, chimie, environnement, sociologie, etc. il est très fréquent de rencontrer ce type de données incomplètes. Pourtant très peu de références existent sur ce sujet. Dans ce cadre, nous proposons tout d'abord un lissage par noyau de l'estimateur de Kaplan-Meier à gauche (KMG). Pour ce nouvel estimateur, la convergence presque-sûre uniforme, la normalité asymptotique et la décomposition asymptotique du biais, de la variance et de l'erreur quadratique moyenne (EQM) sont établies. De nouveaux estimateurs sont construits par approche directe du taux de hasard et de la densité de probabilité pour données censurées à gauche. Pour chacun de ces estimateurs, les résultats de convergence presque-sûre uniforme et de normalité asymptotique sont prouvés. La décomposition asymptotique de l'erreur quadratique intégrée (EQI) ainsi que l'optimalité d'un critère de validation croisée basé sur l'EQI sont établies.

Le second thème porte sur la modélisation non paramétrique pour des observations à valeurs dans un espace mesuré. Des résultats de convergence ponctuelle en probabilité et presque-complète d'estimateurs de type delta-suites de la fonction de régression, de la fonction de densité et de hasard conditionnelles, sont obtenus pour des processus α-mélangeants et φ-mélangeants.

- Assia Ayache, doctorante de l'Université de Constantine, bénéficie de séjours de recherche financés par son université. En collaboration avec Fouad Rahmani son directeur de thèse, j'assure un suivi à distance. Sa thématique de recherche porte sur les réseaux de neurones et, plus précisément, sur les méthodes supervisées et non supervisées dans le cas de données bruitées.

- Iryna Petrovska, doctorante de l'Université SGGW (Warsaw University of life Sciences), a effectué un séjour de recherche à l'Esitpa (Septembre 2013-Février 2014). Ses travaux ont porté sur les facteurs du statu quo (Taïbi et *al.* 2014)

Encadrement d'ingénieurs d'étude et de recherche

Emmanuelle Nieullet ingénieur diplômée de l' Esitpa a été recrutée dans le cadre du projet Campus Paysan (Tamatave-Madagascar). Elle a eu pour mission l'appui et le suivi des phases de ce programme entre 2005 et 2006 (Taïbi et *al.* 2007).

L'encadrement de l'équipe BIOMATH

Les membres constituant le groupe Biomath sont : Jeanne-Chantal Dur (Inra de Versailles), Parice Lepelletier (Esitpa), Jérome Dantan, Laurence Rougé (Université de rennes), Jeanne Bodin (Post-Doc Esitpa). Pendant la durée du programme Bioindicateurs 2, j'ai assuré la coordination du groupe Biomath avec Jeanne-Chantal Dur, relativement aux travaux de recherche tels que les valeurs de référence, la sélection d'indicateurs sensibles à un usage ou à un stress, les scores de sensibilité et les scores d'opérationnalité ou transférabilité vers un public d'utilisateurs.

Encadrement de stagiaires

Les sujets de stage proposés aux étudiants de Master ou Ingénieur émanent de questionnements en lien avec les projets de recherche ou de recherche fondamentale. L'encadrement a été assuré par l'équipe du Lamsad en collaboration avec des chercheurs du LMRS et de l'Insa de Rouen.

1. 2012 Jia Fan, INSA de Rouen / LMRS Projet de Fin d'études 5ème Année Génie Mathématiques/AIMAF2
2. 2011 Christophe Marborough,Stage Technicien Insa de Rouen,
3. 2010 Pierre Parent, P.,INSA-LAMSAD, 2010.
4. 2010 Théophile Chaumont-Frelet, Insa de Rouen, 3ème Année Génie Mathématiques,
5. 2009 Halima Chtioui, Université de Bourgogne Master 2 MIGS,
6. 2007-2008 Aurore Lambert, Insa de Rouen Projet de fin d'études Génie Mathématiques,
7. 2007 Arles Fanampindrainy, LAMSAD- Université de Tamatave,Maîtrise de gestion,
8. 2007 Valentin Vlaaz, Université Université de Galati (Roumanie) Master 2,
9. 2007 Valentina Contantinescu, Université de Galati (Roumanie) Master 2,
10. 2006 Candice Rouen, Insa de Rouen, énie Mathématiques, Stage Ingénieur,
11. 2005 Jawad Alaoui, Insa de Rouen Projet de Fin d'études, Stage Ingénieur,
12. 2004 Mounir Lafkahi, Université de Caen Master 2,
13. 2004 Youssef Kacimi, Université de Caen Master 2,
14. 2003 Mélanie Frémont, Insa de Rouen Projet de Fin d'études Ingénieur,
15. 2002 Valérie Chauvenssy, Université de Rouen. Master 1
16. 2002 Delphine Grancher, Université de Rouen. Master 1

Dans le cadre des projets et stages proposés aux étudiants de l'Esitpa j'ai assuré le suivi de plus de 200 étudiants.

5.4 Travaux de Vulgarisation

5.4.1 Les sciences de la vie mises en équation

La Fête de la Science est un événement annuel visant à promouvoir la science auprès du grand public. Lors de cette manifestation qui eut lieu à l'Esitpa en 2009, j'ai présenté de manière ludique une balade mathématique que j'ai intitulée, "les sciences de la vie mises en équation". L'objet étant de montrer combien la nature a inspiré les mathématiciens. En prenant quelques exemples dans la nature j'ai abordé la notion de fractales, la suite de Fibonacci, les avancées de la statistique (S.R Fisher) pour l'analyse et planification expérimentale en agronomie.

J'ai passé en revue l'histoire de la statistique avec pour exemple le test de William Sealy Gosset plus connu sous le nom de Student et ses recherches pour la brasserie

Guinness, S.R Fisher et ses travaux avec Yates sur les plans de haricots à la station expérimentale de Rothamsted en Angleterre, avec une parenthèse sur les plans en carrés latins et le Sudoku.

5.4.2 Quand mathématiques riment avec développement durable

"30 minutes pour comprendre" sont des conférences de vulgarisation animées par l'Université de Rouen et destinées au grand public. Avec Manasé Bezara actuellement enseignant- chercheur à l'ISTOM, nous avons présenté le 18 Octobre 2010 une conférence intitulée " Quand mathématiques riment avec développement durable".

L'objectif étant de montrer comment deux chercheurs en mathématiques ont lancé le débat sur la question suivante : la pauvreté du monde rural de la région de Tamatave (Madagascar) est-elle une conséquence de la dégradation de la vie socio-communautaire dans les campagnes ou la dégradation de la vie communautaire est-elle une conséquence de cet appauvrissement ? Des réflexions entre mathématiciens, géographes, sociologues et agronomes ont été menées pour cerner les raisons de cette dégradation, et ont débouché sur la nécessité de trouver un modèle explicatif de la baisse du rendement en riz.

Chapitre 6

Responsabilités et activités d'enseignement

6.1 Responsabilités d'enseignement

J'assure actuellement la responsabilité de la Plateforme de Modélisations et Traitements physique, mathématique et informatique à l'Esitpa. J'ai débuté ma carrière d'enseignant-chercheur en qualité de maître de conférences à l'Université d'Annaba (Algérie) le 1er Septembre 1986. J'ai assuré les cours magistraux et travaux dirigés en 1ère Année pour les physiciens, en licence et maîtrise de mathématiques, en bio statistique pour la filière Médecine-Pharmacie. J'ai occupé de 1992 à 1996 un poste d'ATER au sein de l'Université de Rouen et assuré des enseignements de mathématiques générales en Deug Mathématiques et Informatique 1ère et 2ème Année, en Deug Sciences de la Vie et de la Terre et en Statistique 2ème Année SVT. Recrutée en 1998 comme enseignant-chercheur au sein de l'Esitpa (École d'Ingénieurs en Agriculture) avec pour mission la gestion des enseignements de Mathématiques et Statistique. J'ai enseigné sur les cinq années (cycle préparatoire et en cycle ingénieur). En 2001, il y eut la création du Secteur « Mathématiques et Statistique » dont j'ai assuré la responsabilité. J'ai élaboré les programmes des enseignements de : mathématiques générales, algèbre linéaire, statistique fondamentale, calcul des probabilités, analyse des données, plans d'expériences, dispositifs expérimentaux en agronomie, modélisation linéaire et non linéaire, séries chronologiques, sensométrie, plans de sondage, méthodologie d'enquêtes et maîtrise statistique des procédés. En 2006, après la fusion des deux secteurs « Mathématiques et Statistique» et « Informatique » en un département, Biométrie et Informatique, j'en ai assuré l'animation.

J'ai eu donc en charge l'animation de l'équipe permanente de ce département, la gestion du budget, le recrutement et la gestion pédagogique des enseignants vacataires, la veille documentaire, la constitution et la mise à jour de programmes d'enseignements en mathématiques, statistique et méthodes de l'ingénieur, la création de nouveaux enseignements ou modules d'enseignements et la recherche de méthodes pédagogiques innovantes.

6.2 Activités d'enseignement

Depuis mon entrée à l'Esitpa, j'ai favorisé l'apprentissage de la statistique en proposant aux étudiants des études de cas et en mobilisant les connaissances acquises en cours. Les entreprises ou structures de recherche proposaient un sujet. Les étudiants travaillaient par groupe, devaient rendre un compte-rendu écrit et soutenir leur travail par une présentation orale devant un jury pluridisciplinaire. Dans le cadre de plusieurs enseignements aussi tels la méthodologie d'enquête, la stratégie de recherche par les plans d'expérience, la maîtrise des procédés, l'analyse sensorielle, il est proposé aux étudiants de travailler sur un projet. La plus-value étant de rendre les étudiants autonomes sur des problématiques similaires ou connexes rencontrées au cours de leurs stages.

6.3 Contributions pédagogiques

6.3.1 Les compléments de savoir, soutien en mathématiques

J'ai mis en place le projet d'intégration des bacheliers, non issus de terminale Scientifique, en 1ère Année ou de BTS en 2ème Année. L'intégration d'étudiants issus de filières différentes a donc nécessité la mise en place de compléments de savoir. J'ai donc élaboré les programmes de cours intensifs en mathématiques générales pour les étudiants s'inscrivant en 1ère Année et 2ème Année. J'ai initié le programme des Khôlles de mathématiques afin de permettre aux étudiants, d'une part, d'approfondir leurs connaissances et, d'autre part, un rythme de travail continu.

6.3.2 Formation continue, Formation par apprentissage

J'ai intervenu dans des enseignements et tutorat dans le cadre de cursus ingénieur et licences professionnelles par la voie de l'apprentissage (Esitpa, Université du Havre, Université de Rouen, Hortithèque). En 2008, j'ai élaboré le programme pour une session de formation continue en modélisation statistique pour l'équipe de recherche de l'unité PESSAC à l'INRA de Versailles. J'ai aussi assuré des cours en licence Pro CAB, en lien avec la formation continue de l'Université de Rouen. J'ai aussi conçu et mis en place une session de formation continue en maîtrise statistique des procédés pour une équipe d'ingénieurs en sidérurgie de la SNS (Société Nationale Sidérurgique) , en 1988, à Annaba (Algérie).

6.3.3 Formation par la voie VAE

J'ai assuré la responsabilité de la formation par la voie VAE (Validation des Acquis de l'Expérience) de 2007 à 2009. J'ai élaboré le dossier en vue de l'habilitation, par la CTI (Commission des Titres d'Ingénieur) , à délivrer le diplôme d'ingénieurs en Agriculture de l'Esitpa par la voie VAE.

