REPUBLIQUE ALGERIENNE DEMOCRATIQUE&POPULAIRE
MINISTERE DE L'ENSEIGNEMENT SUPERIEUR& DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITE DJILLALILIABES
FACULTE DES SCIENCES EXACTE
SIDI BEL-ABBES

BP 89 SBA 22000 –ALGERIE- TEL/FAX 048-54-43-44

THESE

Présentée par:

BENALLOU Mohamed

Pour obtenir le Diplôme de Doctorat en Sciences

Spécialité : Mathématiques
Option : Statistiques

Intitulée

Sur l'estimation partielle linéaire en statistique spatiale

Thèse soutenue le 07/ 07/ 2021
Devant le jury composé de :

Président :
Mr GHARIBALLAH Abdelkader Professeur à L'Université S.B.A
Directeur de thèse :
Mr BENCHIKH Tawfik, Professeur à L'Université S.B.A
Examinateur :
Mme RAHMANI Sâadia Professeur à L'Université Saida
Mr OUADJED Hakim Maître de Conférence A à l'université de Mascara

Année Universitaire 2020-2021

Remerciement

Après avoir remercié sincèrement le Dieu, Seigneur des mondes, qui m'a rendu les choses faciles, je voudrais exprimer toute ma reconnaissance et toute ma gratitude envers mon directeur de thèse, Professeur Tawfik BENCHIKH, pour avoir encadré ce travail et pour la confiance qu'il m'a accordée. Ses conseils et ses encouragements m'ont beaucoup aidé à progresser. Je le remercie aussi pour toutes les relectures, suggestions et commentaires qui m'ont permis d'améliorer la qualité de ma thèse. Je le remercie aussi pour la gentillesse, la patience et la disponibilité qu'il m'a accordées et sur lesquelles j'ai pu compter.

Je voudrais aussi remercier chaleureusement chacun des membres du jury qui me font le grand honneur d'y participer. Je remercie Monsieur. GHARIBE ALLAH Abdelkader, Professeur à l'Université de S.B.A qui m'a fait l'honneur de présider le jury. Je remercie également Madame. RAHMANI Sâadia, Professeur à l'Université de Saida pour l'honneur qu'il m'a fait en acceptant d'examiner mon travail, je l'en remercie vivement pour leur lecture attentive de mon travail. Un grand merci aussi à Monsieur. OUADJED Hakim, Maître de Conférences à l'Université de Mascara pour l'honneur qu'il m'a fait en acceptant d'examiner mon travail. Je remercie les membres du Laboratoire de Statistique et Processus Stochastiques de l'université Djillali Liabès de Sidi Bel Abbès, j'ai toujours trouvé soutien et encouragement. Mes remerciements spéciaux à Monsieur ATTOUCH Mohammed Kadi, Professeur à l'Université de S.B.A pour ses encouragements et son aide fréquents que j'ai reçus de lui. Je n'oublie pas mon ami FETITAH Omar pour son aide, et je vous remercie également le professeur MECHAB Boubaker pour les encouragements que j'ai reçus de lui.

Liste des travaux et publications

1. M. Benallou, M. K. Attouch, T. Benchikh and O. Fetitah, Asymptotic results of semi-functional partial linear regression estimate under functional spatial dependency. COMMUNICATIONS IN STATISTICS ; THEORY AND METHODS. https ://doi.org/10.1080/03610926.2020.1871021
2. Benallou Mohamed, Benchikh Tawfik, Attouch Mohammed Kadi and Fetitah Omar, Some asymptotic results of semi-functional partial linear regression estimate for spatial data (soumis).
3. semi-functional partial linear regression estimate under functional spatial dependency with responses missing at random (soumis).

Table des matières

Chapitre 1

Introduction Générale

1.1 Généralités sur la statistique spatiale

1.1.1 Données spatiale

Les recherches actuelles dans des nombreux domaines tels l'exploitation minière, les sciences de l'environnement et de la terre, la météorologie, l'océanographie, la biologie, la géographie, l'économie, l'épidémiologie, l'agronomie, la géophysique, traitement d'image et bien d'autres sont souvent confrontées à l'analyse d'importantes quantités d'informations qui présentent une composante **spatiale** (position géographique). Par exemple, en météorologiques lorsque on étudiera les cumuls pluviométriques observé sur une collection de stations météo, ou en environnement lorsque on étudiera la pollution atmosphérique. C'est le cas aussi en épidémiologie pour étudier le pourcentage d'individus atteints de Covid-19 dans les 48 wilayas de l'Algérie selon une variable d'intérêt, ou en biomédical dans le cas d'organisation spatiale des noyaux cellulaires afin d'analyser les différences éventuelles entre les modèles d'implantation ectopique et orthotopique.

Ces informations sont, de nos jours, faciles à recueillir et à gérer, mais elles ne pouvent pas être analysées avec les outils standards de la statistique classique dont l'une des hypothèses de base est l'indépendance entre les observations. Donc, il faut avoir recours à des nouvelles méthodes d'analyse qui prennent en compte **la dépendance spatiale** entre les observations.

Selon les cas, les observations appartiennent à un certain type de données spatiales, à savoir, les données géostatistiques, les données latticielles et les processus ponctuels. Lorsque les données peuvent être mesurées en tout point d'un domaine continu, on se place dans le cadre de la géostatistique. Si les

données sont liées à un réseau, on parle de données latticielles. Le dernier type de données spatiales, les processus ponctuels, survient lorsque c'est l'ensemble des sites où ont lieu les observations qui est étudié. Il n'est pas toujours aisé de déterminer le type de données à traiter. Cependant, le point commun entre ces catégories est la présence de dépendance dans une ou plusieurs directions mais qui s'affaiblit lorsque les sites d'observations sont plus éloignés.

Les méthodes à utiliser pour traiter ces données diffèrent selon leurs types et la modélisation de telles activités nécessite de trouver une sorte de corrélation entre certaines variables aléatoires dans un endroit avec d'autres dans des endroits voisins. C'est une caractéristique importante de l'analyse des données spatiales. Ainsi, les méthodes de statistique spatiale vont permettre, entre autres, l'analyse exploratoire des données, l'étude de la corrélation spatiale, leur modélisation jusqu'à la prédiction d'un phénomène en des sites non-observés.

Dans la statistique spatiale, on est amené à étudie des phénomènes observés sur un ensemble spatial D de sites. Mathématiquement, ces observations sont les réalisations d'un processus (appelé champ aléatoire) $X = \{X_{\mathbf{i}}, \mathbf{i} \in D\}$, indexé par un ensemble spatial D, où $X_{\mathbf{i}}$ sont des variables aléatoires appartenant à un espace d'états $\mathcal{E}$. L'ensemble d'indices est un domaine D de $\in \mathbb{R}^N$, $N > 1$ contenant un rectangle de volume strictement positif et l'indice $\mathbf{i}$ (la localisation) varie donc continument dans cet espace et elle est soit fixée et déterministe, soit aléatoire. Par contre, dans la pratique, les observations du champ sont faites en un nombre fini de points déterministes de D où D est un sous-ensemble bidimensionnel ($D \subseteq \mathbb{R}^2$). Mais D peut aussi être unidimensionnel (chromatographie, essai agronomique en ligne) ou encore être un sous-ensemble de $\mathbb{R}^3$ (prospection minière, science du sol, imagerie $3D$). Dans d'autres domaines, comme la statistique bayésienne ou la planification des expériences numériques, on peut faire appel à des espaces D de dimension $d > 3$.

1.1.2 Régression non paramétrique spatiale

L'un des problèmes fondamentals en statistique spatiale est celui de la reconstruction d'un phénomène sur son domaine à partir d'un ensemble discret de valeurs observées. Les techniques de krigeage sont généralement utilisées pour ce dernier objectif, cependant, leurs résultats reposent sur la validité des hypothèses, de sorte qu'un échec dans les hypothèses peut avoir un effet significatif. Compte tenu de ce qui précède, une alternative est d'utilisé une approche non paramétrique.

Les techniques non paramétriques ont été appliquées pour faire des inférences concernant les caractéristiques de la variable d'intérêt dans divers contextes. Ils sont particulièrement utiles lorsque le modèle de distribution n'est pas connu et produisent, en général, des résultats cohérents en imposant des hypothèses simples. Dans ce context, la méthode du noyau a été largement utilisée dans la littérature statistique, en raison de sa simplicité et de son applicabilité (voir par exemple Silverman (1986) ; Härdle (1990) ; Wand et Jones (1995)). Une certaine attention a été accordée à l'estimation non paramétrique par la méthode de noyau de la fonction de régression.

Cette fonction est utlisée pour décrire l'influence ou le lien entre une variable X (explicative) et une variable Y (réponse).

La relation de régression est modélisée par

$$Y_{\mathbf{i}} = r(X_{\mathbf{i}}) + \epsilon_{\mathbf{i}} \ , \ \mathbf{i} \in \mathbb{Z}^N \, ,$$

où $\{Z_{\mathbf{i}} = (X_{\mathbf{i}}, Y_{\mathbf{i}}) \ , \ \mathbf{i} \in \mathbb{Z}^N\} , (N > 1)$ est un processus spatial strictement stationnaire, défini sur un espace probabilisé $(\Omega, \mathcal{A}, \mathbb{P})$ tel que les variables $Z_{\mathbf{i}}$ sont de même distribution que la variable $Z = (X, Y)$, avec Y est une variable aléatoire réelle bornée et X à valeurs dans $\mathbb{R}^d$ avec $d \geq 1$, $r(\cdot)$ est la fonction de régression qui est inconnue et $\epsilon_{\mathbf{i}}$ sont les observations d'erreurs.

On suppose que ce processus spatial est observable dans une région rectangulaire $\mathcal{I}_{\mathbf{n}} = \{\mathbf{i} = (i_1, \ldots, i_k, \ldots, i_N) \in \mathbb{N}^N \ : \ 1 \leq i_k \leq n_k \ , k = 1, \ldots, N\}$ et on note $\mathbf{n} = (n_1, \ldots, n_N)$. Ces régions sont utilisées dans la littérature pour estimer de manière non paramétrique la densité spatiale (Tran (1990) ; Biau et Cadre (2004) ; Wang et Wang (2009)). Les méthodes proposés reste valable lorsque la région observée à une forme plus générale.

