UNIVERSITÉ CHEIKH ANTA DIOP DE DAKAR (UCAD)

École Doctorale de Mathématiques et Informatiques (EDMI)

Année: 2019-2020 **nº d'ordre: 163**

THÈSE DE DOCTORAT UNIQUE

Présentée pour obtenir le titre de Docteur de l'Université Cheikh Anta Diop de Dakar

Mention : Mathématiques Appliquées

Formation doctorale: Analyse, Statistiques et Applications

Spécialité: Statistiques et Applications

présentée et soutenue publiquement par

Mamadou NDIAYE

le 26 Novembre 2020

Cette thèse a été préparée dans le cadre d'une collaboration entre les laboratoires LMA-CAD, LEM-Lille, CRODT-ISRA et l'IRD-Dakar

Contribution à la statistique spatiale et fonctionnelle : Modélisation spatio-temporelle des ressources halieutiques du Sénégal

Devant le Jury composé de :

Président:	Diaraf SECK,	Professeur Titulaire	UCAD
Rapporteurs:	Aliou DIOP,	Professeur Titulaire	UGB Saint-Louis
	Jean François DUPUY,	Professeur	IRMAR-INSA RENNES
Examinateurs:	Patrice BREHMER,	Chargé de Recherches (HDR)	IRD/Lemar
	Abdoulaye SENE,	Professeur Titulaire	USSEIN Kaolack
Directeurs:	Papa NGOM,	Professeur Titulaire	UCAD Dakar
	Sophie NIANG-DABO,	Professeur	Université de Lille, France
Encadrants:	Massal FALL,	Maître Assistant (CAMES)	USSEIN Kaolack
	Ciré Élimane SALL,	Chargé de Recherches	ISRA-CNRA Bambey

DÉDICACES

Je dédicace, ce travail

A mon cher Papa et guide Cheikh Ahmed Tidjane NDIAYE, qui nous a quitté le 08 janvier 2004. Paix à son âme. Il m'a transmis sa passion pour la quête du savoir. Un homme très modeste, disponible et qui se met toujours au service de sa communauté pour la bonne cause. Il a su se distinguer par sa détermination à éduquer et ramener les "égarés" dans le droit chemin. Je me rappelle toujours de sa rigueur quand il s'agissait d'éduquer ses enfants. Il nous disait ceci "Avant d'entreprendre une action quelconque mesurer d'abord les conséquences. Et si vous êtes convaincus que vous avez pris la bonne décision et que vous êtes sur le droit chemin, quoi qu'il arrive, ne s'en détournez pas". Il nous disait aussi "n'ayez pas peur d'entreprendre même si vous pensez ne pas en avoir les moyens". Il est toujours moralement et spirituellement présent à travers ses sages conseils. Ses paroles retentissent toujours dans mon esprit.

A ma Chère Maman Sokhna Arame NDIAYE, pour sa simplicité, sa sagesse, son sens de l'écoute et sa présence perpétuelle à nos côtés.

A tous mes chers parents, pour tous leurs sacrifices, leurs amours, leurs tendresses, leurs soutiens et leurs prières tout au long de mes études.

A mon Grand Frère Ahmed Mansour NDIAYE, qui a joué pleinement son rôle de père, depuis la disparition de notre Vénéré Père et guide. Un homme très disponible qui ne me répond jamais par la négation.

A mon Oncle Aliou BA, pour son soutien sans faille sur tous les plans. Un homme très sage à qui j'ai beaucoup d'estime.

A mon Oncle, Docteur Oumar DIA, qui m'a beaucoup soutenu lorsque j'étais à l'UNIVERSITÉ GASTON BERGER (UGB).

A mon épouse Gnagna Fal NDIAYE, pour soutien moral et ses conseils. Je te remercie pour ta présence, patience et surtout pour ta discrétion.

A ma tante Corca FAL, pour soutien moral et matériel.

A ma soeur Sokhna Mariama NDIAYE, nous avons partagé tant de choses ensembles.

A mon cousin et Oncle Amar FALL, qui a contribué considérablement à ma formation de base.

A mes chères petites sœurs pour leurs encouragements permanents, et leur soutien moral.

A mes chers petits frères, j'en cite Malick et Ahmad, pour leur appui et leur encouragement.

A ma fille Sokhna Khadidjatou NDIAYE.

A mon fils Cheikh Ahmad Tidjane NDIAYE.

A tous mes enfants, nièces et neveux.

A mon guide sprituel Cheikh Ahmad Tidjane Sy Ibn Cheikh Mouhamadou Mansour SY Boro Daradji, qui nous a pris sous son aile protectrice. Il a su apaisé nos douleurs causées par la perte subite de notre cher papa. Un homme très effacé doté d'une forte personnalité.

A mes amis Docteur Kamarel BA, Ousmane Wéllé et Cheikh Ahmad Tidjane DIOP.

A une personne spéciale professeur Moussa LÔ, le recteur de l'Université Virtuelle du Sénégal. IL m'a tendu la main quand j'en avais besoin, alors qu'il ne me connaissait même pas. Il a peut être oublié le geste noble qu'il a eu à faire pour me faire sortir des gouffres du désespoir. J'étais très affecté par la disparition subite de notre cher papa et le rôle que je devais assumer au sein de la famille m'empêcher de mener correctement les études . Je commençais les années universitaires sans pour autant pouvoir les terminer correctement. Il a accepté d'annuler deux années universitaires, en sa qualité de Directeur de l'U.F.R Science et Technologie de l'UGB, pour que je puisse me réinscrire et continuer à mener à terme les études.

A toute ma famille pour leur soutien tout au long de mon parcours universitaire.

Que ce travail soit l'accomplissement de vos vœux tant allégués, et le fruit de votre soutien infaillible.
Merci d'être toujours là pour moi.

REMERCIEMENTS

Je tiens tout d'abord à exprimer toute ma reconnaissance à Sophie Niang-Dabo, Professeure à l'Université de Lille, ma co-directrice de thèse, qui m'a tellement appris, et ce, depuis mon D.E.A. Elle a supervisé parfaitement l'encadrement tout au long de cette thèse malgré ses diverses responsabilités et son emploi du temps chargé. Je la remercie de m'avoir accordé sa confiance en acceptant de co-diriger ma thèse mais aussi de m'avoir donné la chance de vivre des expériences enrichissantes. Elle m'a facilité beaucoup de voyager d'étude et m'a fait participer à des conférences internationales. J'espère que nous aurons l'opportunité de collaborer durant les prochaines années.

J'exprime, également, ma reconnaissance à Monsieur Papa NGOM, Professeur à l'UCAD, qui a été aussi mon professeur en Statistique Bayésienne lorsque je faisais le D.E.A. Je me rappelle le jour où j'ai frappé à sa porte pour chercher un sujet de thèse.

J'exprime ma reconnaissance à monsieur Patrice BREHMER, Chargé de recherches (HDR) à l'Institut de Recherche pour le Développement (IRD), qui a été toujours présent du début à la fin. Je vous remercie pour votre collaboration et implication dans mes travaux de recherches, durant mes années de thèse. Sa contribution dans ce travail a rendu les résultats des applications simples à être interprétés. Il a beaucoup insisté sur la qualité des cartes et la simplicité dans les interprétations. Il facilité, aussi, le financement des ma dernière année de thèse et subventionné le dernier voyage d'étude. J'espère que nous aurons l'opportunité de collaborer durant les prochaines années.

Je remercie Monsieur Aliou Diop, Professeur à l'UGB, qui m'a beaucoup appris depuis ma licence. Il a encadré mon premier travail de recherche à travers un stage au Centre de Recherches Océanographiques de Dakar-Thiaroye (CRODT). Il m'a conseillé d'aller voir Papa Ngom, pour trouver un sujet de thèse.

Je remercie Messieurs Massal Fall et Ciré Elimane Sall pour avoir accepté de co-encadrer le travail à travers l'Institut Sénégalais de Recherches Agricôles (ISRA). Ils ont facilité les trois premières années d'allocation de recherche à l'ISRA. Ils ont assuré l'encadrement interne dans l'Institut d'accueil (ISRA).

Je remercie Docteur Ndiaga Thiam, Directeur du centre de recherche CRODT, d'avoir accepté de collaborer dans mes publications. Il a mis à ma disposition les ressources nécessaires, support d'application du Chapitre 4. Je te remercie surtout pour votre collaboration dans le Chapitre 5.

Je remercie Docteur Kamarel, chargé de recherche au CRODT pour son implication dans le chapitre 1. Il a mis à ma disposition une synthèse bibliographique récente sur les ressources demersales côtières et la description du milieu d'étude. J'apprécie à juste titre votre collaboration.

Je remercie Docteur Mohamed-Salem Ahmed, enseignant chercheur à l'Université de Lille, pour sa collaboration dans le chapitre 3 et la mise à ma disposition des programmes de simulations qui m'ont beaucoup inspiré dans les applications réelles.

Je remercie Monsieur Diaraf Seck, Professeur Titulaire à l'UCAD d'avoir accepté de présider mon jury thèse.

Je remercie mes rapporteurs de thèse Jean François Dupuy, Professeur Titulaire à Institut de recherche mathématique de Rennes (IRMAR) et Aliou Diop, Professeur Titulaire à l'UGB.

Je remercie les examinateurs Professeur Abdoluaye Sène Professeur Titulaire à Université du Sine Saloum El-Hâdj Ibrahima NIASS (USSEIN) et Docteur Patrice Brehmer chargé de recherches (HDR) à l'IRD.

Je remercie tous les membres du jury.

Je remercie toute l'équipe du laboratoire LEM de l'Université de Lille, Docteur Gharbi Zied, les doctorants Salsabil Yacour, Yoba Kandé , Fahariat BOUKARI, Solange, Alaa Ali hassan, et Moise Basse, Aladji Bassène.

Je remercie toute l'équipe mon centre d'accueil CRODT.

Je tiens à préciser que ce travail est le fruit d'une riche collaboration entre ISRA, UCAD, IRD, UGB et l'Université de Lille. L'ISRA a financé les trois premières années d'allocation de recherche. Il a mobilisé toutes les ressource nécessaires, à travers son centre CRODT, pour l'accomplissement du travail. L'IRD a financé la dernière année, à travers le projet Project Enhancing Prediction of Tropical Atlantic Climate and its Impacts (PREFACE) piloté par Docteur Patrice BREHEMR, Chargé de Recherche (HDR) à l'IRD. Il a subventionné le dernier voyage d'étude. Le consortium CEAMITIC a financé les bourses de mobilités et les participations dans les conférences internationales.

RÉSUMÉ

L'aménagement du secteur de la pêche maritime requiert, entre autres, des connaissances sur le fonctionnement de l'écosystème. En particulier, dans le cadre de la gestion de la ressource marine, une connaissance approfondie de l'environnement marin et des relations inter-spécifiques en son sein est indispensable. L'analyse spatio-temporelle, des facteurs environnementaux ou ceux liés à la pression de pêche qui influencent l'évolution de la ressource ainsi que sa production, s'avère ainsi d'une importance capitale. La présente étude se veut une approche écosystémique prédictive. Elle met en œuvre un modèle de relation explicative entre la ressource halieutique et son environnent marin. Cette étude est basée essentiellement sur une analyse spatiale et fonctionnelle de données concernant les ressources halieutiques, démersales côtières notamment, au large du Sénégal. Elle est réalisée à l'aide de la modélisation non-paramétrique spatiale dans le temps et dans l'espace, ce qui fait appel à des méthodes statistiques spatiales de nature infinie (données fonctionnelles). L'approche non-paramétrique est particulièrement adaptée pour la modélisation des données spatiales de grande dimension, en particulier pour les ressources halieutiques observées dans le temps et à des endroits spécifiques. En effet, elle considère des modèles qui prennent en compte une large classe de processus. Elle permet de caractériser et de quantifier les facteurs influençant l'évolution spatio-temporelle de la ressource au sein de l'écosystème marin, tout en tenant compte de la probabilité de capture de la ressource. Dans un premier temps, nous nous sommes intéressés aux aspects mathématiques des modèles de régressions. Deux approches, d'estimations de la fonction de régression, basées sur la méthode du noyau double ont été proposées. Elles sont définies suivant la structure des données et la nature des co-variables. La particularité de telles approches réside dans le fait de prendre en compte la proximité entre les sites (la structure spatiale) et celle entre les observations, dans un contexte de dépendance spatiale. Nous avons étudié les comportements asymptotiques des estimateurs proposés et établi ainsi la convergence presque complète entrainant celle presque-sûre. La première approche est une méthode finie (non-fonctionnelle) basée sur l'approche des k-plus proches voisins. Elle est très adaptée aux données dont la structure présente une certaine hétérogénéité. Une méthode de prédiction, et de classification en particulier, découle de l'estimateur des k-plus proches voisins proposé. La procédure de classification supervisée appliquée sur les données bio-écologiques donne les meilleures précisions de classement comparée aux méthodes classiques. La seconde, est une méthode de régression fonctionnelle très adaptée à des données de grande dimension. Une méthode de prédiction fonctionnelle et, particulièrement, une nouvelle procédure de classification supervisée découlent de l'estimateur de la régression fonctionnelle. La convergence ponctuelle et celle uniforme ont été établies. L'application de la prédiction fonctionnelle sur les données bi-écologiques donne les erreurs de prédictions plus petites que celles classiques. L'application de la nouvelle règle de classification supervisée sur les données bi-écologiques montre que celle-ci est une alternative par rapport à de récentes méthodes. Les résultats obtenus prouvent que cette méthode donne généralement plus de précisions dans les classements avec des taux de classifications correctes plus grands sur l'ensemble des classes. L'application nous a permis d'analyser la distribution spatio-temporelle des espèces démersales côtières et d'étudier celles-ci par rapport aux paramètres environnementaux qui influencent la variabilité et la distribution de la ressource.

Mots-clefs :
Estimation non-paramétrique, Estimateur à noyau, Régression, k-plus proches voisins, Statistique spatiale, Analyses de données fonctionnelles, Prédiction, Classification supervisée, Machine Learning, estimation de biomasse, prédiction de quantité de biomasse, Distribution spatiale des espèces démersales, Ressource halieutique du Sénégal.

ABSTRACT

Monitoring the maritime fishing sector requires, among others, knowledge of the functioning of the ecosystem. In particular, within the framework of the management of fishery resources, an in-depth knowledge of the marine environment and the inter-specific relationships within it is essential. The spatio-temporal analysis of environmental factors or factors linked to fishing pressure which influence the evolution of the resource as well as its production, is thus of capital importance. This study is intended to be a predictive ecosystem approach. It implements an explanatory relationship model between the fishery resource and its marine environment. This study is essentially based on a spatial and functional analysis of data concerning fishery resources, particularly coastal demersal, off Senegal. It is carried out using spatial non-parametric modeling in time and space, which calls for spatial statistical methods of an infinite nature (functional data). The non-parametric approach is particularly suitable for modeling large-scale spatial data, in particular for fishery resources observed over time and at specific locations. Indeed, it considers very large models making it possible to characterize and quantify the factors influencing the spatio-temporal evolution of the resource within the marine ecosystem, while taking into account the probability of capture of the resource. First, we were interested in the mathematical aspects of regression models. Two approaches to estimating the regression function based on the double kernel method have been proposed. They are defined according to the data structure and the nature of the co-variables. The particularity of such approaches lies in taking into account the proximity between the sites (the spatial structure) and that between the observations, in a context of spatial dependence. The first approach is a finite (non-functional) method based on the k-nearest neighbor approach. It is very suitable for data whose structure presents a certain heterogeneity. We have studied the asymptotic behaviors of the proposed estimators and thus established the almost complete convergence leading to the almost sure one. One prediction and classification method in particular derives from the proposed k-nearest neighbor estimator. In particular, the supervised classification procedure applied to the bi-ecological data gives the best classification precision compared to conventional methods. The second is a functional regression method very suitable for large-dimensional data. A functional prediction method and in particular a new supervised classification procedure derive from the functional regression estimator. Point convergence and uniform convergence have been established. The application of the functional prediction on the bi-ecological data gives the prediction errors smaller than the classical ones. The application of the new supervised classification rule on bi-ecological data shows that it is an alternative to recent methods. The results obtained prove that this method generally gives more precision in the classifications with higher correct classification rates on all the classes. The application allowed us to analyse the spatio-temporal distribution of coastal demersal species and to study them in relation to the environmental parameters that influence the variability and distribution of the resource.

Keywords :
Nonparametric estimation, Kernel estimator, Regression, k-nearest neighbors, Spatial statistics, Functional data analyzes, Prediction, Supervised classification, Machine Learning, biomass estimation, biomass quantity prediction, Spatial distribution of demersal species, Senegalese fishery resource.

TABLE DES MATIÈRES

LISTE DES FIGURES

LISTE DES TABLEAUX

CHAPITRE 1

INTRODUCTION GÉNÉRALE

La pêche maritime occupe la première place du secteur primaire de l'économie sénégalaise. Elle joue un rôle stratégique et de soutien à la croissance de l'économie nationale et contribue de façon déterminante à l'équilibre de la balance commerciale réduisant aussi celles des paiements et du chômage. Ce secteur constitue ainsi une composante principale dans la politique de l'état en matière de création d'emplois et de sécurité alimentaire.
Durant les trois dernières décennies, la pêche s'est intensément développée en Afrique de l'Ouest. Au Sénégal, l'effort de pêche a été multiplié par 2.5 au cours de la période 1981 – 2013, entraînant en même temps, une forte diminution de la biomasse de beaucoup d'espèces, notamment démersales côtières [25; 26; 32; 163; 332; 333; 335; 336; 337]. Les ressources halieutiques exploitées principalement au Sénégal comprennent les pélagiques, les démersaux côtiers et profonds[140]. Les espèces démersales côtières regroupent divers poissons crustacés et mollusques pêchés entre 0 et 200 m de profondeur à l'aide d'engins de fonds industriels et artisanaux [139]. Pour garantir une gestion durable de cette ressource marine, les autorités de l'état s'appuient, entre autres, sur l'expertise scientifique du CRODT. Un centre sous la tutelle de l'ISRA depuis 1979 qui y a domicilié son département de recherches océanographiques et halieutiques. Le centre fixe trois objectifs : l'évaluation et le suivi des ressources halieutiques, la compréhension des dynamiques de systèmes d'exploitations et la fourniture des bases techniques de mesures d'aménagement des pêcheries. Pour atteindre ses objectifs, le CRODT s'appuie sur des programmes de recherches pluridisciplinaires : ressources et milieux, dynamique des systèmes d'exploitation, gestion et aménagement des pêcheries et de leurs milieux. Les recherches mettent en œuvre des méthodes spécifiques de collecte d'informations à travers la pêche artisanale et des campagnes scientifiques menées à Bord du Navire Océanographique (N/O) Itaf Dème. Elles visent, ainsi, à la fois à acquérir l'ensemble des informations pertinentes sur le milieu (voir figure 1.1), la ressource et les systèmes d'exploitation, mais aussi à développer des outils de traitement et d'analyse de ces données dans le contexte particulier de gestion de la pêcherie sénégalaise. Ce dernier point est au coeur des préoccupations de cette présente contribution. Le milieu d'étude concerne l'écosystème marin qui couvre le plateau continental du Sénégal (voir figure 1.1) et la ressource ciblée dans ce travail est celle démersale côtière. Nous donnerons dans ce qui suit une description caractérisant l'interaction entre ce milieu d'étude, la nature de ses fonds, ses conditions hydro-climatiques et la ressource cible.

1.1 Milieu d'étude : Plateau continental du Sénégal

Le Sénégal est un pays sahélien doté d'une superficie terrestre de 200000 km^2 environ et d'un espace maritime de 198000 km^2 (figure 1.1). Il occupe la partie méridionale du bassin sédimentaire sénégalo-mauritanien. Globalement compris entre le rivage et l'isobathe 200 m, son plateau continental d'environ 23600 km^2 est peu accidenté [25; 26; 126]. Il dispose d'une frange littorale longue de 718 km et orientée Nord-Est Sud-Ouest pour la partie située au nord de la presqu'île du Cap Vert, et pour la partie sud, l'orientation est Nord-Ouest Sud-Est [25; 26; 125]. La façade maritime se situe entre 12°20′ et 16°03′N et est large de 240000 milles nautiques [270; 293]. La presqu'île du Cap-Vert, située à la latitude 14°40′N, sépare donc le domaine maritime sénégalais en deux régions aux caractéristiques topographiques distinctes [114; 201; 309]. Au nord de la presqu'île (figure 1.2), le plateau continental est étroit avec la présence d'une fosse au niveau de Kayar. Et au sud, le plateau est large, surtout au sud du pays (12°45′N). Au large de Saint-Louis (16°20′N), l'isobathe 200 m, situé à 27 miles limite le plateau à hauteur de Dakar, l'isobathe 200 m se situe à 5 miles

du trait de côte. En Casamance, l'isobathe nautique des 100 m atteint 54 miles. Dans [76], l'auteur estime la superficie totale du plateau continental à 28700 km^2. Elle est répartie comme suit : 4700 km^2 de fonds de 0 à 10 m, 14200 km^2 de 10 à 50 m et 9800 km^2 de fonds de 50 à 200 m. La configuration du plateau continental Sénégalais offre une diversité topographique diversifiée. Elle influence les comportements des espèces démersales qui l'habitent. Nous donnons dans les sections 1.1.1 et 1.1.2, respectivement, la nature des fonds et les conditions hydro-climatiques de la zone d'étude. Nous rappelons que la nature des fonds et les conditions environnementales jouent des rôles importants sur la distribution spatiale, la croissance et la vulnérabilité des espèces rencontrées dans cette zone.

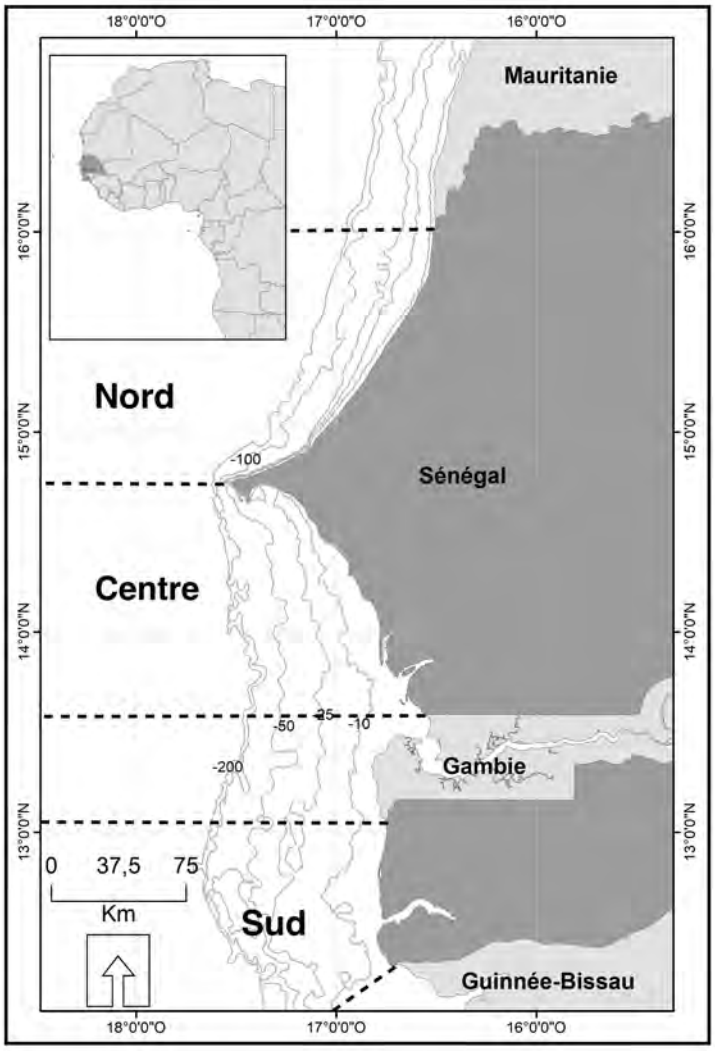

FIGURE 1.1 – Carte de la zone d'étude

1.1.1 Nature des fonds

Au Sénégal, les fonds marins (figure 1.2) ont été décrits en détail par [124; 125]. Les fonds, généralement peu accidentés, permettent une exploitation chalutière sur la plus grande étendue du plateau continental [124]. Le seul accident géographique remarquable est l'ensemble de la fosse Kayar–presqu'île du Cap-Vert qui coupe la région en deux. Cette particularité géographique limite la migration saisonnière vers le sud de certaines espèces démersales à affinité saharienne comme *Epinephelus aeneus, Epinephelus goreensis, Dentex filosus, Dentex canariensis, Pomadasys incisus, Umbrina canariensis* [25; 75; 124]. Enfin, la nature du fond a une influence sur la vulnérabilité des espèces aux engins de pêche : ainsi, sur les fonds rocheux, inaccessibles aux chalutiers, les poissons ne peuvent être capturés qu'à la ligne ou aux filets maillant, alors que sur certains fonds vaseux, des espèces qui s'enfouissent dans le sédiment comme les soles ou la crevette *Penaeus duorarum*, ne sont capturées que par les chaluts de fond munis de dispositifs permettant de fouiller la vase [25; 124].

- **Les fonds durs :** Les véritables fonds durs sont constitués de formations rocheuses. Sur le plateau continental, plusieurs bancs rocheux parallèles à la côte sont dénombrés à des profondeurs de 10 à 20 m au nord du Cap-Vert. Au sud, deux importantes falaises existent aux profondeurs 50 et 70 m, respectivement. Des plateaux rocheux côtiers se situent au niveau du Cap-Vert et entre Joal et Mbour. Les fonds durs sableux avec affleurement rocheux se retrouvent principalement de Dakar à l'embouchure du Saloum [124]. Les zones rocheuses du plateau continental jouent ainsi pour ces espèces, exploitées industriellement, un rôle de protection. Ces fonds durs sont très propices pour les activités de pêche à la ligne et les palangres. Les espèces de fonds durs (*Dentex filosus, Dentex canariensis, Plectorhinchus mediterraneus, Mycteroperca rubra et Epinephelus goreensis*) ont comme biotope normal la roche mais, elles peuvent vivre dispersées à proximité de celle-ci devenant ainsi vulnérables au chalutage. Leurs juvéniles se rassemblent aussi près de la côte, de la baie de Gorée jusqu'à l'embouchure du Saloum et deviennent alors vulnérables aux sennes de plage. De même, certains gros spécimens des espèces de fonds meubles telles que *Pagellus bellottii, Pagrus caeruleostictus* ou *Epinephelus aeneus* peuvent y trouver refuge.
- **Les fonds meubles :** Les fonds meubles correspondent à la couverture sédimentaire et sont constitués de fonds vaseux, sable-vaseux, vase sableuse et sableux propices aux engins de pêche constitués de filets comme le chalutage, les filets dormants de fond et les filets maillants habituellement endommagés par les rochers des fonds durs. Au niveau du littoral (< 20 m de profondeur), dans les fonds meubles de sable, de vase, de sable vaseux et de vase sableuse on peut trouver des espèces telles que *Pomadasys jubelini, Galeoides decadactylus, Ilisha africana, Drepana africana, Pentanemus quinquarius, Sphyraena dubia, Scyris alexandrinus, Argyrosomus regius, Brachydeterus auritus* [25].
  Les biotopes de la partie intermédiaire du plateau continental entre 20 et 80 m contiennent des espèces à affinité pour les fonds meubles et sableux (*Scyris alexandrinus, Brachydeterus auritus, Argyrosomus regius, Cynoglossus canariensis*), des espèces à affinité pour les fonds durs et sableux (juvéniles de *Zeus faber*) et des espèces à affinité pour les deux types de substrat (*Pagellus bellottii, Dentex filosus, Dentex canariensis, Plectorhinchus mediterraneus, Mycteroperca rubra, Pseudupeneus prayensis, Pagrus caeruleostictus, Epinephelus aeneus*) [25]. À la même gamme de profondeur (20 – 80 m), une importante vasière située de part et d'autre de l'embouchure du fleuve Sénégal et dans la zone comprise entre l'embouchure du fleuve Casamance et le large des îles Bissagos (figure 1.2) constitue un habitat idéal pour les juvéniles de la crevette côtière *Farfantepenaeus notialis* [8; 127; 161; 218; 226]. Les processus de fertilisation des eaux et de démarrage de la chaîne trophique s'y manifestent de façon intense du fait de la présence de matière organique (assimilée par le benthos) et de sels minéraux (nécessaire au développement de la mangrove observée entre le complexe du Delta du Saloum et sud du pays). Ces milieux sont considérés comme des nurseries pour les juvéniles de la crevette *Penaeus duorarum* et des espèces de la famille des *Sciaenidae.*
  Les fonds sablo-vaseux situés entre 80 et 200 m dominent la quasi-totalité de la surface du plateau et ceci jusqu'au nord des îles Bissagos (Guinée Bissau), avec des débris coquilliers plus ou moins importants, ainsi qu'une teneur élevée en carbonates (70%) [25; 124]. Ces fonds riches en matière organique sont favorables au développement du benthos (*lamellibranches, gastéropodes, polychètes, amphipodes*, petits crustacés : petits *macroures, stomatopodes, mysidacés* et larves de crustacés). Proie des poissons, ce benthos joue un rôle important dans la répartition et dans l'abondance des espèces démersales de la région.
  Les fonds correspondant au peuplement du rebord du plateau continental peuvent être recouverts de vase, de vase sableuse ou de sable vaseux. Ils regorgent des espèces à affinité pour la vase (*Brotula barbata, Dentex angolensis*) et pour les sables vaseux (*Dentex macrophtalmus, Epinephelus caninus* et *Zeus faber*).
  A Saint-Louis, le fond est constitué de sable contenant des débris coquilliers. De Dakar à l'embouchure du fleuve Casamance, les fonds de 0 et 70 m de profondeur sont constitués de sables grossiers et sont très riches en benthos, principale proie des poissons fouisseurs comme *Pseudupeneus prayensis.*
  Les sables fins rencontrés jusqu'à 50 m de profondeur, de l'embouchure du Saloum jusqu'à celle de la Casamance, sont un milieu trop compact pour permettre l'installation d'une faune interstitielle, d'où sa pauvreté en matière organique, en organismes du benthos [124; 218].
  En somme, nous pouvons dire que la nature du fond influe indirectement sur la répartition des espèces démersales [25; 124].

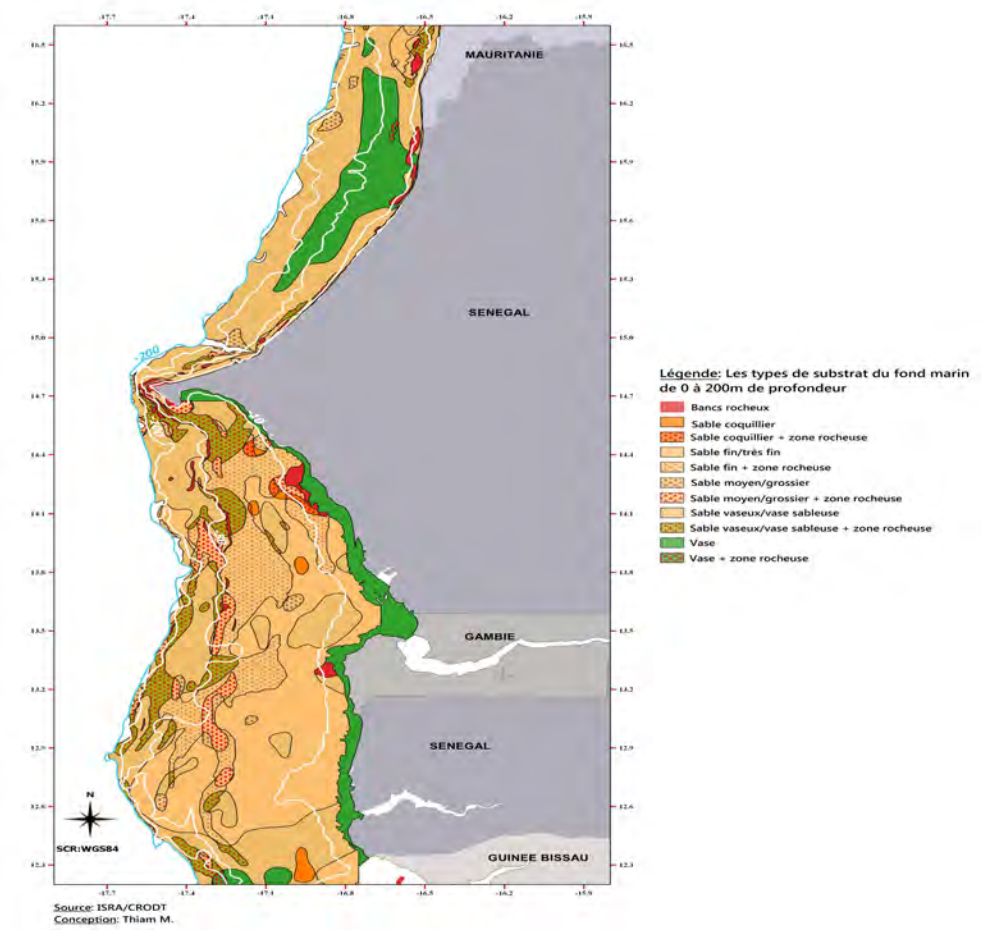

FIGURE 1.2 – Nature des fonds du plateau continental sénégalais.

1.1.2 Hydro-climat

Les caractéristiques de l'hydro climat dans la région ouest africaine ont été présentées dans de nombreux travaux [25; 124; 218; 225; 233; 270; 293]. Ces travaux ont permis de réaliser une synthèse sur l'environnement hydro-climatique de la région.

- **Les conditions océanographiques et hydrologiques :** Il existe deux principaux courants océaniques : le courant froid des Canaries et le contre-courant chaud équatorial. Les effets de ces courants sont variables sur le plateau continental. La température et le mouvement des eaux de surface sont sous l'influence de ces deux courants océaniques. Le courant froid des Canaries est caractérisé par des eaux de température inférieure à 20°C et de salinité comprise entre 35,4 et 36. Sa direction est nord-sud en longeant ainsi la côte par une dérive partielle ouest. Cette masse d'eau, dite Canarienne, recouvre progressivement tout le plateau continental sénégalais de Novembre à Mai. L'intensité de ce courant est maximale durant la période de renforcement des vents alizés.
  Le contre-courant équatorial véhicule deux types de masses d'eau chaude sur le plateau continental : (i) une eau tropicale chaude (plus de 24°C) et salée (salinité de 36) provenant du large et (ii) les eaux guinéennes chaudes et dessalées (moins de 35). La masse d'eau tropicale apparaît à la fin des alizés et est présente sur la presque totalité du plateau de Mai à Août. Les eaux guinéennes font suite aux importants apports d'eau douce par les fleuves de Guinée Bissau et de Guinée. Ces eaux remplacent celles tropicales sur le plateau et sont présentes de septembre à début Novembre. Les effets conjugués du climat et des courants marins permettent de distinguer deux saisons marines et deux périodes de transition :
  - une saison froide (Novembre à Mai) caractérisée par un refroidissement des eaux de surface et une salinité élevée des eaux côtières en raison de la baisse du débit des fleuves ;
  - une saison chaude (Juin à Octobre) caractérisée par l'installation d'une couche de surface isotherme (25 à 27°C) et homogène de 20 à 60 m d'épaisseur ;

— les périodes de transition entre la saison froide et chaude, et vice versa, correspondant à une durée d'environ un mois, entre mi-Mai et mi-Juin, puis entre mi-Octobre et mi-Novembre.

Il convient de noter que les dates d'apparitions des saisons et leurs durées sont caractérisées par une variabilité spatiale et inter-annuelle pouvant être importante. Ainsi, la durée de chaque saison, d'une année à l'autre, peut varier de l'ordre d'un mois.
Lors de la saison chaude, une thermocline de profondeur variable se crée avec la couche profonde plus froide et entraîne une variabilité spatio-temporelle de la biomasse des espèces marines d'une année sur l'autre [25; 74; 213]. Les changements hydrologiques influencent fortement la répartition spatiale et temporelle de nombreuses espèces et la productivité du milieu [218]. En effet, dans les zones tropicales, les apports nutritifs limitent la production primaire. Ainsi, l'upwelling côtier, les apports terrigènes par les alluvions des fleuves et plus faiblement les courants côtiers contribuent à enrichir le milieu marin [124].
L'upwelling côtier est une remontée d'eau froide profonde riche en sels minéraux. Sous l'effet du vent et de la force de Coriolis, une couche superficielle (quelques dizaines de mètres) ou couche d'Ekman se déplace vers le large. Ainsi, un flux vertical d'eau profonde (chargée en sels minéraux) se crée le long du talus continental afin de compenser le déséquilibre. L'intensité de l'upwelling dépend de la vitesse des alizés qui le génèrent et de l'orientation de la côte [94; 237]. Au Sénégal, l'upwelling est plus fort de Janvier à Avril avec une différence d'intensité de part et d'autre de la presqu'île du Cap Vert. La rétention côtière est également un processus local et saisonnier lié à la topographie du plateau continental ouest africain et à l'upwelling [113]. La topographie de sa Petite Côte combinée à un upwelling saisonnier offre un cadre idéal pour former une zone de forte rétention côtière favorable au développement de vie larvaire et juvénile pour beaucoup d'espèces comme *Octapus Vulgarus* [113].

- **Conditions climatiques :** Au Sénégal, il existe de fortes variations climatiques liées à l'influence de trois masses d'air (les alizés, l'harmattan et la mousson). La Zone Intertropicale de Convergence (ZITC) sépare ces masses d'air qui régissent les saisons. Cette zone est la surface de discontinuité en vent, température et humidité [25]. En effet, le mouvement de la ZITC entraîne une alternance très nette des saisons sur l'ensemble du littoral.
  - La saison sèche avec le régime d'alizés de nord s'étend de Novembre à mai. Les vents alizés ont pour origine l'anticyclone des Açores. Ils sont à la source de l'upwelling côtier qui s'étend jusqu'au sud du Sénégal, avec une forte variabilité spatio-temporelle du signal thermique.
  - Durant les périodes d'accalmie des alizés, le vent chaud et sec de l'harmattan en provenance du continent souffle entre Avril et Mai et peut-être fortement chargé en poussière.
  - Le déplacement de la ZITC vers le nord entraîne un régime de vent d'ouest (vent de mousson) chaud et humide, responsable de la pluviométrie de Juin à Octobre. Cette pluviométrie entraîne la baisse de la salinité au niveau des embouchures des fleuves. Par contre, le régime saisonnier des vents régit dans une certaine mesure, la possibilité de chalutage des navires et de sortie en mer des piroguiers.
- **Productivité et écosystèmes à mangroves :** La production primaire dans les eaux ouest africaines est la conséquence directe de l'enrichissement du milieu, de l'ensoleillement important et de la température élevée. Ainsi, les travaux de [124; 233], parmi bien d'autres, citent cette région comme l'une des plus productives au monde. Néanmoins cette production varie saisonnièrement en fonction de la latitude. En zone d'upwelling, la production est maximale en fin de période, soit au mois de mai, du sud de la Mauritanie jusqu'au sud du Sénégal [218]. La turbidité de l'eau limite certes la photosynthèse durant la saison des pluies [277], mais l'apport des eaux continentales (à la sortie des estuaires) riches en matière organique (directement assimilée par le zoo plancton) favorisent la production zoo planctonique.
  Ces habitats (côtiers et estuariens) restreints et instables constituent des nourriceries où de nombreuses espèces passent au cours de leur cycle biologique [123; 356]. Les facteurs explicatifs de l'attrait des juvéniles de beaucoup d'espèces dans les milieux estuariens, lagunaires et de mangroves sont multiples. Ces milieux, souvent turbides et touffus (racine de palétuviers, herbiers marins, zones peu profondes), offrent aux juvéniles et aux larves une nourriture abondante, des biotopes d'eaux calmes et une protection contre les prédateurs [303; 356; 363]. Ces conditions favorables à la survie des larves et juvéniles, selon [46], combinées aux températures relativement élevées dans les estuaires favorisent la croissance des espèces [283].

1.2 Contributions originales de la thèse

Les caractérisations du milieu (zone d'étude : figures 1.1 et 1.2) montrent que la topographie, la nature des fonds et les conditions hydro–climatiques déterminent l'aire de distribution des espèces démersales côtières. Elles influent le cycle biologique de ces dernières et gouvernent leurs migrations verticales et horizontales. Elles agissent sur les différentes phases du processus de maturation des démersaux. La variabilité de la biomasse de la ressource marine, son abondance et sa richesse sont en relation avec les différents facteurs qui définissent les conditions du milieu marin.
Le diagnostic du milieu donne des informations capitales sur l'environnement marin et la relation complexe des espèces et leurs comportements au sein de l'écosystème. En plus de ces informations recueillies sur la ressource, le milieu et leurs interactions, nous rappelons que, les programmes de recherches définis par le centre CRODT, mettent en œuvre des outils de traitement et d'analyse de ces informations pour une meilleure compréhension du comportement de la ressource dans son environnement. Le constat est que les outils de modélisation élaborés ne prennent pas en compte toutes les informations apportées par les données biologiques et environnementales. La dimension spatio-temporelle et l'aspect fonctionnel des données sont ignorés dans les analyses statistiques. Des approches écosystémiques explicatives et prédictives tenant compte des structures spatiales et l'aspect fonctionnel s'avèrent nécessaires pour la pérennisation de la ressource halieutique.
Les données spatiales sont des données pour lesquelles une information géographique est attachée à chaque unité statistique. L'information (localisation) dans notre contexte d'étude est la position de l'unité dans un référentiel spatial ou spatio-temporel. Elles peuvent être de nature finie appartenant à des espaces métriques. Elles peuvent aussi être de dimension éventuellement infinie appartenant à des espaces semi-métriques. Ces dernières sont appelées données fonctionnelles (une courbe est une donnée fonctionnelle). Les images satellites, les séries chronologiques, les courbes spectrométriques, les profils océanographiques de lumière et de chlorophylle ne sont que quelques exemples illustrant le grand nombre et la diversité des données fonctionnelles auxquelles le statisticien peut être confronté. Nous retrouvons ces types de données dans beaucoup de domaines, comme le montrent les travaux de [35] sur des profils océanographiques ainsi que ceux de [104; 263] sur des données biologiques. Les récentes innovations réalisées sur les appareils de mesures et l'utilisation intensive des moyens informatiques permettent souvent de récolter des données discrétisées sur des grilles de plus en plus fines, ce qui les rend intrinsèquement fonctionnelles. De plus, même si les données dont dispose le statisticien ou le biologiste ne sont pas de nature fonctionnelle, il peut être amené à étudier des variables fonctionnelles construites à partir d'un échantillon initial. Un exemple classique est celui où l'on observe plusieurs échantillons de données réelles indépendantes et où l'on est ensuite amené à comparer les densités de ces différents échantillons ou bien à considérer des modèles où elles interviennent (se référer à [289] pour plus de détails). Enfin, les variables multivariées peuvent être perçues comme celles fonctionnelles particulières. C'est pourquoi l'étude des variables fonctionnelles repose en grande partie sur la généralisation de l'étude des méthodes multivariées. Un traitement statistique de telles données qui ignore la dimension spatio-temporelle ou/et l'aspect fonctionnel ou les intègre de façon inadéquate repose sur une perte d'information importante. Cela conduit à faire des erreurs de spécifications ; ce qui peut résulter à construire des estimateurs non convergents et moins efficaces. L'étude paramétrique de la modélisation des données spatiales et fonctionnelles est très abondante. Elle repose principalement sur des hypothèses très restrictives comme la stationnarité stricte et l'hypothèse de la connaissance, à priori, de la loi du mécanisme qui génère les données. Ces hypothèses ne sont pas vérifiées par tous les processus. L'approche non-paramétrique est une alternative plus flexible et assez souple. Elle prend en compte une large classe de processus et elle repose sur des conditions plus générales et moins restrictives. Cette approche ne suppose aucune loi de probabilité au mécanisme qui produit les données. Elle se construit à travers les informations apportées par les observations.
Dans ce travail, la modélisation des données de grande dimension fait appel à des outils mathématiques non-paramétriques pour l'élaboration des estimateurs de régression. Ces derniers sont des supports de base des méthodes de prédictions et *classifications supervisées*. Nous nous intéressons ainsi, principalement, à la prévision spatio-temporelle appliquée à la ressource démersale côtière, dans la zone économique exclusive du Sénégal. Différentes approches ont été proposées selon la nature du support des données et le contexte d'étude. Dans un premier cas, nous étendons dans un contexte d'hétérogénéité spatiale l'estimateur proposé par [98]. L'approche proposée est basée sur la méthode des k-plus proches voisins. En suite, nous proposons la généralisation du même estimateur de [98] dans le contexte fonctionnel. Dans les deux cas, le support d'étude est basé sur le design fixe, c'est à dire les données sont collectées selon un plan d'échantillonnage déterministe. La construction des estimateurs repose sur deux noyaux. L'un contrôle la structure spatiale et l'autre mesure la proximité entre les observations. Les propriétés des estimateurs sont étudiées par le biais de la convergence presque complète ou sûre. L'implémentation des estimateurs est

faite sur des simulations numériques et sur des données réelles.
Ce travail est donc, une contribution, d'une part, à la statistique spatiale et fonctionnelle proposant ainsi des outils mathématiques de modélisations des données spatiales de grande dimension. Il contribue, d'autre part, à une meilleure gestion de la ressource halieutique, en offrant des outils d'évaluation de stock.
Nous exposons dans le Chapitre 2 un résumé de synthèse de la littérature sur la régression spatiale et fonctionnelle, ses applications ainsi que notre contribution dans ce thème de recherche. L'estimateur des k-plus proches voisins est abordé dans le Chapitre 3. L'idée de prendre en compte la structure locale des données est d'incorporer localement, dans la construction de l'estimateur, les sites proches ayant plus de similitudes entre eux. Les fenêtres de lissage sont des paramètres aléatoires appartenant à des ensemble discrets, contrairement à l'estimateur de [98]. La convergence presque complète de l'estimateur est étudiée, entre autres, sous la condition locale stationnaire et celle de α-mélange. Cet estimateur permet de définir un prédicteur qui, dans le cas particulier où la variable réponse est de nature discrète, est une nouvelle méthode de *classification supervisée*.
La méthode d'estimation de la régression fonctionnelle est abordée dans le Chapitre 4. Nous généralisons les travaux de [98] en considérant une co-variable fonctionnelle. La convergence ponctuelle et celle uniforme sont établies avec des vitesses de convergences sous des conditions, entre autres, de régularité du modèle, de la stationnarité locale et de α-mélange. Cet estimateur permet de définir un prédicteur fonctionnel qui, dans le cas particulier où la variable réponse est de nature discrète, est une nouvelle méthode *classification supervisée*. Cette procédure de *discrimination* incorpore la structure spatiale et l'aspect fonctionnel des données.
Le Chapitre 5 est entièrement consacré à l'application des prédicteurs et règles de discrimination définis dans les Chapitres 3 et 4 respectivement. L'application du Chapitre 5 porte sur la prédiction de la distribution spatiale des démersaux ainsi que leurs quantités de biomasse au large du Sénégal.
Le Chapitre 6 est réservé à la conclusion et aux perspectives.

1.2.1 Publications, Séminaires, ateliers et communications écrites et orales

Travaux et Publications

- Nonparametric Prediction for Spatial Dependent Functional Data : Application to Demersal Coastal Fish off Senegal (Chapitre 2 [269] du livre [243]), voir http://www.iste.co.uk/book.php?id=1601.
  Auteurs : Mamadou NDIAYE, Sophie DABO-NIANG, Papa NGOM, Ndiaga THIAM, Massal FALL, Patrice BREMER
- k-nearest neighbors prediction and classification for spatial data Journal de la Société Française de Statistique (arXiv :math.ST/1806.00385v1), en révision à Journal de la SFDS.
  Auteurs : Mohamed-Salem AHMED, Mamadou NDIAYE, Sophie DABO-NIANG
- Nonparametric prediction and supervised classification for spatial dependent functional data under fixed sampling design, Afrika Statistika, A soumettre.
  Auteurs : Mamadou NDIAYE, Sophie DABO-NIANG, Pape NGOM, Massal FALL, Ndiaga THIAM, Patrice BREMER.

Séminaires et Conférences

- Nonparametric spatial model using size and point transect for monitoring density population of demersal fish in coastal Senegalese sea water, Laboratoire Lille Economie Menagement (LEM) de l'Université de Lille, jeudi 24 Novembre 2016.
- Nonparametric prediction of spatial dependent functional data, Lebanese International Conference on Mathematics and Applications (LICMA)'17, from 16 to 19 May 2017, Lebanese University Faculty of Sciences I, Doctorate School of Science and Technology, Beirut, Lebanon http://www.licma.net.
- Nonparametric spatial model using size and point transect for monitoring density population of demersal fish in coastal Senegalese sea water, Journées scientifiques du Comité Scientifique et Technique (CST) de l'ISRA, SESSION 2018,05-09 Février 2018. " Changements climatiques et développement agricole durable : stratégies d'adaptation des acteurs et nouveaux paradigmes de la recherche"
- k-nearest neighbors prediction and classification for spatial data, International Conference on Biomathematics in Senegal (ICBS), from 29 June to 01 July 2018, école polytechnique de thiès, Sénégal.
- Application of functional classification on high resolution oceanographic data in Canaries current large marine ecosystem : toward fine scale analysis, International Conference on Ocean, Climate and

Ecosystems and PREFACE Final General Assembly, Lanzarote, Spain https://preface.w.uib.no/output/presentations/#GA2018

1.2.2 Participation à des ateliers, séminaires et journées d'études

— AIMS-Senegal & SWMA Workshop on Towards a new generation of common pool resource experiments : Using a mobile app to analyse dynamic human harvest behaviour, September, 18th and September, 20th 2015.This event is jointly organized by AIMS-Senegal and the Centre for Tropical Marine Ecology of Bremen in Germany and supported by DAAD in Germany.

— AIMS-Senegal & SWMA Workshop on Financial and Actuarial Mathematics July, 11-15th, 2016 AIMS Senegal, Mbour, Senegal.

— Learning with functional data,University Lille(IUT "A"), October 7th, 2016.

— Workshop on Statistical Methods for Recurrent Data, lundi 7 Novembre, 2016, Maison de la Recherche, l'Université de Lille.

— Africa winter school AIMS/ZMT 2020 Workshop on "Mathematical Modeling of Ecological and Socioeconomic Systems",AIMS-Senegal, Mbour, 27th to 31st January 2020.

— Training Course on Trawling and Acoustic Surveys-Durban, South Africa, 29 June-17 July 2015.

— RTC-SN-OTGA, l'Atelier de formation sur les principes fondamentaux sur la gestion des données Océanographique, 25 au 29 Janvier 2016, Direction Générale de l'ISRA, Sénégal.

— RTC-SN-OTGA, l'Atelier de formation sur la gestion des données sur la biodiversité marine,17 au 20 Juillet 2017, Direction Générale de l'ISRA, Sénégal.

CHAPITRE 2

REVUE DE LA LITTÉRATURE ET CONTRIBUTIONS

2.1 Revue de la littérature

Nous donnons dans ce qui suit un état de l'art sur l'estimation non-paramétrique à noyau dans un cadre spatial et fonctionnel, domaines qui concernent la contribution de la thèse.

2.1.1 Statistique non-paramétrique : Estimation à noyau

La statistique non-paramétrique est l'ensemble des méthodes d'estimations qui ne supposent pas que les lois de probabilités qui régissent les distributions des données étudiées sont paramétriques. Un modèle non-paramétrique repose, en partie, sur des hypothèses de régularités (continuité, dérivabilité,...) du modèle choisi dans une certaine classe $\mathscr{P}$ [345; 361].
Parmi les multiples problèmes non-paramétriques rencontrés dans la littérature, nous mettons l'accent sur l'estimation de la fonction densité et celle de régression. Ses dernières sont étudiées, en partie, à travers des techniques d'interpolation, de lissage ou d'approximation. Elles se distinguent en trois approches : les bases de splines, l'approche par projection et celle par noyau (dite de Parzen-Rozenblatt),... Dans notre travail, nous utilisons la méthode à noyau. Elle est introduite par [305] puis généralisée par [276].

Estimateur à noyau de la fonction densité dans le cadre réel

La fonction de densité de probabilité f d'une variable X est un concept fondamental en probabilité et en statistique. Considérons un échantillon de taille n d'une variable aléatoire réelle X, dont les composantes sont des variables $X_1,...,X_n$ définies dans un espace de probabilité, indépendantes et identiquement distribuées. L'estimateur à noyau de f en un point x est donné par :

$$f_n(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right), \tag{1.1}$$

où h est un paramètre de lissage appelé "fenêtre" et K est une fonction de poids appelée "noyau", K vérifie certaines conditions, comme :

$$\int_{-\infty}^{+\infty} K(x)dx = 1 \text{ et} \int_{-\infty}^{+\infty} K(x)^2 dx < \infty.$$

Estimateur à noyau de la fonction de régression dans le cadre réel

Soient $(X_1,Y_1),...,(X_n,Y_n)$, n un échantillon i.i.d (indépendant et identiquement distribué) de même loi qu'un couple de variables réelles (X,Y). Nous supposons qu'il existe une relation non-linéaire entre les composantes de ces couples, décrite par le modèle suivant :

$$Y_i = r(X_i) + \varepsilon_i, i = 1,...,n,$$

ε_i est le terme aléatoire non expliqué par le modèle. Il est centré et indépendant des variables X_i. La variable X est explicative, co-variable ou régresseur et Y est appelée variable réponse ou à expliquer. La fonction de

lien $r(.)$ détermine le modèle. En effet le modèle est linéaire si la fonction de lien r est une combinaison linéaire des composantes de X_i.
Dans cette étude, nous considérons que $r(.)$ est l'espérance conditionnelle de Y sachant X. L'estimateur de régression consiste tout simplement à estimer la fonction $r(.)$ à travers les observations $(X_1, Y_1), ..., (X_n, Y_n)$. La construction de l'estimateur repose, en partie, sur la fonction de poids K et la fenêtre h qui contrôle le poids alloué à chaque observation qui entre dans la construction. L'estimateur classique est celui de Nadaraya-Watson introduit par [268; 357]. Il est ainsi défini :

$$r_n(x) = \begin{cases} \dfrac{\frac{1}{nh}\sum_{i=1}^{n} Y_i K\left(\frac{x-X_i}{h}\right)}{f_n(x)} \text{ si } f_n(x) \neq 0; \\ \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} Y_i, \text{sinon.} \end{cases} \tag{1.2}$$

Ces estimateurs à noyau (équations (1.1) et (1.2), respectivement) ont été largement étudiés, sous divers contextes, dans la littérature, voir les monographes de [220; 345; 361] pour plus de détails.

2.1.2 Statistique spatiale

La statistique spatiale étudie les phénomènes observés sur un ensemble spatial S. L'ensemble S est généralement constitué de sites géographiques. Les observations spatiales sont des réalisations d'un champ aléatoire Z sur S, c'est à dire une collection $Z = \{Z_{\mathbf{i}}, \mathbf{i} \in S\}$ de variables aléatoires indexées par l'ensemble spatial S. La connaissance des localisations de ces sites d'observations et les valeurs des phénomènes mesurés dans ces sites constituent l'objet d'étude en statistique spatiale. Cette dernière s'exprime dans les champs d'application comme l'océanographie, les sciences halieutiques, l'agronomie, la météorologie, la foresterie, la géologie, l'imagerie, l'épidémiologie, la qualité de l'air, les sciences atmosphériques et celles du sol, etc.

Nous pouvons citer plusieurs situations réelles illustrant des problématiques liées à la modélisation spatiale des données. En science halieutique et océanographie, la connaissance de l'état de la ressource marine et celle de sa distribution spatiale, dans le temps, est nécessaire pour une bonne gestion des ressources halieutiques. Des travaux ont été faits dans ce domaine, voir entre autres les références [24; 78; 85; 111; 166; 221; 267; 273; 300; 301; 324; 326; 338; 360]. Plusieurs approches d'analyse de la répartition spatiale de cette ressource, ont été considérées dans ces travaux; ciblant une large variété d'écosystème marin. Les résultats issus de ces travaux ont contribué à l établissement des bases d'aménagement pour la pérennisation des pêcheries et la conservation de la biodiversité marine. D'autres exemples d' application de la statistique spatiale concernent l'étude de la fertilité des sols, en agronomie [198; 234; 359; 368], la mesure de la pollution de l'air et des sols, en science de l'environnement et celle atmosphérique [67; 159; 202; 231; 244; 349], l'évaluation de gisements miniers et pétroliers, en géologie [62; 95; 229; 238; 294; 295; 315; 331] et l'analyse de la dynamique évolutive des individus infectés d'une maladie transmissible dans une région, en épidémiologie [44; 187; 204; 261; 266; 279; 313; 322]. Récemment, la pandémie covid-19 [1] a soulevé beaucoup de questions pour comprendre la dynamique évolutive de la maladie et ses vitesses de diffusion dans les pays du nord et ceux du sud, faisant appel à la modélisation spatiale [82; 108; 148; 167; 195; 259; 280; 354; 381].
La statistique spatiale peut être subdivisée en trois cadres selon le support de mesure des données et l'objectif fixé. Lorsque le support de mesure des données est continu c'est-à- dire celles-ci peuvent, en principe, être mesurées en tout point d'un domaine continu (par ex. la teneur en matière organique dans un champ agricole, la salinité ou la température des océans, la teneur en gisements de pétrole en mer ou dans le sol, la quantité de butane en mer,...), on utilise le formalisme des champs aléatoires continus; c'est le domaine habituel de la géostatistique. Lorsque le support de mesure des données est lié à un réseau (des images ou mesures récoltées sur des entités administratives, etc...), on parle de données latticielles, c'est le cas des champs de Markov spatiaux. Enfin, lorsque l'on s'intéresse aux coordonnées dont on suppose être porteuses d'informations (position des arbres ou animaux dans une forêt, localisation des personnes infectées dans une région en épidémiologie etc.), on parle de processus ponctuels spatiaux.
La statistique spatiale s'intéresse à la modélisation descriptive, explicative et prédictive des données spatiales. La problématique réelle définie, dans un contexte spatial, fait appel à la méthode utilisée en relation avec le cadre d'étude. L'ensemble spatial S permet au statisticien de définir le cadre c'est-à-dire le type de données (géostatisitique, latticielles, processus ponctuels) dont il dispose et les méthodes spatiales requises pour l'analyse et la modélisation de ces données. Dans le cas de la géostatistique, principalement considérée dans ce travail, S est un sous ensemble fixé d'un espace euclidien de dimension N, soit $\mathbb{R}^N, N > 1$.

1. La maladie à coronavirus 2019, abrégée en COVID − 19 (acronyme anglais signifiant coronavirus disease 2019), est une maladie infectieuse émergente de type zoonose virale, provoquée par le coronavirus SARS − C*o*V − 2 (ex 2019-nCoV), responsable d'une pandémie ayant débuté en décembre 2019 dans la ville de *Wuhan*, capitale de la province du *Hubei*, en Chine centrale

Les sites localisés dans S, sont notés par $\mathbf{i} = (i_1, i_2..., i_N)$. Ainsi, des cartes d'interpolation ou de prédiction peuvent être dessinées à travers différentes méthodes de Krigeage ou co-krigeage.
La définition des contours d'un ensemble de méthodes statistiques dédiées à l'étude et la modélisation spatiale apparaît durant la première moitié du 20^{ieme} siècle. Les premiers résultats ont été établis par [83; 91; 158; 298]. L'approche paramétrique, connue sous le nom de "Krigeage", est la première méthode élaborée en géostatistique. Ce terme provient du nom de famille de l'ingénieur sud-africain *Daniel Gerhardus Krige*. Le krigeage est une méthode de prédiction spatiale formalisée pour la prospection minière, par *Georges Matheron*, à l'écôle des mines de Paris, voir [251]. Nous pourrons nous référer à la monographie de [350] pour une introduction à la modélisation géostatistique (Krigeage). Cependant, l'application des méthodes de Krigeage repose sur des hypothèses assez restrictives, comme celle gaussienne, sur le processus. Cela n'est pas un critère vérifié par tous les processus. En effet, dans le domaine biologie marine, il n'est pas évident que les données de densités de poissons soient générées par un mécanisme gaussien. Certains auteurs ont proposé des méthodes non-paramétriques qui reposent sur des hypothèses plus générales [41; 64; 66; 98; 105; 106; 130; 343]. Ces nouvelles méthodes, alternatives, prennent en compte une large classe de processus observés dans divers champs d'applications. L'estimation de la fonction densité de probabilité d'un champ spatial constitue la première étape dans cette direction.

Estimation non paramétrique spatiale : cadre réel géostatistique

L'approche non-paramétrique spatiale est de nos jours au cœur d'une dynamique de recherche avec de nombreuses applications dans beaucoup de domaines. Les premiers résultats issus de la modélisation non-paramétrique des données spatialement dépendantes sont l'extension des études menées en série chronologique, voir par exemple les travaux de [86; 112; 250; 302; 316; 344; 365; 366]. Ces derniers étudient les processus spatiaux sur des régions rectangulaires, $\mathscr{I}_{\mathbf{n}}$, définies par : $\mathscr{I}_{\mathbf{n}} = \{\mathbf{i} = (i_1, ..., i_n) : 1 \leq i_k \leq n_k, k = 1...N\}$, pour $\mathbf{n}$=$(n_1,...,n_N) \in \mathbb{R}_+^N$. Ces régions sont considérées pour l'estimation non-paramétrique de la fonction densité et celle de la régression spatiale.
Dans ce qui suit, nous supposons que les observations sont collectées dans, $\mathscr{I}_{\mathbf{n}}$, suivant un design fixe et nous posons $\widehat{\mathbf{n}} = n_1 \times n_2, ...n_N$, pour désigner la taille de l'échantillon.

Estimateur à noyau de la fonction densité de probabilité

L'estimateur $f_{\mathbf{n}}$ de la densité f d'une variable aléatoire spatiale $X \in \mathbb{R}^d$ se construit à partir des observations $\{X_{\mathbf{i}}, \mathbf{i} \in \mathscr{I}_{\mathbf{n}}\}$. Soit $x \in \mathbb{R}^d$, nous avons :

$$f_{\mathbf{n}}(x) = \frac{1}{\widehat{\mathbf{n}} h_{\mathbf{n}}^d} \sum_{\mathbf{i} \in \mathscr{I}_{\mathbf{n}}} K\left(\frac{x - X_{\mathbf{i}}}{h_{\mathbf{n}}}\right), \tag{1.3}$$

où K est une fonction noyau et $h_{\mathbf{n}}$ est une séquence de nombres réels positifs qui tendent vers 0 quand $\mathbf{n}$ tend vers l'infini. L'estimateur classique (1.3) de la fonction densité de probabilité d'une variable aléatoire spatiale a été étudié dans le cas discret ($\mathbf{i} \in \mathbb{Z}^d$) et dans le cas continu ($\mathbf{i} \in \mathbb{R}^d$). Les propriétés asymptotiques de l'estimateur ont été établies par [39; 41; 65; 66; 105; 134; 142; 185; 341]. Notons que le choix du paramètre de lissage $h_{\mathbf{n}}$ est déterminant. Il se fait, souvent, suivant la structure des données. Ce paramètre appelé "bandwith" en anglais, joue un rôle important et contrôle, en quelque sorte, la précision des estimations. Il a une influence sur le biais et la variance et contrôle, en partie, les vitesses de convergence des estimateurs. Des approches différentes de sélection du paramètre de lissage sont proposées dans la littérature [109; 193; 205]. Dans certaines situations, quand les données présentent une certaine hétérogénéité locale spatiale, le choix de $h_{\mathbf{n}}$ doit prendre en compte la nature de la structure spatiale des données. Si cela ne se fait pas, le risque de produire des estimateurs moins performants, se poserait. La méthode k-NN est une approche alternative, à l'estimateur de l équation (1.3), pour contourner ce problème. Elle est proposée par [154] et puis généralisée par [240]. Ces auteurs proposent un paramètre de lissage k qui dépend des plus proches voisins de x.
Toutefois, malgré ces efforts, ces méthodes ne couvrent pas toutes les situations possibles. Certaines difficultés demeurent toujours quand le contexte d'étude et d'application des modèles fait appel à de nouveaux paramètres à tenir en compte dans les modélisations. L'estimation par la méthode des noyaux récursifs a été proposée pour améliorer certaine difficultés liées à l'acquisition des données spatiales volumineuses collectées dans divers domaines d'applications. Cette méthode permet, entre autres, de réduire le temps de calcul dont souffrent les méthodes classiques quand il est question de modéliser des données volumineuses. L'approche récursive a d'abord été proposée dans le cas non-spatial (voir [248; 249; 358; 362]), puis étendue dans le cadre spatial [12; 50; 364]. Ces derniers proposent une estimation de la densité par la méthode des noyaux récursifs pour des processus spatiaux. Les travaux récents de [52; 53] établissent une nouvelle approche basée sur l'approximation stochastique.

Une autre approche d'estimation par noyau qui prend en compte la structure spatiale est étudiée par [99]. Les auteurs s'inspirent des travaux de [352] et proposent une méthode d'estimation de la densité de probabilité qui se base sur deux noyaux. L'un des noyaux prend en compte la structure spatiale par la mesure de la proximité entre les valeurs observées du champs et l'autre contrôle la distance entre les observations, soit :

$$f_{\mathbf{n}}(x) = \frac{1}{\widehat{\mathbf{n}} h_{\mathbf{n}}^{d} \rho_{\mathbf{n}}^{\mathrm{N}}} \sum_{\mathbf{i} \in \mathscr{I}_{\mathbf{n}}} \mathrm{K}_1\left(\frac{x - \mathrm{X}_{\mathbf{i}}}{h_{\mathbf{n}}}\right) \mathrm{K}_2\left(\rho_{\mathbf{n}}^{-1} \left\| \frac{\mathbf{i}_0 - \mathbf{i}}{\mathbf{n}} \right\|\right). \tag{1.4}$$

K_1 et K_2 sont des noyaux et $h_{\mathbf{n}}$ et $\rho_{\mathbf{n}}$ des paramètres de lissage qui tendent vers 0 quand $\mathbf{n} \to \infty$. L'auteur propose un estimateur du mode spatial à partir de d'équation (1.4). Il établit la convergence ponctuelle et celle uniforme de l'estimateur de l'équation (1.4), sous les conditions de stationnarité locale. Il en déduit ainsi les propriétés asymptotiques de l'estimateur du mode spatial.

Estimateur à noyau de la fonction de régression spatiale

Le problème de l'estimation de la fonction de régression spatiale se présente quand on intéresse à la relation explicative entre deux variables spatiales, dans le but de prédire, éventuellement, l'une en fonction de l'autre. Cette prédiction devient de la *classification supervisée* ou *discrimination* dans le cas particulier où la variable réponse appartient à un ensemble discret de cardinal fini. Nous supposons que la suite de couples de variables spatiales $\{(\mathrm{X}_{\mathbf{i}}, \mathrm{Y}_{\mathbf{i}})_{\mathbf{i} \in \mathscr{I}_{\mathbf{n}}}\}$ observée sur la région rectangulaire ci-dessus, satisfait le modèle non-paramétrique de régression suivant :

$$\mathrm{Y}_{\mathbf{i}} := r(\mathrm{X}_{\mathbf{i}}) + \varepsilon_{\mathbf{i}}; \mathbf{i} \in \mathscr{I}_{\mathbf{n}}, \tag{1.5}$$

où

$$r(.) = \mathbb{E}(\mathrm{Y}_{\mathbf{i}} | \mathrm{X}_{\mathbf{i}} = .). \tag{1.6}$$

La fonction $r(.)$ est supposée indépendante de $\mathbf{i}$, le bruit $\{\varepsilon_{\mathbf{i}}, \mathbf{i} \in \mathscr{I}_{\mathbf{n}}\}$ est centré, α-mélange et indépendant de $\{\mathrm{X}_{\mathbf{i}}, \mathbf{i} \in \mathscr{I}_{\mathbf{n}}\}$.

Plusieurs travaux ont été réalisés sur l'estimation non-paramétrique du modèle de l'équation (1.6). Des résultats ont été établis dans les cas discret et continu, respectivement, avec l'hypothèse principale de la dépendance spatiale sur les processus.

Dans le cas discret, le processus spatial est indexé par $\mathbf{i} \in (\mathbb{N}^*)^{\mathrm{N}}$ et $\mathrm{X}_{\mathbf{i}} \in \mathbb{R}^d$. L'estimateur $r_{\mathbf{n}}$ de r se construit à travers les observations $\{(\mathrm{X}_{\mathbf{i}}, \mathrm{Y}_{\mathbf{i}})_{\mathbf{i} \in \mathscr{I}_{\mathbf{n}}}\}$ soit :

$$r_{\mathbf{n}}(x) = \frac{\frac{1}{\widehat{\mathbf{n}} h_{\mathbf{n}}^{d}} \sum_{\mathbf{i} \in \mathscr{I}_{\mathbf{n}}} \mathrm{Y}_{\mathbf{i}} \mathrm{K}\left(\frac{x - \mathrm{X}_{\mathbf{i}}}{h_{\mathbf{n}}}\right)}{f_{\mathbf{n}}(x)}, \text{ avec } f_{\mathbf{n}}(x) = \frac{1}{\widehat{\mathbf{n}} h_{\mathbf{n}}^{d}} \sum_{\mathbf{i} \in \mathscr{I}_{\mathbf{n}}} \mathrm{K}\left(\frac{x - \mathrm{X}_{\mathbf{i}}}{h_{\mathbf{n}}}\right). \tag{1.7}$$

Les travaux de [80; 235] étudient l'estimateur (1.7) et traitent respectivement le cas d'anisotropie et celui d'isotropie. Ils obtiennent les propriétés asymptotiques de l'estimateur et établissent les convergences faible et forte. Dans [184], les auteurs étudient (1.7) dans le cadre de la linéarité locale et anisotropique. Ils montrent la normalité asymptotique, sous des conditions générales. Le problème de la prédiction d'un processus spatial indexé sur un ensemble discret est abordé par [41]. Cet auteur se base sur l'estimateur (1.7) pour construire le prédicteur spatial. Il établit la convergence uniforme ainsi que la normalité asymptotique. Une méthode de prédiction basée sur le modèle auto-régressif non-paramétrique est proposé par [64]. Une méthode alternative au noyau (1.7), basée sur la pondération par la technique des k plus proches voisins a été étudiée par [227]. Ce dernier construit ainsi un estimateur dont il montre la normalité asymptotique. Le cas d'estimation de la fonction de régression pour des processus spatiaux continûment indexés par $\mathbf{i} \in \mathbb{R}^{\mathrm{N}}$ est abordé dans les travaux de [105]. Ce dernier considère un estimateur de la fonction de régression basé sur un processus spatial multidimensionnel $\{\mathrm{Z}_{\mathbf{i}} = (\mathrm{X}_{\mathbf{i}}, \mathrm{Y}_{\mathbf{i}}), \mathbf{i} \in \mathbb{R}^{\mathrm{N}}\}$. Sous des conditions suffisantes de mélange et des fenêtres optimales, la convergence faible et celle forte sont établies avec, respectivement, des vitesses optimales et sur-optimales. Ensuite, les auteurs mettent en œuvre une première approche de prédiction spatiale pour le processus spatial multidimensionnel continûment indexé. Les approches citées concernent la situation pour laquelle les observations sont collectées suivant un plan d'échantillonnage fixe. Certains auteurs ont abordé l'estimation non-paramétrique basée sur l'échantillonnage stochastique. Dans [257], les auteurs proposent une méthode de régression à noyau basée sur un processus stochastique spatial à design aléatoire. La précision du prédicteur est mesurée par l'erreur quadratique moyenne qui tend à être négligeable si la taille de l'échantillon augmente, sous des hypothèses moins restrictives. La procédure de validation croisée a été appliquée pour la sélection des fenêtres locale et globale. Le cadre asymptotique intensif est considéré dans ce travail c'est à dire la région considérée est bornée et fixe. Dans [133], les auteurs établissent la normalité asymptotique de l'estimateur de [268; 357] dans un contexte où les données spatiales dépendantes sont supposées espacées de manière irrégulière. Les observations sont collectées dans une région de $\mathbb{Z}^d$. L'estimation de la fonction de régression basée sur les noyaux récursifs

est traitée par [50; 51; 81]. La méthode de régression basée sur les quantiles conditionnelles est étudiée dans les travaux de [1; 54; 103; 107; 211; 228; 254; 314; 369; 376]. Nous constatons que les approches citées précédemment ignorent la structure spatiale dans la construction des estimateurs. Pour mieux prendre en compte la dépendance spatiale dans l'estimateur, [98] propose un modèle de régression basé sur deux noyaux. Comme pour la densité, l'estimateur prend en compte, dans sa construction, la structure spatiale, mesurée par la proximité géographique entre les sites, et la distance entre les observations. La méthode est construite comme suit :

$$r_{\mathbf{n}}(x) = \frac{\sum_{\mathbf{i}\in\mathscr{I}_{\mathbf{n}}} \mathrm{Y}_{\mathbf{i}} \mathrm{K}_1\left(\frac{x-\mathrm{X}_{\mathbf{i}}}{h_{\mathbf{n}}}\right) \mathrm{K}_2\left(\rho_{\mathbf{n}}^{-1}\left\|\frac{\mathbf{i}_0-\mathbf{i}}{\mathbf{n}}\right\|\right)}{\sum_{\mathbf{i}\in\mathscr{I}_{\mathbf{n}}} \mathrm{K}_1\left(\frac{x-\mathrm{X}_{\mathbf{i}}}{h_{\mathbf{n}}}\right) \mathrm{K}_2\left(\rho_{\mathbf{n}}^{-1}\left\|\frac{\mathbf{i}_0-\mathbf{i}}{\mathbf{n}}\right\|\right)}. \tag{1.8}$$

Les travaux de [98] établissent la convergence uniforme et la normalité asymptotique de l'estimateur (1.8), sous la condition de stationnarité locale.