6.3.4 Mise en place d'observatoires

L'observatoire des Jeunes Diplômés

L'Observatoire des Jeunes Diplômés de l'Esitpa a pour objectif le suivi de ses Jeunes Ingénieurs. J'ai eu comme mission de prendre en main et d'améliorer l'observatoire des Jeunes Diplômé(e)s. J'ai donc conçu un nouveau questionnaire, amélioré le mode de sondage qui n'était que par courrier et assuré tous les traitements des résultats. J'ai mis en ligne le questionniare et les tableaux de bords à partir de l'année 2002. Des résultats de suivi des Jeunes Diplômés sur cinq ans ont fait l'objet d'une communication lors du colloque francophone sur les sondages (Taïbi et Lepelletier 2010).

L'observatoire de l'enseignement

En 2004, en collaboration avec l'ensemble de l'équipe pédagogique, j'ai conçu le questionnaire servant à évaluer tous les enseignements suivis par les étudiants (cinq promotions). J'ai élaboré une automatisation pour le dépouillement, de sorte que le rendu final soit assez synthétique et puisse être aisément mis en ligne.

Chapitre 7

Conclusion générale et perspectives de recherche

À travers ce mémoire, j'ai décrit une synthèse de mes activités de recherche et mes contributions en modélisation statistique au terme de plus de vingt années de recherche. Dans une première partie, j'ai mis en évidence les apports théoriques en estimation non paramétrique de fonctionnelles telles que le taux de hasard, la fonction de survie, la densité conditionnelle, le mode contionnel et la fonction de régression. Les estimateurs que nous étudions sont des estimateurs lissés c'est-à-dire utilisant des noyaux ou des fonctions dérivées de noyaux. En effet ces estimateurs ont de bonnes propriétés de convergence asymptotiques. Ces estimateurs utilisent des échantillons d'une variable ou d'un couple de variables aléatoires. Dans notre cas nous avons étendu certains résultats au cas de variables dépendantes et plus précisément au cas de processus mélangeants. Nous avons ainsi énoncé des théorèmes de convergence de l'estimateur de la fonction de hasard dans le cas de la méthode du noyau et des k-points les plus proches et pour des observations complètes.

Il arrive cependant que la durée de vie ne puisse pas être observée de manière complète mais est censurée à droite ou gauche. Nous avons donc développé des estimateurs de la fonction de hasard dans le cas censuré à droite et à gauche. En effet dans de nombreux ouvrages, la censure à gauche est vue comme un problème d'inversion du temps et les travaux s'y rattachant sont juste une adaptation des travaux relatifs à la censure à droite. Cette démarche présente des inconvénients, nous avons développé un nouvel estimateur par approche directe.

Dans un souci de répondre à des problématiques environnementales, agronomiques, économiques, sociétales, biologiques et économiques, nous avons contribué à divers travaux en lien avec des projets de recherche pluridisciplinaires. Ainsi les travaux sur la fonction de régression étendus au cas d'espaces mesurables ont été appliqués au consentement à payer d'un individu (Taïbi-Hassani *et al.* 2015a).

Outre les réponses aux questionnements par les chercheurs, j'ai établi des démarches méthodologiques, des modèles innovants pour l'élaboration d'indices tels que l'indice de la qualité des sols, le consentement à payer, le taux de richesse.

En parallèle, les recherches sur des projets finalisés m'ont amenée à nourrir la recherche fondamentale. Tel est le cas de la modélisation des variables qualitatives dans le cadre d'un projet en économie. J'ai ainsi établi une nouvelle méthode statistique, applicable dans d'autres contextes.

La grande diversité des thèmes que j'ai abordés lors de toutes ces années de recherche ouvrent de nombreuses perspectives de recherche fondamentale et appliquée. Par exemple, nous avons toujours supposé l'indépendance de la variable censure et la variable durée de vie alors qu'en pratique il peut arriver que cette condition soit difficilement satisfaite.

Nous avons vu qu'à travers les travaux présentés ici le sol représente un système complexe. En effet aussi la non connaissance des lois de distribution des phénomènes, les données inobservables en raison d'une censure à gauche ou droite sont autant de paramètres nous amenant à considérer une modélisation non paramétrique.

Dans le cadre du projet Unitwin, dont le thème est la modélisation des systèmes complexes, les échanges entre chercheurs internationaux permettront d'approfondir ces thèmes et d'appliquer nos résultats à d'autres systèmes complexes (Taïbi *al.* 2014 b) .

Les interactions sol et climat sont autant de pistes de recherche pas assez explorées ouvrant des perspectives de recherche en estimation non paramétrique et en modélisation statistique. Notamment à travers les travaux de recherche menés dans le cadre de l' agriculture de précision, qui sont en lien étroit avec les "Big Data", le Data Mining et les méthodes d'estimation non paramétriques.

BIBLIOGRAPHIE

1. Acton D.F., Padbury G.A.,1993. A conceptual framework for soil quality assessment andmonitoring. A Program to Assess and Monitor Soil Quality in Canada. *Soil Quality Evaluation Summary. Res Branch Agric.* Ottawa,Canada, 1993.
2. Adigaw-E-Touck Adigaw S.L.,2013. Modèles non paramétriques de survie pour données incomplètes, Thèse de doctorat, , Université de Rouen, 160 pages, Janvier 2013.
3. Ahmad. I. A.,1976. Uniform strong convergence of a generalized failure rate estimate. *Bull. Math. Statist.*, **17**, pp. 77-82.
4. Amigues J.-P., Boulatoff C., Desaigues B., Gauthier C., Keith J. E., 2002. The benefits and costs of riparian analysis habitat preservation : a willingness to accept/willingness to pay contingent valuation approach, *Ecological Economics* **43**, 17-31.
5. Anderson T.H., Domsch K.H., 1985. Determination of ecophysiological maintenance carbon requirements of soil microorganisms in a dormant state. *Biol. Fertil. Soils* **1**, 81-89.
6. Anderson, T.H., Domsch, K.H., 1990. Application of eco-physiological quotients (qCO2 and qD) on microbial biomasses from of soils from different cropping histories. *Soil Biol. Biochem.* **22**, 251-255.
7. Andrews S.S., Karlen D.L., Mitchel, J.P., 2002a. A comparison of soil quality indexing methods for vegetable production systems in Northern California. *Agric. Ecosyst.Environ.* **90**, 25-45.
8. Andrews S.S., Mitchell J.P., Mancinelli R., Karlen D.L., Hartz T.K., Horwarth W.R., Pettygrove G.S., Scow K.M., Munk D.S., 2002b. On-farm assessment of soil quality in California's central valley. *Agronomic Journal*, **94**, 12-23.
9. Antoch J., Collomb G., Hassani S. 1984. Robustness in parametric and non parametric regression estimation : An investigation by computer simulations, *COMPSTAT*, Physica Verlag, Vienna , 49-54.
10. Ardilly P., 2006. Les techniques de sondage. *Edition : Nouv. éd. actualisée et augm* . 675 pages.
11. Arkoun O., Barbu V., Crastes R, Laroutis D., Jia F., Taïbi-Hassani S. 2012. Sondages et plans fractionnaires appliqués à la méthode des programmes. 7ème Colloque Francophone sur les sondages. Rennes. Nov 2012.
12. Arrow K., Solow R., Leamer E., Portney P., Randner R., Schuman H.,1993. Report of the NOAA Panel on contingent valuations, U. S. Federal Register, 15 January, **10**, 4601-4614.
13. Aspetti G.P., Boccelli R., Ampollini D., Del Re A A M, Capri E., 2010. Assessment of soil-quality index based on microarthropods in corn cultivation in Northern Italy. *Ecological Indicators.* **10** (2), Pages 129-135.
14. Balesdent J., Chenu C., Balabane M., 2000. Relationship of soil organic matter dynamics to physical protection and tillage. *Soil and Tillage Research*, **53**, 215-230.
15. Baize D., 2000. Guide des analyses en pédologie. *Editions Quae*, 257 p.

16. Bastida F., Zsolnay A., Hernández T., García C., 2008. Past, present and future of soil quality indices : A biological perspective. *Geoderma*, **147**, 159-171.

17. Bean S.J., Tsakas C.P., 1980. Developments in non-parametric density estimation, International Statistical Review, **48**, 267-287.

18. Beaumais O., Laroutis D., Chakir R., 2008. Conservation versus conversion des zones humides : Une analyse comparative appliquée à l'estuaire de la Seine, Revue d'Économie Régionale et Urbaine **4**, 565-590.

19. Beck, T.H., 1984. Methods and application of soil microbiological analysis at the Landensanstalt fur Bodenkultur und Pflanzenbau (LBB) for determination of some aspects of soil fertility. Proceedings of the Fifth Symposium on Soil Biology. *Rumanian National Society of Soil Science*, Bucharest, Rumania, 13-20.

20. Bernard S., Heutte L., Adam S., 2007. Using Random Forests for handwritten digit recognition, International Conference on document Analysis and Recognition 1043-1047.

21. Bernard S., Heutte L., Adam S., 2009. Une Étude sur la Paramétrisation des Forêts Aléatoires, XI ème Conférence francophone sur l'Apprentissage Artificiel, Hammamet : Tunisie,2009.

22. Berthier A., Sentilhes L., Taïbi S. , Loisel C. ,Philippe Grise , Marpeau L., 2008. Sexual function in women following the transvaginal tension-free tape procedure for incontinence. *International Journal of Gynecology and Obstetrics* ,**102** (2), 105-109.

23. Billingsley P., 1968. Convergence of probability measures, John Wiley & Sons Inc.

24. Blum J. R., Susarla V., 1980. Maximal deviation theory of density and failure rate function estimates based on censored data. *In Multivariate analysis*, V (Proc. Fifth Internat. Sympos., Univ. Pittsburgh, Pittsburgh, Pa., 1978),213-222. North-Holland, Amsterdam. .

25. Bodin J., Taïbi, S., Thoisy-Dur J.-C.,Dantan J., Lepelletier, P., Rougé L., 2013. Bioindicateurs de la qualité des sols. Démarche d'analyse globale. Rapport d'activités. Ademe- Esitpa.

26. Bouroche J.M., Saporta G., Tenenhaus M., 1977. Some methods of qualitative data analysis - Recent development in Statistics. *North Holland Publishing Company.*

27. Breiman L., 2001. Random Forests, *Machine Learning* **45**, 5-32.

28. Brisson N., Mary B., Ripoche D., Jeuffroy MH., Ruget F., Nicoullaud B., Gate P, Devienne F., Antonioletti R., Dürr C., Richard G., Beaudoin N., Recous S., Tayot X., Plénet D., Cellier P., Machet JM, Meynard JM., Delécolle R., 1998. STICS : a generic model for the simulation of crops and their water and nitrogen balance. I. Theory and parameterization applied to wheat and corn. *Agronomie* **18**, 311-346.

29. Brookes P.C., 2001. The use of microbial parameters in monitoring soil pollution by heavy-metals. *Biol. Fertil. Soils* **19**, 269-279.

30. Brosteaux Y., 2005. Etude du classement par forêts aléatoires d'échantillons perturbés à forte structure d'interaction, thesis, Gembloux University, 178 pages.

31. Burez J., Van den Poel D., 2007. CRM at a pay-TV company : Using analytical models to reduce customer attrition by targeted marketing for subscription services, *Expert Systems with Applications* **32**, 277-288.

32. Bury Q., Giro F., Legras,M. 2005. Trace Element speciation in soils, phytoavailability and distribution in field-grown flax oilseeds as affected by different contamined soils. 9 th FECS Conference on Chemistry and the environment.

33. Calkins P., Larue B., Vezina M., 2002 Willingness to pay for drinking water in the Sahara : the case of Douentza in Mali. *Cahiers d'économie et sociologie rurales*, **64**.

34. Caravaca F., Masciandaro G., Ceccanti B., 2002. Land use in relation to soil chemical and biochemical properties in a semiarid Mediterranean environment. *Soil Tillage Res.***68**, 23-30.