L'estimation de la fonction régression ainsi que prévision spatiale sont des problèmes importants dans de nombreuses applications. Le but est alors d'approximer la moyenne conditionnelle $r(X) := \mathbb{E}(Y|X)$, ce qui peut-être fait essentiellement de deux manières :

- Dans l'approche paramétrique, qui suppose que $r(\cdot)$ est une fonction ayant une forme connue contenant des paramètres inconnues, plusieurs statisticiens ont élaborés des outils statistiques pour estimer ces paramètres (voir Ripley (1981), Cressie (1993), Guyon (1995), Anselin et Florax (1995), Cressie et Wikle (2011) et les références qui y figurent). Cependant, de nos jours, la régression paramétrique est très critiquée du fait que nous sommes souvent confrontés à une mauvaise spécification de

modèle. De plus l'hypothèse de normalité faite sur les erreurs du modèle est très forte pour l'obtention d'estimateurs efficaces et convergents.

- Approche non paramétrique qui consiste à ne supposer aucune hypothèse sur la forme de $r(\cdot)$, le traitement de cette approche est très récent pour des données spatiales. La plupart de ces methodes font appel à l'estimateur à noyau.

La méthode du noyau a été introduite indépendamment par Nadaraya (1964) et Watson (1964) pour estimer la fonction de régression à partir d'observations i.i.d par une moyenne pondérée des valeurs des variables de réponse. Comparée à la modélisation paramétrique, la littérature sur la régression spatiale non paramétrique n'est pas exhaustive. Ce qu'après le travail fondateur de Tran (1990) sur l'estimation de la densité spatiale par la méthode de noyau, que un certain nombre d'articles ont été consacrés à la régression non paramétrique spatiale et à la prédiction en utilisant cette méthode.

Un des premiers résultats sur l'estimation non-paramétriques de la régression spatiale par la méthode de noyau a été développée par Lu et Chen (2004). Ils ont étendu l'estimateur de Nadaraya-Watson de $r(.)$ aux données spatiales :

$$\hat{r}_{\mathbf{n}}(x) = \frac{\sum_{\mathbf{i}\in\mathcal{I}_n} Y_{\mathbf{i}} K\left(d(x, X_{\mathbf{i}})\, h_{\mathbf{n}}^{-1}\right)}{\sum_{\mathbf{i}\in\mathcal{I}_n} K\left(d(x, X_{\mathbf{i}})\, h_{\mathbf{n}}^{-1}\right)},\ x \in \mathbb{R}^d \tag{1.1}$$

où $K(.)$ est une fonction noyau et $h_{\mathbf{n}}$ est une suite positive de fenêtres qui dépend de $\mathbf{n}$. Ils ont établi la convergence en probabilité de cet estimateur. Par la suite, Biau et Cadre (2004) ont obtenus la convergence presque complète et la normalité asymptotique de cet estimateur. Dabo-Niang et Yao (2007) ont contribué à cette problématique en s'intéressant à l'estimation de la régression et la prédiction de champs aléatoires continuement indexés. Les vitesses de convergences faible et forte de l'estimateur à noyau de $r(.)$ sont montrées sous des conditions suffisantes sur les coefficients de mélange et sur les fenêtres. De plus, elles proposent une première approche de prédiction spatiale pour des processus continûment indexés. En 2007, Carbon *et al.* ont ajouté des résultats sur la régression spatiale via le modèle autorégressif. L'estimation de la fonction de régression spatiale robuste a été abordée par Gheriballah et al. (2010). Leurs principaux résultats est d'établir le taux de convergence presque complète de ces estimateurs. Dans Menezes *et al.* (2010), la prédic-

tion non paramétrique construit avec un seul noyau sur les sites est proposé quand les sites d'observations sont supposés aléatoires. Sous des hypothèses assez générales, ils montrent que l'erreur quadratique moyenne du prédicteur tend à être négligeable quand la taille de l'échantillon augmente. Ils proposent également des approches alternatives de validation croisée pour sélectionner les fenêtres locales et globales. El-Mechkouri (2010) a affaibli les conditions de la normalité asymptotique de l'estimateur de Tran (1990) en démontrant la convergence en probabilité uniforme de cet estimateur. En parallèle, Karácsony et Filzmoser (2010) ont obtenu la normalité asymptotique de l'estimateur de la fonction de régression de type Nadaraya-Watson pour des champs aléatoires α-mélangeants. Ils considèrent la structure extensive-intensive d'asymptotique qui comble le fossé entre les modèles continus et discrets. Par la suite, Robinson (2011) a établi la convergence en probabilité et la normalité asymptotique de l'estimateur à noyau de la régression sous des conditions moins restrictives. Récemment, Dabo Niang *et al.* (2016) proposent une nouvelle technique d'estimation non papamétrique en utilisant deux noyaux afin de contrôler à la fois la distance entre les observations et les emplacements spatiaux. Une convergence et une cohérence presque complètes en norme L^q ($q \in \mathbb{N}^*$) de l'estimation sont obtenues lorsque les observations sont α-mélangeants.

On trouve également, dans la littérature spatiale, d'autres formes de régressions basées sur d'autres caractéristiques conditionnelles (quantile et mode) qui utilisent la méthode de noyau (cf. Hallin et al. (2009), Dabo-Niang et Thiam (2010), Ould-Abdi et al. (2010a,2010b), Laksaci et Maref (2009), Dabo-Niang et al. (2011,2012a,2012b), Ould-Abdi et al. (2011)).

Notons ausi que d'autres méthodes d'estimations sont proposées pour estimer la fonction de régression spatiale, on cite par exemple, Hallin *et al.* (2004) pour la méthode des polynômes locaux, Li et Tran (2009) pour la méthode des plus proches voisins (kNN-méthode), Wang et Wang (2009) pour l'estimation locale linéaire.

1.2 Régression fonctionnelles

1.2.1 Données fonctionnelles

Au cours de ces dernières années, la branche de la statistique consacrée à l'analyse des données fonctionnelles (FDA) a connu un réel essor tant en termes des développements théoriques et méthodologiques que de la diversifi-

cation des domaines d'application. Ceci revient aux progrès qu'à connu l'outil informatique au niveau des capacités de stockage qui permettent d'enregistrer des données de plus en plus volumineuses. Les indices boursiers, les courbes de consommation, les enregistrements sonores, les images satellites, ne sont que quelques exemples illustrant le grand nombre et la diversité des donnés de nature fonctionnelle auxquelles le statisticien peut être confronté. De plus, même si les données dont dispose le statisticien ne sont pas de nature fonctionnelle, il peut être amené à étudier des variables fonctionnelles construites à partir de son échantillon initial. Ainsi, une nouvelle branche de la statistique, dénommée statistique fonctionnelle s'est développée pour traiter des observations comme éléments aléatoires fonctionnelles.

Ce domaine de recherche en statistique connaît actuellement un grand succès auprès de la communauté des statisticiens. La preuve de cet intérêt est la publication de nombreus travaux scientifiques sur ce sujet ainsi que les nombreuses applications pratiques auxquelles ces données s'y prêtent. A titre d'exemple, Ferraty et Vieu (2002, 2003) se sont intéressés à des données spectrométriques et à des enregistrements vocaux, Besse et al. (2000) à des données météorologiques, Gasser et al. (1998) ont considéré des données médicales.

Parmi les ouvrages de référence en la matière, on peut citer les monographies de Ramsay et Silverman (2002, 2005) pour les aspects appliqués, Bosq (2000) pour les aspects théorique (dans le cadre de séries chronologiques), Ferraty et Vieu (2006) pour une étude non paramétrique. Pour les dernières contributions dans ce domaine, on peut se référer au livre de Hsing et Eubank (2015) ainsi qu'à quelques enquêtes bibliographiques comme Goia et Vieu (2016).

Dans ce context, les données fonctionnelles avec dépendance spatiale sont un nouveau sujet qui offre la possibilité de combiner les connaissances issues des statistiques spatiales et de l'analyse des données fonctionnelles. En effet, l'analyse des données fonctionnelles inclut des méthodes et la théorie pour les données sous forme de fonctions, et les statistiques fonctionnelles spatiales étendent cette approche pour traiter des échantillons de fonctions enregistrés à différents endroits d'une région (les données fonctionnelles dites spatialement corrélées). Cette combinaison a un avenir très prometteur tant du côté appliqué que du côté théorique de la statistique, car les outils de statistiques spatiales multivariées peuvent être généralisés pour être valides sur les données fonctionnelles. On peut consulter Mateu et Romano (2017) et Giraldo et *et al.* (2018) pour les dernières contributions, des applications et une bibliographique

sur le sujet.

D'après Cressie (1993), pour la définition des processus spatiaux, et Ferraty et Vieu (2006), pour les variables aléatoires fonctionnelles, un processus fonctionnel spatial est défini par :

$$\{X_s : s \in D \subset \mathbb{R}^d\}$$

où s est l'indice spatiale (la localisation) localisé dans un espace Euclidien de dimension d, l'ensemble $D \subset \mathbb{R}^d$ (d est généralement égal à 2) peut être fixer ou alétoire, et X_s sont des variables aléatoires prenant des valeurs dans un espace de dimensionnel infini (ou espace fonctionnel).

Comme cela se produit dans l'analyse de données spatiales univariée ou multivariée, la nature de l'ensemble D permet de classer les données fonctionnelles spatiales.

- Les données fonctionnelles géostatistiques apparaissent lorsque D est un sous-ensemble fixe de $\mathbb{R}^d$ à volume positif et n points $s_1, \ldots, s_n$ dans D sont choisis pour observer les fonctions aléatoires X_{s_i},$i = 1, \ldots, n$. L'exemple le plus courant est l'ensemble de données canadiennes sur la température (pour 35 stations météorologiques, la température quotidienne a été calculée en moyenne sur une période de 30 ans (Ramsay et Silverman (2005)).
- Données latticielles fonctionnelles correspondent au cas où D est un ensemble fixe et dénombrable. Il y a généralement une bijection entre D et une partition d'une zone géographique et, pour tout $s \in D$, X_s est une fonction récapitulative d'un l'événement s'est produit dans la partie de la zone correspondant à s par cette bijection. On peut cité comme exemple les pyramides des âges de 38 quartiers de Barcelone (Espagne). Les 38 zones constituent une division officielle de la municipalité de Barcelone (elles sont appelées zones statistiques). Les données sur la population, classées par sexe et par âge (par intervalles d'un an), ont été utilisées pour construire des pyramides de chaque zone statistique.

Les objectifs de ce type des données fonctionnelles sont similaires à ceux concernant les données univariées ou multivariées. Parmi les sont objectif, on peut cité : la détection de la dépendance spatiale (via des tests d'autocorrélation spatiale), identification de clusters spatiaux, et modélisation de la dépendance spatiale (via des modèles de régression spatiale, par exemple).