2.1.3 Statistique fonctionnelle

Au cours de ces dernières années, la branche de la statistique consacrée à l'analyse des données fonctionnelles a connu un réel essor en termes de développements théoriques et méthodologiques avec des applications dans divers domaines. Les objets fonctionnelles sont des variables aléatoires à valeurs dans un espace, de fonctions par exemple, éventuellement infini. Elles peuvent être considérées comme des courbes [164; 165; 296], des images ou d'autres formes plus complexes. Elles sont parfois supposées être des réalisations d'un processus stochastique unidimensionnel appartenant à un espace Hilbertien. L'apparition de ces types de données remonte aux années 1950 avec les travaux de [176] et ceux de [292; 346], respectivement. À notre connaissance, les travaux de [292; 346] sont les premiers à avoir étudié l'analyse en composantes principales et l'analyse factorielle de ces types de données assez particulières. Plus tard, [287; 290] introduisent la notion de donnée fonctionnelle et amorcent l'adoption des méthodes statistiques multivariées dans le cadre fonctionnel. Ce terme a été repris depuis par les auteurs qui travaillent dans ce domaine. Ces premiers travaux ont donné naissance à des ouvrages de références dans le domaine fonctionnel. [146; 288; 291] ont étudié des méthodes statistiques adaptées aux variables fonctionnelles dans le cadre linéaire et non linéaire. De même, [48] a contribué aux développements des méthodes statistiques permettant l'analyse des variables fonctionnelles dépendantes (processus autorégressif hilbertiens).
L'une des premières questions qui se posent, en statistique fonctionnelle, est comment les observations sont représentées. Soit $\mathscr{E}$ l'espace fonctionnel de dimension éventuellement infinie. L'ensemble $\mathscr{E}$ peut être un espace de fonctions ou celui d'opérateurs linéaires, etc. Pour simplifier, nous posons $\mathscr{E} = \mathrm{L}^2(\mathrm{I})$ avec $\mathrm{I} = [0,..,\mathrm{T}]$ un compact de $\mathbb{R}$. Soit $\mathrm{X}(t), t \in \mathrm{I}$, une variable fonctionnelle appartenant à $\mathscr{E}$ ($\mathrm{X}(t) \in \mathrm{L}^2(\mathrm{I})$ si seulement si $\mathbb{E}(\int_{\mathrm{I}} \mathrm{X}^2(t))dt < \infty$) et $\mathrm{X}_1(t),....,\mathrm{X}_n(t)$ des copies de $\mathrm{X}(t)$.

Représentation d'une variable fonctionnelle

La représentation d'une donnée fonctionnelle nécessite un système de base de fonctions connues $\mathscr{B} = \{\rho_j\}, j \in \mathbb{N}$ de l'espace considéré. La base de fonctions permet en particulier d'approcher chaque objet fonctionnel par une combinaison linéaire de P élémente de $\mathscr{B}$, soit

$$\mathrm{X}(t) = \sum_{j\in\mathbb{N}}^{+\infty} \varepsilon_j \rho_j \approx \sum_{j\in\mathbb{N}}^{\mathrm{P}} \varepsilon_j \rho_j. \tag{1.9}$$

Nous pouvons regrouper les bases de fonctions, couramment utilisées, en deux types :

- Les bases fixes constituées par des fonctions définies sur une grille de nœuds :
  - La base de fonctions trigonométriques appelée base de Fourier utilisée pour représenter les processus qui ont un caractère périodique [289].
  - La base de fonctions polynômes définies sur des sous-intervalles de $\mathbb{R}$ appelée B−spline. Une base flexible et adaptée aux lissages des courbes [69].
  - La base d'ondelette ([14]) utilisée pour la réduction de la dimension des signaux.
- Bases construites à partir des informations apportées par les données.
  - Les Composante principales fonctionnelles (CPF) qui constituent une base de fonctions construites à l'aide des projections séquentielles sur des sous-espaces de $\mathrm{L}^2(\mathrm{I})$ (voir [68]).
  - Les Moindres carrés partiels fonctionnels (MCPF), voir [70].

Outre la difficulté liée à la représentation des données, il se peut que celles-ci soient observées avec un certain bruit ε lié à l'instrument de mesure ou une autre perturbation. Une méthode de lissage appropriée permet d'enlever ces bruits.

Régression Fonctionnelle (FR)

Les variables fonctionnelles étant perçues comme une généralisation (dans une certaine mesure) des données multivariées, les premières méthodes d'analyse fonctionnelle reposent sur l'extension de celles classiques multivariées dont modèle linéaire généralisé (GLM) et modèle additif généralisé (GAM) sont les plus connues. Les auteurs [263] étendent le modèle GLM dans le contexte fonctionnel en combinant des co-variables non fonctionnelles et celles fonctionnelles. Les modèles GAM de [189; 328] ont été étendus, sous différentes approches, dans le cadre fonctionnel. Nous citons les travaux de [6; 72; 77; 136; 138; 141; 143; 147; 171; 217; 230; 264; 371; 379; 380]. Le modèle logistique fonctionnel est étudié par [136]. Dans ces travaux, les auteurs utilisent les bases de MCPF pour la représentation des co-variables fonctionnelles. [70; 203; 263] utilisent les mêmes formes de représentation des variables fonctionnelles avec/et sans pénalisation. [284] tente d'apporter des solutions aux problèmes de la dimension infinie en utilisant des techniques de pénalisation du critère de moindres carrés partiels (MCPF). Les modèles Régression inverse par tranches (RIT) ont été étendus aussi dans le cadre fonctionnel par [6; 77]. Les approches de la régression fonctionnelle abordées dans la littérature dépendent des objectifs que l'on se fixe, du support et de la nature des variables d'entrée et de sortie. La régression est dite fonctionnelle si l'une des variables en relation explicative, au moins, est fonctionnelle.
Nous présentons d'abord le modèle général de régression fonctionnelle qui est le support de la généralisation des autres modèles de régression et de classification.
Soient $Y \in \mathbb{R}$, une variable de réponse, $\mathbf{Z} \in \mathbb{R}^p$, un vecteur de co-variable et $\mathscr{X} = \{X_j(.)\}_{j=1}^{q}$ q co-variables fonctionnelles. Le modèle général peut être représenté comme suit :

$$Y = r(\mathbf{Z}, \mathscr{X}) + \varepsilon. \tag{1.10}$$

La fonction $r = \mathbb{E}(Y/(\mathbf{Z}, \mathscr{X}))$ est supposée être inconnue et ε est l'erreur sur le modèle (1.10).

Régression Fonctionnelle Linéaire (RFL)

Le modèle linéaire classique suppose que r est une fonction linéaire des co-variables.
Dans le cas où une seule co-variable fonctionnelle est observée soit $\mathscr{X} = X(.)$, nous avons :

$$Y = \beta_0 + \langle X, \beta \rangle + \varepsilon = \beta_0 + \int_I X(t)\beta(t)\,dt + \varepsilon. \tag{1.11}$$

Avec β_0 la constante et $\beta(.)$ la fonction des coefficients (paramètres) de régression et ε le bruit aléatoire de variance constante et de moyenne nulle. Ce modèle a été largement étudié par [68; 69; 183] dans le cadre i.i.d. L'idée de base est de chercher la représentation du couple $(X(.), \beta(.))$ à travers une même base de fonctions B-spline, Fourier ou Ondelette (voir [69; 289]). Le couple $(X(.), \beta(.))$ peut aussi être représenté à travers des bases CPF ou celles MCPF (voir [3; 68]).

Dans le cas où on observe $(\mathbf{Z}, \mathscr{X})$ avec $\mathscr{X}$ un vecteur de dimension q, nous avons :

$$Y = \beta_0 + \langle \mathbf{Z}, \theta \rangle + \sum_{j=1}^{p} \int_{I_j} X_j(t)\beta_j(t)\,dt + \varepsilon, \tag{1.12}$$

où θ un vecteur de coefficients de régression de $\mathbf{Z}$. Ce modèle a été étudié, entre autres, par [194].

Modèle additif généralisé fonctionnel (FGAM)

Le modèle présenté dans l'équation (1.12) repose sur l'hypothèse de la linéarité de la relation explicative entre les prédicteurs et la variable réponse. Cela est restrictif, dans certains cas, en particulier pour des processus spatialement dépendants. Le modèle de l'équation (1.13) est plus flexible et prend en compte certains aspects de la non-linéarité éventuelle :

$$Y = \sum_{j=1}^{q} r_j(X_j) + \varepsilon. \tag{1.13}$$

Les fonctions r_j, $j = 1...q$ étant définies dans $\mathscr{E}$.

Modèle linéaire généralisé fonctionnel (FGLM)

Pour prendre en compte des situations diverses suivant la nature de la variable Y, certains auteurs introduisent, dans l'équation (1.13), une fonction de lien notée g (voir [141; 143; 263]). On obtient ainsi le modèle (FGLM) suivant :

$$\mathbb{E}(\mathrm{Y}|\mathrm{X},\mathbf{Z}) = \eta = g^{-1}\left(\beta_0 + \langle \mathbf{Z}, \theta \rangle + \sum_{j=1}^{q} \int_{\mathrm{T}_j} \mathrm{X}_j(t)\beta_j(t)\,dt\right). \tag{1.14}$$

Modèle additif spectral généralisé fonctionnel (MASGF).

Le modèle additif de régression fonctionnelle est défini par :

$$\mathbb{E}(\mathrm{Y}|\mathrm{X},\mathbf{Z}) = \eta = g^{-1}\left(\sum_{k=1}^{p} f_k(\mathrm{Z}^k) + \sum_{j=1}^{q} r_j(\mathrm{X}_j(t))\right) \tag{1.15}$$

Les fonctions $f_k, k = 1...p$ sont définies dans $\mathbb{R}^p$. Les techniques d'estimations des fonctions r_j sont diverses. [264] propose la décomposition spectrale du modèle MASGF, tandis que [143] introduit l'estimation non-paramétrique par noyau correspondant au Modèle additif à noyau généralisé fonctionnel (MANGF).
Nous constatons que ces travaux de recherches portent, généralement, sur les modèles paramétriques ou semi-paramétriques fonctionnels linéaires. Cependant, bien que l'approche linéaire a donné des résultats satisfaisants, il n'en demeure pas moins qu'elle présente des limites; car ne prenant pas en compte certains processus temporellement et/ou spatialement dépendants. Les méthodes non-paramétriques, au cœur d'une dynamique de recherche, constituent une alternative. Elles reposent sur des conditions moins restrictives. Elles s'adaptent dans beaucoup de situations en particulier quand les observations présentent une structure de dépendance non-linéaire. Elles sont connues par leur flexibilité et prennent en compte l'aspect non-linéaire entre le prédicateur et la variable réponse (voir [146]).

Modèle fonctionnel basé sur l'estimation non-paramétrique de la fonction de régression

Supposons que la relation explicative entre le prédicteur fonctionnel et la réponse réelle soit une fonction d'espérance conditionnelle, $\mathbb{E}(\mathrm{Y}|\mathrm{X}) = r(\mathrm{X})$, suivant le modèle $\mathrm{Y} = r(\mathrm{X}) + \varepsilon$; [144; 146] proposent l'estimateur de la fonction de régression fonctionnelle suivante :

$$\widehat{r}(x) = \frac{\sum_{i=1}^{n} \mathrm{Y}_i \mathrm{K}(h^{-1}d(x,\mathrm{X}_i))}{\sum_{i=1}^{n} \mathrm{K}(h^{-1}d(x,\mathrm{X}_i))}, \text{ pour tout } x \in \mathscr{E}. \tag{1.16}$$

K est une fonction noyau et h une fenêtre de lissage, $d(.,.)$ est une métrique, norme ou semi-métrique.

L'estimateur (1.16) a été étudié dans la littérature sous des angles différents. En effet, les premiers résultats sur l'estimation non linéaire de l'opérateur de régression, équation (1.16) ont été établis par [144]. Ils ont été ensuite étendus par [146] en traitant le cas des données dépendantes et en établissant des convergences fortes. [97] a étudié la convergence en norme L^p de l'estimateur de l'opérateur de régression. Ensuite ils ont appliqué leurs résultats aux problèmes de la classification des courbes. [36] ont étudié l'estimateur de la régression en supposant que les erreurs sont corrélées et ont la propriété de longue mémoire. Ils ont également étudié la convergence en probabilité ponctuelle puis celle uniforme de l'estimateur (1.16). Une autre contribution basée sur la construction d'un critère de choix automatique et optimal du paramètre de lissage h a été présentée par [286]. Les travaux de [11; 17; 37; 320] étudient l'estimateur (1.16) sous différentes approches récursives. Cette dernière permet de réduire le temps de calcul dans l'acquisition des données de masses. Dans certaines situations, la présence de valeurs aberrantes peut amener à produire des résultats non pertinents. Dans ce cadre, les méthodes de régression non-paramétrique robuste sont introduites pour résoudre ce problème [9; 196; 241].
La régression fonctionnelle a été étudiée pour des processus spatiaux [21; 196; 246; 275]. Les travaux de [101] font partis des premiers qui ont introduit une méthode non-paramétrique de régression fonctionnelle pour des données spatiales. Ils établissent la convergence faible et celle forte de l'estimateur sous des conditions $\alpha-$mélange. Ils obtiennent ainsi des vitesses de convergence uniforme. Plus tard [330] intègre la structure spatiale dans la construction de l'estimateur de la fonction de régression et incorpore deux noyaux dont l'un mesure la proximité entre les sites tandis que l'autre contrôle la distance entre les observations. Il établit la convergence ponctuelle quand le processus est supposé strictement stationnaire. [4] propose un modèle linéaire fonctionnel dans un cadre spatial qui repose sur un design aléatoire. [169] donne une synthèse sur l'utilisation du cadre fonctionnel dans l'analyse de données spatiales massives (Big-Data). Récemment, [19; 20; 38] introduisent des nouvelles méthodes de modélisation des données fonctionnelles spatiales et

spatio-temporelles dépendantes, basées sur la régression avec régularisation différentielle. Dans ces méthodes, le terme de régularisation permet d'inclure des informations spécifiques au problème sur la variation spatiale ou spatio-temporelle du phénomène étudié, formalisées en termes d'équations aux dérivées partielles (EDP). Ainsi l'estimateur construit dans ces travaux est basé sur les EDP. [20] étudie le biais et la variance de l'estimateur ainsi que l'erreur quadratique moyenne.

Classification supervisée ou Discrimination :

La *classification supervisée* ou le problème de *discrimination* est un outil important utilisé en analyse multivariée et en statistique fonctionnelle. Elle a fait l'objet d'une étude approfondie en dimension finie et particulièrement en série temporelle avec l'hypothèse d'indépendance sur les observations ou celle de la stricte stationnarité sur le processus étudié [42; 118; 120; 157; 172; 188; 190; 274]. En dimension infinie, nous pouvons citer, entre autres, [10; 40; 92; 93; 110; 143; 146; 245; 281; 307; 308; 374]. Cependant, le problème de la *discrimination* reste relativement peu développé en statistique spatiale. Il n'existe presque pas de travaux qui intègrent à la fois les aspects, spatial et fonctionnel, des données liées aux problèmes de *discrimination*. À notre connaissance, les travaux de [373; 374] sont les premiers à avoir traité le cas des données spatiales de nature finie et fonctionnelle respectivement.
De façon générale, la *classification supervisée* vise à attribuer à un individu une étiquette (ou classe) dans l'ensemble des classes pré-définies via une fonction de *discrimination* ou *classeur* g. Supposons que nous travaillons sur la classification binaire $\mathrm{Y} \in \{0,1\}$ et qu'on observe $\mathrm{X} = x$. Nous voulons construire une fonction g qui vise à attribuer 1 ou 0 à Y, sous le "regard" du superviseur x. Posons $\eta(x) = \mathbb{P}\{\mathrm{Y} = 1|\mathrm{X} = x\}$, la fonction η est appelée probabilité postérieure. La performance du *classeur* g est mesurée par la probabilité, de l'erreur d'attribution de classe, définie comme suit :

$$\mathrm{L}(g) = \mathbb{P}\{g(x) \neq \mathrm{Y}\}$$

On se base sur les observations $\mathrm{D}_n = \{(\mathrm{X}_i, \mathrm{Y}_i), i = 1...n\}$, un échantillon i.i.d de (X, Y), pour la construction d'une *règle de classification* $g_{\mathbf{n}}$ qui est un estimateur de g. Ainsi, l'attribution de classe à Y se fait à travers $g_{\mathbf{n}}(x) = g_{\mathbf{n}}\left((\mathrm{X}_i, \mathrm{Y}_i)_{i\in\{1,\ldots,n\}}\right)$. La performance de $g_{\mathbf{n}}$ est mesurée par :

$$\mathrm{L}_n(g) = \mathbb{P}\{g(x) \neq \mathrm{Y}|\mathrm{D}_n\}.$$

L'objectif de la théorie (et de la pratique) de la *classification supervisée* est de construire des estimateurs $g_{\mathbf{n}}$ du *classeur* g dont la probabilité d'erreur est aussi minimale que possible. Plusieurs algorithmes ont été proposés dans la littérature ; voir les ouvrages de [49; 119], pour une introduction approfondie à la théorie de la *classification supervisée*. Les approches de cette dernière peuvent être regroupées en trois grandes familles.

- Les entropies pures et la minimisation du risque empirique [33; 49; 242].
- L'interprétation géométrique appelée en anglais Support Vector Machine (SVM) [45; 325; 348].
- La règle de Bayes [22; 43; 182].

Notons que la *règle de classification* basée sur la procédure des k voisins les plus proches est générale et elle englobe les deux dernières approches. Elle a fait l'objet de plusieurs travaux, voir [2; 27; 61; 88; 116; 117; 128; 149; 155; 173; 174; 175; 177; 178; 179; 181; 190; 208; 232; 323; 327; 353; 367; 377].
Notons, cependant, que l'approche qui consiste à maximiser la probabilité postérieure a aussi fait l'objet de beaucoup de théories et d'applications (voir [146]). Elle est basée sur l'estimation des probabilités postérieures $\eta(x)$. Elle se présente comme suit.
D'une manière générale, considérons D_n un échantillon d'observation, où $\mathrm{Y}_i = m \in \{1, ..., \mathrm{M}\}$ représente une classe associée à une observation X_i.
L'objectif est d'affecter une classe $m \in \{1, ..., \mathrm{M}\}$ à Y, sous la supervision $\mathrm{X} = x$.
Les probabilités postérieures d'appartenance dans chaque groupe sont :

$$\mathrm{P}_m(\mathrm{X}) = \eta(x) = \mathbb{P}(\mathrm{Y} = m|x) = \mathbb{E}\left(\mathbb{1}_{[\mathrm{Y}=m]}|x\right). \tag{1.17}$$

La *règle de discrimination* consiste de construire les estimateurs de $\mathrm{P}_m(\mathrm{X})$, $m \in \{1, ..., \mathrm{M}\}$. Ainsi, la prédiction de classe est donnée par :

$$\widehat{\mathrm{Y}} = \arg\max_{m\in\{1,\ldots,\mathrm{M}\}} \widehat{\mathrm{P}}_m(x). \tag{1.18}$$

Étant donné que le problème de *discrimination* en dimension infinie peut être perçu comme la généralisation de celui en dimension finie ($\mathrm{X} \in \mathbb{R}^d$), nous pouvons nous limiter à présenter, dans ce qui suit, quelques méthodes classiques de *discrimination fonctionnelle* basées sur la maximisation des probabilités postérieures. Ces dernières, reposent, en partie, sur des estimateurs de régression fonctionnelle [146; 156; 203; 262; 375]. Elles sont des cas particuliers de la prédiction.

Régression logistique fonctionnelle :

Le modèle de classification basé sur la régression logistique fonctionnelle est donnée par :

$$\mathrm{P}(\mathrm{Y}=m|\mathrm{X}=x)=\log\left(\frac{\mathrm{P}r(\mathrm{Y}=m|\mathrm{X}=x)}{1-\mathrm{P}r(\mathrm{Y}=m|\mathrm{X}=x)}\right)=\beta_0+\int_{\mathrm{I}}\mathrm{X}(t)\beta(t)dt,\ m\in\{1,...,\mathrm{M}\}, \tag{1.19}$$

où β_0 est une constante et β la fonction de paramètres du prédicteur fonctionnelle X. Nous attribuons à x la classe m si $\mathrm{P}(m|\mathrm{X}=x)>\mathrm{P}(j|x)$, $j\in\{1,...,\mathrm{M}\}$, $j\neq m$.
L'équation (1.19) est une extension du modèle de régression multivarié que l'on peut retrouver dans les récents travaux de [253]). [223] étudie un modèle linéaire généralisé fonctionnel basé sur l'approche d'analyse en composantes principales fonctionnelle (CPF). Il coïncide avec le modèle fonctionnel logistique quand la fonction de lien est le logarithme. Différentes variantes de ce modèle ont été étudiées dans la littérature (voir [15; 252; 355]).

Les Modèles FGAM de discrimination :

$$\mathrm{P}(\mathrm{Y}=m|\mathrm{X}=x)=\log\left(\frac{\mathrm{P}r(\mathrm{Y}=m|\mathrm{X}=x)}{1-\mathrm{P}r(\mathrm{Y}=m|\mathrm{X}=x)}\right)=\beta_0+f_0(x),\quad m\in\{1,...,\mathrm{M}\}, \tag{1.20}$$

où la méthode d'estimation de la fonction f_0 détermine le modèle FGAM utilisé.

La méthode non-paramétrique de classification supervisée :

Cette méthode a été étudiée par [10; 145; 146], soit :

$$\mathrm{P}(\mathrm{Y}=m|\mathrm{X}=x)=\frac{\sum_{i=1}^{n}\mathbb{1}_{[\mathrm{Y}_i=m]}\mathrm{K}\left(\frac{d(x,\mathrm{X}_i)}{h}\right)}{\sum_{i=1}^{n}\mathrm{K}\left(\frac{d(x,\mathrm{X}_i)}{h}\right)},\quad m\in\{1,...,\mathrm{M}\}. \tag{1.21}$$

La méthode k-plus proches voisins fonctionnelle (kPPF) :

La méthode de *classification supervisée* kPPF est basée sur les densités de groupe. Elle est déterminée par le nombre k des plus proches voisins de x au sens d'une semi-métrique $d(.,.)$ donnée et l'ordre de dérivation a, soit $\mathscr{N}_l^{(a)}(x)$ l'ensemble des k plus proches voisins de x, nous avons :

$$\mathrm{P}(\mathrm{Y}=m|\mathrm{X}=x)=\frac{1}{k}\sum_{\left\{j:\mathrm{X}_j^{(a)}\in\mathscr{N}_l^{(a)}(x)\right\}}\mathbb{1}_{[\mathrm{Y}_j=m]}, \tag{1.22}$$

voir les travaux de [152; 153] pour plus de détails.

La méthode DD^{G} : La méthode DD^{G} basée sur l'approche par profondeur a été étudiée par [92], se conférer à ce dernier pour plus de détail.

2.2 Contributions :

Dans cette thèse, nous nous intéressons principalement à l'estimation non-paramétrique de la fonction de régression en nous appuyant sur des données géostatistiques échantillonnées de manière déterministe. Dans le Chapitre 3, nous abordons le cas de la régression spatiale basée sur la méthode des k-plus proches voisins (la co-variable X est de dimension finie). Dans le Chapitre 4, nous étudions la régression spatiale fonctionnelle (X est de dimension infinie). Les méthodes de régressions proposées dans ces Chapitres ont été étendues au cadre de la prédiction et de *discrimination.* Le Chapitre 5 est consacré à l'application des méthodes de prédiction et *classification supervisée* pour analyser la distribution spatiale la ressource démersale côtière du Sénégal.

2.2.1 Contribution à la régression et à la prédiction spatiale

La méthode de régression basée sur les k-plus proches voisins du Chapitre 3 généralise, dans le cadre spatial, les méthodes classiques des k-plus proches voisins [87]. Elle est une alternative à la méthode de régression proposée par [98] qui s'appuie sur deux noyaux dont l'un contrôle la proximité entre les sites

et l'autre contrôle la distance entre les observations. L'approche est particulièrement innovante et adaptée surtout quand le support des données présente une structure spatiale hétérogène. En effet, elle utilise une fenêtre de lissage aléatoire adaptée à une éventuelle hétérogénéité au niveau des réalisations de la variable explicative spatiale observée.
Un des objectifs dans le Chapitre 3 est de construire un prédicteur qui s'adapte quand le processus est local stationnaire et que la structure spatiale présente une certaine hétérogénéité.
La *classification supervisée* est étudiée dans le Chapitre 3. Cette dernière est un cas particulier de la prédiction et correspond au cas où le processus spatial $\{Y_{\mathbf{i}}, \mathbf{i} \in \mathbb{N}^N\}$ prend ses valeurs dans un ensemble discret.
La méthode de régression du Chapitre 4, généralise , dans le cadre fonctionnel, celle proposée par [98]. L'extension se fait, entre autres, sous l'hypothèse de stationnarité locale et nous établissons la convergence ponctuelle et celle uniforme. [330] propose la même méthode mais elle établit seulement la convergence ponctuelle sous la condition de stricte stationnarité. Un prédicteur spatial et fonctionnel est construit à partir de la régression du Chapitre 4. Ce prédicteur devient une nouvelle méthode de *classification supervisée* dans le cas où le processus spatial $\{Y_{\mathbf{i}}, \mathbf{i} \in \mathbb{N}^N\}$ prend ses valeurs dans un ensemble discret.

2.2.2 Contributions en recherche halieutique

Les approches développées dans cette thèse sont innovantes et très adaptées dans beaucoup de situations surtout quand le support d'étude produit des données spatialement hétérogènes. En particulier, lorsque nous modélisons des phénomènes éventuellement continus dans le temps et/ou dans l'espace, des méthodes appropriées permettant de capturer le maximum d'informations s'avèrent nécessaires. Des exemples étudiés dans ce travail sont les phénomènes mesurés par les données d'océanographie halieutiques issues des campagnes scientifiques du CRODT qui ciblent la ressource démersale côtière. Ces données sont décrites, respectivement, dans les Chapitres 1 et 5. Nous avons évoqué dans le Chapitre 1 l'interaction entre la ressource côtière et son milieu marin. Les propriétés physiques des fonds et les conditions environnementales de l'écosystème marin gouvernent le comportement des espèces démersales côtières. Elles sont caractérisées par l'hétérogénéité dans le temps et dans l'espace. Elles agissent sur la distribution spatiale de ces poissons. Ces derniers sont à la recherche d'habitat favorable à leur survie. Cela fait que la ressource halieutique n'est pas uniformément distribuée dans son aire géographique. Des migrations saisonnières vers des zones propices sont notées durant la durée de vie des espèces. Elles concernent des déplacements verticaux et horizontaux suivant les différentes phases de maturation des poissons. Les méthodes paramétriques classiques telles que le Krigeage et le co-krigeage, (voir [91; 299]), sont habituellement appliquées pour l'évaluation de stock et la prédiction d'abondance ou de biomasses des poissons. D'autres méthodes multivariées telles que la fonction K de Ripley [162; 191; 222; 297] et les méthodes Species Distribution Modeling (SDM)/Joint Species Distribution Modeling (JSDM) basées sur les modèles GLM et GAM [236; 282; 372] ont également été utilisées en biologie marine. Cependant, ces méthodes classiques reposent souvent sur des hypothèses assez restrictives comme la distribution gaussienne et de covariance paramétrique. Il faut noter également que lorsque l'échantillon d'intérêt est un ensemble de données volumineuses, les techniques classiques de réduction des dimensions sont des approches courantes. En général, pour résoudre le problème de dimension, plusieurs méthodes de régressions multivariées utilisant un grand nombre de prédicteurs considèrent la dimension comme un paramètre de nuisance. En outre, ces méthodes ne capturent pas les informations supplémentaires provenant du processus qui génère les données.
La modélisation non-paramétrique spatiale et fonctionnelle peut constituer une alternative, aux modèles mathématiques multivariés, pour le traitement et l'analyse des données spatiales massives de grande dimension. Cette modélisation fait intervenir un domaine de recherche récent combinant les branches bien développées de la statistique fonctionnelle et celle spatiale montrant la capacité d'analyser les données complexes. Ainsi, des approches de prédiction et de classification non-paramétriques spatiales et fonctionnelles sont, parallèlement, étudiées (voir les Chapitres 3 et 4). Ces approches sont appliquées dans le Chapitre 5 sur des données d'océanographies halieutiques du Sénégal. Ainsi nous utilisons les procédures de classification des Chapitres 3 et 4 pour prédire la distribution spatiale des poissons démersaux côtiers. Les méthodes de prédiction permettent d'évaluer les quantités de biomasses/abondances sur des sites où les poissons sont présents et les conditions environnementales connues.

CHAPITRE 3

PRÉDICTION ET CLASSIFICATION POUR DONNÉES SPATIALES RÉELLES PAR LA MÉTHODE DES K-PLUS PROCHES VOISINS

Résumé

Dans ce Chapitre, nous proposons une méthode de régression non-paramétrique, d'un processus spatial réel, basée sur l'approche des k-plus proches voisins. La spécificité de l'estimateur proposé est d'incorporer la structure spatiale par la mesure de la proximité entre les sites d'observations et la prise en compte d'une certaine hétérogénéité éventuelle. Le modèle de régression proposé est le support de base d'un prédicteur spatial et une *règle de classification supervisée*, en particulier. Les résultats de convergence presque complète ou presque sûre de l'estimateur et du prédicteur sont obtenus. Les résultats numériques sur des données simulées illustrent la performance de la méthodologie proposée.

3.1 Introduction

La variété des domaines dans lesquels les données spatiales/spatio-temporelles apparaissent, naturellement, montre l'importance de la statistique spatiale. Ces types de données sont retrouvés dans les sciences de l'environnement, celles du sol, la géophysique, l'océanographie, l'économétrie, l'épidémiologie, la foresterie, le traitement d'images et bien d'autres domaines dans lesquels les données d'intérêt sont collectées dans l'espace. La diversité de ces champs d'applications de la modélisation spatiale fait intervenir différents processus spatiaux. Ces derniers ne vérifient pas, tous, certaines hypothèses nécessaires pour l'élaboration des modèles mathématiques multivariés paramétriques comme la régression, la prédiction ou la *discrimination*. Cela fait soulever des problèmes complexes, en statistique spatiale ou spatio-temporelle, dont certains ne sont pas clairement définis encore moins complètement résolus. La résolution de cette complexité pose des défis qui constituent la base des recherches actuelles en mathématique spatiale. Parmi les hypothèses pratiques et classiques, qui influencent les techniques disponibles utilisées dans ces modèles multivariés spatiaux, il y a celles qui supposent l'indépendance des observations, la linéarité et la normalité. La littérature sur la modélisation des données spatiales/spatio-temporelles est abondante (voir par exemple la monographie de [90]). Elle repose, en grande partie, sur ces modèles linéaires paramétriques.

Les variables spatiales se caractérisent, principalement, par la dépendance spatiale. En outre, certains processus spatiaux ne vérifient pas l'hypothèse de normalité et celle de la linéarité. Les modèles linéaires appliqués aux données spatiales capturent uniquement les relations linéaires globales entre observations. Rappelons que, dans de nombreuses situations, la dépendance spatiale n'est pas linéaire. C'est par exemple, le cas classique où l'on traite la modélisation spatiale des événements extrêmes ou l'étude du comportement de la ressource halieutique face aux effets du changement climatique, etc. Un modèle qui prend en compte les aspects qui spécifient et caractérisent les différents types de données spatiales générées par divers champs d'applications est nécessaire (voir dans le Chapitre 2, les différents types de données spatiales). Dans certains champs d'applications de la modélisation spatiale, se pose la situation selon laquelle il est important d'étudier la relation explicative entre deux variables dans le but de prédire, principalement, l'une d'elles à des endroits où les observations ne sont pas disponibles. Par exemple, en biologie marine, il est souvent utile de prendre en compte l'influence des paramètres environnementaux ou écologiques sur la

variabilité de la biomasse des poissons et leur distribution spatio-temporelle. La réponse à tous ces problématiques complexes fait appel aux modélisations non-paramétriques comme méthodes alternatives aux celles paramétriques et linéaires ; quand les approches classiques ne donnent pas de résultats satisfaisants. La littérature sur les techniques d'estimation non-paramétriques, qui intègrent une dépendance spatiale non linéaire, n'est pas très abondante par rapport à celle qui traite la dépendance linéaire. Pour un aperçu des résultats et des applications tenant compte des données spatialement dépendantes pour la densité, l'estimation de régression, la prédiction et la classification, nous nous référons aux travaux suivants [1; 41; 64; 84; 100; 105; 135; 184; 186; 235; 257; 258; 330; 339; 351].
Parmi les méthodes non-paramétriques, nous mettons l'accent, dans ce Chapitre, sur la méthode des k-voisins les plus proches (k-NN). L'estimateur á noyau k-NN (voir [42]) a un avantage significatif sur l'estimation á noyau proposé par [98]. Sa spécificité réside dans le fait qu'il est flexible à une hétérogénéité éventuellement observée sur les variables. Cela lui permet de prendre en compte la structure locale des données. Il utilise dans le choix d'un nombre approprié de voisins, un paramètre aléatoire adapté à la structure de dépendance spatiale. Un autre avantage de la méthode k-NN est la mise en œuvre facile des paramètres de lissage. En effet, dans la méthode á noyau proposé par [98], le paramètre de lissage est un réel fixé, alors que dans la méthode k-NN, les paramètres de lissage appartiennent à un ensemble discret.
L'utilisation de la méthode k-NN est récente pour les données spatiales. [227] a proposé un estimateur de régression des données spatiales basé sur la méthode k-NN. Il a établi les résultats asymptotiques d'un estimateur k-NN appliqué à des données multivariées.
L'objectif, dans ce Chapitre, est de développer des outils, de prédiction et de *classification supervisée*, appliqués à un processus spatial réels. Ils sont basés sur l'estimation non-paramétrique de la régression spatiale k-NN. La dépendance spatiale non linéaire entre les sites d'observations est mesurée par le critère des coefficients de mélange fort [341]. La construction de l'estimateur de régression repose sur l'utilisation de deux noyaux, l'un contrôle la distance entre les observations à l'aide d'une fenêtre aléatoire et l'autre contrôle la structure de dépendance spatiale. Cette idée a été présentée dans les travaux de [98; 257; 330].
Le reste du Chapitre 3 est organisé comme suit. Dans la section 3.2, nous introduisons le modèle de régression qui est le support de base du prédicteur. La section 3.3 est dédiée à la convergence presque complète [1] du prédicteur alors que la section 3.4 applique le modèle de régression à une *règle de classification supervisée* et adapte les résultats asymptotiques du prédicteur. La section 3.5 donne une application à des données simulées pour mettre en évidence les performances de la méthode proposée. La section 3.6 est consacrée à la conclusion. Enfin, les preuves des principaux résultats asymptotiques sont reportés à l'annexe A.

3.2 Modèle et construction du prédicteur

Soit $\{Z_{\mathbf{i}} = (X_{\mathbf{i}}, Y_{\mathbf{i}}) \in \mathbb{R}^d \times \mathbb{R}\,,\ \mathbf{i} \in \mathbb{N}^N, d \geq 1\}$, un processus spatial défini sur un espace probabilisé $(\Omega, \mathscr{A}, \mathbb{P})$, $N \in \mathbb{N}^*$. Nous supposons que ce processus est observé sur l'ensemble discret $\mathscr{I}_{\mathbf{n}} = \{\mathbf{i} = (i_1, \ldots, i_N), 1 \leq i_k \leq n_k,\ k = 1, \ldots, N\}$, $\mathbf{n} = (n_1, \ldots, n_N) \in \mathbb{N}^N$, et $\widehat{\mathbf{n}} = n_1 \times \ldots \times n_N$, on a $\mathbf{n} \to \infty$ si $\min\{n_k\} \to +\infty$, pour une certaine $C > 0$, $n_k / n_i \leq C,\ \forall\ 1 \leq k, i \leq N$. Nous notons par $\|.\|$ la norme euclidienne définie dans $\mathbb{R}^N$ ou dans $\mathbb{R}^d$ et $\mathbb{I}(\cdot)$ désigne la fonction indicatrice. Nous supposons que la régression de $\{Y_{\mathbf{i}}, \mathbf{i} \in \mathbb{N}^N\}$ sur $\{X_{\mathbf{i}}, \mathbf{i} \in \mathbb{N}^N\}$ est définie par le modèle suivant :

$$Y_{\mathbf{i}} = r(X_{\mathbf{i}}) + \varepsilon_{\mathbf{i}},\ \mathbf{i} \in \mathbb{N}^N, \tag{2.1}$$

où

$$r(\cdot) = E\,(Y_{\mathbf{i}} | X_{\mathbf{i}} = \cdot)\,, \tag{2.2}$$

r est supposé être indépendant de $\mathbf{i}$, le bruit $\{\varepsilon_{\mathbf{i}}, \mathbf{i} \in \mathbb{N}^N\}$ est un processus centré vérifiant le critère de dépendance α-mélange (voir la section 3.3 pour une description de cette condition). Il est indépendant de $\{X_{\mathbf{i}}, \mathbf{i} \in \mathbb{N}^N\}$. Nous nous intéressons à la prédiction du processus spatial $\{Y_{\mathbf{i}}, \mathbf{i} \in \mathbb{N}^N\}$ dans des sites où les observations de ce processus ne sont pas disponibles, soit en particulier $\mathbf{i}_0 \in \mathscr{I}_{\mathbf{n}}$; en se basant sur les informations $X_{\mathbf{i}_0}$ et les observations $\{(X_{\mathbf{i}}, Y_{\mathbf{i}})_{\mathbf{i} \in \mathscr{O}_{\mathbf{n}}}\}$. Supposons que $(X_{\mathbf{i}_0}, Y_{\mathbf{i}_0})$ est de même loi que le couple (X, Y).
L'espace $\mathscr{O}_{\mathbf{n}} \subset \mathscr{I}_{\mathbf{n}}$ est l'ensemble spatial sur lequel le processus $\{(X_{\mathbf{i}}, Y_{\mathbf{i}})_{\mathbf{i} \in \mathscr{O}_{\mathbf{n}}}\}$ est observé, avec $\mathbf{i}_0 \notin \mathscr{O}_{\mathbf{n}}$ et $\mathrm{Card}(\mathscr{O}_{\mathbf{n}})$ tend vers ∞ quand $\mathbf{n} \to \infty$.
Nous supposons qu'un nombre suffisant d'observations $\{(X_{\mathbf{i}}, Y_{\mathbf{i}})_{\mathbf{i} \in \mathscr{O}_{\mathbf{n}}}\}$ a la même distribution de probabilité que celle du couple (X, Y). Nous supposons également que $\{Y_{\mathbf{i}}, \mathbf{i} \in \mathbb{N}^N\}$ est intégrable, (X, Y) et que $\{(X_{\mathbf{i}}, Y_{\mathbf{i}})_{\mathbf{i} \in \mathscr{O}_{\mathbf{n}}}\}$ admettent des fonctions densités inconnues par rapport à la mesure de Lebesgue. Soient f et

1. Soit $(z_n)_{n \in \mathbb{N}}$ une séquence de variables aléatoires à valeur réelle. z_n converge presque complètement ($p.c.$) vers zéro si, et seulement si, $\forall \varepsilon > 0$, $\sum_{n=1}^{\infty} P(|z_n| > \varepsilon) < \infty$. De plus, nous disons que la vitesse de convergence presque complète de z_n vers zéro est d'ordre u_n (avec $u_n \to 0$) et nous écrivons $z_n = O_{p.c.}(u_n)$ si, et seulement si, $\exists \varepsilon > 0$ est telle que $\sum_{n=1}^{\infty} P(|z_n| > \varepsilon u_n) < \infty$. Ce type de convergence implique à la fois la convergence presque sûre et la convergence en probabilité.

$f_{\mathrm{X,Y}}$, les fonctions densités, respectives, de X et (X,Y). La méthode de prédiction s'appuie sur l'estimateur de la fonction de régression k-NN suivant :

$$r_{\mathrm{kNN}}(x) = \begin{cases} \dfrac{g_{\mathbf{n}}(x)}{f_{\mathbf{n}}(x)}, \text{si } f_{\mathbf{n}}(x) \neq 0 \\ \\ \dfrac{1}{\widehat{\mathbf{n}}} \displaystyle\sum_{\mathbf{i} \in \mathscr{I}_{\mathbf{n}} \mathbf{i} \neq \mathbf{i}_0} \mathrm{Y}_{\mathbf{i}}, \text{sinon,} \end{cases} \tag{2.3}$$

avec

$$g_{\mathbf{n}}(x) = \frac{1}{\widehat{\mathbf{n}} h_{\mathbf{n},\mathbf{i}_0}^{\mathrm{N}} \mathrm{H}_{\mathbf{n},x}^{d}} \sum_{\mathbf{i} \in \mathscr{I}_{\mathbf{n}}, \mathbf{i} \neq \mathbf{i}_0} \mathrm{K}_1\left(\frac{x - \mathrm{X}_{\mathbf{i}}}{\mathrm{H}_{\mathbf{n},x}}\right) \mathrm{K}_2\left(h_{\mathbf{n},\mathbf{i}_0}^{-1} \left\| \frac{\mathbf{i}_0 - \mathbf{i}}{\mathbf{n}} \right\|\right) \mathrm{Y}_{\mathbf{i}}.$$

$$f_{\mathbf{n}}(x) = \frac{1}{\widehat{\mathbf{n}} h_{\mathbf{n},\mathbf{s}_0}^{\mathrm{N}} \mathrm{H}_{\mathbf{n},x}^{d}} \sum_{\mathbf{i} \in \mathscr{I}_{\mathbf{n}}, \mathbf{i} \neq \mathbf{i}_0} \mathrm{K}_1\left(\frac{x - \mathrm{X}_{\mathbf{i}}}{\mathrm{H}_{\mathbf{n},x}}\right) \mathrm{K}_2\left(h_{\mathbf{n},\mathbf{i}_0}^{-1} \left\| \frac{\mathbf{i}_0 - \mathbf{i}}{\mathbf{n}} \right\|\right).$$

Le prédicteur de $\mathrm{Y}_{\mathbf{i}_0}$ est construit comme suit :

$$\widehat{\mathrm{Y}}_{\mathbf{i}_0} = \frac{\sum_{\mathbf{i} \in \mathscr{O}_{\mathbf{n}}} \mathrm{Y}_{\mathbf{i}} \mathrm{K}_1\left(\frac{\mathrm{X}_{\mathbf{i}_0} - \mathrm{X}_{\mathbf{i}}}{\mathrm{H}_{\mathbf{n},\mathrm{X}_{\mathbf{i}_0}}}\right) \mathrm{K}_2\left(h_{\mathbf{n},\mathbf{i}_0}^{-1} \left\| \frac{\mathbf{i}_0 - \mathbf{i}}{\mathbf{n}} \right\|\right)}{\sum_{\mathbf{i} \in \mathscr{O}_{\mathbf{n}}} \mathrm{K}_1\left(\frac{\mathrm{X}_{\mathbf{i}_0} - \mathrm{X}_{\mathbf{i}}}{\mathrm{H}_{\mathbf{n},\mathrm{X}_{\mathbf{i}_0}}}\right) \mathrm{K}_2\left(h_{\mathbf{n},\mathbf{i}_0}^{-1} \left\| \frac{\mathbf{i}_0 - \mathbf{i}}{\mathbf{n}} \right\|\right)}, \tag{2.4}$$

si le dénominateur n'est pas nul, sinon le prédicteur est égal à la moyenne empirique. Ici, K_1 et K_2 sont deux noyaux de $\mathbb{R}^d$ et $\mathbb{R}$ à valeur dans $\mathbb{R}_+$ respectivement, $\dfrac{\mathbf{i}}{\mathbf{n}} = \left(\dfrac{i_1}{n_1}, \cdots, \dfrac{i_{\mathrm{N}}}{n_{\mathrm{N}}}\right)$,

$$h_{\mathbf{n},\mathbf{i}_0} = \min\left\{ h \in \mathbb{R}_+^* : \sum_{\mathbf{i} \in \mathscr{O}_{\mathbf{n}}} \mathbb{I}_{\left\{ \left\| \frac{\mathbf{i} - \mathbf{i}_0}{\mathbf{n}} \right\| < h \right\}} = k_{\mathbf{n}}^{'} \right\}$$

et

$$\mathrm{H}_{\mathbf{n},\mathrm{X}_{\mathbf{i}_0}} = \min\left\{ h \in \mathbb{R}_+^* : \sum_{\mathbf{i} \in \mathscr{V}_{\mathbf{i}_0}} \mathbb{I}_{\{\|\mathrm{X}_{\mathbf{i}} - \mathrm{X}_{\mathbf{i}_0}\| < h\}} = k_{\mathbf{n}} \right\},$$

où $k_{\mathbf{n}}^{'}$ et $k_{\mathbf{n}}$ sont des suites de nombres entiers positifs et $\mathscr{V}_{\mathbf{i}_0} = \{\mathbf{i} \in \mathscr{O}_{\mathbf{n}},\ \| \frac{\mathbf{i} - \mathbf{i}_0}{\mathbf{n}} \| < h_{\mathbf{n},\mathbf{i}_0}\}$.
Nous rappelons qu'une des spécificités dans cette approche réside sur le fait que la fenêtre de lissage $\mathrm{H}_{\mathbf{n},\mathrm{X}_{\mathbf{i}_0}}$ est une variable aléatoire positive; elle dépend de $\mathrm{X}_{\mathbf{i}_0}$ et des observations $\{\mathrm{X}_{\mathbf{i}}, \mathbf{i} \in \mathscr{O}_{\mathbf{n}}\}$. Cette particularité, entre autres, le rend plus adapté quand l'hétérogénéité se présente (voir [59]). De ce fait, il est plus facile de le mettre en œuvre et plus avantageux que la méthode proposée par [98] ; qui est basée sur une fenêtre fixe $h_{\mathbf{n}}$ (voir équation (2.5)).

$$\widehat{\mathrm{Y}}_{\mathbf{i}_0}^{\mathrm{NW}} = \frac{\sum_{\mathbf{i} \in \mathscr{O}_{\mathbf{n}}} \mathrm{Y}_{\mathbf{i}} \mathrm{K}_1\left(\frac{\mathrm{X}_{\mathbf{i}_0} - \mathrm{X}_{\mathbf{i}}}{h_{\mathbf{n}}}\right) \mathrm{K}_2\left(\rho_{\mathbf{n}}^{-1} \left\| \frac{\mathbf{i}_0 - \mathbf{i}}{\mathbf{n}} \right\|\right)}{\sum_{\mathbf{i} \in \mathscr{O}_{\mathbf{n}}} \mathrm{K}_1\left(\frac{\mathrm{X}_{\mathbf{i}_0} - \mathrm{X}_{\mathbf{i}}}{h_{\mathbf{n}}}\right) \mathrm{K}_2\left(\rho_{\mathbf{n}}^{-1} \left\| \frac{\mathbf{i}_0 - \mathbf{i}}{\mathbf{n}} \right\|\right)}, \tag{2.5}$$

où $h_{\mathbf{n}}$, $\rho_{\mathbf{n}}$ ne sont pas aléatoires.

3.3 Principaux résultats

Nous supposons que le processus $\{\mathrm{Z}_{\mathbf{i}} = (\mathrm{X}_{\mathbf{i}}, \mathrm{Y}_{\mathbf{i}}) \in \mathbb{R}^d \times \mathbb{R}\,,\ \mathbf{i} \in \mathbb{N}^{\mathrm{N}}\}$ satisfait la condition de mélange définie comme suit : il existe une fonction $\varphi(t) \searrow 0$ lorsque $t \to \infty$, telle que

$$\begin{aligned} \alpha\left(\sigma(\mathrm{S}), \sigma\left(\mathrm{S}'\right)\right) &= \sup\left\{ |\mathbb{P}(\mathrm{A} \cap \mathrm{B}) - \mathbb{P}(\mathrm{A})\,\mathbb{P}(\mathrm{B})|,\ \mathrm{A} \in \sigma(\mathrm{S}),\ \mathrm{B} \in \sigma\left(\mathrm{S}'\right) \right\} \\ &\leq \psi\left(\mathrm{Card}(\mathrm{S}), \mathrm{Card}(\mathrm{S}')\right) \varphi\left(\mathrm{dist}(\mathrm{S}, \mathrm{S}')\right), \end{aligned} \tag{3.1}$$

où S et S′ sont deux ensembles finis de sites, Card(S) dénote le cardinal de l'ensemble S, $\sigma(\mathrm{S})$ (resp.$\sigma\left(\mathrm{S}'\right)$) est un σ-algèbre généré par le $\mathrm{Z}_{\mathbf{i}} = \{\mathrm{Z}_{\mathbf{i}}, \mathbf{i} \in \mathrm{S}\}$ (resp.$\mathrm{Z}_{\mathbf{i}} = \left\{\mathrm{Z}_{\mathbf{i}}, \mathbf{i} \in \mathrm{S}'\right\}$), dist(S,S′) est la distance euclidienne entre S et S′, et $\psi(\cdot)$ est une fonction symétrique positive non décroissante.
Nous rappelons que le processus est fortement mélangeant si $\psi \equiv 1$ (voir [129]). Comme d'habitude, nous supposerons que $\varphi(t)$ vérifie :

$$\varphi(t) \leq \mathrm{C} t^{-\theta}, \qquad \theta > 0\,,\ t \in \mathbb{R}_+^*, \mathrm{C} > 0, \text{ a constant,} \tag{3.2}$$

i.e. $\varphi(t)$ tend vers zéro avec une vitesse polynomiale. Les résultats asymptotiques, établis dans ce qui suit, ne concernent que le cas polynomial. Des résultats similaires peuvent être obtenus, facilement, pour le cas exponentiel ($\varphi(t)$ tend vers zéro avec une vitesse exponentielle voir par exemple [129] pour plus de détails.). Avant de donner les principaux résultats, nous donnons les d'hypothèses sur lesquelles nous nous appuyons pour les établir. Tout au long de ce Chapitre, nous fixons un sous-ensemble compact D dans $\mathbb{R}^d$. Lorsqu'aucune confusion n'est possible, nous désignerons par C, une constante générique strictement positive.

3.3.1 Les principales hypothéses

(**H1**) f et $r(\cdot)$ sont des fonctions lipschitziennes définies sur D. En plus, $\inf_{x\in D} f(x) > 0$.

(**H2**) $k_{\mathbf{n}}^{'} \sim \widehat{\mathbf{n}}^{\gamma}$ et $k_{\mathbf{n}} \sim \widehat{\mathbf{n}}^{\tilde{\gamma}}$, où $\gamma, \tilde{\gamma} \in]0.5, 1[$ et $\tilde{\gamma} < \gamma$.

(**H3**) Le noyau K_1 est borné, de support compact et

$$\forall u \in \mathbb{R}^d,\ K_1(u) \leq K_1(tu), \quad \forall\, t \in]0,1[. \tag{3.3}$$

(**H4**) K_2 est une fonction non négative bornée. Et il existe des constantes C_1, C_2 et ρ telles que

$$C_1 \mathbb{I}_{\{\|t\|\leq\rho\}} \leq K_2(\|\ t\ \|) \leq C_2 \mathbb{I}_{\{\|t\|\leq\rho\}}, \qquad \forall\ t \in \mathbb{R}^N,\ 0 < C_1 \leq C_2 < \infty, \rho > 0. \tag{3.4}$$

(**H5**) La fonction densité $f_{X_{\mathbf{i}}X_{\mathbf{j}}}$ du couple $(X_{\mathbf{i}}, X_{\mathbf{j}})$ est bornée dans D et $\left| f_{X_{\mathbf{i}}X_{\mathbf{j}}}(u,v) - f_{X_{\mathbf{i}}}(u) f_{X_{\mathbf{j}}}(v) \right| \leq C$ pour tout $\mathbf{i} \neq \mathbf{j}$ et $(u,v) \in D \times D$.

(**H6**) $\forall n, m \in \mathbb{N} \quad \psi(n,m) \leq C \min(n,m)$ et

$(1 - s(1-\tilde{\gamma}))\theta > N\left((2 + s(2-\tilde{\gamma}))d + 2s(2+\gamma-\tilde{\gamma})\right)$, où $2 < s < \frac{1}{1-\tilde{\gamma}}$.

(**H7**) $\forall n, m \in \mathbb{N} \quad \psi(n,m) \leq C(n+m+1)^{\tilde{\beta}}, \tilde{\beta} \geq 1$ et

$(1 - s(1-\tilde{\gamma}))\theta > N\left(2 + (2 + s(2-\tilde{\gamma}))d + s(4 + 2\tilde{\beta} + 2\gamma - 3\tilde{\gamma})\right)$ où $2 < s < \frac{1}{1-\tilde{\gamma}}$.

(**H8**) les fonctions densités $f_{X_{\mathbf{i}}}$ et $f_{X_{\mathbf{i}},Y_{\mathbf{i}}}$ de $X_{\mathbf{i}}$ et du couple $(X_{\mathbf{i}}, Y_{\mathbf{i}})$, respectivement, sont telles que :

$$\sup_{x\in D, \mathbf{i}\in\mathcal{V}_{\mathbf{i}_0}} |f_{X_{\mathbf{i}}}(x) - f(x)| = o(1), \quad \sup_{x\in D, \mathbf{i}\in\mathcal{V}_{\mathbf{i}_0}} |g_{\mathbf{i}}(x) - g(x)| = o(1) \qquad \text{as} \quad \mathbf{n} \to \infty,$$

avec $g_{\mathbf{i}}(x) = \int y f_{X_{\mathbf{i}},Y_{\mathbf{i}}}(x,y)\,dy$, $g(x) = \int y f_{X,Y}(x,y)\,dy$.

La densité conditionnelle $f_{Y_{\mathbf{i}},Y_{\mathbf{j}}|X_{\mathbf{i}},X_{\mathbf{j}}}$ de $(Y_{\mathbf{i}}, Y_{\mathbf{j}})$ sachant $(X_{\mathbf{i}}, X_{\mathbf{j}})$ et la densité conditionnelle $f_{Y_{\mathbf{i}}|X_{\mathbf{j}}}$ de $Y_{\mathbf{i}}$ sachant $X_{\mathbf{j}}$ existent et

$$f_{Y_{\mathbf{i}},Y_{\mathbf{j}}|X_{\mathbf{i}},X_{\mathbf{j}}}(y,t|u,v) < C \qquad f_{Y_{\mathbf{i}}|X_{\mathbf{j}}}(y|u) < C,$$

pour tout $y, t, u, v, \mathbf{i}, \mathbf{j}$; $(u,v) \in D \times D$.

Remarque 1.

1. *Dans l'hypothèse (**H1**), f est lipschitzienne. Elle intervient, en particulier, dans le terme du biais (voir la condition* (L1) *dans les preuves des lemmes 1 et 2). Elle permet avec l'hypothèse (**H8**) d'établir la vitesse de convergence dans le corollaire 1.*
2. *La condition sur $k_{\mathbf{n}}$ dans l'hypothèse (**H2**) étend celle sur le nombre de voisins supposés par [265], dans le contexte de séries chronologiques fonctionnelles dépendantes. La condition sur $k_{\mathbf{n}}^{'}$ est la même que celle supposée par [98] sur le nombre de voisins du site $\mathbf{i}_0$.*
3. *La condition (3.3) sur le noyau K_1 est requise dans les preuves des lemmes 1 et 2, respectivement (pour plus de détails sur cette condition, voir [87]).*
4. *Les hypothèses (**H4**)-(**H8**) sont utiles dans l'estimation non-paramétrique des données spatiales non stationnaires, voir par exemple [98] pour plus de détails. En particulier, (**H4**) est nécessaire pour simplifier les preuves. Elle est satisfaite, par exemple, par plusieurs noyaux avec support compact tels que : triangular, biweight, triweight, Epanechnikov, Parzen kernels.*

3.3.2 Les résultats asymptotiques

Le théorème suivant donne la convergence presque complète du prédicteur.

Théorème 1. *Sous les hypothèses (***H1***)-(***H5***), (***H8***) et (***H6***) ou (***H7***), nous avons*

$$\left|\widehat{\mathrm{Y}}_{\mathbf{i}_0} - \mathrm{Y}_{\mathbf{i}_0}\right| \underset{\mathbf{n}\to\infty}{\longrightarrow} 0 \quad p.c. \tag{3.5}$$

Si $r(\cdot)$ est lipschitzienne, nous pouvons déduire la vitesse de convergence presque complète dans le corollaire suivant

Corollaire 1. *Sous les hypothèses (***H1***)-(***H5***), (***H8***) et (***H6***) ou (***H7***), lorsque* $\mathbf{n}\to\infty$*, nous avons,*

$$\left|\widehat{\mathrm{Y}}_{\mathbf{i}_0} - \mathrm{Y}_{\mathbf{i}_0}\right| = \mathrm{O}\left(\left(\frac{k_{\mathbf{n}}}{k'_{\mathbf{n}}}\right)^{1/d} + \left(\frac{\log(\widehat{\mathbf{n}})}{k_{\mathbf{n}}}\right)^{1/2}\right) \quad p.c. \tag{3.6}$$

Les résultats du théorème 1 et du corollaire 1 peuvent être facilement démontrés à partir de ceux asymptotiques (indiqués respectivement dans les lemmes 1 et 2) de l'estimateur de la fonction de régression (2.3).

Lemme 1. *Sous les hypothèses (***H1***)-(***H5***), (***H8***) et (***H6***) ou (***H7***), nous avons*

$$\sup_{x\in\mathrm{D}} |r_{\mathrm{kNN}}(x) - r(x)| \underset{\mathbf{n}\to\infty}{\longrightarrow} 0 \quad p.c. \tag{3.7}$$

Lemme 2. *Sous les hypothèses (***H1***)-(***H5***), (***H8***) et (***H6***) or (***H7***) et si* $r(\cdot)$ *est lipschitzienne, lorsque* $\mathbf{n}\to\infty$*, nous avons*

$$\sup_{x\in\mathrm{D}} |r_{\mathrm{kNN}}(x) - r(x)| = \mathrm{O}\left(\left(\frac{k_{\mathbf{n}}}{k'_{\mathbf{n}}}\right)^{1/d} + \left(\frac{\log(\widehat{\mathbf{n}})}{k_{\mathbf{n}}}\right)^{1/2}\right) \quad p.c. \tag{3.8}$$

Les preuves seront reportées à l'annexe A. La principale difficulté dans les preuves des lemmes 1 et 2 vient du caractère aléatoire de la fenêtre $\mathrm{H}_{\mathbf{n},x}$ et le fait que nous n'avons pas dans le numérateur et le dénominateur de $r_{\mathrm{kNN}}(x)$ des sommes de variables identiquement distribuées. Nous utilisons des techniques d'encadrements pour lever certaines difficultés. L'idée est d'encadrer sensiblement $\mathrm{H}_{\mathbf{n},x}$ par deux paramètres réels non aléatoires. Les preuves du théorème 1 et du corollaire 1 peuvent être omises car elles proviennent directement de celles des lemmes 1 et 2.
Dans la section suivante (3.4), nous introduisons une procédure de *classification supervisée* qui est un cas particulier de la méthode de prédiction proposée.

3.4 Application à la discrimination : règle de classement k-NN

Le problème classique de la *classification supervisée* ou *discrimination* consiste, simplement, à prédire la nature inconnue, d'un objet, qui peut prendre deux valeurs possibles par exemple : absence ou présence, malade ou sain, noir ou blanc, un sol fertile ou non-fertile, une terre cultivable ou non-cultivable. L'objet inconnu peut aussi avoir trois modalités ou plus. Un autre exemple, dans le contexte de la pandémie covid-19 [2], on peut s'intéresser à l état de santé d'un patient, qui arrive sous observation (consultation médicale). On attribue, au patient, une classe ou groupe suivant les états suivants : non-suspect, suspect, malade et guéri. Chaque état définit une classe. La nature inconnue de l'objet est appelée une classe et elle est notée par Y. Elle prend ses valeurs dans un ensemble fini $\{1,...,\mathrm{M}\}$. En *discrimination*, on construit une fonction g qui prend ses valeurs dans $\{1,...,\mathrm{M}\}$. Et qui représente l'affectation $g(\mathrm{X})$ de Y sachant X. La fonction g est appelée *classeur*.
Dans cette section, notre objectif est de prédire Y à travers X dans un site donné en utilisant l'échantillon du couple de variables observées dans certains sites. Comme dans la section 3.2, nous supposons que le site de prédiction $\mathbf{i}_0 \in \mathscr{I}_{\mathbf{n}}$ et que $(\mathrm{X}_{\mathbf{i}_0}, \mathrm{Y}_{\mathbf{i}_0})$ a la même distribution que (X,Y) et les observations $\left\{(\mathrm{X}_{\mathbf{i}}, \mathrm{Y}_{\mathbf{i}})_{\mathbf{i}\in\mathscr{O}_{\mathbf{n}}}\right\}$ sont localement identiquement distribuées.
Nous rappelons que la fonction g est définie sur $\mathbb{R}^d$ et elle prend ses valeurs dans $\{1,...,\mathrm{M}\}$. On se trompe sur Y si $g(\mathrm{X}) \neq \mathrm{Y}$, et la probabilité d'erreur, pour un *classeur* g, est donnée par :

$$\mathrm{L}(g) = \mathrm{P}\{g(\mathrm{X}) \neq \mathrm{Y}\}.$$

2. La maladie à coronavirus 2019, abrégée en COVID – 19 (acronyme anglais signifiant coronavirus disease 2019), est une maladie infectieuse émergente de type zoonose virale, provoquée par le coronavirus SARS – CoV – 2 (ex 2019-nCoV), responsable d'une pandémie ayant débuté en décembre 2019 dans la ville de *Wuhan*, capitale de la province du *Hubei*, en Chine centrale

Il est bien connu que le *classeur* de Bayes, défini par,

$$g^* = \underset{g:\mathbb{R}^d \to \{1,\ldots,\mathrm{M}\}}{\arg\min} \; \mathrm{P}\{g(\mathrm{X}) \neq \mathrm{Y}\},$$

est le meilleur *classeur* possible, en ce qui concerne la perte quadratique. La probabilité minimale d'erreur est appelée *Erreur de Bayes* et elle est notée par $\mathrm{L}^* = \mathrm{L}(g^*)$. Notons que g^* dépend de la distribution de (X,Y) qui est inconnue.
Un estimateur $g_{\mathbf{n}}$ de g est basé sur les observations $\left\{(\mathrm{X}_{\mathbf{i}},\mathrm{Y}_{\mathbf{i}})_{\mathbf{i}\in\mathscr{O}_{\mathbf{n}}}\right\}$; Y est prédit par $g_{\mathbf{n}}\left(\mathrm{X};(\mathrm{X}_{\mathbf{i}},\mathrm{Y}_{\mathbf{i}})_{\mathbf{i}\in\mathscr{O}_{\mathbf{n}}}\right)$. La performance de $g_{\mathbf{n}}$ est mesurée par la probabilité conditionnelle d'erreur, soit :

$$\mathrm{L}_{\mathbf{n}} = \mathrm{L}\left(g_{\mathbf{n}}\right) = \mathrm{P}\left\{g_{\mathbf{n}}\left(\mathrm{X};(\mathrm{X}_{\mathbf{i}},\mathrm{Y}_{\mathbf{i}})_{\mathbf{i}\in\mathscr{O}_{\mathbf{n}}}\right) \neq \mathrm{Y}\right\} \geq \mathrm{L}^*.$$

La suite $\left\{g_{\mathbf{n}},\ \mathbf{n}\in\mathbb{N}^{*\mathrm{N}}\right\}$ est la *règle de discrimination*. Elle a été étudiée de manière approfondie dans la littérature, en particulier pour les données indépendantes et en séries chronologiques (voir [42; 118; 120; 190; 274], pour plus de détails). [373] a abordé une *règle de discrimination* par la méthode non-paramétrique à noyau, supportée par un processus spatial multivarié strictement stationnaire $\left\{\mathrm{X}_{\mathbf{i}}\in\mathbb{R}^d\right\}_{\mathbf{i}\in\mathbb{N}^{\mathrm{N}}}$ et un processus spatial binaire $\{\mathrm{Y}_{\mathbf{i}}\in(0,1)\}_{\mathbf{i}\in\mathbb{N}^{\mathrm{N}}}$. À notre connaissance, ce dernier est le premier travail qui traite une *règle de discrimination* sur un processus spatial.
Dans cette section, nous étendons le prédicteur k-NN précédent (2.4) dans le cas où Y appartient à $\{1,\ldots,\mathrm{M}\}$. Le *classeur* de Bayes g^* peut être approximé par la *règle de régression à noyau* $\left\{g_{\mathbf{n}},\ \mathbf{n}\geq 1\right\}$ basée sur l'estimateur k-NN de régression $r_{\mathrm{kNN}}(.)$. Elle est définie comme suit :

$$\sum_{\mathbf{i}\in\mathscr{O}_{\mathbf{n}}} \mathrm{W}_{\mathbf{ni}}(\mathbf{i}_0)\,\mathbb{I}_{\{\mathrm{Y}_{\mathbf{i}}=g_{\mathbf{n}}(\mathbf{i}_0)\}} = \max_{1\leq j\leq \mathrm{M}} \sum_{\mathbf{i}\in\mathscr{O}_{\mathbf{n}}} \mathrm{W}_{\mathbf{ni}}(\mathbf{i}_0)\,\mathbb{I}_{\{\mathrm{Y}_{\mathbf{i}}=j\}}, \tag{4.1}$$

où

$$\mathrm{W}_{\mathbf{ni}}(\mathbf{i}_0) = \frac{\mathrm{K}_1\left(\frac{\mathrm{X}_{\mathbf{i}_0}-\mathrm{X}_{\mathbf{i}}}{\mathrm{H}_{\mathbf{n},\mathrm{X}_{\mathbf{i}_0}}}\right)\mathrm{K}_2\left(h_{\mathbf{n},\mathbf{i}_0}^{-1}\left\|\frac{\mathbf{i}_0-\mathbf{i}}{\mathbf{n}}\right\|\right)}{\sum_{\mathbf{i}\in\mathscr{O}_{\mathbf{n}}}\mathrm{K}_1\left(\frac{\mathrm{X}_{\mathbf{i}_0}-\mathrm{X}_{\mathbf{i}}}{\mathrm{H}_{\mathbf{n},\mathrm{X}_{\mathbf{i}_0}}}\right)\mathrm{K}_2\left(h_{\mathbf{n},\mathbf{i}_0}^{-1}\left\|\frac{\mathbf{i}_0-\mathbf{i}}{\mathbf{n}}\right\|\right)}. \tag{4.2}$$

Une telle *règle de discrimination* ou de *classification supervisée* $g_{\mathbf{n}}$ est appelée un *estimateur de classeur de Bayes*.
Nous disons, qu'une *règle de discrimination* est bonne [118], si elle est *consistante*, c'est à dire si, $\mathrm{L}_{\mathbf{n}}\to\mathrm{L}$ en probabilité ou presque sûrement lorsque $\mathbf{n}\to\infty$.
La convergence presque sûre de la *règle de discrimination* proposée est établie dans le théorème suivant, à travers la convergence presque complète.

Théorème 2. *Sous les hypothèses (***H1***)-(***H5***), (***H8***) et (***H6***) ou (***H7***), lorsque* $\mathbf{n}\to\infty$,

$$\mathrm{L}_{\mathbf{n}} - \mathrm{L}^* \underset{\mathbf{n}\to\infty}{\longrightarrow} 0 \qquad p.c.$$

La preuve de ce théorème consiste à montrer que [Théorème 2.3 dans 180]

$$\int_{\mathrm{D}} |r(x) - r_{\mathrm{kNN}}(x)|\, f(x)\,dx \underset{\mathbf{n}\to\infty}{\longrightarrow} 0, \qquad p.c.$$

Ce dernier provient directement du lemme 1 et de l'intégrabilité de la fonction de densité.

Après avoir vérifié le comportement asymptotique du prédicteur $\widehat{\mathrm{Y}}_{\mathbf{i}_0}$ et son extension à une *règle de classification supervisée* $\left\{g_{\mathbf{n}},\ \mathbf{n}\in\mathbb{N}^{*\mathrm{N}}\right\}$, nous étudions ses caractéristiques pratiques à travers une application sur des données simulées.

3.5 Simulations numériques

Afin d'évaluer l'efficacité du prédicteur k-NN, on calcule sa précision au moyen de la moyenne des erreurs absolues moyennes (AMAE). Nous la comparons avec celle du prédicteur proposé par [98]. L'application est basée sur des données simulées, on observe $(\mathrm{X}_{i,j},\mathrm{Y}_{i,j})$, $1\leq i\leq n_1, 1\leq j\leq n_2$ telle que $\forall\, i,j$:

$$\mathrm{X}_{i,j} = \mathrm{A}_{i,j}\,\mathrm{U}_{i,j}\,\mathrm{T}_{i,j} + (1-\mathrm{A}_{i,j})\left(6+\mathrm{U}_{i,j}\mathrm{Z}_{i,j}\right)$$

et

$$\mathrm{Y}_{i,j} = r(\mathrm{X}_{i,j}) + \varepsilon_{i,j}, \qquad \text{avec} \qquad r(x) = x^2.$$

Les $A_{i,j}$ sont des variables aléatoires de Bernoulli indépendantes de paramètre 0.5,

$$T = (T_{i,j})_{1\le i\le n_1, 1\le j\le n_2} = \text{GRF}(0,5,3),\ \ Z = (Z_{i,j})_{1\le i\le n_1, 1\le j\le n_2} = \text{GRF}(0,\sigma,3)$$

et

$$\varepsilon = (\varepsilon_{i,j})_{1\le i\le n_1, 1\le j\le n_2} = \text{GRF}(0,0.1,3),$$

où on note par $\text{GRF}(\mu,\sigma,s)$ un champ aléatoire gaussien stationnaire de moyenne μ et covariance définie par $C(h) = \sigma^2 \exp\left(-(\|h\|/s)^2\right)$, $h \in \mathbb{R}^2, s > 0, \sigma > 0$. Le processus $U = (U_{i,j})_{1\le i\le n_1, 1\le j\le n_2}$ permet de contrôler la dépendance locale entre l'ensemble des sites, soit :

$$U_{i,j} = \frac{1}{n_1 \times n_2} \sum_{t,m} \exp\left(-\|(i,j)-(m,t)\|/a\right),\ a > 0,$$

Plus a est élevé, plus la dépendance spatiale est faible. Les simulations sont faites avec différentes valeurs de a; $a = 5, 10, 20$, des tailles différentes : $n_1 = 25$, $n_2 = 25$ et $n_1 = 35$, $n_2 = 30$ et deux paramètres de variance σ^2 ($\sigma = 5$ et 0.1). On répète la simulation du modèle 100 fois. Nous choisissons les fonctions noyaux suivantes

$$K_2(x) = \begin{cases} 1-6x^2+6|x|^3, & \text{si } |x| < 0.5; \\ 20(1-|x|)^3, & \text{si } 0.5 \le |x| \le 1; \\ 0, & \text{sinon,} \end{cases}$$

et

$$K_1(x) = 0.75(1-x^2)\mathbb{I}_{\{|x|<1\}}$$

suivant l'hypothèse (**H4**). Les paramètres de lissage sont calculés grâce à la validation croisée suivant la même procédure que celle de [98], en utilisant l'erreur absolue moyenne

$$\text{MAE} = \frac{1}{n_1 \times n_2} \sum_{\mathbf{i}\in\mathscr{I}_\mathbf{n}} \left|Y_\mathbf{i} - \widehat{Y}_\mathbf{i}\right| \quad \text{avec} \quad \widehat{Y}_\mathbf{i} = \tilde{Y}_\mathbf{i} \quad \text{ou} \quad \tilde{Y}_\mathbf{i}^{\text{NW}},$$

où

$$\tilde{Y}_\mathbf{i} = \frac{\sum_{\mathbf{j}\neq\mathbf{i}} Y_\mathbf{j} K_1\left(\frac{X_\mathbf{i}-X_\mathbf{j}}{H_{\mathbf{n},X_\mathbf{i}}}\right) K_2\left(h_{\mathbf{n},\mathbf{i}}^{-1}\left\|\frac{\mathbf{i}-\mathbf{j}}{\mathbf{n}}\right\|\right)}{\sum_{\mathbf{j}\neq\mathbf{i}} K_1\left(\frac{X_\mathbf{i}-X_\mathbf{j}}{H_{\mathbf{n},X_\mathbf{i}}}\right) K_2\left(h_{\mathbf{n},\mathbf{i}}^{-1}\left\|\frac{\mathbf{i}-\mathbf{j}}{\mathbf{n}}\right\|\right)}$$

et

$$\tilde{Y}_\mathbf{i}^{\text{NW}} = \frac{\sum_{\mathbf{j}\neq\mathbf{i}} Y_\mathbf{j} K_1\left(\frac{X_\mathbf{i}-X_\mathbf{j}}{h_\mathbf{n}}\right) K_2\left(\rho_\mathbf{n}^{-1}\left\|\frac{\mathbf{i}-\mathbf{j}}{\mathbf{n}}\right\|\right)}{\sum_{\mathbf{j}\neq\mathbf{i}} K_1\left(\frac{X_\mathbf{i}-X_\mathbf{j}}{h_\mathbf{n}}\right) K_2\left(\rho_\mathbf{n}^{-1}\left\|\frac{\mathbf{i}-\mathbf{j}}{\mathbf{n}}\right\|\right)}. \tag{5.1}$$

Nous notons par NW-prédiction, la méthode de [98], définie dans l'équation (5.1).

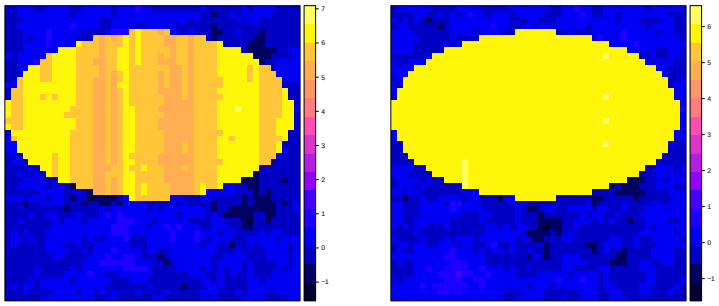

FIGURE 3.1 – Champs aléatoires simulés avec dépendance spatiale mesurée par $a = 10$, sur une grille 50×50.$\sigma = 5$ (gauche) et $\sigma = 0,1$ (droite) .