35. Collomb G., 1976. Estimation non paramétrique de la régression par la méthode du noyau. Thèse de doctorat -Université Paul Sabatier, Toulouse, France,

36. Collomb G., 1981. Estimation non paramétrique de la régression : Revue bibliographique. *International Statistical Review*, **49**, 75-93.

37. Collomb G., 1984. Propriétés presque complète du prédicteur à noyau. *Z. Wahrsch. Verw. Gebiete* **66**, 448-460.

38. Collomb G., Hassani S., Sarda P., Vieu P., 1985a. Estimation non paramétrique de la fonction de hasard pour des observations dépendantes. *Statistique et Analyse des Données*, **10** (13),42-49.

39. Collomb G., Hassani S., Vieu P., Sarda P., 1985b. Convergence uniforme d'estimateurs de la fonction de hasard pour des observations dépendantes : méthodes du noyau et des k-points les plus proches. *Comptes-Rendus de l'Académie des Sciences de Paris* **301**, série 1,(12), 653-656.

40. Collomb G., Härdle W., Hassani S.,1986. A note on prediction via estimation of the conditionnal mode function, *Journal of Statistical Planning and Inference*, **15**, 227-236.

41. Colombo S., Hanley, N., Calatrava-Requena, J., 2005.Designing Policy for reducing the off-farm 482 effects of soil erosion using choice experiment.*Journal of agricultural economics* **56** (1), 81-95.

42. Coussement K., Van den Poel D., 2009. Improving customer attrition prediction by integrating emotions from clients/company interaction emails and evaluating multiple classifiers, *Expert Systems with Applications* **36**, 6127-6134.

43. Crastes R., Beaumais O., Arkoun O., Laroutis D., Mahieu P.A, Rulleau B., Hassani-Taïbi S., Barbu V., Gaillard D., 2014. Erosive runoff events in the European Union : using discrete choice experiment to assess the benefits of integrated management policies when preferences are heterogeneous, *Ecological Economics***102**, 105-112.

44. Crooker J.R., Herriges J.A., 2004. Parametric and Semi-Nonparametric Estimation of Willingness To Pay in a Contingent Valuation Framework, *Staff General Research Papers* **11156**.

45. Csörgo S.and Horvaath.L., 1980. Random censorship from the left. *Studia Sci. Math.Hungar.*, **15**(4),397-401.

46. Csörgo S., and Horvaath L., 1983 The baboons come down from the trees quite normally.*In Mathematical statistics and applications*, **B** (Bad Tatzmannsdorf, 1983) , 95-106. Reidel, Dordrecht.

47. Cutler D.R., Edwards T.C., Beard K.H., Cutler A., Hess K.T., Gibson J., Lawler J., 2007. Random Forests for classification in ecology, *Ecology* **88**,2783-2792.

48. Czajkowski, M., Buszko-Briggs, M., Hanley, N., 2009. Methods Valuing changes in forest biodiversity,*Ecological Economics*, **68**, 2910-2917.

49. Da Silva Fc, Vendite L.L., Bergamasco A.F., 2000. Heavy metal transportation modelling in the soilsugarcane system under fertilization with urban waste compost. *Proceedings of international Canegro Workshop mount Edgecombe*. South Africa.

50. Dale V. H., Beyeler S. C, 2001. Challenges in the development and use of ecological indicators. *Ecological Indicators*, **1**, 3-10.

51. Dale V. H., Polasky S., 2007. Measures of the effects of agricultural practices on ecosystem services, *Ecological Economics* **64**, 286-296.

52. Dantan J., Pollet Y., Taïbi S., 2012. A KDD Process to retrieve and aggregate data from relational databases. *In proceedings of IADIS International Conference Information Systems*. Berlin, Germany, 443-445.

53. Dantan, J., Pollet, Y., Taïbi, S., 2013. The G.O.A.L. Approach. *In proceedings of ENASE International Conference on Evaluation of Novel Approaches to Software Engineering* ,173-180. Angers, France.

54. Dantan J., Pollet Y, Taîbi S., 2015a. Combination of Imperfect Data in Fuzzy and Probabilistic Extension Classes, *Journal of Environmental Accounting and Management*,**3**, (2), 123-150.

55. Dantan, J., Pollet, Y., Taïbi, S. 2015b. A formal model to compute uncertain continuous data. In proceedings of CCS 2015 (international Conference on Complex Systems) - CS-DC'15 World e-conference (Complex Systems Digital Campus) UNITWIN/UNESCO. September 28 - October 2, 2015.

56. Deheuvels P., 1977. Estimation non paramétrique de la densité par histogramme généralisés. *Revue de statistique Appliquée*, **25** (3).

57. Dilly O., Blume H.P., Sehy U., Jiménez M., Munich J.C., 2003. Variation of stabilised, microbial and biologically active carbon and nitrogen soil under contrasting land use and agricultural management practices. *Chemosphere* **52**, 557-569.

58. Doob J., 1953. Stochastic Processes. Wiley, New-York.

59. Drosbeke J.J, , Fichet B., Tassi P., 1989. Analyse Statistique des Durées de Vie. *Economica* 282 p.

60. Everitt B., 1992. The Analysis of Contingency Tables, Chapman & Hall/CRC.

61. Eryern J., 1977. Revue of some non-parametric methods of density estimation. *Journal Inst. Math. Applic*, **20**, 335-354.

62. Garcia C., Hernández T., Costa F., 1994. Microbial activity in soils under Mediterranean environmental conditions. *Soil Biol. Biochem.* **26**, 1185-1191.

63. Garzón M.B., Blazk R., Neteler M., Sánchez de Dios R, Ollero H.S., Furlanello C., 2006. Predicting habitat suitability with machine learning models : The potential area of Pinus sylvestris L. in the Iberian Peninsula, *Ecological Modelling* **197**, 383-393.

64. Ghanem A., Bados P., Estaun R.A, Felippe L.de Alencastro, Taïbi S., Einhorn J., Mougin C., 2007. Concentrations and specific loads of glyphosate, diuron, atrazine, nonylphenol and metabolites thereof in French urban sewage sludge. *Chemosphere.*, **69** ,1368-1373.

65. Gomez G., O. Julia, and F. Utzet , 1994. Asymptotic properties of the left Kaplan-Meier estimator. *Comm. Statist. Theory Methods*, **23**(1), pp.123-135.

66. Goupy J., 2006. Introduction aux plans d'expériences, 3ème édition. Dunod, Paris.

67. Gurler U., Wang J. L., 1993. Nonparametric estimation of hazard functions and their derivatives under truncation model. *Ann. Inst. Statist. Math.*, **45**(2),249-264.

68. Halvorson J.J., Smith J.L., Papendick R.I., 1996. Integration of multiple soil parameters to evaluate soil quality : a field experiment example. *Biol. Fertil. Soils*, **21**, 207-214.

69. Hanley N., Adamowicz W., Wright, R.E., 2005. Price vector effects in choice experiments : an empirical test, *Resource and Energy Economics* **27**, 227-234, .

70. Hanley N., Wright R.E., Adamowicz V., 1998. Using choice experiments to value environment design Issues, current experience and future prospects, *Environmental and Ressource Economics*, **11**, 413-428.

71. Härdle W., Marron J. S., 1985. Optimal bandwidth selection in nonparametric regression function estimation. *Ann. Statist.*, **13**(4),1465-1481, .

72. Harris R.F., Karlen D.L., Mulla D.J., 1996. A conceptual framework for assessment and management of soil quality and health. In : Doran, J.W., Jones, A.J. (Eds.), Methods for Assessing Soil Quality, . SSSA, Madison, Wisconsin, SSSA Spec Publ., **49**, 61-82.

73. Hassani S., 1985. Sur quelques problèmes d'estimation et de prédiction non paramétriques. Thèse de 3ème Cycle. Université Paul Sabatier Toulouse.

74. Hassani S., Sarda P. Vieu. P., 1986. Approche non paramétrique en théorie de la fiabilité : revue bibliographique, *Revue de Statistique Appliquée*, **35**, (4), 27-41.

75. Hedde M., Peres G., Villenave C., Gattin I., Leguedard M., Harris-Hellal J., Dequiedt S., De Vaufleury A., Taïbi S., Grand C., Bispo A., 2014. Comment calculer les services écosystémiques rendus par les sols : un essai sur la base des données du programme « Bioindicateurs de qualité des sols » de l'ADEME,12èmes journées d'étude des sols, Chambéry.

76. Heutte L.,Bernard S., Adam S., Oliveira E., 2008. De la sélection d'arbres de décision dans les forêts aléatoires. Colloque International Francophone sur l'Ecrit et le Document,163-168.

77. Hoyos D., 2010. The state of the art of environmental valuation with discrete choice experiment, Ecological Economics, **69**,(8),1595-1603.

78. Iverson L.R., Prasad A.M., Liaw A., 2004. New machine tools for predictive vegetation mapping after climate change : Bagging and Random Forest perform better than Regression Tree Analysis, *Landscape ecology of trees forest*, 317-320.

79. Insam H., Domsch K.H., 1988. Relationship between soil organic-carbon and microbial biomass on chronosequences of reclamation sites. *Microb. Ecol.* **15**, 177-188.

80. Insam H., Haselwandter K., 1989. Metabolic quotient of the soil microflora in relation to plant succession. Oecologia 79, 174-178.

81. Jayasuriya R.T, 2003. Measurement of the scarcity of soil in agriculture. *Resources Policy* **29**, 119-129.

82. Jenkinson D.S., Ladd J.N., 1981. Microbial biomass in soil : measurement and turnover. In : Paul, E.A., Ladd, J.N. (Eds.), *Soil Biochemistry. Marcel Dekker*, New York,415-471.

83. Joergensen R.G., Brookes P.C., Jenkinson D.S., 1990. Survival of the microbial biomass at elevated-temperatures. *Soil Biol. Biochem.* **22**, 1129-1136.

84. Kang G.S., Beri V., Sidhu B.S., Rupela O.P., 2005. A new index to assess soil quality and sustainability of wheat-based cropping systems. Biol. Fertil. Soils **41**, 389-398.

85. Kaplan E. L., Meier.P., 1958. Non parametric estimation from incomplete observations. *Amer J.. Statist. Assoc.*, **53**, 457-481.

86. Karlen D.L.,Wollenhaupt N.C., Erbach D.C., Berry, E.C., Swan, J.B., Eash, N.S., Jordhal, J.L., 1994. Crop residue effects on soil quality following 10-years of no-till corn. Soil Tillage Res. **31**, 149-167.

87. Karlen D.L., Mausbach M.J., Doran, J.W., Cline, R.G., Harris, R.F., Schuman, G.E., 1997. Soil quality : a concept, definition, and framework for evaluation. Soil Sci. Soc. Am. J. **61**, 4-10.

88. Knoepp J. D., Coleman D. C., Crossley D. A. , Clark J. S., 2000. Biological indices of soil quality : an ecosystem case study of their use. *For. Ecol. Manage.* **138**,357-68.

89. Koper, J., Piotrowska, A., 2003. Application of biochemical index to define soil fertility depending on varied organic and mineral fertilization. *Electron. J. Pol. Agric. Univ.*

90. Kosz M., 1996. Valuing riverside wetlands : the case of the Donau-Auen national park, *Ecological Economics*,**16**, 109-127.

91. Ladenburg J., Lundhede T., Olsen S.B.,2008. The discovered preference hypothesis - an empirical test. Article présenté à la Conférence *European Association of Environmental and Resource Economists* , Gotenberg, Sweden.

92. Laroutis D., Taïbi-Hassani S., 2011. Discriminant Analysis Versus Random Forests on Qualitative Data : Contingent Valuation Method Applied to the Seine Estuary Wetlands. *International Journal of Ecological Economics and Statistics*,**20**, (11), 1-13.