Étant donné que la statistique pour les données fonctionnelles est un su-

jet relativement nouveau, la littérature sur les statistiques spatiales pour les données fonctionnelles n'est pas exhaustive pour le moment (Ramsay (2008) inclut ce sujet dans la liste des problèmes qui représentent les opportunités de recherche les plus intéressantes à l'analyse des données fonctionnelles). Certaines contributions sur la géostatistique, les processus ponctuels et les données latticielles avec observations fonctionnelles ont été présentées, cependant, la géostatistique est actuellement le sujet le plus développé.

1.2.2 Estimation par la méthode à noyau de la fonction du régression fonctionnelle

En tant qu'outil important de la FDA, la régression fonctionnelle visent à modéliser la relation entre la réponse fonctionnelle (ou scalaire) et les covariables fonctionnelles. Historiquement, les premiers résultats concernent l'estimation de la fonction de régression par la méthode de noyau (dans des espaces vectoriels semi-normé) ont été élaborés par Ferraty et Vieu (2000) où ils ont établi la convergence presque complète dans le cas i.i.d., étendue en (2002), par les mêmes auteurs, pour la prévision des série chronologique et généralisés par la suite pour les données α-mélangeant par Ferraty et Vieu (2004) en utilisant la théorie des probabilités de petites boules. En plus, ils ont exploité l'importance de la modélisation non paramétrique des données fonctionnelles en appliquant leur étude à la discrimination des courbes et à la prévision. Dabo-Niang et Rhomari (2003) étudient l'estimateur à noyau de la régression et prouve la convergence presque sûre ainsi que et la convergence en norme L^p. Cependant, la normalité asymptotique de l'estimateur à noyau de la fonction de régression pour des observations dépendantes (α-mélangeant) a été établi par Masry (2005). Ferraty et al. (2007), considèrent le cas où la variable explicative est à valeurs dans un espace de Banach séparable. Ils ont démontré la convergence en moyenne quadratique de l'estimateur. Ce résultat a été utilisé par Rachdi et Vieu (2007) pour présenter un critère de sélection automatique du paramètre de lissage basé sur la validation croisée. Tandis que Benhenni et al. (2007) construit la version locale de ce critère.

Le cas où la variable réponse est fonctionnelle a été considéré par Dabo-Niang et Rhomari (2009). Ils ont établi la convergence en norme L^p de l'estimateur alors que Ferraty et al (2011) ont obtenu la convergence presque complète et uniforme de l'estimateur. La généralisation de ces résultats pour les don-

nées dépendantes a été faite par Ferraty et al (2012a). Ensuite, Ferraty *et al.* (2012b) ont établi la normalité asymptotique de cet estimateur. Ling et al. (2015) ont étudié les propriétés asymptotiques de l'estimateur de la fonction de régression quand les données sont fonctionnelles et stationnaires ergodiques avec des valeurs manquant au hasard (MAR) Récemment, Kara-Zaitri et al. (2017) ont démontré la convergence uniforme par rapport au paramètre de lissage.

Pour une discution complète avec une état de l'art sur la modélisation nonparamétrique par régression fonctionnelle avec la méthode de noyau, nous vous envoyons à enquêtes bibliographique de Ling et Vieu (2018).

D'autres approches ont été proposés pour l'estimation de la fonction de régression. On cite par exemple, la méthode des k plus proches voisins utiliser par Burba et al. (2008), les techniques robustes proposées et étudiées par Azzidine et al. (2008), Attouch et al. (2009) et Crambes et al. (2008), et l'estimation par la méthode des polynômes locaux par Barrientos-Marin et al. (2010).

L'estimation de la fonction de régression pour les données fonctionnelles et spatialement dépendante a été considérée par Dabo-Niang et al. (2011). Les auteurs ont établit la vitesse de convergence presque sûre de la version spatiale de l'estimateur à noyau. Attouch et al. (2011) étendre l'un des résultats de Gheriballah et al. (2010) quand les covariables sont de nature fonctionnelles et ils ont démontré la convergence presque complète de ces estimateurs en donnant leurs vitesses de convergence. Tandis que la normalité asymptotique de la régression robuste a été obtenue par Attouch et al. (2012).

Pour les autres méthodes d'estimation de la fonction de régression spatiale, on peut cité le travail de Chouaf et Laksaci (2012) par la méthode locale linéaire. Ils ont démontré la convergence presque complète de la version spatiale de l'estimateur de Barrientos-Marin et al. (2010). On peut également cité les travaux de Giraldo (2011) et Nerini al. (2010) pour l'étude des modèles linéaires et ces applications en statistique fonctionnelle spatiale.

1.3 Régression semi-paramétrique partiellement linéaire

En régression, les modèles semi-paramétriques ont été introduits afin d'équilibrer la faible flexibilité de la régression linéaire et la grande sensibilité aux effets dimensionnels des approches non paramétriques. Leurs intérêt principal

est de prendre en compte à priori, qu'il existe une certaine relation linéaire afin de réduire le cout de l'estimation qu'aurait un modèle non paramétrique, tout en gardant l'effet sous-jacente au modèle non paramétrique pour expliquer les autres relations. Ces modèles ont démontré leur utilité dans de nombreux domaines des sciences appliquées, comme l'économie, les études environnementales, la médecine, ect ... (voir Härdle *et al.* (2000) pour une large enquête).

Dans la littérature, nous distinguons trois types de régression semi-paramétrique exploitant l'approche du noyau : les régressions partiellement linéaires, à coefficients variables et sur indice. L'estimation partiellement linéaire reste la méthode la plus utilisée. Il s'agit des modèles où la variable réponse Y est généralement expliquée par la somme d'une combinaison linéaire inconnue des composantes d'un vecteur aléatoire multivarié X et une transformation inconnue d'une autre variable explicative multivariée Z,

Statistiquement, le modèle de régression semi-paramétrique partiellement linéaire (RPL) est un modèle de régression qui peut être exprimée comme :

$$Y = X^T\beta + g(Z) + \varepsilon\,,$$

où X et Z sont des vecteurs de covariable à valeur dans $\mathbb{R}^p$ et $\mathbb{R}^d$ respectivement, X^T représente la transposée du vecteur X, β est un paramètre inconnu p-dimensionnel, $g(.)$ est une fonction nonlinéaire (lisse) à plusieurs variables à valeurs dans $\mathbb{R}$ et ε est un vecteur de erreurs aléatoire de moyenne nulle et de variance fini σ. On suppose généralement une condition supplémentaire d'indépendance de la varaible ε par rapport au vecteur aléatoire (X, Z).

Historiquement, et depuis les travaux pionniers d'Engle et al. (1986), les techniques d'estimation sur les modèles partiellement linéaire (PLR) est devenu très populaire et plusieurs articles ont été publiés dans le cadre des données i.i.d. (voir, par exemple, Speckman(1988), Robinson (1988), Stock (1989), Linton (1995), ou Liang (2000)) ainsi que pour les données dépendantes (voir, par exemple, Gao (1995) ou Aneiros-Pérez *et al.* (2004)). Dans ces articles, on trouve des résultats asymptotiques (cohérence, normalité asymptotique) sur les estimateurs de chaque composante du modèle PLR, ainsi que des résultats sur les sélecteurs de la bande passante et les tests d'hypothèses pour ces estimateurs.

D'autres méthodes d'estimation, comme la méthode des polynômes locaux (Hamilton et Truong (1997) ; Aneiros-Pérez et Vilar-Fernández (2008)) et lissage par les ondelettes (Chang et Qu, 2004) sont également appliqués aux modèles partiellement linéaire.

Le développement de ce type de méthode pour les données spatiales est faible, mais il est toujours nécessaire d'explorer et de caractériser les relations de dépendance non linéaires. Cressie (1993) fait le premier à souligne la possibilité d'essayer de tels modèles pour les données spatiales. Gao *et al.* (2006) propose une approche de régression spatiale semi-paramétrique basée sur la combinaison de la technique d'intégration dite marginale avec une approche linéaire locale. La convergence asymptotique et les taux de convergence sont établis dans certaines conditions. Brown *et al.* (2016) considère le cas de conception fixes et aléatoires.

En 2006, Aneiros et Vieu ont étendu ce modèle au cas où Z est à valeurs dans un espace semi-métrique éventuellement de dimension infinie avec des données i.i.d. Ils le nomme la régression semi fonctionnelle partiellement linéaire (Semi-functional partially linear regression : SFPLR). Les estimateurs de β et $g(.)$ sont obtenus successivement à l'aide de la méthode des moindres carrés et celle de Nadaraya-Watson. Ils ont prouvé la normalité asymptotique de l'estimateur de β et obtenu le taux de convergence de l'estimateur de $g(.)$. Les même auteurs en 2008, ont étendu le modèle au domaine des séries chronologiques et ils ont établi les propriétés asymptotique de ces estimateurs. Par ailleurs, Aneiros-Pérez et Vieu (2011) ont proposé et étudié les sélections de bande passante par validation croisée. Parallèlement, Lian (2011) a étendu le modèle SFPLR au cas où la composante linéaire est également de caractéristique fonctionnelle. Des tests d'hypothèses sur la composante paramétrique du modèle (SFPLR) ont été abordés par Aneiros-Pérez et Vieu (2013). En 2014, Shang utilise une approches bayésiennes, tand que Aneiros et al. (2015a) propose la méthode des moindres carrés pénalisés pour étudier le problème de la sélection des variables. Aneiros-Pérez *et al.* (2015b) ont construit un estimateur de la variance d'erreur du modèle et ils ont établi la normalité asymptotique ainsi que la loi du logarithme itéré de l'estimateur. Récemment, Ling *et al.* (2017) ont proposé une procédure k-plus proche-voisins (kNN) et ont dérivé les performances asymptotiques des estimateurs. En parallèle, Boente *et al (2017)* ont considéré une estimation robuste de β et $g(.)$. Germán *et al.* (2017) propose des procédures de bootstrap naïves et sauvages pour approximer la distribution des estimateurs et leurs validités asymptotiques sont obtenues. Des applications de ces procédures sur des données simulées et réelles sont présenter. On peut se référer à Lin it et al. (2020) et ses références pour plus de modèles et de discussion.

Dans cette thèse, c'est ce modèle qu'on va l'utilisée pour les données spatiales et le comparé avec d'autres méthodes non paramétriques.