TABLEAU 3.1 – Tableau de comparaison de la performance des méthodes de prédictions NW-prédiction et k-NN -prédiction, respectivement

			NW-prédiction		k-NN prédiction		
$n_1 \times n_2$	σ	a	Mean	SD	Mean	SD	$p-value$
25×25	5	5	0.303	0.0047	0.241	0.001	**
		10	0.548	0.0308	0.396	0.017	**
		20	0.747	0.0471	0.579	0.024	**
	0.1	5	0.289	0.0045	0.149	0.0001	***
		10	0.428	0.0052	0.198	0.0001	***
		20	0.629	0.0006	0.289	0.0012	***
35×30	5	5	0.235	0.002	0.208	0.0002	**
		10	0.367	0.009	0.288	0.0023	**
		20	0.476	0.010	0.405	0.0065	**
	0.1	5	0.169	0.001	0.141	0.00001	***
		10	0.271	0.002	0.178	0.00004	***
		20	0.482	0.004	0.241	0.00023	***

Le tableau 3.1 donne les moyennes (Mean) et les écarts types (SD) des erreurs absolues moyennes des deux méthodes, sur les 100 réplications du modèle. La colonne intitulée $p-value$ donne, pour chaque cas considéré, $lap-value$ d'un t-test effectué afin de déterminer si l'erreur absolue moyenne de NW-prédiction est significativement supérieure à celle de la méthode k-NN. On remarque que les erreurs produites par k-NN prédiction sont plus petites que celles produites par NW-prédiction dans tous les cas du paramètre de dépendance spatiale a et du paramètre d'écart type σ. En particulier, la méthode k-NN est significativement plus efficace que NW-prédiction avec une valeur de $p-value$ très faible lorsque l'écart est petit. Cela signifie que la méthode k-NN est plus adaptée à une structure hétérogène locale des données. Nous notons que les valeurs des $p-values$ sont très proches de 0. Elles sont remplacées par le symbole (*). Plus on a des (*) plus la $p-value$ est proche de 0.

3.6 Conclusion

Dans ce Chapitre, nous avons proposé une méthode de régression non-paramétrique basée sur l'approche des voisins les plus proches. L'estimateur permet de définir un prédicteur spatial non-paramétrique et une *règle de discrimination* appliqués sur un processus spatial non strictement stationnaire. L'originalité de la méthode proposée est de prendre en compte à la fois la distance entre les sites et celle entre les observations. Dans le cadre de la prédiction, ce travail est une extension de [98] sur le prédicteur spatial du noyau d'un processus stationnaire multivarié. Nous étendons également la *règle de discrimination à noyau* de [373] qui considère un processus spatial strictement stationnaire multivarié et deux classes. La méthode de classification proposée pourrait être appliquée à des ensembles contenant trois classes ou plus. Les résultats asymptotiques des méthodes élaborées ont été établis. Cette *règle de discrimination* sera appliquée à un problème de prédiction grâce à un ensemble de données environnementales sur les poissons démersaux (Chapitre 5). Les résultats produits par la simulation numérique montrent que la méthode k-NN proposée surpasse la méthode à noyau de [98]. En effet, *k-NN prédiction* offre des prédictions plus précises que celles produites par *NW-prédiction*. Nous pouvons dire que la méthodologie proposée est une bonne alternative, comparée à l'approche classique de prédiction appliquée sur les données spatiales, surtout quand ces dernières présentent une structure spatiale hétérogène.

CHAPITRE 4

PRÉDICTION ET CLASSIFICATION SUPERVISÉES POUR DES PROCESSUS FONCTIONNELS SPATIALEMENT DÉPENDANTS

Résumé

Dans ce Chapitre, nous considérons une prédiction non-paramétrique et une *classification supervisée*, d'un processus spatial et fonctionnel observé suivant un plan d'échantillonnage non aléatoire. Le prédicteur proposé est basé sur la régression fonctionnelle et dépend de deux noyaux dont l'un contrôle la structure spatiale et l'autre mesure la proximité entre les observations fonctionnelles. L'erreur quadratique moyenne et la convergence presque complète (ou sûre) sont étudiées lorsque l'échantillon considéré est une séquence α-mélange localement stationnaire. Des études numériques ont été réalisées afin d'illustrer le comportement de notre prédicteur. Cette application par simulation d'un modèle numérique montre que la méthode de prédiction proposée est plus performante que celle classique qui ne prend pas en compte la structure spatiale.

4.1 Introduction

L' analyse des données fonctionnelles (ADF) est la branche mathématique qui traite la théorie fonctionnelle et la mise en pratique de cette dernière pour la modélisation des données qui se présentent sous la forme de fonctions, de courbes, d'images ou d'objets généralement plus complexes. Elle modélise les phénomènes, mesurés (et continus) dans le temps ou dans l'espace/ espace-temps, appelés simplement données fonctionnelles. Au cours de la dernière décennie, l'ADF a connu un développement important avec des applications dans beaucoup de domaines scientifiques tels que l'écologie [370], la médecine [321], et les sciences de l'environnement [104; 137; 168; 200; 340], entre autres. Elle a été appliquée en statistique spatiale, une branche mathématique qui étudie les processus spatialement dépendants.
L'analyse et le traitement d'informations spatialement distribuées et qui sont mesurées continument dans le temps ou espace/espace-temps font appel à la modélisation fonctionnelle. Ils ont contribué aux développements de nouvelles théories mathématiques en statistique fonctionnelle ; avec l'apparition des données continues et spatialement dépendantes. Ils ont donné naissance à une nouvelle branche mathématique qui est une combinaison de la statistique spatiale et celle fonctionnelle. Les travaux de [310] vont dans ce sens. Ces derniers ont montré la nécessité de prendre en compte l'aspect fonctionnel dans la modélisation des données spatialement corrélées. Ils ont étendu le modèle spatial auto-régressif et le modèle moyenne mobile spatial aux processus stochastiques prenant leurs valeurs dans les espaces de Hilbert. La méthodologie de projection basée sur les fonctions propres de l'opérateur d'auto-covariance a été utilisée dans les travaux de [311; 312]. Le problème de régression évoqué dans le Chapitre 3 peut se poser, lorsque, la situation, selon laquelle il s'avère utile d'étudier le lien explicatif entre des variables dont certaines sont fonctionnelles, se présente. Par exemple, en biologie marine, il est important de prendre en compte l'influence des paramètres environnementaux ou écologiques sur la variabilité de la biomasse des poissons et sur leur distribution spatio-temporelle. En recherche pétrolière, on peut s'intéresser à la prédiction des paramètres physiques des couches pétrolifères en tenant compte d'autres paramètres disponibles dans les champs pétroliers (voir [31]). La relation explicative entre variables (variable de réponse et co-variables) est largement étudiée, dans des problèmes de prédiction et de classification, à l'aide des méthodes classiques

de l'ADF. L'analyse fonctionnelle en statistique spatiale permet ainsi de construire, entre autres, des méthodes de régression, de prédiction et de classification. Cependant, cette étude est relativement limitée du point de vue théorique et pratique. En plus, elle porte, en grande partie, sur des modèles paramétriques/et semi-paramétriques [255; 256; 378]. La grande dimension de ces types de données pose des défis à la fois théoriques et pratiques qui font que les modèles paramétriques/et semi-paramétriques ne soient point adaptés dans ce nouveau contexte spatio-fonctionnel. Par conséquent, la modélisation de ces problèmes complexes fait appel aux modèles de régressions non-paramétriques comme méthodes alternatives surtout lorsque la relation entre les deux variables d'intérêt n'est pas linéaire.
Cette contribution étudie dans un cadre fonctionnel les problématiques soulevées dans le Chapitre 3. Ainsi notre objectif, dans ce Chapitre, est de proposer une nouvelle approche de prédiction spatiale non-paramétrique dans un cadre fonctionnel. Elle prend en compte des processus plus généraux qui reposent sur des hypothèses moins restrictives. Le prédicteur proposé est basé sur le fait que la co-variable est de nature fonctionnelle, il n'y a pas de modèle de corrélation paramétrique sur le terme d'erreur et les observations sont supposées être localement identiquement distribuées, contrairement à l'hypothèse de [330]. Nous incorporons, dans la construction du prédicteur, l'aspect non linéaire de la dépendance spatiale. L'originalité du prédicteur proposé est de tirer profit des travaux antérieurs (voir [96; 150; 151; 330]) résumés dans le Chapitre 2 et de les adapter dans un contexte plus général. En effet, nous tenons compte de nouveaux paramètres apportés par les divers champs d'applications de la statistique spatiale et fonctionnelle dont nous rappelons comme suit : la médecine, l'épidémiologie, la géologie, l'écologie, l'agronomie, les sciences halieutiques, l'océanographie, l'étude de la qualité du sol, les sciences de l'environnement, etc. Chaque domaine d'application se caractérise par des paramètres qui définissent le processus spatial et les types d'hypothèses nécessaires pour répondre aux problématiques des questions qui s'y posent. Par conséquent, notre prédicteur dépend de deux noyaux et repose sur des conditions plus générales à savoir la stationnarité locale, la dépendance spatiale non linéaire entre les sites d'observations et le design fixe de l'échantillonnage. L'idée d'incorporer un deuxième noyau, qui mesure la proximité entre les sites, a été présentée par [96; 330], pour étudier l'estimation de la densité et de la régression, respectivement, portant sur des processus spatio-fonctionnels strictement stationnaires. Nous notons que l'incorporation d'une structure de corrélation spatiale dans un estimateur de régression non-paramétrique en supposant que le terme d'erreur est un processus stationnaire de second ordre avec un modèle de corrélation paramétrique, a été utilisée dans les travaux de [150; 151]. Ces auteurs ont utilisé un estimateur de régression linéaire locale et un critère de validation croisée générale pour l'effet de la corrélation spatiale. Malgré ces efforts, une méthode de prédiction qui prend en compte explicitement la structure spatiale dans sa construction n'est toujours pas disponible. Une nouvelle méthode de *classification supervisée* (ou *règle de discrimination*) que nous appelons Méthode Fonctionnelle et Spatiale de Discrimination (MFSD) découle du prédicteur. Elle correspond au cas particulier où la variable réponse Y appartient à un ensemble discret.
Le reste du Chapitre 4 est organisé comme suit. Dans la section 4.2, nous introduisons le modèle de régression qui est le support de base du prédicteur. La section 4.3 est dédiée aux hypothèses et aux résultats asymptotiques (convergence presque complète [1]) du prédicteur alors que la section 4.4 applique le modèle de régression à une règle de *classification supervisée*. Cette dernière est un cas particulier du prédicteur. La section 4.5 donne une application à des données simulées pour mettre en évidence les performances de la méthode proposée. La section 4.6 est consacrée à la conclusion. Enfin, les preuves des principaux résultats asymptotiques sont reportées à l'annexe B.

4.2 Modèle de régression et construction du prédicteur

Soit $\{Z_{\mathbf{i}} = (X_{\mathbf{i}}, Y_{\mathbf{i}}), \mathbf{i} \in \mathbb{Z}^N, N \geq 1\}$, un processus spatial défini dans l'espace probabilisé $(\Omega, \mathscr{A}, \mathbb{P})$, avec $X_{\mathbf{i}}$ une variable qui prend ses valeurs dans un espace semi-métrique séparable $(\mathscr{E}, d(.,.))$ de dimension éventuellement infinie (i.e. $X_{\mathbf{i}}$ est une variable fonctionnelle et $d(.,.)$ est une semi-métrique) et $Y_{\mathbf{i}}$ prend ses valeurs dans $\mathbb{R}$. Dans ce qui suit, $\|.\|$ désigne une norme quelconque dans $\mathbb{R}^d$ ou $\mathbb{R}^N$. Les paramètres C et C$'$ désignent des constantes positives arbitraires qui peuvent varier d'une ligne à l'autre. Pour chaque réel u, nous notons par $\lfloor u \rfloor$ sa partie entière. Par ailleurs, $u_{\mathbf{n}} = O(v_{\mathbf{n}})$ signifie que $\exists C > 0$ telle que $|u_{\mathbf{n}}/v_{\mathbf{n}}| \leq C$ quand $v_{\mathbf{n}} \to \infty$ et $u_{\mathbf{n}} = o(v_{\mathbf{n}})$ signifie que $|u_{\mathbf{n}}/v_{\mathbf{n}}| \to 0$ quand $v_{\mathbf{n}} \to \infty$, où $\mathbf{n} = (n_1, \ldots, n_N) \in \mathbb{R}^N$. Nous posons $\mathscr{I}_{\mathbf{n}} = \{\mathbf{i} = (i_1, \ldots, i_N), 1 \leq i_k \leq n_k, k = 1, \ldots, N\}$.

1. Soit $(z_n)_{n\in\mathbb{N}}$ une séquence de variables aléatoires à valeur réelle. z_n converge presque complètement (*p.c.*) vers zéro si, et seulement si, $\forall \varepsilon > 0$, $\sum_{n=1}^{\infty} P(|z_n| > \varepsilon) < \infty$. De plus, nous disons que la vitesse de convergence presque complète de z_n vers zéro est d'ordre u_n (avec $u_n \to 0$) et nous écrivons $z_n = O_{p.c.}(u_n)$ si, et seulement si, $\exists \varepsilon > 0$ est telle que $\sum_{n=1}^{\infty} P(|z_n| > \varepsilon u_n) < \infty$. Ce type de convergence implique à la fois la convergence presque sûre et la convergence en probabilité.

Le processus $\{Z_{\mathbf{i}}\ \mathbf{i} \in \mathbb{Z}^N\}$ est observable sur le domaine rectangulaire $\mathscr{I}_{\mathbf{n}}$ et nous posons $\widehat{\mathbf{n}} = n_1 \times \ldots \times n_N$ pour désigner la taille de l'échantillon d'observation. Désormais, nous supposons, par souci de simplicité, que $n_1 = n_2 = \ldots = n_N = n$ (voir [132; 134; 135]), mais les résultats suivants peuvent être étendus à un cadre plus général. Nous avons $\mathbf{n} \to \infty$ si $n \to \infty$. Pour chaque site $\mathbf{i}_0$, on note par $k_{\mathbf{n}} = k_{\mathbf{n},\mathbf{i}_0} = \sum 1_{\{\|\mathbf{i}-\mathbf{i}_0\| \leq d_{\mathbf{n}}\|\}}$ le nombre de sites $\mathbf{i}$ voisins de $\mathbf{i}_0$ pour lesquelles la distance entre $\mathbf{i}_0$ et $\mathbf{i}$ est inférieure ou égale à $d_{\mathbf{n}} > 0$. Ainsi, nous pouvons dire que $d_{\mathbf{n}} \to \infty$ quand $\mathbf{n} \to \infty$. Cela suppose que la distance entre les sites de mesures augmentent (éventuellement) à mesure que la taille de l'échantillon augmente. Nous posons $k_{\mathbf{n}} = C_N d_{\mathbf{n}}^N + O(d_{\mathbf{n}}^\beta)$ quand $d_{\mathbf{n}} \to +\infty$, avec $0 < \beta < N$ et C_N est une constante qui dépend de N. Soit le couple (X,Y) de même distribution que $(X_{\mathbf{i}_0}, Y_{\mathbf{i}_0})$.
Nous considérons que les variables $(X_{\mathbf{i}}, Y_{\mathbf{i}})_{\mathbf{i} \in \mathscr{I}_{\mathbf{n}}}$ sont localement identiquement distribuées (voir par exemple [212] qui a considéré l'estimation de la densité pour des séries chronologiques localement identiquement distribuées et le Chapitre précédent). Et de plus, $\mathbb{E}|Y_{\mathbf{i}}| < \infty$. Nous remplaçons $X_{\mathbf{i}_0} = x_{\mathbf{i}_0}$ par x dans ce qui suit par abus de notation.
Nous supposons que le processus spatial satisfait le modèle non-paramétrique de régression suivant :

$$Y_{\mathbf{i}} := r(X_{\mathbf{i}}) + \varepsilon_{\mathbf{i}}, \tag{2.1}$$

où la fonction $r(.) = \mathbb{E}(Y_{\mathbf{i}}|X_{\mathbf{i}} = .)$, est supposée être indépendante de $\mathbf{i}$, le bruit $\varepsilon_{\mathbf{i}}$ est centré, α-mélange et indépendant de $X_{\mathbf{i}}$.
L'estimateur de régression $r_{\mathbf{n}}(.)$ permet de définir le prédicteur de l'équation (3.8). Il est défini comme suit :

$$r_{\mathbf{n}}(x) = \begin{cases} \dfrac{g_{\mathbf{n}}(x)}{f_{\mathbf{n}}(x)}, \text{si } f_{\mathbf{n}}(x) \neq 0 \\ \dfrac{1}{\widehat{\mathbf{n}}} \displaystyle\sum_{\mathbf{i} \in \mathscr{I}_{\mathbf{n}}} Y_{\mathbf{i}}, \text{ sinon.} \end{cases} \tag{2.2}$$

Où $f_{\mathbf{n}}$ et $g_{\mathbf{n}}$ sont des fonctions définies, respectivement, par :

$$f_{\mathbf{n}}(x) = \frac{1}{a_{\mathbf{n}}} \sum_{\mathbf{i} \in \mathscr{I}_{\mathbf{n}}} K_1\left(\frac{d(x,X_{\mathbf{i}})}{b_{\mathbf{n}}}\right) K_{2,\rho_{\mathbf{n}}}(\|\mathbf{i}_0 - \mathbf{i}\|),$$

$$g_{\mathbf{n}}(x) = \frac{1}{a_{\mathbf{n}}} \sum_{\mathbf{i} \in \mathscr{I}_{\mathbf{n}}} Y_{\mathbf{i}} K_1\left(\frac{d(x,X_{\mathbf{i}})}{b_{\mathbf{n}}}\right) K_{2,\rho_{\mathbf{n}}}(\|\mathbf{i}_0 - \mathbf{i}\|).$$

Avec $a_{\mathbf{n}} = \displaystyle\sum_{\mathbf{i} \in \mathscr{I}_{\mathbf{n}}} K_{2,\rho_{\mathbf{n}}}(\|\mathbf{i}_0 - \mathbf{i}\|)\, \mathbb{E}\left[K_1\left(\frac{d(x,X_{\mathbf{i}})}{b_{\mathbf{n}}}\right)\right]$,
où $K_{2,\rho_{\mathbf{n}}}(\|\mathbf{i}_0 - \mathbf{i}\|) = K_2\left(\frac{\|(\mathbf{i}_0-\mathbf{i})/\mathbf{n}\|}{\rho_{\mathbf{n}}}\right) = K_2\left(\frac{\|\mathbf{i}_0-\mathbf{i}\|}{n\rho_{\mathbf{n}}}\right)\left(\frac{\mathbf{i}}{\mathbf{n}} = (\frac{i_1}{n}, \frac{i_2}{n}, \ldots \frac{i_N}{n})\right)$. $b_{\mathbf{n}}$ et $\rho_{\mathbf{n}}$ sont des paramètres de lissages qui tendent vers zéro ; K_1 et K_2 sont des noyaux, définis dans l'hypothèse **H1**.
Nous remarquons que, si nous prenons la distance euclidienne avec N = 2 (grille carrée), alors $k_{\mathbf{n}} \leq 4d_{\mathbf{n}}^2 - 4d_{\mathbf{n}} + 4$. Ce qui conduit à $k_{\mathbf{n}} = O(d_{\mathbf{n}}^2) = O(\widehat{\mathbf{n}}\rho_{\mathbf{n}}^2)$.
L'objectif principal de l'application de l' estimateur de régression $r_{\mathbf{n}}(x)$ défini dans l'équation (2.2) est la prédiction du processus $\{Y_{\mathbf{i}}, \mathbf{i} \in \mathbb{Z}^N\}$ dans certains endroits où le processus n'est pas observé. Plus particulièrement pour prédire la valeur non observée $Y_{\mathbf{i}_0}$ à un site $\mathbf{i}_0 \in \mathbb{Z}^N$, on utilise un nombre suffisant d'observations $(X_{\mathbf{i}}, Y_{\mathbf{i}})$ disponible dans $\mathscr{O}_{\mathbf{n}} = \mathscr{I}_{\mathbf{n}} \setminus \{\mathbf{i}_0\}$. L'ensemble $\mathscr{O}_{\mathbf{n}}$ est dans le domaine rectangulaire $\mathscr{I}_{\mathbf{n}}$. La prédiction de $Y_{\mathbf{i}_0}$ dans le site $\mathbf{i}_0 \in \mathbb{Z}^N$ est calculée avec des fenêtres optimales $b_{\mathbf{n}}^\sharp$ et $\rho_{\mathbf{n}}^\sharp$ détaillées dans la procédure de prédiction de la section 4.5, soit :

$$\widehat{Y}_{\mathbf{i}_0}^\sharp = \frac{\sum_{\mathbf{i} \in \mathscr{O}_{\mathbf{n}}} Y_{\mathbf{i}} K_1\left(\frac{d(x,X_{\mathbf{i}})}{b_{\mathbf{n}}^\sharp}\right) K_{2,\rho_{\mathbf{n}}^\sharp}(\|\mathbf{i}_0 - \mathbf{i}\|)}{\sum_{\mathbf{i} \in \mathscr{O}_{\mathbf{n}}} K_1\left(\frac{d(x,X_{\mathbf{i}})}{b_{\mathbf{n}}^\sharp}\right) K_{2,\rho_{\mathbf{n}}^\sharp}(\|\mathbf{i}_0 - \mathbf{i}\|)}, \tag{2.3}$$

si le dénominateur n'est pas nul sinon la prédiction est égale à la moyenne empirique. La précision de cette prédiction (2.3) est calculée et comparée à celle qui ne prend pas en compte la structure spatiale définie comme suit :

$$\widehat{Y}_{\mathbf{i}_0}^\star = \frac{\sum_{\mathbf{i} \in \mathscr{O}_{\mathbf{n}}} Y_{\mathbf{i}} K_1\left(\frac{d(x,X_{\mathbf{i}})}{b_{\mathbf{n}}^\star}\right)}{\sum_{\mathbf{i} \in \mathscr{O}_{\mathbf{n}}} K_1\left(\frac{d(x,X_{\mathbf{i}})}{b_{\mathbf{n}}^\star}\right)}. \tag{2.4}$$

Avec $b_{\mathbf{n}}^\star$ la fenêtre optimale détaillée dans la sous section 4.5.1.
Remarquons que l'équation (2.4) est basée sur l'estimation de régression non-paramétrique qui ne prend

pas en compte la structure spatiale [101] :

$$r_{\mathbf{n}}^{cl}(x) = \frac{\sum_{\mathbf{i}\in\mathscr{I}_{\mathbf{n}}} Y_{\mathbf{i}} K_1\left(\frac{d(x,X_{\mathbf{i}})}{b_{\mathbf{n}}}\right)}{\sum_{\mathbf{i}\in\mathscr{I}_{\mathbf{n}}} K_1\left(\frac{d(x,X_{\mathbf{i}})}{b_{\mathbf{n}}}\right)}. \tag{2.5}$$

4.3 Hypothèses et propriétés asymptotiques

Nous introduisons l'hypothèse α-mélange. Elles prend en compte la dépendance spatiale, nous supposons que le processus $\{Z_{\mathbf{i}} = (X_{\mathbf{i}}, Y_{\mathbf{i}}),\ \mathbf{i} \in \mathbb{Z}^N\}$ satisfait la condition de mélange définie (dans [66]) comme suit : il existe une fonction $\chi(t) \searrow 0$ quand $t \to \infty$, telle que

$$\begin{aligned} \alpha(\sigma(S),\sigma(S')) &= \sup\{|\mathbb{P}(A\cap B) - \mathbb{P}(A)\mathbb{P}(B)|, A \in \sigma(S), B \in \sigma(S')\}, \\ &\leq \psi(\mathrm{Card}(S),\mathrm{Card}(S'))\chi(\mathrm{dist}(S,S')), \end{aligned}$$

où $\mathrm{dist}(S,S')$ est la distance euclidienne entre les deux ensembles de sites S et S′, Card(S) est le cardinal de S, σ(S) (resp. σ(S′)) est le σ-algèbre engendré par $\{Z_{\mathbf{i}}, \mathbf{i} \in S\}$ (resp. $\{Z_{\mathbf{i}}, \mathbf{i} \in S'\}$) et ψ(·) est une fonction symétrique positive non décroissante. Nous rappelons que le processus est fortement mélange si $\psi \equiv 1$ [129]. Nous supposons que le mélange fort vérifie l'une des deux conditions suivantes sur $\chi(i)$:

$$\chi(i) \leq C i^{-\theta}, \text{ for some } \theta > 0, \tag{3.1}$$

i.e. que $\chi(i)$ tend vers zéro à la vitesse polynomiale, ou

$$\chi(i) \leq C\exp(-si), \text{pour une constante } s > 0,$$

i.e. que $\chi(i)$ tend vers zéro a la vitesse exponentielle. Concernant la fonction χ(·), par souci de simplicité, nous allons étudier uniquement le cas où χ(·) tend vers zéro à la vitesse polynomiale. Cependant, des résultats asymptotiques similaires pourraient être établis en considérant le cas où χ(·) tend vers zéro à la vitesse exponentielle (ce qui implique le cas polynomial). Tout au long du Chapitre, on supposera que ψ satisfait

$$\forall n, m \in \mathbb{N}, \quad \psi(n,m) \leq C\min(n,m), \tag{3.2}$$

où

$$\psi(n,m) \leq C(n+m+1)^{\kappa}. \tag{3.3}$$

Pour une certaine $C > 0$, et $\kappa \geq 1$. La fonction $\psi(n,m)$ peut être trouvée, par exemple, dans les travaux de [41; 66; 106; 184; 342].
Soit $u_{\mathbf{n}} = \prod_{i=1}^{N} (\log n_i)(\log\log n_i)^{1+\varepsilon}$ pour $\varepsilon > 0$, alors $\sum_{\mathbf{n}\in\mathbb{N}} 1/\hat{\mathbf{n}} u_{\mathbf{n}} < \infty$.
Nous notons par $p_{\mathbf{i}}$ la probabilité de distribution de $X_{\mathbf{i}}$ et par $p_{\mathbf{i},\mathbf{j}}$ la distribution de probabilité jointe du couple $(X_{\mathbf{i}}, X_{\mathbf{j}})$, pour tout **i** et **j**. Les probabilités de petites boules sont notées par $\varphi_{\mathbf{i},x}(h) = \mathbb{P}[X_{\mathbf{i}} \in B(x,h)]$. Rappelons que $\varphi_{\mathbf{i},x}(h)$ tend vers zéro quand h tend vers zéro (voir [146] pour plus détails).
Pour toute variable aléatoire Z et $p \in \mathbb{N}^*$, $\|Z\|_p = (\mathbb{E}[|Z|^p])^{1/p}$.
Des résultats asymptotiques de $r_{\mathbf{n}}(.)$ ont été établis, sous des hypothèses sur r, le noyau, la fenêtre et la condition de dépendance locale.

-**H1** :
— K_1 est une fonction définie de $\mathbb{R}^+$ à $\mathbb{R}^+$, et il existe deux constantes C_{11} et C_{12} avec $0 < C_{11} < C_{12} < \infty$, telles que :

$$C_{11}\mathbb{I}_{[0,1]}(t) \leq K_1(t) \leq C_{12}\mathbb{I}_{[0,1]}(t).$$

— K_2 est une fonction non négative bornée, et nous supposons qu'il existe deux constantes C_{21}, C_{22} et ρ telles que :

$$C_{21}\mathbb{I}_{\{\|s\|\leq\rho\}} \leq K_2(\|\ s\ \|) \leq C_{22}\mathbb{I}_{\{\|s\|\leq\rho\}},\ \forall\ s \in \mathbb{R}^N,\ 0 < C_{21} \leq C_{22} < \infty, \rho > 0. \tag{3.4}$$

- (**H2**) : r est une fonction lipschitzienne, telle que $r \in Lip_{\mathscr{E}}$, où

$$Lip_{\mathscr{E}} = \{f : \mathscr{E} \to \mathbb{R}, \exists C_3 \in \mathbb{R}_*^+, \forall x, x' \in \mathscr{E}, |f(x) - f(x')| < C_3 d(x,x')\}.$$

- (**H3**) : (i) **Condition de dépendance locale :** Pour tout $\mathbf{i} \neq \mathbf{j} \in \mathbb{N}^N$, la probabilité distribution jointe $p_{\mathbf{i},\mathbf{j}}$ de $X_{\mathbf{i}}$ et $X_{\mathbf{j}}$ vérifie

$$\exists \varepsilon \in (0,1], p_{\mathbf{i},\mathbf{j}}(B(x,b_{\mathbf{n}}) \times B(x,b_{\mathbf{n}})) \leq C_4(\varphi_{\mathbf{i},x}(b_{\mathbf{n}})\varphi_{\mathbf{j},x}(b_{\mathbf{n}}))^{\frac{1+\varepsilon}{2}},$$
$$\sup_{\mathbf{i}\in\mathcal{V}_{\mathbf{i}_0}} |\varphi_{\mathbf{i},x}(b_{\mathbf{n}}) - \varphi_{\mathbf{i}_0,x}(b_{\mathbf{n}})| = o(1),$$

avec

$$\mathcal{V}_{\mathbf{i}_0} = \left\{ \mathbf{i} \in \mathcal{J}_{\mathbf{n}}, \left\| \frac{\mathbf{i}-\mathbf{i}_0}{\mathbf{n}} \right\| < \rho_{\mathbf{n},\mathbf{i}} \right\}$$

(ii) **Les probabilités de petites boules :** Pour tout $\mathbf{i}$ et x, il existe deux constantes positives C_1' et C_2' et une fonction $\varphi_x(h)$ qui tend vers zéro lorsque h tend vers zéro, de telle sorte que

$$0 < C_1'\varphi_x(h) \leq \varphi_{\mathbf{i}_0,x}(h) \leq C_2'\varphi_x(h).$$

Remarque 2. *Ces hypothèses sont classiques et habituellement utilisées dans le contexte de la modélisation spatiale non-paramétrique. En effet, les hypothèses (**H1**) et (**H2**) permettent de contrôler, entre autres, le biais de l'estimateur. L'hypothèse (**H1**) est satisfaite, par plusieurs noyaux, par exemple : triangular (Bartlett), biweight, triweight, Epanechnikove, Parzen. La condition lipschitzienne (**H2**) permet d'établir la vitesse de convergence. La condition de dépendance locale (**H3**)(i) est une hypothèse classique dans l'estimation non-paramétrique à noyau sur des processus dépendants non nécessairement strictement stationnaire (voir, [47; 66; 247]).*

Afin de contrôler les contraintes sur les fenêtres de lissage dûes aux coefficients de mélange décroissants à une vitesse polynomiale (3.1), nous définissons les paramètres suivants :

$$\gamma_1 = \frac{2N-\theta}{4N-\theta} \text{ et } \gamma_1^* = \frac{N-\theta}{N(3+2\kappa)-\theta}.$$

Nous établissons le résultat de convergence suivant, de $r_{\mathbf{n}}$.

Théorème 3. *Supposons que les hypothèses **H1**-**H3** sont vérifiées avec* $|Y_{\mathbf{i}}| \leq M$

1. *Si l'hypothèse* (3.2) *est vérifiée et*

$$\widehat{\mathbf{n}}\rho_{\mathbf{n}}^N\varphi_x(b_{\mathbf{n}})^{\gamma_1}\rho_{\mathbf{n}}^{N\gamma_1}\left(\log\widehat{\mathbf{n}}\right)^{-\gamma_1} \to \infty \text{ avec } \theta > 4N,$$

   où

2. *si l'hypothèse* (3.3) *est vérifiée et*

$$\widehat{\mathbf{n}}\rho_{\mathbf{n}}^N\varphi_x(b_{\mathbf{n}})^{\gamma_1^*}\rho_{\mathbf{n}}^{N\gamma_1^*}\left(\log\widehat{\mathbf{n}}\right)^{-\gamma_1^*} \to \infty \text{ avec } \theta > (3+2\kappa)N,$$

   alors

$$\|r_{\mathbf{n}}(x) - r(x)\|_2 \quad = \quad O\left(b_{\mathbf{n}} + \sqrt{\frac{1}{\widehat{\mathbf{n}}\rho_{\mathbf{n}}^N\varphi_x(b_{\mathbf{n}})}}\right).$$

Précisément, nous avons

$$\|r_{\mathbf{n}}(x) - r(x)\|_2 \quad = \quad C_3 \times b_{\mathbf{n}}$$
$$+ \left(2C(2MC_2 + 2M\sqrt{C_4} + C_0) + 4M\right) \times \sqrt{\frac{1}{\widehat{\mathbf{n}}\rho_{\mathbf{n}}^N\varphi_x(b_{\mathbf{n}})}},$$

où C_0 *est une constante qui dépend de celle qui apparait dans le Lemme de [66].*

Remarque 3. *Les conditions sur la fenêtre dans le théorème 3 sont des hypothèses techniques classiques, qui apparaissent (dans les calculs lors de l'étude du comportement asymptotique de l'estimateur) dans le cas particulier où le coefficient de mélange est tel que* χ *tende vers zéro à une vitesse polynomiale (voir [271] et [304] pour quelques exemples). Chacune de ces conditions est liée à un cas spécifique de mélange dans le contexte spatial et elles sont utilisées dans les travaux de [271; 329].*

Convergence uniforme presque complète

Nous considérons un ensemble $\mathscr{D}$ tel que $\mathscr{D} \subset \bigcup_{k=1}^{\nu_{\mathbf{n}}} \mathrm{B}_k$, où $\mathrm{B}_k = \mathrm{B}(x_k, \ell_{\mathbf{n}})$ (notons qu'un tel ensemble peut toujours être construit), $\nu_{\mathbf{n}} > 0$ est une constante, $x_k \in \mathscr{E}$, $k = 1, ..., \nu_{\mathbf{n}}$, et $\ell_{\mathbf{n}} > 0$. Nous supposons que :

-H4 Il existe une fonction positive non croissante Γ telle que,

(i) $\lim_{\mathbf{n}\to\infty} \Gamma(b_{\mathbf{n}}) = 0$ et

$$0 < \mathrm{C}_1'' \Gamma(b_{\mathbf{n}}) \leq \varphi_{\mathbf{i}_0,x}(b_{\mathbf{n}}) \leq \mathrm{C}_2'' \Gamma(b_{\mathbf{n}}), \text{pour tout } x \in \mathscr{D},$$

où C_1'' et C_2'' sont des constantes.

(ii) $\lim_{\mathbf{n}\to\infty} \frac{\widehat{\mathbf{n}} \rho_{\mathbf{n}}^{\mathrm{N}} \Gamma(b_{\mathbf{n}})}{\log \widehat{\mathbf{n}}} \to \infty$.

(iii) $\nu_{\mathbf{n}} = \widehat{\mathbf{n}}^{\beta}$ pour un certain réel $\beta > 0$.

-H5 La condition de dépendance locale : Pour tout $\mathbf{i} \neq \mathbf{j} \in \mathbb{N}^{\mathrm{N}}$, nous supposons que la distribution de probabilité jointe $p_{\mathbf{i},\mathbf{j}}$ de $\mathrm{X}_{\mathbf{i}}$ et $\mathrm{X}_{\mathbf{j}}$ satisfait

$$\exists \varepsilon \in (0,1], p_{\mathbf{i},\mathbf{j},\mathbf{i}\in\mathscr{V}_{\mathbf{i}_0},\mathbf{j}\in\mathscr{V}_{\mathbf{i}_0}} (\mathrm{B}(x, b_{\mathbf{n}}) \times \mathrm{B}(x, b_{\mathbf{n}})) \leq \mathrm{C}_3'' (\Gamma(b_{\mathbf{n}}))^{1+\varepsilon}, \text{ pour tout } x \in \mathscr{D}.$$

-H6 Il existe $s > 2$ et $\mathrm{C} > 0$ telles que :

(i) $\sup_{\mathbf{i}\in\mathscr{V}_{\mathbf{i}_0}} \mathbb{E}(|\mathrm{Y}_{\mathbf{i}}|^s \mid \mathrm{X}_{\mathbf{i}}) < \mathrm{C}$.

(ii) $\sup_{\mathbf{i},\mathbf{j},\mathbf{i}\in\mathscr{V}_{\mathbf{i}_0},\mathbf{j}\in\mathscr{V}_{\mathbf{i}_0}} \mathbb{E}\left(\left|\mathrm{Y}_{\mathbf{i}}\mathrm{Y}_{\mathbf{j}}\right| | \mathrm{X}_{\mathbf{i}}, \mathrm{X}_{\mathbf{j}}\right) < \mathrm{C}$ pour une constante $\mathrm{C} > 0$.

Introduisons les coefficients de mélange qui sont liés aux conditions soumises aux paramètres de lissage et au moment de la co-variable fonctionnelle :

$$\begin{aligned}
\theta_1 &= \frac{2s(\mathrm{N}-\theta)}{2\mathrm{N}s(\beta+2)+\theta(2-s)}, \quad \theta_2 = \frac{(\theta-2\mathrm{N})s}{2\mathrm{N}s(\beta+2)+\theta(2-s)}. \\
\theta_3 &= \frac{2(\mathrm{N}s+\theta)}{2\mathrm{N}s(\beta+2)+\theta(2-s)} \quad \theta_1^* = \frac{s(-\mathrm{N}-\theta)}{\mathrm{N}(2s\beta+2s\kappa+s+2)+\theta(2-s)}. \\
\theta_2^* &= \frac{s(\theta-\mathrm{N})}{\mathrm{N}(2s\beta+2s\kappa+s+2)+\theta(2-s)} \quad \theta_3^* = \frac{2(\mathrm{N}+\theta)}{\mathrm{N}(2s\beta+2s\kappa+s+2)+\theta(2-s)}.
\end{aligned}$$

Le théorème suivant établit la convergence uniforme presque sûre de l'estimateur de régression.

Théorème 4. *Supposons que les hypothèses **H1–H6** sont vérifiées.*

(i) Si la condition (3.2) *est vérifiée et*

$$\widehat{\mathbf{n}} \Gamma(b_{\mathbf{n}})^{\theta_1} \rho_{\mathbf{n}}^{\mathrm{N}\theta_1} \left(\log \widehat{\mathbf{n}}\right)^{\theta_2} u_{\mathbf{n}}^{\theta_3} \to \infty \, avec \, \theta > 2\mathrm{N}s(\beta+2)/(s-2). \tag{3.5}$$

(ii) ou bien si l'hypothèse (3.3) *est vérifiée et*

$$\widehat{\mathbf{n}} \Gamma(b_{\mathbf{n}})^{\theta_1^*} \rho_{\mathbf{n}}^{\mathrm{N}\theta_1^*} \left(\log \widehat{\mathbf{n}}\right)^{\theta_2^*} u_{\mathbf{n}}^{\theta_3^*} \to \infty \, avec \, \theta > \mathrm{N}(2s\beta+2s\kappa+s+2)/(s-2), \tag{3.6}$$

alors

$$\sup_{x\in\mathscr{D}} |r_{\mathbf{n}}(x) - r(x)| = \mathrm{O}\left(b_{\mathbf{n}} + \sqrt{\frac{\log \widehat{\mathbf{n}}}{\widehat{\mathbf{n}} \rho_{\mathbf{n}}^{\mathrm{N}} \Gamma(b_{\mathbf{n}})}}\right) \; p.c.$$

Rappelons que [101] a donné une borne limite de la convergence uniforme presque sûre d'un estimateur de régression sur un ensemble spécifique $\mathscr{C}$, soit : $\mathrm{O}\left(b_{\mathbf{n}}^{\star} + \sqrt{\frac{\log \widehat{\mathbf{n}}}{\Gamma(b_{\mathbf{n}}^{\star})\widehat{\mathbf{n}}}}\right)$ avec $\Gamma(b_{\mathbf{n}}^{\star}) = \sup_{x\in\mathscr{C}} \varphi_x(b_{\mathbf{n}})^{\star}$, sous l'hypothèse que le processus considéré soit strictement stationnaire.

Corollaire 2. *À partir de ces résultats, sous les conditions du théorème 4, on peut déduire la convergence presque sûre du prédicteur, soit.*

$$\left|\widehat{\mathrm{Y}}_{\mathbf{i}_0} - \mathrm{Y}_{\mathbf{i}_0}\right| \underset{\mathbf{n}\to\infty}{\longrightarrow} 0 \quad \textit{presque sûrement,} \tag{3.7}$$

où

$$\widehat{\mathrm{Y}}_{\mathbf{i}_0} = \frac{\sum_{\mathbf{i}\in\mathscr{O}_{\mathbf{n}}} \mathrm{Y}_{\mathbf{i}} \mathrm{K}_1\left(\frac{d(x,\mathrm{X}_{\mathbf{i}})}{b_{\mathbf{n}}}\right) \mathrm{K}_{2,\rho_{\mathbf{n}}}(\|\mathbf{i}_0 - \mathbf{i}\|)}{\sum_{\mathbf{i}\in\mathscr{O}_{\mathbf{n}}} \mathrm{K}_1\left(\frac{d(x,\mathrm{X}_{\mathbf{i}})}{b_{\mathbf{n}}}\right) \mathrm{K}_{2,\rho_{\mathbf{n}}}(\|\mathbf{i}_0 - \mathbf{i}\|)}, \tag{3.8}$$

définit le prédicteur de $\mathrm{Y}_{\mathbf{i}_0}$.

4.4 Application à la discrimination ou classification supervisée

L'analyse discriminante est une méthode qui a été largement étudiée et étendue à différents contextes depuis sa découverte par Fisher. Les divers domaines d'application dans lesquels nous retrouvons les problèmes de *discrimination* expliquent sans doute son succès. Elle se définit comme une méthode de *classification supervisée* qui consiste à classer des individus sur la base de variables explicatives et d'un échantillon d'apprentissage pour lequel à la fois ces variables et l'affectation aux classes sont connues. La tache principale en *discrimination* ou *classification supervisée* est, donc, d'affecter à Y une classe $m \in \{1,...,M\}$, quand la variable explicative X est observée. Nous notons par $\{1,...,M\}$ l'ensemble fini d'entiers qui définit le nombre de classes. Lorsque $Card(\{1,...,M\}) = 2$, le problème devient une classification binaire. Cela peut se produire lorsque nous voulons modéliser, par exemple, l'absence ou la présence d'une espèce de poissons, abondance ou non, surpêche ou non. Lorsque $Card(\{1,...,M\}) > 2$, nous avons un problème de classification catégorielle. Cela peut se produire par exemple lorsqu'on veut qualifier l'état d'un stock de poissons en trois états : équilibre, abondance et surabondance. Dans ce cadre, la décision des politiques publiques, pour une bonne gestion des pêcheries, devrait s'appuyer sur l'avis de scientifiques. Si la ressource est en état d'équilibre, elle aura besoin de repos pour son renouvellement avec l'entrée, dans le stock de nouveaux recrues. Dans ce cas, une licence de pêche ne peut être accordée pour ce stock de poissons. Si le stock est en abondance le surplus est autorisé à être prélevé par une licence de pêche qui limite la quantité à prélever suivant l'état d'abondance du stock en question (abondance et surabondance).
Notre but est de prédire cette *classe* Y à partir d'une nouvelle co-variable fonctionnelle X qui vient d'être observée dans un site $\mathbf{i}_0 \in \mathscr{I}_\mathbf{n}$ en s'appuyant sur l'échantillon $(X_\mathbf{i}, Y_\mathbf{i})_{\mathbf{i}\in\mathscr{O}_\mathbf{n}}$, $\mathscr{O}_\mathbf{n} \subset \mathscr{I}_\mathbf{n}\backslash\mathbf{i}_0$. Dans ce qui suit, nous décrivons la méthodologie proposée pour construire une *règle de classification fonctionnelle spatiale non-paramétrique*. Elle se construit à l'aide de la régression $r_\mathbf{n}$, définie dans l'équation (2.2).

Règle non-paramétrique de discrimination spatiale et fonctionnelle : Nous supposons que le couple $(X_{\mathbf{i}_0}, Y_{\mathbf{i}_0})$ a la même distribution q'un couple (X,Y), nous rappelons qu'en *classification supervisée*, nous construisons une fonction g définie dans $\mathscr{E}$ et qui prend ses valeurs dans $\{1,...,M\}$. Cette fonction définit la relation d'affectation $g(X) \in \{1,...,M\}$ à Y pour une observation X. La fonction g est appelée *classeur*.
On se trompe sur l'affectation si $g(X) \neq Y$, et la probabilité d'erreur pour un *classeur* g est donnée par

$$L(g) = P\{g(X) \neq Y\}.$$

Il est bien connu que le *classeur* de Bayes, défini par,

$$g^* = \underset{g:\mathscr{E}\to\{1,...,M\}}{\operatorname{argmin}} \; P\{g(X) \neq Y\},$$

est le meilleur *classeur* possible, en ce qui concerne la perte quadratique. La probabilité minimale d'erreur est appelée *Erreur de Bayes* et elle est notée par $L^* = L(g^*)$. Notons que g^* dépend de la distribution de (X,Y) qui est inconnue.
Un estimateur $g_\mathbf{n}$ de g est basée sur les observations $\{(X_\mathbf{i}, Y_\mathbf{i})_{\mathbf{i}\in\mathscr{O}_\mathbf{n}}\}$; Y est prédit par $g_\mathbf{n}\left(X_{\mathbf{i}_0}; (X_\mathbf{i}, Y_\mathbf{i})_{\mathbf{i}\in\mathscr{O}_\mathbf{n}}\right)$. La performance de $g_\mathbf{n}$ est mesurée par la probabilité conditionnelle d'erreur

$$L_\mathbf{n} = L\left(g_\mathbf{n}\right) = P\left\{g_\mathbf{n}\left(X; (X_\mathbf{i}, Y_\mathbf{i})_{\mathbf{i}\in\mathscr{O}_\mathbf{n}}\right) \neq Y\right\} \geq L^*.$$

La séquence $\{g_\mathbf{n},\ \mathbf{n} \in \mathbb{N}^{*N}\}$ est la *règle de discrimination*. Le *classeur* de Bayes g^* peut être approximé par la *règle de régression à noyau* $\{g_\mathbf{n},\ \mathbf{n} \in \mathbb{N}^{*N}\}$ basée sur l'estimateur $r_\mathbf{n}$ de régression (voir, l'équation (2.2)) elle est définie comme suit :

$$\sum_{\mathbf{i}\in\mathscr{O}_\mathbf{n}} W^\sharp_{\mathbf{n},\mathbf{i}_0}(x)\, \mathbb{I}_{\{Y_\mathbf{i} = g_\mathbf{n}(X_{\mathbf{i}_0})\}} = \max_{1\leq j\leq M} \sum_{\mathbf{i}\in\mathscr{O}_\mathbf{n}} W^\sharp_{\mathbf{n},\mathbf{i}_0}(x)\, \mathbb{I}_{\{Y_\mathbf{i} = j\}}, \tag{4.1}$$

où

$$W^\sharp_{\mathbf{n},\mathbf{i}_0}(x) = \frac{K_1\left(\frac{d(x,X_\mathbf{i})}{b_\mathbf{n}}\right) K_{2,\rho_\mathbf{n}}(\|\mathbf{i}_0 - \mathbf{i}\|)}{\sum_{\mathbf{i}\in\mathscr{O}_\mathbf{n}} K_1\left(\frac{d(x,X_\mathbf{i})}{b_\mathbf{n}}\right) K_{2,\rho_\mathbf{n}}(\|\mathbf{i}_0 - \mathbf{i}\|)}. \tag{4.2}$$

Nous rappelons que le problème de *discrimination* peut être considéré comme celui de la prédiction car il s'agit d'estimer l'espérance conditionnelle de la variable indicatrice (une par *classe*), voir [146]. Ainsi, tous les résultats asymptotiques (par exemple le théorème 4) établis dans le cas de la prédiction restent valides dans le contexte de la *discrimination*. Les travaux de [5; 118; 120; 146] montrent qu'une règle est bonne si elle est *consistante*, c'est-à-dire si, $L_\mathbf{n} \to L^*$ en probabilité (ou presque sûrement) quand $\mathbf{n} \to \infty$. Nous pouvons obtenir le théorème suivant qui établit la convergence presque complète /ou sûre de la règle proposée.

Théorème 5. *Sous les conditions du théorème 4*

$$\mathrm{L}_{\mathbf{n}} - \mathrm{L}^{*} \underset{\mathbf{n}\to\infty}{\longrightarrow} 0 \qquad p.c. \tag{4.3}$$

Après avoir vérifié le comportement asymptotique du prédicteur $\widehat{\mathrm{Y}}_{\mathbf{i}_0}$ et son extension à une *règle de classification supervisée* $\{g_{\mathbf{n}},\ \mathbf{n} \in \mathbb{N}^{*\mathrm{N}}\}$, nous étudions ses caractéristiques pratiques à travers une application sur des données simulées.

4.5 Simulations numériques

Dans cette section, nous étudions la performance du prédicteur proposé, à travers une simulation numérique, à fin de mettre en évidence son importance. Il est comparé à la méthode classique à noyau qui ne prend pas en compte la dépendance spatiale (voir [41; 105]). Décrivons d'abord la procédure de prédiction.

4.5.1 Procédure de prédiction

Le choix des paramètres de lissage (ou fenêtres) est une question cruciale en estimation non-paramétrique. Nous avons parlé de son importance dans le Chapitre 2. Nous proposons de choisir les paramètres optimaux de lissage (fenêtres optimales) par validation croisée.

Étape 1

Soient S_1 et S_2 deux sous ensembles de $\mathbb{R}_+^*$ sur lesquels appartiennent les paramètres de lissage $b_{\mathbf{n}}$ et $\rho_{\mathbf{n}}$ respectivement. $b_{\mathbf{n}}$ est la fenêtre du noyaux K_1 et $\rho_{\mathbf{n}}$ celle du noyau K_2.

Étape 2

Pour chaque $b_{\mathbf{n}} \in \mathrm{S}_1$, $\rho_{\mathbf{n}} \in \mathrm{S}_2$ et $\mathbf{i}_0 \in \mathscr{I}_{\mathbf{n}}$, nous calculons l'équation (2.2).

Étape 3

Puis, nous évaluons les paramètres optimaux $b_{\mathbf{n}}^{\sharp}$ et $\rho_{\mathbf{n}}^{\sharp}$ par la procédure de la validation croisée sur le produit cartésien $\mathrm{S}_1 \times \mathrm{S}_2$. Plus précisément, nous considérons le problème de minimisation, équation (5.1). Nous déterminons le couple $(b_{\mathbf{n}}, \rho_{\mathbf{n}})$ qui minimise l'erreur quadratique moyenne sur les $\hat{\mathbf{n}}$ sites (solution de l'équation (5.1)) :

$$\min_{(b_{\mathbf{n}},\rho_{\mathbf{n}})\in \mathrm{S}_1\times\mathrm{S}_2} \frac{1}{\hat{\mathbf{n}}} \sum_{\mathbf{i}_0 \in \mathscr{I}_{\mathbf{n}}} (\widehat{\mathrm{Y}}_{\mathbf{i}_0} - \mathrm{Y}_{\mathbf{i}_0})^2. \tag{5.1}$$

La solution de l'équation (5.1) sont notées par $(b_{\mathbf{n}}^{\sharp}, \rho_{\mathbf{n}}^{\sharp})$.
La même procédure est appliquée à l'équation (2.5) pour calculer $b_{\mathbf{n}}^{\star}$ en remplaçant $r_{\mathbf{n}}(.)$ par $r_{\mathbf{n}}^{cl}(.)$ dans l'équation (5.1) de minimisation sur S_1.

Étape 4

Pour chaque site $\mathbf{i}_0$, on prédit $\mathrm{Y}_{\mathbf{i}_0}$ comme suit :

— Nous calculons $\widehat{\mathrm{Y}}_{\mathbf{i}_0}^{\sharp}$ avec $b_{\mathbf{n}}^{\sharp}$ et $\rho_{\mathbf{n}}^{\sharp}$, voir l'équation (2.3).

— En suite nous calculons $\widehat{\mathrm{Y}}_{\mathbf{i}_0}^{*}$ avec $b_{\mathbf{n}}^{\star}$ via le prédicteur qui ne prend pas en compte la proximité spatiale, voir équation (2.4).

4.5.2 Études de simulation

Dans la suite, nous prenons $\mathrm{N} = 2$, pour illustrer le comportement du prédicteur, nous avons simulé des observations numériques, soient $(\mathrm{X}_{(i,j)}, \mathrm{Y}_{(i,j)}), 0 \le i, j \le 25$, basées sur la relation explicative et les modèles suivants :

$$\mathrm{Y}_{(i,j)} = r(\mathrm{X}_{(i,j)}) + \varepsilon_{(i,j)}. \tag{5.2}$$

$$= 4\mathrm{A}_{(i,j)}^2 + \varepsilon_{(i,j)}. \tag{5.3}$$

et pour $t \in [0,1]$, $\mathrm{X}_{(i,j)}(t)$ est :

$$\mathrm{X}_{i,j}(t) = \mathrm{A}_{i,j}^2 * (t - 0.5)^2, \tag{5.4}$$

où $A = (A_{i,j})$ et $\varepsilon = (\varepsilon_{i,j})$ sont des variables aléatoires qui seront spécifiées plus tard. Les courbes $X_{(i,j)}(t)$ sont représentées sur la figure **4.1** (voir le panel de gauche). Un exemple de fonction $r(.)$ peut être $r(X) = 2X''$ (où f'' est la dérivée seconde de la fonction f).

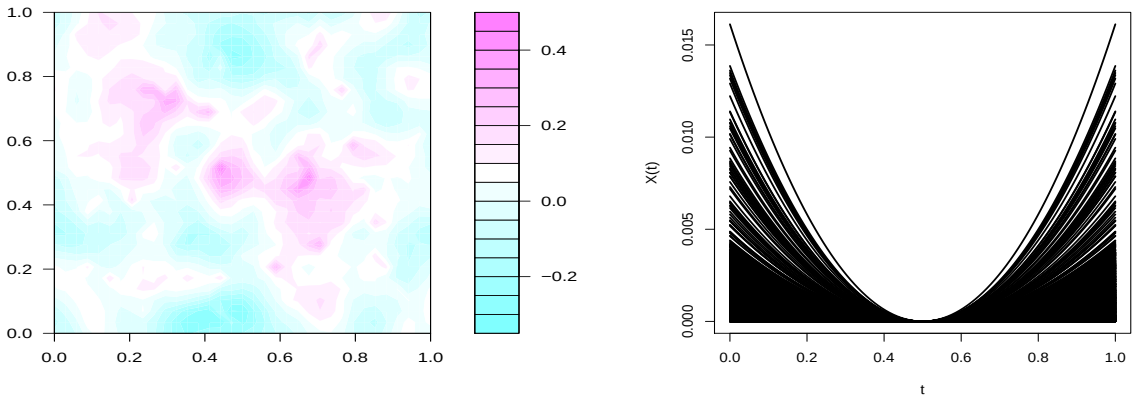

FIGURE 4.1 – Exemple ($a = 5$) de simulation du champs Y (panel de droite) ; et du modèle 5.4 (panel de gauche)

Le modèle 5.2 est simulé en considérant la dépendance spatiale. Par la suite, nous notons par $GRF(m, \sigma^2, s)$ un champ aléatoire gaussien stationnaire de moyenne m et de fonction de covariance définie par $C(h) = \sigma^2 \exp(-(\frac{\|h\|}{s}))$, $h \in \mathbb{R}^2$ et $s > 0$. Nous définissons $A = (A_{i,j})$ et $\varepsilon_{i,j}$, respectivement, comme suit.

$$A_{i,j} = D_{i,j}(\sin(2G_{i,j}) + 2\exp(-16G_{i,j}^2)) \ , \ \varepsilon_{i,j} = GRF(0,.1,5) \text{ avec } G_{i,j} = GRF(0,5,3).$$

$$\text{Nous posons } D_{\mathbf{i}} = \frac{1}{\widehat{\mathbf{n}}} \sum_{\mathbf{j}} \exp\left(-\frac{\left\|\mathbf{i}-\mathbf{j}\right\|}{a}\right). \text{ Soit } D_{(i,j)} = \frac{1}{\widehat{\mathbf{n}}} \sum_{m,t} \exp\left(-\frac{\left\|(i,j)-(m,t)\right\|}{a}\right).$$

La fonction D assure et contrôle la condition de dépendance spatiale locale mesurée par les coefficients de mélange. En effet, notre modèle vérifie la condition de coefficient mélange avec $\alpha(h) \to 0$ à une vitesse exponentielle. Notons que plus a est grand, plus faible est la dépendance spatiale. De plus, si $a \to \infty$, $D_{\mathbf{i}} \to 1$. Les simulations sont faites suivant différentes valeurs de a ($a = 5, 10, 20$, plus a est petite plus forte est la dépendance locale) et une taille de grille ($\widehat{\mathbf{n}} = 35 \times 30 = 1050$).
Les prédictions spatiales sont calculées en utilisant plusieurs noyaux K_1 (pour les observations) et K_2 (pour les sites), respectivement. Le choix de la semi-métrique $d(.,.)$ est important et dépend, en pratique des informations apportées par les données. Dans ce travail, nous considérons une semi-métrique entre les courbes (observations) basée sur les dérivées $q = 2$ (voir [146]).
Pour évaluer la performance de notre prédicteur $\widehat{Y}_{\mathbf{i}_0}$ basée sur le régresseur $r_{\mathbf{n}}^{\sharp}(.)$, on calcule sa précision par l'intermédiaire de la moyenne des erreurs quadratiques moyennes réalisées sur les prédictions $Y_{\mathbf{n}}^{\sharp}(.)$ de $Y_{\mathbf{i}_0}$. Nous comparons cette dernière à celle obtenue par le prédicteur qui ne prend pas en compte la dépendance spatiale, soit $Y_{\mathbf{n}}^{\star}(.)$. Le modèle numérique est simulé 50 fois à fin de tester le niveau de "significativité" de la comparaison. Rappelons que les prédictions $Y_{\mathbf{n}}^{\sharp}(.)$ et $Y_{\mathbf{n}}^{\star}(.)$ de $Y_{\mathbf{i}_0}$ sont, respectivement, définies par :

$$Y_{\mathbf{n}}^{\sharp}(X_{\mathbf{j}}) = \frac{\sum_{\substack{\mathbf{i}\in\mathscr{I}_{\mathbf{n}} \\ \mathbf{i}\neq\mathbf{j}}} Y_{\mathbf{i}} K_1\left(\frac{d(X_{\mathbf{j}},X_{\mathbf{i}})}{b_{\mathbf{n}}^{\sharp}}\right) K_{2,\rho_{\mathbf{n}}^{\sharp}}(\|\mathbf{j}-\mathbf{i}\|)}{\sum_{\substack{\mathbf{i}\in\mathscr{I}_{\mathbf{n}} \\ \mathbf{i}\neq\mathbf{j}}} K_1\left(\frac{d(X_{\mathbf{j}},X_{\mathbf{i}})}{b_{\mathbf{n}}^{\sharp}}\right) K_{2,\rho_{\mathbf{n}}^{\sharp}}(\|\mathbf{j}-\mathbf{i}\|)} \text{ et } Y_{\mathbf{n}}^{\star}(X_{\mathbf{j}}) = \frac{\sum_{\substack{\mathbf{i}\in\mathscr{I}_{\mathbf{n}} \\ \mathbf{i}\neq\mathbf{j}}} Y_{\mathbf{i}} K_1\left(\frac{d(X_{\mathbf{j}},X_{\mathbf{i}})}{b_{\mathbf{n}}^{\star}}\right)}{\sum_{\substack{\mathbf{i}\in\mathscr{I}_{\mathbf{n}} \\ \mathbf{i}\neq\mathbf{j}}} K_1\left(\frac{d(X_{\mathbf{j}},X_{\mathbf{i}})}{b_{\mathbf{n}}^{\star}}\right)}. \tag{5.5}$$

Pour chaque réplication k, nous calculons l'erreur quadratique moyenne sur les $\widehat{\mathbf{n}}$ sites. Les fenêtres utilisées à chaque réplication sont celles obtenues à l'aide de la procédure précédente 4.5.1. Pour la réplication k^{th}, nous définissons l'erreur quadratique moyenne ($MSE_{(k)}^{(+)}$) par :

$$MSE_{(k)}^{(+)} = \frac{1}{\widehat{\mathbf{n}}} \sum_{\mathbf{j}\in\mathscr{I}_{\mathbf{n}}} (Y_{\mathbf{n},opt}^{+}(X_{\mathbf{j}}) - Y_{\mathbf{j}})^2 \text{ avec } Y_{\mathbf{n},opt}^{+} = Y_{\mathbf{n}}^{\sharp} \text{ ou } Y_{\mathbf{n}}^{\star}. \tag{5.6}$$

Les résultats obtenus sont résumés dans les tableaux 4.1, 4.2 et 4.3. Ces derniers donnent la moyenne des erreurs quadratiques moyennes (AMSE), la variance, les coefficients de détermination moyens $R^{2\sharp}$ et $R^{2\star}$,

respectivement. La dernière colonne donne la $p-value$ d'un t-test effectué pour déterminer si AMSE$^\sharp$ est significativement inférieur à AMSE* (l'hypothèse alternative est alors H1 : AMSE$^\sharp$ < AMSE*). Nous notons que les valeurs des $p-values$ sont très proches de 0. Elles sont remplacées par le symbole (*). Plus on a des (*) plus la $p-value$ est proche de 0. La qualité de l'estimation est mesurée par le coefficient de détermination. Nous rappelons que plus la valeur de R^2 est proche de 1 plus est fiable le modèle.
Nous remarquons, globalement, que l'estimateur $r^\sharp_{\mathbf{n}}(.)$ donne de meilleures précisions. En effet, tous les résultats des tableaux 4.1, 4.2 et 4.3 montrent que AMSE$^\sharp$ est significativement inférieure à AMSE*. D'autre part, nous notons que la variance de AMSE$^\sharp$ est plus petit que celle de AMSE*, pour tous les cas considérés. De plus, même si la dépendance spatiale devient faible, AMSE* reste plus élevée et relativement constante, alors que AMSE$^\sharp$ varie constamment. Cela montre que la méthode basée sur $r^\sharp_{\mathbf{n}}(.)$ est plus adaptée à la structure spatiale des données et à une stationnarité locale. Enfin, nous notons que $R^{2\sharp}$, est supérieur à $R^{2\star}$ pour tous les cas considérés, mais la différence entre eux diminue à mesure que la valeur de a augmente (moins de dépendance spatiale).

TABLEAU 4.1 – Résultats de simulations basés sur différents noyaux, avec $a = 5$

K_1	K_2	$AMSE^\sharp$	$Var(AMSE^\sharp)$	$AMSE^\star$	$Var(AMSE^\star)$	$R^{2\sharp}$	$R^{2\star}$	P-value
Trianglaire	**Triangular**	0,00018	$1,1\,10^{-09}$	0,0080	$2,2\,10^{-06}$	0,9998	0,9905	****
	Biweight	0,00017	$1,3\,10^{-09}$	0,0081	$2,9\,10^{-06}$	0,9998	0,9902	****
	Triweight	0,00011	$3,7\,10^{-10}$	0,0079	$2,5\,10^{-06}$	0,9999	0,9903	****
	Parzen	0,00006	$1,0\,10^{-10}$	0,0081	$3,2\,10^{-06}$	0,9999	0,9906	****
	Epanechnikov	0,00032	$2,9\,10^{-09}$	0,0077	$1,8\,10^{-06}$	0,9996	0,9907	****
	Gauss	0,00195	$9,8\,10^{-09}$	0,0079	$2,2\,10^{-06}$	0,9976	0,9904	****
Biweight	**Triangular**	0,00037	$4\,10^{-09}$	0,0080	$2,4\,10^{-06}$	0,9996	0,9907	****
	Biweight	0,00020	$1,2\,10^{-09}$	0,0083	$2,3\,10^{-06}$	0,9998	0,9902	****
	Triweight	0,00019	$9,1\,10^{-10}$	0,0078	$1,8\,10^{-06}$	0,9998	0,9910	****
	Parzen	0,00012	$4,8\,10^{-10}$	0,0081	$2,4\,10^{-06}$	0,9999	0,9903	****
	Epanechnikov	0,00218	$1,6\,10^{-07}$	0,0083	$2,6\,10^{-06}$	0,9975	0,9904	****
	Gauss	0,00007	$2,0\,10^{-10}$	0,0084	$3,5\,10^{-06}$	0,9999	0,9898	****
Triweight	**Triangular**	0,00032	$3,6\,10^{-09}$	0,0081	$2,4\,10^{-06}$	0,9996	0,9903	****
	Biweight	0,00018	$9,8\,10^{-10}$	0,0077	$2\,10^{-06}$	0,9998	0,9911	****
	Triweight	0,00017	$9,7\,10^{-10}$	0,0080	$2,9\,10^{-06}$	0,9998	0,9908	****
	Parzen	0,00010	$4\,10^{-10}$	0,0082	$2,2\,10^{-06}$	0,9999	0,9903	****
	Epanechnikov	0,00183	$9,6\,10^{-09}$	0,0079	$2,1\,10^{-06}$	0,9979	0,9910	****
	Gauss	0,00006	$2,1\,10^{-10}$	0,0081	$2\,10^{-06}$	0,9999	0,9902	****
Parzen	**Triangular**	0,00027	$3,3\,10^{-09}$	0,0077	$2\,10^{-06}$	0,9997	0,9908	****
	Biweight	0,00015	$1,3\,10^{-09}$	0,0080	$5\,10^{-06}$	0,9998	0,9908	****
	Triweight	0,00014	$5,5\,10^{-10}$	0,0078	$3\,10^{-06}$	0,9998	0,9909	****
	Parzen	0,00009	$3\,10^{-10}$	0,0083	$2\,10^{-06}$	0,9999	0,9904	****
	Epanechnikov	0,00162	$1,1\,10^{-07}$	0,0078	$2\,10^{-06}$	0,9980	0,9903	****
	Gauss	0,00005	$1\,10^{-10}$	0,0081	$2,9\,10^{-06}$	0,9999	0,9905	****
Epanechnikov	**Triangular**	0,00044	$6,3\,10^{-09}$	0,0083	$3,2\,10^{-06}$	0,9995	0,9895	****
	Biweight	0,00025	$2,4\,10^{-09}$	0,0086	$3,5\,10^{-06}$	0,9997	0,9900	****
	Triweight	0,00024	$1,8\,10^{-09}$	0,0084	$2,7\,10^{-06}$	0,9997	0,9901	****
	Parzen	0,00015	$6\,10^{-10}$	0,0083	$2\,10^{-06}$	0,9998	0,9903	****
	Epanechnikove	0,00247	$1\,10^{-07}$	0,0083	$1\,10^{-06}$	0,9971	0,9903	****
	Gauss	0,00008	$2,8\,10^{-10}$	0,0083	$2\,10^{-06}$	0,999	0,9905	****
Gauss	**Triangular**	0,00067	$6,6\,10^{-09}$	0,0091	$2\,10^{-06}$	0,9992	0,9887	****
	Biweight	0,00041	$3,8\,10^{-09}$	0,0086	$3\,10^{-06}$	0,9995	0,9896	****
	Triweight	0,00040	$3,5\,10^{-09}$	0,0090	$3\,10^{-06}$	0,9995	0,9892	****
	Parzen	0,00026	$1,1\,10^{-09}$	0,0089	$3\,10^{-06}$	0,9997	0,9896	****
	Epanechnikove	0,00319	$4\,10^{-07}$	0,0085	$3\,10^{-06}$	0,996	0,990	****
	Gauss	0,00015	$5,9\,10^{-10}$	0,0088	$2,8\,10^{-06}$	0,9998	0,9895	****

TABLEAU 4.2 – Résultats de simulations basés sur différents noyaux, avec $a = 10$

K_1	K_2	$AMSE^\sharp$	$Var(AMSE^\sharp)$	$AMSE^\star$	$Var(AMSE^\star)$	$R^{2\sharp}$	$R^{2\star}$	P-value
Triangular	**Triangular**	0,0005	$6,4\,10^{-09}$	0,0085	$2,2\,10^{-06}$	0,990	0,840	****
	Biweight	0,0005	$3,8\,10^{-09}$	0,0087	$1,5\,10^{-06}$	0,991	0,837	****
	Triweight	0,0003	$2,0\,10^{-09}$	0,0085	$1,3\,10^{-06}$	0,994	0,845	****
	Epanechnikove	0,0009	$1,5\,10^{-08}$	0,0085	$1,5\,10^{-06}$	0,983	0,839	****
	Gaussian	0,0035	$2,6\,10^{-07}$	0,0086	$2,2\,10^{-06}$	0,936	0,842	****
	Parzen	0,0002	$7,0\,10^{-10}$	0,0085	$1,5\,10^{-06}$	0,997	0,839	****
Biweight	**Triangular**	0,0006	$9,0\,10^{-09}$	0,0086	$2,9\,10^{-06}$	0,989	0,841	****
	Parzen	0,0002	$9,9\,10^{-10}$	0,0088	$1,9\,10^{-06}$	0,996	0,841	****
	Biweight	0,0006	$6,5\,10^{-09}$	0,0089	$1,7\,10^{-06}$	0,989	0,829	****
	Triweight	0,0004	$2,1\,10^{-09}$	0,0088	$1,7\,10^{-06}$	0,993	0,836	****
	Epanechnikove	0,0010	$1,6\,10^{-08}$	0,0086	$1,5\,10^{-06}$	0,981	0,838	****
	Gaussian	0,0037	$2,3\,10^{-07}$	0,0088	$1,8\,10^{-06}$	0,932	0,838	****
Triweight	**Triangular**	0,0005	$7,1\,10^{-09}$	0,0085	$2,8\,10^{-06}$	0,990	0,842	****
	Parzen	0,0002	$5,0\,10^{-10}$	0,0085	$1,5\,10^{-06}$	0,997	0,839	****
	Biweight	0,0005	$5,4\,10^{-09}$	0,0088	$1,6\,10^{-06}$	0,991	0,831	****
	Parzen	0,0002	$5,0\,10^{-10}$	0,0085	$1,5\,10^{-06}$	0,997	0,839	****
	Gaussian	0,0034	$4,4\,10^{-07}$	0,0085	$2,8\,10^{-06}$	0,937	0,842	****
	Triweight	0,0003	$1,6\,10^{-09}$	0,0085	$1,5\,10^{-06}$	0,994	0,839	****
Epanechnikov	**Triangular**	0,0007	$1,2\,10^{-08}$	0,0087	$2,9\,10^{-06}$	0,987	0,838	****
	Biweight	0,0007	$8,5\,10^{-09}$	0,0091	$1,7\,10^{-06}$	0,987	0,827	****
	Triweight	0,0004	$3,5\,10^{-09}$	0,0089	$1,7\,10^{-06}$	0,992	0,833	****
	Parzen	0,0002	$1,4\,10^{-09}$	0,0089	$1,9\,10^{-06}$	0,996	0,838	****
	Epanechnikove	0,0012	$2,2\,10^{-08}$	0,0087	$1,6\,10^{-06}$	0,978	0,835	****
	Gaussian	0,0040	$2,9\,10^{-07}$	0,0089	$1,8\,10^{-06}$	0,925	0,835	****
Parzen	**Triangular**	0,0004	$5,1\,10^{-09}$	0,0083	$2,7\,10^{-06}$	0,992	0,845	****
	Biweight	0,0004	$4,1\,10^{-09}$	0,0087	$1,6\,10^{-06}$	0,992	0,834	****
	Triweight	0,0003	$1,1\,10^{-09}$	0,0085	$1,6\,10^{-06}$	0,995	0,841	****
	Parzen	0,0001	$5,7\,10^{-10}$	0,0085	$1,8\,10^{-06}$	0,997	0,845	****
	Epanechnikov	0,0008	$1,0\,10^{-08}$	0,0083	$1,4\,10^{-06}$	0,986	0,842	****
	Gaussian	0,0030	$1,6\,10^{-07}$	0,0085	$1,7\,10^{-06}$	0,944	0,842	****
Gaussian	**Triangular**	0,0010	$2,4\,10^{-08}$	0,0090	$3,1\,10^{-06}$	0,981	0,833	****
	Biweight	0,0009	$1,4\,10^{-08}$	0,0093	$1,8\,10^{-06}$	0,982	0,822	****
	Triweight	0,0006	$5,7\,10^{-09}$	0,0092	$1,8\,10^{-06}$	0,988	0,829	****
	Parzen	0,0004	$2,6\,10^{-09}$	0,0091	$2,0\,10^{-06}$	0,993	0,834	****
	Epanechnikov	0,0016	$3,9\,10^{-08}$	0,0090	$1,7\,10^{-06}$	0,970	0,831	****
	Gaussian	0,0048	$3,8\,10^{-07}$	0,0091	$1,9\,10^{-06}$	0,910	0,831	****

TABLEAU 4.3 – Résultats de simulations basés sur différents noyaux, avec $a = 20$

K_1	K_2	$AMSE^\sharp$	$Var(AMSE^\sharp)$	$AMSE^\star$	$Var(AMSE^\star)$	$R^{2\sharp}$	$R^{2\star}$	P-value
Epanechnikov	**Triangular**	0,0007	$1,39\,10^{-08}$	0,0087	$3,00\,10^{-06}$	0,998	0,981	****
	Biweight	0,0007	$8,58\,10^{-09}$	0,0091	$1,69\,10^{-06}$	0,999	0,980	****
	Triweight	0,0004	$2,71\,10^{-09}$	0,0089	$1,80\,10^{-06}$	0,999	0,980	****
	Parzen	0,0002	$1,36\,10^{-09}$	0,0089	$2,01\,10^{-06}$	0,997	0,981	****
	Epanechnikov	0,0012	$2,05\,10^{-08}$	0,0087	$1,64\,10^{-06}$	0,991	0,981	****
	Gauss	0,0041	$2,98\,10^{-07}$	0,0089	$1,94\,10^{-06}$	0,997	0,981	****
Triangular	**Triangular**	0,0005	$7,90\,10^{-09}$	0,0085	$2,86\,10^{-06}$	0,991	0,981	****
	Biweight	0,0005	$5,31\,10^{-09}$	0,0089	$1,62\,10^{-06}$	0,998	0,980	****
	Triweight	0,0003	$1,54\,10^{-09}$	0,0087	$1,72\,10^{-06}$	0,998	0,981	****
	Parzen	0,0002	$7,90\,10^{-10}$	0,0085	$2,86\,10^{-06}$	0,999	0,981	****
	Gauss	0,0036	$2,26\,10^{-07}$	0,0085	$1,55\,10^{-06}$	0,992	0,981	****
	Epanechnikov	0,0009	$1,27\,10^{-08}$	0,0085	$1,55\,10^{-06}$	0,998	0,981	****
Biweight	**Triangular**	0,0006	$9,71\,10^{-09}$	0,0086	$2,90\,10^{-06}$	0,999	0,981	****
	Biweight	0,0006	$6,45\,10^{-09}$	0,0089	$1,64\,10^{-06}$	0,999	0,981	****
	Triweight	0,0004	$1,90\,10^{-09}$	0,0088	$1,74\,10^{-06}$	0,997	0,980	****
	Parzen	0,0002	$9,46\,10^{-10}$	0,0088	$1,94\,10^{-06}$	0,992	0,981	****
	Epanechnikov	0,0010	$1,51\,10^{-08}$	0,0086	$1,57\,10^{-06}$	0,998	0,981	****
	Gauss	0,0037	$2,48\,10^{-07}$	0,0088	$1,89\,10^{-06}$	0,997	0,981	****
Triweight	**Triangular**	0,0005	$7,94\,10^{-09}$	0,0085	$2,24\,10^{-06}$	0,992	0,981	****
	Biweight	0,0005	$3,62\,10^{-09}$	0,0088	$1,52\,10^{-06}$	0,998	0,981	****
	Triweight	0,0003	$2,02\,10^{-09}$	0,0086	$1,39\,10^{-06}$	0,999	0,981	****
	Parzen	0,0002	$6,66\,10^{-10}$	0,0082	$1,62\,10^{-06}$	0,999	0,982	****
	Epanechnikov	0,0009	$1,55\,10^{-08}$	0,0085	$1,56\,10^{-06}$	0,998	0,982	****
	Gauss	0,0034	$2,54\,10^{-07}$	0,0086	$2,29\,10^{-06}$	0,992	0,981	****
Parzen	**Triangular**	0,0004	$4,24\,10^{-09}$	0,0085	$1,81\,10^{-06}$	0,998	0,981	****
	Biweight	0,0004	$4,00\,10^{-09}$	0,0084	$1,50\,10^{-06}$	0,999	0,982	****
	Triweight	0,0003	$1,95\,10^{-09}$	0,0084	$2,17\,10^{-06}$	0,999	0,982	****
	Parzen	0,0001	$4,89\,10^{-10}$	0,0084	$1,35\,10^{-06}$	0,999	0,982	****
	Epanechnikov	0,0008	$9,41\,10^{-09}$	0,0084	$1,46\,10^{-06}$	0,998	0,981	****
	Gauss	0,0032	$2,21\,10^{-07}$	0,0086	$1,48\,10^{-06}$	0,993	0,981	****
Gauss	**Triangular**	0,0011	$1,50\,10^{-08}$	0,0090	$1,69\,10^{-06}$	0,997	0,980	****
	Biweight	0,0010	$1,96\,10^{-08}$	0,0090	$3,14\,10^{-06}$	0,997	0,980	****
	Triweight	0,0007	$6,94\,10^{-09}$	0,0093	$1,79\,10^{-06}$	0,998	0,989	****
	Parzen	0,0004	$1,34\,10^{-09}$	0,0092	$2,05\,10^{-06}$	0,999	0,980	****
	Epanechnikov	0,0016	$3,57\,10^{-08}$	0,0090	$1,69\,10^{-06}$	0,996	0,980	****
	Gauss	0,0050	$4,67\,10^{-07}$	0,0092	$1,86\,10^{-06}$	0,989	0,980	****

4.6 Conclusion

Dans ce Chapitre, nous avons proposé un nouveau prédicteur spatial non-paramétrique dans un cadre fonctionnel, basé sur un processus spatial localement stationnaire. Le préditeur proposé donne une nouvelle méthode de *classification supervisée* lorsque la variable de réponse Y appartient à un ensemble discret. L'originalité de la méthode proposée est de prendre en compte à la fois la distance entre les sites et celle entre les observations fonctionnelles. Dans le cadre de la prédiction, nous donnons une extension du prédicteur spatial d'un processus multivarié localement stationnaire établi par [98]. L'approche de classification qui en découle est une extension aux *règles de discriminations* établies par [5; 146; 373]. Nous avons établi les propriétés asymptotiques du prédicteur par le biais de la convergence presque complète. Les résultats numériques montrent que la méthode de prédiction proposée donne des prédictions plus précises

que celles du prédicteur classique à noyau de [146].