93. Larson J.S., Adamus P.R., Clairian E.J., 1989.Functional assessment of freshwater wetlands : A manual and training outline. WWF and Environmental Institute, University of Massachusetts, Amherst, U.S.A.

94. Laval K., Mougin C., Akpa-Vinceslas M., Barray S., Dur J.C., Gangneux C., Lebrun J., Legras M., Lepelletier P., Plassart P., Taïbi S., Trinsoutrot-Gattin I., 2008. Nouvelles avancées vers la compréhension des données biologiques, *Étude et Gestion des Sols*, **16**, 275-287.

95. Lejeune M., Sarda P., 1992. Smooth estimation of distribution and density function. *Comp. Statist. Data Analysis* **14**, 457-471.

96. Leng.W., He H.S., Bu R., Dai L., Hu Y., Wang X., 2008. Predicting the distributions of suitable habitat for three larch species under climate warning in Northeastern China, *Forest Ecology and Management* **254**, 420-428 .

97. Liao M., Xiao X.M., 2007. Effect of heavy metals on substrate utilization pattern, biomass, and activity of microbial communities in a reclaimed mining wasteland of red soil area. *Ecotoxicol. Environ. Saf*, **66**, 21-223.

98. Loftsgaarden D.O., Quesenberry C.P., 1965. A nonparametric estimate of a mutivariate density function,. *Ann. Math. Stat.*, **36**,1049-1051.

99. Louviere,J., Hensher D.A., Swait J., 2000. Stated Choice Methods. *Analysis and Applications*, Cambridge University Press, Cambridge, UK.

100. Marron J. S.,W.Härdle., 1986. Random approximations to some measures ofaccuracy in nonparametric curve estimation. *Journal of Multivariate Analysis.*,**20**(1), 91-113.

101. Marron J. S., Padgett.W. J., 1987. Asymptotically optimal bandwidth selection for kernel density estimators from randomly right-censored samples. *Annals of Statistics*, 1520-1535 .

102. Maruyama T., Takimoto H., 2008. An economic evaluation of Kanazawa and Shichika irrigation water's multifonctional roles using CVM. *Paddy and Water Environment*, **6** 3, 309-318.

103. Masciandaro G., Ceccanti B., Gallardo-Lancho J.F., 1998. Organic matter properties in cultivated versus set-aside arable soils. Agric. Ecosyst. Environ. 67, 267-274.

104. Mohanty M., Painuli D.K., Misra A.K., Ghosh, P.K., 2007. Soil quality effects of tillage under rice-wheat cropping on a Vertisol in India. Soil Tillage Res. **92**, 243-250.

105. Moore, D.S., Yackell J.W., 1977. Consistency prpoperties of nearest neighbour density functions. *Annals of Statistics*,**5**, 143-154.

106. Muntean S., Legras M, Llorens J-M, Giro F., Allaoui J., Taïbi S. , 2007. Estimation of rates of uptake of trace elements from the soil to seeds of oilseed flax, *Bulletin of the University of Agricultural Sciences and Veterinary Medicine Cluj-Napoca*, **337**, 63-64 .

107. Murthy V.K, 1965. Estimation of jumps, reliability and hazard rate. *Annals of Statistics* **36**, 1032-1040.

108. Nannipieri, P., Grego, S., Ceccanti, B., 1990. Ecological significance of the biological activity in soils. In : Bollag, J.M., Stotzky, G. (Eds.), Soil Biochemistry. Marcel Dekker, New York, 293-355.

109. OFEV, Office Fédéral de l'Environnement, 2010. Méthodes d'analyse et d'appréciation des cours d'eau. *Office fédéral de l'environnement Berne*, p.61.

110. Ouvry J.F., 1992. L'évolutuion de la grande culture et l'érosion des terres dans le Pays de Caux. *Bulletin de l'Association des Géographes Français* **2**, 107-113.

111. Parisi V., Menta C., Gardi C. , Jacomini C. , Mozzanica E., 2005. Microarthropod communities as a tool to assess soil quality and biodiversity : a new approach in Italy, *Agriculture, Ecosystems & Environment* **105**, (1-2), 323-333.

112. Parr J.F., Papendick R.I., Hornik S.B., Meyer, R.E., 1992. Soil quality : attributes and relationship to alternative and sustainable agriculture. *Am. J. Altern. Agric.* **7**, 5-10.

113. Parzen E., 1962. On estimation of a probability density function and mode. *Ann.Math. Statist.*, **33**, 1065-1076.

114. Patil P.N., 1993a. Bandwidth choice for nonparametric hazard rate estimation. J.Statist. Plann. Inference, 35(1),15-30.

115. Patil P.N., 1993b.On the least squares cross validation bandwidth in hazard rate estimation, Annals od Statistics,21, 1792-1810.

116. Pérès G., Bispo A., Grand C., Gattin I., Hedde M., Harris-Hellal J., Leguedard, M. Ruiz, N., Alaphilippe A., Beguiristain T., Douay F., Faure O., Hitmi A., Houot S., Legras M., Guernion M., Vian J.F., Conil S., Rougé L., Lepelletier P., Taïbi S., Dur J.C., Cluseau D., 2012. Soil bioindicators for soil monitoring, risk assessment and soil characterization. Results from the French national "Bioindicators Programme". , 4th EUROSOIL , Bari.

117. Perucci P., 1992. Enzyme-activity and microbial biomass in a field soil amended with municipal refuse. *Biol. Fertil. Soils* **14**,54-60, .

118. Peters J., De Baerts B., Verhoest N.E.C., Samson R., Degroeve S., De Becker P., Huybrechts W., 2007. Random forests as a tool for ecohydrological distribution modelling, Ecological Modelling **207**, 304-318.

119. Peters J., Verhoest N.E.C., Samson R., Boeckx P., De Baets B., 2008. Wetlands vegetation distribution modeling for the identification of constraining environmental variables, Landscape Ecology **23**, 1049-1065, .

120. Powlson, D.S., Jenkinson, D.S., 1981. A comparison of the organic matter, biomass, adenosine-triphosphate, and mineralizable nitrogen contents of ploughed and direct drilled soils. *J. Agric. Sci.* **97**, 713-721.

121. Prasad A.M., Iverson L.R., Liaw A., 2006. Newer Classification and Regression Tree Techniques : Bagging and Random Forests for Ecological Prediction, *Ecosystems* **9**, 181-199.

122. Puglisi E., Del Re A.A.M., Rao M.A., Gianfreda L., 2006. Development and validation of numerical indices integrating enzyme activities of soils. *Soil Biol. Biochem.* **38**, 1673-1681.

123. Quigging J., 1998. Existence value and the contingent valuation method, *Australian Economic Papers* **37**, 312-329.

124. Ramlau-Hansen H., 1983. Smoothing counting process intensities by means of kernel functions. *Ann. Statist.*, **11**(2) :453-466.

125. Reiss R. D. 1981. Nonparametric estimation of smooth distribution functions. Scand.J. Statist., **8**(2),116-119.

126. Reyes C., Due E., 2009. Les faits, une arme contre la pauvreté. Le système de suivi communautaire de la pauvreté. Centre de recherches pour le développement international, 124 p.

127. Riffaldi, R., Saviozzi, A., Levi-Minzi, R., Cardelli, R., 2002. Biochemical properties of a Mediterranean soil as affected by long-term crop management systems. *Soil Tillage Res.* **69**, 109-114.

128. Ritz K., Helaina I.J, Colin D., Campbell, Harris J.A, Wood C., 2008. Selecting biological indicators for monitoring soils : a framework for balancing scientific and technical opinion to assist policy development. *Ecological Indicators*, **9**, 1212-1221.

129. Rosenblatt M., 1956. Remarks on some nonparametric estimates of a density function. *Annals of Mah. Stat*, **27**, 832-837.

130. Rosenblatt M., 1971.Markov Processes, structure and asymptotic behavior. Springer Berlin.

131. Sandro T., Bovo E., Fiore A.R., Guzzinati S., Monetti D., Stocco C.F., Zambon P., 2009. Probabilistic classifiers and automated cancer registration : An explanatory application. *Journal of Biomedical Informatics* **42**, 1-10.

132. Saporta G.,1977. Une méthode et un programme d'analyse discriminante pas à pas sur variables qualitatives, *INRIA Analyse des données et informatique*, **1**, 201-210, 1977.

133. Saviozzi A., Levi-Minzi, R., Cardelli, R., Riffaldi, R., 2001. A comparison of soil quality in adjacent cultivated, forest and native grassland soils. Plant Soil **233**, 251-259.

134. Schimmerling, P., Sisson, J.C., Zaïdi, A., 1998. Pratique des plans d'expériences, *Technique et Documentation*, Paris.

135. Singpurwalla N. D., Wong M. Y., 1983. Estimation of the failure rate :a survey of nonparametric methods. Part I : Non-Bayesian Methods. *Comm. Statist. Theory Methods*, **12**(5),559-588.

136. Smith J.L., Halvorson, J.J., Papendick, R.I., 1993. Using multivariable-indicator kriging for evaluating soil quality. Soil Sci. Soc. Am. J. **57**, 743-749.

137. Sojka R.E.et. Upchurch D.R., 1999. Reservations regarding the soil quality concept. Soil Sci. Soc. Am. J.**63**, 1039-1054.

138. Soliño M. , Alvarez-Farizo B. , Campos P., 2009. The influence of home-site factors on residents' willingness to pay : An application for power generation from scrubland in Galicia,Spain. Energy Policy, **37** (10), pp. 4055-4065.

139. Stefanic F., Ellade G., Chirnageanu J., 1984. Researches concerning a biological index of soil fertility. In : Nemes, M.P., Kiss, S., Papacostea, P., Stefanic, C., Rusan, M. (Eds.),Fifth Symposium on Soil Biology. Romanian National Society of Soil Science, Bucharest, 35-45.

140. Sun L.,1997. Bandwidth choice for hazard rate estimators from left truncated andright censored data. Statist. Probab. Lett., **36**(2),101-114.

141. Taïbi-Hassani S., Youndjé E., 1997. Estimation lisse d'une fonction de hasard : Choix optimal de la fenêtre pour des observations censurées. Comptes Rendus de l'Académie des Sciences de Paris. **324**, (I), pp. 481-484.

142. Taïbi-Hassani S., Youndjé E., 2003. Validation croisée pour l'estimateur lissé de la fonction de hasard : cas des données censurées, *Revue de Statistique Appliquée*, **LI**(I),73-86.

143. Taïbi S., Lemaire A.S., 2005. Estimation du taux de transfert des éléments trace du sol vers les graines de lin oléagineux- ; *Statistique des Processus*. Angers, France.

144. Taïbi, S., Bezara M., Nieullet E., Nodjirim D., 2006. Mise en place d'un modèle de développement durable dans la province de Tamatave, Rapport d'activités, Esitpa, Région Haute Normandie, Université de Tamatave.

145. Taïbi S. Roche D., 2009. Rapport d'évaluation du projet "Campus Paysan". Région Haute Normandie - Université de Tamatave (Madagascar), Esitpa.

146. Taïbi S., Adigaw-E-Touck S.L, 2011 Validation croisée pour un estimateur lisse de la fonction de hasard sous données censurées à gauche 44èmes Journées de Statistique de la SFDS. Gammarth 23-27 Mai, 2011a.

147. Taïbi S., Lepelletier P., G Perez, Rougé L., Dur J-C, Bispo A.,2011b. Démarche en vue d'élaborer un indice d'état du sol. 44èmes Journées de Statistique de la SFDS. Gammarth-Tunisie.