1.4 Régression avec les données manquante

1.4.1 Données manquantes : motivation

Au cours des dernières décennies, le problème des données manquantes a attiré beaucoup d'attention au sein de la communauté des statisticiens. On les trouve souvent dans les recherches en économie, en sociologie, en science politique (c'est le cas par exemple, où les gouvernements ne peuvent pas ou elle choisissent de ne pas communiquer certaines statistiques critiques), ou dans les études médicales (par xemple, dans les essais cliniques, les temps de survie et d'autres covariables peuvent être manquants en raison de l'abandon du patient) ou dans l'environnement (c'est le cas par exemple, ou les données ne sont pas enregistrées en raison de défaillances d'appareils ou d'instruments de mesure), ect. Différents types de données manquantes existent, ce qui impacte de différentes manières la validité des conclusions d'une recherche. Dans leur livre fondateur, Little et Rubin (2002) identifient trois mécanismes possibles de manque des données. Si la probabilité d'être manquant est la même pour toutes les observations, alors on peut supposer que la distribution des données manquantes ne dépend d'aucune des variables observées ou manquantes. Dans ce cas, on dit que les données sont complètement manquantes au hasard (Missing Completely At Random ou MCAR). Si la distribution des données manquantes dépend des variables complètement observées (et ne dépend pas de celles manquantes), les données sont dites manquantes au hasard (Missing At Random ou MAR). Enfin, les données manquantes par omission prévisible (Missing Not At Random (MNAR)), également appelée données de non-réponses ou données non-ignorables, qui ne sont ni MAR ni MCAR. En ce sens, la valeur de la variable manquante est liée à la raison pour laquelle elle est manquante. Le problème de gestion des données manquantes est un vaste sujet et il ne peut être ignoré lors d'une analyse statistique et elles peuvent avoir un effet significatif sur l'inférence, les performances de prédiction ou toute autre utilisation faite avec les données.

1.4.2 Régression non paramétrque à noyau avec des données manquantes

L'analyse statistique de données incomplètes est un ancien sujet d'étude en statistique. Elle est largement couverte dans le cas multivarié (voir, par exemple, Schafer (2002) ; Little et Rubin (2002) ; Molenberghs *et al.* (2015) ; Trivellore (2015) ; Graham (2012)). Les méthodes courante pour traiter les données manquantes dans un ensemble de données volumineux consiste à imputer (c'est-à-dire à remplacer) une valeur plausible pour chaque donnée manquante, puis à analyser le résultat comme s'il était complet. Les méthodes d'imputation couramment utilisées pour les réponses manquantes comprennent l'imputation par régression linéaire (Yates (1993) ; Healy et Westmacott (1956)), l'imputation par régression par noyau (Cheng (1994)), l'imputation par un ratio (Rao (1996)) et entre autres.

De nombreux travaux sur les données manquantes et ses inférences statistiques pour le modèle de régression peuvent être trouvées dans la littérature statistique lorsque les variables explicatives sont de dimension fini. On peut citer les travaux d'Ali et Abu-Salih (1988) et Siepman et Yang (1994) pour la régression paramétrique, Chen (1994), Chu et Cheng (1995) pour la régression non paramétrique avec un noyau, dont ils ont établi la normalité asymptotique. Toujours pour la régression non paramétrique à noyau, Nittner (2003) et Efromovich (2011 a,b) ont considéré le cas où certaines observations sur les covariables sont MAR mais que les observations variable réponse sont complètement observées. La régression avec la variable réponse et/ou les prédicteurs (covariables) sont MAR a été considéré par Efromovich (2014). Tandis que Boente *et al.* (2009) ont étudié le modèle de régression robuste avec des données manquantes. D'autres méthode d'estimation ont été proposé et étudié, on cite, Chen et Shao (2000) pour l'imputation par la méthode du plus proche voisin, Aerts *et al.* (2002) pour l'imputation multiple non paramétrique. Wang et Rao (2002a) ont développé des approches d'impultaion avec de maximum de vraisemblance pour construire des intervalles de confiance de fonction de régression. Pérez-González et al. (2009) pour l'estimation par les polynômes locaux lorsque les erreurs sont corrélés, Mason *et al.* (2012) pour les approches bayésiennes afin d'ajuster plusieurs covariables manquantes dans les études longitudinales ou transversales. Concernant les modèles semi-paramétrique on trouve par exemple, les travaux de Wang *et al.* (2004) et Ling *et al.* (2007)

pour les modèles partiellement linéaire quand la variable réponse est MAR et les erreur sont corrélés avec les covariables. Pour les dernières contributions dans ce domaine, on peut se référer au livre de Little et Rubin (2019) et pour une référence large on cite Tsiatis (2006).

Lorsque les variables explicatives sont de nature fonctionnelle, très peu de littérature a été rapportée pour étudier les propriétés statistiques de régression non paramétrique.Ferraty et al. (2013) ont proposé, pour la première fois, d'estimer la moyenne d'une variable réponse scalaire basée sur un échantillon i.i.d. dans lequel des variables explicatives sont fonctionnelles et complètement observées, tandis que la variable réponse est MAR. Ils généralisent les résultats de Cheng (1994). Ils ont établi les propriétés asymptotiques de l'estimateur de la fonction de la régression. Tandis que Ling et al. (2015) ont prouvé la normalité asymptotique de l'estimateur proposé par Ferraty et al. (2013) pour des données ergodiques stationnaire et fonctionnelles (séries chronologiques). Rachdi *et al.* (2020a) propose de construire un nouvel estimateur de la fonction de régression en combinant la méthode de k plus proche voisin (kNN) et l'approche d'estimation linéaire locale lorsque le régresseur est de type fonctionnel et que la variable de réponse est un scalaire mais observée avec quelques observations manquantes au hasard (MAR). Nous citons également le travail de Kraus (2015) qui a développé une méthodologie pour l'analyse d'échantillons fonctionnels incomplets où chaque courbe peut être observée sur un sous-ensemble du domaine et non observée ailleurs. De même, Ling et al. (2016) pour l'estimation par noyau du mode conditionnel pour les données fonctionnelle ergodique avec réponses MAR et Ibrahim *et al.* (2020) pour les données de substitution. Rachdi *et al.* (2020b) utilisent les mêmes techniques d'estimation en combinant la méthode de k plus proche voisin (kNN) et l'approche d'estimation linéaire locale pour estimer efficacement la fonction de répartition conditionnelle d'une variable de réponse scalaire, avec des données manquantes au hasard, étant donné une covariable fonctionnelle. On cite aussi le travail de Müller (2009) pour les modèles nonlinéaires fonctionnelle. Pour les modèles semi fonctionnels partiellement linéaire, ling *et al* (2018) construire les estimateurs de la partie paramétrique et non paramétrique respectivement avec variable réponse MAR. Ils établissent quelques propriétés asymptotiques des estimateurs telles que la convergence presque sûres et le taux de convergence de la composante non paramétrique et les lois asymptotique de la partie paramétriques qu'ils sont obtenus sous certaines conditions.

1.5 Contribution de la thèse

1.5.1 Problématique

Les données fonctionnelles avec dépendance spatiale sont un nouveau sujet qui offre la possibilité de combiner les connaissances issues des statistiques spatiales et de l'analyse des données fonctionnelles. En effet, l'analyse des données fonctionnelles inclut des méthodes et la théorie pour les données sous forme de fonctions, et les statistiques fonctionnelles spatiales étendent cette approche pour traiter des échantillons de fonctions enregistrés à différents endroits d'une région (les données fonctionnelles dites spatialement corrélées). Cette combinaison a un avenir très prometteur tant du côté appliqué que du côté théorique de la statistique car les outils de statistiques spatiales multivariées peuvent être généralisés pour être valides pour les données fonctionnelles. D'autre part, les modèles partiellement linéaire, qui incarnent un compromis entre un modèle non paramétrique et un modèle entièrement paramétrique, sont envisagés pour détourner la malédiction de la dimensionnalité (lorsqu'il s'agit d'étudier un grand nombre de Co-variables). Leurs principales avantage par rapport à la régression non paramétrique est la convergence plus rapide des estimateurs. Le but de cette thèse est de combiner la flexibilité d'une modélisation partiellement linéaire avec la méthodologie récente de traitement non paramétrique des données fonctionnelles spatialement dépendantes. Cela nous amène au modèle de régression semi-fonctionnelle et partiellement linéaire pour les données spatiales.

1.5.2 Plan de la thèse

Notre travail est structurée en cinq chapitres :

Le premier chapitre est introductif, où nous avons exposé un bref descriptif et bibliographique sur les données spatiales, les données fonctionnelles, les données manquantes, la régression non paramétrique à noyau dans le cas multivarié et fonctionnel et les modèles de régression pariellement linéaire.

Dans le deuxième chapitre, les modèles de régression partiellement linéaires lorsque les variables sont fonctionnelles et spatialement dépendantes est introduite. Nous construisons les estimateurs et nous établissons leurs propriétés asymptotiques. Une étude pratique sur des données simulées et réelles est présentée à la fin.

Le troisième chapitre est consacré à la régression spatiale fonctionnelle à noyau dans le cas où la variable réponse est " Missing at random ". Sous des

conditions assez générales, on construit un estimateur à noyau pour la fonction de régression on étudions la convergence en probabilité. Nous présentons à la fin une application à travers des données simulées.

Dans le chapitre quatre, on étendre les résultats obtenus dans le premier chapitre dans le cas où les valeurs de la variable réponse sont incomplètes de manière aléatoire (Missing at random).

Enfin, on achève notre travail en dernier chapitre, par des conclusions sur les résultats obtenus et on donnera également quelques perspectives de recherche qu'on peut les envisager d'entreprendre dans le futur.

1.5.3 Présentation des estimateurs étudiés dans la thèse

Soit $\mathbb{Z}^2$ le réseau entier des points (dans l'espace euclidien à deux dimensions). Un point en gras $\mathbf{i} = (i_1, i_2)$ fait référence à un site. On considère le champ aléatoire $\Gamma_{\mathbf{i}} = (Y_{\mathbf{i}}, X_{\mathbf{i}}, Z_{\mathbf{i}})\,, \mathbf{i} \in \mathbb{Z}^2$ à valeurs dans $\mathbb{R} \times \mathbb{R}^p \times \mathcal{E}$, où $(\mathcal{E}, d\,(.,.))$ est un espace semi-métrique de dimension éventuellement infinie.

Par ailleurs, nous supposons que le champ aléatoire est observé sur l'ensemble $\mathcal{I}_{\mathbf{n}} = \{\mathbf{i} = (i_1, i_2) \in \mathbb{N}^2, 1 < i_1 < n_1, 1 < i_2 < n_2\}$, avec $\mathbf{n} = (n_1, n_2)$, et la taille d'échantillon est $\widehat{\mathbf{n}} = n_1 n_2$. Nous allons examiner l'approximation de la fonction de régression $M(X_{\mathbf{i}}, Z_{\mathbf{i}}) = \mathbb{E}(Y_{\mathbf{i}} | X_{\mathbf{i}}, Z_{\mathbf{i}})$ par une fonction semi-paramétrique (partiellement linéaire) du forme $M_0(X_{\mathbf{i}}, Z_{\mathbf{i}}) = X_{\mathbf{i}}^T \beta + g(Z_{\mathbf{i}})$.