CHAPITRE 5

APPLICATIONS À LA MODÉLISATION DE RESSOURCES HALIEUTIQUES DU SÉNÉGAL

Résumé

La caractérisation de la distribution spatiale des poissons est au cœur d'une dynamique de recherche autour de l'évaluation des stocks halieutiques. La science halieutique prend part à la mise en œuvre des méthodes mathématiques dans le but d'identifier la meilleure manière d'analyser et de prédire la distribution de la ressource côtière, en particulier. Elle se base, généralement sur des méthodes de type krigeage, co-kirigeage et la modélisation SDM, qui reposent, en grande partie, sur des modèles statistiques paramètriques et parfois sémi-paramétriques et des hypothèses de normalité sur les observations. Cela peut parfois être contraignant et exclut une large classe de processus qui ne vérifient pas ces hypothèses. Dans la contribution de ce Chapitre, nous proposons des alternatives à ces méthodes paramétriques. Nous appliquons ainsi les méthodes de régression développées dans les Chapitres 3 et 4, respectivement. De manière plus précise, nous nous intéressons à la prédiction non-paramétrique de la distribution spatiale des poissons démersaux côtiers au large du Sénégal ainsi que leurs quantités de biomasse. Les prédicteurs proposés tiennent compte de la distribution spatiale des poissons et des conditions environnementales telles que la salinité et la température.

5.1 Introduction

La description à petite échelle de l'habitat démersal est essentielle pour mieux cerner le fonctionnement de l'écosystème qui abrite les ressources halieutiques. Comprendre le fonctionnement de ce type de système complexe présente un intérêt particulier pour une gestion durable de la ressource. L'analyse écosystémique des pêcheries a montré comment l'environnement et les paramètres écologiques et physiques affectent la variabilité des stocks de poissons, en étudiant les informations (données) recueillies, à différents endroits, pendant une longue période. Les travaux de [236; 278; 372] ont montré que les conditions du milieu et les paramètres physiques gouvernent la distribution spatio-temporelle des espèces (voir également la description faite, du comportement de la ressource démersale côtière, sur le plateau continental sénégalais dans les sections 1.1.2 et 1.1.1 du Chapitre1). Cependant les caractéristiques physiques, les conditions environnementales et la dynamique structurelle de l'espace ne sont pas prises en compte, convenablement, dans les modélisations. Dans les travaux de [63; 170], les auteurs ont souligné l'importance de prendre en compte la dynamique spatiale dans l'évaluation des stocks. Dans ce contexte, pour une meilleure gestion des ressources halieutiques, il s'avère utile d'étudier le lien explicatif entre les variables environnementales, la structure spatiale et la variabilité de la densité/biomasse de la ressource dans une approche écosystémique prédictive. Cette étude doit se faire à l'aide des données massives et complexes (sur le mieux et la ressource) qui présentent une composante dynamique spatiale et/ou temporelle. Les données spatiales/spatio-temporelles abondent dans de nombreux domaines, en particulier dans la description des systèmes océanologiques où l'étude des relations explicatives entre les variables constituées de vecteurs spatio-temporels complexes et de grande dimension (données fonctionnelles) est envisagée pour comprendre le fonctionnement de l'écosystème marin. Ces données collectées dans des structures très denses sont souvent basées sur des informations observées continûment dans l'espace et/ou le temps.

La complexité de ces données (de hautes résolutions informatives et massives) suscite l'intérêt des scientifiques. Ces derniers ont essayé d'élaborer de nouvelles méthodes mathématiques qui peuvent prendre en compte toutes les informations apportées par les données et permettant ainsi de fournir des réponses adéquates aux différentes problématiques. Ces théories mathématiques aboutissent à la naissance de nouvelles branches en modélisation statistique et définit ainsi de nouveaux axes de recherches.
Les techniques statistiques multivariées paramétriques telles que le Krigeage et co-Krigeage [91; 300], ont été souvent appliquées dans le but d'évaluer l'abondance des stocks (de poissons). Ces modèles reposent sur la linéarité entre la variable de réponse et les prédicteurs et nécessitent des hypothèses comme la distribution gaussienne et la covariance paramétrique, entre autres. Des méthodes telles que la fonction *K-Ripley* [162; 191; 222; 297] et SDM basé conventionnellement sur les modèles linéaires généralisés et les modèles additifs généralisés [55; 160; 236; 282; 372] ont également été utilisés en biologie marine, pour prédire la quantité de biomasse des poissons et analyser leur distribution. Notons toutefois que, lorsque l'échantillon d'intérêt est un ensemble de données massives, le problème du fléau de la dimension se pose et les techniques de réduction de la dimension sont des approches courantes. En général, pour résoudre le problème de dimension, plusieurs méthodes de régression multivariées utilisant un grand nombre de prédicteurs, considèrent la dimension comme un paramètre de nuisance. De plus, ces méthodes de régression ne capturent pas toutes les informations provenant du processus qui génère les données.
Une alternative aux modèles multivariés paramétriques, quand on modélise les données spatiales massives observées dans l'espace et/ou le temps, peut être l'analyse des données fonctionnelles (ADF) spatiales, un domaine de recherche récent combinant les branches bien développées de la statistique fonctionnelle et celle spatiale. Elle a montré son adaptabilité dans l'analyse des données spatiales et fonctionnelles.
L'ADF transforme des objets de grande dimension en des données fonctionnelles, c'est-à-dire, elle transforme des objets, tels que des courbes, formes, images ou objet plus complexe, considérés comme la réalisation d'un processus stochastique, en des données issues d'espaces de fonctions de dimension éventuellement infinie. Un échantillon de données multidimensionnelles observées sur plusieurs unités spatiales dépendantes peut être analysé par des méthodes de l'ADF, en le considérant comme une réalisation d'un processus stochastique spatial à valeur dans un espace fonctionnel (semi-métrique, de Hilbert, Banach,...).
Dans ce Chapitre, nous appliquons les méthodes mathématiques développées dans les Chapitres 3 et 4, respectivement. Elles permettent de prédire la distribution spatiale des poissons démersaux. Elles fournissent des outils de prédiction des quantités de biomasse à des endroits donnés où des conditions environnementales sont connues. Les approches tiennent compte des paramètres de l'environnement marin, dans un cadre spatial et fonctionnel.
Le reste du Chapitre est organisé comme suit. Dans la section 5.2, nous introduisons la description des données supports d'applications des méthodes développées dans ce travail de thèse. Dans les sections 5.3 et 5.4, nous appliquons les *règles de classification supervisées* développées, respectivement, dans les Chapitres 3 et 4. Nous donnons une prédiction de la distribution spatiale des poissons démersaux au large du Sénégal. La section 5.5 est réservée à la prédiction de la quantité de biomasse en utilisant l'approche développée dans le Chapitre 4. La section 5.6 est consacrée à une conclusion des résultats des applications.

5.2 Description des données

Les données proviennent des campagnes scientifiques menées par le CRODT durant les saisons froides et chaudes au large des côtes sénégalaises de 2001 à 2016. Les stocks ciblés sont les démersaux côtiers pêchés dans les fonds de 10 à 200 m. Les sites de pêche ou stations sont échantillonnés suivant une double stratification par zone (Nord Centre et Sud) et par profondeur ou tranche bathymétrique ($0-50$ m, $50-100$ m, $100-150$ m et $150-200$ m). Dans la pratique les sites sont visités suivant un plan d'échantillonnage fixe. Cela permet de supposer que les données sont collectées suivant un plan déterministe. Cette hypothèse est en phase avec le cadre théorique du présent travail. Les données sont composées de 496 observations (stations/sites) décrites par l'identification de l'observation (numéro de campagne et station ou de chalutage), la période (date, année, saison, la durée du chalutage), le lieu ou les coordonnées géographiques (latitude, longitude, zone, profondeur, strate bathymétrique), les paramètres biologiques (capture par unité d'effort, richesse ou nombre d'espèces capturées sur un site) et les caractéristiques environnementales (température et salinité) à différentes profondeurs.
Un aperçu sur la variation spatiale de la température et la salinité dans la zone d'étude, est donné par la figure 5.1. Le panel (a) montre la salinité de fond (SBS), c'est la valeur de la salinité à la profondeur $t = 200$ m. Le panel (b) montre la salinité de surface (SSS), c'est la valeur de la salinité à $t = 0$ m. Le panel (c) montre la température du fond (SBT), la température en profondeur $t = 200$ m. Le panel (d) montre la température de surface (SST), à savoir à $t = 0$ m. Nous constatons qu'il existe globalement une hétérogénéité spatiale dans le fond et à la surface, de la température et de la salinité.

Les histogrammes des variables quantitatives biologiques et environnementales (température, salinité, capture, richesse) sont donnés dans la figure 5.3. Ils donnent un aperçu sur leurs distributions. Ils montrent une variabilité importante des variables SBS, SSS, SBT et SST. Ce qui nécessite une transformation des observations. Nous l'avons faite en utilisant le logarithme (voir figure 5.4).
Sur certains sites d'observation, nous avons des mesures de température et de salinité par variation d'un pas de 1 m, à partir de la surface jusqu'au fond de 200 m. Ces mesures de paramètres environnementaux (température et salinité) sont modélisées comme des observations d'une variable fonctionnelle $X(t), t \in [0,200]$. Un pré-traitement appliqué à ces observations permet de transformer les mesures discrétisées de température et de salinité en des variables fonctionnelles. Nous avons utilisé le *package rainbow* de [318] pour visualiser les données, détecter éventuellement des valeurs aberrantes et transformer les données brutes en fonctionnelles. Nous enlevons les bruits et données aberrantes. Les figures 5.5 et 5.6 montrent, respectivement, les courbes de température et de salinité suivant la profondeur (données fonctionnelles). Les panels de droite (des figures 5.5 et 5.6) représentent les données brutes. Ils montrent que les données ont été observées avec un certain bruit. Ce dernier peut être lié à la précision de l'instrument de mesure ou une autre perturbation liée aux conditions de chalutage. Après avoir enlevé les bruits et éventuelles valeurs aberrantes, nous représentons ces données reconstruites sur des bases B-Spline, voir la figure 5.7. Les panels supérieurs (de la figure 5.7) donnent les courbes lissées (données fonctionnelles). Les panels inférieurs (de la figure 5.7) montrent les moyennes des courbes. Le panel (a) (respectivement le panel (b)) montre une hétérogénéité de la salinité (respectivement de la température) entre les sites. Le panel (c) (respectivement le panel (d)) montre des variations importantes de la salinité (respectivement de la température) suivant la profondeur (bathymétrie).

5.2.1 Répartition spatiale des espèces démersales

Notons que nous avons enregistré la présence (**label 1**)/ ou absence (**label 0**) de 7 espèces qui présentent des intérêts économiques et/ou écologiques, soient *Dentex angolensis*, *Epinephelus aeneus*, *Octapus Vulgarus*, *Pagellus bellottii*, *Pagrus caeruleosticus*, *Galeoides decadactylus*, *Pseudupeneus prayensis* (voir la figure 5.9) sur les 495 stations d'observation (la figure 5.8 donne la répartition spatiale des stations de pêche). Nous décrirons leurs répartitions spatiales en fonction des conditions environnementales.
Le panel (a) de la figure 5.9 montre la répartition de *Dentex angolensis* sur presque toute la zone d'étude avec une forte présence dans le centre et le nord.
Le panel (d) de la figure 5.9 montre la présence de *Pagellus bellottii* sur presque toute la zone d'étude avec une concentration importante au centre.
Les panels à gauche des figures 5.10 et 5.11 montrent que les courbes bleues sont au dessus de celles rouges. Les panels à droite de ces précédentes figures montrent que les courbes bleues sont en dessous. Cela pourrait nous amener à supposer que nous pouvons espérer trouver plus de *Dentex angolensis* et de *Pagellus bellottii* dans les zones salées et de températures relativement basses.
Le panel (c) de la figure 5.9 montre une répartition de *Octopus vulgaris* relativement homogène sur presque toute la zone.
La figure 5.12 montre que les courbes bleues (présence de l'espèce) dominent (sont au dessus) celles rouges (absence de l'espèce) . Cela laisse supposer qu'on peut trouver plus de *Octapus Vulgarus* dans les zones salées et relativement tempérées.
Le panel (e) de la figure 5.9 montre que la présence de *Pagrus caeruleosticus* est relativement abondante dans le centre et au sud des côtes sénégalaises.
Le panel à gauche de la figure 5.14 montre que la courbe bleue est au dessus de celle en rouge à partir de la surface jusqu'à la profondeur de 125 m. La courbe rouge est au dessus de celle en bleue à partir de la profondeur 125 m jusqu'au fond de 200 m. Le panel à droite de la figure 5.14 montre que la courbe rouge domine celle en bleue à partir de la surface jusqu'au profondeur de 125 m. La courbe rouge est en dessous de celle bleue à partir de la profondeur 125 m jusqu'au fond de 200 m. Cela laisse supposer que la probabilité de trouver *Pagrus caeruleosticus* dans les zones salées et de température relativement faible est forte sur la tranche bathymétrique $0-125$ m. Sur la tranche $125-200$ m, la probabilité de trouver *Pagrus caeruleosticus* est forte dans les zones moins salées et relativement tempérées.
Le panel (b) de la figure 5.9 montre une faible présence de *Epinephelus aeneus* sur toute la zone d'étude.
Le panel à gauche de la figure 5.13 montre que la courbe bleue est au dessus de celle rouge. Le panel à droite de la figure 5.13 montre que les courbes sont superposées de la surface jusqu'à 75 m de profondeur et la courbe bleue domine sur tout le reste du profil bathymétrique. Cela nous amène à supposer que *Epinephelus aeneus* est présent dans les zones salées et relativement tempérées.
Le panel (g) de la figure 5.9 montre une forte présence de *Pseudupeneus prayensis* au centre et dans le sud.
Le panel à gauche de la figure 5.15 montre que la courbe bleue est au-dessus de celle rouge sur presque tout

le profil bathymétrique. Le panel à droite de la figure 5.15 montre que la courbe bleue est en dessous de celle rouge. Cela pourrait nous amener à supposer qu'on peut trouver plus de *Pseudupeneus Prayensis* dans les habitats salés et à basse température (eaux froides).
Le panel (f) de la figure 5.9 montre la présence de *Galeoides decadactylus* dans le sud et une absence quasi-totale au nord et dans le centre.
La figure 5.16 montre que les courbes bleues dominent (sont au dessus) de celles en rouge sur tout le profil bathymétrique. Cela laisse supposer que nous avons une forte probabilité de trouver *Galeoides decadactylus* dans les zones salées et tempérées.
Dans le paragraphe précédent (sous-section 5.2.1), nous avons décrit, à travers les informations apportées par les données, comment la répartition spatiale des espèces démersales est sensible aux variations de la température et de la salinité. Il est important de résumer l'écologie de ces espèces pour voir s'il est conforme avec cette description de leurs répartitions spatiales.

5.2.2 Bio-écologie des 7 espèces démersales

Dentex angolensis est la plus profonde des espèces de la famille de Sparidae. Nous les retrouvons dans des zones de facteurs environnementaux relativement stables avec des températures de 14 à 15°C et de salinité de l'ordre de 36. Sa présence est signalée dans des fonds riches en lutites et carbonates et à des profondeurs allant jusqu'à 200 m. Elles se nourrissent principalement de crustacés ; de poissons, parfois de mollusques et de vers [23; 25; 73; 272].
Pagellus bellottii est une espèce intertropicale qui appartient à la communauté des Sparidés ; un seul stock est signalé au Sénégal, s'étendant de l'embouchure du fleuve Sénégal au sud de la Gambie avec une forte abondance du sud de Dakar à la Gambie. Son habitat est constitué d'eaux relativement froides ($19-20°C$) entre 10 à 90 m sur les fonds durs et sableux. Nous notons de fortes concentrations de ces juvéniles en saison froide entre 5 et 30 m et des adultes entre 15 et 65 m. Ces juvéniles sont rencontrés dans les zones d'affleurement rocheux et de sables grossiers. Son reproduction est intermittente dès la deuxième année, en saison chaude avec deux pics : en Juin et en Octobre. Son lieu de ponte se trouve près de la côte sur les fonds de 50 m ; principale nurserie est la petite Côte. *Pagellus bellottii* est une espèce omnivore à prédominance carnivore. Ses proies sont des invertébrés benthiques, petits poissons, crustacés, mollusques et céphalopodes. Régime alimentaire variable selon les saisons : polychètes en printemps (80%), amphioxus en été (64%), et céphalopodes en automne (50%) [25; 30; 34; 199; 207; 334].
Octopus vulgaris est une espèce de la famille des Octopodidae. Elle possède une large aire de distribution incluant les eaux tropicales, subtropicales et tempérées (Océans Atlantique, Indien et Pacifique Ouest), à l'exception des régions polaires et sub−polaires. Un seul stock est noté au Sénégal, situé au sud de Dakar. Son habitat est constitué de fonds rocheux, sableux et boueux depuis les zones intertidales et sub−tidales jusqu'au bord du plateau continental de 10 à 200 m. Ses principales périodes de ponte sont fin Septembre-début Décembre et Février-début Mai. Elle meurt généralement après la ponte (femelle) et la couvaison (mâle). Ces embryons sont planctoniques et les adultes benthiques. Ses proies sont poissons, mollusques, crabes et crevettes ([25; 192; 209; 210]).
Pagrus caeruleosticus est une espèce démersale appartenant à la communauté des Sparidés. Son habitat est constitué des eaux froides ($< 15°$) des fonds rocheux, sableux ou sablo-vaseux. Nous trouvons un seul stock au Sénégal avec une abondance notée au sud de Dakar en des profondeurs entre 15 à 35 m. Les individus âgés migrent vers la partie plus profonde de son aire de répartition et nous retrouvons les jeunes dans des zones côtières. La ponte continue avec un pic en Juin sur les fonds de 20 à 35 m. *Pagrus caeruleosticus* une espèce omnivore à prédominance carnivore ; son régime alimentaire est varié : crustacés, céphalopodes, mollusques bivalves, petits poissons, amphioxus et vers, voir [25; 71].
Epinephelus aeneus est une espèce de la famille des Serranidae ; appartenant à la communauté démersale des sparidés, espèces à affinité saharienne. Elle est présente dans les eaux froides à fond rocheux et sablonneux. Ses juvéniles sont rencontrés dans les lagunes côtières et les estuaires. Nous notons une abondance élevée en remontée avec un seul stock au Sénégal. Sa reproduction continue avec des pics de Mai à Juin et de Juillet à Septembre. Son régime alimentaire est constitué de poissons (58%), gastéropodes (21%), crustacés (10%) et céphalopodes (10%). Ses proies préférées sont *Sardinella aurita*, celles secondaires sont *Octopus vulgaris* et *Sepia officinalis*, celles occasionnelles sont *Callinectes amincola*, voir [25; 122; 124; 206].
Pseudupeneus prayensis est une espèce intertropicale appartenant à la communauté des Sparidae. Nous avons noté un stock au Sénégal avec une forte abondance au sud de Dakar. Son habitat est caractérisé par des eaux froides entre 20 et 70 m, sur fonds rocheux ou sableux. La ponte continue avec des pics en saison chaude (Mai à Septembre). Elle est carnivore, se nourrissant principalement de vers, polychètes, crustacés, mollusques, copépodes et amphipodes [25; 214; 219; 347].
Galeoides decadactylus est une espèce de la famille des Polynemidae appartenant à la communauté des

Sciaenidae. Son habitat est constitué des eaux marines chaudes et côtières sur des fonds sablo-vaseux (10 et 20 m), parfois saumâtres, estuaires et lagunes. Nous avons noté un seul stock au Sénégal avec des migrations perpendiculaires à la côte pour échapper aux eaux pauvres en oxygène. Sa reproduction continue avec un pic en saison chaude. Elle est carnivore, se nourrissant principalement de crustacés et des petits poissons, voir [215].
Rappelons que les recherches halieutiques montrent, dans une approche écosystèmique, comment l'environnement, les conditions hydro-climatiques et les paramètres écologiques affectent la variabilité de la biomasse des poissons et leurs distributions spatiales dans le milieu marin, en étudiant les données à différents endroits de capture pendant une longue période [236; 278; 372]. Cette interaction entre le milieu marin et la ressource halieutique côtière est décrite également dans le Chapitre1. Le résumé précédent de l'écologie des espèces et la description des données dans les figures 5.10, 5.11, 5.12, 5.13, 5.14, 5.15 et 5.16, respectivement, montrent que les variables environnementales, entre autres, telles que salinité et température jouent un rôle important dans la croissance et la migration des espèces démersales citées ci-dessus. Le fait que la salinité et la température soient importantes en elles-mêmes pour décrire la répartition des espèces démersales n'est pas surprenant, car nombreuses études ont démontré leurs influences sur la structure de l'assemblage des poissons et leurs distributions spatiales [13; 16; 30; 34; 199; 207; 260; 334]. Notons que la salinité varie très peu, à la limite pendant la saison des pluies à une gamme très faible et également en profondeur où la variabilité induite par l'eau douce a très peu d'impact. Nous pouvons s'attendre à ce que cette variation à petite échelle n'influence pas, en quelle que manière que se soit, la ressource. Cependant, il est intéressant d'analyser la variation de la salinité car les micro-organismes qui jouent un rôle important sur la chaine trophique sont sensibles à cette petite variation. Ces organismes peuvent être des paramètres écologiques indiquant, en partie, la répartition verticale et/ou horizontale des poissons de manière générale des démersaux en particulier. Cela suffit pour montrer l'importance d'analyser le comportement des poissons démersaux en fonction de la salinité.
En somme, nous pouvons dire que l'approche écosystémique prédictive, de la distribution spatiale et celle de l'abondance des poissons démersaux en fonction des facteurs environnementaux, montre toute son importance ; pour une meilleure gestion des pêcheries.
Nous nous intéressons, d'abord, à la prédiction de la distribution spatiale de certaines espèces qui ont un d'intérêt économique, tenant compte des caractéristiques ponctuelles de l'environnement (SBT, SST, SBS, SSS). La prédiction se fait à travers l'application de la *classification supervisée* basée sur l'approche des k-plus proches voisins (voir la section 5.3).

5.3 Prédiction de la distribution spatiale de trois espèces démersales par la méthode des k-plus proches voisins

Nous appliquons la procédure de *classification supervisée* qui découle de la régression des k-plus proches voisins établie dans le Chapitre 3. Ainsi dans cette première application, nous modélisons par la variable binaire $Y \in \{0,1\}$; $Y = 1$ si l'espèce cible est présent dans le site sinon $Y = 0$. L'équation (4.1) est mise en œuvre en utilisant les variables environnementales ponctuelles SBT, SST, SBS et SSS (décrite dans la figure 5.1). Ces dernières supervisent la classification. Nous comparons les résultats de notre *règle de discrimination k-NN method* avec celle par noyau *kernel-method* qui découlent de la régression proposée par [98] et d'autres méthodes classiques décrites dans ce qui suit :

- La méthode de classification du package `caret`, *k-nn classic*, où la dépendance spatiale est ignorée et le nombre de voisins choisis par la Validation Croisée (CV). Notez qu'en plus des co-variables précédentes, les coordonnées géographiques (longitude et latitude) sont prises en compte comme variables explicatives.
- SVM (support vecteur machine) avec fonction de base radiale définie dans le package `caret`. Ce modèle est simulé avec les mêmes co-variables que le classificateur *k-nn classic*.
- Modèle Logit. Nous sélectionnons le meilleur modèle en terme de Critères D'information D'Akaike (AIC). Il s'agit d'un *modèle logit* avec longitude, latitude, SBS, SSS, SST, SBT comme variables explicatives.

Pour chaque espèce considérée, l'échantillon est divisé en deux avec des tailles 80% et 20%, respectivement. Le premier est utilisé pour l'apprentissage et le second pour la prédiction. Nous précisons que les deux échantillons sont sélectionnés aléatoirement et ils ont la même partition en terme de représentativité de groupe que l'échantillon initial. La validation croisée est appliquée pour, entre autres, choisir les paramètres de réglage. Nous avons utilisé des noyaux différents, certains d'entre eux ne sont pas compacts comme le suppose les hypothèses théoriques sur les noyaux.

Toutes les méthodes ont été validées avec l'échantillon d'apprentissage, nous présentons les performances des classifications réalisées sur l'échantillon de prédiction. Nous évaluons ainsi les TCC (taux de classifications correctes), respectivement, sur tous les groupes confondus, sur le premier groupe (Y = 1) et sur le deuxième (Y = 0). Les tableaux suivants 5.1, 5.2 et 5.3 donnent les résultats de TCC correspondants aux classifications des trois espèces : *Dentex angolensis, Pagrus caeruleostictus* et *Galeoides decadactylus* .

- *Dentex angolensis*

TABLEAU 5.1 – Résultats de TCC correspondants au cas de *Dentex angolensis*

		k-NN method			*kernel method*		
		TCC pour :			TCC pour :		
K_1	K_2	tous	Y = 1	Y = 0	tous	Y = 1	Y = 0
Biweight	Biweight	0.765	0.717	0.822	0.653	0.453	0.889
	Epanechnikov	0.735	0.679	0.800	0.653	0.453	0.889
	Gaussian	0.776	0.830	0.711	0.653	0.453	0.889
	Indicator	0.745	0.698	0.800	0.673	0.528	0.844
	Parzen	0.724	0.623	0.844	0.673	0.528	0.844
	Triangular	0.694	0.604	0.800	0.673	0.528	0.844
Epanechnikov	Biweight	0.745	0.755	0.733	0.612	0.396	0.867
	Epanechnikov	0.694	0.566	0.844	0.612	0.396	0.867
	Gaussian	0.765	0.792	0.733	0.551	0.265	0.889
	Indicator	0.714	0.623	0.822	0.633	0.453	0.844
	Parzen	0.745	0.698	0.800	0.643	0.509	0.800
	Triangular	0.724	0.755	0.689	0.633	0.453	0.844
Gaussian	Biweight	0.837	0.849	0.822	0.745	0.774	0.711
	Epanechnikov	**0.847**	0.86	0.822	**0.745**	0.736	0.75
	Gaussian	0.786	0.830	0.733	0.694	0.698	0.689
	Indicator	0.786	0.830	0.733	0.694	0.698	0.689
	Parzen	0.806	0.868	0.733	0.724	0.660	0.800
	Triangular	0.837	0.849	0.822	0.735	0.736	0.733
Triangular	Biweight	0.684	0.547	0.844	0.653	0.509	0.822
	Epanechnikov	0.714	0.642	0.800	0.653	0.509	0.822
	Gaussian	0.745	0.755	0.733	0.643	0.453	0.867
	Indicator	0.776	0.717	0.844	0.643	0.453	0.867
	Parzen	0.735	0.660	0.822	0.653	0.509	0.822
	Triangular	0.72	0.69	0.75	0.653	0.47	0.86
Autres méthodes							
k-nn classic		0.776	0.811	0.733			
SVM		0.816	**0.887**	0.733			
modèle logit		0.714	0.774	0.644			

- *Pagrus caeruleostictus*

TABLEAU 5.2 – Résultats de TCC correspondants au cas de *Pagrus caeruleostictus*

		k-NN method			*kernel method*		
		TCC pour :			TCC pour :		
K_1	K_2	tous	Y = 1	Y = 0	tous	Y = 1	Y = 0
Biweight	Biweight	0.724	0.452	0.851	0.704	0.097	0.985
	Epanechnikov	0.694	0.355	0.851	0.704	0.097	0.985
	Gaussian	0.755	0.581	0.836	0.684	0.00	1.00
	Indicator	0.745	0.452	0.881	0.714	0.323	0.896
	Parzen	0.663	0.258	0.851	0.694	0.032	1.00
	Triangular	0.663	0.419	0.776	0.704	0.097	0.985
Epanechnikov	Biweight	0.694	0.419	0.821	0.694	0.032	1.00
	Epanechnikov	0.714	0.323	0.896	0.694	0.032	1.00
	Gaussian	**0.796**	0.581	0.896	0.714	0.226	0.940
	Indicator	0.694	0.323	0.866	0.745	0.484	0.866
	Parzen	0.724	0.387	0.881	0.755	0.419	0.910
	Triangular	0.694	0.452	0.806	0.694	0.032	1.00
Gaussian	Biweight	0.735	0.613	0.791	0.776	0.645	0.836
	Epanechnikov	0.745	0.581	0.821	0.786	0.452	0.940
	Gaussian	0.724	0.452	0.851	**0.827**	0.581	0.940
	Indicator	0.745	0.581	0.821	0.786	0.452	0.940
	Parzen	0.714	0.548	0.791	0.786	0.452	0.940
	Triangular	0.765	0.581	0.851	0.776	0.6452	0.836
Triangular	Biweight	0.724	0.516	0.821	0.755	0.452	0.896
	Epanechnikov	0.673	0.387	0.806	0.714	0.129	0.985
	Gaussian	0.745	0.581	0.821	0.745	0.387	0.91
	Indicator	0.714	0.323	0.896	0.745	0.355	0.925
	Parzen	0.704	0.387	0.851	0.694	0.032	1.00
	Triangular	0.745	0.581	0.821	0.714	0.129	0.985
Autres méthodes							
k-nn classic		0.735	0.516	0.836			
SVM		0.786	0.452	0.940			
modèle Logit		0.704	0.258	0.910			

- *Galeoides decadactylus*

TABLEAU 5.3 – Résultats de TCC correspondants au cas de *Galeoides decadactylus*

		k-NN method			*kernel method*		
		TCC pour :			TCC pour :		
K_1	K_2	tous	Y = 1	Y = 0	tous	Y = 1	Y = 0
Biweight	Biweight	0.837	0.45	0.936	0.847	0.25	1.00
	Epanechnikov	0.867	0.60	0.936	0.847	0.45	0.949
	Gaussian	**0.959**	0.85	0.987	0.827	0.35	0.949
	Indicator	0.867	0.45	0.974	0.857	0.45	0.962
	Parzen	0.857	0.50	0.949	0.827	0.25	0.974
	Triangular	0.939	0.75	0.987	0.827	0.25	0.974
Epanechnikov	Biweight	0.898	0.70	0.949	0.868	0.50	0.962
	Epanechnikov	**0.960**	0.80	1.00	0.847	0.45	0.949
	Gaussian	0.929	0.70	0.987	0.827	0.25	0.974
	Indicator	0.949	0.80	0.987	0.827	0.15	1.00
	Parzen	0.949	0.80	0.987	0.837	0.20	1.00
	Triangular	0.939	0.75	0.987	0.837	0.35	0.962
Gaussian	Biweight	0.878	0.70	0.923	0.878	0.65	0.936
	Epanechnikov	0.888	0.60	0.962	0.878	0.65	0.936
	Gaussian	0.929	0.70	0.987	0.888	0.55	0.974
	Indicator	0.908	0.65	0.974	**0.898**	0.70	0.949
	Parzen	0.888	0.75	0.923	0.878	0.65	0.936
	Triangular	0.908	0.65	0.974	0.878	0.65	0.936
Triangular	Biweight	0.929	0.80	0.962	0.867	0.50	0.962
	Epanechnikov	0.918	0.70	0.974	0.796	0.00	1.00
	Gaussian	0.939	0.75	0.987	0.847	0.40	0.962
	Indicator	0.929	0.70	0.987	0.847	0.35	0.974
	Parzen	0.908	0.75	0.949	0.867	0.50	0.962
	Triangular	0.929	0.75	0.974	0.857	0.50	0.949
Autres méthodes							
k-nn classic		0.857	0.4	0.974			
SVM		0.908	0.75	0.95			
modèle Logit		0.878	0.75	0.91			

5.3.1 Quelques commentaires sur les résultats

Les procédures de classifications appliquées sur les trois espèces donnent des résultats différents. Concernant le *Dentex angolensis* (voir le tableau 5.1), la méthode proposée *k-NN method* donne, globalement, les meilleurs TCC, soient 85% sur les deux groupes confondus, 87% sur le premier groupe (Y = 1) et 82% sur le second (Y = 0). Notons que les autres méthodes donnent des résultats relativement bons. Par exemple, SVM donne un TCC égal à 82% sur les deux groupes confondus avec 89% et 73% de TCC, respectivement, sur le premier et le second groupe. La méthode *kernel method* donne un TCC égal à 74% sur tous les groupes confondus, 73% pour le premier groupe et 75% sur le second. Le *modèle Logit* donne un TCC égal à 71% sur les deux groupes confondus avec 77% et 64% de TCC, respectivement sur le premier et le second groupe. Le *k*-nn classic donne un TCC égal à 77% sur les deux groupes confondus avec 81% et 73% de TCC, respectivement sur le premier et le second groupe.
Pour l'espèce *Pagrus caeruleostictus* (voir tableau 5.2), la méthode proposée *k-NN method* donne un TCC égal à 79% sur le total avec 58% et 89% pour les deux groupes, respectivement. Notons que les autres méthodes ne sont pas très performantes sur la classification du premier groupe.
Pour l'espèce *Galeoides decadactylus* (voir tableau 5.3), toutes les méthodes donnent, globalement, de bons résultats avec des TCC autour de 90%. La méthode proposée *k-NN method* donne plus de précision dans les classifications avec un TCC égal à 96% sur tous les groupes confondus avec 98% et 85% sur les deux groupes respectivement.
Dans cette application, les variables SBS, SSS, SST et SBT, supervisent (ensemble) la classification. Les résultats montrent que la méthode proposée *k-NN method* peut être une bonne alternative aux méthodes qui

ne prennent pas en compte la structure spatiale. Les autres méthodes de classifications donnent des TCC moins élevés surtout quand il s'agit de calibrer le second groupe (Y = 0). Cela laisse supposer que l'absence des espèces n'est pas bien prédite par les méthodes classiques. Précisons toutefois que l'absence d'une espèce dans un site donné est relative au fait que cette dernière n'est pas capturée dans ce site.
Dans une deuxième étape dans cette direction, les profils bathymétriques de la salinité et ceux de la température (décrits dans la figure 5.7) considérés comme des variables fonctionnelles supervisent la classification. Ainsi nous nous intéressons à la prévision de la distribution spatiale de trois espèces démersales. La prédiction est définie par la classification spatiale et fonctionnelle supervisée, MFSD, définie dans la section 4.4 du Chapitre 4. La nouveauté dans cette deuxième application est que les variables fonctionnelles supervisent, individuellement, les affectations.

5.4 Prédiction de la distribution spatiale de trois espèces démersales par la méthode non-paramétrique spatiale et fonctionnelle

Nous appliquons la procédure de *discrimination* qui découle de la régression fonctionnelle établie dans le Chapitre 4. Ainsi, dans cette deuxième application, nous modélisons par la variable binaire $Y \in \{0,1\}$ la présence/absence d'une espèce ; Y = 0 si l'espèce est présente dans le site sinon Y = 0. Les variables fonctionnelles supervisent la classification (voir dans les figures 5.5 et 5.6). L'échantillon initial est divisé en deux de tailles 50% et 50%, respectivement. Le premier est pour l'apprentissage et le second pour mesurer la précision de la prédiction. Nous précisons que les deux échantillons ont la même partition en terme de représentativité de groupe que l'échantillon initial. Supposons que $\bar{\mathcal{O}}_{\mathbf{n}}$ contient l'échantillon d'apprentissage et $\mathcal{O}_{\mathbf{n}}$ contient l'échantillon de prédiction. Pour chaque site $\mathbf{i}_0 \in \bar{\mathcal{O}}_{\mathbf{n}}$, nous estimons la classe $Y_{\mathbf{i}_0}$, soit $\widehat{Y}_{\mathbf{i}_0}$ par validation croisée. La qualité de notre modèle est mesurée à l'aide de l'erreur d'estimation (ME)

$$\mathrm{ME} = \frac{1}{card\bar{\mathcal{O}}_{\mathbf{n}}} \sum_{\mathbf{i}_0 \in \bar{\mathcal{O}}_{\mathbf{n}}} 1_{[Y_{\mathbf{i}_0} \neq \widehat{Y}_{\mathbf{i}_0}]}. \tag{4.1}$$

L'échantillon d'apprentissage est utilisé pour évaluer la fiabilité des méthodes par ME et le paramètre de réglage. L 'échantillon de prédiction est utilisé pour évaluer les TCC.
La même procédure est effectuée pour la méthode fonctionnelle proposée par [146] et les autres qui ne prennent pas en compte la structure spatiale.

- Modèle fonctionnel de régression non linéaire, Méthode Fonctionnelle Classique (MFC) (voir [146])
- Modèles linéaires généralisés fonctionnels (approche de classification fonctionnelle basée sur la régression) ou régression binaire fonctionnelle, FGLM (voir [28]).
- Régression fonctionnelle additive, MASGF ([28]).
- La méthode des k-plus proches voisins fonctionnelle, kPPF (voir [28]).
- La méthode DD^G (DD-Classifier Basé sur DD-plot, voir [92])

Nous donnons dans les résultats qui suivent les ME et TCC calculés à travers les méthodes de *discrimination* citées précédement (MFSD, MFC, FGLM, MASGF, kPPF et DD^G). Nous rappelons qu'elles sont appliquées sur les espèces suivantes : *Dentex angolensis, Pagellus bellottii* et *Octapus Vulgarus.*

Cas du *Dentex angolensis*

TABLEAU 5.4 – ME et TCC correspondant à la variable fonctionnelle température

		MFSD				MFC			
		TCC pour :				TCC pour :			
K_1	K_2	ME	tous	Y=1	Y=0	ME	tous	Y=1	Y=0
Triangular	Triangular	0,32	0,64	0,63	0,67	0,40	0,46	0,70	0,52
	Epanechnikov	0,28	0,68	0,69	0,67	0,40	0,46	0,70	0,52
	Biweight	0,08	0,71	0,75	0,67	0,40	0,46	0,70	0,52
	Quad	0,28	0,68	0,72	0,63	0,40	0,46	0,70	0,52
Epanechnikov	Triangular	0,29	0,54	0,53	0,56	0,40	0,46	0,70	0,52
	Epanechnikov	0,23	0,66	0,66	0,67	0,40	0,46	0,70	0,52
	Biweight	0,04	0,69	0,75	0,63	0,40	0,46	0,70	0,52
	Quad	0,23	0,69	0,72	0,67	0,40	0,46	0,70	0,52
Biweight	Triangular	0,34	0,64	0,69	0,59	0,40	0,46	0,70	0,52
	Epanechnikov	0,30	0,64	0,63	0,67	0,40	0,46	0,70	0,52
	Biweight	0,24	0,71	0,75	0,67	0,40	0,46	0,70	0,52
	Quad	0,30	0,68	0,72	0,63	0,40	0,46	0,70	0,52
Gaussian	Triangular	0,07	0,61	0,53	0,70	0,40	0,46	0,70	0,52
	Epanechnikov	0,02	0,66	0,63	0,70	0,40	0,46	0,70	0,52
	Biweight	0,00	0,73	0,81	0,63	0,40	0,46	0,70	0,52
	Quad	0,02	0,71	0,72	0,70	0,40	0,46	0,70	0,52
Quad	Triangular	0,29	0,56	0,56	0,56	0,40	0,46	0,70	0,52
	Parzen	0,23	0,68	0,66	0,70	0,40	0,46	0,70	0,52
	Biweight	0,04	0,69	0,72	0,67	0,40	0,46	0,70	0,52
	Quad	0,23	0,66	0,66	0,67	0,40	0,46	0,70	0,52
Tukey	Triangular	0,36	0,69	0,75	0,63	0,40	0,46	0,70	0,52
	Parzen	0,29	0,75	0,81	0,67	0,40	0,46	0,70	0,52
	Biweight	**0,25**	**0,76**	**0,81**	**0,70**	**0,40**	**0,46**	**0,70**	**0,52**
	Quad	0,29	0,71	0,75	0,67	0,40	0,46	0,70	0,52
Autres méthodes									
FGLM		0,40	0,47	0,52	0,39				
MASGF		0,43	0,44	0,55	0,30				
kPPF		0,49	0,51	0,67	0,30				
DD^G		0,28	0,45	0,50	0,39				

TABLEAU 5.5 – ME et TCC correspondant à la variable fonctionnelle salinité

		MFSD				MFC			
		TCC pour :				TCC pour :			
K_1	K_2	ME	tous	Y=1	Y=0	ME	tous	Y=1	Y=0
Triangular	Triangular	0,21	0,53	0,69	0,33	0,34	0,44	0,83	0,37
	Epanechnikov	0,20	0,61	0,69	0,52	0,34	0,44	0,83	0,37
	Biweight	**0,09**	**0,75**	**0,78**	**0,70**	**0,34**	**0,44**	**0,83**	**0,37**
	Quad	0,20	0,73	0,75	0,70	0,34	0,44	0,83	0,37
Epanechnikov	Triangular	0,20	0,61	0,69	0,52	0,34	0,44	0,83	0,37
	Epanechnikov	0,16	0,64	0,63	0,67	0,34	0,44	0,83	0,37
	Biweight	0,08	0,71	0,69	0,74	0,34	0,44	0,83	0,37
	Quad	0,16	0,71	0,72	0,70	0,34	0,44	0,83	0,37
Biweight	Triangular	0,23	0,59	0,59	0,59	0,34	0,44	0,83	0,37
	Epanechnikov	0,22	0,76	0,66	0,89	0,34	0,44	0,83	0,37
	Biweight	0,13	0,73	0,75	0,70	0,34	0,44	0,83	0,37
	Quad	0,22	0,75	0,66	0,85	0,34	0,44	0,83	0,37
Gaussian	Triangular	0,13	0,51	0,53	0,48	0,34	0,44	0,83	0,37
	Epanechnikov	0,11	0,68	0,59	0,78	0,34	0,44	0,83	0,37
	Biweight	0,02	0,76	0,72	0,81	0,34	0,44	0,83	0,37
	Quad	0,11	0,73	0,63	0,85	0,34	0,44	0,83	0,37
Quad	Triangular	0,20	0,61	0,66	0,56	0,34	0,44	0,83	0,37
	Parzen	0,16	0,64	0,53	0,78	0,34	0,44	0,83	0,37
	Biweight	0,08	0,75	0,72	0,78	0,34	0,44	0,83	0,37
	Quad	0,16	0,71	0,69	0,74	0,34	0,44	0,83	0,37
Tukey	Triangular	0,24	0,69	0,69	0,70	0,34	0,44	0,83	0,37
	Parzen	0,23	0,75	0,78	0,70	0,34	0,44	0,83	0,37
	Biweight	0,12	0,71	0,78	0,63	0,34	0,44	0,83	0,37
	Quad	0,23	0,75	0,81	0,67	0,34	0,44	0,83	0,37
Autres méthodes									
FGLM		0,44	0,59	0,98	0,14				
MASGF		0,36	0,64	0,90	0,34				
kPPF		0,44	0,53	1,00	0,00				
DD^G		0,27	0,57	0,75	0,37				

Cas du *Pagellus bellottii*

TABLEAU 5.6 – ME et TCC correspondant à la variable fonctionnelle température

		MFSD				MFC			
			TCC pour :				TCC pour :		
K_1	K_2	ME	tous	Y=1	Y=0	ME	tous	Y=1	Y=0
Triangular	Triangular	0,26	0,59	0,67	0,53	0,26	0,42	0,62	0,52
	Epanechnikov	0,27	0,56	0,70	0,44	0,26	0,42	0,62	0,52
	Biweight	0,21	0,58	0,74	0,44	0,26	0,42	0,62	0,52
	Quad	0,25	0,56	0,70	0,44	0,26	0,42	0,62	0,52
Epanechnikov	Triangular	0,26	0,53	0,67	0,41	0,26	0,42	0,62	0,52
	Epanechnikov	0,25	0,56	0,74	0,41	0,26	0,42	0,62	0,52
	Biweight	0,20	0,56	0,67	0,47	0,26	0,42	0,62	0,52
	Quad	**0,23**	**0,63**	**0,89**	**0,41**	**0,26**	**0,42**	**0,62**	**0,52**
Biweight	Triangular	0,24	0,58	0,63	0,53	0,26	0,42	0,62	0,52
	Epanechnikov	0,25	0,53	0,67	0,41	0,26	0,42	0,62	0,52
	Biweight	0,23	0,53	0,59	0,47	0,26	0,42	0,62	0,52
	Quad	0,24	0,54	0,70	0,41	0,26	0,42	0,62	0,52
Gaussian	Triangular	0,23	0,53	0,93	0,19	0,26	0,42	0,62	0,52
	Epanechnikov	0,18	0,49	0,78	0,25	0,26	0,42	0,62	0,52
	Biweight	0,04	0,49	0,78	0,25	0,26	0,42	0,62	0,52
	Quad	0,18	0,49	0,74	0,28	0,26	0,42	0,62	0,52
Quad	Triangular	0,26	0,51	0,63	0,41	0,26	0,42	0,62	0,52
	Parzen	0,24	0,49	0,63	0,38	0,26	0,42	0,62	0,52
	Biweight	0,20	0,54	0,67	0,44	0,26	0,42	0,62	0,52
	Quad	0,24	0,53	0,74	0,34	0,26	0,42	0,62	0,52
Tukey	Triangular	0,21	0,54	0,63	0,47	0,26	0,42	0,62	0,52
	Parzen	0,26	0,54	0,78	0,34	0,26	0,42	0,62	0,52
	Biweight	0,20	0,59	0,70	0,50	0,26	0,42	0,62	0,52
	Quad	0,20	0,53	0,74	0,34	0,26	0,42	0,62	0,52
Autres méthodes									
FGLM		0,33	0,49	0,62	0,33				
MASGF		0,33	0,47	0,6	0,30				
kPPF		0,35	0,48	0,62	0,30				
DD^G		0,21	0,51	0,48	0,55				

TABLEAU 5.7 – ME et TCC correspondant à la variable fonctionnelle salinité

		MFSD				MFC			
		TCC pour :				TCC pour :			
K_1	K_2	ME	tous	Y=1	Y=0	ME	tous	Y=1	Y=0
Triangular	Triangular	0,22	0,64	0,83	0,45	0,22	0,42	0,55	0,64
	Epanechnikov	0,23	0,64	0,87	0,41	0,22	0,42	0,55	0,64
	Biweight	0,13	0,73	0,83	0,62	0,22	0,42	0,55	0,64
	Quad	0,23	0,63	0,83	0,41	0,22	0,42	0,55	0,64
Epanechnikov	Triangular	0,23	0,63	0,87	0,38	0,22	0,42	0,55	0,64
	Epanechnikov	0,21	0,61	0,77	0,45	0,22	0,42	0,55	0,64
	Biweight	0,09	0,63	0,70	0,55	0,22	0,42	0,55	0,64
	Quad	0,21	0,59	0,77	0,41	0,22	0,42	0,55	0,64
Biweight	Triangular	0,18	0,64	0,80	0,48	0,22	0,42	0,55	0,64
	Epanechnikov	0,20	0,61	0,73	0,48	0,22	0,42	0,55	0,64
	Biweight	0,15	0,59	0,67	0,52	0,22	0,42	0,55	0,64
	Quad	0,20	0,63	0,70	0,55	0,22	0,42	0,55	0,64
Gaussian	Triangular	0,10	0,63	0,80	0,45	0,22	0,42	0,55	0,64
	Epanechnikov	**0,07**	**0,64**	**0,83**	**0,45**	**0,22**	**0,42**	**0,55**	**0,64**
	Biweight	0,00	0,63	0,77	0,48	0,22	0,42	0,55	0,64
	Quad	0,07	0,63	0,80	0,45	0,22	0,42	0,55	0,64
Quad	Triangular	0,23	0,64	0,83	0,45	0,22	0,42	0,55	0,64
	Parzen	0,21	0,61	0,73	0,48	0,22	0,42	0,55	0,64
	Biweight	0,09	0,61	0,67	0,55	0,22	0,42	0,55	0,64
	Quad	0,21	0,61	0,70	0,52	0,22	0,42	0,55	0,64
Tukey	Triangular	0,18	0,58	0,73	0,41	0,22	0,42	0,55	0,64
	Parzen	0,22	0,61	0,70	0,52	0,22	0,42	0,55	0,64
	Biweight	0,16	0,64	0,70	0,59	0,22	0,42	0,55	0,64
	Quad	0,22	0,64	0,70	0,59	0,22	0,42	0,55	0,64
Autres méthodes									
FGLM		0,48	0,40	0,19	0,69				
MASGF		0,44	0,44	0,28	0,66				
kPPF		0,49	0,41	0,00	0,97				
DD^G		0,28	0,44	0,40	0,50				

Cas du *Octapus Vulgarus*

TABLEAU 5.8 – ME et TCC correspondant à la variable fonctionnelle température

		MFSD				MFC			
			TCC pour :				TCC pour :		
K_1	K_2	ME	tous	Y=1	Y=0	ME	tous	Y=1	Y=0
Triangular	Triangular	0,27	0,56	0,97	0,04	0,26	0,34	0,77	0,16
	Epanechnikov	0,27	0,58	1,00	0,04	0,26	0,34	0,77	0,16
	Biweight	0,23	0,59	1,00	0,08	0,26	0,34	0,77	0,16
	Quad	0,27	0,58	1,00	0,04	0,26	0,34	0,77	0,16
Epanechnikov	Triangular	0,28	0,59	1,00	0,08	0,26	0,34	0,77	0,16
	Epanechnikov	0,26	0,59	1,00	0,08	0,26	0,34	0,77	0,16
	Biweight	0,23	0,63	1,00	0,15	0,26	0,34	0,77	0,16
	Quad	0,26	0,59	1,00	0,08	0,26	0,34	0,77	0,16
Biweight	Triangular	0,25	0,54	0,94	0,04	0,26	0,34	0,77	0,16
	Epanechnikov	0,25	0,58	0,94	0,12	0,26	0,34	0,77	0,16
	Biweight	0,25	0,58	0,94	0,12	0,26	0,34	0,77	0,16
	Quad	0,25	0,59	0,97	0,12	0,26	0,34	0,77	0,16
Gaussian	Triangular	0,26	0,61	1,00	0,12	0,26	0,34	0,77	0,16
	Epanechnikov	0,25	0,61	1,00	0,12	0,26	0,34	0,77	0,16
	Biweight	**0,09**	**0,64**	**1,00**	**0,19**	**0,26**	**0,34**	**0,77**	**0,16**
	Quad	0,25	0,63	1,00	0,15	0,26	0,34	0,77	0,16
Quad	Triangular	0,28	0,56	0,97	0,04	0,26	0,34	0,77	0,16
	Parzen	0,26	0,56	0,97	0,04	0,26	0,34	0,77	0,16
	Biweight	0,23	0,61	0,97	0,15	0,26	0,34	0,77	0,16
	Quad	0,26	0,56	0,97	0,04	0,26	0,34	0,77	0,16
Tukey	Triangular	0,25	0,54	0,97	0,00	0,26	0,34	0,77	0,16
	Parzen	0,25	0,56	1,00	0,00	0,26	0,34	0,77	0,16
	Biweight	0,25	0,58	1,00	0,04	0,26	0,34	0,77	0,16
	Quad	0,25	0,56	1,00	0,00	0,26	0,34	0,77	0,16
Autres méthodes									
FGLM		0,36	0,63	0,98	0,00				
MASGF		0,31	0,60	0,88	0,11				
kPPF		0,35	0,64	1,00	0,00				
DD^G		0,32	0,56	0,77	0,19				

TABLEAU 5.9 – ME et TCC correspondant à la variable fonctionnelle salinité

		MFSD				MFC			
		TCC pour :				TCC pour :			
K_1	K_2	ME	tous	Y=1	Y=0	ME	tous	Y=1	Y=0
Triangular	Triangular	0,35	0,64	0,92	0,14	0,32	0,39	0,74	0,18
	Epanechnikov	0,35	0,69	0,92	0,29	0,32	0,39	0,74	0,18
	Biweight	0,28	0,69	0,82	0,48	0,32	0,39	0,74	0,18
	Quad	0,35	0,69	0,89	0,33	0,32	0,39	0,74	0,18
Epanechnikov	Triangular	0,35	0,68	0,97	0,14	0,32	0,39	0,74	0,18
	Epanechnikov	0,35	0,61	0,87	0,14	0,32	0,39	0,74	0,18
	Biweight	0,27	0,71	0,89	0,38	0,32	0,39	0,74	0,18
	Quad	0,35	0,66	0,89	0,24	0,32	0,39	0,74	0,18
Biweight	Triangular	0,33	0,56	0,82	0,10	0,32	0,39	0,74	0,18
	Epanechnikov	0,33	0,56	0,71	0,29	0,32	0,39	0,74	0,18
	Biweight	0,27	0,59	0,84	0,14	0,32	0,39	0,74	0,18
	Quad	0,33	0,54	0,68	0,29	0,32	0,39	0,74	0,18
Gaussian	Triangular	0,33	0,64	1,00	0,00	0,32	0,39	0,74	0,18
	Epanechnikov	0,28	0,66	0,97	0,10	0,32	0,39	0,74	0,18
	Biweight	**0,14**	**0,71**	**0,95**	**0,29**	**0,32**	**0,39**	**0,74**	**0,18**
	Quad	0,28	0,68	0,97	0,14	0,32	0,39	0,74	0,18
Quad	Triangular	0,35	0,61	0,89	0,10	0,32	0,39	0,74	0,18
	Parzen	0,35	0,58	0,82	0,14	0,32	0,39	0,74	0,18
	Biweight	0,27	0,64	0,82	0,33	0,32	0,39	0,74	0,18
	Quad	0,35	0,63	0,84	0,24	0,32	0,39	0,74	0,18
Tukey	Triangular	0,33	0,64	0,92	0,14	0,32	0,39	0,74	0,18
	Parzen	0,33	0,66	0,92	0,19	0,32	0,39	0,74	0,18
	Biweight	0,28	0,63	0,79	0,33	0,32	0,39	0,74	0,18
	Quad	0,33	0,68	0,92	0,24	0,32	0,39	0,74	0,18
Autres méthodes									
FGLM		0,35	0,67	1,00	0,04				
MASGF		0,33	**0,68**	0,98	0,12				
kPPF		0,36	0,65	1,00	0,00				
DD^G		0,23	0,55	0,71	0,23				

Quelques commentaires sur les résultats

- *Dentex angolenis*
  - Le tableau 5.4 donne les résultats de la classification correspondante à la variable fonctionnelle température. Nous notons que MFSD donne la plus petite d'erreur d'estimation, ME, pour presque tous les cas considérés. Par exemple, si nous considérons les noyaux Tukey et Biweight pour K_1 et K_2, respectivement, nous constatons que MFSD, donne ME égal à 0,25 ; MFC, donne ME = 0.40. Pour FGLM, ME est égal à 0,40 ; MASGF, donne ME égal à 0.43 ; kPPF, donne ME égal à 0.49 ; DD^G, donne ME égal à 0,28 ; Nous remarquons que tous les ME des méthodes précédentes sont inférieurs à 0,5.
    Pour les colonnes correspondantes aux TCC, MFSD donne les meilleurs TCC, soient 76% sur la classification de tous les groupes combinés, 81% et 70% sur le premier groupe (Y = 1) et le deuxième groupe (Y = 0), respectivement. MFC donne des TCC égaux à 46% sur tous les groupes combinés, 70% et 52% sur le premier groupe (Y = 1) et le deuxième groupe (Y = 0), respectivement. FGLM donne des TCC égaux à 47 % sur tous les groupes combinés, 52% et 39%

pour le premier groupe (Y = 1) et le deuxième groupe (Y = 0), respectivement. MASGF donne des TCC égaux à 44% sur tous les groupes combinés, 55% et 30 % sur le premier groupe (Y = 1) et le deuxième groupe (Y = 0), respectivement. kPPF donne des TCC égaux à 51% sur tous les groupes combinés, 67% et 30% sur le premier groupe (Y = 1) et le deuxième groupe (Y = 0), respectivement. DD^G donne des TCC égaux à 45% sur tous les groupes combinés, 50% et 39% sur le premier groupe (Y = 1) et le deuxième groupe (Y = 0), respectivement.

— Le tableau 5.5 donne les résultats de classification correspondante à la variable fonctionnelle salinité (de l'eau). Nous notons que MFSD donne la plus petite d'erreur d'estimation ME pour presque tous les cas considérés. Par exemple, considérons les noyaux triangular et Biweight pour K_1 et K_2, respectivement. MFSD donne ME égal à 0,09 ; MFC donne ME = 0,34. FGLM produit ME = 0.44 ; MASGF donne ME égal à 0.36 ; kPPF donne ME égal à 0.44 ; ME, pour DD^G est égal à 0,27.
Pour les colonnes correspondantes aux TCC, MFSD donne les meilleurs TCC, soient 75% sur la classification de tous les groupes combinés, 78% et 70% sur le premier groupe (Y = 1) et le second groupe (Y = 0), respectivement. MFC donne des TCC égaux à 44% sur tous les groupes combinés, 83% et 37% sur le premier groupe (Y = 1) et le deuxième groupe (Y = 0), respectivement. FGLM donne des TCC égaux à 59 % sur tous les groupes combinés, 98% et 14% pour le premier groupe (Y = 1) et le deuxième groupe (Y = 0), respectivement. MASGF donne des TCC égaux à 64% sur tous les groupes combinés, 90% et 34% sur le premier groupe (Y = 1) et le deuxième groupe (Y = 0), respectivement. kPPF donne des TCC égaux à 53% sur tous les groupes combinés, 100% et 00% sur le premier groupe (Y = 1) et le deuxième groupe (Y = 0), respectivement. DD^G donne des TCC égaux à 57% sur tous les groupes combinés, 70% et 37% sur le premier groupe (Y = 1) et le deuxième groupe (Y = 0), respectivement.

- *Pagellus bellottii*
  — Le tableau 5.6 donne les résultats de la classification correspondante à la variable fonctionnelle température. Nous notons que MFSD donne la plus petite d'erreur d'estimation ME pour presque tous les cas considérés. Par exemple, si nous considérons les noyaux Epanechnikov et Quadratic pour K_1 et K_2, respectivement, nous constatons que MFSD produit ME = 0,23 ; MFC donne ME = 0.26. Pour FGLM, ME est égal à 0,33 ; MASGF donne ME égal à 0.33 ; kPPF donne ME égal à 0.35 ; DD^G donne ME égal à 0,21. Nous remarquons que tous les ME inférieurs à 0,5. Pour les colonnes correspondantes aux TCC, MFSD donne les meilleurs TCC, soient 63% sur la classification de tous les groupes combinés, 89% et 41% sur le premier groupe (Y = 1) et le deuxième groupe (Y = 0), respectivement. MFC donne des TCC égaux à 42% sur tous les groupes combinés, 62% et 52% sur le premier groupe (Y = 1) et le deuxième groupe (Y = 0), respectivement. FGLM donne des TCC égaux à 49% sur tous les groupes combinés, 62% et 33% pour le premier groupe (Y = 1) et le deuxième groupe (Y = 0), respectivement. MASGF donne des TCC égaux à 47% sur tous les groupes combinés, 60% et 30% sur le premier groupe (Y = 1) et le deuxième groupe (Y = 0), respectivement. kPPF donne des TCC égaux à 48% sur tous les groupes combinés, 62% et 30% sur le premier groupe (Y = 1) et le deuxième groupe (Y = 0), respectivement. DD^G donne des TCC égaux à 51% sur tous les groupes combinés, 48% et 55% sur le premier groupe (Y = 1) et le deuxième groupe (Y = 0), respectivement.
  — Le tableau 5.7 donne les résultats de classification correspondante à la variable fonctionnelle salinité (de l'eau). Nous notons que MFSD donne la plus petite d'erreur d'estimation ME pour presque tous les cas considérés. Par exemple, considérons les noyaux Gaussian et Epanechnikov pour K_1 et K_2, respectivement. MFSD donne ME égal à 0,07 ; MFC donne ME = 0,22. FGLM produit ME = 0.48 ; MASGF donne ME égal à 0.44 ; kPPF donne ME égal à 0.49 ; ME, pour DD^G est égal à 0,28.
  Pour les colonnes correspondantes aux TCC, MFSD donne les meilleurs TCC, soient 64% sur la classification de tous les groupes combinés, 83% et 45% sur le premier groupe (Y = 1) et le deuxième groupe (Y = 0), respectivement. MFC donne des TCC égaux à 42% sur tous les groupes combinés, 55% et 40% sur le premier groupe (Y = 1) et le deuxième (Y = 0), respectivement. FGLM donne des TCC égaux à 49% sur tous les groupes combinés, 19% et 69% pour le premier groupe (Y = 1) et le deuxième groupe (Y = 0), respectivement. MASGF donne des TCC égaux à 44% sur tous les groupes combinés, 28% et 66% sur le premier groupe (Y = 1) et le deuxième groupe (Y = 0), respectivement. kPPF donne des TCC égaux à 41% sur tous les groupes combinés, 00% et 97% sur le premier groupe (Y = 1) et le deuxième groupe (Y = 0), respectivement. DD^G donne des TCC égaux à 44% sur tous les groupes combinés, 40% et 50% sur le premier groupe (Y = 1) et le deuxième groupe (Y = 0), respectivement.
- *Octopus Vulgarus*

— Le tableau 5.8 donne les résultats de la classification correspondante à la variable fonctionnelle température. Nous notons que MFSD donne la plus petite d'erreur d'estimation ME pour presque tous les cas considérés. Par exemple, si nous considérons les noyaux Gaussian et Biweight pour K_1 et K_2, respectivement, nous constatons que MFSD produit ME = 0,09 ; MFC donne ME = 0.26. Pour FGLM, ME est égal à 0,36 ; MASGF donne ME égal à 0.31 ; kPPF donne ME égal à 0.35 ; DD^G donne ME égal à 0,32. Nous remarquons que tous les ME inférieurs à 0,5. Pour les colonnes correspondantes aux TCC, MFSD donne les meilleurs TCC, soient 64% sur la classification de tous les groupes combinés, 100% et 0.19% sur le premier groupe (Y = 1) et le deuxième groupe (Y = 0), respectivement. MFC donne des TCC égaux à 34% sur tous les groupes combinés, 77% et 16% sur le premier groupe (Y = 1) et le deuxième groupe (Y = 0), respectivement. FGLM donne des TCC égaux à 63% sur tous les groupes combinés, 98% et 00% pour le premier groupe (Y = 1) et le deuxième groupe (Y = 0), respectivement. MASGF donne des TCC égaux à 60% sur tous les groupes combinés, 88% et 11% sur le premier groupe (Y = 1) et le deuxième groupe (Y = 0), respectivement. kPPF donne des TCC égaux à 64% sur tous les groupes combinés, 100% et 00% sur le premier groupe (Y = 1) et le deuxième groupe (Y = 0), respectivement. DD^G donne des TCC égaux à 56% sur tous les groupes combinés, 77% et 19% sur le premier groupe (Y = 1) et le deuxième groupe (Y = 0), respectivement.

— Le tableau 5.9 donne les résultats de classification correspondante à la variable fonctionnelle salinité (de l'eau). Nous notons que MFSD donne la plus petite d'erreur d'estimation ME pour presque tous les cas considérés. Par exemple, considérons les noyaux Gaussian et Biweight pour K_1 et K_2, respectivement. MFSD donne ME égal à 0,14 ; MFC donne ME = 0,32. FGLM produit ME = 0.35 ; MASGF donne ME égal à 0.33 ; kPPF donne ME égal à 0.36 ; ME, pour DD^G est égal à 0,23.
Pour les colonnes correspondantes aux TCC, MFSD donne les meilleurs TCC, soient 64% sur la classification de tous les groupes combinés, 83% et 45% sur le premier groupe (Y = 1) et le deuxième (Y = 0), respectivement. MFC donne des TCC égaux à 39% sur tous les groupes combinés, 74% et 18% sur le premier groupe (Y = 1) et le deuxième groupe (Y = 0), respectivement. FGLM donne des TCC égaux à 67% sur tous les groupes combinés, 100% et 0.04% pour le premier groupe (Y = 1) et le deuxième groupe (Y = 0), respectivement. MASGF donne des TCC égaux à 68% sur tous les groupes combinés, 98% et 12% sur le premier groupe (Y = 1) et le deuxième groupe (Y = 0), respectivement. kPPF donne des TCC égaux à 65% sur tous les groupes combinés, 100% et 00% sur le premier groupe (Y = 1) et le deuxième groupe (Y = 0), respectivement. DD^G donne des TCC égaux à 55% sur tous les groupes combinés, 71% et 23% sur le premier groupe (Y = 1) et le deuxième groupe (Y = 0), respectivement.

Interprétation

Les procédures de classification appliquées sur les trois espèces donnent des résultats différents.
Nous notons que, sur l'échantillon d'apprentissage, toutes les méthodes appliquées sont fiables. En effet, elles produisent des erreurs d'estimations de classes ME très petite (< 0,5). La plus petite erreur ME est produite par MFSD.
Sur l'échantillon de prédiction, nous remarquons, globalement, que les meilleures précisions de classement, sur tous les groupes combinés, ont été données par MFSD. Les autres méthodes qui ne prennent pas en compte la structure spatiale ne parviennent pas à classer correctement ces deux groupes ensemble, sauf pour le cas *Octapus Vulgarus*. En effet, pour cette espèce si, nous prenons le cas de la variable fonctionnelle salinité comme superviseur de la *règle de discrimination*, nous voyons que la méthode MASGF donne le TCC le plus élevé.
Cependant, bien que MFSD donne le meilleur TCC dans l'ensemble, ses performances diminuent, si nous considérons séparément le premier groupe (Y = 1) et le second groupe (Y = 0), respectivement. En effet, pour ces espèces, selon la variable fonctionnelle (température pour la plus part) qui supervise la *règle de classification*, MFSD donne une bonne qualité de classification sur le premier groupe (Y = 1). Pour le *Dentex angolensis*, lorsque la salinité supervise la *discrimination* les autres méthodes sont plus performantes sur la prédiction du premier groupe (Y = 1).
En ce qui concerne les deux espèces *Pagellus bellottii* et *Octapus Vulgarus*, la méthode proposée MFSD ne semble pas être bien adaptée pour la prédiction du second groupe (Y = 0) . Cependant, nous rappelons que, le fait qu'une espèce ne soit pas capturée sur un site donné ne signifie pas forcément qu'elle ne se trouve pas dans ce site. Les évitements développés par certaines espèces peuvent poser le problème de capturabilité, une notion bien connue en sciences halieutiques. Cela explique peut être le fait que la MFSD prédit moins bien le second groupe (Y = 0).

Les *règles de discrimination* appliquées dans les sections précédentes permettent de prédire la répartition spatiale des démersaux, en ce sens qu'elles affirment, conditionnellement, la présence/ou absence d'une espèce démersale donnée dans les sites où des conditions environnementales sont connues.
Nous nous intéresserons maintenant à la prédiction des quantités de biomasse des poissons démersaux, CPUE, dans les sites où ces espèces sont présentes (section 5.5).

5.5 Prédiction de quantité de biomasse de poissons démersaux par la méthode non-paramétrique spatiale et fonctionnelle

La prédiction des CPUE sur les sites où les conditions environnementales (dont la température et salinité) sont connues se fait en utilisant la méthode de prédiction fonctionnelle développée dans le Chapitre 4. Nous utilisons les données fonctionnelles lissées sans valeurs aberrantes (voir les panels (a) et (b) de la figure 5.7) et les CPUE associées. Soit $\mathscr{I}_{\mathbf{n}}$ l'ensemble des sites et $\widehat{\mathbf{n}}$ la taille de l'échantillon. Nous souhaitons prédire la CPUE en un site donné $\mathbf{i_0}$ dans $\mathscr{I}_{\mathbf{n}}$ où nous supposons que $Y_{\mathbf{i}_0}$ (CPUE sur le site $\mathbf{i_0}$) n'est pas mesuré. La prédiction de CPUE dans le site $\mathbf{i_0}$ (soit $\widehat{Y}^{\sharp}_{\mathbf{i}_0}$ défini dans l'équation (2.3)) se base sur les observations $(Y_{\mathbf{i}}, X_{\mathbf{i}})_{\mathbf{i}\in\mathscr{O}_{\mathbf{n}}}$ disponibles sur les $\widehat{\mathbf{n}}-1$ sites ($\mathscr{O}_{\mathbf{n}} = \mathscr{I}_{\mathbf{n}} \setminus \{\mathbf{i}_0\}$) et la co-variable fonctionnelle (salinité ou température) $X_{\mathbf{i}_0}(.)$. Afin de mettre en évidence les performances de ce prédicteur, nous le comparons, par le critère de l'erreur quadratique moyenne (MSE), avec le prédicteur à noyau qui ne prend pas en compte la proximité spatiale entre les sites, soit $\widehat{Y}^{\star}_{\mathbf{i}_0}$. Ce prédicteur est défini dans l'équation (2.4).
Nous calculons l'erreur quadratique moyenne $\mathrm{MSE}^{\sharp}$ et $\mathrm{MSE}^{\star}$ pour, respectivement, $\widehat{Y}^{\sharp}_{\mathbf{i}_0}$ et $\widehat{Y}^{\star}_{\mathbf{i}_0}$. L'erreur quadratique moyenne est notée $\mathrm{MSE}^{(+)}$:

$$\mathrm{MSE}^{(+)} = \frac{1}{\widehat{\mathbf{n}}} \sum_{\mathbf{i}_0\in\mathscr{O}} (\widehat{Y}^{+}_{\mathbf{i}_0}(X_{\mathbf{i}_0}) - Y_{\mathbf{i}_0})^2 \text{ avec } \widehat{Y}^{+}_{\mathbf{i}_0}(X_{\mathbf{i}_0}) = \widehat{Y}^{\sharp}_{\mathbf{i}_0} \text{ ou } \widehat{Y}^{\star}_{\mathbf{i}_0}. \tag{5.1}$$

TABLEAU 5.10 – Résultats des Erreurs quadratiques moyennes des prédictions $\widehat{Y}^{\sharp}_{\mathbf{i}_0}$ et $\widehat{Y}^{\star}_{\mathbf{i}_0}$, respectivement, correspondantes à la variable fonctionnelle salinité.

K_1	K_2	$MSE^{\sharp}$	$MSE^{\star}$
Triangle	Epanechnikov	1.2426774	1.2845134
	Biweight	1.2467063	1.2845134
	Triweight	1.2424457	1.2845134
	Gaussian	1.2425748	1.2845134
	Parzen	1.2353714	**1.2845134**
	Quad	1.2426774	1.2845134
	Silverman	1.2437404	1.2845134
	Tukey	1.2447179	1.2845134
Epanechnikov	Epanechnikov	1.4477781	1.4679125
	Biweight	1.3818974	1.4679125
	Triweight	1.3370865	1.4679125
	Gaussian	1.4484400	1.4679125
	Parzen	1.4167447	1.4679125
	Quad	1.4477781	1.4679125
	Silverman	1.4484608	1.4679125
	Tukey	1.4446595	1.4679125
Parzen	Epanechnikov	1.623466	1.627702
	Biweight	1.626234	1.627702
	Triweight	1.579487	1.627702
	Gaussian	1.618070	1.627702
	Parzen	1.566932	1.627702
	Quad	1.623466	1.627702
	Silverman	1.617559	1.627702
	Tukey	**1.228367**	1.627702
p.value	$\mathbf{3.702191 \times 10^{-21}}$		

TABLEAU 5.11 – Résultats des Erreurs quadratiques moyennes des prédictions $\widehat{Y}^{\sharp}_{\mathbf{i}_0}$ et $\widehat{Y}^{\star}_{\mathbf{i}_0}$, respectivement, correspondantes à la variable fonctionnelle température.