148. Taïbi, S., Bezara M., 2011c. La méthode des Forêts aléatoires appliquée à l'Observatoire de la ruralité à Tamatave. Pratiques et méthodes de sondage. Dunod, Collection Cours et Cas Pratiques, 382 p.

149. Taïbi S., Rougé L., Thoisy-Dur J.-C., Bodin J., Lepelletier P., Dantan J., Pérès G., Grand C.,Bispo A., 2012. « Gestion et traitement des données du programme. Approche statistique de sélection d'Indicateurs et de biomarqueurs dans la surveillance de la qualité des sols et l'évaluation des risques. Journées Techniques Nationales, Bioindicateurs pour la caractérisation des sols, ADEME, Paris, 10 p.

150. Taïbi S., Lepelletier P., Dantan J., Bodin J., Thoisy-Dur J-C., Rougé L., 2012. Gestion et traitement des données du programme. Approche statistique de sélection d'Indicateurs et de biomarqueurs dans la surveillance de la qualité des sols et l'évaluation des risques. Rapport Final Ademe , Esitpa, 101 p.

151. Taïbi-Hassani, S., Thoisy-Dur, J-C., Lepelletier, P., Bodin, J., Bennegadi-Laurent, N., Bessoule, J-J., Bispo, A.,Bodilis, J., Chaussod, R., Cheviron, N., Cortet, J., Criquet, S., Dantan, J Dequiedt, , A., Faure, O., Gangneux, C., Harris-Hellal J., Hedde, M., Hitmi, A., Le Guedard, M., Legras, M., Pérès, G., Repinçay, Rougé, L., C., Ruiz, N., Trinsoutrot-Gattin, I., Villenave, C., 2013. Démarche statistique pour la sélection des indicateurs par Random Forests pour la surveillance de la qualité des sols. *Etude et Gestion des Sols* **20** (2), 127-136.

152. Taïbi-Hassani S., Lepelletier P., Dantan J., Thoisy-Dur J.C.,Bodin J., 2014a. A Statistical approach for soil monitoring, risk assessment and soil characterization, *e-Kickoff ICCSA'14, Complex Systems Digital Campus*, UNITWIN-UNESCO June 23-26th.

153. Taïbi-Hassani S., Petrovska I., Laroutis D., 2014b. Status Quo and willingness to pay for reduction of risk of erosive runoff for longitudinal data. *: Problems of World Agriculture*, **14** (4).173-177.

154. Taïbi-Hassani S., Laroutis D., Adigaw-E-Touck S., 2015a. Pointwise Convergence of a nonparametric estimator of regression in a measurable space used in Contingent Valuation Method. *Journal of Mathematics and System Science*, **5**, 188-195.

155. Taïbi-Hassani, S., Lepelletier, P., Blot, A., Thoisy-Dur, J-C., 2015b. A statistical approach to the evaluation and modelling of contamination in an agro-ecosystem. *International Journal of Ecology Economics and Statistics (IJEES)*, **36**, (1),83-97.

156. Taïbi-Hassani S., Adigaw E-Touck S., 2015c. A direct approach of nonparametric estimation of the hazard rate with left censored data. Soumis.

157. Taglioni C., Cavicchi A., Torquati B., Scarpa R.,2001. Influence of Brand Equity on Milk Choice : A Choice Experiment Survey, *Int. J. Food*
System Dynamics, **2** (3), 305-325.

158. Tanner M. A.,Wong W. H., 1983. The estimation of the hazard function from randomly censored data by the kernel method. *Annals of Statistics*, **11**(3), 989-993.

159. Tapia R.A, Thompson J.R, 1978. Non-parametric probability density estimation. Baltimore, London : Johns Hopkins University Press.

160. Thiria S., Lechevallier Y., Gascuel O., Canu S., 1997. Statistique et méthodes neuronales, Dunod, Paris.

161. Trèves, F., 1964. Topological vector spaces, distribution and kernels. *Academic press.*.

162. Uzunogullari U., Wang J. L., 1992. A comparison of hazard rate estimators for left truncated and right censored data. Biometrika, **79**(2),297-310.

163. Van Haluwyn, Chantal, Lerond, M., 1993. Guide des lichens pour le diagnostic écholichénique de la qualité de l'air , Ed. LeChevallier, Paris.

164. Vebeke G., Molenlebergh G., 2000. Linear Mixed Models for longitudinal data. Spinger-Verlag, New York.

165. Vieu P., 1991. Quadratic errors for nonparametric estimates under dependence. Journal of Multivariate Analysis, **39**(2),324-347.

166. Villanneau E., Perry-Giraud C., Saby N., Jolivet C., Marot F., Maton D., Floch-Barneaud A., V. Antoni et Arrouays D., 2008. Détection de valeurs anomaliques d'éléments traces métalliques dans les sols à l'aide du Réseau de Mesure de la Qualité des Sols. Etude et Gestion des Sols, **15** (3), 183-200.

167. Wardle D.A, Ghani A., 1995. A critique of the microbial metabolic quotient (qCO2)) as a bioindicator of disturbance and ecosystem development. Soil Biol Biochem **27**, 1601-1610.

168. Ware J. H., Demets D. L., 1976. Reanalysis of some baboon descent data. Biometrics, **32**, 459-463.

169. Watson G. S., Leadbetter M. R., 1964. Hazard analysis. II. *Sankhya* Ser. A,**26**, 101-116.

170. Watson G. S.,Leadbetter M. R., 1964. Hazard analysis. I. *Biometrika*, **51**, 175-184.

171. Wei Y., Davidson B., Chen D., White R., Li B., Zhang J., 2007. Can Contingent Valuation be Used to Measure the in Situ Value of Groundwater on the North China Plain ? *Water Ressource Management* **21**, 1735-1749.

172. Wertz W., 1981. Nonparametric density estiùatorsin abstract and homogeneous spaces.*Lecture notes in Mathematics*. University of Technology, Wien.

173. Yandell B.S., 1983. Nonparametric inference for rates with censored survival data. *Annals of Statistics*, **11**(4),1119-1135.

174. Youndjé E., Sarda P. , Vieu P., 1996. Optimal smooth hazard estimates. *Test*, **5**(2),379-394.

ANNEXE 1 Curriculum Vitae

TAIBI SALIMA née HASSANI Nationalité Française - Mariée 2 enfants
Docteure en Mathématiques Appliquées
Responsable de Service
Esitpa - École d'ingénieurs en Agriculture
staibi@esitpa.fr
Tel : 00 33 (0)664116918

FORMATION

Doctorat de Mathématiques appliquées de l'Université Paul Sabatier Toulouse (Toulouse III).

Intitulé de la thèse : «Sur quelques problèmes d'estimation et de prédiction non paramétriques» Septembre 1985. Direction Gérard Collomb.

DEA de Mathématiques, Université Paul Sabatier, Toulouse, (Toulouse III) Juillet 1982.

Missions et Responsabilités

Depuis Septembre 2014 Responsable de la Plateforme traitements et modélisations physiques, mathématiques et informatiques - Esitpa

De 2006 à Sept 2014 Responsable du Département de Biométrie et Informatique Esitpa

- Enseignement Recherche Encadrement - Coordination et animation d'une équipe d'enseignants et enseignants-chercheurs - Enseignement en Mathématiques Appliquées et Statistique - Conception des programmes de mathématiques, statistique et sciences pour l'ingénieur - Suivi des étudiants : Khôlles en mathématiques, compléments de savoir, soutien - Suivi et encadrement de stagiaires ingénieurs et master - Conception et gestion du budget du département - Formation continue : INRA , Université de Rouen ;,

Responsable du laboratoire de modélisation statistique Lamsad (De 2001 à Avril 2013)et membre du LMRS

- Animation d'une équipe de recherche - Chercheure associée au Laboratoire de Mathématiques LMRS - Animation et participation à des programmes de recherche - Mise en place d'un projet de développement à Madagascar - Encadrement de Post-doctorants : Université de Rouen, Université de Nancy, Université de Cluj Napoca (Roumanie), Tamatave (Madagascar) - Encadrement de doctorants (Université de Rouen, Université de Tizi Ouzou, Université de Constantine) - Encadrement de stagiaires (INSA, Université de Caen, Université du Havre, Université de Rouen, Université de Dijon, Esitpa).

2006-2009 Responsable VAE Esitpa

- Gestion des dossiers VAE. - Rédaction du Dossier CTI pour l'habilitation à délivrer le Diplôme par la voie VAE.

1998-2009 Responsable de l'Observatoire des Jeunes Diplômés Esitpa - Gestion de l'observatoire des Jeunes Diplômés - Analyse des données, synthèse des résultats - Mise en ligne de l'enquête : questionnaire, tableaux de bords (Sphinxonline)

2001-2006 Responsable du Secteur Mathématiques et Statistique

- Enseignement Encadrement - Coordination et animation d'une équipe denseignants et enseignants-chercheurs - Enseignement en Mathématiques Appliquées et Statistique - Animation de projets pédagogiques - Participation à la mise en place du projet d'établissement - Conception des programmes de mathématiques, statistique et sciences pour l'ingénieur - Formation / Conception de concours : Esitpa, EXIA-CESI, ISTOM, Université du Havre (Master Arômes Parfums Chimie Fine et Cosmétologie), Université de Rouen (Licence pro Contrôles Agroalimentaires et Biotechnologies) - Projet d'intégration des Bacheliers STAV, ES- Conception des programmes pour les compléments de savoir, et admis directs en 2ème Année.

1998-2001 Enseignant chercheur Esitpa

- Enseignement en Mathématiques Appliquées et Statistique - Coordination des enseignements de mathématiques et statistique - Encadrement de mémoires de fin d'études et de stages d'initiation à la recherche - Recherche en modélisation et statistique fondamental **1995-1997 Enseignante Esitpa**

- Enseignement (1ère à 5ème Année) - Conception des programmes de 4ème Année Esitpa - Encadrant Scientifique pour les mémoires d'ingénieur Esitpa

1993 - 2000 Qualification par le CNU au poste de Maître de Conférences.

1992-1996 ATER Université de ROUEN

- Deug MI (Math /Informatique) 1ère et 2ème A. Math Probabilités Statistique - DEUG SVT (Sciences de la Vie et de la Terre) 1ère et 2ème A.