Tout d'abord, dans le cas où les observations sont complètes, on contruit un estimateur par noyau de la fonction $g(.)$ et un estimateur du paramètre β par méthode de moindre carrée sur les observations du processus $\Gamma_{\mathbf{i}}$ sur $I_{\mathbf{n}}$:

$$\widehat{\beta} = \left(\sum_{i_1=1}^{n_1} \sum_{i_2=1}^{n_2} \widetilde{X}_{\mathbf{i}} (\widetilde{X}_{\mathbf{i}})^T \right)^{-1} \left(\sum_{i_1=1}^{n_1} \sum_{i_2=1}^{n_2} \widetilde{Y}_{\mathbf{i}} \widetilde{X}_{\mathbf{i}} \right)$$

et

$$\widehat{g}_{\mathbf{n}}(z) = \sum_{i_1=1}^{n_1} \sum_{i_2=1}^{n_2} w_{\mathbf{n}}(z, Z_{\mathbf{i}}) \left(Y_{\mathbf{i}} - X_{\mathbf{i}}^T \widehat{\beta} \right)$$

où

$$\widetilde{Y}_{\mathbf{i}} = Y_{\mathbf{i}} - \sum_{j_1=1}^{n_1} \sum_{j_2=1}^{n_2} Y_{\mathbf{j}} w_{\mathbf{n}}(Z_{\mathbf{i}}, Z_{\mathbf{j}}) \text{ et } \widetilde{X}_{\mathbf{i}} = X_{\mathbf{i}} - \sum_{j_1=1}^{n_1} \sum_{j_2=1}^{n_2} w_{\mathbf{n}}(Z_{\mathbf{i}}, Z_{\mathbf{j}}) X_{\mathbf{j}}$$

avec

$$w_{\mathbf{n}}(z, Z_{\mathbf{i}}) = \frac{K\left(d(z, Z_{\mathbf{i}})h_{\mathbf{n}}^{-1}\right)}{\sum_{i_1=1}^{n_1}\sum_{i_2=1}^{n_2} K\left(d(z, Z_{\mathbf{i}})h_{\mathbf{n}}^{-1}\right)},$$

où K est un noyau de $\mathbb{R}^+$ à $\mathbb{R}^+$, et $h_{\mathbf{n}}$ est une suite de nombres réels positifs tendant vers zéro lorsque $\mathbf{n}$ tend vers l'infini,

Deuxièment, dans le cas où la variable réponse est « Missing at random», nous observant un échantillon incomplète $\{(Y_{\mathbf{i}}X_{\mathbf{i}}Z_{\mathbf{i}}, \delta_{\mathbf{i}}), \mathbf{i} \in I_{\mathbf{n}}\}$, où δ est un variable aléatoire de Bernoulli, telle que $\delta_{\mathbf{i}} = 1$ si $Y_{\mathbf{i}}$ est observable, et $\delta_{\mathbf{i}} = 0$ autrement. On déduit l'estimateurs de β et $g(.)$ par

$$\widehat{\beta} = \left(\sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} \widetilde{X}_{\mathbf{i}}(\widetilde{X}_{\mathbf{i}})^T\right)^{-1}\left(\sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} \widetilde{Y}_{\mathbf{i}}\widetilde{X}_{\mathbf{i}}\right)$$

et

$$\widehat{g}_{\mathbf{n}}(t) = \sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} w_{\mathbf{n}}(t, T_{\mathbf{i}})\left(Y_{\mathbf{i}} - X_{\mathbf{i}}^T\widehat{\beta}\right)$$

où

$$\widetilde{Y}_{\mathbf{i}} = Y_{\mathbf{i}} - \sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} Y_{\mathbf{i}} w_{\mathbf{n}}(T_{\mathbf{i}}, T_{\mathbf{j}}) \text{ and } \widetilde{X}_{\mathbf{i}} = X_{\mathbf{i}} - \sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} w_{\mathbf{n}}(T_{\mathbf{i}}, T_{\mathbf{j}})X_{\mathbf{i}}$$

avec

$$w_{\mathbf{n}}(T_{\mathbf{i}}, T_{\mathbf{j}}) = \frac{K\left(d(T_{\mathbf{i}}, T_{\mathbf{j}})h_{\mathbf{n}}^{-1}\right)}{\sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} K\left(d(T_{\mathbf{i}}, T_{\mathbf{j}})h_{\mathbf{n}}^{-1}\right)}$$

1.5.4 Présentation des résultats obtenus dans la thèse

Résultat 1 : Pour les modèles SFPLR ; on a établi la normalité asymptotique et la convergence presque sûr de l'estimateur de la partie paramétrique, ainsi que la convergence et la vitesse de convergence presque sûr de l'estimateur de la partie non paramétrique.

Théorème 1 Sous la condition de mélange et les hypothèses de concentration de la loi conjointe de variables spatiales, si $\mathbf{n}$ tend vers l'infini, alors on a :

$$\sqrt{\widehat{\mathbf{n}}}\left(\widehat{\beta}-\beta\right)\xrightarrow{\mathcal{D}}N\left(0,Q\right),$$

avec $Q=\mathbf{B}^{-1}C\left(\mathbf{B}^{-1}\right)^{T}$ est définies dans les hypothèses.

De plus on a :

$$\limsup_{\mathbf{n}\longrightarrow\infty}\sqrt{\frac{\widehat{\mathbf{n}}}{4\log\log(\widehat{\mathbf{n}})}}\left|\widehat{\beta}^{(s)}-\beta^{(s)}\right|\longrightarrow\sigma_{s}\quad a.s$$

où $\sigma_{s}=\left(q_{ss}\right)^{1/2}$ with $q_{ss}=(Q)_{ss}$.

Théorème 2 : Sous les conditions des théorème précédente, on a

$$|\widehat{g}_{\mathbf{n}}(z)-g(z)|\rightarrow 0\ a.s. \tag{3}$$

Et

$$|\widehat{g}_{\mathbf{n}}(z)-g(z)|=O\left(h_{\mathbf{n}}+\sqrt{\frac{\log(\widehat{\mathbf{n}})}{\widehat{\mathbf{n}}(p^{z}_{h_{\widehat{\mathbf{n}}}})}})\right)\quad a.s.$$

Résultat 2 : Concernant la régression spatiale fonctionnelle avec réponse MAR ; on a obtenu la convergence en probabilité de l'estimateur à noyau.

Théorème 3 Sous la condition de mélange et les hypothèses de concentration de la loi conjointe de variables spatiales, comme $\mathbf{n}$ va à l'infini, nous avons :

$$\sqrt{\widehat{\mathbf{n}}(\phi^{x}_{h_{\mathbf{n}}}/\log(\widehat{\mathbf{n}})}\left|\widetilde{r}_{\mathbf{n}}\left(x\right)-r(x)-B_{\mathbf{n}}\right|\rightarrow 0_{p}. \tag{5}$$

De plus, si

$$\widehat{\mathbf{n}}h^{2\alpha}_{\mathbf{n}}(\phi^{x}_{h_{\mathbf{n}}}/\log(\widehat{\mathbf{n}}))\rightarrow 0$$

on à

$$\sqrt{\widehat{\mathbf{n}}(\phi^{x}_{h_{\mathbf{n}}}/\log(\widehat{\mathbf{n}})}\left|\widetilde{r}_{\mathbf{n}}\left(x\right)-r(x)\right|\rightarrow 0_{p}$$

Résultat 3 : Pour les modèles SFPLR spatiale avec réponse MAR, on a établi la normalité asymptotique de l'estimateur de la partie paramétrique et la convergence en probabilité de l'estimateur de la partie non paramétrique

Théorème 4

Sous la condition de mélange et les hypothèses de concentration de la loi conjointe de variables spatiales, si $\mathbf{n}$ va à l'infini, nous avons :

$$\sqrt{\widehat{\mathbf{n}}}\left(\widehat{\beta}-\beta\right) \xrightarrow{\mathcal{D}} N\left(0, \Sigma^{-1} C \left(\Sigma^{-1}\right)^{T}\right)$$

Théorème 5 Sous les condition du théorème 4, si de plus on a

$$\widehat{\mathbf{n}}(\phi^{x}_{h_{\mathbf{n}}} / \log(\widehat{\mathbf{n}}) \to \infty$$

alors,

$$|\widehat{g}_{\mathbf{n}}(t) - g(t)| \to 0_p$$

1.5.5 Outils utilisés en statistique spatiale

Dans le cas des données géostatistiques, principalement considérées dans ce travail, D est un sous-ensemble fixé de $\mathbb{R}^N$ avec $N > 1$. On dénote par $\mathbf{i} = (i_1, \ldots, i_N)$ un site localisé dans un espace Euclidien de dimension N.

Les propriétés asymptotiques des estimateurs peuvent être obtenues en étudiant leurs comportements lorsque $n \to \infty$ selon deux situations. La première correspond au cas le plus restrictif pour lequel on dit que la région D s'étend à l'infini à la même vitesse dans toutes les directions (divergences isotropiques). Dans ce cas, les conditions suivantes sont requises : $n \to \infty$ si $min_{1 \leq k \leq N} n_k \to \infty$ et $\frac{n_j}{n_k} < C$ pour une constante C telle que $0 < C < 1$ et $1 \leq j, k \leq N$. Dans l'autre situation (divergences non-isotropiques), la région D ne s'étend pas à l'infini à la même vitesse dans toutes les directions et seulement $n \to \infty$ si $min_{1 \leq k \leq N} n_k \to \infty$. De plus, notons qu'il existe généralement deux structures possibles pour l'étude asymptotique spatiale. Tout d'abord, l'asymptotique extensive (increasing domain asymptotics) qui concerne la situation où le domaine d'échantillonnage augmente avec le nombre de données et la distance entre deux emplacements d'échantillonnage est délimitée à partir de 0. La seconde asymptotique permet de traiter les situations où les observations se densifient dans un domaine D fixé et borné, on parle alors d'asymptotique de remplissage ou intensive (infill asymptotics).