K_1	K_2	$MSE^{\sharp}$	$MSE^{\star}$
Biweight	Triangle	1.4833015	1.6148783
	Epanechnikov	1.517042	1.6148783
	Biweight	1.4620797	1.6148783
	Triweight	1.561286	1.6148783
	Gaussian	1.529699	1.6148783
	Quad	1.517042	1.6148783
	Silverman	1.534799	1.6148783
	Tukey	1.4597212	1.6148783
Gaussian	Triangle	1.2663868	1.3601111
	Epanechnikov	1.3038538	1.3601111
	Biweight	1.3262059	1.3601111
	Triweight	1.3392725	1.3601111
	Gaussian	1.2340943	1.3601111
	Quad	1.3038538	1.3601111
	Silverman	1.2548942	1.3601111
	Tukey	**1.1825773**	**1.3601111**
Parzen	Triangle	1.3567657	1.4435814
	Epanechnikov	1.3885122	1.4435814
	Biweight	1.4135013	1.4435814
	Triweight	1.2471777	1.4435814
	Gaussian	1.4313367	1.4435814
	Quad	1.3885122	1.4435814
	Silverman	1.4072615	1.4435814
	Tukey	1.3599626	1.4435814
p.value	$\mathbf{2.177332 \times 10^{-20}}$		

Les résultats de la prédiction sont présentés dans les tableaux 5.10 et 5.11. Nous avons utilisé différentes combinaisons de types des noyaux. Globalement, le prédicteur spatial et fonctionnel proposé produit des prédictions plus précises que celui qui ne prend pas en compte la structure spatiale. En effet, pour les tableaux 5.10 et 5.11, $MSE^{\sharp}$ est significativement plus petite que $MSE^{\star}$, comme le confirment les très petites valeurs des **p.values** des t-tests de comparaison des deux types d'erreur pour déterminer si $MSE^{\sharp}$ est significativement inférieur à $MSE^{\star}$ (l'hypothèse alternative est alors H1 : $MSE^{\sharp} < MSE^{\star}$).
Le tableau 5.10 montre que $MSE^{\star} = 1,2845134$ et $MSE^{\sharp} = 1,2353714$, si nous prenons les noyaux Triangular et Parzen respectivement pour K_1 et K_2. Nous précisons également que les moyennes des erreurs quadratiques moyennes pour $\widehat{Y}^{\sharp}_{\mathbf{i}_0}$ et $\widehat{Y}^{\star}_{\mathbf{i}_0}$ sont $AMSE^{\sharp} = 1,35$ et $AMSE^{\star} = 1,46$, respectivement.
Pour la prédiction avec la température (tableau 5.11), nous voyons que sur tous les cas des noyaux considérés, respectivement pour K_1 et pour K_2, on a $MSE^{\star} > MSE^{\sharp}$. En effet, $MSE^{\star} = 1,3601111$ et $MSE^{\sharp} = 1,1825773$, si nous prenons les noyaux Gaussian et Tukey respectivement pour K_1 et K_2. Nous précisons également que les moyennes des erreurs quadratiques moyennes pour $\widehat{Y}^{\sharp}_{\mathbf{i}_0}$ et $\widehat{Y}^{\star}_{\mathbf{i}_0}$ sont $AMSE^{\sharp} = 1,38$ et $AMSE^{\star} = 1,47$, respectivement.
Par conséquent, $\widehat{Y}^{\sharp}_{\mathbf{i}_0}$ produit de meilleurs résultats de prédiction que $\widehat{Y}^{\star}_{\mathbf{i}_0}$ qui ne prend pas en compte la distance entre les sites.
Les résultats montrent que le prédicteur proposé constitue une bonne alternative à la méthode non-paramétrique classique pour le cas des données côtières du Sénégal. En comparant le prédicteur de salinité (tableau 5.10) avec celui de la température (tableau 5.11), on remarque que la salinité produit la plus petite erreur quadratique moyenne.

5.6 Conclusion

Ce Chapitre a été consacré à l'application des prédicteurs non-paramétriques sur des données spatiales et fonctionnelles.
Les *règles de classification supervisées* appliquées, comparée aux méthodes classiques, donnent les meilleures précisions de la prédiction présence/absence de l'espèce cible en fonction des conditions environnementales.
La méthode de prédiction non-paramétrique de quantité de biomasse d'un stock de poissons proposée est meilleure que la celle usuelle qui ne prend pas en compte la structure spatiale.

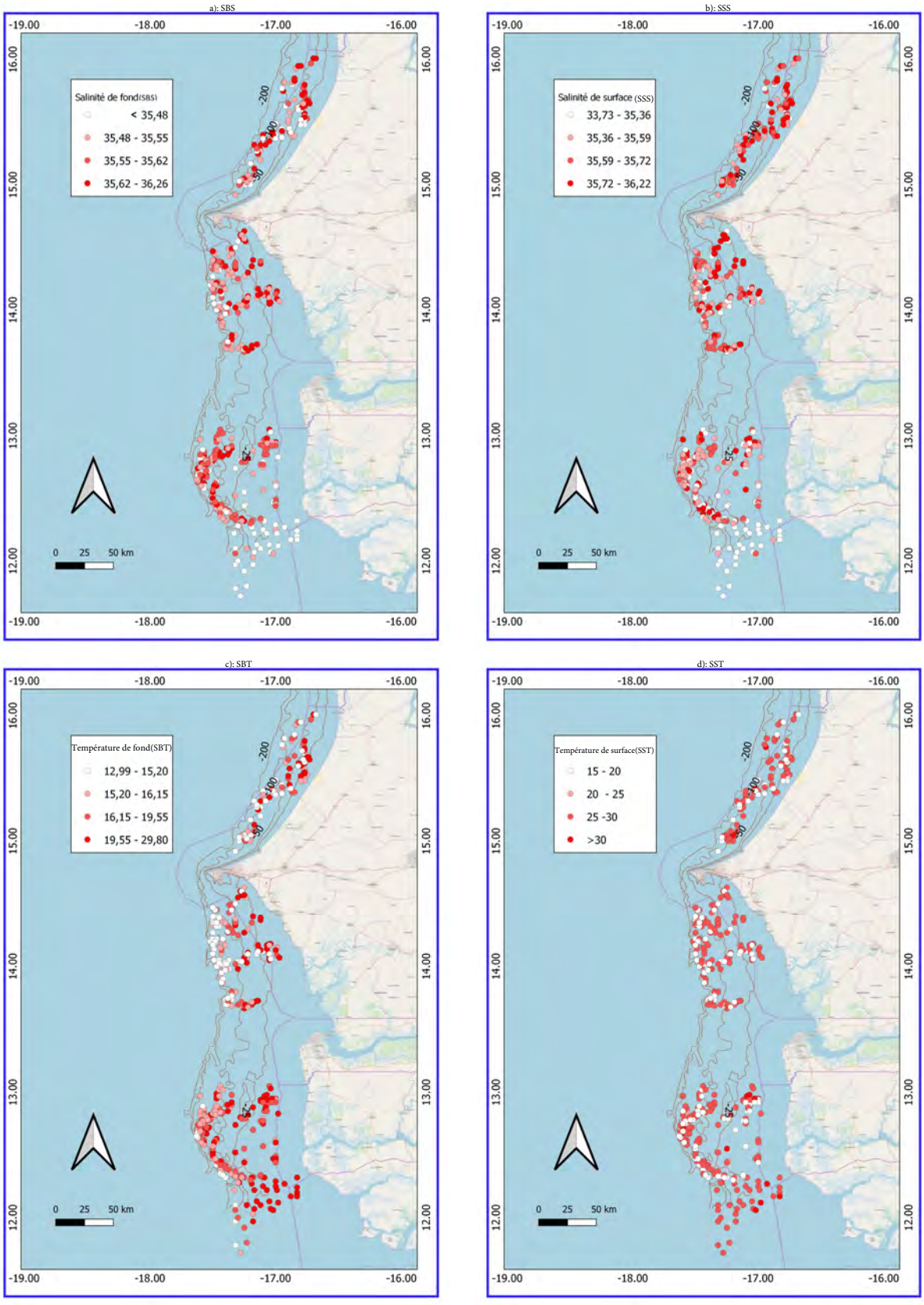

FIGURE 5.1 – Variation spatiale de la température et de la salinité dans le fond et la surface.

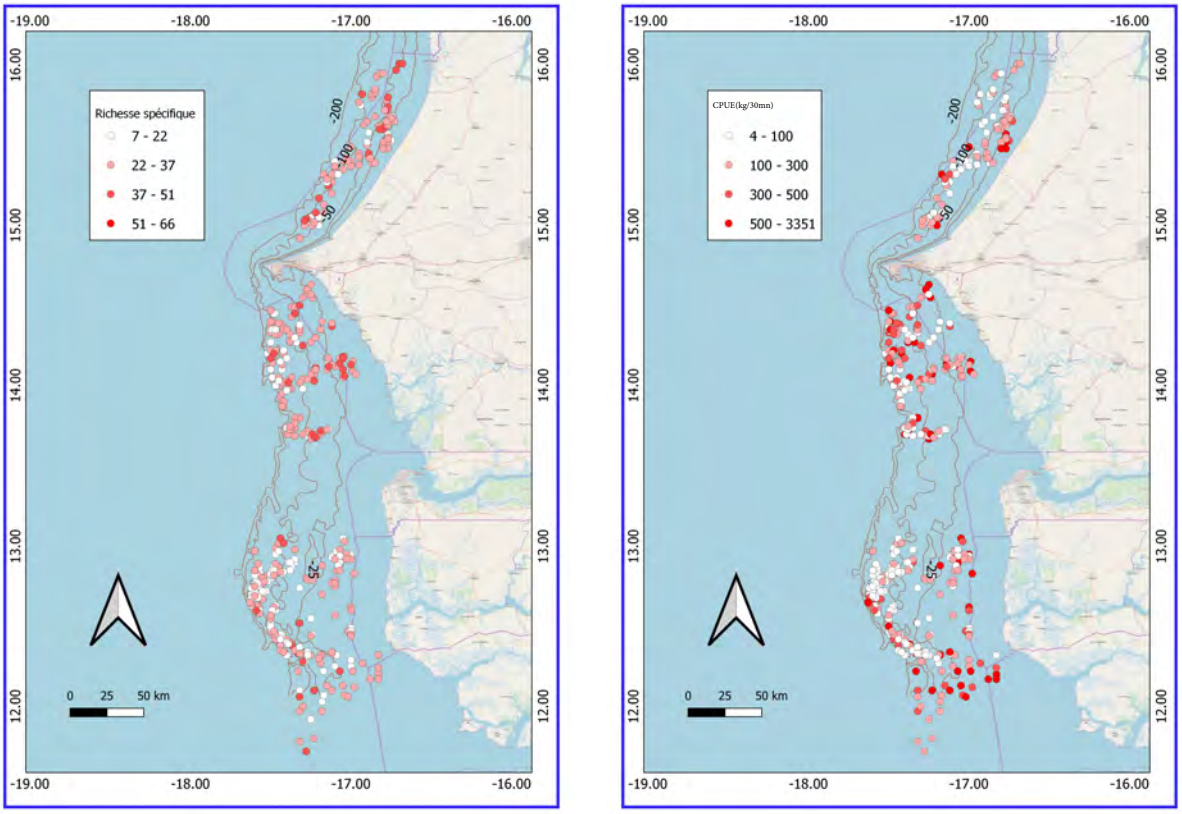

FIGURE 5.2 – Les captures par unité d'effort sur les stations : le panel de gauche est la richesse et le panel de droite est la biomasse

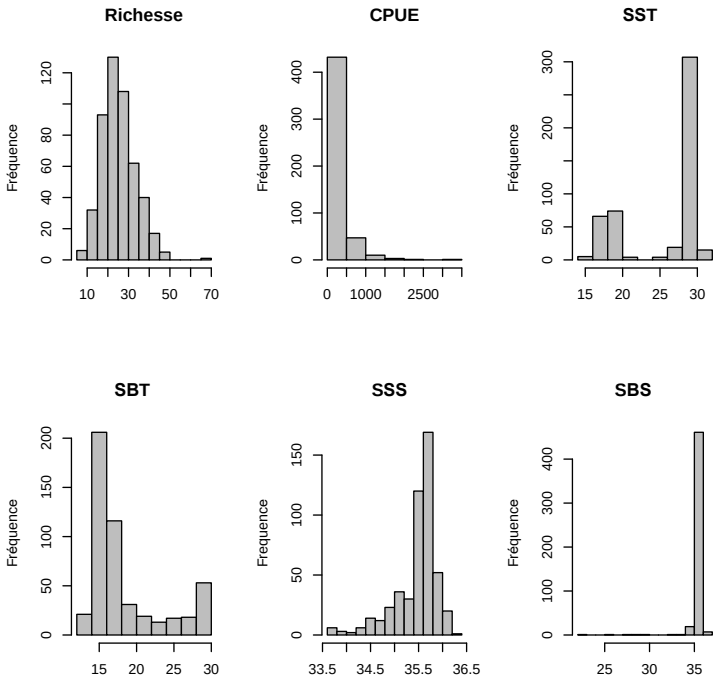

FIGURE 5.3 – Histogrammes des données quantitatives brutes

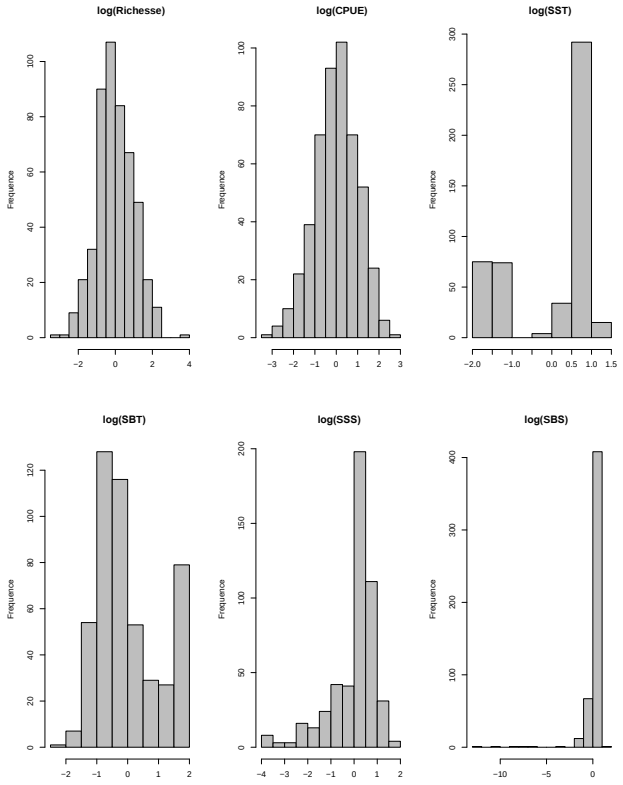

FIGURE 5.4 – Histogrammes de valeurs quantitatives transformées.

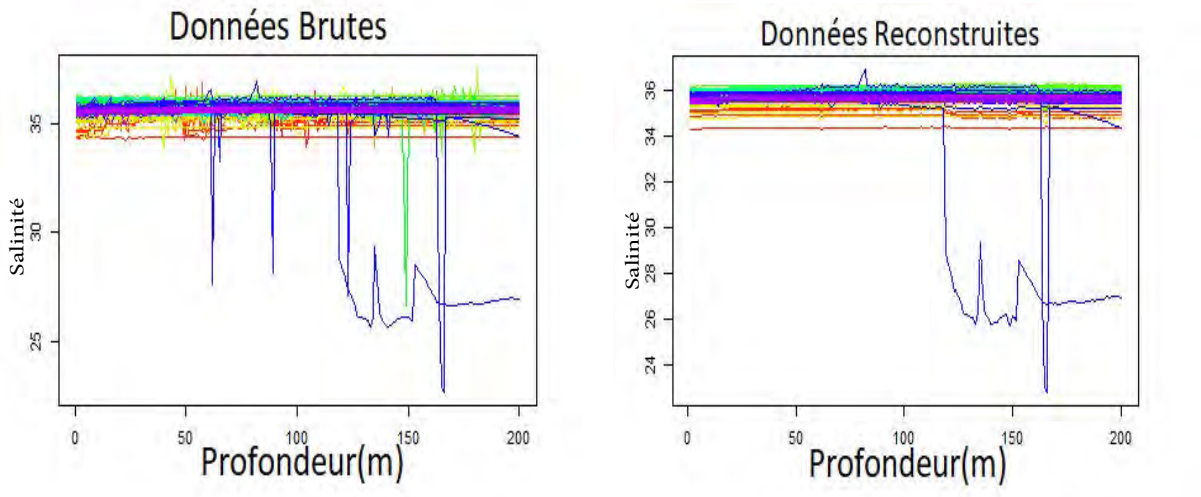

FIGURE 5.5 – Courbes des profils bathymétriques de salinité : données brutes (gauche) et reconstruites (à droite).

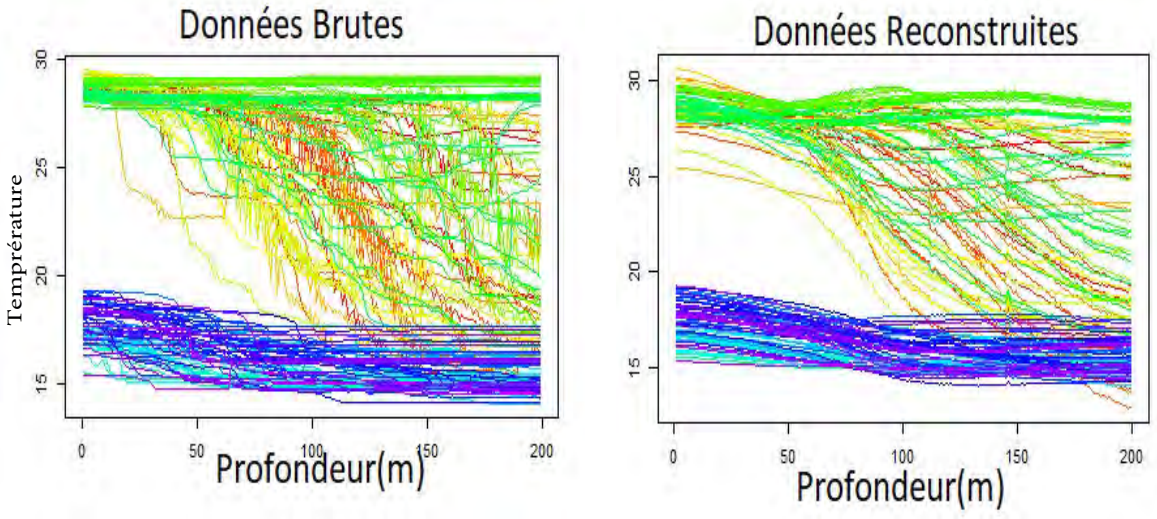

FIGURE 5.6 – Courbes des profils bathymétriques de température : données brutes (gauche) et reconstruites (à droite).

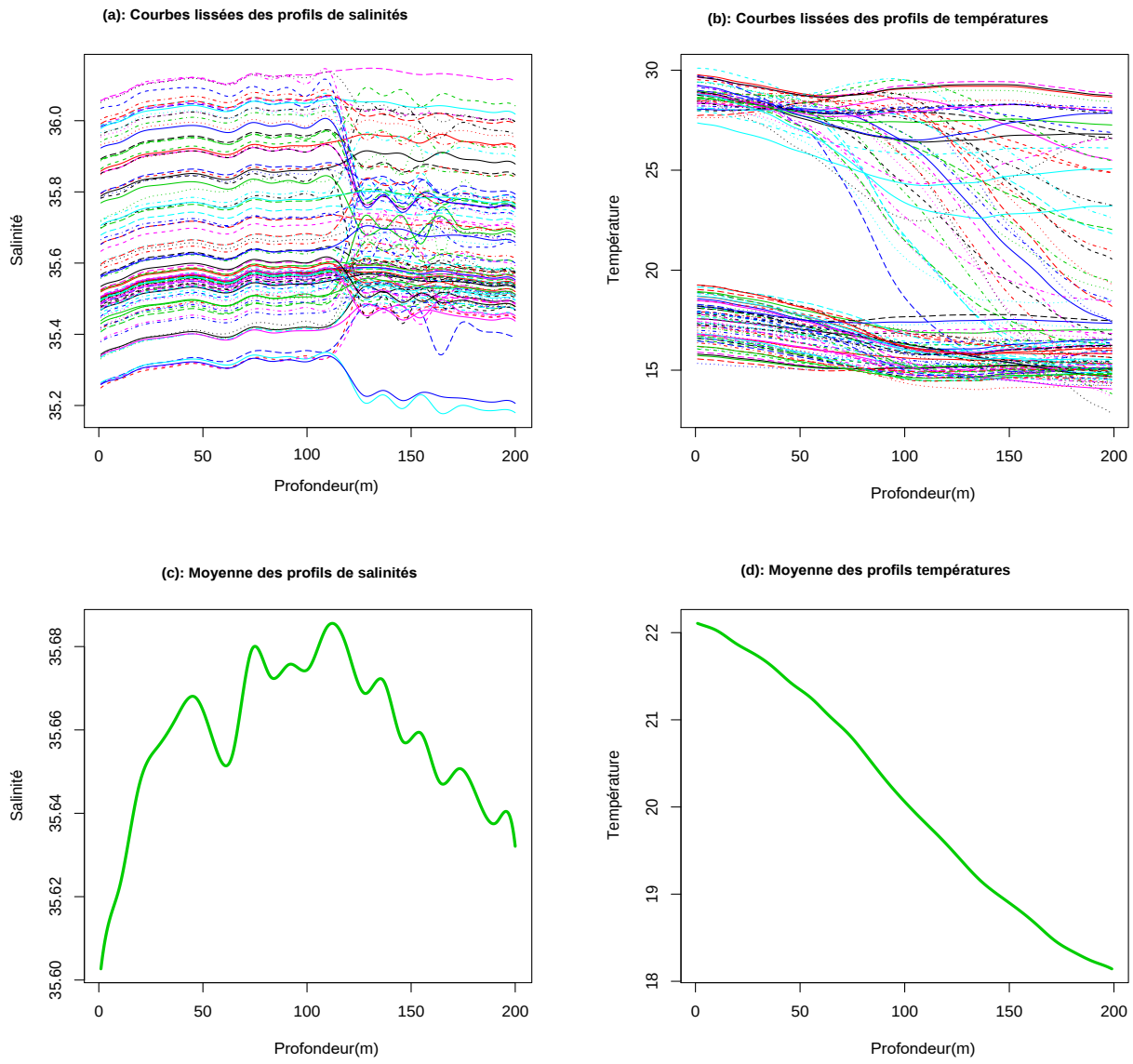

FIGURE 5.7 – Courbes environnementales lissées sans valeurs aberrantes selon le profil bathymétrique. Toutes les courbes sont lissées en utilisant la base B-spline.

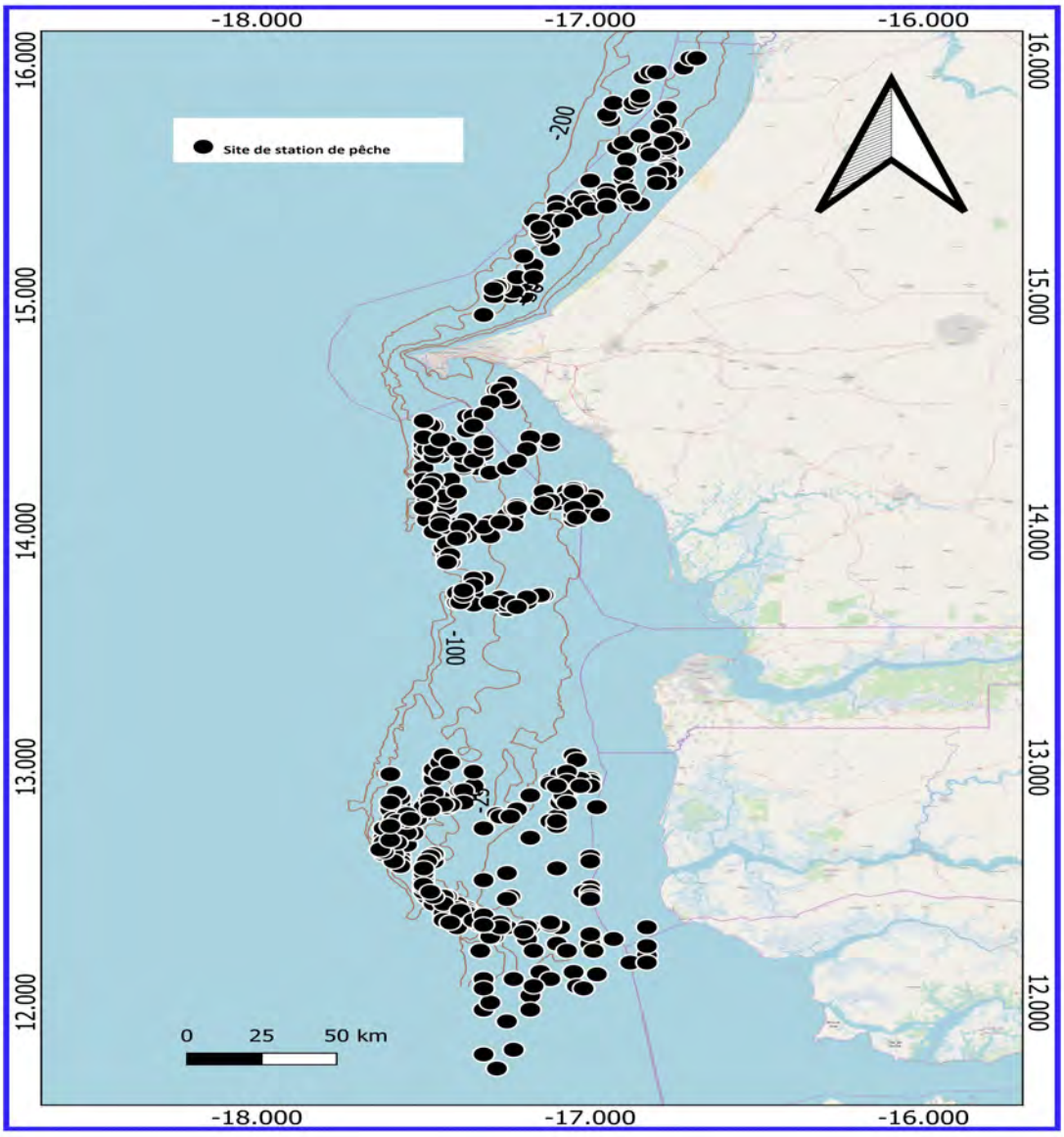

FIGURE 5.8 – Carte des stations de pêche au large du Sénégal sur le plateau continental.

La figure 5.9 donne la distribution spatiale des poissons cibles

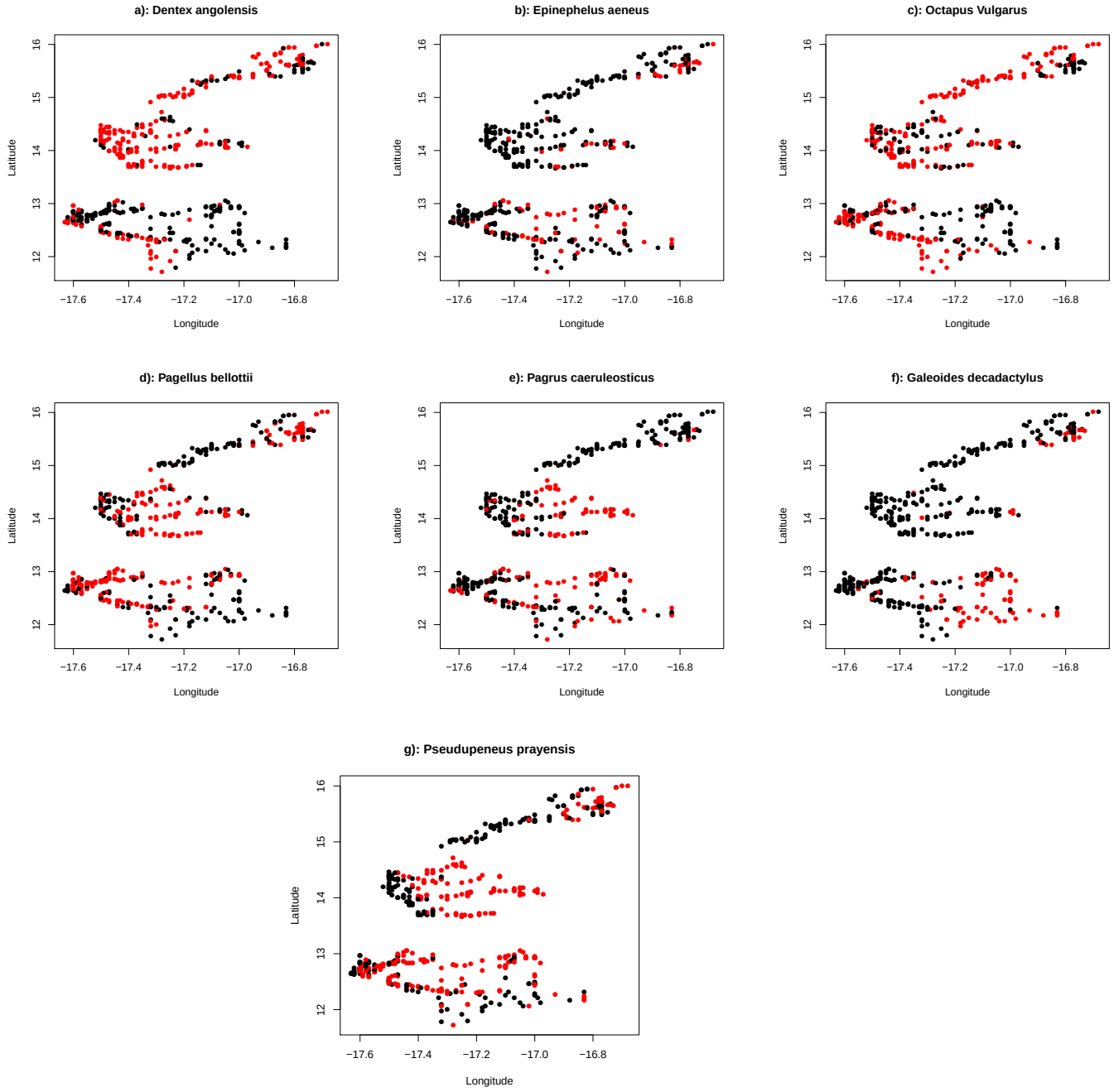

FIGURE 5.9 – Distribution spatiale des 7 espèces considérées. Le point rouge indique la présence du poisson démersal côtier et le point noir indique son absence

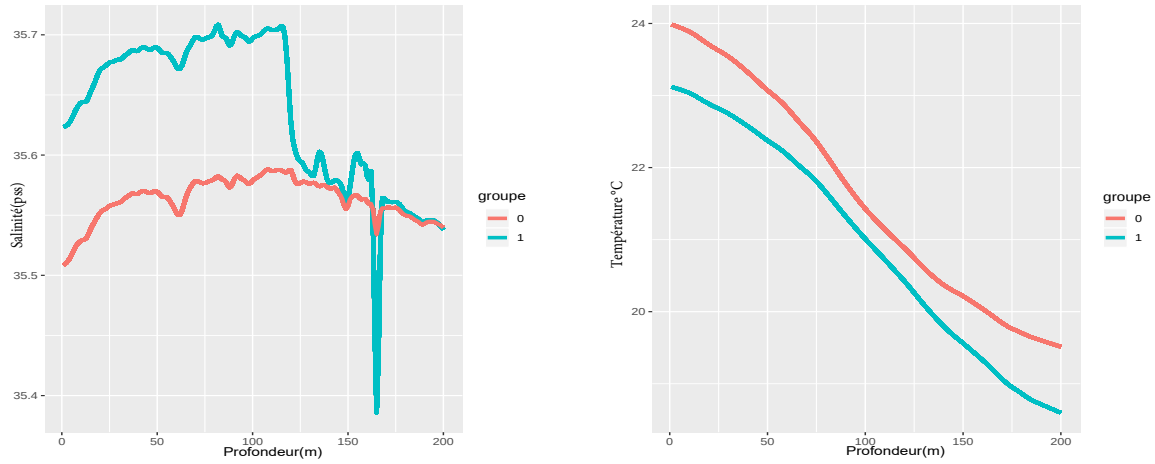

FIGURE 5.10 – Courbes moyennes de la salinité (panel de gauche) el la température (panel de droite) correspondant à la répartition spatiale de *Dentex angolensis* : présence (**label 1**) et absence (**label 0**)

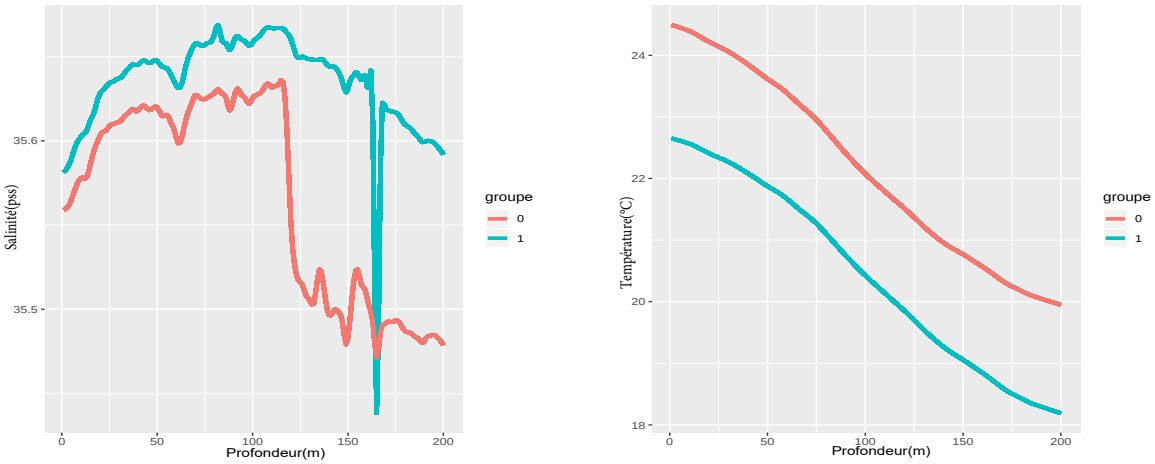

FIGURE 5.11 – Courbes moyennes de la salinité (panel de gauche) et la température (panel de droite) correspondant à la répartition spatiale de *Pagellus bellottii* : présence (**label 1**) et absence (**label 0**).

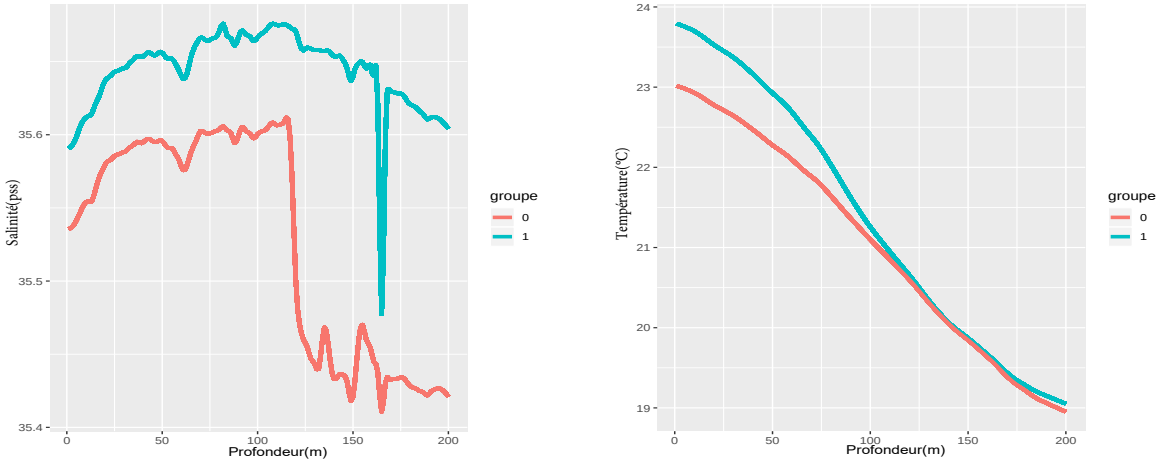

FIGURE 5.12 – Courbes moyennes de la salinité (panel de gauche) et la température (panel de droite) correspondant à la répartition spatiale de *Octapus Vulgaris* : présence (**label 1**) et absence (**label 0**)

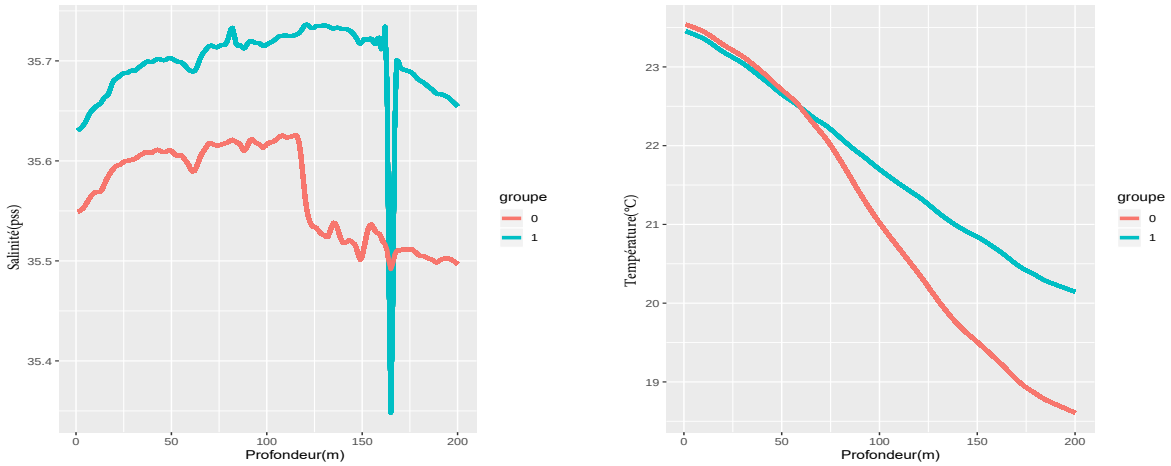

FIGURE 5.13 – Courbes moyennes de la salinité (panel de gauche) et la température (panel de droite) correspondant à la répartition spatiale de *Epinephelus aeneus* : présence (**label 1**) et absence (**label 0**)

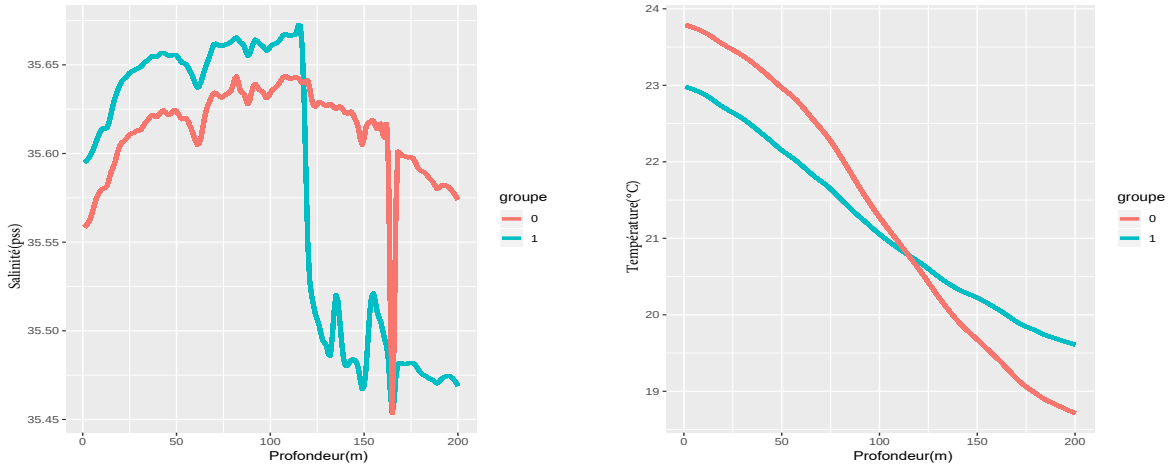

FIGURE 5.14 – Courbes moyennes de la salinité (panel de gauche) et la température (panel de droite) correspondant à la répartition spatiale de *Pagrus caeruleosticus* : présence (**label 1**) et absence (**label 0**)

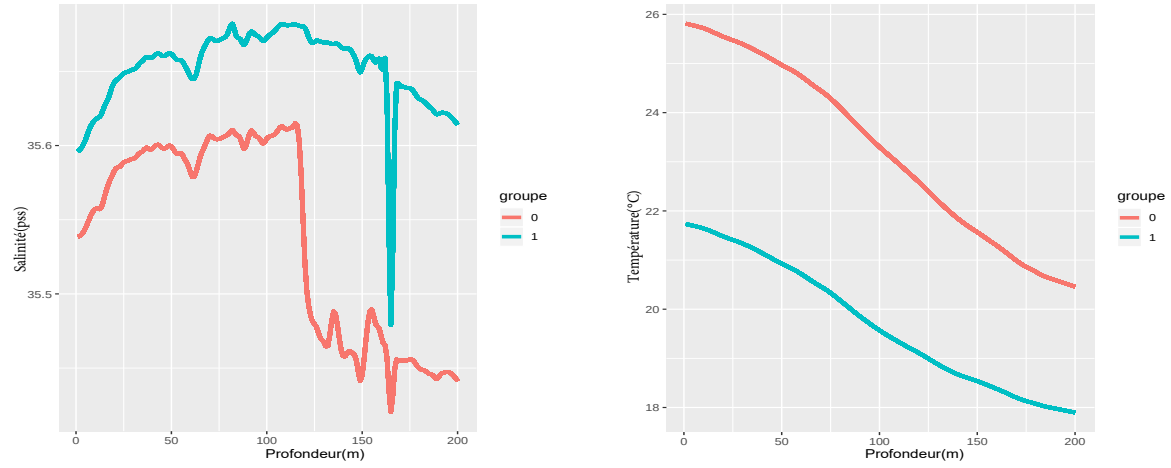

FIGURE 5.15 – Courbes moyennes de la salinité (panel de gauche) et la température (panel de droite) correspondant à la répartition spatiale de *Pseudupeneus priensis* : présence (**label 1**) et absence (**label 0**)

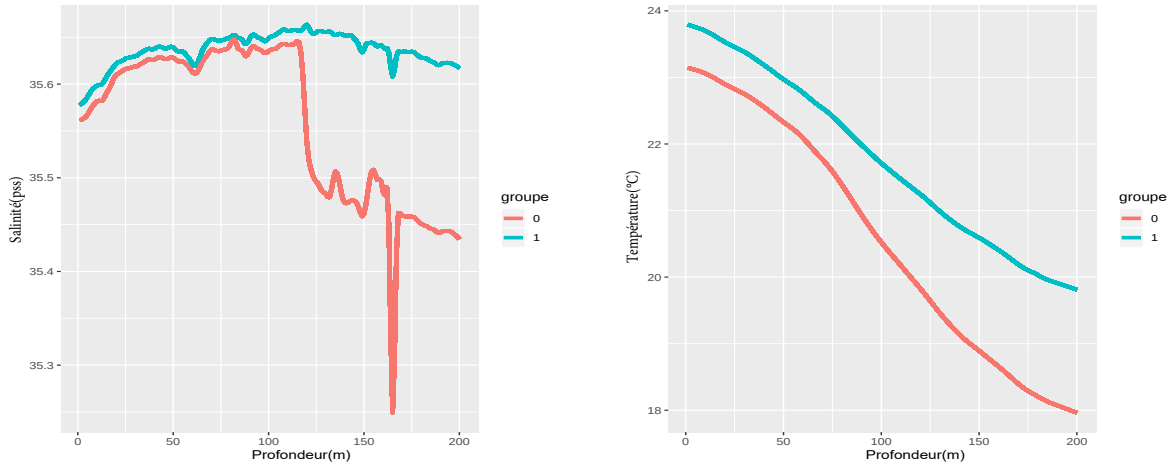

FIGURE 5.16 – Courbes moyennes de la salinité (panel de gauche) et la température (panel de droite) correspondant à la répartition spatiale de *Galeoides decadactylus* : présence (**label 1**) et absence (**label 0**)

CHAPITRE 6

CONCLUSION GÉNÉRALE ET PERSPECTIVES

6.1 Conclusion

Dans ce travail, nous avons proposé deux méthodes de modélisation spatiale/et spatio-fonctionnelle. Elles reposent sur des modèles de régression non-paramétrique à deux noyaux. La prédiction et la *classification supervisée* sont les deux outils mathématiques d'applications qui découlent de ces approches non-paramétriques de régression. Nous étudions le prédicteur de [98] dans un contexte d'hétérogénéité spatiale, d'une part. Nous proposons l'extension du même prédicteur dans un cadre fonctionnel, d'autre part. Cette extension de l'estimateur dans le cadre fonctionnel est perçue comme une contribution à la ADF (FDA en anglais). Les résultats asymptotiques ont été établis par le biais de la convergence presque complète. Un mode de convergence simple à établir et qui entraine celui presque-sûre. Ils reposent, entre autres, sur des hypothèses telles que la dépendance spatiale, la non-linéarité et la stationnarité locale du processus. Les théories spatiale et fonctionnelle (prédiction et/ ou classification) développées dans cette thèse ont été appliquées à des données simulées pour illustrer le comportement des estimateurs. Elle sont aussi été appliquées à des données réelles halieutiques. L'originalité des méthodes proposées est le fait de prendre en compte la structure spatiale, l'effet des co-variables et la continuité éventuelle des phénomènes modélisés par les co-variables fonctionnelles. En effet, la méthode prend en compte la proximité géographique des sites et la distance entre les observations sur ces sites, dans un cadre spatial et fonctionnel. Ces données ont un support de mesure spatiale (coordonnées géographiques) et certaines sont de nature fonctionnelle. Les modèles sont adaptés en fonction de la structure du support de mesure des données spatiales et en fonction de la nature de celles-ci. La première approche de prédiction et de classification est appliquée à des données spatiales réelles (Chapitre 3), alors que, la seconde approche est appliquée à des données spatiales qui sont de nature fonctionnelle (Chapitre 4). Les *règles de classification supervisée* sont utilisées pour prédire la distribution spatiale des espèces halieutiques cibles. Les méthodes de prédictions permettent d'évaluer la quantité, espérée, de biomasse des poissons démersaux. Les modèles sont construits sur la base de toutes les informations fournies par les données. Cela permet de construire des estimateurs qui minimisent la perte éventuelle d'information pour mieux prendre en compte toute la réalité du phénomène étudié. À notre connaissance, ces méthodes constituent de nouvelles approches qui peuvent être particulièrement utiles pour la modélisation des données issues de la campagne scientifique du CRODT. Les applications sur les données du CRODT montrent que les méthodes de prédiction et de *classifications supervisées* proposées dans nos travaux, sont une bonne alternative aux méthodes classiques utilisées au sein du CRODT. Les résultats montrent que les performances des méthodes proposées surpassent celles des prédicteurs classiques. Ils montrent l'importance de prendre en compte la structure spatiale et l'effet de l'environnement dans la prédiction de la distribution des poissons et l'évaluation de leurs densités. La méthodologie peut permettre d'améliorer la compréhension de la répartition des poissons démersaux côtiers sénégalais. Elle facilite l'élaboration des modèles de gestion des stocks pour rendre l'exploitation des pêcheries plus durable. Cela peut être utile pour identifier les zones de concentration potentielle d'espèces côtières pendant les saisons chaudes et froides au Sénégal. Ce travail offre, donc, des outils de gestion et suivi des pêcheries pour une meilleure pérennisation des ressources démersales dans un contexte du changement climatique. Il faut noter que ces campagnes scientifiques interviennent dans une période marquée par l'effet des changement climatiques dans tout l'écosystème marin. La méthode peut aider à se décider sur la réduction des efforts de pêche et les coups financiers en visitant moins de stations, stratégiques,

lors des campagnes scientifiques du CRODT. Dans cette contribution, nous n'utilisons que les captures par unité d'effort des données sur les poissons démersaux côtiers sénégalais observées en 2016. Travailler avec toutes les données disponibles sur la période $2001-2016$ doit être fait en tenant compte d'une certaine variabilité spatio-temporelle et l'impact du changement climatique sur la répartition spatiale et la variabilité de la biomasse de la ressource. Le problème qui se posera est le manque de données dans certains endroits, ce qui pose la question comment gérer l'irrégularité de l'espace-temps, dans la modélisation. D'autre part, la variation de la durée des cycles biologiques des espèces démersaux pose le problème de la variabilité de la biomasse inter-annuelle de la ressource. Ces problèmes peuvent être résolus grâce aux progrès de la ADF sur la modélisation des données éparses et celles spatio-temporelle. Pour aller plus loin dans cette application marine, un modèle de prédiction incluant, simultanément, plus d'un paramètre environnemental pourrait améliorer les résultats. Ce modèle prendra en compte l'irrégularité de l'acquisition des informations dans le temps et dans l'espace. Un autre modèle basé sur l'introduction d'un troisième noyau qui prend en compte l'aspect temporel permettrait de voir l'effet de la variation inter/ ou intra-annuelle sur le comportement de la ressource.

6.2 Perspectives

Quelques questions sont restées sans réponse et peuvent alimenter nos perspectives de recherches. Nous donnons dans la suite quelques pistes d'investigation futures.

6.2.1 Modèle non-paramétrique spatial d'évaluation de stock de poisson utilisant un transect ponctuel

L'approche d'estimation spatiale k-NN proposée dans le Chapitre 3 sera adaptée dans un cadre particulier d'évaluation de stock de poisson. Ainsi, nous considérons un modèle spatial non-paramétrique pour l'estimation de la densité de population dérivée de la méthode des transects (distance sampling).
Les méthodes des transects (distance sampling) sont utilisées pour estimer la densité de population d'objets dans une zone donnée. Ces objets sont généralement des animaux ou groupes d'animaux; ils peuvent être aussi des arbres. Dans cette extension proposée, les objets sont des poissons, un banc de poissons, ou des bancs de poissons. Dans l'échantillonnage par transect ponctuel, les objets sont enregistrés à partir d'un certain nombre de stations $\hat{\mathbf{n}}$ fixes appelées transects ponctuels. Les données principales sont les distances d'éloignement r entre les points de référence (stations) et les positions des objets détectés. Plusieurs auteurs ont proposé des méthodes paramétriques et non-paramétriques permettant d'estimer la densité de population par la méthode des transects, voir [7; 18; 58; 89; 131; 239]. Nous proposons une méthode non-paramétrique d'estimation de densité de poissons qui prend en compte la nature de la structure spatiale. La particularité de notre méthode est de prendre en compte la proximité entre les transects et dévaluer la densité de population des démersaux qui ne peuvent pas être détectés dans les eaux côtières avec précision. L'estimateur proposé dépend du nombre de stations visitées et prend deux noyaux, l'un contrôle la distance entre les observations et l'autre contrôle la proximité géographique des points transects (ou station).

Dans l'évaluation de densité de population par la méthode des transects, le processus d'estimation suppose que chaque objet est associé à une probabilité de détection g, qui dépend de la distance d'éloignement r de l'objet et de sa taille s. Nous précisons que plus l'objet est loin moins il est détectable et la taille s de l'objet peut être le poids ou la longueur. Si l'objet est un banc de poissons, s peut être le nombre d'individus ou le poids du banc. Si l'objet est constitué par des bancs de poissons, s peut être le nombre de bancs. g dépend aussi de la taille s de l'objet car un objet plus grand est plus facilement détectable qu'un objet plus petit ou plus court, situé à la même distance. La probabilité de détection devient donc une fonction bi-variée $g(r,s)$. Dans ce contexte, la densité D est la taille moyenne agrégée (de l'objet) par unité de surface. Si l'objet est un groupe de poissons (par exemple des groupes d'espéces de poissons), alors D peut être le nombre moyen d'individus par unité de surface. Nous nous occuperons également d'estimer la densité des objets, le nombre moyen d'objets par unité de surface Δ, c'est-à-dire la taille moyenne μ de chaque objet. Un exemple que nous pouvons donner pour éviter de confondre ces paramètres est une opération de chalutage : D est le poids moyen d'une capture, Δ le nombre d'individus qui la composent et μ, le poids moyen de chaque individu. La base mathématique de l'évaluation de densité de population par transect (distance sampling) a été développée par [60; 317]. Le livre de [56] expose une théorie complète et des cas d' applications des échantillonnages par *distance sampling*. Les techniques utilisées incluent la modélisation paramétrique et non-paramétrique de la fonction $g(r)$ ou $g(r,s)$ (appelée la fonction de détection)

Nous rappelons que l'estimation non-paramétrique à noyau a fait l'objet de recherches intensives au cours des sept dernières décennies depuis qu'elle a été utilisée pour la première fois par Rosenblatt [306] pour estimer une fonction de densité de probabilité $f(x)$. Les travaux antérieurs dans ce sens incluent

ceux de [57], qui a implémenté l'algorithme de Silverman basé sur le noyau gaussien. Dans les travaux de [79; 285; 319], les auteurs ont appliqué la méthode du noyau dans l'estimation de la densité des poissons.

Notre objectif, tout d'abord, est de fournir une base théorique rigoureuse d'une nouvelle méthode non-paramétrique d'estimation de densité des poissons par transect. Deuxièmement, cette méthode sera appliquée pour évaluer la densité de population des poissons démersaux côtiers au large des côtes sénégalaises.

6.2.2 Régression fonctionnelle non linéaire pour les réponses fonctionnelles appartenant à des espaces de Hilbert auto-reproduisant

La méthode de prédiction établie aux Chapitres 3 et 4 est construite sous deux noyaux lorsque la variable de réponse est réelle et la co-variable est de nature fonctionnelle. Nous voulons l'étendre en prenant la variable de réponse comme une variable fonctionnelle. Ainsi, nous étudions un estimateur de régression non-paramétrique lorsque la variable de réponse appartient à un espace de Banach séparable et la variable explicative prend ses valeurs dans un espace semi-métrique séparable.
Soit $\{Z_{\mathbf{i}} = (X_{\mathbf{i}}, Y_{\mathbf{i}}) \in \mathscr{H} \times \mathscr{Y}_I\ ,\ \mathbf{i} \in \mathbb{N}^N\}$, un processus spatial défini sur un espace probabilisé $(\Omega, \mathscr{A}, \mathbb{P})$, $N \in \mathbb{N}^*$.

$$Y_{\mathbf{i}}(t) = R(X_{\mathbf{i}}(t)) + \varepsilon_{\mathbf{i}}(t), \tag{2.1}$$

où $\varepsilon(t)$ est le terme d'erreur.
Nous considérons le problème d'inférence et de prédiction pour le modèle 2.1. Nous nous intéressons à construire l'estimateur de $R_\Psi(x) = \mathbb{E}(\Psi(Y)|X = x)$, soit $R_{\Psi,\mathbf{n}}(x)$. La construction de ce dernier repose, en partie, sur deux noyaux K_1 et K_2 et sur une fonction $\Psi : \mathscr{Y}_I \to \mathscr{Y}$. K_1 prend en compte les observations et K_2 les sites d'observations. Le noyau K_1 est défini dans un espace Hilbert auto-reproduisant $\mathscr{H}$. Les ensembles $\mathscr{Y}_I$ et $\mathscr{Y}$ sont des espaces de Banach séparables. Sur la base de l'observation du processus dans $\mathscr{I}_{\mathbf{n}}$, nous avons :

$$R_{\Psi,\mathbf{n}}(.) = \frac{1}{a_{\mathbf{n}}(.)} \sum_{\mathbf{i}\in\mathscr{I}_{\mathbf{n}}} K_1(X_{\mathbf{i}}, .)\, K_{2,\rho_{\mathbf{n}}}(\|\mathbf{i}_0 - \mathbf{i}\|)\, \Psi(Y_{\mathbf{i}}). \tag{2.2}$$

Avec $a_{\mathbf{n}}(.) = \sum_{\mathbf{i}\in\mathscr{I}_{\mathbf{n}}} K_{2,\rho_{\mathbf{n}}}(\|\mathbf{i}_0 - \mathbf{i}\|)\, \mathbb{E}[K_1(X_{\mathbf{i}}, .)]$ et $K_{2,\rho_{\mathbf{n}}}(\|\mathbf{i}_0 - \mathbf{i}\|) = K_2\left(\frac{\|\mathbf{i}_0 - \mathbf{i}\|}{n\rho_{\mathbf{n}}}\right)\left(\frac{\mathbf{i}}{\mathbf{n}} = (\frac{i_1}{n}, \frac{i_2}{n}, \ldots \frac{i_N}{n})\right)$. Certaines difficultés se posent à ce stade. Premièrement la construction de K_1, en général, est un challenge. Le choix de la fonction Ψ sera aussi importante. Le modèle développé dans la section 6.2.2 sera appliqué aux données ScanFish.

6.2.3 Données ScanFish

Les données ScanFish sont issues des campagnes acoustiques internationales réalisées en 2014 " AWA " à bord du navire de recherche Thalassa (Ifremer, Brest) le long des côtes de la Mauritanie et du Sénégal. Des matériaux comme un sondeur scientifique multifréquence et un Scanfish ont été simultanément utilisées pour recueillir, entre autres, l'intensité acoustique (Sv), turbidité, fluorescence, température, salinité, conductivité. Les deux systèmes permettent une acquisition continue de données de qualité à haute résolution spatiale et temporelle sur longue distance. Plusieurs fréquences (38khz, 333khz, 200khz, 120khz, ...) et 12 radiales ont été utilisées. La figure 6.1 donne un exemple de données ScanFish correspondantes à la fréquence 200khz et à la radiale 2

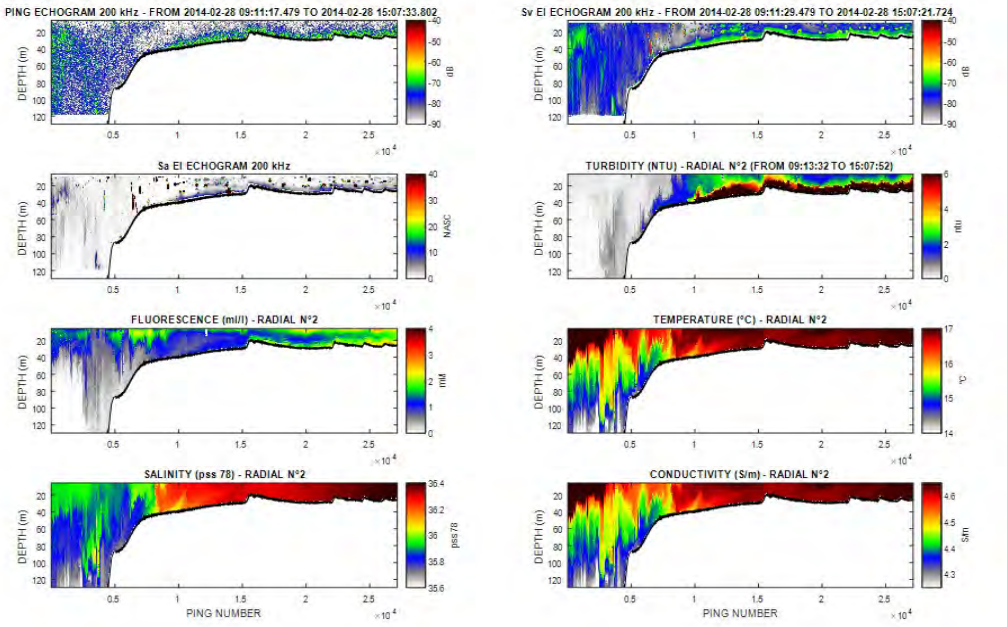

FIGURE 6.1 – Données ScanFish correspondantes à la fréquence 200khz et à la radiale 2.

On s'intéressera à l'intensité acoustique (Sv) qui est susceptible d'être la réponse acoustique d'un organisme marin comme zoo-plancton, micronecton, etc. Ces organismes sont importants pour caractériser l'habitat des petits pélagiques [121]. Nous effectuons un échantillonnage sur 652 sites d'observations (Esu : Echograme sampling unit). Nous disposons de la variable Sv sur tout le profil bathymétrique jusqu'à 120 m de profondeur. La figure 6.2 donne des Esus effectués durant le jour et la nuit.

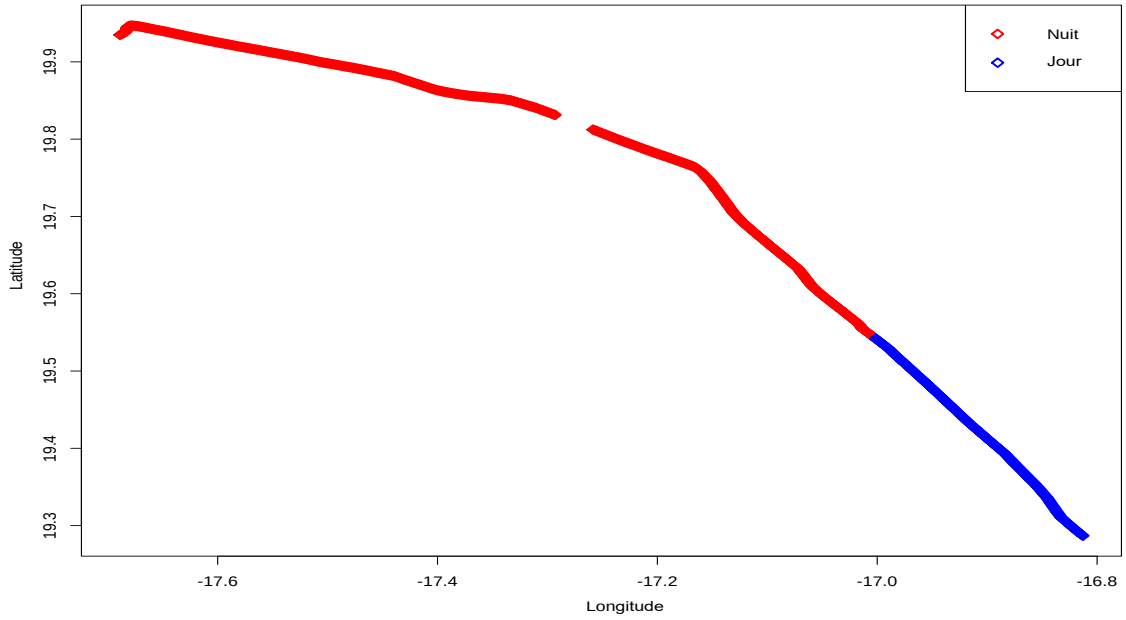

FIGURE 6.2 – Sv correspondantes à la fréquence 200 khz et à la radiale 2.

Une procédure de classification non-supervisée fonctionnelle appliquée au Sv, échantillonnée pendant la journée, montre une variation verticale mais aussi horizontale [29; 104; 224].

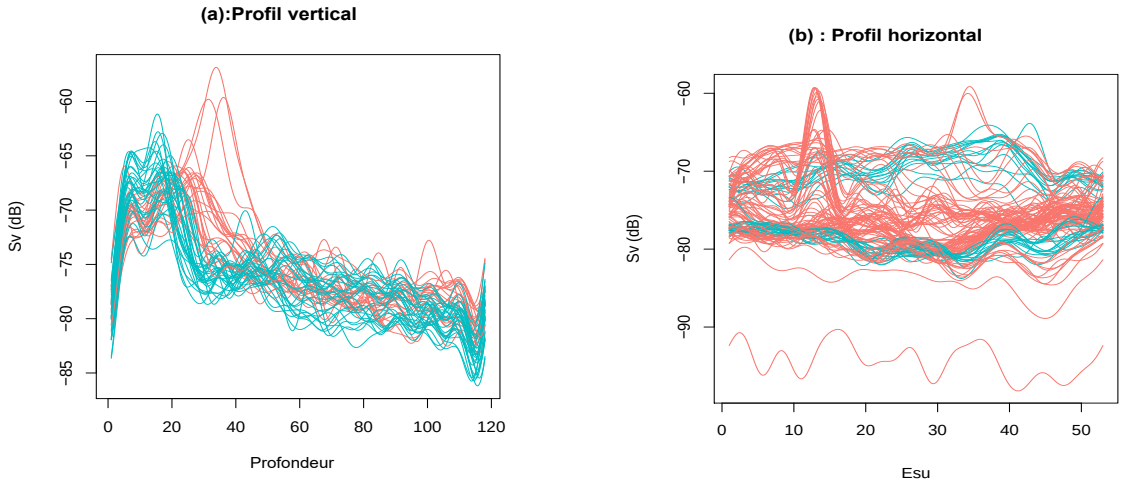

FIGURE 6.3 – Profils vertical (Panel a) et horizontal (Panel b) des intensités acoustiques (Sv).

La figure 6.3 montre deux classes de la variable Sv sur la verticale et sur l'horizontale. La variation des Sv peut être expliquée par les paramètres environnements et aussi par la nature de l'organisme marin qui renvoie le son acoustique. Une méthode de *classification supervisée*, permettant de classifier les organismes marins suivant la supervision des paramètres environnementaux, serait utile. Les méthodes qui seront développées dans la section 6.2.2 pourrait aider à classifier les organismes marins selon leurs intensités acoustiques. C'est à dire si nous observons un Sv on peut décider à priori, sur la base d'une *classification supervisée*, qu'il porte la signature d'un organisme marin donné.