1986 à 1992 **Maître-assistante /Chargée de Cours- Université d'Annaba** - Enseignement :

Licence et Maîtrise de Mathématiques-Biomédical -Technologie - Mise en place d'une unité de recherche en Mathématiques et Statistique - Formation continue en maîtrise des procédés pour les ingénieurs en Sidérurgie (SNS)

Production scientifique

Articles dans des revues internationales et nationales avec comité de lecture

1. Taïbi-Hassani S., Laroutis D., Adigaw-E-Touck S., 2015. Pointwise Convergence of a nonparametric estimator of regression in a measurable space used in Contingent Valuation Method. *Journal of Mathematics and System Science*, **5**, 188-195.
2. Taïbi-Hassani S., Lepelletier P., Blot A., Thoisy-Dur J-C., 2015. A statistical approach to the evaluation and modelling of contamination in an agro-ecosystem. *International Journal of Ecology Economics and Statistics (IJEES)*, **36**, (1),83-97.
3. Dantan J., Pollet Y., Taîbi S., 2015. Combination of Imperfect Data in Fuzzy and Probabilistic Extension Classes, *Journal of Environmental Accounting and Management*,**3**, (2), 123-150.
4. Taïbi-Hassani S., Adigaw E-Touck S.,2015. A direct approach of nonparametric estimation of the hazard rate with left censored data. Soumis
5. Taïbi S., Petrovska I., Laroutis D., 2014. Status Quo and willingness to pay for reduction of risk of erosive runoff. *Scientific Journal Warsaw University of Life Sciences : Problems of World Agriculture*, **14** (XXIX) n°4,173-177.
6. Crastes R., Beaumais O., Arkoun O., Laroutis D., Mahieu P.A, Rulleau B., Hassani-Taïbi S. Barbu V., Gaillard D., 2014. Erosive runoff events in the European Union : using discrete choice experiment to assess the benefits of integrated management policies when preferences are heterogeneous.*Ecological Economics*,**102**, 105-112.
7. Taïbi-Hassani S., Thoisy-Dur J-C, Lepelletier P., Bodin J., Bennegadi-Laurent N., Bessoule J-J., Bispo A., Bodilis J., Chaussod R., Cheviron N., Cortet J., Criquet S., Dantan J., Dequiedt S., Faure O., Gangneux C., Harris-Hellal J., Hedde M., Hitmi A., Le Guedard M,, Legras M., Pérès G., Repinçay C., Rougé L. , Ruiz N., Trinsoutrot-Gattin I. ,Villenave C., 2013. Démarche statistique pour la sélection des indicateurs par Random Forests pour la surveillance de la qualité des sols, *Etude et Gestion des Sols*, **20**(2), 127-135.
8. Laroutis D., Taïbi-Hassani S., 2011. Discriminant Analysis Versus Random Forests on Qualitative Data : Contingent Valuation Method Applied to the Seine Estuary Wetlands. *International Journal of Ecological Economics & Statistics* ,**20** (11), 1-13.
9. Berthier A., Sentilhes L., Taïbi S. , Loisel C. ,Philippe Grise , Marpeau L., 2008. Sexual function in women following the transvaginal tension-free tape procedure for incontinence. *International Journal of Gynecology and Obstetrics*, **102**(2),105-109 .
10. Laval K., Mougin C., Akpa-Vinceslas M., Barray S., Dur J.C., Gangneux C., Lebrun J., Legras M., Lepelletier P., Plassart P., Taïbi S., Trinsoutrot-Gattin I.,2008. Nouvelles avancées vers la compréhension des données biologiques,*Etude et Gestion des Sols* ,**16**, 275-287.
11. Ghanem A., Bados P., Estaun R.A, Felippe L.de Alencastro, Taïbi S., Einhorn J., Mougin C.,2007. Concentrations and specific loads of glyphosate, diuron, atrazine, nonylphenol and metabolites thereof in French urban sewage sludge. *Chemosphere*, **69** , 1368-1373.

12. Muntean S., Legras M., Llorens J.M, GIRO F., Allaoui J., Taïbi S., 2007. Estimation of rates of uptake of trace elements from the soil to seeds of oilseed flax. *USAMV-CN,Bulletin of the University of Agricultural Sciences and Veterinary Medicine Cluj-Napoca*, **63** 337.
13. Taïbi-Hassani S.Youndjé E., 2003. Validation croisée pour l'estimateur lissé de la fonction de hasard : cas des données censurées. *Revue de Statistique Appliquée*, **LI**(I),73-86.
14. Taïbi-Hassani S., Youndjé E., 1997. Estimation lisse d'une fonction de hasard : Choix optimal de la fenêtre pour des observations censurées. *Comptes Rendus de l'Académie des Sciences de Paris*. Tome 324, Série I, 481-484.
15. Collomb G. Härdle W., Hassani S., 1987. A note on prediction via estimation of the conditionnal mode function, *Journal of Statistical Planning and Inference*, **15** ; 227-236.
16. Hassani S., Collomb G.,Sarda P. Vieu. P., 1986. Approche non paramétrique en théorie de la fiabilité : revue bibliographique, Revue de Statistique Appliquée, **35** (4) 27-41.
17. Collomb G., Hassani S., Sarda P., Vieu P.,1985. Estimation non paramétrique de la fonction de hasard pour des observations dépendantes., *Statistique et Analyse des Données*, **10** (13) ; 42-49. .
18. Collomb.G., Hassani S.,Vieu P.,Sarda P., 1985. Convergence uniforme d'estimateurs de la fonction de hasard pour des observations dépendantes : méthodes du noyau et des k-points les plus proches. *Comptes-Rendus de l'Académie des Sciences de Paris tome 301*, **série 1** (12), 653-656..
19. Antoch J., Collomb G., Hassani S., 1984. Robustness in parametric and non parametric regression estimation : An investigation by computer simulations. *COMPSTAT, Physica Verlag,Vienna* , 49-54.

Chapitre d'ouvrage

1. Taïbi S., Bezara M., 2011. La méthode des Forêts aléatoires appliquée à l'Observatoire de la ruralité à Tamatave. Pratiques et méthodes de sondage. Dunod, Collection Cours et Cas Pratiques, 382 p.

Conférences données à l'invitation du comité d'organisation dans des congrès ou colloques

1. Taïbi S., Petrovska I., Laroutis D., 2014. Status Quo and willingness to pay for reduction of risk of erosive runoff. 11th International Science Conference on Global Problems of Agriculture, forestry and food economy Warsaw.
2. Taïbi, S., Lepelletier, P., Dantan J., Thoisy-Dur, J.-C.,Bodin, J. A., 2014. Statistical approach for soil monitoring, risk assessment and soil characterization, e-Kickoff ICCSA'14 , Complex Systems Digital Campus, UNITWIN-UNESCO.
3. Taïbi, S., Rougé, L., Thoisy-Dur, J.-C., Bodin, J., Lepelletier, P., Dantan, J., Pérès, G., Grand, C. and Bispo, A., 2012. « Gestion et traitement des données du programme. Approche statistique de sélection d'Indicateurs et de biomarqueurs dans la surveillance de la qualité des sols et l'évaluation des risques. Journées Techniques Nationales, Bioindicateurs pour la caractérisation des sols, ADEME, Paris, 10 p.
4. Taïbi S., 2007. Statistical modelling and sustainable development. Ells University SGGW Varsow.Conference Erasmus Mundus Warsaw Poland.
5. Taïbi S., Problèmes d'estimation et de prédiction non paramétriques sous censure aléatoire à droite, Université de Dijon Juin 2004.

Communications avec comité de lecture dans des congrès internationaux

1. Dantan, J., Pollet, Y., Taïbi, S. 2015. A formal model to compute uncertain continuous data. In proceedings of CCS 2015 (international Conference on Complex Systems) - CS-DC'15 World e-conference (Complex Systems Digital Campus) UNITWIN/UNESCO. September 28 - October 2, 2015.
2. Dantan, J., Pollet, Y., Taïbi, S. 2015. A systemic meta-model for socio-environmental systems. In proceedings of the Sixth International Conference on Complex Systems Design Management, CSDM 2015. Editors : Auvray, G., Bocquet, J.-C., Bonjour, E., Krob, D. (Eds.). November 23-25, 2015. P. 307. Paris, France
3. Pauget B., Rougé L., Bispo A., Grand C., Beguiristain T., Bessoule J.-J., Bodilis J., Chaussod R., Cheviron N., Coeurdassier M., CortetJ., Criquet S., Dequiedt S., Faure O., Gangneux C., Gattin I., le Guedard M., Hitmi A., Laurent N., Legras M., Nélieu S., Ruiz N., Taïbi S., Vandenbulcke, F., de Vaufleury, A., Villenave C. and Pérès G. 2015. « Soil bioindicators : how soil properties influence their responses and how to select them in function of the site issues ? ». SETAC Europe 25th Annual Meeting. 3-7 May, Barcelona, Spain.
4. Dantan J., Pollet Y., Taïbi S. 2014. Taking account of uncertain, imprecise and incomplete data in sustainability assessments in agriculture. In proceedings of Computational Science and Its Applications - ICCSA 2014 - 14th International Conference, Part III, Lecture Notes in Computer Science LNCS 8581, ISBN 978-3-319-09149-5, pp. 625639. Springer International Publishing Switzerland. Guimarães, Portugal, June 30 - July 3, 2014. Communications oralese*t al.* ICCSA-CLASS 2014.pdf
5. Dantan J., Pollet Y., Taïbi S. 2014. A goal-oriented meta-model for scientific research. In proceedings of Computational Science and Its Applications - ICCSA 2014 - 14th International Conference, Part V, Lecture Notes in Computer Science LNCS 8583, ISBN 978-3-319-09155-6, pp. 762774. Springer International Publishing Switzerland. Guimarães, Portugal, June 30 - July 3, 2014. ICCSA-AEIDSS 2014.pdf
6. Pauget B., Rougé L., Bispo A., Grand C., Beguiristain T., Bessoule J.-J., Bodilis J., Chaussod R., Cheviron N., Coeurdassier M., Cortet J., Criquet S., Dequiedt S., Faure O., Gangneux C., Gattin I., le Guedard M., Hitmi A., Laurent N., Legras M., Nélieu S., Ruiz N., Taïbi S., Vandenbulcke F., de Vaufleury A., Villenave C., Cluzeau D., Pérès G., 2014. Soil bioindicators : how soil properties influence their responses and how to select them in function of the site issues ? 1er GSBI. 3-5 Décembre 2014, Dijon, France.
7. Pérès G., Pauget B., De Vaufleury A., Coeurdassier M., Leguedard M., Bessoule J.J., Dequiedt S., Chaussod R., Ranjard L., Cluzeau D., Guernion M., Rougé L., Hedde M., Cheviron N., Dur J.C., Nelieu S., Mougin C., Gattin I., Gangneux C., Laurent N., Legras M., Laval K., Lepelletier P., Taïbi S., Villenave C., Faure O., Hellal J., Cortet J., Beguiristain T., Leyval C., Bodilis J., Criquet S., Hitm A., Ruiz N., Vandenbulcke F., Grand C., Galsomies L., Bispo A. 2014. Which bioindicators are suitable for soil quality monitoring and risk assessment ? From relevance study to transfer tool development. 1er GSBI. 3-5 Décembre 2014, Dijon, France.
8. Taïbi S., Thoisy-Dur J.C., Bodin J., Rougé L., Dantan J., Lepelletier P., Michaud A., Houot S., Pérès G., Bispo A., 2013.A statistical approach to assess soil biodiversity and biological activity responses to repeated organic amendment applications in cultivated soils - Relationships with soil functions. *RAMIRAN , 15th international conference*, Versailles, France.
9. Dantan J., Pollet Y., Taïbi S. 2013. The G.O.A.L. Approach. In proceedings of ENASE International Conference on Evaluation of Novel Approaches to Software Engineering. ,173-180. Angers, France, July 4-6, 2013.
10. Pérès G., Bispo, A., Grand C., Cluzeau D., Gattin I., Hedde M., Cheviron N., Harris-Hellal J., LeGuedard M., Bessoule J.J., RuizN., Pauget B., de Vaufleury A., Beguiristain T., Dequiedt S., Chaussod R., Faure. O., Hitmi A., Criquet S., Legras, M., Laurent N., Vandenbulcke F., Coeurdassier M., Ponton S., Cortet J., Villenave C., Bodillis J., Lepelletier P., Taïbi S., Dur J.-C., Bodin J. 2013. Application of soil bioindicators for risk assessment, monitoring and soil characterization in contaminated soils. Results from the French national "Bioindicators Programme".

12th International UFZ-Deltares Conference on Groundwater-Soil-Systems and Water Resource Management (AquaConSoil). Barcelona, Spain.