Mesures de dépendance spatiale

Comme cela se produit souvent dans l'analyse de données spatiales dépendantes, il faut définir le type de dépendance. Nous considérons ici les deux mesures de dépendance suivantes :

Condition de dépendance locale : On suppose aussi que pour tout $\mathbf{i} \neq \mathbf{j} \in \mathbb{Z}^N$ la distribution de probabilité conjointe $\nu_{\mathbf{i},\mathbf{j}}$ de $X_{\mathbf{i}}$ et $X_{\mathbf{j}}$ satisfait

$$\exists \varepsilon_1 \in]0;1] \,,\, \nu_{\mathbf{ij}} \left(B(z_1, h_{(n,m)}) \times B(z_2, h_{(n,m)}) \right) = (p_{h_{\mathbf{n}}}^{z_1} p_{h_{\mathbf{n}}}^{z_2})^{\frac{1+\varepsilon_1}{2}},$$

où $p_{h_{\mathbf{n}}}^{z} = P(Z \in B(z, h_{\mathbf{n}})) = \mu\left(B(z, h_{\mathbf{n}}),\ z \in \mathcal{E}\right.$, appelé probabilité de petite boules dans la littérature (voir Ferraty et Vieu (2006)).

Une telle condition de dépendance locale est nécessaire pour atteindre le même taux de convergence que dans le cas **i.i.d**.

Conditions de mélange

Une autre condition de dépendance complémentaire concernait la condition de mélange qui mesure la dépendance au moyen d'un paramètre de mélange α. Nous supposons que $(Z_{\mathbf{i}}\,, \mathbf{i} \in \mathbb{Z}^N)$ satisfait la condition de mélange suivante : il existe une fonction $\phi(t) \downarrow 0$ quant $t \to \infty$, telle que pour S, S' sous-ensembles de $(\mathbb{Z})^N$ de cardinaux finis, notés par $Card(S), Card(S^{'})$ respectivement, on a :

$$\alpha(\mathfrak{B}(S), \mathfrak{B}(S')) = \sup_{\{A \in \mathfrak{B}(S), B \in \mathfrak{B}(S')\}} \{|P(AB) - P(A)P(B)|\}$$

$$\leq \psi(Card(S), Card(S'))\varphi(dist(S, S')),$$

où $\mathcal{B}(S)$ (*resp.* $\mathcal{B}(S^{'})$) désigne la σ-algèbre engendrée par $(Z_{\mathbf{i}},\ \mathbf{i} \in S)$ (*resp.* $\left(Z_{\mathbf{i}},\ \mathbf{i} \in S^{'}\right)$, dist$(S, S')$ la distance euclidienne entre S and S' , et $\psi : \mathbb{N}^2 \longrightarrow \mathbb{R}^+$ est une fonction symétrique, positive et croissante sur chaque variable.

On suppose que la fonction ψ vérifie l'une des conditions :

$$\forall a, b \in \mathbb{N},\ \psi(a, b) \leq C \min(a, b)\,,$$

ou

$$\forall a, b \in \mathbb{N},\ \psi(a, b) \leq C\,(a + b + 1)^{\lambda}\ ,$$

pour une certain $C > 0$ et $\lambda \geq 1$.

Concernant la fonction $\varphi(.)$, nous n'étudierons que le cas où $\varphi(t)$ vérifie la condition suivante :

$$\varphi(t) \leq Ct^{-\theta}, \quad \text{for some} \quad \theta > 0.$$

Lemma 1.5.1. *(Carbon et al. (1997)). On suppose que $E_1, \cdots, E_r$ des ensembles contenant m sites avec $dist(E_i, E_j) \geq \delta_0$ pour tout $i \neq j$ où $1 \leq i, j \leq r$, . On suppose que $V_1, \cdots V_r$ est une suite de variables aléatoires réelles respectivement des σ-algèbre $B(E_1), \cdots, B(E_r)$ et V_i prend ses valeurs dans $[a; b]$. Alors, il existe une suite de variables aléatoires réelles $V_1^*, \cdots V_r^*$ indépendantes telle que V_i^* à la même distribution que V_i et vérifiant :*

$$\sum_{i=1}^{r} \mathbb{E}\left|V_i - V_i^*\right| \leq 2r(b-a)\psi((r-1)m, m)\varphi(\delta_0)\,.$$

Lemma 1.5.2. ***Theorem 4.1, Bulinski, A., et Shashkin, A. (2006)).*** *On suppose que Υ est un champ aléatoire satisfaisant tous les conditions cidessous :*

1) $\sup_{\mathbf{j}\in\mathbb{Z}^d} \mathbb{E}\left|\Upsilon_{\mathbf{j}}\right|^p < \infty$ *for somme* $p > 2$

2) $\sigma^2 := \sum_{\mathbf{j}\in\mathbb{Z}^d} cov(\Upsilon_{\mathbf{0}}, \Upsilon_{\mathbf{j}}) \neq 0$, *with* $\mathbf{0}$ *is the site* $(0, \ldots, 0)$.

3) Pour toute paire d'ensembles finis disjoints $\mathbf{I}, \mathbf{J} \subset \mathbb{Z}^{\mathbf{d}}$,, *et toute paire de Fonctions borné Lipschitz* $f : \mathbb{R}^{|\mathbf{I}|} \to \mathbb{R}$ *and* $g : \mathbb{R}^{|\mathbf{J}|} \to \mathbb{R}$ *on a*

$$|cov(f(\Upsilon_{\mathbf{i}}, \mathbf{i} \in \mathbf{I}), (f(X_{\mathbf{j}}, \mathbf{j} \in \mathbf{J}))| \leq Lip(f)Lip(g)(|\mathbf{I}| \wedge |\mathbf{J}|)\theta_r.$$

$|V|$ *est le cardinal de l'ensemble fini* V, $r = dist(\mathbf{I}, \mathbf{J}) = \min\{\|\mathbf{i} - \mathbf{j}\| : \mathbf{i} \in \mathbf{I}, \mathbf{j} \in \mathbf{J}\}$, *avec la norme* $\|z\| = \max_{i=1,\cdots,d} |z_i|$, $z = (z_1, \cdots, z_d) \in \mathbb{Z}^d$, *et, pour* $F : \mathbb{R}^n \to \mathbb{R}$,

$$Lip(F) = \sup_{x \neq y} \frac{|F(x) - F(y)|}{|x_1 - y_1| + \cdots + |x_n - y_n|}$$

and $\theta_r \leq c_0 r^{-\lambda}$ *or* $\theta_r \leq c_0 e^{-\lambda r}$, $r \in \mathbb{N}$, *for somme* $c_0 > 1$ *and* $\lambda > 0$.

Ensuite nous avons :

$$\limsup_{\mathbf{n}\to\infty} \frac{S_{\mathbf{n}}}{\sqrt{2d\ Var(S_{\mathbf{n}}) \log\log \hat{\mathbf{n}}}} = 1 \quad a.s.,$$

où $S_{\mathbf{n}} = \sum_{\mathbf{i}\in G_\tau} \Upsilon_{\mathbf{i}}$, $\mathbf{n} = (n_1 \ldots, n_d) \in G_\tau$ *et* $\hat{\mathbf{n}} = n_1 \times \ldots \times n_d$, *avec* $G_\tau = \bigcap_{s=1}^{d} \left\{ j \in \mathbb{N}^d : j_s \geq \left(\prod_{s' \neq s} j_{s'} \right)^{\tau} \right\}$ *pout tout* $\tau \in (0, 1/(d-1))$.

Bibliographie

[1] Aerts, M., Claeskens, G., Hens, N., Molenberghs, G. (2002). Local multiple imputation. Biometrika 89(2) :375-388.

[2] Ali, M. A., Abu-Salih, M. S. (1988). On estimation of missing observations in linear regression models, Sankhya. Indian J. Statist. 50 (B),404-411.

[3] Anselin, L. and Florax, R. J. G. M. (1995). New Directions in Spatial Econometrics. Advances in Spatial Science. Springer.

[4] Aneiros-Pérez G, González-Manteiga W., Vieu. P. (2004). Estimation and testing in a partial regression model under long-memory dependence, Bernoulli 10 :49–78.

[5] Aneiros-Pérez, G., Ferraty, F. and Vieu, P. (2015a). Variable selection in partial linear regression with functional covariate, Statistics, Vol. 49, No. 6, pp. 1322–1347.

[6] Aneiros-Pérez, G., Ling, N. and Vieu P., (2015b). Error variance estimation in semi-functional partially linear regression models, J Nonparametr Stat, Vol 27, No. 3, pp. 316–330.

[7] Aneiros-Pérez G. and Vilar-Fernández J.M (2008). Local polynomial estimation in partial linear regression models under dependence. Computational statistics et data analysis, 52 (5), 2757-2777.

[8] Aneiros-Pérez G. and Vieu P. (2006). Semi-functional partial linear regression. Stat. Probab. Lett., 76, (11), 1102-1110.

[9] Aneiros-Pérez G. and Vieu P. (2008). Nonparametric time series prediction. A semi-functional partial linear modeling. Journal of Multivariate Analysis, 99, 834-857.

[10] Aneiros-Pérez, G. and Vieu, P. (2011). Automatic estimation procedure in partial linear model with functional data. Stat.Pap., 52, (4), 751-771.

[11] Aneiros-Pérez, G. and Vieu, P. (2013). Testing linearity in semi-parametric functional data analysis, Computational Statistics, Vol 28,No. 2, pp. 413–434.

[12] Attouch, M., Chouaf, B. and Laksaci, A. (2012) Nonparametric M-estimation for functional spatial data , Communication of the Korean Statistical society, 19, 193-211.

[13] Attouch, M. K., Gheriballah, A., and Laksaci, A. (2011) Robust nonparametric estimation for functional spatial regression. In Ferraty, F., editor, Recent Advances in Functional Data Analysis and Related Topics, Contributions to Statistics, 27-31. Physica-Verlag HD.

[14] Attouch, M., Laksaci, A. and Ould-Said, E. Asymptotic distribution of robust estimator for functional nonparametric models. Comm. Statist. Theory Methods., 38 (2009), no. 8-10, 1317-1335.

[15] Azzedine, N., Laksaci, A. and Ould-Said, E. On robust nonparametric, regression estimation for a functional regressor. Statist. Probab. Lett., 78 (2008), no. 18, 3216-3221

[16] Barrientos-Marin, J., Ferraty, F. and Vieu, P. Locally modelled regression and functional data. J. Nonparametr. Stat., 22, (2010), no. 5-6,617-632

[17] Benhenni, K., Ferraty, F., Rachdi, M. and Vieu, P. Local smoothing regression with functional data. Comput. Statist., 22 (2007), no. 3, 353-369.

[18] Besse, P., Cardot, H. et Stephenson D. (2000) Autoregressive Forecasting of Some Functional Climatic Variations. Scandinavian Journal of Statistics 27 ; 673-687.

[19] Biau, G. and Cadre, B. (2004). Nonparametric spatial prediction. Statistical Inference for Stochastic Processes, 7(3) :327-349.