Bibliographie

[1] Ahmedoune Ould Abdi, Aliou Diop, Sophie Dabo-Niang, and Sidi Ali Ould Abdi. Estimation non paramétrique du mode conditionnel dans le cas spatial. *Comptes Rendus Mathematique*, 348(13-14) :815–819, 2010. 13, 20

[2] Haneen Arafat Abu Alfeilat, Ahmad BA Hassanat, Omar Lasassmeh, Ahmad S Tarawneh, Mahmoud Bashir Alhasanat, Hamzeh S Eyal Salman, and VB Surya Prasath. Effects of distance measure choice on k-nearest neighbor classifier performance : A review. *Big data*, 2019. 16

[3] Ana M Aguilera, Manuel Escabias, Cristian Preda, and Gilbert Saporta. Using basis expansions for estimating functional pls regression : applications with chemometric data. *Chemometrics and Intelligent Laboratory Systems*, 104(2) :289–305, 2010. 14

[4] Mohamed Salem Ahmed, Laurence Broze, Sophie Dabo-Niang, and Zied Gharbi. Functional linear spatial autoregressive models, 2017. 15

[5] Mohamed Salem Ahmed, Mamadou Ndiaye, Mohamed Attouch, and Sophie. Dabo-Niang. k-nearest neighbors prediction and classification for spatial data. *preprinted*, 2019. 33, 39

[6] Ahmed Ait-Saïdi, Frederic Ferraty, Rabah Kassa, and Philippe Vieu. Cross-validated estimations in the single-functional index model. *Statistics*, 42(6) :475–494, 2008. 14

[7] Mohammad Al-Bassam and Omar Eidous. Combination of parametric and nonparametric estimators for population abundance using line transect sampling. *Journal of Information and Optimization Sciences*, 39(7) :1449–1462, 2018. 74

[8] FAYE Alioune, SARR Alassane, Malick DIOUF, Modou Thiaw, FALL Jean, Ismaïla NDIAYE, BA Kamarel, and Najih LAZAR. Contribution to the study of the size structure, the length-weight relationship, the condition factor and the sex-ration of shrimp farfantepenaeus notialis (pérez farfante, 1967) in the estuary of sine-saloum (senegal). *American Scientific Research Journal for Engineering, Technology, and Sciences (ASRJETS)*, 14(2) :97–112, 2015. 3

[9] Ibrahim M Almanjahie, Mohammed Kadi Attouch, Zoulikha Kaid, and Hayat Louhab. Robust equivariant non parametric regression estimators for functional ergodic data. *Communications in Statistics-Theory and Methods*, pages 1–17, 2020. 15

[10] Javier Álvarez-Liébana and M Dolores Ruiz-Medina. Functional statistical classification of non-linear dynamics and random surfaces roughness in control systems. *Int. J. Math. Models Methods Appl. Sciences*, 9 :1–20, 2015. 16, 17

[11] Aboubacar Amiri, Christophe Crambes, and Baba Thiam. Recursive estimation of nonparametric regression with functional covariate. *Computational Statistics & Data Analysis*, 69 :154–172, 2014. 15

[12] Aboubacar Amiri, Sophie Dabo-Niang, and Mohamed Yahaya. Nonparametric recursive density estimation for spatial data. *Comptes Rendus Mathematique*, 354(2) :205–210, 2016. 11

[13] Marti J Anderson and Russell B Millar. Spatial variation and effects of habitat on temperate reef fish assemblages in northeastern new zealand. *Journal of Experimental Marine Biology and Ecology*, 305(2) :191–221, 2004. 45

[14] Anestis Antoniadis and Theofanis Sapatinas. Wavelet methods for continuous-time prediction using hilbert-valued autoregressive processes. *Journal of Multivariate Analysis*, 87(1) :133–158, 2003. 13

[15] Yuko Araki, Sadanori Konishi, Shuichi Kawano, and Hidetoshi Matsui. Functional logistic discrimination via regularized basis expansions. *Communications in Statistics -Theory and Methods*, 38(16-17) :2944–2957, 2009. 17

[16] Miguel B Araújo, Richard G Pearson, Wilfried Thuiller, and Markus Erhard. Validation of species–climate impact models under climate change. *Global Change Biology*, 11(9) :1504–1513, 2005. 45

[17] Fatima Zohra Ardjoun, Larbi Ait Hennani, and Ali Laksaci. A recursive kernel estimate of the functional modal regression under ergodic dependence condition. *Journal of Statistical Theory and Practice*, 10(3) :475–496, 2016. 15

[18] Osama H Arif and Omar Eidous. Fourth-order kernel method for simple linear degradation model. *Communications in Statistics-Simulation and Computation*, 47(1) :16–29, 2018. 74

[19] Eleonora Arnone, Laura Azzimonti, Fabio Nobile, and Laura M Sangalli. Modeling spatially dependent functional data via regression with differential regularization. *Journal of Multivariate Analysis*, 170 :275–295, 2019. 15

[20] Eleonora Arnone, Alois Kneip, Fabio Nobile, and Laura M Sangalli. Some first results on the consistency of spatial regression with partial differential equation regularization. Technical report, 2020. 15, 16

[21] Mohammed K Attouch, Abdelkader Gheriballah, and Ali Laksaci. Robust nonparametric estimation for functional spatial regression. In Frédéric Ferraty, editor, *Recent Advances in Functional Data Analysis and Related Topics*, Contributions to Statistics, pages 27–31. Physica-Verlag HD, 2011. 15

[22] Jean-Yves Audibert, Alexandre B Tsybakov, et al. Fast learning rates for plug-in classifiers. *The Annals of statistics*, 35(2) :608–633, 2007. 16

[23] Christopher Mulanda Aura, Rashid Oketch Anam, Safina Musa, and Edward Ndirui Kimani. The length-weight relationship and condition factor (k constant) of the sparidae (dentex marocannus, valenciennes 1830) of malindi, kenya. *Western Indian Ocean Journal of Marine Science*, 12(1) :79–83, 2013. 44

[24] Otilio Avendaño, Iván Velázquez-Abunader, Carlos Fernández-Jardón, Luis Enrique Ángeles-González, Alvaro Hernández-Flores, and Ángel Guerra. Biomass and distribution of the red octopus (octopus maya) in the north-east of the campeche bank. *Journal of the Marine Biological Association of the United Kingdom*, 99(6) :1317–1323, 2019. 10

[25] Kamarel BA. *Dynamique des ressources démersales côtières et aménagement de la pêcherie du poulpe au Sénégal*. PhD thesis, ÉCOLE DOCTORALE : SCIENCES DE LA VIE, DE LA SANTÉ ET DE L'ENVIRONNEMENT FACULTÉ DES SCIENCES ET TECHNIQUES, 2018. 1, 2, 3, 4, 5, 44

[26] Kamarel Ba, Modou Thiaw, Massal Fall, Ndiaga Thiam, Beyah Meissa, Didier Jouffre, Omar Thiom Thiaw, and Didier Gascuel. Long-term fishing impact on the senegalese coastal demersal resources : diagnosing from stock assessment models. *Aquatic Living Resources*, 31 :8, 2018. 1

[27] Akshay Balsubramani, Sanjoy Dasgupta, Yoav Freund, and Shay Moran. An adaptive nearest neighbor rule for classification. *arXiv preprint arXiv :1905.12717*, 2019. 16

[28] Manuel Febrero Bande, Manuel Oviedo de la Fuente, Pedro Galeano, Alicia Nieto, Eduardo Garcia-Portugues, and Maintainer Manuel Oviedo de la Fuente. *Package 'fda. usc'*, 2019. 49

[29] Sanghamitra Bandyopadhyay and Sriparna Saha. *Unsupervised classification : similarity measures, classical and metaheuristic approaches, and applications*. Springer Science & Business Media, 2012. 76

[30] Rafael Banon, David Barros-Garcia, Gonzalo Mucientes, and Alejandro De Carlos. Northernmost records of pagrus auriga (actinopterygii : Perciformes : Sparidae) and pomadasys incisus (actinopterygii : Perciformes : Haemulidae) in the eastern atlantic. *Acta Ichthyologica et Piscatoria*, 44(4), 2014. 44, 45

[31] Rafik Baouche. *Prédiction des Paramètres Physiques des Couches Pétroliferes par Analyse des Réseaux de Neurones et Analyse Faciologique*. PhD thesis, FACULTE DES SCIENCES PHYSIQUES, LABORATOIRE LIMOSE, Université M'hamed Bougara, Boumerdés, 2015. 27

[32] Mariama Dalanda Barry, Martial Laurans, Djiga Thiao, and Didier Gascuel. Diagnostic de l'état d'exploitation de cinq espèces démersales côtières sénégalaises. In *Pêcheries maritimes, écosystèmes et sociétés en Afrique de l'Ouest : un demi siècle de changement. Coll. Rap. Actes du Symposium international DAKAR Juin*, pages 183–194, 2002. 1

[33] Peter L Bartlett and Shahar Mendelson. Empirical minimization. *Probability theory and related fields*, 135(3) :311–334, 2006. 16

[34] ML Bauchot. Check-list of the fishes of the eastern tropical atlantic (clofeta). sparidae in : Jc quero, jc hureau, c karrer, a post and l saldanha jnict-portugal, sei-france, 1990. 44, 45

[35] Séverine Bayle, Pascal Monestiez, and David Nerini. Modèle linéaire de prédiction fonctionnelle sur données environnementales : choix de modélisation. *Journal de la Société Française de Statistique*, 155(2) :121–137, 2014. 6

[36] Karim Benhenni, Sonia Hedli-Griche, Mustapha Rachdi, and Philippe Vieu. Consistency of the regression estimator with functional data under long memory conditions. *Statistics & Probability Letters*, 78(8) :1043–1049, 2008. 15

[37] Bernard Bercu, Sami Capderou, and Gilles Durrieu. Nonparametric recursive estimation of the derivative of the regression function with application to sea shores water quality. *Statistical Inference for Stochastic Processes*, 22(1) :17–40, 2019. 15

[38] Mara S Bernardi, Laura M Sangalli, Gabriele Mazza, and James O Ramsay. A penalized regression model for spatial functional data with application to the analysis of the production of waste in venice province. *Stochastic Environmental Research and Risk Assessment*, 31(1) :23–38, 2017. 15

[39] G Biau. Spatial kernel density estimation. *Mathematical methods of Statistics*, 12(4) :371–390, 2003. 11

[40] Gérard Biau, Florentina Bunea, and Marten H Wegkamp. Functional classification in hilbert spaces. *IEEE Transactions on Information Theory*, 51(6) :2163–2172, 2005. 16

[41] Gérard Biau and Benoit Cadre. Nonparametric spatial prediction. *Statistical Inference for Stochastic Processes*, 7(3) :327–349, 2004. 11, 12, 20, 30, 34

[42] Gérard Biau and Luc Devroye. *Lectures on the nearest neighbor method*. Springer, 2015. 16, 20, 24

[43] Peter J Bickel, Ya'acov Ritov, et al. Nonparametric estimators which can be" plugged-in". *The Annals of Statistics*, 31(4) :1033–1053, 2003. 16

[44] Alberto Bisin and Andrea Moro. Learning epidemiology by doing : The empirical implications of a spatial sir model with behavioral responses. 2020. 10

[45] Gilles Blanchard, Olivier Bousquet, Pascal Massart, et al. Statistical performance of support vector machines. *The Annals of Statistics*, 36(2) :489–531, 2008. 16

[46] JHS Blaxter and JR Hunter. The biology of the clupeoid fishes. In *Advances in marine biology*, volume 20, pages 1–223. Elsevier, 1982. 5

[47] D. Bosq. *Nonparametric Statistics for Stochastic Processes : Estimation and prediction*, volume 110 of *Lecture Notes in Statist.* Springer-Verlag, New York, 2nd edition, 1998. 31

[48] Denis Bosq. Modelization, nonparametric estimation and prediction for continuous time processes. In *Nonparametric functional estimation and related topics*, pages 509–529. Springer, 1991. 13

[49] Stéphane Boucheron, Olivier Bousquet, and Gábor Lugosi. Theory of classification : A survey of some recent advances. *ESAIM : probability and statistics*, 9 :323–375, 2005. 16

[50] Salim Bouzebda and Yousri Slaoui. Nonparametric recursive method for kernel-type function estimators for spatial data. *Statistics & Probability Letters*, 139 :103–114, 2018. 11, 13

[51] Salim Bouzebda and Yousri Slaoui. Large and moderate deviation principles for recursive kernel estimators of a regression function for spatial data defined by stochastic approximation method. *Statistics & Probability Letters*, 151 :17–28, 2019. 13

[52] Salim Bouzebda and Yousri Slaoui. Bandwidth selector for nonparametric recursive density estimation for spatial data defined by stochastic approximation method. *Communications in Statistics-Theory and Methods*, 49(12) :2942–2963, 2020. 11

[53] Salim Bouzebda and Yousri Slaoui. Large and moderate deviation principles for recursive kernel estimators for spatial data. *Journal of Stochastic Analysis*, 1(1) :7, 2020. 11

[54] Antonio Bracale, Pierluigi Caramia, Pasquale De Falco, and Tao Hong. Multivariate quantile regression for short-term probabilistic load forecasting. *IEEE Transactions on Power Systems*, 35(1) :628–638, 2019. 13

[55] Stephanie J Brodie, James T Thorson, Gemma Carroll, Elliott L Hazen, Steven Bograd, Melissa A Haltuch, Kirstin K Holsman, Stan Kotwicki, Jameal F Samhouri, Ellen Willis-Norton, et al. Trade-offs in covariate selection for species distribution models : a methodological comparison. *Ecography*, 43(1) :11–24, 2020. 42

[56] ST Buckland, DR Aanderson, and KP Burnham. Distance sampling. *Estimating abundance of biological populations. Chapman &l-lall, London*, 1993. 74

[57] Stephen T Buckland. Fitting density functions with polynomials. *Journal of the Royal Statistical Society : Series C (Applied Statistics)*, 41(1) :63–76, 1992. 75

[58] Stephen T Buckland and David L Miller. 6 distance sampling. *Quantitative Analyses in Wildlife Science*, page 97, 2019. 74

[59] Florent Burba, Frédéric Ferraty, and Philippe Vieu. k-nearest neighbour method in functional nonparametric regression. *Journal of Nonparametric Statistics*, 21(4) :453–469, 2009. 21, I

[60] KP Burnham and DR Anderson. Mathematical models for nonparametric inferences from line transect data. *Biometrics*, 32(2) :325–336, 1976. 74

[61] Jose M Cadenas, M Carmen Garrido, Raquel Martínez, Enrique Muñoz, and Piero P Bonissone. A fuzzy k-nearest neighbor classifier to deal with imperfect data. *Soft Computing*, 22(10) :3313–3330, 2018. 16

[62] Jef Caers. *Petroleum geostatistics*. Society of Petroleum Engineers Richardson, 2005. 10

[63] Jie Cao, James T Thorson, André E Punt, and Cody Szuwalski. A novel spatiotemporal stock assessment framework to better address fine-scale species distributions : Development and simulation testing. *Fish and Fisheries*. 41

[64] Michel Carbon, Christian Francq, and Lanh Tat Tran. Kernel regression estimation for random fields. *Journal of Statistical Planning and Inference*, 137(3) :778–798, 2007. 11, 12, 20

[65] Michel Carbon, Marc Hallin, and Lanh Tat Tran. Kernel density estimation for random fields : the l 1 theory. *Journal of nonparametric Statistics*, 6(2-3) :157–170, 1996. 11

[66] Michel Carbon, Lanh Tat Tran, and Berlin Wu. Kernel density estimation for random fields (density estimation for random fields). *Statistics & Probability Letters*, 36(2) :115–125, 1997. 11, 30, 31, VI, XI

[67] Monica Cardarilli, Mara Lombardi, and Angelo Corazza. Landslide risk management through spatial analysis and stochastic prediction for territorial resilience evaluation. *International Journal of Safety and Security Engineering*, 9(2) :109–120, 2019. 10

[68] Hervé Cardot, Frédéric Ferraty, and Pascal Sarda. Functional linear model. *Statistics & Probability Letters*, 45(1) :11–22, 1999. 13, 14

[69] Hervé Cardot, Frédéric Ferraty, and Pascal Sarda. Spline estimators for the functional linear model. *Statistica Sinica*, pages 571–591, 2003. 13, 14

[70] Hervé Cardot and Pacal Sarda. Estimation in generalized linear models for functional data via penalized likelihood. *Journal of Multivariate Analysis*, 92(1) :24–41, 2005. 13, 14

[71] KE Carpenter and N De Angelis. The living marine resources of the eastern central atlantic. vol. 4, bony fishes part 2 (perciformes to tetraodontiformes) and sea turtles. *FAO Species Identification Guide for Fishery Purposes, FAO, Rome*, 2016. 44

[72] Raymond J Carroll, Arnab Maity, Enno Mammen, and Kyusang Yu. Nonparametric additive regression for repeatedly measured data. *Biometrika*, 96(2) :383–398, 2009. 14

[73] P Cayre and A Fontana. [deep sea stocks [shrimps (parapenaeus longirostris, aristeus viridens, plesiopenaeus edwardsianus), sea bream (dentex angolensis), hakes (merluccius polli), squids (loligo sp., sepia officinalis), crabs (geryon quinquedens)]]. *Travaux et Documents de l'ORSTOM (France)*, 1981. 44

[74] Jacques Chabanne. *Le peuplement des fonds durs et sableux du plateau continental sénégambien : étude de sa pêcherie chalutière : biologie et dynamique d'une espèce caractéristique : le rouget (Pseudupeneus prayensis)*. Paris (France) Eds de L'ORSTOM, 1987. 5

[75] Christian Champagnat and François Domain. Migrations des poissons démersaux le long des côtes ouest-africaines de 10 à 24 de latitude nord. *Cahiers ORSTOM. Série Océanographie*, 16(3-4) :239–261, 1978. 2

[76] Pierre Chavance, M Ba, Didier Gascuel, JM Vakily, and D Pauly. *Pêcheries maritimes, écosystèmes et sociétés en Afrique de l'Ouest : un demi-siècle de changement : Actes du symposium international, Dakar, Sénégal, 24-28 juin 2002*. Institut de Recherche pour le Développement (IRD) ; Commission européenne, 2004. 2

[77] Dong Chen, Peter Hall, Hans-Georg Müller, et al. Single and multiple index functional regression models with nonparametric link. *The Annals of Statistics*, 39(3) :1720–1747, 2011. 14

[78] Shuai Chen, Hongliang Huang, and Peng Zhang. Research on antarctic krill resources survey and assessment methods. In *IOP Conference Series : Earth and Environmental Science*, volume 170, page 022089. IOP Publishing, 2018. 10

[79] Song Xi Chen. Studying school size effects in line transect sampling using the kernel method. *Biometrics*, pages 1283–1294, 1996. 75

[80] Xing Chen et al. Spatial nonparametric regression estimation : Non-isotropic case. *Acta Mathematicae Applicatae Sinica*, 18(4) :641–656, 2002. 12

[81] Yun Chen and Hui Yang. Sparse modeling and recursive prediction of space–time dynamics in stochastic sensor networks. *IEEE Transactions on Automation Science and Engineering*, 13(1) :215–226, 2015. 13

[82] Wen-Hao Chiang, Xueying Liu, and George Mohler. Hawkes process modeling of covid-19 with mobility leading indicators and spatial covariates. *medRxiv*, 2020. 10

[83] Jean-Paul Chilés and Pierre Delfiner. *Geostatistics : Modeling Spatial Uncertainty*. Wiley Series in Applied Probability and Statistics. John Wiley & Sons, Inc, 1999. 11

[84] Joydeep Chowdhury, Probal Chaudhuri, et al. Nonparametric depth and quantile regression for functional data. *Bernoulli*, 25(1) :395–423, 2019. 20

[85] Francesco Colloca, Valerio Bartolino, Giovanna Jona Lasinio, Luigi Maiorano, Paolo Sartor, and Giandomenico Ardizzone. Identifying fish nurseries using density and persistence measures. *Marine Ecology Progress Series*, 381 :287–296, 2009. 10

[86] Gbbabb Collomb. Jfon parametric time series analysis and prediction : uniform almost sure convergence of the window and jt-nn autoregression estimates. *Statistics : A Journal of Theoretical and Applied Statistics*, 16(2) :297–307, 1985. 11

[87] Gérard Collomb. Estimation de la régression par la méthode des k points les plus proches avec noyau : quelques propriétés de convergence ponctuelle. *Statistique non Paramétrique Asymptotique*, pages 159–175, 1980. 17, 22, I, II

[88] Thomas Cover and Peter Hart. Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1) :21–27, 1967. 16

[89] BR Crain, KP Burnham, DR Anderson, and JL Lake. Nonparametric estimation of population density for line transect sampling using fourier series. *Biometrical Journal*, 21(8) :731–748, 1979. 74

[90] Noel Cressie and Christopher K Wikle. *Statistics for spatio-temporal data*. John Wiley & Sons, 2015. 19

[91] Noel A C Cressie. *Statistics for Spatial Data*, volume 110 of *Wiley Series in Probability and Statistics*. Wiley-Interscience, revised edition, 1993. 11, 18, 42

[92] Juan A Cuesta-Albertos, Manuel Febrero-Bande, and M Oviedo de la Fuente. The dd^{G}-classifier in the functional setting. *Test*, 26(1) :119–142, 2017. 16, 17, 49

[93] Antonio Cuevas, Manuel Febrero, and Ricardo Fraiman. Robust estimation and classification for functional data via projection-based depth notions. *Computational Statistics*, 22(3) :481–496, 2007. 16

[94] Philippe Cury and Claude Roy. Migration saisonnière du thiof (epinephelus aeneus) au sénégal : influence des upwellings. *Oceanologica acta*, 11(1) :25–36, 1988. 5

[95] Jean-Jerôme Da Costa, Fabien Chainet, Benoît Celse, Marion Lacoue-NeÌĂgre, Cyril Ruckebusch, Noémie Caillol, and Didier Espinat. Kriging modeling to predict viscosity index of base oils. *Energy & fuels*, 32(2) :2588–2597, 2018. 10

[96] S. Dabo-Niang, L. Hamdad, C. Ternynck, and Anne-Fran c coise. Yao. A kernel spatial density estimation allowing for the analysis of spatial clustering : application to Monsoon Asia Drought Atlas data. *Stoch. Environ. Res. Risk Assess*, 28(8) :2075–2099, 2014. 28

[97] S Dabo-Niang and N Rhomari. Kernel regression estimation when the regressor takes values in metric space. *COMPTES RENDUS MATHEMATIQUE*, 336(1) :75–80, 2003. 15

[98] S. Dabo-Niang, C. Ternynck, and Anne-Francoise. Yao. Nonparametric prediction of spatial multivariate data. *Nonparametric Statistics., 2., 428-458*, 2016. 6, 7, 11, 13, 17, 18, 20, 21, 22, 24, 25, 26, 39, 45, 73

[99] Sophie Dabo-Niang, Leila Hamdad, Camille Ternynck, and Anne-Françoise Yao. A kernel spatial density estimation allowing for the analysis of spatial clustering. application to monsoon asia drought atlas data. *Stochastic environmental research and risk assessment*, 28(8) :2075–2099, 2014. 12

[100] Sophie Dabo-Niang, Zoulikha Kaid, and Ali Laksaci. Spatial conditional quantile regression : Weak consistency of a kernel estimate. *Rev. Roumaine Math. Pures Appl.*, 57(4) :311–339, 2012. 20

[101] Sophie Dabo-Niang, Mustapha Rachdi, and Anne-Francoise Yao. Kernel regression estimation for spatial functional random variables. *Far East Journal of Theoretical Statistics*, 37(2) :77–113, 2011. 15, 30, 32

[102] Sophie Dabo-Niang, Camille Ternynck, and Anne-Françoise Yao. Nonparametric prediction of spatial multivariate data. *Journal of Nonparametric Statistics*, 28(2) :428–458, 2016. VIII, IX, X

[103] Sophie Dabo-Niang and Baba Thiam. Robust quantile estimation and prediction for spatial processes. *Statistics & probability letters*, 80(17-18) :1447–1458, 2010. 13

[104] Sophie Dabo-Niang, Anne-Fran c coise Yao, Laura Pischedda, Philippe Cuny, and Franck Gilbert. Spatial mode estimation for functional random fields with application to bioturbation problem. *Stochastic Environmental Research and Risk Assessment*, 24(4) :487–497, 2010. 6, 27, 76

[105] Sophie. Dabo-Niang and Anne-Francoise. Yao. Kernel regression estimation for continuous spatial processes. *Mathematical Methods of Statistics*, 16(4) :298–317, 2007. 11, 12, 20, 34

[106] Sophie Dabo-Niang and Anne-Francoise Yao. Kernel spatial density estimation in infinite dimension space. *Metrika*, 76(1) :19–52, 2013. 11, 30

[107] Xiaowen Dai, Erqian Li, and Maozai Tian. Quantile regression for varying coefficient spatial error models. *Communications in Statistics-Theory and Methods*, pages 1–16, 2019. 13

[108] Leon Danon, Ellen Brooks-Pollock, Mick Bailey, and Matt J Keeling. A spatial model of covid-19 transmission in england and wales : early spread and peak timing. *MedRxiv*, 2020. 10

[109] Tilman M Davies and Andrew B Lawson. An evaluation of likelihood-based bandwidth selectors for spatial and spatiotemporal kernel estimates. *Journal of Statistical Computation and Simulation*, 89(7) :1131–1152, 2019. 11

[110] Manuel Oviedo de la Fuente. *Advances in functional regression and classification models*. PhD thesis, Universidade de Santiago de Compostela, 2019. 16

[111] Allan J Debertin. Estimating the biomass of a mixed species complex using hydroacoustics and catch data from the bay of fundy and scotian shelf summer ecosystem survey. *Canadian Journal of Fisheries and Aquatic Sciences*, (999) :1–16, 2020. 10

[112] Laurent Delsol. Advances on asymptotic normality in non-parametric functional time series analysis. *Statistics*, 43(1) :13–33, 2009. 11

[113] Hervé DEMARCQ and Valérie FAURE. Coastal upwelling and associated retention indices derived from satellite sst. application to octopus vulgaris recruitment. *Oceanologica acta*, 23(4) :391–408, 2000. 5

[114] Itaf Deme-Gningue, Claude Roy, and Diafara Touré. Variabilité spatio-temporelle de la température, des nitrates et de la chlorophylle devant les côtes du sénégal. *CRODT*, 1990. 1

[115] Chandrakant M Deo. A note on empirical processes of strong-mixing sequences. *The Annals of Probability*, pages 870–875, 1973. III

[116] Luc Devroye. A universal k-nearest neighbor procedure in discrimination. *Nearest Neighbor (NN) Norms : NN Pattern Classification Techniques*, pages 101–106, 1978. 16

[117] Luc Devroye et al. On the asymptotic probability of error in nonparametric discrimination. *The Annals of Statistics*, 9(6) :1320–1327, 1981. 16

[118] Luc Devroye, Laszlo Gyorfi, Adam Krzyzak, and Gabor Lugosi. On the strong universal consistency of nearest neighbor regression function estimates. *The Annals of Statistics*, 1994. 16, 24, 33

[119] Luc Devroye, László Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media, 2013. 16

[120] Luc Devroye and Terry J Wagner. 8 nearest neighbor methods in discrimination. *Handbook of Statistics*, 1982. 16, 24, 33

[121] Ndague Diogoul, Patrice Brehmer, Yannick Perrot, Maik Tiedemann, Salaheddine El Ayoubi, Anne Mouget, Chloe Migayrou, Oumar Sadio, Abdoulaye Sarre, et al. Fine-scale vertical structure of sound-scattering layers over an east border upwelling system and its relationship to pelagic habitat characteristics. *Ocean Science*, 16(1) :65–81, 2020. 76

[122] JN Diouf, BS Toguebaye, et al. Study of some marine fish coccidia of the genus eimeria schneider, 1815 (apicomplexa, coccidia) from senegal coasts. *Acta Protozoologica*, 33(4) :239–250, 1994. 44

[123] PS Diouf and I Deme-Gningue. Bio-écologie et structure des peuplements de poissons de l'estuaire du sine-saloum. *Rapp. Scient. CRODT/ORSTOM*, 1992. 5

[124] F Domain. *Contribution à la connaissance de l'écologie des espèces démersales du plateau continental sénégalo-mauritanien : les ressources démersales dans le contexte du golfe de Guinée*. PhD thesis, Thèse de doctorat, université Paris VI, 1980. 2, 3, 4, 5, 44

[125] François Domain. *Carte sédimentologique du plateau continental sénégambien : extension à une partie du plateau continental de la Mauritanie et de la Guinée Bissau*. 1977. 1, 2

[126] François Domain. Potentialités comparées des différentes zones de pêche d'espèces démersales du golfe de guinée ($19^\circ n$ à $6^\circ s$). *CRODT*, 1978. 1

[127] François Domain, Didier Jouffre, and Alain Caverivière. Growth of octopus vulgaris from tagging in senegalese waters. *Journal of the Marine Biological Association of the United Kingdom*, 80(4) :699–705, 2000. 3

[128] Maik Döring, László Györfi, and Harro Walk. Rate of convergence of k-nearest-neighbor classification rule. *The Journal of Machine Learning Research*, 18(1) :8485–8500, 2017. 16

[129] Paul Doukhan. *Mixing*, volume 85 of *Lecture Notes in Statistics*. Springer-Verlag, New York, 1994. Properties and examples. 21, 22, 30

[130] Naihua Duan. Smearing estimate : a nonparametric retransformation method. *Journal of the American Statistical Association*, 78(383) :605–610, 1983. 11

[131] Omar Eidous and Fahid Al-Eibood. A bias-corrected histogram estimator for line transect sampling. *Communications in Statistics-Theory and Methods*, 47(15) :3675–3686, 2018. 74

[132] M. El Machkouri. Nonparametric regression estimation for random fields in a fixed-design. *Stat. Inference Stoch. Process.*, 10(1) :29–47, 2007. 29

[133] M El Machkouri, X Fan, and L Reding. On the nadaraya–watson kernel regression estimator for irregularly spaced spatial data. *Journal of Statistical Planning and Inference*, 2019. 12

[134] Mohamed El Machkouri. Asymptotic normality of the Parzen–Rosenblatt density estimator for strongly mixing random fields. *Statistical Inference for Stochastic Processes*, 14(1) :73–84, 2011. 11, 29

[135] Mohamed El Machkouri and Radu Stoica. Asymptotic normality of kernel estimates in regression model for random fields. *Nonparametric Statistics, 22, 955-971*, 2010. 20, 29

[136] M Escabias, AM Aguilera, and MJ Valderrama. Principal component estimation of functional logistic regression : discussion of two different approaches. *Journal of Nonparametric Statistics*, 16(3-4) :365–384, 2004. 14

[137] M Escabias, AM Aguilera, and MJ Valderrama. Modeling environmental data by functional principal component logistic regression. *Environmetrics : The official journal of the International Environmetrics Society*, 16(1) :95–107, 2005. 27

[138] Manuel Escabias, Ana M Aguilera, and Mariano J Valderrama. Functional pls logit regression model. *Computational Statistics & Data Analysis*, 51(10) :4891–4902, 2007. 14

[139] Massal Fall et al. Pêcherie démersale côtière au sénégal. essai de modélisation de la dynamique de l'exploitation des stocks. B.S. thesis, 2009. 1

[140] Massal Fall and Farokh Niass. Analyse de données de campagnes scientifiques relatives à la brotule brotula barbata et aux saintpierre zeus faber mauritanicus et zenopsis conchifer des côtes sénégalaises. *Journal of Applied Biosciences*, 83(1) :7506–7519, 2014. 1

[141] Yingying Fan, Natasha Foutz, Gareth M James, Wolfgang Jank, et al. Functional response additive model estimation with online virtual stock markets. *The Annals of Applied Statistics*, 8(4) :2435–2460, 2014. 14, 15

[142] István Fazekas and Alexey Chuprunov. Asymptotic normality of kernel type density estimators for random fields. *Statistical inference for stochastic processes*, 9(2) :161–178, 2006. 11

[143] Manuel Febrero-Bande and Wenceslao González-Manteiga. Generalized additive models for functional data. *Test*, 22(2) :278–292, 2013. 14, 15, 16

[144] Frédéric Ferraty and Philippe Vieu. The functional nonparametric model and application to spectrometric data. *Computational Statistics*, 17(4) :545–564, 2002. 15

[145] Frédéric Ferraty and Philippe Vieu. Curves discrimination : a nonparametric functional approach. *Computational Statistics & Data Analysis*, 44(1-2) :161–173, 2003. 17

[146] Frédéric Ferraty and Philippe Vieu. *Nonparametric functional data analysis : theory and practice*. Springer Science & Business Media, 2006. 13, 15, 16, 17, 30, 33, 35, 39, 40, 49

[147] Frédéric Ferraty and Philippe Vieu. Additive prediction and boosting for functional data. *Computational Statistics & Data Analysis*, 53(4) :1400–1413, 2009. 14

[148] Gentile Francesco Ficetola and Diego Rubolini. Climate affects global patterns of covid-19 early outbreak dynamics. *medRxiv*, 2020. 10

[149] Evelyn Fix and Joseph L Hodges Jr. Discriminatory analysis-nonparametric discrimination : consistency properties. Technical report, California Univ Berkeley, 1951. 16

[150] Mario Francisco-Fernández and Jean D. Opsomer. Smoothing parameter selection methods for non-parametric regression with spatially correlated errors. *Canad. J. Statist.*, 33(2) :279–295, 2005. 28

[151] Mario Francisco-Fernández, Alejandro Quintela-del Río, and Rubén Fernández-Casal. Nonparametric methods for spatial regression. an application to seismic events. *Environmetrics*, 23(1) :85–93, 2012. 28

[152] Karen Fuchs, Jan Gertheiss, and Gerhard Tutz. Nearest neighbor ensembles for functional data with interpretable feature selection. *Chemometrics and Intelligent Laboratory Systems*, 146 :186–197, 2015. 17

[153] Karen Fuchs, Wolfgang Pößnecker, and Gerhard Tutz. Classification of functional data with k-nearest-neighbor ensembles by fitting constrained multinomial logit models. *arXiv preprint arXiv :1612.04710*, 2016. 17

[154] Keinosuke Fukunaga and L Hostetler. Optimization of k nearest neighbor density estimates. *IEEE Transactions on Information Theory*, 19(3) :320–326, 1973. 11

[155] Marteau Gadat, Klein. Classification with the nearest neighbor rule in general finite dimensional spaces. *The Annals of Statistics*, 44(3) :982–1009, 2016. 16

[156] Sébastien Gadat, Sébastien Gerchinovitz, Clément Marteau, et al. Optimal functional supervised classification with separation condition. *Bernoulli*, 26(3) :1797–1831, 2020. 16

[157] Sébastien Gadat, Thierry Klein, and Clément Marteau. Classification with the nearest neighbor rule in general finite dimensional spaces : necessary and sufficient conditions. *arXiv preprint arXiv :1411.0894*, 2014. 16

[158] Carlo Gaetan and Xavier Guyon. *Modélisation et statistique spatiales*, volume 63. Springer, 2008. 11

[159] Antonio F Castro Gámez, José Miguel Rodríguez Maroto, and Iñaki Vadillo Pérez. Quantification of methane emissions in a mediterranean landfill (southern spain). a combination of flux chambers and geostatistical methods. *Waste Management*, 87 :937–946, 2019. 10

[160] Isabel García-Barón, Matthieu Authier, Ainhoa Caballero, José A Vázquez, M Begoña Santos, José Luis Murcia, and Maite Louzao. Modelling the spatial abundance of a migratory predator : A call for transboundary marine protected areas. *Diversity and Distributions*, 25(3) :346–360, 2019. 42

[161] Eva García-Isarch and Isabel Muñoz. Biodiversity and biogeography of decapods crustaceans in the canary current large marine ecosystem. 2015. 3

[162] Beth Gardner, Patrick J Sullivan, Stephen J Morreale, and Sheryan P Epperly. Spatial and temporal statistical analysis of bycatch data : patterns of sea turtle bycatch in the north atlantic. *Canadian Journal of Fisheries and Aquatic Sciences*, 65(11) :2461–2470, 2008. 18, 42

[163] Didier Gascuel. Cinquante ans d'évolution des captures et biomasses dans l'atlantique centre-est : analyse par les spectres trophiques de captures et de biomasses. In *Pêcheries maritimes, écosystèmes et sociétés en Afrique de l'Ouest : un demi-siècle de changement. Actes du symposium Dakar (Senegal), Luxembourg, Office des publications officielles des comm. européennes*, 2004. 1

[164] Theo Gasser and Alois Kneip. Searching for structure in curve samples. *Journal of the american statistical association*, 90(432) :1179–1188, 1995. 13

[165] Theo Gasser, Hans-Georg Muller, Walter Kohler, Luciano Molinari, and Andrea Prader. Nonparametric regression analysis of growth curves. *The Annals of Statistics*, pages 210–229, 1984. 13

[166] Sven Gastauer, Ben Scoulding, and Miles Parsons. Estimates of variability of goldband snapper target strength and biomass in three fishing regions within the northern demersal scalefish fishery (western australia). *Fisheries Research*, 193 :250–262, 2017. 10

[167] Marino Gatto, Enrico Bertuzzo, Lorenzo Mari, Stefano Miccoli, Luca Carraro, Renato Casagrandi, and Andrea Rinaldo. Spread and dynamics of the covid-19 epidemic in italy : Effects of emergency containment measures. *Proceedings of the National Academy of Sciences*, 117(19) :10484–10491, 2020. 10

[168] R Giraldo, Pedro Delicado, and Jorge Mateu. Ordinary kriging for function-valued spatial data. *Environmental and Ecological Statistics*, 18(3) :411–426, 2011. 27

[169] Ramón Giraldo, Sophie Dabo-Niang, and Sergio Martinez. Statistical modeling of spatial big data : An approach from a functional data analysis perspective. *Statistics & Probability Letters*, 136 :126–129, 2018. 15

[170] Daniel R Goethel, Terrance J Quinn, and Steven X Cadrin. Incorporating spatial structure in stock assessment : movement modeling in marine fish population dynamics. *Reviews in Fisheries Science*, 19(2) :119–136, 2011. 41

[171] Aldo Goia. A functional linear model for time series prediction with exogenous variables. *Statistics & Probability Letters*, 82(5) :1005–1011, 2012. 14

[172] AD Gordon. Classification, chapman & hall. *CRC, London*, 1999. 16

[173] Jianping Gou, Hongxing Ma, Weihua Ou, Shaoning Zeng, Yunbo Rao, and Hebiao Yang. A generalized mean distance-based k-nearest neighbor classifier. *Expert Systems with Applications*, 115 :356–372, 2019. 16

[174] Jianping Gou, Wenmo Qiu, Zhang Yi, Xiangjun Shen, Yongzhao Zhan, and Weihua Ou. Locality constrained representation-based k-nearest neighbor classification. *Knowledge-Based Systems*, 167 :38–52, 2019. 16

[175] Jianping Gou, Taisong Xiong, and Yin Kuang. A novel weighted voting for k-nearest neighbor rule. *JCP*, 6(5) :833–840, 2011. 16

[176] Ulf Grenander. Stochastic processes and statistical inference. *Arkiv för matematik*, 1(3) :195–277, 1950. 13

[177] Nilgun Guler Bayazit and Ulug Bayazit. Fuzzy k-nn classification with weights modified by most informative neighbors of nearest neighbors. *Journal of Intelligent & Fuzzy Systems*, 36(6) :6717–6729, 2019. 16

[178] Hyukjun Gweon, Matthias Schonlau, and Stefan H Steiner. The k conditional nearest neighbor algorithm for classification and class probability estimation. *PeerJ Computer Science*, 5 :e194, 2019. 16

[179] L Gyorfi. The rate of convergence of k_n-nn regression estimates and classification rules (corresp.). *IEEE Transactions on Information Theory*, 27(3) :362–364, 1981. 16

[180] L Devroye L Gyorfi, Gabor Lugosi, and L Devroye. A probabilistic theory of pattern recognition, 1996. 24

[181] László Gyorfi. On the rate of convergence of nearest neighbor rules (corresp.). *IEEE Transactions on Information Theory*, 24(4) :509–512, 1978. 16

[182] László Györfi, Michael Kohler, Adam Krzyzak, and Harro Walk. *A distribution-free theory of nonparametric regression*. Springer Science & Business Media, 2006. 16

[183] Peter Hall, Joel L Horowitz, et al. Methodology and convergence rates for functional linear regression. *The Annals of Statistics*, 35(1) :70–91, 2007. 14

[184] Marc Hallin, Zudi Lu, and Lanh T Tran. Local linear spatial regression. *The Annals of Statistics*, 32(6) :2469–2500, 2004. 12, 20, 30

[185] Marc Hallin, Zudi Lu, Lanh Tat Tran, et al. Density estimation for spatial linear processes. *Bernoulli*, 7(4) :657–668, 2001. 11

[186] Marc Hallin, Zudi Lu, and Keming Yu. Local linear spatial quantile regression. *Bernoulli*, 15(3) :659–686, 2009. 20

[187] Miki Hamada and Fugo Takasu. Equilibrium properties of the spatial sis model as a point pattern dynamics-how is infection distributed over space ? *Journal of theoretical biology*, 468 :12–26, 2019. 10

[188] David J Hand. *Construction and assessment of classification rules*. Wiley, 1997. 16

[189] T Hastie and R Tibshirani. Generalized additive models statistical science. 1986. 14

[190] Trevor Hastie and Robert Tibshirani. Discriminant adaptive nearest neighbor classification and regression. *Advances in Neural Information Processing Systems*, 1996. 16, 24

[191] SELINA S Heppell, LARRY B Crowder, and TODD R Menzel. Life table analysis of long-lived marine species with implications for conservation and management. In *American Fisheries Society Symposium*, volume 23, pages 137–148, 1999. 18, 42

[192] Consuelo Hermosilla, Francisco Rocha, and Vasilis D Valavanis. Assessing octopus vulgaris distribution using presence-only model methods. *Hydrobiologia*, 670(1) :35–47, 2011. 44

[193] Alexander Hohl and Peilin Chen. Spatiotemporal simulation : local ripley's k function parameterizes adaptive kernel density estimation. In *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on GeoSpatial Simulation*, pages 16–23, 2019. 11

[194] Zonghui Hu, Naisyin Wang, and Raymond J Carroll. Profile-kernel versus backfitting in the partially linear models for longitudinal/clustered data. *Biometrika*, 91(2) :251–262, 2004. 14

[195] Rui Huang, Miao Liu, and Yongmei Ding. Spatial-temporal distribution of covid-19 in china and its prediction : A data-driven modeling analysis. *The Journal of Infection in Developing Countries*, 14(03) :246–253, 2020. 10

[196] Tingting Huang, Gilbert Saporta, Huiwen Wang, and Shanshan Wang. A robust spatial autoregressive scalar-on-function regression with t-distribution. *Advances in Data Analysis and Classification*, pages 1–25, 2020. 15

[197] I. A. Ibragimov and Yu. V. Linnik. *Independent and stationary sequences of random variables*. Wolters-Noordhoff Publishing, Groningen, 1971. With a supplementary chapter by I. A. Ibragimov and V. V. Petrov, Translation from the Russian edited by J. F. C. Kingman. III

[198] Ousseini Zakaria Ibrahim, Abdourahamane Tankari Dan-Badjo, Yadji Guero, Farida Maissoro Malan Idi, Cyril Feidt, Thibault Sterckeman, and Guillaume Echevarria. Distribution spatiale des éléments traces métalliques dans les sols de la zone aurifère de komabangou au niger. *International Journal of Biological and Chemical Sciences*, 13(1) :557–573, 2019. 10

[199] Samuel P Iglésias and Pascal Lorance. First record of pagellus bellottii (teleostei : Sparidae) in the bay of biscay, france. *Marine Biodiversity Records*, 9(1) :16, 2016. 44, 45

[200] Rosaria Ignaccolo, Stefania Ghigo, and Stefano Bande. Functional zoning for air quality. *Environmental and ecological statistics*, 20(1) :109–127, 2013. 27

[201] Ndour Ismaïla, Baldé Assiatou, Thiam Ndiaga, Thiaw Modou, Faye Saliou, and Fall Massal. Identification and characterization of critical sites for small pelagic fish in the coastal marine area of senegal, west africa. 1

[202] Mohammad Jalali, Shawgar Karami, and Ahmad Fatehi Marj. On the problem of the spatial distribution delineation of the groundwater quality indicators via multivariate statistical and geostatistical approaches. *Environmental monitoring and assessment*, 191(2) :323, 2019. 10

[203] Gareth M James. Generalized linear models with functional predictors. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 64(3) :411–432, 2002. 14, 16

[204] Peng Jia, Weihua Dong, Shujuan Yang, Zhicheng Zhan, La Tu, and Shengjie Lai. Spatial lifecourse epidemiology and infectious disease research. *Trends in Parasitology*, 36(3) :235–238, 2020. 10

[205] Zhenyu Jiang, Nengxiang Ling, Zudi Lu, Dag Tj, Qiang Zhang, et al. On bandwidth choice for spatial data density estimation. *Journal of the Royal Statistical Society Series B*, 82(3) :817–840, 2020. 11

[206] Didier Jouffre, Gilles Domalain, Alain Caveriviere, and Mamadou Diallo. Typology of the demersal communities off senegal : an approach combining gis and multifactorial analysis. *GIS/Spatial Analyses in Fishery and Aquatic Sciences*, 2 :129–145, 2004. 44

[207] Didier Jouffre and Cheikh A Inejih. Assessing the impact of fisheries on demersal fish assemblages of the mauritanian continental shelf, 1987–1999, using dominance curves. *ICES Journal of Marine Science*, 62(3) :380–383, 2005. 44, 45

[208] Lydia-Zaitri Kara, Ali Laksaci, Mustapha Rachdi, and Philippe Vieu. Data-driven knn estimation in nonparametric functional data analysis. *Journal of Multivariate Analysis*, 153 :176–188, 2017. 16

[209] Stelios Katsanevakis and George Verriopoulos. Abundance of octopus vulgaris on soft sediment. *Scientia Marina*, 68(4) :553–560, 2004. 44

[210] Stelios Katsanevakis and George Verriopoulos. Den ecology of octopus vulgaris cuvier, 1797, on soft sediment : availability and types of shelter. *Scientia Marina*, 68(1) :147–157, 2004. 44

[211] Clay King and Joon Jin Song. Bayesian spatial quantile regression for areal count data, with application on substitute care placements in texas. *Journal of Applied Statistics*, 46(4) :580–597, 2019. 13

[212] Jussi Klemelä. Density estimation with locally identically distributed data and with locally stationary data. *J. Time Ser. Anal.*, 29(1) :125–141, 2008. 29

[213] Kouassi Sylvain Konan, Yao Laurent Alla, Moustapha Diaby, and Konan N'Da. Reproduction et mortalité naturelle des poissons capitaines : Polydactylus quadrifilis (cuvier, 1829), galeoides decadactylus (bloch, 1795) et pentanemus quinquarius (linné, 1758) de la pêcherie artisanale maritime de grand-lahou (côte d'ivoire). *International Journal of Innovation and Applied Studies*, 26(2) :477–485, 2019. 5

[214] KA Koranteng. 14 fish species assemblages on the continental shelf and upper slope off ghana. In *Large Marine Ecosystems*, volume 11, pages 173–187. Elsevier, 2002. 44

[215] S Konan Kouassi, A Kone, CMA Akadje, M Diaby, and K N'Da. Ecologie des poissons capitaines : Polydactylus quadrifilis (cuvier, 1829), galeoides decadactylus (bloch, 1795) et pentanemus quinquarius (linné, 1758) de la pêcherie artisanale maritime de grand-lahou (côte d'ivoire). *Tropicultura*, 31(3), 2013. 45

[216] Nadia L Kudraszow and Philippe Vieu. Uniform consistency of knn regressors for functional variables. *Statistics & Probability Letters*, 83(8) :1863–1870, 2013. I

[217] Randy CS Lai, Hsin-Cheng Huang, Thomas CM Lee, et al. Fixed and random effects selection in nonparametric additive mixed models. *Electronic Journal of Statistics*, 6 :810–842, 2012. 14

[218] M Laurans. Evaluation des ressources halieutiques en afrique de l'ouest : dynamique des populations et variabilité écologique. *These pour l'obtention du Diplôme de docteur de l'École Nationale Supérieure Agronomique de Rennes, mention Halieutique, Rennes, France*, 2005. 3, 4, 5

[219] Gaël Le Croizier, Gauthier Schaal, Regis Gallon, Massal Fall, Fabienne Le Grand, Jean-Marie Munaron, Marie-Laure Rouget, Eric Machu, François Le Loc'H, Raymond Laë, et al. Trophic ecology influence on metal bioaccumulation in marine fish : Inference from stable isotope and fatty acid analyses. *Science of the Total Environment*, 573 :83–95, 2016. 44

[220] Jean-Pierre Lecoutre and Philippe Tassi. *Statistique non paramétrique et robustesse*, volume 2. Economica, 1987. 10

[221] Pierre Leenhardt, Matthew Lauer, Rakamaly Madi Moussa, Sally J Holbrook, Andrew Rassweiler, Russell J Schmitt, and Joachim Claudet. Complexities and uncertainties in transitioning small-scale coral reef fisheries. *Frontiers in Marine Science*, 3 :70, 2016. 10

[222] Riwal Lefort, Ronan Fablet, Laurent Berger, and J-M Boucher. Spatial statistics of objects in 3-d sonar images : application to fisheries acoustics. *IEEE Geoscience and Remote Sensing Letters*, 9(1) :56–59, 2011. 18, 42

[223] Xiaoyan Leng and Hans-Georg Müller. Time ordering of gene coexpression. *Biostatistics*, 7(4) :569–584, 2006. 17

[224] T Leon, G Ayala, M Gaston, and F Mallor. Using mathematical morphology for unsupervised classification of functional data. *Journal of Statistical Computation and Simulation*, 81(8) :1001–1016, 2011. 76

[225] Frank Lhomme. Biologie et dynamique de penaeus (farfante penaeus) notiolis (perez farfante 1967) au sénégal. 1981. 4

[226] Frank Lhomme and Serge Garcia. Biologie et exploitation de la crevette penaide au sénégal. 1984. 3

[227] Jiexiang Li and Lanh Tat Tran. Nonparametric estimation of conditional expectation. *Journal of Statistical Planning and Inference*, 139(2) :164–175, 2009. 12, 20

[228] Linchao Li, Bin Ran, Jiasong Zhu, and Bowen Du. Coupled application of deep learning model and quantile regression for travel time and its interval estimation using data in different dimensions. *Applied Soft Computing*, page 106387, 2020. 13

[229] Shengli Li, Ye Zhang, Y Zee Ma, Christopher Dorion, Colin Daly, and Tuanfeng Zhang. A comparative study of reservoir modeling techniques and their impact on predicted performance of fluvial-dominated deltaic reservoirs : Discussion. *AAPG Bulletin*, 102(8) :1659–1663, 2018. 10

[230] Xihong Lin and Daowen Zhang. Inference in generalized additive mixed modelsby using smoothing splines. *Journal of the royal statistical society : Series b (statistical methodology)*, 61(2) :381–400, 1999. 14

[231] Yue Liu, Qiuming Cheng, Emmanuel John M Carranza, and Kefa Zhou. Assessment of geochemical anomaly uncertainty through geostatistical simulation and singularity analysis. *Natural Resources Research*, 28(1) :199–212, 2019. 10

[232] Zhun-Ga Liu, Quan Pan, and Jean Dezert. A new belief-based k-nearest neighbor classification method. *Pattern Recognition*, 46(3) :834–844, 2013. 16

[233] Alan R Longhurst. *An ecological survey of the West African marine benthos*. HM Stationery Office, 1958. 4, 5

[234] Macoumba Loum, Alioune Badara Dieye, Mar Ndiaye, François Mendy, Samba Sow, and Pape Nekhou Diagne. Remote sensing and sustainable management of soc in the sahelian area. In *Sustainable Agriculture Reviews 29*, pages 111–124. Springer, 2019. 10

[235] Zudi Lu and Xing Chen. Spatial kernel regression estimation : weak consistency. *Statistics & probability letters*, 68(2) :125–136, 2004. 12, 20

[236] Jing Luan, Chongliang Zhang, Binduo Xu, Ying Xue, and Yiping Ren. Modelling the spatial distribution of three portunidae crabs in haizhou bay, china. *PloS one*, 13(11) :e0207457, 2018. 18, 41, 42, 45

[237] MAGNANI LUANA, Porto San Paolo Dive Center, Tavolara-Punta Coda, Parc Valrose ECOSEAS, France CoNISM Nice, and Interuniversity CoNISMa. Copyright© 2019 mediterranean marine science. *Science*, 20(2) :326–330, 2019. 5

[238] YZ Ma. Quantitative geosciences : Data analytics, geostatistics, reservoir characterization and modeling, 2019. 10

[239] YP Mack and Pham X Quang. Kernel methods in line and point transect sampling. *Biometrics*, pages 606–619, 1998. 74

[240] YP Mack and Murray Rosenblatt. Multivariate k-nearest neighbor density estimates. *Journal of Multivariate Analysis*, 9(1) :1–15, 1979. 11

[241] Hamdy FF Mahmoud, Byung-Jun Kim, and Inyoung Kim. Robust nonparametric derivative estimator. *Communications in Statistics-Simulation and Computation*, pages 1–21, 2020. 15

[242] Enno Mammen, Alexandre B Tsybakov, et al. Smooth discrimination analysis. *The Annals of Statistics*, 27(6) :1808–1829, 1999. 16

[243] Solym Mawaki Manou-Abi, Sophie Dabo-Niang, and Jean-Jacques Salone. *Mathematical Modeling of Random and Deterministic Phenomena*. Wiley Online Library, 2020. 7

[244] Rodrigo Lilla Manzione and Annamaria Castrignanò. A geostatistical approach for multi-source data fusion to predict water table depth. *Science of The Total Environment*, page 133763, 2019. 10

[245] Belen Martin-Barragan, Rosa Lillo, and Juan Romo. Interpretable support vector machines for functional data. *European Journal of Operational Research*, 232(1) :146–155, 2014. 16

[246] Israel Martínez-Hernández, Marc G Genton, et al. Recent developments in complex and spatially correlated functional data. *Brazilian Journal of Probability and Statistics*, 34(2) :204–229, 2020. 15

[247] E. Masry. Nonparametric regression estimation for dependent functional data : asymptotic normality. *Stochastic Process. Appl.*, 115(1) :155–177, 2005. 31

[248] Elias Masry. Recursive probability density estimation for weakly dependent stationary processes. *IEEE Transactions on Information Theory*, 32(2) :254–267, 1986. 11

[249] Elias Masry. Almost sure convergence of recursive density estimators for stationary mixing processes. *Statistics & probability letters*, 5(4) :249–254, 1987. 11

[250] Elias Masry and Dag Tjøstheim. Nonparametric estimation and identification of nonlinear arch time series strong convergence and asymptotic normality : Strong convergence and asymptotic normality. *Econometric theory*, 11(2) :258–289, 1995. 11

[251] G Matheron. Présentation des variables régionalisées. *Journal de la société française de statistique*, 107 :263–275, 1966. 11

[252] Hidetoshi Matsui, Takamitsu Araki, and Sadanori Konishi. Multiclass functional discriminant analysis and its application to gesture recognition. *Journal of classification*, 28(2) :227–243, 2011. 17

[253] Peter McCullagh. *Generalized linear models*. Routledge, 2019. 17

[254] Daniel P McMillen. Linear and nonparametric quantile regression. In *Quantile Regression for Spatial Data*, pages 13–27. Springer, 2013. 13

[255] Alessandra Menafoglio, Pigoli Davide, Piercesare Secchi, et al. Mathematical foundations of functional kriging in hilbert spaces and riemannian manifolds. 2019. 28

[256] Alessandra Menafoglio, Piercesare Secchi, Matilde Dalla Rosa, et al. A universal kriging predictor for spatially dependent functional data of a hilbert space. *Electronic Journal of Statistics*, 7 :2209–2240, 2013. 28

[257] Raquel Menezes, Pilar García-Soidán, and Célia Ferreira. Nonparametric spatial prediction under stochastic sampling design. *Journal of Nonparametric Statistics*, 22(3) :363–377, 2010. 12, 20

[258] Josephine Merhi Bleik. Fully bayesian estimation of simultaneous regression quantiles under asymmetric laplace distribution specification. *Journal of Probability and Statistics*, 2019, 2019. 20

[259] Abolfazl Mollalo, Behzad Vahedi, and Kiara M Rivera. Gis-based spatial modeling of covid-19 incidence rate in the continental united states. *Science of The Total Environment*, page 138884, 2020. 10

[260] Cordelia H Moore, Euan S Harvey, and Kimberly P Van Niel. Spatial prediction of demersal fish distributions : enhancing our understanding of species–environment relationships. *ICES Journal of Marine Science*, 66(9) :2068–2075, 2009. 45

[261] Erinn M Muller, Constance Sartor, Nicholas I Alcaraz, and Robert van Woesik. Spatial epidemiology of the stony-coral-tissue-loss disease in florida. *Frontiers in Marine Science*, 7 :163, 2020. 10

[262] HANS-GEORG MÜLLER. Functional modelling and classification of longitudinal data. *Scandinavian Journal of Statistics*, 32(2) :223–240, 2005. 16

[263] Hans-Georg Müller, Ulrich Stadtmüller, et al. Generalized functional linear models. *the Annals of Statistics*, 33(2) :774–805, 2005. 6, 14, 15

[264] Hans-Georg Müller and Fang Yao. Functional additive models. *Journal of the American Statistical Association*, 103(484) :1534–1544, 2008. 14, 15

[265] Simon Muller and Jurgen Dippon. k-nn kernel estimate for nonparametric functional regression in time series analysis. *Fachbereich Mathematik, Fakultat Mathematik und Physik (Pfaffenwaldring 57)*, 14 :2011, 2011. 22

[266] R Murakami, N Matsuo, K Ueda, and M Nakazawa. Epidemiological and spatial factors for tuberculosis : a matched case-control study in nagata, japan. *The International Journal of Tuberculosis and Lung Disease*, 23(2) :181–186, 2019. 10

[267] Hiroto Murase, Hiroshi Nagashima, Shiroh Yonezaki, Ryuichi Matsukura, and Toshihide Kitakado. Application of a generalized additive model (gam) to reveal relationships between environmental factors and distributions of pelagic fish and krill : a case study in sendai bay, japan. *ICES Journal of Marine Science*, 66(6) :1417–1424, 2009. 10

[268] Elizbar A Nadaraya. On estimating regression. *Theory of Probability & Its Applications*, 9(1) :141–142, 1964. 10, 12

[269] Mamadou n'diaye, Sophie Dabo-Niang, Papa Ngom, Ndiaga Thiam, Massal Fall, and Patrice Brehmer. Nonparametric prediction for spatial dependent functional data : Application to demersal coastal fish off senegal. *Mathematical Modeling of Random and Deterministic Phenomena*, pages 31–51, 2020. 7

[270] Siny Ndoye. *Fonctionnement dynamique du centre d'upwelling sud-sénégalais : approche par la modélisation réaliste et l'analyse d'observations satellite de température de surface de la mer*. PhD thesis, 2016. 1, 4

[271] C. C. Neaderhouser. Convergence of block spins defined by a random field. *J. Statist. Phys.*, 22(6) :673–684, 1980. 31

[272] Hannah Omoloye Omogoriola, Akanbi Bamikole Williams, Oyeronke Mojisola Adegbile, Fisayo Christie Olakolu, Stella Ukamaka Ukaonu, and Emmanuel Friday Myade. Length-weight relationships, condition factor (k) and relative condition factor (kn) of sparids, dentex congoensis (maul, 1954) and dentex angolensis (maul and poll, 1953), in nigerian coastal water. *International Journal of Biological and Chemical Sciences*, 5(2), 2011. 44

[273] María Pacheco, Jorge Paramo, and Arturo Acero. First evidence of spatial structure and morphometric relationships of dwarf dory zenion hololepis (goode and bean, 1896)(zeiformes : Zeniontidae) : A deepsea fish in the colombian caribbean. *Boletín Científico. Centro de Museos. Museo de Historia Natural*, 23(1) :219–234, 2019. 10

[274] Roberto Paredes and Enrique Vidal. Learning weighted metrics to minimize nearest-neighbor classification error. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (7) :1100–1110, 2006. 16, 24

[275] Yeonjoo Park, Xiaohui Chen, and Douglas G Simpson. Robust m-estimation for partially observed functional data. *arXiv preprint arXiv :2002.08560*, 2020. 15

[276] Emanuel Parzen. On estimation of a probability density function and mode. *The annals of mathematical statistics*, 33(3) :1065–1076, 1962. 9

[277] Olivier Pezennec. L'environnement hydro-climatique de la guinée. *Pêche côtière en Guinée : Ressources et Exploitation. CNSHB, Conakry and IRD, Paris*, pages 7–25, 1999. 5

[278] Benjamin Planque and Laure Buffaz. Quantile regression models for fish recruitment–environment relationships : four case studies. *Marine Ecology Progress Series*, 357 :213–223, 2008. 41, 45

[279] Tomasz Podgórski, Tomasz Borowik, Magdalena Łyjak, and Grzegorz Woźniakowski. Spatial epidemiology of african swine fever : host, landscape and anthropogenic drivers of disease occurrence in wild boar. *Preventive veterinary medicine*, page 104691, 2019. 10

[280] Canelle Poirier, Wei Luo, Maimuna S Majumder, Dianbo Liu, Kenneth Mandl, Todd Mooring, and Mauricio Santillana. The role of environmental factors on transmission rates of the covid-19 outbreak : An initial assessment in two spatial scales. *Available at SSRN 3552677*, 2020. 10

[281] Oleksii Pokotylo, Pavlo Mozharovskyi, and Rainer Dyckerhoff. Depth and depth-based classification with r-package ddalpha. *arXiv preprint arXiv :1608.04109*, 2016. 16

[282] Laura J Pollock, Reid Tingley, William K Morris, Nick Golding, Robert B O'Hara, Kirsten M Parris, Peter A Vesk, and Michael A McCarthy. Understanding co-occurrence by modelling species simultaneously with a joint species distribution model (jsdm). *Methods in Ecology and Evolution*, 5(5) :397–406, 2014. 18, 42

[283] Ian C Potter, Lynnath E Beckley, Alan K Whitfield, and Rodney CJ Lenanton. Comparisons between the roles played by estuaries in the life cycles of fishes in temperate western australia and southern africa. In *Alternative life-history styles of fishes*, pages 143–178. Springer, 1990. 5

[284] Cristian Preda, Gilbert Saporta, and Caroline Lévéder. Pls classification of functional data. *Computational Statistics*, 22(2) :223–235, 2007. 14

[285] Pham Xuan Quang. Nonparametric estimators for variable circular plot surveys. *Biometrics*, pages 837–852, 1993. 75

[286] Mustapha Rachdi and Philippe Vieu. Nonparametric regression for functional data : automatic smoothing parameter selection. *Journal of Statistical Planning and Inference*, 137(9) :2784–2801, 2007. 15

[287] James O Ramsay and CJ Dalzell. Some tools for functional data analysis. *Journal of the Royal Statistical Society : Series B (Methodological)*, 53(3) :539–561, 1991. 13

[288] James O Ramsay and Bernard W Silverman. *Applied functional data analysis : methods and case studies*. Springer, 2007. 13

[289] Jim O Ramsay and Bernard W Silverman. *Functional Data Analysis*. Springer Series in Statistics. Springer, second edition, 2005. 6, 13, 14

[290] JO Ramsay. *When the data are functions*, volume 47. Springer, 1982. 13

[291] JO Ramsay and BW Silverman. Principal components analysis for functional data. *Functional data analysis*, pages 147–172, 2005. 13

[292] C Radhakrishna Rao. Some statistical methods for comparison of growth curves. *Biometrics*, 14(1) :1–17, 1958. 13

[293] Jean-Paul Rebert. Hydrologie et dynamique des eaux du plateau continental sénégalais. 1982. 1, 4

[294] Didier Renard, Christian Lajaunie, Simon Lopez, Cécile Allanic, Gabriel Courrioux, Bernard Bourgine, and Philippe Calcagno. La géostatistique au service de la modélisation géologique 3d. In *Annales des Mines-Responsabilite et environnement*, number 2, pages 30–33. FFE, 2019. 10

[295] JD Rhoades, PAC Raats, and RJ Prather. Effects of liquid-phase electrical conductivity, water content, and surface conductivity on bulk soil electrical conductivity 1. *Soil Science Society of America Journal*, 40(5) :651–655, 1976. 10

[296] John A Rice and Bernard W Silverman. Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of the Royal Statistical Society : Series B (Methodological)*, 53(1) :233–243, 1991. 13

[297] BD Ripley. Spatial point pattern analysis in ecology. In *Develoments in Numerical Ecology*, pages 407–429. Springer, 1987. 18, 42

[298] Brian D Ripley. *Spatial Statistics*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc, 1981. 11

[299] J. Rivoirard. *Geostatistics for Estimating Fish Abundance*. Sparks Computer Solutions Ltd, Oxford, 2000. 18

[300] Jacques Rivoirard, J Simmonds, KG Foote, P Fernandes, and N Bez. *Geostatistics for estimating fish abundance*. John Wiley & Sons, 2008. 10, 42

[301] Rubén Roa-Ureta and Edwin Niklitschek. Biomass estimation from surveys with likelihood-based geostatistics. *ICES Journal of Marine Science*, 64(9) :1723–1734, 2007. 10

[302] Peter M Robinson. Nonparametric estimators for time series. *Journal of Time Series Analysis*, 4(3) :185–207, 1983. 11

[303] Derek Roff. *Evolution of life histories : theory and analysis*. Springer Science & Business Media, 1993. 5

[304] M. Rosenblatt. *Stationary sequences and random fields*. Birkhauser, Boston, 1985. 31

[305] Murray Rosenblatt. A central limit theorem and a strong mixing condition. *Proceedings of the National Academy of Sciences of the United States of America*, 42(1) :43, 1956. 9

[306] Murray Rosenblatt. Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, 27(3) :832–837, 1956. 74

[307] Fabrice Rossi and Nathalie Villa. Support vector machine for functional data classification. *Neurocomputing*, 69(7-9) :730–742, 2006. 16

[308] Fabrice Rossi and Nathalie Villa. Recent advances in the use of svm for functional data classification. In *Functional and operatorial statistics*, pages 273–280. Springer, 2008. 16

[309] Claude Roy. Fluctuations des vents et variabilité de l'upwelling devant les côtes du sénégal. *Oceanologica Acta*, 12(4) :361–369, 1989. 1

[310] María Dolores Ruiz-Medina. Spatial autoregressive and moving average hilbertian processes. *Journal of Multivariate Analysis*, 102(2) :292–305, 2011. 27

[311] María Dolores Ruiz-Medina, Vo V Anh, Rosa M Espejo, José Miguel Angulo, and Maria Pilar Frias. Least-squares estimation of multifractional random fields in a hilbert-valued context. *Journal of Optimization Theory and Applications*, 167(3) :888–911, 2015. 27

[312] Espejo R Ruiz-Medina M. Spatial autoregressive functional plug-in prediction of ocean surface temperature. *Stoch Environ Res Risk Assess 26(3) :335-344*, 2012. 27

[313] Norraisha Md Sabtu, Mohamad Hafis Izran Ishak, and Nurul Hawani Idris. The spatial epidemiology of jackfruit pest and diseases : A review. *International Journal of Built Environment and Sustainability*, 6(1-2) :169–175, 2019. 10

[314] Luis Sánchez, Víctor Leiva, Manuel Galea, and Helton Saulo. Birnbaum-saunders quantile regression models with application to spatial data. *Mathematics*, 8(6) :1000, 2020. 13

[315] Bhabesh C Sarkar. Geostatistics in groundwater modelling. In *Groundwater Development and Management*, pages 147–169. Springer, 2019. 10

[316] Olivier Scaillet and Jean-David Fermanian. Nonparametric estimation of copulas for time series. *FAME Research paper*, (57), 2002. 11

[317] George AF Seber. A review of estimating animal abundance. *Biometrics*, pages 267–292, 1986. 74

[318] Han Lin Shang, Rob J Hyndman, and Maintainer Han Lin Shang. Package 'rainbow'. *R packages*, 2019. 43

[319] Bernard W Silverman. *Density estimation for statistics and data analysis*. Routledge, 2018. 75

[320] Yousri Slaoui. Recursive nonparametric regression estimation for independent functional data. *Statistica Sinica*, 30(1) :417–37, 2020. 15

[321] Helle Sørensen, Jeff Goldsmith, and Laura M Sangalli. An introduction with medical applications to functional data analysis. *Statistics in medicine*, 32(30) :5222–5240, 2013. 27

[322] Marc Souris. *Epidemiology and Geography : Principles, Methods and Tools of Spatial Analysis*. John Wiley & Sons, 2019. 10

[323] Roberto Souza, Letícia Rittner, and Roberto Lotufo. A comparison between k-optimum path forest and k-nearest neighbors supervised classifiers. *Pattern recognition letters*, 39 :2–10, 2014. 16

[324] GB Sreekanth, SK Chakraborty, AK Jaiswar, Bappa Das, and EB Chakurkar. Application of deterministic and stochastic geo-statistical tools for analysing spatial patterns of fish density in a tropical monsoonal estuary. *Aquatic ecology*, 53(1) :49–60, 2019. 10

[325] Ingo Steinwart. Consistency of support vector machines and other regularized kernel classifiers. *IEEE Transactions on Information Theory*, 51(1) :128–142, 2005. 16

[326] Sarah Stienessen, Taina Honkalehto, Nathan Lauffenburger, Patrick Henry Ressler, and Robert Russell Lauth. Acoustic vessel-of-opportunity (avo) index for midwater bering sea walleye pollock, 2016-2017. 2019. 10

[327] Charles J Stone. Consistent nonparametric regression. *The annals of statistics*, pages 595–620, 1977. 16

[328] Charles J Stone et al. Additive regression and other nonparametric models. *The annals of Statistics*, 13(2) :689–705, 1985. 14

[329] H. Takahata. On the rates in the central limit theorem for weakly dependent random fields. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 64(4) :445–456, 1983. 31

[330] C. Ternynck. Spatial regression estimation for functional data with spatial dependency. *SFDS,155, 2*, 2014. 15, 18, 20, 28

[331] Shaymaa Riyadh Thanoon. Spatial prediction for real data using kriging technique. *Iraqi Journal of Statistical Science*, 27 :1–11, 2018. 10

[332] Djiga Thiao. Un système d'indicateurs de durabilité des pêcheries côtières comme outil de gestion intégrée des ressources halieutiques sénégalaises. B.S. thesis, 2009. 1

[333] Djiga Thiao, Christian Chaboud, Alassane Samba, Francis Laloë, and PM Cury. Economic dimension of the collapse of the 'false cod' epinephelus aeneus in a context of ineffective management of the small-scale fisheries in senegal. *African Journal of Marine Science*, 34(3) :305–311, 2012. 1

[334] Djiga Thiao and Francis Laloë. A system of indicators to understand the socioeconomic and ecological interactions and manage the fisheries sustainability. 2010. 44, 45

[335] M Thiaw, Didier Gascuel, D Thiao, OT Thiaw, and Didier Jouffre. Analysing environmental and fishing effects on a short-lived species stock : the dynamics of the octopus octopus vulgaris population in senegalese waters. *African Journal of Marine Science*, 33(2) :209–222, 2011. 1

[336] Modou Thiaw et al. *Dynamique des ressources halieutiques à durée de vie courte : cas des stocks de poulpe et de crevettes exploités au Sénégal*. PhD thesis, 2010. 1

[337] Modou Thiaw, Didier Gascuel, Didier Jouffre, and Omar Thiom Thiaw. A surplus production model including environmental effects : Application to the senegalese white shrimp stocks. *Progress in Oceanography*, 83(1-4) :351–360, 2009. 1

[338] James T Thorson, Andrew O Shelton, Eric J Ward, and Hans J Skaug. Geostatistical delta-generalized linear mixed models improve precision for estimated abundance indices for west coast groundfishes. *ICES Journal of Marine Science*, 72(5) :1297–1310, 2015. 10

[339] Léonard Torossian, Victor Picheny, Robert Faivre, and Aurélien Garivier. A review on quantile regression for stochastic computer experiments. *arXiv preprint arXiv :1901.07874*, 2019. 20

[340] J Martínez Torres, PJ Garcia Nieto, L Alejano, and AN Reyes. Detection of outliers in gas emissions from urban areas using functional data analysis. *Journal of hazardous materials*, 186(1) :144–149, 2011. 27

[341] Lanh Tat Tran. Kernel density estimation on random fields. *Journal of Multivariate Analysis*, 34(1) :37–53, 1990. 11, 20, V

[342] Lanh Tat Tran. Kernel density estimation on random fields. *Journal of Multivariate Analysis*, 34(1) :37–53, 1990. 30, XI, XV

[343] Lanh Tat Tran. Nonparametric function estimation for time series by local average estimators. *The Annals of Statistics*, pages 1040–1057, 1993. 11

[344] Young K Truong. Nonparametric curve estimation with time series errors. *Journal of Statistical Planning and Inference*, 28(2) :167–183, 1991. 11

[345] Alexandre B Tsybakov. Introduction to nonparametric estimation, 2009. *URL https ://doi.org/10.1007/b13794. Revised and extended from the*, 2004. 9, 10

[346] Ledyard R Tucker. Determination of parameters of a functional relation by factor analysis. *Psychometrika*, 23(1) :19–23, 1958. 13

[347] Franz Uiblein. Goatfishes (mullidae) as indicators in tropical and temperate coastal habitat monitoring and management. *Marine Biology Research*, 3(5) :275–288, 2007. 44

[348] Vladimir Vapnik and Vlamimir Vapnik. Statistical learning theory, 1998. 16

[349] Emmanouil A Varouchakis, Panagiota G Theodoridou, and George P Karatzas. Decision-making tool for groundwater level spatial distribution and risk assessment using geostatistics in r. *Journal of Hazardous, Toxic, and Radioactive Waste*, 24(1) :04019031, 2019. 10

[350] H. Wackernagel. *Multivariate Geostatistics : An Introduction with Applications*. Springer-Verlag, 2003. 11

[351] Hongxia Wang and Jinde Wang. Estimation of the trend function for spatio-temporal models. *Journal of Nonparametric Statistics*, 21(5) :567–588, 2009. 20

[352] Hongxia Wang, Jinde Wang, and Bo Huang. Prediction for spatio-temporal models with autoregression in errors. *Journal of Nonparametric Statistics*, 24(1) :217–244, 2012. 12

[353] Jigang Wang, Predrag Neskovic, and Leon N Cooper. Improving nearest neighbor rule with a simple adaptive distance measure. *Pattern Recognition Letters*, 28(2) :207–213, 2007. 16

[354] Li Wang, Guannan Wang, Lei Gao, Xinyi Li, Shan Yu, Myungjin Kim, Yueying Wang, and Zhiling Gu. Spatiotemporal dynamics, nowcasting and forecasting of covid-19 in the united states. *arXiv preprint arXiv :2004.14103*, 2020. 10

[355] Xiaohui Wang, Shubhankar Ray, and Bani K Mallick. Bayesian curve classification using wavelets. *Journal of the American Statistical Association*, 102(479) :962–973, 2007. 17

[356] K Warburton. Community structure, abundance and diversity of fish in a mexican coastal lagoon system. *Estuarine and coastal marine science*, 7(6) :497–519, 1978. 5

[357] Geoffrey S Watson. Smooth regression analysis. *Sankhyā : The Indian Journal of Statistics, Series A*, pages 359–372, 1964. 10, 12

[358] Edward J Wegman and HI Davies. Remarks on some recursive estimators of a probability density. *The Annals of Statistics*, pages 316–327, 1979. 11

[359] ME Wigwe, MC Watson, A Giussani, E Nasir, S Dambani, et al. Application of geographically weighted regression to model the effect of completion parameters on oil production–case study on unconventional wells. In *SPE Nigeria Annual International Conference and Exhibition*. 2019. 10

[360] Mathieu Woillez, Jacques Rivoirard, and Paul G Fernandes. Evaluating the uncertainty of abundance estimates from acoustic surveys using geostatistical simulations. *ICES Journal of Marine Science*, 66(6) :1377–1383, 2009. 10

[361] Jacob Wolfowitz. Additive partition functions and a class of statistical hypotheses. *The Annals of Mathematical Statistics*, 13(3) :247–279, 1942. 9, 10

[362] Charles T Wolverton and Terry J Wagner. Recursive estimates of probability densities. *IEEE Transactions on Systems Science and Cybernetics*, 5(3) :246–247, 1969. 11

[363] Robert J Wootton. The evolution of life histories : theory and analysis. *Reviews in Fish Biology and Fisheries*, 3(4) :384–385, 1993. 5

[364] Mohamed Yahaya. *Extension au cadre spatial de l'estimation non paramétrique par noyaux récursifs*. PhD thesis, Université Charles de Gaulle-Lille III, 2016. 11

[365] Sid Yakowitz. Nearest neighbor regression estimation for null-recurrent markov time series. *Stochastic Processes and their Applications*, 48(2) :311–318, 1993. 11

[366] Sidney Yakowitz. Nonparametric density and regression estimation for markov sequences without mixing assumptions. *Journal of Multivariate Analysis*, 30(1) :124–136, 1989. 11

[367] Miin-Shen Yang and Chien-Hung Chen. On the edited fuzzy k-nearest neighbor rule. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 28(3) :461–466, 1998. 16

[368] QuanHe Yang, YongLong An, et al. Comprehensive evaluation of soil fertility in yujiawu town of tongzhou district using geostatistics and gis. *Southwest China Journal of Agricultural Sciences*, 32(4) :882–891, 2019. 10

[369] Yunwen Yang and Xuming He. Quantile regression for spatially correlated data : an empirical likelihood approach. *Statistica Sinica*, pages 261–274, 2015. 13

[370] Jian DL Yen, James R Thomson, David M Paganin, Jonathan M Keith, and Ralph Mac Nally. Function regression in ecology and evolution : Free. *Methods in Ecology and Evolution*, 6(1) :17–26, 2015. 27

[371] Jinhong You and Haibo Zhou. Two-stage efficient estimation of longitudinal nonparametric additive models. *Statistics & Probability Letters*, 77(17) :1666–1675, 2007. 14

[372] Mary Young and Mark H Carr. Application of species distribution models to explain and predict the distribution, abundance and assemblage structure of nearshore temperate reef fishes. *Diversity and Distributions*, 21(12) :1428–1440, 2015. 18, 41, 42, 45

[373] Ahmad Younso. On the consistency of a new kernel rule for spatially dependent data. *Statistics & Probability Letters*, 2017. 16, 24, 26, 39

[374] Ahmad Younso. On the consistency of kernel classification rule for functional random field. *Journal de la Société Française de Statistique*, 159(1) :68–87, 2018. 16

[375] Ahmad Younso et al. Nonparametric discrimination of areal functional data. *Brazilian Journal of Probability and Statistics*, 34(1) :112–126, 2020. 16

[376] Yixiao Yu, Xueshan Han, Ming Yang, and Jiajun Yang. Probabilistic prediction of regional wind power based on spatiotemporal quantile regression. *IEEE Transactions on Industry Applications*, 2020. 13

[377] Yong Zeng, Yupu Yang, and Liang Zhao. Pseudo nearest neighbor rule for pattern classification. *Expert Systems with Applications*, 36(2) :3587–3595, 2009. 16

[378] Haozhe Zhang. Topics in functional data analysis and machine learning predictive inference. 2019. 28

[379] Xiaoke Zhang, Byeong U Park, and Jane-ling Wang. Time-varying additive models for longitudinal data. *Journal of the American Statistical Association*, 108(503) :983–998, 2013. 14

[380] Xiaoke Zhang and Jane-Ling Wang. Varying-coefficient additive models for functional data. *Biometrika*, 102(1) :15–32, 2014. 14

[381] Chenghu Zhou, Fenzhen Su, Tao Pei, An Zhang, Yunyan Du, Bin Luo, Zhidong Cao, Juanle Wang, Wen Yuan, Yunqiang Zhu, et al. Covid-19 : Challenges to gis with big data. *Geography and Sustainability*, 2020. 10

ANNEXE A

PREUVES DU CHAPITRE 3

Les preuves sont rédigées en anglais.