11. Crastes R., Beaumais, O., Arkoun, O., Laroutis, D., Mahieu, P.A., Rulleau, B., Hassani-Taïbi, S., Barbu, V.S., Gaillard, D., 2013. Erosive Runoff Events in the European Union : Using Discrete Choice Experiment to Assess the Benefits of Integrated Management Policies when Preferences are Heterogeneous. Workshop on non-market valuation. , Nantes, France..
12. Arkoun O., Barbu V., Crastes R, Laroutis D., Jia F., Taïbi-Hassani S., 2012. Sondage et plans fractionnaires appliqués à la méthode des programmes. 7ème Colloque Francophone sur les sondages. Rennes.
13. Thoisy-Dur J.-C., Lepelletier P., Taïbi S., Rougé L., Dantan J., Pérès G., Grand C. and Bispo A., 2012. Statistical approach to select soil bioindicators for soil monitoring, risk assessment and soil characterization ». Results from the French national Programme Bioindicators. 6th SETAC World Congress/SETAC Europe 22nd Annual Meeting, Berlin.
14. Bodin J., Dur J.-C., Rougé L., Dantan J., Lepelletier P., Grand C., Pérès G., Bispo A. Taïbi S., 2013. Soil bioindicators to assess soil biodiversity and activity responses to land-use practices. Final results of the research project Bioindicators ». International Interdisciplinary Conference on Land Use and Water Quality : Reducing Effects of Agriculture, The Hague, Netherland.
15. Gaillard D, Bonnet E., Bensaïd A., Arkoun O., Barbu V., Beaumais O., Crastes R., Laroutis D., Mahieu P.A., Rulleau B., Taïbi S., 2012. Analyse spatialisée de la perception du risque de ruissellement érosif. Modélisation et consentement à payer. Intérêt et apports d'une approche pluridisciplinaire », 2ème séminaire international euro-méditerranée sur l'Aménagement du Territoire la Gestion des risques et la Sécurité civile, Algérie.
16. Pérès G., Bispo A., Grand C., Gattin I., Hedde M., Harris-Hellal J., Leguedard M., Ruiz N., Alaphilippe A., Beguiristain T., Douay F., Faure O., Hitmi A., Houot S., Legras M., Guernion M., Vian J.F., Conil S., Rougé L., Lepelletier P., Taïbi S., Dur J.C., Cluseau D., 2012. Soil bioindicators for soil monitoring, risk assessment and soil characterization. Results from the French national "Bioindicators Programme". , 4th EUROSOIL , Bari, Italy.
17. Dantan J., Pollet Y., Taïbi S., 2012. Semantic indexation of Web services for collaborative expert activities. In proceedings of IADIS International Conference Information Systems. March 10-12, , 57-64, Berlin, Germany.
18. Dantan J., Pollet Y., Taïbi S., 2012. A KDD Process to retrieve and aggregate data from relational databases. *In proceedings of IADIS International Conference Information Systems.*, 443-445, Berlin, Germany.
19. Pérès G., Grand C., GattinI., Hedde M., Harris-Hellal J., Leguedard M., Ruiz N., Alalaphilippe A., Beguiristain T., Pruvot,C., FaureO., Hitmi A., Houot S., Legras M., Guernion M., Vian J.F., Conil S., Rougé L., Taïbi, S., Cluzeau , D., 2011. A national research programme to validate a battery of soil bioindicators for impact and risk assessment in urban soils. SUITMA 6, Marrakech, Morocco.
20. Taïbi S., Adigaw-E-Touck S.,2011. Validation croisée pour un estimateur lisse de la fonction de hasard sous données censurées à gauche 44èmes Journées de Statistique de la SFDS. Gammarth, Tunisie.
21. Taïbi S., Lepelletier P., G Perez, Rougé L., Dur J-C, Bispo A., 2011. Démarche en vue d'élaborer un indice d'état du sol. 44èmes Journées de Statistique de la SFDS. Gammarth-Tunisie.
22. Pérès G., Ruiz N., Hedde M. , Le Guedard M., Gattin I., d'Hugues P., Beguiristain T., Douay F. , Houot S. , Vian J.F., Faure O., Hitmi A., Alalaphilippe A., Dubs F., Rougé L., Taïbi S., Bispo A. , Grand C., Galsomies L. , Cluzeau D., 2011. Development and relevance assessment of bioindicators for soil monitoring, characterization and risk assessment. Example of a Bioindicator Program developed at National scale . *EJSB* , France.
23. Laroutis D., Taïbi S., 2010. Discriminant Analysis versus random forests on qualitative data : Contingent Valuation Method applied to the Seine estuary wetlands. 44th Annual Conference of the Canadian Economics Association, Quebec Canada.

24. Taïbi S., Bezara M., Lepelletier P., Nodjirim D., 2010. Mise en place de l'Observatoire de la ruralité dans le cadre du projet Campus Paysan à Madagascar, 6ème Colloque Francophone sur les Sondages, Tanger, Maroc.
25. Taïbi S, Lepelletier P., 2010. L'Observatoire des Jeunes Diplômés en Agriculture. 6ème Colloque Francophone sur les Sondages, Tanger, Maroc.
26. Taïbi S., Laroutis, D., 2009. Discriminant analysis versus random forests on qualitative data : Contingent Valuation Method applied to the Seine estuary wetlands. Applied Statistics International Conference, Ribno Slovenia.
27. Adigaw Touk S., Laroutis D, Taïbi S., 2009. Estimation non paramétrique de la régression : cas du consentement à payer. Journées d'études en statistique , Bordeaux.
28. Laroutis D. Taïbi S.,2009. Analyse discriminante versus forêts aléatoires : méthode d'évaluation contingente appliquée à l'estuaire de la Seine. Journées d'études en statistique , Bordeaux.
29. Taïbi S., Rouen C., Bezara M., 2008 Modélisation du rendement du riz à partir de données longitudinales. Congrès SFDS SSC, Montréal, Canada.
30. Taïbi S., Lambert A., Lepelletier P., Laval K., Mougin C., 2008. Elaboration d'un indice de la qualité des sols. Congrès Statistical Society of Canada (SSC) et Société Française de Statistique Montréal, Canada.
31. Mougin C. Dur J-C , Ridreau C., Huard E., Taïbi S., Tessier D., 2008. High levels of enzymatic activities are measured in soils managed under no-tillage leading to acidification and increased bioavailability of toxic metals. SETAC Europe Varsaw .
32. Dur J.C., Legras M., Gangneux C., Gattin I., Bailleul C., Akpa M., Plassart P., Barray S., Taïbi S., Massignam A., Pandolfo C., Lebrun J., Hedde M., Mougin C., Laval K., 2007. Interest in the Development on one Risk Indicators in Soil Ecotoxicology. Soil and Wetland Ecotoxicology, SOWETOX Barcelona.
33. Duval C., Debandt V, Eveillé J-P, Mahieu D., Lepelletier P., Taïbi S., LLorens J.M., 2007a.Influence de l'hétérogénéité pédoclimatique en Haute-Normandie sur la variabilité intraparcellaire des rendements en blé et en colza,Journées d'études en statistique, Angers.
34. Duval C., Debandt V., Eveillé J-P., Mahieu D., Taïbi S., LLorens J-M., 2007b. Influence of the pedo-climatic variability in Haute Normandie (NW France) on the intra field spatial variability on yields of wheat and oilseed rape.6th European Conference on Precision Agriculture and the 3rd Euro pean Conference on Precision Livestock Farming Skiathos, Greece, 87-94 .
35. Mougin C., Laval K., Legras M., Barray S., Taïbi S., Tessier, D., 2007. Chemical contamination versus non chemical stressors : the case study of agricultural soils. SETAC Europe 17th Annual Meeting, Porto, Portugal.
36. Muntean S., Legras M., Llorens J.M., Giro F., Allaoui J., Taïbi S., 2007. Estimation of rates of uptake of trace elements from the soil to seeds of oilseed flax, 6th Iinternational Symposium "Prospects for the 3rd millennium agriculture", University of Agricultural Sciences and Veterinary Medicine, Cluj-Napoca, Romania, 4-6 October 2007.
37. Legras M., Bailleul C., Tessier D.,Dur J.C,, Gangneux C., Taïbi S., Laval K., 2006. Effect of physicochemical characteristics of agricultural soils on fungal biomass. Impact of copper, EMEC 7, European Meeting on Environmental Chemistry, Brno, République Tchèque, 6-10 december 2006.
38. Legras M., Gangneux C., Tessier D., Dur J.C, Bailleul C., Taïbi S., Laval K. Effect of physicochemical characteristics of agricultural soils on fungal biomass, ISME-11, 11th International Symposium on Microbial Ecology, Vienna (Autriche), 20-25 august 2006.
39. Mougin C ; Laval, K. Taïbi S., Tessier, D. Lemaire A-S.and Barray S.,2005. Towards an index of biological state of the soil as a new tool for ecotoxicological studies, SETAC Europe 15th Annual Meeting, Lille.
40. Taïbi-Hassani S., Youndjé E., 1996. Estimation lisse d'une fonction de hasard : Choix optimal de la fenêtre pour des observations censurées. XVIIèmes Rencontres Franco-Belges de Statisticiens. Marne-La-Vallée, France.

41. Hassani S., 1984. Régression non paramétrique pour des variables aléatoires à valeurs dans un espace mesurable. Journées de Statistique. A.S.U. Montpellier.

Communications avec comité de lecture dans des congrès nationaux

1. Hedde M., Peres G., Villenave C., Gattin I,. Leguedard M., Harris-Hellal J., Dequiedt S., de Vaufleury A., Taïbi S., Grand C. Bispo A., 2014. Comment calculer les services écosystémiques rendus par les sols : un essai sur la base des données du programme « Bioindicateurs de qualité des sols » de l'ADEME. *Les 12èmes Journées d'Etudes sur les sols.*
2. Crastes R., BeaumaisO., LaroutisD., ArkounO., MahieuP.A., Rulleau B., Taïbi S., 2012. Valuing the reduction of risks provoked by erosive runoffs using choice experiment. *Journée d'étude en Econométrie Appliquée.* Le Havre, France.
3. Taïbi S., Dur J.C Lepelletier P., Rougé L., Dantan J., Bispo A, Grand, C., G Perez., Approche statistique de sélection d'Indicateurs et de Biomarqueurs dans la surveillance de la qualité des sols et l'évaluation des risques. Résultats du programme national ADEME, Bioindicateurs II, JES Versailles, Mars 2012.
4. Taïbi S., Lemaire A.S., 2005. Estimation du taux de transfert des éléments trace du sol vers les graines de lin oléagineux, Statistique des Processus. Angers, France.

Communications sans actes dans des congrès ou des colloques nationaux

1. Laroutis D., Taïbi S.,2009. Analyse discriminante versus Forêts Aléatoires sur des données qualitatives : Méthode d'évaluation contingente appliquée aux zones humides de l'estuaire de la Seine, *XLVIème Colloque de l'Association de Science Régionale De Langue Française (ASRDLF)*, Entre enjeux locaux de développement et globalisation de l'économie : quels équilibres pour les espaces régionaux ?, Clermont-Ferrand, France 6-8 juillet .
2. Barray S., Laval K., Legras M., Mougin C., Taïbi S., Tessier D.,2006. Elaboration et validation d'un indice d'état biologique des sols, *3ème Séminaire d'Ecotoxicologie de l'INRA*, Dinard, France .
3. Taïbi S., 2004. Statistique et analyse sensorielle. Évaluation de la performance des juges dans le cadre d'une épreuve de profils sensoriels. Université de Rouen Janvier .
4. Taïbi-Hassani S.,2004. Statistique et analyse sensorielle. Évaluation de la performance des juges dans le cadre d'une épreuve de profils sensoriels. Séminaire du laboratoire LMRS, Université de Rouen.

Ouvrages de vulgarisation

1. Taïbi S., Bezara M., « Quand mathématiques riment avec développement durable ». Conférence "30 minutes pour comprendre", Université de Rouen. Octobre 2010.
2. Taïbi, S., Les sciences de la vie mises en équation. Conférence Fête de la Science, 17-23 novembre 2008, Rouen, France.