[20] Boente, G., Gonzalez-Manteiga, W. and Perez-Gonzalez, A. (2009). Robust nonparametric estimation with missing data. J. Statist. Plann. Inference, 139, 571-592.

[21] Boente, G. and Vahnovan, A. (2017). Robust estimators in semi-functional partial linear regression models, Journal of Multivariate Analysis, Vol 154, No. C, pp. 59–84.

[22] Bosq, D. Linear Processes in Function Spaces : Theory and applications. Lecture Notes in Statistics, 149, Springer. (2000).

[23] Brown, L. D., Levine, M., and Wang, L. (2016). A semiparametric multivariate partially linearmodel : A difference approach.Journal of Statistical Planning and Inference, 178 :99-111.

[24] Bulinski, A., and Shashkin, A., (2006). Strong invariance principle for dependent random fields. Dynamics and Stochastics, 128-143.

[25] Burba, F., Ferraty, F. and Vieu, P. k-nearest neighbour method in functional nonparametric regression. J. Nonparametr. Stat., 21 (2009), no. 4, 453-469.

[26] Carbon, M., Tran, L. T., and Wu, B., (1997). Kernel density estimation for random fields (density estimation for random fields). Stat Probab Lett, 36, (2), 115-125.

[27] Carbon, M., Francq, C., and Tran, L. T. (2007). Kernel regression estimation for random fields. Journal of Statistical Planning and Inference, 137 (3),778-798.

[28] Chang, X.W. et Qu, L. (2004). Wavelet estimation of partially linear models. Computationnal Statistics and Data Analysis. 47(1), 31-48.

[29] Chen, J. et Shao, J. (2000). Nearest Neighbor Imputation for Survey Data. Journal of Official Statistics, 16, 113-131.

[30] Cheng, P. (1994). Nonparametric estimation of mean functionals with data missing at random, J. Amer. Statist. Assoc., 89, 81-87.

[31] Chouaf, A. and Laksaci, A. On the functional local linear estimate for spatial regression. Stat. Risk Model., 29 (2012), no. 3, 189-214.

[32] Chu, C. K., Cheng, P. E. (1995). Nonparametric regression estimation with missing data. J. Statist. Planning Inference. 48,85-99.

[33] Crambes, C., Delsol, L. and Laksaci, A. (2008). Robust nonparametric estimation for functional data. J. Nonparametr. Stat., 20 (7), 573-598.

[34] Cressie, N. A. C. (1993). Statistics for Spatial Data. Wiley, New York.

[35] Cressie, N. and Wikle, C. K. (2011). Statistics for Spatio-Temporal Data. Wiley Series in Probability and Statistics. John Wiley and Sons.

[36] Dabo-Niang, S., Kaid, Z., and Laksaci, A. (2011b). Sur la régression quantile pour variable explicative fonctionnelle : Cas des données spatiales. CRAS, 349(23) :1287-1291.

[37] Dabo-Niang, S., Kaid, Z., and Laksaci, A., (2012a). On spatial conditional mode estimation for a functional regressor. Stat Probab Lett, 82, (7), 1413-1421.

[38] Dabo-Niang, S., Kaid, Z., and Laksaci, A., (2012b). Spatial conditional quantile regression : Weak consistency of a kernel estimate, Rev. Roumaine Math. Pures Appl 57, 311-339.

[39] Dabo-Niang, S., Rachdi, M., and Yao, A.F., (2011). Kernel regression estimation for spatial functional random variables. Far East Journal of Theoretical Statistics, 37, (2), 77-113.

[40] Dabo-Niang, S. and Rhomari, N. (2003). Estimation non paramétrique de la régression avec variable explicative dans un espace métrique. Comptes Rendus Mathematique, 336(1) :75-80.

[41] Dabo-Niang, S. and Rhomari, N. (2009). Kernel regression estimation in a Banach space. J. Statistical Planning and Inference, 139, 1421-1434.

[42] Dabo-Niang, S., Ternynck, C. and Yao, A.-F. (2016). Nonparametric prediction of spatial multivariate data. Journal of Nonparametric Statistics, 28 , 428-458.

[43] Dabo-Niang, S. and Thiam, B. (2010). Robust quantile estimation and prediction for spatial processes. Stat. Probab. Letters, 80(17) :1447-1458.

[44] Dabo-Niang, S. and Yao, A.-F. (2007). Kernel regression estimation for continuous spatial processes. Mathematical Methods of Statistics, 16(4) :298-317.

[45] Dabo-Niang, S., Yao, A.-F., Pischedda, L., Cuny, P., and Gilbert, F. (2010). Spatial mode estimation for functional random fields with application to bioturbation problem. Stochastic Environmental Research and Risk Assessment, 24, (4), 487-497.

[46] Efromovich, S., (2011b) Nonparametric Regression with Predictors Missing at Random, J. Amer. Statist. Assoc., 106 : 306-319.

[47] Efromovich, S., (2011a). Nonparametric regression with responses missing at random. J. Statist. Plann. Inference 141, 3744-3752.

[48] Efromovich, S Nonparametric regression with missing data, Wiley Interdisciplinary Reviews Computational Statistics, 6 (4) 2014, 265-275.

[49] El Machkouri, M. and Stoica, R. (2010). Asymptotic normality of kernel estimates in a regression model for random fields. Journal of Nonparametric Statistics, 22(8) :955-971.

[50] Engle R., Granger C., Rice J. and Weiss A., (1986). Nonparametric estimates of the relation between weather and electricity sales, J. Amer. Statist. Assoc., 81, 310-320.

[51] Ferraty, F., Goia, A. and Vieu, P. (2002). Functional nonparametric model for time series : a fractal approach for dimension reduction. Test, 11 (2), 317-344.

[52] Ferraty, I. Van Keilegom, P. Vieu.(2012b). Regression when both response and predictor are functions, J. Multivar.Anal.109, 10-28.

[53] Ferraty, A. Laksaci, A. Tadj, P. Vieu.(2011). Kernel regression with functional response.Electronic. J. Stat.5 159-171.

[54] Ferraty, A. Laksaci, A. Tadj, P. Vieu.(2012a). Estimation de la fonction de régression pour variable explicative et réponses fonctionnelles dépendante.C. R. Acad. Sci.Paris, Ser.I 350, 717-720.

[55] Ferraty, F., Mas, A. and Vieu, P. (2007). Nonparametric regression on functional data : inference and practical aspects. Aust. N. Z. J. Stat., 49 (3), 267-286.

[56] Ferraty, F., Sued, M, Vieu, P. (2013). Mean estimation with data missing at random for functional covariables, Statistics, 47, 688-706.

[57] Ferraty, F. and Vieu, P. (2000) Dimension fractale et estimation de la régression dans des espaces vectoriels semi-normés. C.R. Math. Acad. Sci. Paris., 330 (2), 139-142.

[58] Ferraty, F. et Vieu, P. (2002) The functional nonparametric model and application to spectrometric data. Comput. Statist. 17 (4) 545-564.

[59] Ferraty, F. and Vieu, P. (2003) Curves discrimination : a nonparametric functional approach Computational Statistics and Data Analysis 44 (1-2) 161-173.

[60] Ferraty, F., Vieu, P. (2004). Nonparametric models for functional data, with application in regression times series prediction and curves discrimination. J. Nonparametric Statist., 16, 111-127.

[61] Ferraty, F. and Vieu, P. (2006). Nonparametric functional data analysis. Theory and Practice. Springer Series in Statistics. New York.

[62] Ferraty, F., Vieu, P. (2011). Kernel regression estimation for functional data.Inthe Oxford Handbook of Functional Data Analysis (Ed. F. Ferraty and Y. Romain).Oxford University Press

[63] Gao, J.T. (1995). The laws of the iterated logarithm of some estimates in partly linear models. Statist. Probab. Lett., 25, 153-162.

[64] Gao, J. T., Lu, Z. and Tjϕstheim D., (2006). Estimation in semiparametric spatial regression. The Annals of Statistics, 34, (3), 1395-1435.

[65] Gasser, T., Hall, P. et Presnell, B. (1998) Nonparametric estimation of the mode of a distribution of random curves. J. R. Statomptes. Soc. Ser. B Stat. Methodol. 60 (4) 681-691.

[66] Gheriballah, A., Laksaci, A., and Rouane, R. (2010). Robust nonparametric estimation for spatial regression. Journal of Statistical Planning and Inference, 140(7) :1656 - 1670.

[67] Germán A., Raña P., Vieu p., et Vilar J. (2017). Bootstrap in semi-functional partial linear regression under dependence. TEST. DOI 10.1007/s11749-017-0566-y.

[68] Giraldo, R., Delicado, P. and Mateu, J. (2011). Geostatistics with infinite dimensional data : a generalization of cokriging and multivariable spatial prediction. Matematica : ICM-ESPOL, 9 (1), 16-21.

[69] Giraldo, R., Dabo-Niang, S. and Martinez, S. (2018). Statistical modeling of spatial big data : An approach from a functional data analysis perspective. Statistics and Probability Letters, 136, 126-129.

[70] Goia, A., Vieu, P., (2016). An introduction to recent advances in high/infinite dimensional statistics. J. Multivariate Anal., 146, 1-6.

[71] Graham, J. W., Missing data analysis and design. NY : Springer, New York, 2012.

[72] Guyon, X. (1995).Random Fields on a Network - Modeling, Statistics, and Applications, Springer, New-York

[73] Hallin, M., Lu, Z., and Tran, L. T. (2004). Local linear spatial regression. The Annals of Statistics, 32(6), 2469-2500.

[74] Hallin, M., Lu, Z. and Yu, K. (2009). Local linear spatial quantile regression, Bernouilli. 15, 659-686.

[75] Hamilton, S. A. and Truong, Y. K. (1997). Local linear estimation in partly linear models. J. Multivariate Anal., 60(1), 1-19.

[76] Härdle, W. (1990), Applied Nonparametric Regression, Econometric Society Monographs, Cambridge : Cambridge University Press.

[77] Härdle W, Liang H. and Gao J (2000). Partially linear models. Physica-Verlag, Heidelberg.

[78] Healy, M.J.R. and Westmacott,M.(1956). Missing values in experiments analyzed on automatic computers. Appl. Statist

[79] Hsing, T. and Eubank, R.L., (2015). Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators. John Wiley and Sons.

[80] Ibrahim F, Ali-Hajj H, Demongeot J, et al. Regression model for surrogate data in high dimensional statistics. Commun Stat Theory Methods. Pages 3206-3227 49, 2020 - Issue 13

[81] Kara-Zaitri, L., Laksaci A., Rachdi M. et Vieu P. (2016). Uniform in bandwidth consistency for various kernel estimators involving functional data,Journal of Nonparametric Statistics, DOI : 10.1080/10485252.2016.1254780

[82] Karácsony, Z. et Filzmoser, P. (2010). Asymptotic normality of kernel type regression estimators for random fields. Journal of Statistical Planning and Inference, 140 : 872-886.