Technical Lemmas

We start by the following technical lemmas that are helpful to handle the difficulties induced by the random bandwidth $\mathrm{H}_{\mathbf{n},x}$ in $r_{\mathrm{kNN}}(x)$. They are adaptation of the results given in [87] (for independent multivariate data) and their generalized version by [59], [216] (for independent functional data).
For any random positive variable T, $\mathbf{n} \in \mathbb{N}^{*\mathrm{N}}$, and $x \in \mathrm{D}$, we define

$$c_{\mathbf{n}}(\mathrm{T}) = \frac{\sum_{\mathbf{i}\in\mathscr{I}_{\mathbf{n}},\mathbf{s_0}\neq\mathbf{i}} \mathrm{Y_i K_1}\left(\frac{x-\mathrm{X_i}}{\mathrm{T}}\right)\mathrm{K_2}\left(h_{\mathbf{n},\mathbf{s_0}}^{-1}\left\|\frac{\mathbf{s_0}-\mathbf{i}}{\mathbf{n}}\right\|\right)}{\sum_{\mathbf{i}\in\mathscr{I}_{\mathbf{n}},\mathbf{s_0}\neq\mathbf{i}} \mathrm{K_1}\left(\frac{x-\mathrm{X_i}}{\mathrm{T}}\right)\mathrm{K_2}\left(h_{\mathbf{n},\mathbf{s_0}}^{-1}\left\|\frac{\mathbf{s_0}-\mathbf{i}}{\mathbf{n}}\right\|\right)}.$$

Let us set the following sequences, for all $\mathbf{n} \in \mathbb{N}^{*\mathrm{N}}$

$$v_{\mathbf{n}} = \left(\frac{k_{\mathbf{n}}}{k_{\mathbf{n}}'}\right)^{1/d} + \left(\frac{\log(\hat{\mathbf{n}})}{k_{\mathbf{n}}}\right)^{1/2},$$

and for all $\beta \in]0,1[$ and $x \in \mathrm{D}$

$$\mathrm{D}_{\mathbf{n}}^{-}(\beta,x) = \left(\frac{k_{\mathbf{n}}}{c f(x) k_{\mathbf{n}}'}\right)^{1/d}\beta^{1/2d}, \qquad \mathrm{D}_{\mathbf{n}}^{+}(\beta,x) = \left(\frac{k_{\mathbf{n}}}{c f(x) k_{\mathbf{n}}'}\right)^{1/d}\beta^{-1/2d}, \tag{0.1}$$

where c is the volume of the unit sphere in $\mathbb{R}^d$. It is clear that

$$\forall\, \mathbf{n} \in \mathbb{N}^{*\mathrm{N}}, \forall x \in \mathrm{D} \qquad \mathrm{D}_{\mathbf{n}}^{-}(\beta,x) \leq \mathrm{D}_{\mathbf{n}}^{+}(\beta,x).$$

Lemma 1. *If the following conditions are verified :*

(L_1) $\mathbb{I}_{\{\mathrm{D}_{\mathbf{n}}^{-}(\beta,x)\leq \mathrm{H}_{\mathbf{n},x}\leq \mathrm{D}_{\mathbf{n}}^{+}(\beta,x),\,\forall x\in\mathrm{D}\}} \longrightarrow 1 \quad a.co.$

(L_2) $\sup_{x\in\mathrm{D}}\left|\frac{\sum_{\mathbf{i}\in\mathscr{I}_{\mathbf{n}},\mathbf{s_0}\neq\mathbf{i}} \mathrm{K_1}\left(\frac{x-\mathrm{X_i}}{\mathrm{D}_{\mathbf{n}}^{-}(\beta,x)}\right)\mathrm{K_2}\left(h_{\mathbf{n},\mathbf{s_0}}^{-1}\left\|\frac{\mathbf{s_0}-\mathbf{i}}{\mathbf{n}}\right\|\right)}{\sum_{\mathbf{i}\in\mathscr{I}_{\mathbf{n}},\mathbf{s_0}\neq\mathbf{i}} \mathrm{K_1}\left(\frac{x-\mathrm{X_i}}{\mathrm{D}_{\mathbf{n}}^{+}(\beta,x)}\right)\mathrm{K_2}\left(h_{\mathbf{n},\mathbf{s_0}}^{-1}\left\|\frac{\mathbf{s_0}-\mathbf{i}}{\mathbf{n}}\right\|\right)} - \beta\right| \longrightarrow 0 \quad a.co.$

(L_3) $\sup_{x\in\mathrm{D}}\left|c_{\mathbf{n}}\left(\mathrm{D}_{\mathbf{n}}^{-}(\beta,x)\right) - r(x)\right| \longrightarrow 0 \quad a.co.,\ \sup_{x\in\mathrm{D}}\left|c_{\mathbf{n}}\left(\mathrm{D}_{\mathbf{n}}^{+}(\beta,x)\right) - r(x)\right| \longrightarrow 0 \quad a.co,$

then we have $\sup_{x\in\mathrm{D}}|c_{\mathbf{n}}\left(\mathrm{H}_{\mathbf{n},x}\right) - r(x)| \longrightarrow 0 \qquad a.co.$

Lemma 2. *Under the following conditions :*

(L_1) $\mathbb{I}_{\{\mathrm{D}_{\mathbf{n}}^{-}(\beta,x)\leq \mathrm{H}_{\mathbf{n},x}\leq \mathrm{D}_{\mathbf{n}}^{+}(\beta,x),\,\forall x\in\mathrm{D}\}} \longrightarrow 1 \quad a.co.$

(L_2') $\sup_{x\in\mathrm{D}}\left|\frac{\sum_{\mathbf{i}\in\mathscr{I}_{\mathbf{n}},\mathbf{s_0}\neq\mathbf{i}} \mathrm{K_1}\left(\frac{x-\mathrm{X_i}}{\mathrm{D}_{\mathbf{n}}^{-}(\beta,x)}\right)\mathrm{K_2}\left(h_{\mathbf{n},\mathbf{s_0}}^{-1}\left\|\frac{\mathbf{s_0}-\mathbf{i}}{\mathbf{n}}\right\|\right)}{\sum_{\mathbf{i}\in\mathscr{I}_{\mathbf{n}},\mathbf{s_0}\neq\mathbf{i}} \mathrm{K_1}\left(\frac{x-\mathrm{X_i}}{\mathrm{D}_{\mathbf{n}}^{+}(\beta,x)}\right)\mathrm{K_2}\left(h_{\mathbf{n},\mathbf{s_0}}^{-1}\left\|\frac{\mathbf{s_0}-\mathbf{i}}{\mathbf{n}}\right\|\right)} - \beta\right| = \mathscr{O}(v_{\mathbf{n}}) \qquad a.co.$

$$(L_3') \quad \sup_{x\in D}\left|c_{\mathbf{n}}\left(D_{\mathbf{n}}^{-}(\beta,x)\right)-r(x)\right|=\mathcal{O}(v_{\mathbf{n}}) \qquad a.co,$$
$$\sup_{x\in D}\left|c_{\mathbf{n}}\left(D_{\mathbf{n}}^{+}(\beta,x)\right)-r(x)\right|=\mathcal{O}(v_{\mathbf{n}}) \qquad a.co,$$
$$\textit{we have,} \sup_{x\in D}\left|c_{\mathbf{n}}\left(H_{\mathbf{n},x}\right)-r(x)\right|=\mathcal{O}(v_{\mathbf{n}}) \qquad a.co.$$

The proof of Lemma 2 is similar as in [87] and is therefore omitted. Lemma 1 is a particular case of the proof of Lemma 2 when we take $v_{\mathbf{n}} = 1$ and C = 1.

Proofs of Lemma 1 and Lemma 2

Since the proof of Lemma 1 is based on the result of Lemma 1, it is sufficient to check conditions (L_1), (L_2) and (L_3). For the proof of Lemma 2, it suffices to check conditions (L_2') and (L_3').
To check the condition (L_1), we need the following two lemmas.

Lemma 3. *([197] or [115])*

i) We assume that the mixing condition (3.1) is satisfied. We denote by $\mathscr{L}_r(\mathscr{F})$ the class of $\mathscr{F}-$mesurable random variables X *satisfying*

$$\|X\|_r := \left(E\left(|X|^r\right)\right)^{1/r} < \infty.$$

Let $X \in \mathscr{L}_r(\mathscr{B}(E))$, $Y \in \mathscr{L}_s\left(\mathscr{B}(E')\right)$ *and* $1 \le r,s,t \le \infty$ *such that* $\frac{1}{r}+\frac{1}{s}+\frac{1}{t}=1$, *then*

$$|Cov(X,Y)| \le \|X\|_r \|Y\|_s \left\{\psi\left(Card(E), Card(E')\right)\varphi\left(dist(E,E')\right)\right\}^{1/t}. \tag{0.2}$$

ii) For random variables X, Y *bounded with probability* 1, *we have*

$$|Cov(X,Y)| \le C\psi\left(Card(E), Card(E')\right)\varphi\left(dist(E,E')\right). \tag{0.3}$$

Lemma 4. *Under assumptions of Theorem 1, we have*

$$S_{\mathbf{n}} + R_{\mathbf{n}} = \mathscr{O}\left(k'_{\mathbf{n}}\delta_{\mathbf{n}}\right),$$

where

$$S_{\mathbf{n}} = \sum_{\mathbf{i}\in\mathscr{V}_{\mathbf{s_0}}} Var(\Lambda_{\mathbf{i}}) \text{ and } \quad R_{\mathbf{n}} = \sum_{\mathbf{i}\in\mathscr{V}_{\mathbf{s_0}}} \sum_{\substack{\mathbf{j}\in\mathscr{V}_{\mathbf{s_0}} \\ \mathbf{j}\neq\mathbf{i}}} \left|Cov\left(\Lambda_{\mathbf{i}},\Lambda_{\mathbf{j}}\right)\right|,$$

$$\Lambda_{\mathbf{i}} = \mathbb{I}_{B(x,D_{\mathbf{n}})}(X_{\mathbf{i}}), \quad \mathbf{i}\in\mathscr{I}_{\mathbf{n}}, \quad \delta_{\mathbf{n}} = \mathbb{P}\left(\|X-x\| < D_{\mathbf{n}}\right), \quad D_{\mathbf{n}}^d = \mathscr{O}\left(\frac{k_{\mathbf{n}}}{k'_{\mathbf{n}}}\right),$$

$B(x,\varepsilon)$ *denotes the closed ball of* $\mathbb{R}^d$ *with center* x *and radius* ε.

Proof of Lemma 4

Let $\delta_{\mathbf{n},\mathbf{i}} = \mathbb{P}\left(\|X_{\mathbf{i}}-x\| < D_{\mathbf{n}}\right)$, we can deduce that

$$S_{\mathbf{n}} = \sum_{\mathbf{i}\in\mathscr{V}_{\mathbf{s_0}}} Var\left(\Lambda_{\mathbf{i}}\right) = \sum_{\mathbf{i}\in\mathscr{V}_{\mathbf{s_0}}} \delta_{\mathbf{n},\mathbf{i}}(1-\delta_{\mathbf{n},\mathbf{i}}) = \mathscr{O}\left(k'_{\mathbf{n}}\delta_{\mathbf{n}}\right),$$

by the following results.
Firstly, under the Lipschitz condition of f (assumption (**H1**)), we have

$$\begin{aligned}
\delta_{\mathbf{n}} &= \mathbb{P}\left(\|X-x\| < D_{\mathbf{n}}\right) \\
&= f(x)\int_{B(x,D_{\mathbf{n}})} du + \int_{B(x,D_{\mathbf{n}})} (f(u)-f(x))\,du \\
&= cf(x)D_{\mathbf{n}}^d + \mathscr{O}\left(D_{\mathbf{n}}^{d+1}\right).
\end{aligned} \tag{0.4}$$

Secondly

$$\begin{aligned}
\delta_{\mathbf{n},\mathbf{i}} - \delta_{\mathbf{n}} &= \int_{B(x,D_{\mathbf{n}})} \left(f_{\mathbf{i}}(u)-f(u)\right)(u)\,du \\
&= \sup_u \left|f_{\mathbf{i}}(u)-f(u)\right| \mathscr{O}\left(\frac{k_{\mathbf{n}}}{k'_{\mathbf{n}}}\right).
\end{aligned} \tag{0.5}$$

Thus, the local stationarity assumption (H_8) implies

$$\sum_{\mathbf{i}\in\mathscr{V}_{\mathbf{s_0}}} \left(\delta_{\mathbf{n},\mathbf{i}} - \delta_{\mathbf{n}}\right) = o(k_{\mathbf{n}}). \tag{0.6}$$

Now for $R_{\mathbf{n}}$, it should be noted that by (H5) and for each $\mathbf{j}\neq\mathbf{i}$

$$\begin{aligned}
\left|Cov\left(\Lambda_{\mathbf{i}},\Lambda_{\mathbf{j}}\right)\right| &= \Big|\mathbb{P}\left(\|X_{\mathbf{i}}-x\| < D_{\mathbf{n}}, \|X_{\mathbf{j}}-x\| < D_{\mathbf{n}}\right) \\
&\qquad - \mathbb{P}\left(\|X_{\mathbf{i}}-x\| < D_{\mathbf{n}}\right)\mathbb{P}\left(\|X_{\mathbf{j}}-x\| < D_{\mathbf{n}}\right)\Big| \\
&\le \int_{B(x,D_{\mathbf{n}})\times B(x,D_{\mathbf{n}})} \left|f_{X_{\mathbf{i}}X_{\mathbf{j}}}(u,v) - f_{\mathbf{i}}(u)f_{\mathbf{j}}(v)\right| du\,dv \\
&\le CD_{\mathbf{n}}^{2d} \le C\delta_{\mathbf{n}}^2,
\end{aligned} \tag{0.7}$$

since by (0.4)

$$\frac{\mathrm{D}_{\mathbf{n}}^{d}}{\delta_{\mathbf{n}}} \to \frac{1}{cf(x)}, \qquad \text{as} \quad \mathbf{n} \to \infty.$$

Using Lemma 3 and (0.4), we can write for $r = s = 4$

$$\begin{aligned} \left|\mathrm{Cov}\left(\Lambda_{\mathbf{i}}, \Lambda_{\mathbf{j}}\right)\right| &\leq \mathrm{C}\left[\mathrm{E}\left(\Lambda_{\mathbf{i}}^{4}\right)\mathrm{E}\left(\Lambda_{\mathbf{j}}^{4}\right)\right]^{1/4}\left(\psi(1,1)\varphi\left(\|\mathbf{i}-\mathbf{j}\|\right)\right)^{1/2} \\ &\leq \mathrm{C}\delta_{\mathbf{n}}^{1/2}\varphi\left(\|\mathbf{i}-\mathbf{j}\|\right)^{1/2}. \end{aligned} \tag{0.8}$$

Let $q_{\mathbf{n}}$ be a sequence of real numbers defined by $q_{\mathbf{n}}^{\mathrm{N}} = \mathscr{O}\left(\frac{k_{\mathbf{n}}^{'}}{k_{\mathbf{n}}}\right)$. Using the later, we define $\mathrm{S} = \{\mathbf{i}, \mathbf{j} \in \mathcal{V}_{\mathbf{s_0}},\ 0 < \|\mathbf{i}-\mathbf{j}\| \leq q_{\mathbf{n}}\}$ and S^{c} its complementary in $\mathcal{V}_{\mathbf{s_0}}$, and rewrite

$$\mathrm{R}_{\mathbf{n}} = \sum_{\mathbf{i},\mathbf{j}\in\mathrm{S}}\left|\mathrm{Cov}\left(\Lambda_{\mathbf{i}}, \Lambda_{\mathbf{j}}\right)\right| + \sum_{\mathbf{i},\mathbf{j}\in\mathrm{S}^{c}}\left|\mathrm{Cov}\left(\Lambda_{\mathbf{i}}, \Lambda_{\mathbf{j}}\right)\right| = \mathrm{R}_{\mathbf{n}}^{(1)} + \mathrm{R}_{\mathbf{n}}^{(2)}.$$

Firstly, according to the definitions of $q_{\mathbf{n}}$ and S, and equation (0.7), we have

$$\mathrm{R}_{\mathbf{n}}^{(1)} \leq \sum_{\mathbf{i},\mathbf{j}\in\mathrm{S}}\mathrm{C}\delta_{\mathbf{n}}^{2} \leq \mathrm{C}\delta_{\mathbf{n}}^{2}k_{\mathbf{n}}^{'}q_{\mathbf{n}}^{\mathrm{N}} = \mathscr{O}\left(k_{\mathbf{n}}^{'}\delta_{\mathbf{n}}\right),$$

since $\delta_{\mathbf{n}} = \mathscr{O}(q_{\mathbf{n}}^{-\mathrm{N}})$ by (0.4).
Secondly, by (3.2) and (0.8), we get

$$\begin{aligned} \mathrm{R}_{\mathbf{n}}^{(2)} &\leq \mathrm{C}\delta_{\mathbf{n}}^{1/2}\sum_{\mathbf{i},\mathbf{j}\in\mathrm{S}^{c}}\varphi\left(\|\mathbf{i}-\mathbf{j}\|\right)^{1/2} = \mathrm{C}\delta_{\mathbf{n}}^{1/2}k_{\mathbf{n}}^{'}\sum_{\mathbf{i}\in\mathrm{S}^{c}}\varphi\left(\|\mathbf{i}\|\right)^{1/2} \\ &= \mathrm{C}\delta_{\mathbf{n}}k_{\mathbf{n}}^{'}\delta_{\mathbf{n}}^{-1/2}\sum_{\mathbf{i}\in\mathrm{S}^{c}}\varphi\left(\|\mathbf{i}\|\right)^{1/2} \leq \mathrm{C}\delta_{\mathbf{n}}k_{\mathbf{n}}^{'}\left(\frac{k_{\mathbf{n}}}{k_{\mathbf{n}}^{'}}\right)^{-1/2}\sum_{\mathbf{i}\in\mathrm{S}^{c}}\varphi\left(\|\mathbf{i}\|\right)^{1/2} \\ &\leq \mathrm{C}\delta_{\mathbf{n}}k_{\mathbf{n}}^{'}\sum_{\mathbf{i}\in\mathrm{S}^{c}}\|\mathbf{i}\|^{(\mathrm{N}-\theta)/2} = \mathscr{O}\left(\delta_{\mathbf{n}}k_{\mathbf{n}}^{'}\right), \end{aligned}$$

because under assumptions (**H6**) and (**H7**), we have $\theta > (1 + \frac{2\gamma}{\gamma-\bar{\gamma}})\mathrm{N}$, thus

$$\sum_{\mathbf{i}\in\mathrm{S}^{c}}\|\mathbf{i}\|^{(\mathrm{N}-\theta)/2} \leq k_{\mathbf{n}}^{'}q_{\mathbf{n}}^{(\mathrm{N}-\theta)/2} = o(1).$$

Finally, the result follows :

$$\mathrm{R}_{\mathbf{n}} = \mathscr{O}\left(k_{\mathbf{n}}^{'}\delta_{\mathbf{n}}\right) \text{ and } \mathrm{S}_{\mathbf{n}} + \mathrm{R}_{\mathbf{n}} = \mathscr{O}\left(k_{\mathbf{n}}^{'}\delta_{\mathbf{n}}\right).$$

Verification of (L_1)

Let $\varepsilon_{\mathbf{n}} = \frac{1}{2}\varepsilon_0\left(k_{\mathbf{n}}/k_{\mathbf{n}}^{'}\right)^{1/d}$ with $\varepsilon_0 > 0$ and let $\mathrm{N}_{\varepsilon_{\mathbf{n}}} = \mathscr{O}(\varepsilon_{\mathbf{n}}^{-d})$ be a positive integer. Since D is compact, one can cover it by $\mathrm{N}_{\varepsilon_{\mathbf{n}}}$ closed balls in $\mathbb{R}^{d}$ of centers $x_i \in \mathrm{D}$, $i = 1, \ldots, \mathrm{N}_{\varepsilon_{\mathbf{n}}}$ and radius $\varepsilon_{\mathbf{n}}$. Let us show that

$$\mathbb{I}_{\{\mathrm{D}_{\mathbf{n}}^{-}(\beta,x)\leq \mathrm{H}_{\mathbf{n},x}\leq \mathrm{D}_{\mathbf{n}}^{+}(\beta,x),\,\forall x\in\mathrm{D}\}} \longrightarrow 1 \qquad a.co,$$

which can be written as, $\forall\ \eta > 0$,

$$\sum_{\mathbf{n}\in\mathbb{N}^{*\mathrm{N}}}\mathbb{P}(|\,\mathbb{I}_{\{\mathrm{D}_{\mathbf{n}}^{-}(\beta,x)\leq \mathrm{H}_{\mathbf{n},x}\leq \mathrm{D}_{\mathbf{n}}^{+}(\beta,x),\,\forall x\in\mathrm{D}\}} - 1\,| > \eta) < \infty.$$

We have

$$\begin{aligned} &\mathbb{P}(|\,\mathbb{I}_{\{\mathrm{D}_{\mathbf{n}}^{-}(\beta,x)\leq \mathrm{H}_{\mathbf{n},x}\leq \mathrm{D}_{\mathbf{n}}^{+}(\beta,x),\,\forall x\in\mathrm{D}\}} - 1\,| > \eta) \\ &\leq \mathbb{P}\left(\sup_{x\in\mathrm{D}}\left(\mathrm{H}_{\mathbf{n},x} - \mathrm{D}_{\mathbf{n}}^{-}(\beta,x)\right) < 0\right) + \mathbb{P}\left(\inf_{x\in\mathrm{D}}\left(\mathrm{H}_{\mathbf{n},x} - \mathrm{D}_{\mathbf{n}}^{+}(\beta,x)\right) > 0\right) \\ &\leq \mathbb{P}\left(\max_{1\leq i\leq \mathrm{N}_{\varepsilon_{\mathbf{n}}}}\left(\mathrm{H}_{\mathbf{n},x_i} - \mathrm{D}_{\mathbf{n}}^{-}(\beta,x_i)\right) < 2\varepsilon_{\mathbf{n}}\right) + \mathbb{P}\left(\min_{1\leq i\leq \mathrm{N}_{\varepsilon_{\mathbf{n}}}}\left(\mathrm{H}_{\mathbf{n},x_i} - \mathrm{D}_{\mathbf{n}}^{+}(\beta,x_i)\right) > -2\varepsilon_{\mathbf{n}}\right) \\ &\leq \mathrm{N}_{\varepsilon_{\mathbf{n}}}\max_{1\leq i\leq \mathrm{N}_{\varepsilon_{\mathbf{n}}}}\mathbb{P}\left(\mathrm{H}_{\mathbf{n},x_i} < \mathrm{D}_{\mathbf{n}}^{-}(\beta,x_i) + 2\varepsilon_{\mathbf{n}}\right) \\ &\qquad + \mathrm{N}_{\varepsilon_{\mathbf{n}}}\max_{1\leq i\leq \mathrm{N}_{\varepsilon_{\mathbf{n}}}}\mathbb{P}\left(\mathrm{H}_{\mathbf{n},x_i} > \mathrm{D}_{\mathbf{n}}^{+}(\beta,x_i) - 2\varepsilon_{\mathbf{n}}\right). \end{aligned} \tag{0.9}$$

Let us evaluate the first term in the right-hand side of (0.9), without ambiguity we ignore the i index in x_i. As justified in the following

$$\begin{aligned}
\mathbb{P}\left(\mathrm{H}_{\mathbf{n},x} < \mathrm{D}_{\mathbf{n}}^{-}(\beta,x)+2\varepsilon_{\mathbf{n}}\right) &\leq \mathbb{P}\left(\sum_{\mathbf{i}\in\mathcal{V}_{\mathbf{s_0}}} \mathbb{I}_{\mathrm{B}(x,\mathrm{D}_{\mathbf{n}}^{-}(\beta,x)+2\varepsilon_{\mathbf{n}})}(\mathrm{X}_{\mathbf{i}}) > k_{\mathbf{n}}\right) &(0.10)\\
&\leq \mathbb{P}\left(\sum_{\mathbf{i}\in\mathcal{V}_{\mathbf{s_0}}} \xi_{\mathbf{i}} > k_{\mathbf{n}} - k_{\mathbf{n}}^{'}\delta_{\mathbf{n}}\right) &(0.11)\\
&\leq \mathbb{P}\left(\sum_{\mathbf{i}\in\mathcal{V}_{\mathbf{s_0}}} \xi_{\mathbf{i}} > \mathrm{C}k_{\mathbf{n}}(1-\beta^{1/2})\right) := \mathrm{P}_{1,\mathbf{n}} &(0.12)
\end{aligned}$$

where $\xi_{\mathbf{i}} = \Lambda_{\mathbf{i}} - \delta_{\mathbf{n},\mathbf{i}}$ is centered and $\Lambda_{\mathbf{i}}$ is defined in Lemma 4 when we replace $\mathrm{D}_{\mathbf{n}}$ by $\mathrm{D}_{\mathbf{n}}^{-}+2\varepsilon_{\mathbf{n}}$. From (0.10), we get (0.11) by (0.6) while result (0.11) permits to get (0.12) by the help of the following.
Actually, according to the definition of $\mathrm{D}_{\mathbf{n}}^{-}$ in (0.1) and replacing $\mathrm{D}_{\mathbf{n}}$ by $\mathrm{D}_{\mathbf{n}}^{-}+2\varepsilon_{\mathbf{n}}$ in (0.4), we get

$$k_{\mathbf{n}}^{'}\delta_{\mathbf{n}} - k_{\mathbf{n}}\left(\varepsilon_0(cf(x))^{1/d}+\beta^{1/2d}\right)^{d} = o(k_{\mathbf{n}}), \tag{0.13}$$

therefore, for all $\varepsilon_1 > 0$,

$$k_{\mathbf{n}} - k_{\mathbf{n}}^{'}\delta_{\mathbf{n}} > k_{\mathbf{n}}\left(1-\left(\varepsilon_0(cf(x))^{1/d}+\beta^{1/2d}\right)^{d}-\varepsilon_1\right).$$

Then, for ε_1 and ε_0 very small such that $1-\left(\varepsilon_0(cf(x))^{1/d}+\beta^{1/2d}\right)^{d}-\varepsilon_1 > 0$, we can find some constant $\mathrm{C} > 0$ such that

$$k_{\mathbf{n}} - k_{\mathbf{n}}^{'}\delta_{\mathbf{n}} > \mathrm{C}k_{\mathbf{n}}(1-\beta^{1/2}). \tag{0.14}$$

For the second term in the right-hand side of (0.9),

$$\begin{aligned}
\mathbb{P}\left(\mathrm{H}_{\mathbf{n},x} > \mathrm{D}_{\mathbf{n}}^{+}(\beta,x)-2\varepsilon_{\mathbf{n}}\right) &\leq \mathbb{P}\left(\sum_{\mathbf{i}\in\mathcal{V}_{\mathbf{s_0}}} \mathbb{I}_{\mathrm{B}(x,\mathrm{D}_{\mathbf{n}}^{+}(\beta,x)-2\varepsilon_{\mathbf{n}})}(\mathrm{X}_{\mathbf{i}}) < k_{\mathbf{n}}\right) &(0.15)\\
&\leq \mathbb{P}\left(\sum_{\mathbf{i}\in\mathcal{V}_{\mathbf{s_0}}} \Delta_{\mathbf{i}} > k_{\mathbf{n}}^{'}\delta_{\mathbf{n}} - k_{\mathbf{n}}\right) &(0.16)\\
&\leq \mathbb{P}\left(\sum_{\mathbf{i}\in\mathcal{V}_{\mathbf{s_0}}} \Delta_{\mathbf{i}} > \mathrm{C}k_{\mathbf{n}}\left(\beta^{-1/2}-1\right)\right) := \mathrm{P}_{2,\mathbf{n}} &(0.17)
\end{aligned}$$

where $\Delta_{\mathbf{i}} = \delta_{\mathbf{n},\mathbf{i}} - \Lambda_{\mathbf{i}}$ is centered and $\Lambda_{\mathbf{i}}$ is defined in Lemma 4 replacing $\mathrm{D}_{\mathbf{n}}$ by $\mathrm{D}_{\mathbf{n}}^{+}-2\varepsilon_{\mathbf{n}}$. Result (0.16) is obtained by (0.6) while that of (0.17) is obtained by replacing $\mathrm{D}_{\mathbf{n}}$ in (0.4) by $\mathrm{D}_{\mathbf{n}}^{+}-2\varepsilon_{\mathbf{n}}$. Then, we get

$$k_{\mathbf{n}}^{'}\delta_{\mathbf{n}} - k_{\mathbf{n}}\left(\beta^{-1/2d}-\varepsilon_0(cf(x))^{1/d}\right)^{d} = o(k_{\mathbf{n}}). \tag{0.18}$$

Thus for all $\varepsilon_2 > 0$, it is easy to see that

$$k_{\mathbf{n}}^{'}\delta_{\mathbf{n}} - k_{\mathbf{n}} > k_{\mathbf{n}}\left(\left(\beta^{-1/2d}-\varepsilon_0(cf(x))^{1/d}\right)^{d}-1-\varepsilon_2\right),$$

so for ε_2 and ε_0 small enough such that $\left(\left(\beta^{-1/2d}-\varepsilon_0(cf(x))^{1/d}\right)^{d}-1-\varepsilon_2\right) > 0$, then there exists $\mathrm{C} > 0$ such that

$$k_{\mathbf{n}}^{'}\delta_{\mathbf{n}} - k_{\mathbf{n}} > \mathrm{C}k_{\mathbf{n}}\left(\beta^{-1/2}-1\right). \tag{0.19}$$

Now, it suffices to prove that

$$\sum_{\mathbf{n}\in\mathbb{N}^{*\mathrm{N}}} \mathrm{N}_{\varepsilon_{\mathbf{n}}}\mathrm{P}_{1,\mathbf{n}} < \infty \quad \text{and} \quad \sum_{\mathbf{n}\in\mathbb{N}^{*\mathrm{N}}} \mathrm{N}_{\varepsilon_{\mathbf{n}}}\mathrm{P}_{2,\mathbf{n}} < \infty.$$

Let us consider $\mathrm{P}_{1,\mathbf{n}}$

This proof is based on the classical spatial block decomposition of the sum on $\xi_{\mathbf{i}}$ in $\mathcal{V}_{\mathbf{s_0}}$ similarly to [341]. Let $\mathcal{G}_{\mathbf{n}} \subset \mathcal{I}_{\mathbf{n}}$ be the smallest rectangular grid of center $\mathbf{s_0}$ containing $\mathcal{V}_{\mathbf{s_0}}$. Without loss of generality, we assume that $\mathcal{G}_{\mathbf{n}}$ is defined via some $\mathbf{k} = (k_1,\ldots,k_{\mathrm{N}})$ where $1 \leq k_j \leq n_j$, $j = 1,\ldots,\mathrm{N}$. However, by construction

$\mathscr{G}_{\mathbf{n}}$ is of cardinal $\hat{\mathbf{k}} = k_1 \times \cdots \times k_{\mathrm{N}}$ satisfying $k'_{\mathbf{n}} = \mathscr{O}(\hat{\mathbf{k}})$. In addition, we assume that $k_l = 2bq_l$, $l = 1, \ldots, \mathrm{N}$, where q_l and b are positive integers. Then the decomposition can be presented as follows

$$\begin{aligned}
\mathrm{U}(1,\mathbf{k},\mathbf{j}) &= \sum_{\substack{i_l=2j_lb+1,\\ k=1,\ldots,\mathrm{N}.}}^{(2j_l+1)b} \xi_{\mathbf{i}} \\
\mathrm{U}(2,\mathbf{k},\mathbf{j}) &= \sum_{\substack{i_l=2j_lb+1,\\ l=1,\ldots,\mathrm{N}-1.}}^{(2j_l+1)b} \sum_{i_{\mathrm{N}}=(2j_{\mathrm{N}}+1)b+1,}^{2(j_{\mathrm{N}}+1)b} \xi_{\mathbf{i}} \\
\mathrm{U}(3,\mathbf{k},\mathbf{j}) &= \sum_{\substack{i_l=2j_lb+1,\\ l=1,\ldots,\mathrm{N}-2.}}^{(2j_l+1)b} \sum_{i_{\mathrm{N}-1}=(2j_{\mathrm{N}-1}+1)b+1,}^{2(j_{\mathrm{N}-1}+1)b} \sum_{i_{\mathrm{N}}=2j_{\mathrm{N}}b+1,}^{(2j_{\mathrm{N}}+1)b} \xi_{\mathbf{i}} \\
\mathrm{U}(4,\mathbf{k},\mathbf{j}) &= \sum_{\substack{i_l=2j_lb+1,\\ l=1,\ldots,\mathrm{N}-2.}}^{(2j_l+1)b} \sum_{i_{\mathrm{N}-1}=(2j_{\mathrm{N}-1}+1)b+1,}^{2(j_{\mathrm{N}-1}+1)b} \sum_{i_{\mathrm{N}}=(2j_{\mathrm{N}}+1)b+1,}^{(2j_{\mathrm{N}}+1)b} \xi_{\mathbf{i}} \\
&\ldots
\end{aligned}$$

Note that

$$\mathrm{U}(2^{\mathrm{N}-1},\mathbf{k},\mathbf{j}) = \sum_{\substack{i_l=(2j_l+1)b+1,\\ l=1,\ldots,\mathrm{N}-1.}}^{2(j_l+1)b} \sum_{i_{\mathrm{N}}=2j_{\mathrm{N}}b+1,}^{(2j_{\mathrm{N}}+1)b} \xi_{\mathbf{i}}$$

and that

$$\mathrm{U}(2^{\mathrm{N}},\mathbf{k},\mathbf{j}) = \sum_{\substack{i_l=(2j_l+1)b+1,\\ l=1,\ldots,\mathrm{N}.}}^{2(j_l+1)b} \xi_{\mathbf{i}}.$$

For each integer $1 \le l \le 2^{\mathrm{N}}$, let

$$\mathrm{T}(\mathbf{k},l) = \sum_{\substack{j_l=0\\ l=1,\ldots,\mathrm{N}}}^{q_l-1} \mathrm{U}(l,\mathbf{k},\mathbf{j}).$$

Therefore, we have

$$\sum_{\mathbf{i}\in\mathscr{V}_{\mathbf{s_0}}} \xi_{\mathbf{i}} = \sum_{l=1}^{2^{\mathrm{N}}} \mathrm{T}(\mathbf{k},l). \tag{0.20}$$

It follows that

$$\mathrm{P}_{1,\mathbf{n}} = \mathbb{P}\left(\sum_{l=1}^{2^{\mathrm{N}}} \mathrm{T}(\mathbf{k},l) > \mathrm{C}k_{\mathbf{n}}(1-\sqrt{\beta})\right) \le 2^{\mathrm{N}}\mathbb{P}\left(\mid \mathrm{T}(\mathbf{k},1) \mid > \frac{\mathrm{C}k_{\mathbf{n}}(1-\sqrt{\beta})}{2^{\mathrm{N}}}\right).$$

We enumerate in an arbitrary manner the $\hat{\mathbf{q}} = q_1 \times \ldots \times q_{\mathrm{N}}$ terms $\mathrm{U}(1,\mathbf{k},\mathbf{j})$ of the sum $\mathrm{T}(\mathbf{k},1)$ and denote them $\mathrm{W}_1, \ldots, \mathrm{W}_{\hat{\mathbf{q}}}$. Notice that, $\mathrm{U}(1,\mathbf{k},\mathbf{j})$ is measurable with respect to the field generated by the $\mathrm{Z}_{\mathbf{i}}$ with $\mathbf{i} \in \mathbf{I}(\mathbf{k},\mathbf{j}) = \{\mathbf{i} \in \mathscr{G}_{\mathbf{n}} \mid 2j_lb+1 \le i_l \le (2j_l+1)b,\ l = 1, \ldots, \mathrm{N}\}$, the set $\mathbf{I}(\mathbf{k},\mathbf{j})$ contains b^{N} sites and $\mathrm{dist}(\mathbf{I}(\mathbf{k},\mathbf{j}),\mathbf{I}(\mathbf{k},\mathbf{j}'')) > b$. In addition, we have $\mid \mathrm{W}_l \mid \le b^{\mathrm{N}}$.
According to Lemma 4.5 of [66], one can find a sequence of independent random variables $\mathrm{W}_1^*, \ldots, \mathrm{W}_{\hat{\mathbf{q}}}^*$ where W_l has the same distribution as W_l^* and :

$$\sum_{l=1}^{\hat{\mathbf{q}}} \mathbb{E}(\mid \mathrm{W}_l - \mathrm{W}_l^* \mid) \le 4\hat{\mathbf{q}}b^{\mathrm{N}}\psi((\hat{\mathbf{q}}-1)b^{\mathrm{N}}, b^{\mathrm{N}})\varphi(b).$$

Then, we can write

$$\begin{aligned}
\mathrm{P}_{1,\mathbf{n}} &\le 2^{\mathrm{N}}\mathbb{P}\left(\mid \mathrm{T}(\mathbf{n},1) \mid > \frac{\mathrm{C}k_{\mathbf{n}}(1-\sqrt{\beta})}{2^{\mathrm{N}}}\right) \\
&\le 2^{\mathrm{N}}\mathbb{P}\left(\mid \sum_{l=1}^{\hat{\mathbf{q}}} \mathrm{W}_l \mid > \frac{\mathrm{C}k_{\mathbf{n}}(1-\sqrt{\beta})}{2^{\mathrm{N}}}\right) \\
&\le 2^{\mathrm{N}}\mathbb{P}\left(\sum_{l=1}^{\hat{\mathbf{q}}} \mid \mathrm{W}_l - \mathrm{W}_l^* \mid > \frac{\mathrm{C}k_{\mathbf{n}}(1-\sqrt{\beta})}{2^{\mathrm{N}+1}}\right) \\
&\quad + 2^{\mathrm{N}}\mathbb{P}\left(\sum_{l=1}^{\hat{\mathbf{q}}} \mid \mathrm{W}_l^* \mid > \frac{\mathrm{C}k_{\mathbf{n}}(1-\sqrt{\beta})}{2^{\mathrm{N}+1}}\right).
\end{aligned}$$

Let $\mathrm{P}_{11,\mathbf{n}} = \mathbb{P}\left(\sum_{l=1}^{\hat{\mathbf{q}}} \mid \mathrm{W}_l - \mathrm{W}_l^* \mid > \frac{\mathrm{C}k_{\mathbf{n}}(1-\sqrt{\beta})}{2^{\mathrm{N}+1}}\right)$ and $\mathrm{P}_{12,\mathbf{n}} = \mathbb{P}\left(\sum_{l=1}^{\hat{\mathbf{q}}} \mid \mathrm{W}_l^* \mid > \frac{\mathrm{C}k_{\mathbf{n}}(1-\sqrt{\beta})}{2^{\mathrm{N}+1}}\right)$.
It suffices to show that $\sum_{\mathbf{n}\in\mathbb{N}^{*\mathrm{N}}} \mathrm{P}_{11,\mathbf{n}} < \infty$ and $\sum_{\mathbf{n}\in\mathbb{N}^{*\mathrm{N}}} \mathrm{P}_{12,\mathbf{n}} < \infty$.

Let us consider first $\mathrm{P}_{11,\mathbf{n}}$

Using Markov's inequality, we get

$$\begin{aligned}
\mathrm{P}_{11,\mathbf{n}} &= \mathbb{P}\left(\sum_{l=1}^{\hat{\mathbf{q}}} \mid \mathrm{W}_l - \mathrm{W}_l^* \mid > \frac{\mathrm{C}k_{\mathbf{n}}(1-\sqrt{\beta})}{2^{\mathrm{N}+1}}\right) \\
&\leq \frac{2^{\mathrm{N}+3}}{\mathrm{C}k_{\mathbf{n}}(1-\sqrt{\beta})}\hat{\mathbf{q}}b^{\mathrm{N}}\psi((\hat{\mathbf{q}}-1)b^{\mathrm{N}}, b^{\mathrm{N}})\varphi(b) \\
&\leq \frac{\mathrm{C}}{k_{\mathbf{n}}(1-\sqrt{\beta})}k_{\mathbf{n}}^{'}\psi((\hat{\mathbf{q}}-1)b^{\mathrm{N}}, b^{\mathrm{N}})\varphi(b),
\end{aligned}$$

because $\hat{\mathbf{k}} = 2^{\mathrm{N}}\hat{\mathbf{q}}b^{\mathrm{N}}$ by definition and $k_{\mathbf{n}}^{'} = \mathscr{O}(\hat{\mathbf{k}})$
Let us consider that

$$b^{\mathrm{N}} = \mathscr{O}\left(\hat{\mathbf{n}}^{2(1-s(1-\tilde{\gamma}))/a}\right), \tag{0.21}$$

where $a = 2 + (2 + s(2-\tilde{\gamma}))d + s(4 + 2\tilde{\beta} + 2\gamma - 3\tilde{\gamma})$.
Under the assumption on the function $\psi(n,m)$, we distinguish the following two cases :

Case 1

$$\psi(n,m) \leq \mathrm{C}\min(n,m) \text{ with } (1-s(1-\tilde{\gamma}))\theta > \mathrm{N}\left\{(2+s(2-\tilde{\gamma}))d + 2s(2+\gamma-\tilde{\gamma})\right\}, \text{ and } 2 < s < \frac{1}{1-\tilde{\gamma}}.$$

In this case, we have

$$\mathrm{P}_{11,\mathbf{n}} \leq \mathrm{C}\frac{k_{\mathbf{n}}^{'}}{k_{\mathbf{n}}}b^{\mathrm{N}}\varphi(b) \leq \mathrm{C}\frac{k_{\mathbf{n}}^{'}}{k_{\mathbf{n}}}b^{\mathrm{N}-\theta}.$$

Then by using (0.21) and the definition of $\mathrm{N}_{\varepsilon_{\mathbf{n}}}$, we have

$$\mathrm{N}_{\varepsilon_{\mathbf{n}}}\mathrm{P}_{11,\mathbf{n}} \leq \mathrm{C}\hat{\mathbf{n}}^{-2\left(1-\frac{3+s(2\tilde{\beta}-1)}{a}\right)}.$$

One can show that $a > 2(3 + s(2\tilde{\beta}-1))$ and then $\sum_{\mathbf{n}\in\mathbb{N}^{*\mathrm{N}}} \mathrm{N}_{\varepsilon_{\mathbf{n}}}\mathrm{P}_{11,\mathbf{n}} < \infty$.

Case 2

$$\psi(n,m) \leq \mathrm{C}(n+m+1)^{\tilde{\beta}} \text{ with } (1-s(1-\tilde{\gamma}))\theta > \mathrm{N}\left\{2 + (2+s(2-\tilde{\gamma}))d + s(4+2\tilde{\beta}+2\gamma-3\tilde{\gamma})\right\} \text{ and}$$

$2 < s < \frac{1}{1-\tilde{\gamma}}$. In this case, we have

$$\mathrm{P}_{11,\mathbf{n}} \leq \mathrm{C}\frac{k_{\mathbf{n}}^{'}}{k_{\mathbf{n}}}(k_{\mathbf{n}}^{'}b^{\mathrm{N}})^{\tilde{\beta}}\varphi(b) \leq \mathrm{C}\frac{k_{\mathbf{n}}^{'}}{k_{\mathbf{n}}}k_{\mathbf{n}}^{'\tilde{\beta}}b^{-\theta} \leq \mathrm{C}\hat{\mathbf{n}}^{-(2-\gamma\tilde{\beta})}.$$

Then, it follows that $\sum_{\mathbf{n}\in\mathbb{N}^{\mathrm{N}}} \mathrm{N}_{\varepsilon_{\mathbf{n}}}\mathrm{P}_{11,\mathbf{n}} < \infty$ when $\tilde{\beta} < 1/\gamma$.

Let us consider $\mathrm{P}_{12,\mathbf{n}}$

Applying Markov's inequality, we have for $t > 0$:

$$\begin{aligned}
\mathrm{P}_{12,\mathbf{n}} &= \mathbb{P}\left(\sum_{l=1}^{\hat{\mathbf{q}}} \mid \mathrm{W}_l^* \mid > \frac{\mathrm{C}k_{\mathbf{n}}(1-\sqrt{\beta})}{2^{\mathrm{N}+1}}\right) \\
&\leq \exp\left(-t\frac{\mathrm{C}k_{\mathbf{n}}(1-\sqrt{\beta})}{2^{\mathrm{N}+1}}\right)\mathbb{E}\left(\exp\left(t\sum_{l=1}^{\hat{\mathbf{q}}}\mathrm{W}_l^*\right)\right) \\
&\leq \exp\left(-t\frac{\mathrm{C}k_{\mathbf{n}}(1-\sqrt{\beta})}{2^{\mathrm{N}+1}}\right)\prod_{l=1}^{\hat{\mathbf{q}}}\mathbb{E}\left(\exp\left(t\mathrm{W}_l^*\right)\right),
\end{aligned}$$

since the variables $W_1^*,\dots,W_{\hat{\mathbf{q}}}^*$ are independent.
Let $r>0$, for $t=\dfrac{r\log(\hat{\mathbf{n}})}{k_{\mathbf{n}}}$, $l=1,\dots,\hat{\mathbf{q}}$, by using (0.21), we can easily get

$$\begin{aligned} t\,|\,W_l^*\,| &\leq \frac{r\log(\hat{\mathbf{n}})}{k_{\mathbf{n}}}b^{N}\leq C\frac{\log(\hat{\mathbf{n}})}{k_{\mathbf{n}}}\hat{\mathbf{n}}^{2(1-s(1\tilde{\gamma}))/a}\\ &\leq C\frac{\log(\hat{\mathbf{n}})}{\hat{\mathbf{n}}^{\tilde{a}/a}}, \end{aligned}$$

where $\tilde{a}=a\tilde{\gamma}-2(1-s(1-\tilde{\gamma}))>0$ and $\tilde{a}>0$. However, we have $t\,|\,W_l^*\,|<1$ for $\mathbf{n}$ large enough. So, $\exp\left(tW_l^*\right)\leq 1+tW_l^*+t^2W_l^{*2}$ then

$$\mathbb{E}\left(\exp\left(tW_l^*\right)\right)\leq 1+\mathbb{E}\left(t^2W_l^{*2}\right)\leq\exp\left(\mathbb{E}\left(t^2W_l^{*2}\right)\right).$$

Therefore,

$$\prod_{l=1}^{\hat{\mathbf{q}}}\mathbb{E}\left(\exp\left(tW_l^*\right)\right)\leq\exp\left(t^2\sum_{l=1}^{\hat{\mathbf{q}}}\mathbb{E}\left(W_l^{*2}\right)\right).$$

As W_l^* and W_l have the same distribution, we have

$$\sum_{l=1}^{\hat{\mathbf{q}}}\mathbb{E}\left((W_l^*)^2\right)=Var\left(\sum_{l=1}^{\hat{\mathbf{q}}}W_l^*\right)=Var\left(\sum_{l=1}^{\hat{\mathbf{q}}}W_l\right)\leq S_{\mathbf{n}}+R_{\mathbf{n}}.$$

From Lemma 4, we obtain

$$\prod_{l=1}^{\hat{\mathbf{q}}}\mathbb{E}\left(\exp\left(tW_l^*\right)\right)\leq\exp\left(Ct^2k_{\mathbf{n}}\right)\leq\exp\left(Cr^2\frac{\log(\hat{\mathbf{n}})^2}{k_{\mathbf{n}}}\right)\longrightarrow 1,$$

because $\log(\hat{\mathbf{n}})^2/k_{\mathbf{n}}\to 0$ as $\mathbf{n}\to\infty$.
Then, we deduce that

$$\begin{aligned} P_{12,\mathbf{n}} &\leq C\exp\left(-t\frac{Ck_{\mathbf{n}}(1-\sqrt{\beta})}{2^{N+1}}\right)\\ &\leq C\exp\left(-\frac{rC(1-\sqrt{\beta})}{2^{N+1}}\log(\hat{\mathbf{n}})\right)\leq C\hat{\mathbf{n}}^{-\frac{rC(1-\sqrt{\beta})}{2^{N+1}}}. \end{aligned}$$

Then, we have

$$N_{\varepsilon_{\mathbf{n}}}P_{12,\mathbf{n}}<C\hat{\mathbf{n}}^{\gamma-\tilde{\gamma}-\frac{rC(1-\sqrt{\beta})}{2^{N+1}}}.$$

Therefore, for some $r>0$ such that $\dfrac{rC(1-\sqrt{\beta})}{2^{N+1}}\tilde{\gamma}-\gamma>1$, we get

$$\sum_{\mathbf{n}\in\mathbb{N}^N}N_{\varepsilon_{\mathbf{n}}}P_{12,\mathbf{n}}<\infty.$$

By combining the two results on $P_{11,\mathbf{n}}$ and $P_{12,\mathbf{n}}$, we get $\sum_{\mathbf{n}\in\mathbb{N}^N}N_{\varepsilon_{\mathbf{n}}}P_{1,\mathbf{n}}<\infty$.
Using similar arguments, note that $\sum_{\mathbf{n}\in\mathbb{N}^N}N_{\varepsilon_{\mathbf{n}}}P_{2,\mathbf{n}}<\infty$.

Now the check of conditions (L_2), (L_3), (L_2') and (L_3') is based on Theorem 3.1 in [102]. We need to show that $D_{\mathbf{n}}^-(\beta,x)$, $D_{\mathbf{n}}^+(\beta,x)$ satisfy assumptions (**H6**) and (**H7**) used by these authors for all $(\beta,x)\in]0,1[\times D$. This is proved in the following lemmas where without ambiguity $D_{\mathbf{n}}$ will denote $D_{\mathbf{n}}^-(\beta,x)$ or $D_{\mathbf{n}}^+(\beta,x)$.

Lemma 5. *Under assumption (**H2**) and (**H6**) on $\psi(.)$, we have*

$$\hat{\mathbf{n}}D_{\mathbf{n}}^{d\theta_0}h_{\mathbf{n},\mathbf{s}_0}^{N\theta_1}\log(\hat{\mathbf{n}})^{-\theta_2}u_{\mathbf{n}}^{-\theta_3}\to\infty$$

with

$$\theta_0=\frac{s(\theta+N(d+2))}{\theta-N(s(d+4)+2d)};\qquad \theta_1=\frac{s(\theta+Nd)}{\theta-N(s(d+4)+2d)},$$
$$\theta_2=\frac{s(\theta-N(d+2))}{\theta-N(s(d+4)+2d)};\qquad \theta_3=\frac{2(\theta+N(d+s))}{\theta-N(s(d+4)+2d)},$$

and $u_{\mathbf{n}}=\prod_{i=1}^{N}\left(\log(\log(n_i))\right)^{1+\varepsilon}\log(n_i)$ *for all* $\varepsilon>0$.

Proof of Lemma 5

By the definition of $\mathrm{D}_{\mathbf{n}}$ in Lemma 4, hypotheses (**H2**) and (**H6**), we have

$$\begin{aligned}\hat{\mathbf{n}}\mathrm{D}_{\mathbf{n}}^{d\theta_0} h_{\mathbf{n},\mathbf{s}_0}^{\mathrm{N}\theta_1} \log(\hat{\mathbf{n}})^{-\theta_2} u_{\mathbf{n}}^{-\theta_3} &\geq \mathrm{C}\hat{\mathbf{n}} \left(\frac{k_{\mathbf{n}}}{k'_{\mathbf{n}}}\right)^{\theta_0} \left(\frac{k'_{\mathbf{n}}}{\hat{\mathbf{n}}}\right)^{\theta_1} \log(\hat{\mathbf{n}})^{-\theta_2} u_{\mathbf{n}}^{-\theta_3} \\ &\geq \mathrm{C}\frac{\hat{\mathbf{n}}^{1-(\gamma-\tilde{\gamma})\theta_0-(1-\gamma)\theta_1}}{\log(\hat{\mathbf{n}})^{\theta_2} u_{\mathbf{n}}^{\theta_3}}.\end{aligned}$$

Note that $u_{\mathbf{n}} \leq \log(\tilde{n})^{\mathrm{N}(2+\varepsilon)} \Rightarrow \frac{1}{u_{\mathbf{n}}^{\theta_3}} \geq \frac{1}{\log(\tilde{n})^{(2+\varepsilon)\mathrm{N}\theta_3}}$, where $\tilde{n} = \max_{k=1,\ldots,\mathrm{N}} n_k$, and

$$\log(\hat{\mathbf{n}}) \leq \mathrm{C}\log(\tilde{n}) \Rightarrow \frac{1}{\log(\hat{\mathbf{n}})^{\theta_2}} \geq \mathrm{C}\frac{1}{\log(\tilde{n})^{\theta_2}}.$$

Since $\frac{n_k}{n_i} \leq \mathrm{C},\ \forall\ 1 \leq k, i \leq \mathrm{N}$, we deduce that $\hat{\mathbf{n}} \geq \mathrm{C}\tilde{n}^{\mathrm{N}}$ and

$$\hat{\mathbf{n}}\mathrm{D}_{\mathbf{n}}^{d\theta_0} h_{\mathbf{n},\mathbf{s}_0}^{\mathrm{N}\theta_1} \log(\hat{\mathbf{n}})^{-\theta_2} u_{\mathbf{n}}^{-\theta_3} \geq \mathrm{C}\frac{\tilde{n}^{\mathrm{N}\left(1-(\gamma-\tilde{\gamma})\theta_0-(1-\gamma)\theta_1\right)}}{\log(\tilde{n})^{\theta_2+\mathrm{N}\theta_3(\varepsilon+2)}} \qquad \to +\infty$$

because $(1-s(1-\tilde{\gamma}))\theta > \mathrm{N}\big((2+s(2-\tilde{\gamma}))d + 2s(2+\gamma-\tilde{\gamma})\big)$.

Lemma 6. *Under assumption (**H2**) and (**H7**) on $\psi(.)$, we have*

$$\hat{\mathbf{n}}\mathrm{D}_{\mathbf{n}}^{d\theta'_0} h_{\mathbf{n},\mathbf{s}_0}^{\mathrm{N}\theta'_1} \log(\hat{\mathbf{n}})^{-\theta'_2} u_{\mathbf{n}}^{-\theta'_3} \to \infty,$$

with

$$\theta'_0 = \frac{s(\theta+\mathrm{N}(d+3))}{\theta-\mathrm{N}\left(s(d+3+2\tilde{\beta})+2(d+1)\right)};\quad \theta'_1 = \frac{s(\theta+\mathrm{N}(d+1))}{\theta-\mathrm{N}\left(s(d+3+2\tilde{\beta})+2(d+1)\right)}$$

$$\theta'_2 = \frac{s(\theta-\mathrm{N}(d+1))}{\theta-\mathrm{N}\left(s(d+3+2\tilde{\beta})+2(d+1)\right)};\quad \theta'_3 = \frac{2\,(\theta+\mathrm{N}(s+d+1))}{\theta-\mathrm{N}\left(s(d+3+2\tilde{\beta})+2(d+1)\right)}.$$

The proof of this lemma is the same as the one of Lemma 5 and is omitted.

Verification of (L_2)

Let

$$f_{\mathbf{n}}\left(x, \mathrm{D}_{\mathbf{n}}^{-}(\beta,x)\right) = \frac{1}{\hat{\mathbf{n}} h_{\mathbf{n},\mathbf{s}_0}^{\mathrm{N}} \left(\mathrm{D}_{\mathbf{n}}^{-}(\beta,x)\right)^{d}} \sum_{\mathbf{i}\in\mathscr{I}_{\mathbf{n}},\mathbf{i}\neq\mathbf{s}_0} \mathrm{K}_1\left(\frac{x-\mathrm{X}_{\mathbf{i}}}{\mathrm{D}_{\mathbf{n}}^{-}(\beta,x)}\right) \mathrm{K}_2\left(h_{\mathbf{n},\mathbf{s}_0}^{-1}\left\|\frac{\mathbf{s}_0-\mathbf{i}}{\mathbf{n}}\right\|\right),$$

and

$$f_{\mathbf{n}}\left(x, \mathrm{D}_{\mathbf{n}}^{+}(\beta,x)\right) = \frac{1}{\hat{\mathbf{n}} h_{\mathbf{n},\mathbf{s}_0}^{\mathrm{N}} \left(\mathrm{D}_{\mathbf{n}}^{+}(\beta,x)\right)^{d}} \sum_{\mathbf{i}\in\mathscr{I}_{\mathbf{n}},\mathbf{i}\neq\mathbf{s}_0} \mathrm{K}_1\left(\frac{x-\mathrm{X}_{\mathbf{i}}}{\mathrm{D}_{\mathbf{n}}^{+}(\beta,x)}\right) \mathrm{K}_2\left(h_{\mathbf{n},\mathbf{s}_0}^{-1}\left\|\frac{\mathbf{s}_0-\mathbf{i}}{\mathbf{n}}\right\|\right).$$

Under the hypotheses of Lemma 1 and the results of Lemma 5 and Lemma 6 (see [102]), we have

$$\sup_{x\in\mathrm{D}} \left|f_{\mathbf{n}}\left(x, \mathrm{D}_{\mathbf{n}}^{-}(\beta,x)\right) - f(x)\right| \longrightarrow 0 \qquad a.co.$$

$$\sup_{x\in\mathrm{D}} \left|f_{\mathbf{n}}\left(x, \mathrm{D}_{\mathbf{n}}^{+}(\beta,x)\right) - f(x)\right| \longrightarrow 0 \qquad a.co,$$

then,

$$\sup_{x\in\mathrm{D}} \left| \frac{\sum_{\mathbf{i}\in\mathscr{I}_{\mathbf{n}},\mathbf{i}\neq\mathbf{s}_0} \mathrm{K}_1\left(\frac{x-\mathrm{X}_{\mathbf{i}}}{\mathrm{D}_{\mathbf{n}}^{-}(\beta,x)}\right) \mathrm{K}_2\left(h_{\mathbf{n},\mathbf{s}_0}^{-1}\left\|\frac{\mathbf{s}_0-\mathbf{i}}{\mathbf{n}}\right\|\right)}{\sum_{\mathbf{i}\in\mathscr{I}_{\mathbf{n}},\mathbf{i}\neq\mathbf{s}_0} \mathrm{K}_1\left(\frac{x-\mathrm{X}_{\mathbf{i}}}{\mathrm{D}_{\mathbf{n}}^{+}(\beta,x)}\right) \mathrm{K}_2\left(h_{\mathbf{n},\mathbf{s}_0}^{-1}\left\|\frac{\mathbf{s}_0-\mathbf{i}}{\mathbf{n}}\right\|\right)} - \beta \right| = \beta \sup_{x\in\mathrm{D}} \left| \frac{f_{\mathbf{n}}\left(x, \mathrm{D}_{\mathbf{n}}^{-}(\beta,x)\right)}{f_{\mathbf{n}}\left(x, \mathrm{D}_{\mathbf{n}}^{+}(\beta,x)\right)} - 1 \right| \to 0 \quad a.co.$$

Verification of (L_3)

Under assumptions of Lemma 1, Lemma 5 and 6, it follows that (see [102])

$$\sup_{x\in\mathrm{D}} |c_{\mathbf{n}}\left(\mathrm{D}_{\mathbf{n}}^{-}(\beta,x)\right) - r(x)| \to 0 \quad a.co \text{ and } \sup_{x\in\mathrm{D}} |c_{\mathbf{n}}\left(\mathrm{D}_{\mathbf{n}}^{+}(\beta,x)\right) - r(x)| \to 0 \quad a.co.$$

Proof of Lemma 2

The proof of this lemma is based on the results of Lemma 2. It suffices to check the conditions (L_2') and (L_3'). Clearly, similar arguments as those involved to prove (L_2) and (L_3) can be used to obtain the requested conditions.

Verification of (L_2')

Under assumptions of Corollary 1 and Lemmas 5, 6, we have

$$\begin{aligned}\sup_{x\in D}\left|f_{\mathbf{n}}\left(x,D_{\mathbf{n}}^{-}(\beta,x)\right)-f(x)\right| &= \mathcal{O}\left(D_{\mathbf{n}}^{-}(\beta,x)\right)+\mathcal{O}\left(\left(\frac{\log(\hat{\mathbf{n}})}{\hat{\mathbf{n}}(D_{\mathbf{n}}^{-}(\beta,x))^{d}h_{\mathbf{n},\mathbf{s_0}}^{N}}\right)^{1/2}\right) a.co.\\ &= \mathcal{O}\left(\left(\frac{k_{\mathbf{n}}}{k_{\mathbf{n}}'}\right)^{1/d}+\left(\frac{\log(\hat{\mathbf{n}})}{k_{\mathbf{n}}}\right)^{1/2}\right) a.co.,\end{aligned}$$

$$\begin{aligned}\sup_{x\in D}\left|f_{\mathbf{n}}\left(x,D_{\mathbf{n}}^{+}(\beta,x)\right)-f(x)\right| &= \mathcal{O}\left(D_{\mathbf{n}}^{+}(\beta,x)\right)+\mathcal{O}\left(\left(\frac{\log(\hat{\mathbf{n}})}{\hat{\mathbf{n}}(D_{\mathbf{n}}^{+}(\beta,x))^{d}h_{\mathbf{n},\mathbf{s_0}}^{N}}\right)^{1/2}\right) a.co.\\ &= \mathcal{O}\left(\left(\frac{k_{\mathbf{n}}}{k_{\mathbf{n}}'}\right)^{1/d}+\left(\frac{\log(\hat{\mathbf{n}})}{k_{\mathbf{n}}}\right)^{1/2}\right). a.co.\end{aligned}$$

We conclude that

$$\begin{aligned}&\sup_{x\in D}\left|\frac{\sum_{\mathbf{i}\in\mathcal{J}_{\mathbf{n}},\mathbf{i}\neq\mathbf{s_0}}K_1\left(\frac{x-X_{\mathbf{i}}}{D_{\mathbf{n}}^{-}(\beta,x)}\right)K_2\left(h_{\mathbf{n},\mathbf{s_0}}^{-1}\left\|\frac{\mathbf{s_0}-\mathbf{i}}{\mathbf{n}}\right\|\right)}{\sum_{\mathbf{i}\in\mathcal{J}_{\mathbf{n}},\mathbf{i}\neq\mathbf{s_0}}K_1\left(\frac{x-X_{\mathbf{i}}}{D_{\mathbf{n}}^{+}(\beta,x)}\right)K_2\left(h_{\mathbf{n},\mathbf{s_0}}^{-1}\left\|\frac{\mathbf{s_0}-\mathbf{i}}{\mathbf{n}}\right\|\right)}-\beta\right|\\ &\qquad=\beta\sup_{x\in D}\left|\frac{f_{\mathbf{n}}\left(x,D_{\mathbf{n}}^{-}(\beta,x)\right)}{f_{\mathbf{n}}\left(x,D_{\mathbf{n}}^{+}(\beta,x)\right)}-1\right|=\mathcal{O}\left(\left(\frac{k_{\mathbf{n}}}{k_{\mathbf{n}}'}\right)^{1/d}+\left(\frac{\log(\hat{\mathbf{n}})}{k_{\mathbf{n}}}\right)^{1/2}\right) a.co.\end{aligned}$$

Verification of (L_3')

It is relatively easy to deduce from Lemmas 5 and 6 ([102]) that

$$\sup_{x\in D}\left|c_{\mathbf{n}}\left(D_{\mathbf{n}}^{-}(\beta,x)\right)-r(x)\right|=\mathcal{O}\left(\left(\frac{k_{\mathbf{n}}}{k_{\mathbf{n}}'}\right)^{1/d}+\left(\frac{\log(\hat{\mathbf{n}})}{k_{\mathbf{n}}}\right)^{1/2}\right) a.co.$$

$$\sup_{x\in D}\left|c_{\mathbf{n}}\left(D_{\mathbf{n}}^{+}(\beta,x)\right)-r(x)\right|=\mathcal{O}\left(\left(\frac{k_{\mathbf{n}}}{k_{\mathbf{n}}'}\right)^{1/d}+\left(\frac{\log(\hat{\mathbf{n}})}{k_{\mathbf{n}}}\right)^{1/2}\right) a.co.$$

This yields the proof.

ANNEXE B

PREUVES DU CHAPITRE 4

Nous rédigeons les preuves des théorèmes 3, 4 et 5, en anglais.

Some preliminary results for the proofs

Lemma 7. *([66]) Let the sets* $S_1, S_2, ..., S_k$ *containing each* m *sites and such that, for all* $i \neq j$*, and for* $1 \leq i, j \leq k$*,* $dist(S_i, S_j) \geq \delta_0$*. Let* $W_1, W_2, \ldots, W_k$ *a sequence of random variables with real values and measurable respectively with respect to* $\mathscr{B}(S_1), \ldots, \mathscr{B}(S_k)$*. Let be* W_l *with values in* $[a, b]$*. There exists a sequence of independent random variables* $W_1^*, W_2^*, \ldots, W_k^*$ *such that* W_l^* *has the same distribution as* W_l *and satisfies :*

$$\sum_{l=1}^{k} \mathbb{E}|W_l - W_l^*| \leq 2k(b-a)\psi((k-1)m, m)\chi(\delta_0).$$

Lemma 8. *([342]) Denote by* $\mathscr{L}_r(\mathscr{F})$ *the class of* $\mathscr{F}$*-measurable random variables* X *which satisfy :* $\|X\|_r = (\mathbb{E}|X|^r)^{1/r} < \infty$*. Suppose that* $X \in \mathscr{L}_r(\mathscr{B}(E))$*,* $Y \in \mathscr{L}_r(\mathscr{B}(E'))$*,* $1 \leq r, s, t < \infty$ *and* $\frac{1}{r} + \frac{1}{s} + \frac{1}{t} = 1$*. Then,*

$$|\mathbb{E}XY - \mathbb{E}X\mathbb{E}Y| \quad \leq \quad C\|X\|_r \|Y\|_s \{\psi(Card(E), Card(E'))\chi(dist(E, E'))\}^{1/t}.$$

For bounded random variables with probability 1*, we have :*

$$|\mathbb{E}XY - \mathbb{E}X\mathbb{E}Y| \quad \leq \quad C\{\psi(Card(E), Card(E'))\chi(dist(E, E'))\}.$$

In the following, we will often use the notation $K_{\mathbf{i}}(x) = K_{1\mathbf{i}}K_{2\mathbf{i}}$ and $W_{\mathbf{ni}}(x) = \dfrac{K_{\mathbf{i}}(x)}{\sum_{\mathbf{j}\in\mathscr{O}_{\mathbf{n}}} K_{\mathbf{j}}(x)}$ with $K_{1\mathbf{i}} = K_1\left(\frac{d(x,X_{\mathbf{i}})}{b_{\mathbf{n}}}\right)$ and $K_{2\mathbf{i}} = K_{2,\rho_{\mathbf{n}}}(\|\mathbf{i}_0 - \mathbf{i}\|)$. By convention, we set $0/0 = 0$, then $\sum_{\mathbf{i}\in\mathscr{I}_{\mathbf{n}}} W_{\mathbf{ni}}(x) = 0$ or 1. Thus, we have

$$r_{\mathbf{n}}(x) \quad = \quad \begin{cases} \sum_{\mathbf{i}\in\mathscr{I}_{\mathbf{n}}} W_{\mathbf{ni}}(x) Y_{\mathbf{i}} & \text{if} \quad \sum_{\mathbf{i}\in\mathscr{O}_{\mathbf{n}}} W_{\mathbf{ni}}(x) = 1; \\ \frac{1}{\widetilde{\mathbf{n}}} \sum_{\mathbf{i}\in\mathscr{O}_{\mathbf{n}}} Y_{\mathbf{i}} & \text{otherwise.} \end{cases}$$

Let us use the following decomposition :

$$r_{\mathbf{n}}(x) - r(x) = \frac{1}{f_{\mathbf{n}}(x)}\left[(g_{\mathbf{n}}(x) - \mathbb{E}(g_{\mathbf{n}}(x))) - (r(x) - \mathbb{E}(g_{\mathbf{n}}(x)))\right] \tag{0.1}$$

$$-\frac{r(x)}{f_{\mathbf{n}}(x)}\left[f_{\mathbf{n}}(x) - 1\right]$$

Lemma 9. *Under hypotheses* ***H1-H3****, we have*

$$\mathbb{E}^{1/2}\left[\sum_{\mathbf{i}\in\mathscr{V}_{\mathbf{i}_0}} W_{\mathbf{ni}}(x)\mathbb{E}(Y_{\mathbf{i}}|X_{\mathbf{i}}) - r(x)\right]^2 \quad = \quad O(b_{\mathbf{n}}).$$

Lemma 10. *Under the conditions of Theorem 3, we have*

$$\mathbb{E}^{1/2}\left[\sum_{\mathbf{i}\in\mathcal{V}_{\mathbf{i}_0}} \mathrm{W}_{\mathbf{ni}}(x)(\mathrm{Y}_{\mathbf{i}}-\mathbb{E}(\mathrm{Y}_{\mathbf{i}}|\mathrm{X}_{\mathbf{i}}))\right]^2 = \mathrm{O}\left(\frac{1}{\widehat{\mathbf{n}}\rho_{\mathbf{n}}^{\mathrm{N}}\varphi_x(b_{\mathbf{n}})}\right)^{1/2}.$$

Lemma 11. *Under the conditions of Theorem 3, we have*

$$\mathbb{E}^{1/2}\left[\frac{1}{\widehat{\mathbf{n}}}\sum_{\mathbf{i}\in\mathcal{V}_{\mathbf{i}_0}} \mathrm{Y}_{\mathbf{i}}-r(x)\right]^2 = \mathrm{O}\left(\frac{1}{\widehat{\mathbf{n}}\rho_{\mathbf{n}}^{\mathrm{N}}\varphi_x(b_{\mathbf{n}})}\right)^{1/2}.$$

Define

$$\Lambda_{\mathbf{i}}(x)=\frac{1}{a_{\mathbf{n}}}\left[\mathrm{K}_{\mathbf{i}}(x)-\mathbb{E}(\mathrm{K}_{\mathbf{i}}(x))\right],$$

$$\mathrm{I}_{\mathbf{n}}(x)=\sum_{\mathbf{i}\in\mathscr{O}_{\mathbf{n}}}\mathbb{E}\left[(\Lambda_{\mathbf{i}}(x))^2\right] \text{ and } \mathrm{R}_{\mathbf{n}}(x)=\sum_{\mathbf{i},\mathbf{k}\in\mathscr{O}_{\mathbf{n}}}\sum_{\mathbf{i}\neq\mathbf{k}}\left|\mathbb{E}\left[\Lambda_{\mathbf{i}}(x)\Lambda_{\mathbf{k}}(x)\right]\right|.$$

Lemma 12. *Under the conditions of Theorem 3, we have*

$$\mathrm{I}_{\mathbf{n}}(x)+\mathrm{R}_{\mathbf{n}}(x) = \mathrm{O}\left(\frac{1}{\widehat{\mathbf{n}}\rho_{\mathbf{n}}^{\mathrm{N}}\varphi_x(b_{\mathbf{n}})}\right).$$

Proofs

Proof of Theorem 3

Since one of main hypothesis is local stationnarity defined in assumption (**H3**) the decomposition of quantity $r_{\mathbf{n}}(x)-r(x)$ will be doing localy over vicinity $\mathcal{V}_{\mathbf{i}_0}$. Thus we have

$$\begin{aligned} r_{\mathbf{n}}(x)-r(x) &= \left(\sum_{\mathbf{i}\in\mathcal{V}_{\mathbf{i}_0}}\mathrm{W}_{\mathbf{ni}}(x)\mathbb{E}(\mathrm{Y}_{\mathbf{i}}|\mathrm{X}_{\mathbf{i}})-r(x)\right)\mathbf{1}_{\left\{\sum_{\mathbf{i}\in\mathcal{V}_{\mathbf{i}_0}}\mathrm{W}_{\mathbf{ni}}(x)=1\right\}} \\ &\quad+\left(\sum_{\mathbf{i}\in\mathcal{V}_{\mathbf{i}_0}}\mathrm{W}_{\mathbf{ni}}(x)(\mathrm{Y}_{\mathbf{i}}-\mathbb{E}(\mathrm{Y}_{\mathbf{i}}|\mathrm{X}_{\mathbf{i}}))\right)\mathbf{1}_{\left\{\sum_{\mathbf{i}\in\mathcal{V}_{\mathbf{i}_0}}\mathrm{W}_{\mathbf{ni}}(x)=1\right\}} \\ &\quad+\left(\frac{1}{\widehat{\mathbf{n}}}\sum_{\mathbf{i}\in\mathcal{V}_{\mathbf{i}_0}}\mathrm{Y}_{\mathbf{i}}-r(x)\right)\mathbf{1}_{\left\{\sum_{\mathbf{i}\in\mathcal{V}_{\mathbf{i}_0}}\mathrm{W}_{\mathbf{ni}}(x)=0\right\}} := \mathbf{A}+\mathbf{B}+\mathbf{C}. \end{aligned}$$

Applying Minkowski's inequality, we get

$$\|r_{\mathbf{n}}(x)-r(x)\|_2 \leq \mathbb{E}^{1/2}[\mathbf{A}]^2+\mathbb{E}^{1/2}[\mathbf{B}]^2+\mathbb{E}^{1/2}[\mathbf{C}]^2. \tag{0.2}$$

Therefore, Theorem 3 follows from (0.2) and Lemmas 9, 10 and 11. □

Proof of Lemma 9

By the Lipschitz condition on Assumption (**H2**), there exists a constant $\mathrm{C}_3>0$ such that

$$\begin{aligned} \mathbb{E}^{1/2}[\mathbf{A}]^2 &\leq \mathbb{E}^{1/2}\left[\left(\sum_{\mathbf{i}\in\mathcal{V}_{\mathbf{i}_0}}\mathrm{W}_{\mathbf{ni}}(x)|r(\mathrm{X}_{\mathbf{i}})-r(x)|\right)\mathbf{1}_{\left\{\sum_{\mathbf{i}\in\mathcal{V}_{\mathbf{i}_0}}\mathrm{W}_{\mathbf{ni}}(x)=1\right\}}\right]^2 \\ &\leq \mathbb{E}^{1/2}\left[\left(\sum_{\mathbf{i}\in\mathcal{V}_{\mathbf{i}_0}}\mathrm{W}_{\mathbf{ni}}(x)(\mathrm{C}_3\times d(\mathrm{X}_{\mathbf{i}},x))\right)\mathbf{1}_{\left\{\sum_{\mathbf{i}\in\mathcal{V}_{\mathbf{i}_0}}\mathrm{W}_{\mathbf{ni}}(x)=1\right\}}\right]^2 \\ &\leq \mathrm{C}_3\mathbb{E}^{1/2}\left[\times\sum_{\mathbf{i}\in\mathcal{V}_{\mathbf{i}_0}}\mathrm{W}_{\mathbf{ni}}(x)b_{\mathbf{n}}\right]^2 \end{aligned}$$

Thus, the local stationarity assumption **H3** implies

$$\mathrm{C}_3\mathbb{E}^{1/2}\left[\times\sum_{\mathbf{i}\in\mathcal{V}_{\mathbf{i}_0}}\mathrm{W}_{\mathbf{ni}}(x)b_{\mathbf{n}}\right]^2 = \mathrm{O}(b_{\mathbf{n}}).\ \square$$

Proof of Lemma 10

Define

$$\begin{aligned} \mathrm{G}(x) &= \left(\sum_{\mathbf{i}\in\mathcal{V}_{\mathbf{i}_0}} \mathrm{W}_{\mathbf{ni}}(x)[\mathrm{Y}_{\mathbf{i}} - \mathbb{E}(\mathrm{Y}_{\mathbf{i}}|\mathrm{X}_{\mathbf{i}})]\right)\mathbf{1}_{\left\{\sum_{\mathbf{i}\in\mathcal{V}_{\mathbf{i}_0}} \mathrm{W}_{\mathbf{ni}}(x)=1\right\}} \\ &:= \frac{e_{\mathbf{n}}(x)}{f_{\mathbf{n}}(x)}\mathbf{1}_{\left\{\sum_{\mathbf{i}\in\mathcal{V}_{\mathbf{i}_0}} \mathrm{W}_{\mathbf{ni}}(x)=1\right\}}, \end{aligned}$$

where

$$e_{\mathbf{n}}(x) = \frac{1}{a_{\mathbf{n}}}\sum_{\mathbf{i}\in\mathcal{V}_{\mathbf{i}_0}} \mathrm{K}_{\mathbf{i}}(x)[\mathrm{Y}_{\mathbf{i}} - \mathbb{E}(\mathrm{Y}_{\mathbf{i}}|\mathrm{X}_{\mathbf{i}})] \quad \text{and} \quad f_{\mathbf{n}}(x) = \frac{1}{a_{\mathbf{n}}}\sum_{\mathbf{i}\in\mathcal{V}_{\mathbf{i}_0}} \mathrm{K}_{\mathbf{i}}(x).$$