Projets de recherche

1. Projet UNITWIN UNESCO depuis 2013. Membre référente pour l'Esitpa (partenaire).
2. Programme National ADEME - Bioindicateurs II - Bioindicateurs de la qualité des Sols-Coordination et animation du groupe Biomath «Gestion et traitement des données du programme. Approche statistique de sélection d'indicateurs et de biomarqueurs dans la surveillance de la qualité des sols et l'évaluation des risques» 2009-2013 (coordination).

3. Projet « Identification des freins et leviers de l'agriculture intégrée » - Chambre d'agriculture Régionale de Normandie, Oct 2011/ Octobre 2013(partenariat).
4. Projet EMIRE I et II - Impacts sur le ruissellement érosif dans la Vallée du Commerce 2010-2013 (coordination).
5. Programme Ademe Bioindicateurs I 2005-2008 (collaboration).
6. Projet Campus Paysan 2005-2010 (coordination).
7. Projet Agriculture de Précision 2006-2007 (participation).
8. Projet ET LIN 2003-2005, (collaboration).

Responsabilités et animations

1. Responsabilité de l'équipe du Lamsad Laboratoire de modélisation statistique, 2002-2013 Esitpa
2. Animation du groupe Biomath,2010-2013 Projet Bioindicatieurs II

Encadrement

Encadrement de post-doctorants

1. Jeanne Bodin (Février 2012-Mars 2013), Docteur en Ecologie de l'Université de Nancy et Berlin .Projet Bioindicateurs de la qualité des sols(Bodin et *al.* 2013, Taïbi et *al.* 2012).
2. Ouerdia Arkoun (2011-2012) docteure en mathématiques de l'Université de Rouen . Projet EMIRE (Arkoun et *al.* 2012, Crastes et *al.* 2014).
3. Manassé Bezara(2005-2006, 2009) Enseignant-Chercheur , Université de Tamatave. Projet Campus Paysan (Taïbi et *al.* 2012, Bezara et *al.* 2010,Taïbi et 2009, Taïbi et *al.* 2010, Taïbi et *al.* 2007).
4. Sorin Muntean (2004-2005), Docteur 3ème cycle, Université des Sciences Agricoles et Médecine Vétérinaire de Cluj-Napoca (Roumanie) et Assistant-Professeur à la Chaire de Phytotechnie. Projet ETlin (Muntean et *al.* 2006).

Encadrement de doctorants

1. Saturnin Adigaw-E-Touck, Thèse soutenue le 11 Janvier 2013 intitulée "Modèles non paramétriques de survie pour données incomplètes" , Université de Rouen.
2. Iryna Petrovska, doctorante de l'Université SGGW (Warsaw University of life Sciences), a effectué un séjour de recherche à l'Esitpa (Septembre 2013-Février 2014). Ses travaux ont porté sur les facteurs du statu quo (Taïbi et *al.* 2014)
3. Jérôme Dantan, Doctorant CNAM-Esitpa. Directeur de Thèse Yann Pollet.
4. Assia Ayache, doctorante de l'Université de Constantine, bénéficie de séjours de recherche financés par son université. En collaboration avec son directeur de thèse, Fouad Rahmani, responsable de l'ecole doctorale de mathématiques de l'Université de Constatine, j'assure un suivi à distance. Sa thématique de recherche porte sur les réseaux de neurones, et plus précisément sur les méthodes supervisées et non supervisées dans le cas de données bruitées (2013)

-Encadrement d'ingénieurs d'étude et de recherche

Emmanuelle Nieullet, 2006-2007 Ingénieur en Agriculture et titulaire d'un Master de l'Université de Montpellier. Projet Campus Paysan à Madagascar (Taïbi et *al.* 2007).

- Encadrement de stagiaires

1. 2012 Jia Fan, INSA de Rouen / LMRS Projet de Fin d'études 5ème Année Génie Mathématiques/AIMAF2
2. 2011 Christophe Marborough,Stage Technicien Insa de Rouen,
3. 2010 Pierre Parent, P.,INSA-LAMSAD.
4. 2010 Théophile Chaumont-Frelet, Insa de Rouen, 3ème Année Génie Mathématiques,
5. 2009 Halima Chtioui, Université de Bourgogne Master 2 MIGS,
6. 2008 Sébastien Bellet,3ème Année Génie Mathématiques Insa de Rouen,
7. 2007-2008 Aurore Lambert, Insa de Rouen Projet de fin d'études Génie Mathématiques,
8. 2007 Arles Fanampindrainy, LAMSAD- Université de Tamatave,Maîtrise de gestion,
9. 2007 Valentin Vlaaz, Université Université de Galati (Roumanie) Master 2,
10. 2007 Valentina Contantinescu, Université de Galati (Roumanie) Master 2,
11. 2006 Candice Rouen, Insa de Rouen, Génie Mathématiques, Stage Ingénieur,
12. 2005 Jawad Alaoui, Insa de Rouen Projet de Fin d'études, Stage Ingénieur,
13. 2004 Mounir Lafkahi, Université de Caen Master 2,
14. 2004 Youssef Kacimi, Université de Caen Master 2,
15. 2003 Mélanie Frémont, Insa de Rouen Projet de Fin d'études Ingénieur,
16. 2002 Valérie Chauvenssy, Université de Rouen. Master 1
17. 2002 Delphine Grancher, Université de Rouen. Master 1

Rapporteur pour des revues

1. Journal of Applied Statistics.
2. Scientific Journal of the Warsaw University of Life Sciences.
3. Environmental Chemistry Letters.

Représentations

Membre du Conseil Scientifique - Esitpa, 2010-2013.

Rapports de recherche

1. Bodin J., Taïbi, S., Thoisy-Dur J.C.,Dantan J., Lepelletier, P., Rougé L., Bioindicateurs de la qualité des sols. Démarche d'analyse globale. Rapport d'activités. Ademe- Esitpa Fév. 2013.
2. Taïbi S., Lepelletier P., Dantan J., Bodin J., Thoisy-Dur J.C., Rougé L., 2012. Gestion et traitement des données du programme. Approche statistique de sélection d'Indicateurs et de biomarqueurs dans la surveillance de la qualité des sols et l'évaluation des risques. Rapport Final Ademe -Esitpa, 101 pp.
3. Taïbi, S., Rougé, L., Thoisy-Dur, J.C., Bodin, J., Lepelletier, P., Dantan J.2011. Gestion et traitement des données du programme. Approche statistique de sélection d'Indicateurs et de biomarqueurs dans la surveillance de la qualité des sols et l'évaluation des risques, Rapport Intermédiaire Ademe- Esitpa .
4. Taïbi S. Roche D. Rapport d'évaluation du projet « Campus Paysan ».- Région Haute Normandie - Université de Tamatave (Madagascar), Déc. Esitpa, 2009.
5. , Taïbi, S., Lepelletier P., Barray, S., Plassart, P., Brault, A., Dur, J.C., Huard, E., Lebrun, J., Mougin, C., Tessier, D., 2008. "Elaboration et validation d'un indice d'état biologique des sols", Rapport Final ADEME.

6. Barray S., Laval K., Lemaire A-S., Mougin C., Taïbi S.,2008. Rapport d'activité Programme Bioindicateurs I - Ademe.
7. Taïbi, S., Bezara M., Nieullet E., Nodjirim D., 2007. Mise en place du Projet Campus dans la province de Tamatave.
8. Taïbi, S., Bezara M., Nodjirim D.,2006 Mise en place d'un modèle de développement durable dans la province de Tamatave.

Rapports de stagiaires

1. Marborough, C., Bioindicateurs et indices de la qualité des sols INSA Génie Mathématiques-LAMSAD,2011.
2. Chaumont-Frelet T., Analyse statistique d'une base de données concernant l'efficience énergétique et économique des exploitations agricoles de polyculture-élevage de Haute-Normandie. INSA- LAMSAD, 2010.
3. Parent, P., Travaux préliminaires à l'établissement d'un indice de qualité des sols. Génie Mathématiques INSA -LAMSAD, 2010.
4. Chtioui H., Etude de la qualité d'indicateurs socio-économiques et biologiques à partir de méthodes de classement, Mémoire de Master 2, MIGS Uiversité de Dijon LAMSAD, 50pp., 2009.
5. Lambert, A., Mise en place d'un bio-indicateur de la qualité des sols. Projet de fin d'études,5ème année Génie Mathématiques INSA, LAMSAD, 2008.
6. Bellet S. 3ème annnée Génie Mathématiques INSA- LAMSAD, 2008.
7. Constantinescu, V.,. La mise en place d'un site web illustrant le projet -Campus Paysan de Madagascar,Lamsad Esitpa 2007.
8. Fanampindrainy, A., Constitution de la base de données en vue de la mise en place du Campus Paysan à Madagascar. LAMSAD- Université de Tamatave, 2007.
9. Vlaaz, V., Analyse des données issues de l'enquête Observatoire pour le projet Campus Paysan à Madagascar. Lamsad- 2007.
10. Rouen, C. Etude du rendement des exploitations, dans la province de Tamatave, dans le cadre du projet Campus Paysan. Projet de fin d'études Génie Mathématiques, 5ème année INSA-LAMSAD, 2006.
11. Alaoui J. , Transfert des métaux lourds dans le lin. Travaux de simulation. Projet de fin d'études Génie Mathématiques, 5ème année INSA-LAMSAD, 2005.
12. Kacimi Y.,Méthodes de discrimination dans le cadre d'une épreuves sensorielles, . Master 2 Université de Caen,2004.
13. Lafkahi M., Les séries chronologiques et données climatiques. Master 2 Université de Caen, 2004.
14. Frémont M. Statistique des procédés. Cartes de contrôles non paramétriques. Projet de fin d'études Génie Mathématiques, 5ème année INSA-LAMSAD, 2003.
15. Chauvenssy V., Analyse des données pluriannuelles recueillies auprès de l'observatoire de la qualité de l'air en Seine Maritime. Problèmes d'estimation et de prédiction. Mémoire Maîtrise d'Ingénierie Mathématique, Université de Rouen, 58pp., 2002.
16. Grancher D., Simulations en modélisation paramétrique et non paramétrique- Mémoire Maîtrise d'Ingénierie Mathématique Lamsad-Université de Rouen,2002.

ANNEXE 2

LISTE DES ABRÉVIATIONS

– ACM : Analyse des Correspondances Multiples
– ADEME : Agence de l'Environnement et de la Maîtrise de l'Énergie
– AIC : Akaike Information Criterion
– AFD : Analyse Factorielle Discriminante
– CAP : Consentement à payer
– CART : Classification Regression Tree
– CE : Choice Experiment
– CHAID : CHi-squared Automatic Interaction Detector
– CSP : Catégorie Socio-Professionnelle
– EMIRE : Évaluation monétaire de l'impact du ruissellement érosif
– ET : Eléments Traces
– ETM : Eléments Traces Métalliques
– GPN : Gain de Pureté du Noeud
– GRR : Grands Réseaux de Recherche
– IncMSE :Increased Mean Square Error
– INRA : Institut National de Recherche Agronomique
– INSEE : Institut National de la Statistique et des Études Économiques
– ISE : Integrated Squared Error
– ITL : Institut Technique du Lin
– LDA : Linear Discriminant Analysis
– LMRS : Laboratoire de Mathématiques Raphaël Salem
– MEC : Méthode d'Évaluation Contingente
– MISE : Mean Integrated Squared Error
– MO : Matière Organique
– OM : Organic Matter
– OOB : Out-Of-Bag
– PESSAC : Physicochimie et Ecotoxicologie des SolS d'Agrosystèmes Contaminés
– PLS : Partial Least Squares
– RF : Random Forests
– STICS : Simulateur multidisciplinaire pour les cultures standard
– SVM : Support Vector Machine
– VA : Variable Aléatoire
– VA : Validation des Acquis de l'Expérience
– WTP : Willingness To Pay