Note that, since $\mathrm{Y}_{\mathbf{i}}$ is bounded, we have $\forall \mathbf{i}$, $0 \le |\mathrm{Y}_{\mathbf{i}} - \mathbb{E}(\mathrm{Y}_{\mathbf{i}}|\mathrm{X}_{\mathbf{i}})| \le 2\mathrm{M}$. It follows that $|\mathrm{G}(x)| \le 2\mathrm{M}$ and

$$\begin{aligned} |\mathrm{G}(x)| &= |\mathrm{G}(x)|\mathbf{1}_{\left\{\sum_{\mathbf{i}\in\mathcal{V}_{\mathbf{i}_0}} \mathrm{K}_{\mathbf{i}}(x)>c\right\}} + |\mathrm{G}(x)|\mathbf{1}_{\left\{\sum_{\mathbf{i}\in\mathcal{V}_{\mathbf{i}_0}} \mathrm{K}_{\mathbf{i}}(x)\le c\right\}} \\ &\le \frac{|e_{\mathbf{n}}(x)|}{f_{\mathbf{n}}(x)}\mathbf{1}_{\left\{\sum_{\mathbf{i}\in\mathcal{V}_{\mathbf{i}_0}} \mathrm{K}_{\mathbf{i}}(x)>c\right\}} + 2\mathrm{M}\times\mathbf{1}_{\left\{\sum_{\mathbf{i}\in\mathcal{V}_{\mathbf{i}_0}} \mathrm{K}_{\mathbf{i}}(x)\le c\right\}}, \end{aligned}$$

where c is a given constant. Let us take $c = \frac{a_{\mathbf{n}}}{2}$, if $\sum_{\mathbf{i}\in\mathcal{V}_{\mathbf{i}_0}} \mathrm{K}_{\mathbf{i}}(x) > c = \frac{a_{\mathbf{n}}}{2}$ then $f_{\mathbf{n}}(x) > \dfrac{a_{\mathbf{n}}}{2a_{\mathbf{n}}} > \dfrac{1}{2}$. It follows that

$$\|\mathrm{G}(x)\|_2 \le 2\|e_{\mathbf{n}}(x)\|_2 + 2\mathrm{M}\left(\mathbb{P}\left[\sum_{\mathbf{i}\in\mathcal{V}_{\mathbf{i}_0}} \mathrm{K}_{\mathbf{i}}(x) \le \frac{a_{\mathbf{n}}}{2}\right]\right)^{1/2},$$

and

$$\|e_{\mathbf{n}}(x)\|_2 = \frac{1}{a_{\mathbf{n}}}\left[\mathbb{E}\left(\sum_{\mathbf{i}\in\mathcal{V}_{\mathbf{i}_0}} \xi_{\mathbf{i}}\right)^2\right]^{1/2},$$

where

$$\xi_{\mathbf{i}} = \mathrm{K}_{\mathbf{i}}(x)\,[\mathrm{Y}_{\mathbf{i}} - \mathbb{E}(\mathrm{Y}_{\mathbf{i}}|\mathrm{X}_{\mathbf{i}})]\,.$$

To prove Lemma 10, we have to show that

$$\|e_{\mathbf{n}}(x)\|_2 = \mathrm{O}(\widehat{\mathbf{n}}\rho_{\mathbf{n}}^{\mathrm{N}}\varphi_x(b_{\mathbf{n}}))^{-1/2}, \tag{0.3}$$

and

$$\mathbb{P}\left[\sum_{\mathbf{i}\in\mathcal{V}_{\mathbf{i}_0}} \mathrm{K}_{\mathbf{i}}(x) \le \frac{a_{\mathbf{n}}}{2}\right] \le \mathrm{O}\left(\widehat{\mathbf{n}}\rho_{\mathbf{n}}^{\mathrm{N}}\varphi_x(b_{\mathbf{n}})\right)^{-1/2}. \tag{0.4}$$

Observe that, by Assumptions **H1** and **H3**, we have

$$\begin{aligned} \sum_{\mathbf{i}\in\mathcal{V}_{\mathbf{i}_0}} \mathbb{E}\left[\xi_{\mathbf{i}}^2\right] \le \sum_{\mathbf{i}\in\mathcal{V}_{\mathbf{i}_0}} \mathbb{E}\left[\mathrm{K}_{\mathbf{i}}^2(x)\,[\mathrm{Y}_{\mathbf{i}} - \mathbb{E}(\mathrm{Y}_{\mathbf{i}}|\mathrm{X}_{\mathbf{i}})]^2\right] &= 4\mathrm{M}^2\sum_{\mathbf{i}\in\mathcal{V}_{\mathbf{i}_0}} \mathrm{K}_{2\mathbf{i}}^2\mathbb{E}[\mathrm{K}_{1\mathbf{i}}]^2 \le 4\mathrm{M}^2\mathrm{C}_2^{'2}k_{\mathbf{n}}\varphi_x(b_{\mathbf{n}}) \\ &= \mathrm{O}(\widehat{\mathbf{n}}\rho_{\mathbf{n}}^{\mathrm{N}}\varphi_x(b_{\mathbf{n}})). \end{aligned}$$

Now, let $d_{\mathbf{n}}$ be a sequence of real numbers tending to ∞ as $\mathbf{n}\to\infty$ and set

$$\mathrm{S} = \{(\mathbf{i},\mathbf{k}) \in \mathcal{V}_{\mathbf{i}_0}^2,\ \|\mathbf{i}-\mathbf{k}\| \le d_{\mathbf{n}}\} \text{ and } \mathrm{S}^c = \{(\mathbf{i},\mathbf{k}) \in \mathcal{V}_{\mathbf{i}_0}^2,\ \|\mathbf{i}-\mathbf{k}\| > d_{\mathbf{n}}\}.$$

$$\text{First, see that } \mathbb{E}\left(\sum_{\mathbf{i}\in\mathcal{V}_{\mathbf{i}_0}} \xi_{\mathbf{i}}\right)^2 = \sum_{\mathbf{i}\in\mathcal{V}_{\mathbf{i}_0}} \mathbb{E}[\xi_{\mathbf{i}}^2] + \sum_{\mathbf{i},\mathbf{k}\in\mathrm{S}} \mathbb{E}\,[\xi_{\mathbf{i}}\xi_{\mathbf{k}}] + \sum_{\mathbf{i},\mathbf{k}\in\mathrm{S}^c} \mathbb{E}\,[\xi_{\mathbf{i}}\xi_{\mathbf{k}}]$$

Using Assumption **H3**, we have

$$\begin{aligned}
\sum_{\mathbf{i},\mathbf{k}\in S}\mathbb{E}[\xi_{\mathbf{i}}\xi_{\mathbf{k}}] &\leq 4M^2\sum_{\mathbf{i},\mathbf{k}\in S}\mathbb{E}[K_{\mathbf{i}}(x)K_{\mathbf{k}}(x)]\\
&\leq 4M^2\sum_{\mathbf{i},\mathbf{k}\in S}K_{2\mathbf{i}}K_{2\mathbf{k}}\mathbb{P}[(X_{\mathbf{i}},X_{\mathbf{k}})\in B(x,b_{\mathbf{n}})\times B(x,b_{\mathbf{n}})]\\
&\leq 4M^2C_4\sum_{\mathbf{i},\mathbf{k}\in S}\mathbf{1}_{[0,1]}\left(\rho_{\mathbf{n}}^{-1}\left\|\frac{\mathbf{i}_0-\mathbf{i}}{\mathbf{n}}\right\|\right)\mathbf{1}_{[0,1]}\left(\rho_{\mathbf{n}}^{-1}\left\|\frac{\mathbf{i}_0-\mathbf{k}}{\mathbf{n}}\right\|\right)\varphi_x(b_{\mathbf{n}})^{1+\varepsilon}\\
&\leq 4M^2C_4\sum_{\mathbf{i},\mathbf{k}\in\mathcal{V}_{\mathbf{i}_0}}\mathbf{1}_{[0,1]}\left(\rho_{\mathbf{n}}^{-1}\left\|\frac{\mathbf{i}_0-\mathbf{i}}{d_{\mathbf{n}}}\right\|\right)\varphi_x(b_{\mathbf{n}})^{1+\varepsilon}\\
&\leq 4M^2C_4\sum_{\mathbf{i}\in\mathcal{V}_{\mathbf{i}_0}}\sum_{\mathbf{i}-\mathbf{u}\in\mathcal{V}_{\mathbf{i}_0}}\mathbf{1}_{\{\mathbf{u};\|\mathbf{u}\|\leq d_{\mathbf{n}}\}}\left(\rho_{\mathbf{n}}^{-1}\left\|\frac{\mathbf{i}_0-\mathbf{i}}{d_{\mathbf{n}}}\right\|\right)\varphi_x(b_{\mathbf{n}})^{1+\varepsilon}\\
&\leq 4M^2C_4k_{\mathbf{n}}d_{\mathbf{n}}^N\varphi_x(b_{\mathbf{n}}))^{1+\varepsilon}.
\end{aligned}$$

Since K_1 and K_2 are bounded, applying Lemma 8, we get

$$\begin{aligned}
\sum_{\mathbf{i},\mathbf{k}\in S^c}\mathbb{E}[\xi_{\mathbf{i}}\xi_{\mathbf{k}}] &\leq C\sum_{\mathbf{i},\mathbf{k}\in S^c}\{\psi(1,1)\chi(\|\mathbf{i}-\mathbf{k}\|)\}\leq C\sum_{\mathbf{i},\mathbf{k}\in S^c\cap\mathcal{V}_{\mathbf{i}_0}}\chi(\|\mathbf{i}-\mathbf{k}\|)\leq C2^N\sum_{\mathbf{k}\in\mathcal{V}_{\mathbf{i}_0}}\sum_{\substack{\mathbf{k}-\mathbf{u}\in\mathcal{V}_{\mathbf{i}_0}\\ \|\mathbf{u}\|>d_{\mathbf{n}}}}\chi(\|\mathbf{i}\|)\\
&\leq Ck_{\mathbf{n}}\sum_{\|\mathbf{i}\|>d_{\mathbf{n}}}\chi(\|\mathbf{i}\|).
\end{aligned}$$

Since $\sum_{\|\mathbf{i}\|>d_{\mathbf{n}}}\chi(\|\mathbf{i}\|)\leq C\sum_{\|\mathbf{i}\|>d_{\mathbf{n}}}\|\mathbf{i}\|^{-\theta}\leq C\sum_{\|\mathbf{i}\|>d_{\mathbf{n}}}\|\mathbf{i}\|^{-\theta}\|\mathbf{i}\|^{-N}\|\mathbf{i}\|^{N}$, and $\|\mathbf{i}\|>d_{\mathbf{n}}$, $\|\mathbf{i}\|^{-N}\leq(d_{\mathbf{n}})^{-N}$, we have

$$C\sum_{\|\mathbf{i}\|>d_{\mathbf{n}}}\|\mathbf{i}\|^{-\theta}\|\mathbf{i}\|^{-N-\varepsilon}\|\mathbf{i}\|^{N+\varepsilon} \leq Cd_{\mathbf{n}}^{-N-\varepsilon}\sum_{\|\mathbf{i}\|>d_{\mathbf{n}}}\|\mathbf{i}\|^{N+\varepsilon-\theta}.$$

Then,

$$\sum_{\mathbf{i},\mathbf{k}\in S^c}\mathbb{E}[\xi_{\mathbf{i}}\xi_{\mathbf{k}}]\leq Ck_{\mathbf{n}}d_{\mathbf{n}}^{-N-\varepsilon}\sum_{\|\mathbf{i}\|>d_{\mathbf{n}}}\|\mathbf{i}\|^{N+\varepsilon-\theta}.$$

Choosing $d_{\mathbf{n}}=(\varphi_x(b_{\mathbf{n}}))^{\frac{-\varepsilon}{N}+a}$ with $a>0$ such that $Na\leq\varepsilon-\frac{N}{N+\varepsilon}$ lead to

$$d_{\mathbf{n}}^{-(N+\varepsilon)} = \varphi_x(b_{\mathbf{n}})(\varphi_x(b_{\mathbf{n}}))^{\frac{-(N+\varepsilon)(Na-\varepsilon)-N}{N}}=O\left(\varphi_x(b_{\mathbf{n}})\right),$$

Since $\frac{-(N+\varepsilon)(Na-\varepsilon)-N}{N}>0$, Moreover, this choice of $d_{\mathbf{n}}$ implies that

$$\begin{aligned}
\sum_{\mathbf{i},\mathbf{k}\in S}\mathbb{E}[\xi_{\mathbf{i}}\xi_{\mathbf{k}}] &\leq 4M^2C_4k_{\mathbf{n}}d_{\mathbf{n}}^N(\varphi_x(b_{\mathbf{n}}))^{1+\varepsilon}\\
&\leq 4M^2C_4k_{\mathbf{n}}(\varphi_x(b_{\mathbf{n}}))^{1+Na}=O(\widehat{\mathbf{n}}\rho_{\mathbf{n}}^N\varphi_x(b_{\mathbf{n}})).
\end{aligned}$$

Then, we deduce that

$$\mathbb{E}\left(\sum_{\mathbf{i}\in\mathcal{V}_{\mathbf{i}_0}}\xi_{\mathbf{i}}\right)^2 = \sum_{\mathbf{i}\in\mathcal{V}_{\mathbf{i}_0}}\mathbb{E}\left[\xi_{\mathbf{i}}^2\right]+\sum_{\mathbf{i},\mathbf{k}\in S}\mathbb{E}[\xi_{\mathbf{i}}\xi_{\mathbf{k}}]+\sum_{\mathbf{i},\mathbf{k}\in S^c}\mathbb{E}[\xi_{\mathbf{i}}\xi_{\mathbf{k}}]=O\left(\widehat{\mathbf{n}}\rho_{\mathbf{n}}^N\varphi_x(b_{\mathbf{n}})\right).$$

Consequently, $\left[\mathbb{E}\left(\sum_{\mathbf{i}\in\mathcal{V}_{\mathbf{i}_0}}\xi_{\mathbf{i}}\right)^2\right]^{1/2}=O(\widehat{\mathbf{n}}\rho_{\mathbf{n}}^N\varphi_x(b_{\mathbf{n}}))^{1/2}$ and $\|e_{\mathbf{n}}(x)\|_2=O\left(\widehat{\mathbf{n}}\rho_{\mathbf{n}}^N\varphi_x(b_{\mathbf{n}})\right)^{-1/2}$ since by Assumptions **H1** and **H3**(ii), $a_{\mathbf{n}}\geq C_1'C_1k_{\mathbf{n}}\varphi_x(b_{\mathbf{n}})$.
Next, for (0.4), define

$$S_{\mathbf{n}}(x)=\sum_{\mathbf{i}\in\mathcal{V}_{\mathbf{i}_0}}\Lambda_{\mathbf{i}}(x)=\frac{1}{a_{\mathbf{n}}}\left[f_{\mathbf{n}}(x)-\mathbb{E}(f_{\mathbf{n}}(x))\right].$$

Then, we have

$$\begin{aligned}
\mathbb{P}\left[\sum_{\mathbf{i}\in\mathcal{V}_{\mathbf{i}_0}}K_{\mathbf{i}}(x)\leq\frac{a_{\mathbf{n}}}{2}\right] &= \mathbb{P}\left[\sum_{\mathbf{i}\in\mathcal{V}_{\mathbf{i}_0}}(K_{\mathbf{i}}(x)-\mathbb{E}(K_{\mathbf{i}}(x)))\leq\frac{-a_{\mathbf{n}}}{2}\right]\\
&\leq \mathbb{P}\left[\frac{1}{a_{\mathbf{n}}}\left|\sum_{\mathbf{i}\in\mathcal{V}_{\mathbf{i}_0}}(K_{\mathbf{i}}(x)-\mathbb{E}(K_{\mathbf{i}}(x)))\right|\geq\frac{1}{2}\right]\\
&\leq \mathbb{P}\left[|S_{\mathbf{n}}(x)|\geq\epsilon\right], \text{for } \mathbf{n} \text{ large enough.}
\end{aligned}$$

We will now introduce the spatial blocks decomposition introduced by [342] which will be useful afterwards. Without loss of generality, we suppose that $n_k = 2bq_k$, for $1 \leq k \leq \mathrm{N}$. The random variables $\Lambda_{\mathbf{i}}(x)$ can be grouped into $2^{\mathrm{N}} q_1 \dots q_{\mathrm{N}}$ cubic blocks of side b. Let,

$$\begin{aligned}
\mathrm{U}(1,\mathbf{n},x,\mathbf{j}) &= \sum_{\substack{i_k=2j_kb+1,\\ k=1,\dots,\mathrm{N}.}}^{(2j_k+1)b} \Lambda_{\mathbf{i}}(x),\\
\mathrm{U}(2,\mathbf{n},x,\mathbf{j}) &= \sum_{\substack{i_k=2j_kb+1,\\ k=1,\dots,\mathrm{N}-1.}}^{(2j_k+1)b} \sum_{i_{\mathrm{N}}=(2j_{\mathrm{N}}+1)b+1}^{2(j_{\mathrm{N}}+1)b} \Lambda_{\mathbf{i}}(x),\\
\mathrm{U}(3,\mathbf{n},x,\mathbf{j}) &= \sum_{\substack{i_k=2j_kb+1,\\ k=1,\dots,\mathrm{N}-2.}}^{(2j_k+1)b} \sum_{i_{\mathrm{N}-1}=(2j_{\mathrm{N}-1}+1)b+1}^{2(j_{\mathrm{N}-1}+1)b} \sum_{i_{\mathrm{N}}=2j_{\mathrm{N}}b+1}^{(2j_{\mathrm{N}}+1)b} \Lambda_{\mathbf{i}}(x),\\
\mathrm{U}(4,\mathbf{n},x,\mathbf{j}) &= \sum_{\substack{i_k=2j_kb+1,\\ k=1,\dots,\mathrm{N}-2.}}^{(2j_k+1)b} \sum_{i_{\mathrm{N}-1}=(2j_{\mathrm{N}-1}+1)b+1}^{2(j_{\mathrm{N}-1}+1)b} \sum_{i_{\mathrm{N}}=(2j_{\mathrm{N}}+1)b+1}^{(2j_{\mathrm{N}}+1)b} \Lambda_{\mathbf{i}}(x)
\end{aligned}$$

and so on. Noting that

$$\begin{aligned}
\mathrm{U}(2^{\mathrm{N}-1},\mathbf{n},x,\mathbf{j}) &= \sum_{\substack{i_k=(2j_k+1)b+1,\\ k=1,\dots,\mathrm{N}-1.}}^{2(j_k+1)b} \sum_{i_{\mathrm{N}}=2j_{\mathrm{N}}b+1}^{(2j_{\mathrm{N}}+1)b} \Lambda_{\mathbf{i}}(x)\\
\mathrm{U}(2^{\mathrm{N}},\mathbf{n},x,\mathbf{j}) &= \sum_{\substack{i_k=(2j_k+1)b+1,\\ k=1,\dots,\mathrm{N}.}}^{2(j_k+1)b} \Lambda_{\mathbf{i}}(x)
\end{aligned}$$

for each integer $1 \leq l \leq 2^{\mathrm{N}}$, we define $\mathrm{T}(\mathbf{n},x,l) = \sum_{\substack{j_k=0\\ k=1,\dots,\mathrm{N}.}}^{q_k-1} \mathrm{U}(l,\mathbf{n},x,\mathbf{j})$. We obtain $\mathrm{S}_{\mathbf{n}}(x) = \sum_{l=1}^{2^{\mathrm{N}}} \mathrm{T}(\mathbf{n},x,l)$. For $\epsilon > 0$, $\mathrm{P} \leq \mathbb{P}\left(\left|\sum_{l=1}^{2^{\mathrm{N}}} \mathrm{T}(\mathbf{n},x,l)\right| > \epsilon\right) \leq 2^{\mathrm{N}}\mathbb{P}\left(|\mathrm{T}(\mathbf{n},x,1)| > \frac{\epsilon}{2^{\mathrm{N}}}\right)$. We enumerate in arbitrary manner the $\widehat{q} = q_1 \times \dots \times q_{\mathrm{N}}$ terms $\mathrm{U}(1,\mathbf{n},x,\mathbf{j})$ of the sum $\mathrm{T}(\mathbf{n},x,1)$, and refer to them as $\mathrm{W}_1,\dots,\mathrm{W}_{\widehat{q}}$. Note that $\mathrm{U}(1,\mathbf{n},x,\mathbf{j})$ is a measurable σ-algebra generated by $\mathrm{X}_{\mathbf{i}}$, with $\mathbf{i}$ such that $2j_kb+1 \leq i_k \leq (2j_k+1)b$, $k = 1,\dots,\mathrm{N}$. For all $l = 1,\dots,\widehat{q}$, the sets of the sites in W_l are separated by a distance of at least equal to b. In addition, since K_1 and K_2 write $|\mathrm{W}_l| \leq \mathrm{C}\frac{b^{\mathrm{N}}}{a_{\mathbf{n}}}$ with $\mathrm{C} = \|\mathrm{K}_1\|_\infty \|\mathrm{K}_2\|_\infty$ (where $\|\cdot\|_\infty$ is the sup norm). Lemma 7 insures the existence of some random variables $\mathrm{W}_1^*, \mathrm{W}_2^*, \dots, \mathrm{W}_{\widehat{q}}^*$ such that

$$\begin{aligned}
\sum_{l=1}^{\widehat{q}} \mathbb{E}|\mathrm{W}_l - \mathrm{W}_l^*| &\leq 2\widehat{q}\mathrm{C}\frac{b^{\mathrm{N}}}{a_{\mathbf{n}}}\psi((\widehat{q}-1)b^{\mathrm{N}}, b^{\mathrm{N}})\chi(b)\\
&\leq 2\mathrm{C}\frac{\widehat{\mathbf{n}}}{2^{\mathrm{N}}b^{\mathrm{N}}}\frac{b^{\mathrm{N}}}{a_{\mathbf{n}}}\psi(\widehat{\mathbf{n}}, b^{\mathrm{N}})\chi(b).
\end{aligned}$$

Markov inequality allows us to write

$$\mathbb{P}\left(\sum_{l=1}^{\widehat{q}} |\mathrm{W}_l - \mathrm{W}_l^*| > \frac{\epsilon}{2^{\mathrm{N}+1}}\right) \leq 2\mathrm{C}\frac{\widehat{\mathbf{n}}}{2^{\mathrm{N}}b^{\mathrm{N}}}\frac{b^{\mathrm{N}}}{a_{\mathbf{n}}}\psi(\widehat{\mathbf{n}}, b^{\mathrm{N}})\chi(b)2^{\mathrm{N}+1}\epsilon^{-1},$$

and by Bernstein inequality, we have

$$\mathbb{P}\left(\sum_{l=1}^{\widehat{q}} |\mathrm{W}_l^*| > \frac{\epsilon}{2^{\mathrm{N}+1}}\right) \leq 2\exp\left\{\frac{-\epsilon^2/(2^{\mathrm{N}+1})^2}{4\sum_{l=1}^{\widehat{q}}\mathbb{E}(\mathrm{W}_l^{*2}) + \frac{2\epsilon}{2^{\mathrm{N}+1}}\frac{b^{\mathrm{N}}}{a_{\mathbf{n}}}\mathrm{C}}\right\}$$

which leads to

$$\begin{aligned}
\mathbb{P}\left[|\mathrm{S}_{\mathbf{n}}(x)| \geq \epsilon\right] &\leq 2^{\mathrm{N}+1}\exp\left\{\frac{-\epsilon^2/(2^{\mathrm{N}+1})^2}{4\sum_{l=1}^{\widehat{q}}\mathbb{E}(\mathrm{W}_l^{*2}) + 2^{-\mathrm{N}}\mathrm{C}\epsilon\frac{b^{\mathrm{N}}}{a_{\mathbf{n}}}}\right\}\\
&\quad + 2^{\mathrm{N}+1}\mathrm{C}\frac{\widehat{\mathbf{n}}}{2^{\mathrm{N}}b^{\mathrm{N}}}\frac{b^{\mathrm{N}}}{a_{\mathbf{n}}}\psi(\widehat{\mathbf{n}}, b^{\mathrm{N}})\chi(b)2^{\mathrm{N}+1}\epsilon^{-1}.
\end{aligned}$$

Let $\delta > 0$, $\epsilon = \epsilon_{\mathbf{n}} = \delta\left(\frac{\log\widehat{\mathbf{n}}}{\widehat{\mathbf{n}}\rho_{\mathbf{n}}^{\mathrm{N}}\varphi_x(b_{\mathbf{n}})}\right)^{1/2}$ and $b = \left(\frac{\widehat{\mathbf{n}}\rho_{\mathbf{n}}^{\mathrm{N}}\varphi_x(b_{\mathbf{n}})}{\log\widehat{\mathbf{n}}}\right)^{\frac{1}{2\mathrm{N}}}$. Since the variables W_l and W_l^* have the same distributions, we have $\sum_{l=1}^{\widehat{q}}\mathbb{E}\mathrm{W}_l^{*2} = \sum_{l=1}^{\widehat{q}}\mathrm{var}(\mathrm{W}_l^*) = \sum_{l=1}^{\widehat{q}}\mathrm{var}(\mathrm{W}_l) \leq \mathrm{I}_{\mathbf{n}}(x) + \mathrm{R}_{\mathbf{n}}(x)$, and according to Lemma 12, we have $\sum_{l=1}^{\widehat{q}}\mathbb{E}\mathrm{W}_l^{*2} \leq \mathrm{O}\left([\widehat{\mathbf{n}}\rho_{\mathbf{n}}^{\mathrm{N}}\varphi_x(b_{\mathbf{n}})]^{-1}\right)$. Then,

$$\begin{aligned}\mathbb{P}\left[|\mathrm{S}_{\mathbf{n}}(x)| \geq \epsilon\right] &\leq 2^{\mathrm{N}+1}\exp\left\{\frac{-\epsilon^2}{2^{2\mathrm{N}+2}\left(4\frac{\mathrm{C}}{\widehat{\mathbf{n}}\rho_{\mathbf{n}}^{\mathrm{N}}\varphi_x(b_{\mathbf{n}})} + \mathrm{C}2^{-\mathrm{N}}\epsilon\frac{b^{\mathrm{N}}}{a_{\mathbf{n}}}\right)}\right\} \\ &\quad + 2^{\mathrm{N}+2}\mathrm{C}\frac{\widehat{\mathbf{n}}}{a_{\mathbf{n}}}\psi(\widehat{\mathbf{n}}, b^{\mathrm{N}})b^{-\theta}\epsilon^{-1}.\end{aligned}$$

Since $\mathrm{C}_1'' k_{\mathbf{n}}\varphi_x(b_{\mathbf{n}}) \leq a_{\mathbf{n}} \leq \mathrm{C}_2'' k_{\mathbf{n}}\varphi_x(b_{\mathbf{n}})$, where C_1''andC_2'' are positive constant and $k_{\mathbf{n}} = \mathrm{O}(\widehat{\mathbf{n}}\rho_{\mathbf{n}}^{\mathrm{N}})$, we have

$$\begin{aligned}\mathbb{P}\left[|\mathrm{S}_{\mathbf{n}}(x)| \geq \epsilon_{\mathbf{n}}\right] &\leq 2^{\mathrm{N}+1}\exp\left\{\frac{-\delta^2\frac{\log\widehat{\mathbf{n}}}{\widehat{\mathbf{n}}\rho_{\mathbf{n}}^{\mathrm{N}}\varphi_x(b_{\mathbf{n}})}}{\frac{2^{2\mathrm{N}+4}\mathrm{C}}{\widehat{\mathbf{n}}\rho_{\mathbf{n}}^{\mathrm{N}}\varphi_x(b_{\mathbf{n}})} + \frac{\mathrm{C}2^{\mathrm{N}+2}\delta}{\widehat{\mathbf{n}}\rho_{\mathbf{n}}^{\mathrm{N}}\varphi_x(b_{\mathbf{n}})}}\right\} \\ &\quad + 2^{\mathrm{N}+2}\mathrm{C}\frac{\widehat{\mathbf{n}}}{a_{\mathbf{n}}}\psi(\widehat{\mathbf{n}}, b^{\mathrm{N}})b^{-\theta}\delta^{-1}\left(\frac{\widehat{\mathbf{n}}\rho_{\mathbf{n}}^{\mathrm{N}}\varphi_x(b_{\mathbf{n}})}{\log\widehat{\mathbf{n}}}\right)^{1/2} \\ &\leq \mathrm{C}2^{\mathrm{N}+1}\exp\left\{\log\widehat{\mathbf{n}}^{-a}\right\} \\ &\quad + 2^{\mathrm{N}+2}\mathrm{C}\delta^{-1}\frac{\widehat{\mathbf{n}}}{a_{\mathbf{n}}}\psi(\widehat{\mathbf{n}}, b^{\mathrm{N}})\left(\frac{\widehat{\mathbf{n}}\rho_{\mathbf{n}}^{\mathrm{N}}\varphi_x(b_{\mathbf{n}})}{\log\widehat{\mathbf{n}}}\right)^{\frac{\mathrm{N}-\theta}{2\mathrm{N}}} \\ &\leq \mathrm{C}\widehat{\mathbf{n}}^{-a} + 2^{\mathrm{N}+2}\mathrm{C}\delta^{-1}\frac{\widehat{\mathbf{n}}}{a_{\mathbf{n}}}\psi(\widehat{\mathbf{n}}, b^{\mathrm{N}})\left(\frac{\widehat{\mathbf{n}}\rho_{\mathbf{n}}^{\mathrm{N}}\varphi_x(b_{\mathbf{n}})}{\log\widehat{\mathbf{n}}}\right)^{\frac{\mathrm{N}-\theta}{2\mathrm{N}}} \\ &:= \mathrm{C}\widehat{\mathbf{n}}^{-a} + \mathrm{C}2^{\mathrm{N}+2}\delta^{-1}\mathrm{D}_{\mathbf{n}},\end{aligned}$$

with $a = \dfrac{\delta^2}{2^{2\mathrm{N}+4}\mathrm{C} + \mathrm{C}2^{\mathrm{N}+2}\delta} > 0$. Note that $\widehat{\mathbf{n}}^{1-a}\widehat{\mathbf{n}}\rho_{\mathbf{n}}^{\mathrm{N}}\varphi_x(b_{\mathbf{n}})$ tends to 0 for $a > 1$ and then $\mathrm{C}\widehat{\mathbf{n}}^{-a} = o\left([\widehat{\mathbf{n}}\rho_{\mathbf{n}}^{\mathrm{N}}\varphi_x(b_{\mathbf{n}})]^{-1}\right)$. Moreover $a > 1$ if and only if $\delta > 2^{\mathrm{N}+1}\mathrm{C}(1+\sqrt{4\mathrm{C}}) > 2^{\mathrm{N}+1}\mathrm{C}$ (with $\delta > 0$). Now, we treat the second term.
When (3.2) is satisfied, *i.e.* $\psi(n,m) \leq \mathrm{C}\min(n,m)$, $\forall n, m \in \mathbb{N}$, we have

$$\begin{aligned}\widehat{\mathbf{n}}\rho_{\mathbf{n}}^{\mathrm{N}}\varphi_x(b_{\mathbf{n}})2^{\mathrm{N}+2}\mathrm{C}\delta^{-1}\mathrm{D}_{\mathbf{n}} &\leq \widehat{\mathbf{n}}\rho_{\mathbf{n}}^{\mathrm{N}}2^{\mathrm{N}+2}\mathrm{C}\delta^{-1}\frac{\widehat{\mathbf{n}}}{a_{\mathbf{n}}}\left(\frac{\widehat{\mathbf{n}}\rho_{\mathbf{n}}^{\mathrm{N}}\varphi_x(b_{\mathbf{n}})}{\log\widehat{\mathbf{n}}}\right)^{\frac{2\mathrm{N}-\theta}{2\mathrm{N}}} \\ &\leq \widehat{\mathbf{n}}\rho_{\mathbf{n}}^{\mathrm{N}}2^{\mathrm{N}+2}\mathrm{C}\delta^{-1}\frac{1}{\rho_{\mathbf{n}}^{\mathrm{N}}}\left(\frac{\widehat{\mathbf{n}}\rho_{\mathbf{n}}^{\mathrm{N}}\varphi_x(b_{\mathbf{n}})}{\log\widehat{\mathbf{n}}}\right)^{\frac{2\mathrm{N}-\theta}{2\mathrm{N}}} \\ &\leq \mathrm{C}\left[\widehat{\mathbf{n}}\left(\rho_{\mathbf{n}}^{\mathrm{N}}\varphi_x(b_{\mathbf{n}})\right)^{\frac{2\mathrm{N}-\theta}{4\mathrm{N}-\theta}}\left(\log\widehat{\mathbf{n}}\right)^{\frac{\theta-2\mathrm{N}}{4\mathrm{N}-\theta}}\right]^{\frac{4\mathrm{N}-\theta}{2\mathrm{N}}}\end{aligned}$$

which tends to 0 as $\mathbf{n} \to 0$ since $\theta > 4\mathrm{N}$.
When (3.3) is satisfied, *i.e.* $\psi(n,m) \leq \mathrm{C}(n+m+1)^{\kappa}$, $\forall n, m \in \mathbb{N}$, and note that $\psi(\widehat{\mathbf{n}}, b^{\mathrm{N}}) \leq \mathrm{C}(\widehat{\mathbf{n}} + b^{\mathrm{N}} + 1)^{\kappa} \leq \mathrm{C}\widehat{\mathbf{n}}^{\kappa}$, we have

$$\begin{aligned}\widehat{\mathbf{n}}\rho_{\mathbf{n}}^{\mathrm{N}}\varphi_x(b_{\mathbf{n}})\mathrm{C}2^{\mathrm{N}+2}\delta^{-1}\frac{\widehat{\mathbf{n}}}{a_{\mathbf{n}}}\widehat{\mathbf{n}}^{\kappa}\left(\frac{\widehat{\mathbf{n}}\rho_{\mathbf{n}}^{\mathrm{N}}\varphi_x(b_{\mathbf{n}})}{\log\widehat{\mathbf{n}}}\right)^{\frac{2\mathrm{N}-\theta}{2\mathrm{N}}} &\leq \widehat{\mathbf{n}}\rho_{\mathbf{n}}^{\mathrm{N}}2^{\mathrm{N}+2}\mathrm{C}\delta^{-1}\frac{1}{\rho_{\mathbf{n}}^{\mathrm{N}}}\widehat{\mathbf{n}}^{\kappa}\left(\frac{\widehat{\mathbf{n}}\rho_{\mathbf{n}}^{\mathrm{N}}\varphi_x(b_{\mathbf{n}})}{\log\widehat{\mathbf{n}}}\right)^{\frac{2\mathrm{N}-\theta}{2\mathrm{N}}} \\ &\leq \mathrm{C}\left[\widehat{\mathbf{n}}\left(\rho_{\mathbf{n}}^{\mathrm{N}}\varphi_x(b_{\mathbf{n}})\right)^{\frac{\mathrm{N}-\theta}{\mathrm{N}(3+2\kappa)-\theta}}\left(\log\widehat{\mathbf{n}}\right)^{\frac{\theta-\mathrm{N}}{\mathrm{N}(3+2\kappa)-\theta}}\right]^{\frac{\mathrm{N}(3+2\kappa)-\theta}{2\mathrm{N}}}\end{aligned}$$

which tends to 0 as $\mathbf{n} \to$ since $\theta > \mathrm{N}(3+2\kappa)$. Therefore, (0.4) follows, which conclude the proof of Lemma 10.
□

Proof of Lemma 11

Since $Y_{\mathbf{i}}$ and r are bounded, we have

$$\begin{aligned}
\mathbb{E}^{1/2}[\mathbf{C}] &\leq \mathbb{E}^{1/2}\left[\left|\frac{1}{\widehat{\mathbf{n}}}\sum_{\mathbf{i}\in\mathcal{V}_{\mathbf{i}_0}} Y_{\mathbf{i}} - r(x)\right| \mathbf{1}_{\left\{\sum_{\mathbf{i}\in\mathcal{V}_{\mathbf{i}_0}} W_{\mathbf{ni}}(x)=0\right\}}\right] \\
&\leq 2M\mathbb{E}^{1/2}\left[\mathbf{1}_{\left\{\sum_{\mathbf{i}\in\mathcal{V}_{\mathbf{i}_0}} W_{\mathbf{ni}}(x)=0\right\}}\right] = 2M\left(\mathbb{P}\left[\sum_{\mathbf{i}\in\mathcal{V}_{\mathbf{i}_0}} K_{\mathbf{i}}(x)=0\right]\right)^{1/2} \\
&\leq 2M\left(\mathbb{P}\left[\sum_{\mathbf{i}\in\mathcal{V}_{\mathbf{i}_0}} K_{\mathbf{i}}(x)\leq \frac{a_{\mathbf{n}}}{2}\right]\right)^{1/2} = O\left(\frac{1}{\widehat{\mathbf{n}}\rho_{\mathbf{n}}^{N}\varphi_x(b_{\mathbf{n}})}\right)^{1/2},
\end{aligned}$$

by Lemma 10. □

Proof of Lemma 12

Firstly, we deal with $I_{\mathbf{n}}(x) = \sum_{\mathbf{i}\in\mathcal{V}_{\mathbf{i}_0}} \mathbb{E}\left[\left(\frac{1}{a_{\mathbf{n}}}K_{\mathbf{i}}(x)\right)^2\right] - \sum_{\mathbf{i}\in\mathcal{V}_{\mathbf{i}_0}}\left(\frac{1}{a_{\mathbf{n}}}\mathbb{E}(K_{\mathbf{i}}(x))\right)^2$.

$$\begin{aligned}
\sum_{\mathbf{i}\in\mathcal{V}_{\mathbf{i}_0}} \mathbb{E}\left[\left(\frac{1}{a_{\mathbf{n}}}K_{\mathbf{i}}(x)\right)^2\right] &\leq C\frac{1}{a_{\mathbf{n}}^2}\sum_{\mathbf{i}\in\mathcal{V}_{\mathbf{i}_0}} K_{2\mathbf{i}}^2\mathbb{E}\left[K_{1\mathbf{i}}^2(x)\right] \\
&\leq C\frac{1}{a_{\mathbf{n}}^2}\sum_{\mathbf{i}\in\mathcal{V}_{\mathbf{i}_0}} k_{\mathbf{n}}\varphi_x(b_{\mathbf{n}}) \\
&\leq \frac{C}{k_{\mathbf{n}}\varphi_x(b_{\mathbf{n}})} = O\left([\widehat{\mathbf{n}}\rho_{\mathbf{n}}^{N}\varphi_x(b_{\mathbf{n}})]^{-1}\right),
\end{aligned}$$

for $\mathbf{n}$ sufficiently large.
Then, we have $I_{\mathbf{n}}(x) = O\left([\widehat{\mathbf{n}}\rho_{\mathbf{n}}^{N}\varphi_x(b_{\mathbf{n}})]^{-1}\right)$. We now treat the term $R_{\mathbf{n}}(x)$. Since the functions $K_1(.)$ and $K_2(.)$ are bounded, applying Lemma 7, we get

$$|\mathbb{E}[\Lambda_{\mathbf{i}}(x)\Lambda_{\mathbf{k}}(x)]| \leq C\frac{K_{2\mathbf{i}}K_{2\mathbf{k}}}{a_{\mathbf{n}}^2}\psi(1,1)\gamma(\|\mathbf{i}-\mathbf{k}\|).$$

Let $E_{\mathbf{n}}$ be a sequence of real numbers tending to ∞ as $\widehat{\mathbf{n}}\to\infty$. Set $T = \{\mathbf{i},\mathbf{k}\in\mathcal{V}_{\mathbf{i}_0}, \|\mathbf{i}-\mathbf{k}\|\leq E_{\mathbf{n}}\}$ and denote by T^c the complementary of T. Let $R_{\mathbf{n}}^{(1)} = \sum_{\mathbf{i},\mathbf{k}\in T}|\mathbb{E}[\Lambda_{\mathbf{i}}(x)\Lambda_{\mathbf{k}}(x)]|$ and $R_{\mathbf{n}}^{(2)} = \sum_{\mathbf{i},\mathbf{k}\in T^c}|\mathbb{E}[\Lambda_{\mathbf{i}}(x)\Lambda_{\mathbf{k}}(x)]|$. Hence, $R_{\mathbf{n}}(x)\leq R_{\mathbf{n}}^{(1)}+R_{\mathbf{n}}^{(2)}$. Moreover, using the same arguments as in the proof of Lemma 10, we have $I_{\mathbf{n}}(x)+R_{\mathbf{n}}(x) = O\left([\widehat{\mathbf{n}}\rho_{\mathbf{n}}^{N}\varphi_x(b_{\mathbf{n}})]^{-1}\right)$. □

Proof of Theorem 4

Recall that $K_{\mathbf{i}}(x) = K_{1\mathbf{i}}K_{2\mathbf{i}}$. Set $T_{\mathbf{n}} = (\hat{\mathbf{n}}u_{\mathbf{n}})^{1/s}$ where $u_{\mathbf{n}} = \prod_{i=1}^{N}(\log n_i)(\log\log n_i)^{1+\epsilon}$, and define

$$\begin{aligned}
g_{\mathbf{n}}(x) &= \frac{1}{a_{\mathbf{n}}}\sum_{\mathbf{i}\in\mathcal{V}_{\mathbf{i}_0}} Y_{\mathbf{i}}K_{\mathbf{i}}(x), \qquad f_{\mathbf{n}}(x) = \frac{1}{a_{\mathbf{n}}}\sum_{\mathbf{i}\in\mathcal{V}_{\mathbf{i}_0}} K_{\mathbf{i}}(x), \\
\widetilde{g}_{\mathbf{n}}(x) &= \frac{1}{a_{\mathbf{n}}}\sum_{\mathbf{i}\in\mathcal{V}_{\mathbf{i}_0}} Y_{\mathbf{i}}\mathbb{1}_{\{Y_{\mathbf{i}}\leq T_{\mathbf{n}}\}}K_{\mathbf{i}}(x).
\end{aligned}$$

Then, we can write

$$r_{\mathbf{n}}(x) - r(x) = -\frac{r(x)}{f_{\mathbf{n}}(x)}A_1(x) + \frac{1}{f_{\mathbf{n}}(x)}\left[A_2(x)+A_3(x)+A_4(x)\right], \tag{0.5}$$

where

$$\begin{aligned}
A_1(x) &= f_{\mathbf{n}}(x)-1, \\
A_2(x) &= \mathbb{E}\left(\widetilde{g}_{\mathbf{n}}(x)\right) - r(x), \\
A_3(x) &= \widetilde{g}_{\mathbf{n}}(x) - \mathbb{E}\left(\widetilde{g}_{\mathbf{n}}(x)\right), \\
A_4(x) &= g_{\mathbf{n}}(x) - \widetilde{g}_{\mathbf{n}}(x).
\end{aligned}$$

Therefore Theorem 4 follows from (0.5) and Lemmas 13, 14, 15, 18. □

Lemma 13. *Under assumptions **H1-H4** and* **H6**,

$$\sup_{x\in\mathcal{D}}\left|\mathbb{E}\left(\widetilde{g}_{\mathbf{n}}(x)\right) - r(x)\right| = O\left(b_{\mathbf{n}} + \sqrt{\frac{\log\widehat{\mathbf{n}}}{\widehat{\mathbf{n}}\rho_{\mathbf{n}}^{N}\Gamma(b_{\mathbf{n}})}}\right).$$

Proof of Lemma 13

Since

$$\begin{aligned}
&\mathbb{E}\left(\widetilde{g}_{\mathbf{n}}(x)\right) - r(x) \\
&= \frac{1}{a_{\mathbf{n}}\varphi_x(b_{\mathbf{n}})} \sum_{\mathbf{i}\in\mathcal{V}_{\mathbf{i}_0}} \mathbb{E}\left[\left(\mathrm{Y}_{\mathbf{i}} - \mathrm{Y}_{\mathbf{i}}\mathbb{1}_{\{|\mathrm{Y}_{\mathbf{i}}|>\mathrm{T}_{\mathbf{n}}\}}\right)\mathrm{K}_{\mathbf{i}}(x)\right] - r(x) \\
&= \frac{1}{a_{\mathbf{n}}} \sum_{\mathbf{i}\in\mathcal{V}_{\mathbf{i}_0}} \mathbb{E}\left[\mathbb{E}\left(\mathrm{Y}_{\mathbf{i}}|\mathrm{X}_{\mathbf{i}}\right)\mathrm{K}_{\mathbf{i}}(x)\right] - \frac{1}{a_{\mathbf{n}}} \sum_{\mathbf{i}\in\mathcal{V}_{\mathbf{i}_0}} \mathbb{E}\left[\mathrm{Y}_{\mathbf{i}}\mathbb{1}_{\{|\mathrm{Y}_{\mathbf{i}}|>\mathrm{T}_{\mathbf{n}}\}}\mathrm{K}_{\mathbf{i}}(x)\right] - r(x) \\
&= \frac{1}{a_{\mathbf{n}}} \sum_{\mathbf{i}\in\mathcal{V}_{\mathbf{i}_0}} \mathbb{E}\left[\left(r(\mathrm{X}_{\mathbf{i}}) - r(x)\right)\mathrm{K}_{\mathbf{i}}(x)\right] - \frac{1}{a_{\mathbf{n}}} \sum_{\mathbf{i}\in\mathcal{V}_{\mathbf{i}_0}} \mathbb{E}\left[\mathrm{Y}_{\mathbf{i}}\mathbb{1}_{\{|\mathrm{Y}_{\mathbf{i}}|>\mathrm{T}_{\mathbf{n}}\}}\mathrm{K}_{\mathbf{i}}(x)\right],
\end{aligned}$$

we have

$$\begin{aligned}
\left|\mathbb{E}\left(\widetilde{g}_{\mathbf{n}}(x)\right) - r(x)\right| &\leq \frac{1}{a_{\mathbf{n}}} \sum_{\mathbf{i}\in\mathcal{V}_{\mathbf{i}_0}} \mathbb{E}\left[\left|r(\mathrm{X}_{\mathbf{i}}) - r(x)\right|\mathrm{K}_{\mathbf{i}}(x)\right] \\
&+ \frac{1}{a_{\mathbf{n}}} \sum_{\mathbf{i}\in\mathcal{V}_{\mathbf{i}_0}} \mathbb{E}\left[|\mathrm{Y}_{\mathbf{i}}|\,\mathbb{1}_{\{|\mathrm{Y}_{\mathbf{i}}|>\mathrm{T}_{\mathbf{n}}\}}\mathrm{K}_{\mathbf{i}}(x)\right] := \mathrm{I} + \mathrm{II}.
\end{aligned}$$

Using assumptions (**H1**) and (**H2**), we have

$$|r(\mathrm{X}_{\mathbf{i}}) - r(x)| \leq \sup_{u\in\mathrm{B}(x,b_{\mathbf{n}})} |r(x) - r(u)| = \mathrm{O}(b_{\mathbf{n}}), \text{ so that } \mathrm{I} = \mathrm{O}\left(b_{\mathbf{n}}\right).$$

For II, since $s > 2$, using Assumption **H4** and **H6**, we can write

$$\begin{aligned}
\mathrm{II} &\leq \frac{\mathrm{T}_{\mathbf{n}}^{1-s}}{a_{\mathbf{n}}} \sum_{\mathbf{i}\in\mathcal{V}_{\mathbf{i}_0}} \mathbb{E}\left[|\mathrm{Y}_{\mathbf{i}}|^s\mathrm{K}_{\mathbf{i}}(x)\right] \leq \frac{\mathrm{T}_{\mathbf{n}}^{1-s}}{a_{\mathbf{n}}} \sum_{\mathbf{i}\in\mathcal{V}_{\mathbf{i}_0}} \mathbb{E}\left[\mathbb{E}\left(|\mathrm{Y}_{\mathbf{i}}|^s|\mathrm{X}_{\mathbf{i}}\right)\mathrm{K}_{\mathbf{i}}(x)\right] \\
&\leq \mathrm{C}\mathrm{T}_{\mathbf{n}}^{1-s} = o\left((\widehat{\mathbf{n}}u_{\mathbf{n}})^{-1/2}\right) = o\left(\sqrt{\frac{\log\widehat{\mathbf{n}}}{\widehat{\mathbf{n}}\rho_{\mathbf{n}}^{\mathrm{N}}\Gamma(b_{\mathbf{n}})}}\right),
\end{aligned}$$

which conclude the proof of Lemma 13. □

Lemma 14. *If Assumption **(H6)** (i) holds, then*

$$\sup_{x\in\mathscr{D}} \left|g_{\mathbf{n}}(x) - \widetilde{g}_{\mathbf{n}}(x)\right| = 0$$

for sufficiently large $\mathbf{n}$.

Proof of Lemma 14

Recall that $\mathrm{T}_{\mathbf{n}} = (\widehat{\mathbf{n}}u_{\mathbf{n}})^{1/s}$ and note that

$$g_{\mathbf{n}}(x) - \widetilde{g}_{\mathbf{n}}(x) = \frac{1}{a_{\mathbf{n}}} \sum_{\mathbf{i}\in\mathcal{V}_{\mathbf{i}_0}} \mathrm{Y}_{\mathbf{i}}\mathbb{1}_{\{|\mathrm{Y}_{\mathbf{i}}|>\mathrm{T}_{\mathbf{n}}\}}\mathrm{K}_{\mathbf{i}}(x).$$

By the Markov inequality, $\mathbb{P}\left(|\mathrm{Y}_{\mathbf{i}}| > \mathrm{T}_{\mathbf{n}}\right) \leq \mathrm{T}_{\mathbf{n}}^{-s}\mathbb{E}|\mathrm{Y}_{\mathbf{i}}|^s$ for any $\mathbf{i} \in \mathbb{Z}^{\mathrm{N}}$. Therefore

$$\sum_{\mathbf{n}\in\mathbb{Z}^{\mathrm{N}}} \mathbb{P}\left(|\mathrm{Y}_{\mathbf{n}}| > \mathrm{T}_{\mathbf{n}}\right) \leq \mathrm{C} \sum_{\mathbf{n}\in\mathbb{Z}^{\mathrm{N}}} \frac{1}{\widehat{\mathbf{n}}u_{\mathbf{n}}} < \infty.$$

The Borel-Cantelli lemma ensures that almost surely $|\mathrm{Y}_{\mathbf{i}}| \leq \mathrm{T}_{\mathbf{n}}$ for sufficiently large $\mathbf{n}$. Since $\mathrm{T}_{\mathbf{n}} \to \infty$ as $\mathbf{n} \to \infty$, we have almost surely $|\mathrm{Y}_{\mathbf{i}}| < \mathrm{T}_{\mathbf{n}}$ for all $\mathbf{i} \in \mathcal{V}_{\mathbf{i}_0}$ and for $\mathbf{n}$ sufficiently large enough, and thus the conclusion follows. □

Lemma 15. *Under the assumptions of Theorem 4,*

$$\sup_{x\in\mathscr{D}} \left|\widetilde{g}_{\mathbf{n}}(x) - \mathbb{E}\left(\widetilde{g}_{\mathbf{n}}(x)\right)\right| = \mathrm{O}\left(\left(\frac{\log\widehat{\mathbf{n}}}{\widehat{\mathbf{n}}\rho_{\mathbf{n}}^{\mathrm{N}}\Gamma(b_{\mathbf{n}})}\right)^{1/2}\right) a.s$$

Define

$$\widetilde{\Lambda}_{\mathbf{i}}(x) \quad = \quad \mathrm{Y}_{\mathbf{i}}\mathbb{1}_{\{|\mathrm{Y}_{\mathbf{i}}|\leq \mathrm{T}_{\mathbf{n}}\}}\mathrm{K}_{\mathbf{i}}(x) - \mathbb{E}\left(\mathrm{Y}_{\mathbf{i}}\mathbb{1}_{\{|\mathrm{Y}_{\mathbf{i}}|\leq \mathrm{T}_{\mathbf{n}}\}}\mathrm{K}_{\mathbf{i}}(x)\right),$$

$$\widetilde{\mathrm{I}}_{\mathbf{n}}(x) = \frac{1}{a_{\mathbf{n}}^2}\sum_{\mathbf{i}\in\mathcal{V}_{\mathbf{i}_0}} \mathbb{E}\left(\widetilde{\Lambda}_{\mathbf{i}}(x)^2\right) \text{ and } \widetilde{\mathrm{R}}_{\mathbf{n}}(x) = \frac{1}{a_{\mathbf{n}}^2}\sum_{\mathbf{i}\neq\mathbf{j}} \left|\mathbb{E}\left[\widetilde{\Lambda}_{\mathbf{i}}(x)\widetilde{\Lambda}_{\mathbf{j}}(x)\right]\right|. \tag{0.6}$$

Then, arguing as in the proof of Lemma 12 with $\varphi_x(b_{\mathbf{n}})$ replacing by $\Gamma(b_{\mathbf{n}})$, one can prove under assumptions (**H1**)-(**H2**), (**H4**)–(**H6**) that,

$$\widetilde{\mathrm{I}}_{\mathbf{n}}(x) + \widetilde{\mathrm{R}}_{\mathbf{n}}(x) = \mathrm{O}\left(\frac{1}{\widehat{\mathbf{n}}\rho_{\mathbf{n}}^{\mathrm{N}}\Gamma(b_{\mathbf{n}})}\right) \quad \text{for any } x\in\mathscr{D}. \tag{0.7}$$

Let us define

$$\Omega_{\mathbf{n}} = \sqrt{\frac{\log\widehat{\mathbf{n}}}{\widehat{\mathbf{n}}\rho_{\mathbf{n}}^{\mathrm{N}}\Gamma(b_{\mathbf{n}})}} \text{ and choose } \ell_{\mathbf{n}} \leq \mathrm{C}\Omega_{\mathbf{n}}\varphi_x(b_{\mathbf{n}})\rho_{\mathbf{n}}^{\mathrm{N}}\Gamma(b_{\mathbf{n}})\mathrm{T}_{\mathbf{n}}^{-1} \text{ for some constant } \mathrm{C}>0.$$

We suppose that the compact set $\mathscr{D}$ is covered with $v_{\mathbf{n}}$ cubes B_k having sides of length $\ell_{\mathbf{n}}$ and centered at x_k. We have

$$\sup_{x\in\mathscr{D}}\left|\widetilde{g}_{\mathbf{n}}(x) - \mathbb{E}\left(\widetilde{g}_{\mathbf{n}}(x)\right)\right| \leq \mathrm{Q}_{1\mathbf{n}} + \mathrm{Q}_{2\mathbf{n}} + \mathrm{Q}_{3\mathbf{n}}, \tag{0.8}$$

where

$$\begin{aligned}
\mathrm{Q}_{1\mathbf{n}} &= \max_{1\leq k\leq v_{\mathbf{n}}}\sup_{x\in\mathrm{B}_k}\left|\widetilde{g}_{\mathbf{n}}(x) - \widetilde{g}_{\mathbf{n}}(x_k)\right|,\\
\mathrm{Q}_{2\mathbf{n}} &= \max_{1\leq k\leq v_{\mathbf{n}}}\sup_{x\in\mathrm{B}_k}\left|\mathbb{E}\left(\widetilde{g}_{\mathbf{n}}(x_k)\right) - \mathbb{E}\left(\widetilde{g}_{\mathbf{n}}(x)\right)\right|,\\
\mathrm{Q}_{3\mathbf{n}} &= \max_{1\leq k\leq v_{\mathbf{n}}}\sup_{x\in\mathrm{B}_k}\left|\widetilde{g}_{\mathbf{n}}(x_k) - \mathbb{E}\left(\widetilde{g}_{\mathbf{n}}(x_k)\right)\right|.
\end{aligned}$$

Lemma 16. *Under Assumptions (***H1***), (***H2***) and (***H4***),* $\mathrm{Q}_{1\mathbf{n}} = \mathrm{O}(\Omega_{\mathbf{n}})$ *and* $\mathrm{Q}_{2\mathbf{n}} = \mathrm{O}(\Omega_{\mathbf{n}})$ *a.s.*

Proof of Lemma 16

By Assumptions (**H1**), (**H2**) and (**H4**), for all $x\in\mathrm{B}_k$,

$$\left|\widetilde{g}_{\mathbf{n}}(x) - \widetilde{g}_{\mathbf{n}}(x_k)\right| \leq a_{\mathbf{n}}^{-1}\varphi_x(b_{\mathbf{n}})^{-1}\rho_{\mathbf{n}}^{-\mathrm{N}}\Gamma(b_{\mathbf{n}})^{-1}\mathrm{T}_{\mathbf{n}}\|x-x_k\| \leq \mathrm{C}\varphi_x(b_{\mathbf{n}})^{-1}\rho_{\mathbf{n}}^{-\mathrm{N}}\Gamma(b_{\mathbf{n}})^{-1}\mathrm{T}_{\mathbf{n}}\ell_{\mathbf{n}} = \mathrm{O}(\Omega_{\mathbf{n}}) \ a.s.$$

and Lemma 16 follows. □

Next, we have to show that

$$\mathrm{Q}_{3\mathbf{n}} = \mathrm{O}(\Omega_{\mathbf{n}}) \text{ a.s.} \tag{0.9}$$

Define

$$\widetilde{\mathrm{S}}_{\mathbf{n}}(x) = a_{\mathbf{n}}^{-2}\varphi_x(b_{\mathbf{n}})^{-2}\sum_{\mathbf{i}\in\mathcal{V}_{\mathbf{i}_0}}\widetilde{\Lambda}_{\mathbf{i}}(x) = \widetilde{g}_{\mathbf{n}}(x) - \mathbb{E}\left(\widetilde{g}_{\mathbf{n}}(x)\right).$$

Define also $\widetilde{\mathrm{U}}(i,\mathbf{n},x,\mathbf{j})$ and $\widetilde{\mathrm{T}}(\mathbf{n},x,i)$ to be the same as $\mathrm{U}(i,\mathbf{n},\mathbf{j},x)$ and $\mathrm{T}(\mathbf{n},i,x)$ in the proof of Lemma 10 except with $\Lambda_{\mathbf{j}}$ replacing by $\widetilde{\Lambda}_{\mathbf{j}}$. Arguing that $\widetilde{\mathrm{S}}_{\mathbf{n}}$ is a finite sum of the $\widetilde{\mathrm{T}}(\mathbf{n},x,i)$, then showing (0.9) is equivalent to show that

$$\max_{1\leq k\leq v_{\mathbf{n}}}\left|\widetilde{\mathrm{T}}(\mathbf{n},x_k,1)\right| = \mathrm{O}(\Omega_{\mathbf{n}}) \text{ a.s.} \tag{0.10}$$

By same arguments as in Lemma 10, $\widetilde{\mathrm{T}}(\mathbf{n},1,x)$ is the sum of $\widehat{q} = q_1\times\cdots\times q_{\mathrm{N}}$ of the $\tilde{\mathrm{U}}(i,\mathbf{n},\mathbf{j},x)$'s which are measurable with σ-field generated by $\mathrm{X}_{\mathbf{i}}$, where $\mathbf{i}$ belong to the set of sites which are separated by a distance at least p. Enumerate these random variables as $\mathrm{Z}_1,\ldots,\mathrm{Z}_{\widehat{q}}$ and approximate them by the independent random variables $\mathrm{Z}_1^*,\ldots,\mathrm{Z}_{\widehat{q}}^*$ as was done in Lemma 7. Define

$$p\sim\Omega_{\mathbf{n}}^{-1/\mathrm{N}}\mathrm{T}_{\mathbf{n}}^{-1/\mathrm{N}},$$

and

$$\widetilde{\beta}_{\mathbf{n}} = \mathrm{T}_{\mathbf{n}}\rho_{\mathbf{n}}^{-\mathrm{N}}\Gamma(b_{\mathbf{n}})^{-1}\psi(\widehat{\mathbf{n}},p^{\mathrm{N}})p^{-\theta}\Omega_{\mathbf{n}}^{-1}.$$

Lemma 17. *Under assumptions of Theorem 4, there exist two positive constants* A *and* C *such that, for any* $\lambda>0$,

$$\mathbb{P}\left(\max_{1\leq k\leq v_{\mathbf{n}}}\left|\widetilde{\mathrm{T}}(\mathbf{n},x_k,i)\right| > \lambda\Omega_{\mathbf{n}}\right) \leq \mathrm{C}\widehat{\mathbf{n}}^{\beta}\left[\widehat{\mathbf{n}}^{-\mathrm{A}} + \widetilde{\beta}_{\mathbf{n}}\right].$$

Proof of Lemma 17

Since $\widetilde{\mathrm{T}}(\mathbf{n},x,i)=\sum_{i=1}^{\widehat{q}}\mathrm{Z}_i$, we have, for any $\lambda>0$,

$$\mathbb{P}\left(\left|\widetilde{\mathrm{T}}(\mathbf{n},x,i)\right|>\lambda\Omega_{\mathbf{n}}\right)\leq\mathbb{P}\left(\sum_{i=1}^{\widehat{q}}\left|\mathrm{Z}_i-\mathrm{Z}_i^*\right|>\lambda\Omega_{\mathbf{n}}/2\right)+\mathbb{P}\left(\left|\sum_{i=1}^{\widehat{q}}\mathrm{Z}_i^*\right|>\lambda\Omega_{\mathbf{n}}/2\right).$$

By the boundedness of the functions K_1 and K_2 respectively, we have

$$|\mathrm{Z}_{\mathbf{i}}|\leq Cp^{\mathrm{N}}\mathrm{T}_{\mathbf{n}}a_{\mathbf{n}}^{-1}\varphi_x(b_{\mathbf{n}})^{-1}\leq C\mathrm{T}_{\mathbf{n}}p^{\mathrm{N}}\left(\widehat{\mathbf{n}}\rho_{\mathbf{n}}^{\mathrm{N}}\Gamma(b_{\mathbf{n}})\right)^{-1}.$$

Note that $\widehat{\mathbf{n}}=2^{\mathrm{N}}p^{\mathrm{N}}\widehat{q}$. Therefore Markov inequality gives : for any $\lambda>0$,

$$\mathbb{P}\left(\sum_{i=1}^{\widehat{q}}\left|\mathrm{Z}_i-\mathrm{Z}_i^*\right|>\lambda\Omega_{\mathbf{n}}\right)\leq 2\widehat{q}p^{\mathrm{N}}\mathrm{T}_{\mathbf{n}}\left(\widehat{\mathbf{n}}\rho_{\mathbf{n}}^{\mathrm{N}}\Gamma(b_{\mathbf{n}})\right)^{-1}\psi(\widehat{\mathbf{n}},p^{\mathrm{N}})\chi(p)\lambda^{-1}\Omega_{\mathbf{n}}^{-1}\leq C\widetilde{\beta}_{\mathbf{n}}.$$

By Lemma 0.7, we get, for any $\lambda>0$, there exists a constant $C>0$ such that

$$\mathbb{P}\left(\left|\sum_{i=1}^{\widehat{q}}\mathrm{Z}_i^*\right|>\lambda\Omega_{\mathbf{n}}\right)\leq C\widehat{\mathbf{n}}^{-\mathrm{A}},$$

and the conclusion follows. □

Proof of Lemma 15 Note that by the Fubini's theorem, it can be seen that $\sum_{\mathbf{n}\in\mathbb{Z}^{\mathrm{N}}}1/(\widehat{\mathbf{n}}u_{\mathbf{n}})<\infty$. By (0.8), Lemma 16 , and Lemma 17, proving Lemma 15 is equivalent to show that

$$\widehat{\mathbf{n}}u_{\mathbf{n}}\widehat{\mathbf{n}}^{\beta-\mathrm{A}}\to 0 \text{ and } \widehat{\mathbf{n}}u_{\mathbf{n}}\widehat{\mathbf{n}}^{\beta}\widetilde{\beta}_{\mathbf{n}}\to 0 \text{ as } \mathbf{n}\to\infty. \tag{0.11}$$

Note that, the first part of (0.11) holds by choosing A such that $\mathrm{A}>\beta+2$. For its second part, when (3.2) is satisfied, $\psi(\widehat{\mathbf{n}},p^{\mathrm{N}})=p^{\mathrm{N}}$ for $\mathbf{n}$ large enough. Then

$$\begin{aligned}
\widehat{\mathbf{n}}^{\beta+1}u_{\mathbf{n}}\widetilde{\beta}_{\mathbf{n}} &\leq C\widehat{\mathbf{n}}^{\beta}\left(\widehat{\mathbf{n}}u_{\mathbf{n}}\right)^{1/s+1}\rho_{\mathbf{n}}^{-\mathrm{N}}\Gamma(b_{\mathbf{n}})^{-1}\Omega_{\mathbf{n}}^{(\theta-2\mathrm{N})/\mathrm{N}}\left(\widehat{\mathbf{n}}u_{\mathbf{n}}\right)^{(\theta-\mathrm{N})/s\mathrm{N}}\\
&= C\widehat{\mathbf{n}}^{\beta+1/s+1+(\theta-\mathrm{N})/(s\mathrm{N})+(2\mathrm{N}-\theta)/(2\mathrm{N})}\rho_{\mathbf{n}}^{-\frac{\theta}{2}}\Gamma(b_{\mathbf{n}})^{\frac{-\theta}{2\mathrm{N}}}\left(\log\widehat{\mathbf{n}}\right)^{\frac{\theta-2\mathrm{N}}{2\mathrm{N}}}u_{\mathbf{n}}^{\frac{s\mathrm{N}+\theta}{s\mathrm{N}}}\\
&= C\left[\widehat{\mathbf{n}}\rho_{\mathbf{n}}^{\mathrm{N}\theta_1}\Gamma(b_{\mathbf{n}})^{\theta_1}\left(\log\widehat{\mathbf{n}}\right)^{\theta_2}u_{\mathbf{n}}^{\theta_3}\right]^{\frac{2s\mathrm{N}(\beta+2)+\theta(2-s)}{2s\mathrm{N}}},
\end{aligned}$$

which goes to zero when $\theta>\left(2\mathrm{N}s(\beta+2)\right)/(s-2)$.
Similarly, when (3.3) is satisfied, we have $\psi(\widehat{\mathbf{n}},p^{\mathrm{N}})\leq C\widehat{\mathbf{n}}^{\kappa}$ for $\mathbf{n}$ large enough. Then,

$$\begin{aligned}
\widehat{\mathbf{n}}^{\beta+1}u_{\mathbf{n}}\widetilde{\beta}_{\mathbf{n}} &\leq C\widehat{\mathbf{n}}^{\beta+\kappa}\rho_{\mathbf{n}}^{-\mathrm{N}}\Gamma(b_{\mathbf{n}})^{-1}\mathrm{T}_{\mathbf{n}}^{1+\theta/\mathrm{N}}\Omega_{\mathbf{n}}^{\frac{\theta-\mathrm{N}}{\mathrm{N}}}\\
&= C\widehat{\mathbf{n}}^{\beta+\kappa+(\mathrm{N}+\theta)/(s\mathrm{N})+(\mathrm{N}-\theta)/(2\mathrm{N})}\left(\rho_{\mathbf{n}}^{\mathrm{N}}\Gamma(b_{\mathbf{n}})\right)^{\frac{-\mathrm{N}-\theta}{2\mathrm{N}}}\left(\log\widehat{\mathbf{n}}\right)^{\frac{\theta-\mathrm{N}}{2\mathrm{N}}}u_{\mathbf{n}}^{\frac{\mathrm{N}+\theta}{s\mathrm{N}}}\\
&= C\left[\widehat{\mathbf{n}}\left(\rho_{\mathbf{n}}^{\mathrm{N}}\Gamma(b_{\mathbf{n}})\right)^{\theta_1^*}\left(\log\widehat{\mathbf{n}}\right)^{\theta_2^*}u_{\mathbf{n}}^{\theta_3^*}\right]^{\frac{\mathrm{N}(2s\beta+2s\kappa+s+2)+\theta(2-s)}{2s\mathrm{N}}},
\end{aligned}$$

which goes to zero when $\theta>\left(\mathrm{N}(2s\beta+2s\kappa+s+2)\right)/(s-2)$ and Lemma 15 follows. □

Lemma 18. *Under Assumptions (**H1**), (**H2**), (**H4**) and (**H5**),*

1. *if* (3.2) *is satisfied and*

$$\widehat{\mathbf{n}}\left(\rho_{\mathbf{n}}^{\mathrm{N}}\Gamma(b_{\mathbf{n}})\right)^{\theta_4}\left(\log\widehat{\mathbf{n}}\right)^{\theta_5}u_{\mathbf{n}}^{\theta_6}\to\infty \textit{ with } \theta>2\mathrm{N}(\beta+2),$$

2. *or if* (3.3) *is satisfied and*

$$\widehat{\mathbf{n}}\left(\rho_{\mathbf{n}}^{\mathrm{N}}\Gamma(b_{\mathbf{n}})\right)^{\theta_4^*}\left(\log\widehat{\mathbf{n}}\right)^{\theta_5^*}u_{\mathbf{n}}^{\theta_6^*}\to\infty \textit{ with } \theta>\mathrm{N}(2\beta+2\kappa+3),$$

then,

$$\sup_{x\in\mathscr{D}}\left|f_{\mathbf{n}}(x)-1\right|=\mathrm{O}\left(\left(\frac{\log\widehat{\mathbf{n}}}{\widehat{\mathbf{n}}\rho_{\mathbf{n}}^{\mathrm{N}}\Gamma(b_{\mathbf{n}})}\right)^{1/2}\right) \textit{ a.s,}$$

where

$$\begin{aligned}
\theta_4 &= \frac{\theta}{\theta-2\mathrm{N}(\beta+2)} \quad \theta_5=\frac{\theta-2\mathrm{N}}{2\mathrm{N}(\beta+2)-\theta} \quad \theta_6=\frac{2\mathrm{N}}{2\mathrm{N}(\beta+2)-\theta},\\
\theta_4^* &= \frac{-\mathrm{N}-\theta}{\mathrm{N}\left(2\beta+2\kappa+3\right)-\theta} \quad \theta_5^*=\frac{\theta-\mathrm{N}}{\mathrm{N}\left(2\beta+2\kappa+3\right)-\theta} \quad \theta_6^*=\frac{2\mathrm{N}}{\mathrm{N}\left(2\beta+2\kappa+3\right)-\theta}.
\end{aligned}$$

Proof of Lemma 18

To prove Lemma 18, just adapt the arguments considered in the proof of Lemma 15 to the case where $Y_{\mathbf{i}} \equiv 1$ and $T_{\mathbf{n}} = 1$.

Proof of Theorem 5

This result is derived directely from the proof of Theorem 4

ANNEXE C

LISTE DES ACRONYMES

AIC Critères D'information D'Akaike. 71

TCC Taux de Classification Correct. ix, 71–84
MFC Méthode Fonctionnelle Classique. 75–83
CRODT Centre de Recherches Océanographiques de Dakar-Thiaroye. 1, 27, 89, 90
CV Validation Croisée. 71

EDMI École Doctorale de Mathématiques et Informatiques. i

ADF analyse des données fonctionnelles. 39, 56, 89
FGAM Modèle additif généralisé fonctionnel. 20, 22
MANGF Modèle additif à noyau généralisé fonctionnel. 20
FGLM Modèle linéaire généralisé fonctionnel. 20, 75–83
MASGF Modèle additif spectral généralisé fonctionnel. 20, 75–81
CPF Composante principales fonctionnelles. 18, 20, 22
MCPF Moindres carrés partiels fonctionnels. 19, 20
RFL Régression Fonctionnelle Linéaire. 20

GAM modèle additif généralisé. 19, 27
GLM modèle linéaire généralisé. 19, 27

ICBS International Conference on Biomathematics in Senegal. 8
ISRA Institut Sénégalais de Recherches Agricôles. 1, 8, 9

JSDM Joint Species Distribution Modeling. 27, 40

kPPF k-plus proches voisins fonctionnelle. 23, 75–83

LEM Lille Economie Menagement. 8
LICMA Lebanese International Conference on Mathematics and Applications. 8
LMA Laboratoire de Mathématiques Appliquées. i

N/O Navire Océanographique. 1, 57

PREFACE Project Enhancing Prediction of Tropical Atlantic Climate and its Impacts. 9

SDM Species Distribution Modeling. 27, 40, 55
RIT Régression inverse par tranches. 19
MFSD Méthode Fonctionnelle et Spatiale de Discrimination. 75–84, 86
SVM Support Vector Machine. 22, 71–74

UCAD UNIVERSITÉ CHEIKH ANTA DIOP DE DAKAR. i

ZITC Zone Intertropicale de Convergence. 5, 6

ANNEXE D

GLOSSAIRE

benthos l'ensemble des organismes aquatiques (marins ou dulcicoles) vivant à proximité du fond des mers et océans, des lacs et cours d'eau. 3, 4

chalutage Engin de pêche. 3, 6, 57, 64, 91

démersales Les organisme démersaux se dit généralement des espèces nageuses aquatiques qui, dans la colonne d'eau, vivent juste en dessous de la zone pélagique et au-dessus du fond benthique. 1–4, 6

filets engin de pêche. 3

pélagiques Les espèces pélagiques vivent en haute mer et passent leur vie en pleine mer, à proximité de la surface et sont essentiellement planctonophages. 1

ANNEXE E

LISTE DES SYMBOLES

$\mathscr{A}$ σ-Algèbre de sous-ensemnle de Ω. 23, 41

B(x,h) Boule ouverte de centre X et de rayon h. 13, 14, 16, 23

Card cardinal de l'ensemnle..... 46, 47

DD$^{\mathrm{G}}$ Méthode de classification supéviser qui généralise celle des profondeurs. Elle prend en compte plus de deux classes. 23, 41, 75–84

D$_n$ un échnatillon d'observations $\{(\mathrm{X}_i,\mathrm{Y}_i), i = 1...n\}$ quelconques de taille n. 21, 22

$(\mathscr{E}, d(.,.))$ espace fonctionnel générique et sa sémi-métrique $d(.,.)$. 26, 41

O $u_{\mathbf{n}} = \mathrm{O}(v_{\mathbf{n}})$ signifie qu'il existe une constante c telle que $u_{\mathbf{n}} \leq cv_{\mathbf{n}}$. 24, 26, 30, 33, 42, 43, 46

o $u_{\mathbf{n}} = o(v_{\mathbf{n}})$ signifie que $|\frac{u_{\mathbf{n}}}{v_{\mathbf{n}}}| \to 0$. 32, 42

M Un entier naturel supérieur strictement à 1. 22, 23, 25, 27, 34, 46, 47

$\mathbb{N}$ ensemble des entiers naturels :$0,1,2,....$ 27, 41

$\mathbb{N}^{\mathrm{N}}$ espace euclidien d'entier naturel de dimension N. 23–27, 30–32, 35, 41

$\|.\|$ La norme euclidienne. 13, 30, 42

Ω ensemble non vide. 23, 41

$p.c$ presque complètement. 24–27, 30, 33, 35, 46, 48

$\mathbb{P}$ mesure de probabilité sur $\mathscr{A}$. 23, 41

$\mathbb{R}$ ensemble des nombres réels :$]-\infty;+\infty[$. 12, 13, 15, 18, 19, 24, 26, 30–32, 35, 42, 44, 47, 49

$\mathbb{R}_+$ ensemble des nombres réels positifs:$[0;+\infty[$. 24, 31

$\mathbb{R}^d$ espace euclidien réel de dimension d. 16, 17, 22–24, 30–32, 34, 35, 41, 42

$\mathbb{R}^{\mathrm{N}}$ espace euclidien réel de dimension N. 15, 17, 30, 32, 42, 44

$\mathbb{R}^p$ espace euclidien réel de dimension p. 12, 13, 19

$\mathbb{R}_+^*$ ensemble des nombres réels strictement positifs:$]0;+\infty[$. 31, 32

$d(.,.)$ semi-métrique sur un espace fonctionnel $\mathscr{E}$. 13, 23, 26, 41

σ tribu engendrée. 32

□ Indique la fin d'une preuve. 128, 133–137

$\mathbb{Z}$ ensemble des entiers relatifs :...., $-2,-1,0,1,2,....$ 41