[83] Kraus D., (2015). Components and completion of partially observed functional data. J R Stat Soc B 77 :777-801

[84] Laksaci, A. and Maref, F. (2009). Estimation non paramétrique de quantiles conditionnels pour des variables fonctionnelles spatialement dépendantes. Comptes Rendus Mathématique, 347(17-18) :1075-1080.

[85] Laksaci, A. and Mechab, B. (2010). Estimation non paramétrique de la fonction de hasard avec variable explicative fonctionnelle : cas des données spatiales. Revue Roumaine deMathématiques Pures et Appliquées, 55(1) :35-51.

[86] Li, J. and Tran, L. T. (2009). Nonparametric estimation of conditional expectation. J. Statist. Plann. Inference. 139, 164-175.

[87] Lian H., (2011). Functional partial linear model. J Nonparametr Stat, 23(1),115-128.

[88] Liang ; H. (2000). Asymptotic normality of parametric part in partially linear models with measurement error in the nonparametric part. J. Statist. Plann. Inference, 86(1),51-62.

[89] Liang ; H, Wang S, Carroll R (2007) Partially linear models with missing response variables and error-prone covariates. Biometrika 94 :185-198.

[90] Ling ; N., Liang LL, Vieu P (2015) Nonparametric regression estimation for functional stationary ergodic data with missing at random. J Stat Plan Inference 162 :75-87.

[91] Ling ; N., Liu Y, Vieu P (2016) Conditional mode estimation for functional stationary ergodic data with responses missing at random. Statistics 50 :1-23.

[92] Ling ; N. and Vieu P., (2018). Nonparametric modelling for functional data : selected survey and tracks for future. Statistics, 52, (4), 934-949.

[93] Ling, N., Aneiros-Pérez, G. and Vieu, P., (2017), knn estimation in functional partial linear modeling, Statist. Papers, Vol 61, No. 1, pp. 423–444.

[94] Ling ; N. and Vieu P., (2020). On semiparametric regression in functional data analysis. WIRES Computational Statistics, 12, (6), 20-30.

[95] Linton, O. (1995). Second order approximation in the partially linear regression model, Econometrica, 63, 1079-1112.

[96] Little, R, Rubin, D. : Statistical Analysis with Missing Data, Second Edition, Wiley, New York, (2002)

[97] Little, R. J. A. and Rubin, D. B. (2020) ; Statistical Analysis with Missing Data, 3rd Edition ; Wiley Series in Probability and Statistics

[98] Lu, Z. and Chen, X. (2004). Spatial kernel regression estimation : weak concistency, Stat. and Probab. Lett., 68, pp. 125-136.

[99] Mason, A., Richardson, S., Plewis, I. and Best, N. (2012) Strategy for modelling nonrandom missing data mechanisms in observational studies using Bayesian methods. Journal of Official Statistics, 28, 279-302

[100] Masry, E. (2005). Nonparametric regression estimation for dependent functional data : Asymptotic normality. Stoch. Proc. and their Appl., 115, 155-177.

[101] Mateu, J. and Romano, E., (2017). Advances in spatial functional statistics. Stoch Environ Res Risk Assess, 31, 1-6.

[102] Menezes, R., García-Soidán, P., and Ferreira, C. (2010). Nonparametric spatial prediction under stochastic sampling design. Journal of Nonparametric Statistics, 22(3), 363-377.

[103] Molenberghs G., Fitzmaurice G., Kenward M.K., Tsiatis A., Verbeke G., (2015). Handbook of Missing Data Methodology.

[104] Müller U.U., (2009). Estimating linear functionals in nonparametrique regression with responses missing at random. The Annals of Statistics. 37 (5A), 2245-2277.

[105] Nadaraya, E. A. (1964). On estimating regression. Theory of Probability and Its Applications, 9(1) :141-142.

[106] Nerini, D., Monestiez, P., Manté, C., 2010. Cokriging for spatial functional data. J. Multivariate Anal. 101 (2), 409-418.

[107] Nittner T (2003) Missing at random (MAR) in nonparametric regression-a simulation experiment. Stat Methods Appl 12 :195-210.

[108] Ould Abdi A., Diop A., Dabo-Niang S. et Ould Abdi S.A. (2010a), Estimation non paramétrique du mode conditionnel dans le cas spatial Nonparametric estimation of conditional mode in the spatial case CRAS ; 348 (13), 815-819.

[109] Ould Abdi A., Diop A., Dabo-Niang S. et Ould Abdi S.A., (2010b), Consistency of a Nonparametric Conditional Quantile Estimator for Random Fields, Mathematical Methods of Statistics, 19 (1) :1-21.

[110] Perez-Gonzalez, A., Vilar-Fernandez and J. M. Gonzalez-Manteiga, W. (2009). Asymptotic properties of local polynomial regression with missing data and correlated errors. Ann. Inst. Statist. Math., 61, 85-109.

[111] Rachdi, M., Laksaci, A., Almanjahie, I.M. and Chikr-Elmezouar, Z. (2020). FDA :theoretical and practical efficiency of the local linear estimation based on the kNN smoothing of the conditional distribution when there are missing data. J. Stat. Comput.Simul., 90, 1479-1495.

[112] Rachdi M.,, Laksaci A., Kaid Z., Benchiha A., Fahimah A. Al Awadhi, 2021. k-Nearest neighbors local linear regression for functional and missing data at random, Statistica Neerlandica, Netherlands Society for Statistics and Operations Research, 75(1), 42-65.

[113] Rachdi, M., Vieu, P. (2007). Nonparametric regression functional data : Automatic smoothing parameter selection. J. Statist. Plan. Inf., 137, 2784-2801.

[114] Ramsay, J. (2008). Fda problems that i like to talk about. Personal communication.

[115] Ramsay, J. and Silverman, B. (1997) Functional Data Analysis Springer-Verlag, New York.

[116] Ramsay, J. and Silverman, B. (2002) Applied functional data analysis : Methods and case studies Spinger-Verlag, New York.

[117] Ramsay, J. and Silverman, B. (2005) Functional Data Analysis (Second Edition) Spinger-Verlag, New York.

[118] Rao, J.N.K. (1996). On variance estimation with imputed survey data. Journal of the American Statistical Association, 91, 499-506.

[119] Ripley, B. D. (1981). Spatial Statistics. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc.

[120] Robinson, P. M. (1988). Root-N-consistent semiparametric regression. Econometrica, 56, 931–954.

[121] Robinson, P. M. (2011). Asymptotic theory for nonparametric regression with spatial data. Journal of Econometrics, 165(1) :5-19.

[122] Schafer J.L. and Graham J.W. (2002). Missing Data : Our View of the State of the Art. Psychological Methods, 7 (2), 147-177.

[123] Shang H., (2014). Bayesian bandwidth estimation for a semi-functional partial linear regression model with unknown error density. Comput. Stat., 29 (3-4), 829-848

[124] Siepman, H. R., Yang, S.-S. (1994). Generalized least squares estimation of multivariate nonlinear models with missing data. Commun. Statist., 23(6), 1565-1579.

[125] Silverman, B.W. (1986), Density Estimation for Statistics and Data Analysis, in Monographs on Statistical Subjects, London : Chapman and Hall.

[126] Speckman P. (1988). Kernel smoothing in partial linear models. J. Roy. Statist. Soc. Ser. B, 50, 413-436.

[127] Stock, C. J. (1989). Nonparametric policy analysis. J. Amer. Statist. Assoc., 89, 567-575.

[128] Tran, L. T. (1990). Kernel density estimation on random fields. Journal of Multivariate Analysis, 34(1) :37-53.

[129] Tran, L. T. and Yakowitz, S. (1993). Nearest neighbor estimators for random fields. Journal of Multivariate Analysis, 44(1) :23-46.

[130] Trivellore R. ; (2015). Missing Data Analysis in Practice. Chapman and Hall, Taylor and Francis Group.

[131] Tsiatis A (2006) Semiparametric theory and missing data. Springer, New York

[132] Wand, M.P., and Jones, M.C. (1995), Kernel Smoothing, in Monographs on Statistics and Applied Probability, London : Chapman and Hall.

[133] Wang QH, Linton O, Wolfgang H (2004) Semiparametric regression analysis with missing response at random. J Am Stat Assoc 99 :334-345.

[134] Wang, QH., Rao, JNK (2002). Empirical likelihood-based inference under imputation for missing response data. The Annals of Statistics, 30, 896-924.

[135] Wang QH, Sun ZH(2007) Estimation in partially linear models with missing responses at random. J Multivar Anal 98 :1470-1493

[136] Wang, H.,Wang, J., (2009). Estimation of the trend function for spatio-temporal models. Journal of Nonparametric Statistics, 21 : 567-588.

[137] Watson, G. S. (1964). Smooth regression analysis. Sankhya : The Indian Journal of Statistics, Series A, pages 359-372.

[138] Xu, R., Wang, J., 2008.L1-estimation for spatial nonparametric regression. Nonparametric Statist. 20, 523-537

[139] Yates, F. (1933). The analysis of replicated experiments when the field results are incomplete. Emporium J. Exp. Agriculture, 1, 129-142.

Chapitre 2

Régression semi-fonctionnelle partiellement linéaire pour les données spatiales

Dans ce chapitre, on présente les modèles de régression partiellement linéaires lorsque les variables sont fonctionnelles et spatialement dépendantes. Tout d'abord ; on commence par donner la version spatiale des deux estimateurs à noyau pour les deux composantes (linéaire et non paramétrique). Les propriétés asymptotiques de ces deux estimateurs sont établies. On démontre la convergence presque sûr et la normalité asymptotique de l'estimateur de la partie paramétrique et la convergence presque sûr de l'estimateur de la partie non-paramétrique. La performance des estimateurs est illustré par des données simulées et réelles.

Ce chapitre fait l'objet d'une publication dans "Communication in Statistics-Theory and Methods.

2.1 Introduction.

The spatial indexing, which provides geographical reference of data is naturally arise in many applied fields, for example, soil science, geology, oceanography, econometrics, epidemiology, environmental science, forestry, and many others. The literature on spatial models is relatively abundant, however, the nonparametric treatment of such data is relatively recent compared to the parametric case. Note that, the most of nonparametric spatial contributions deal with univariate or multivariate data (see, Tran (1990)), whereas recent advances of real-time measurement instruments and data storage resources led

to the emergence of functional data. This kind of data (functional data analysis, FDA) has been enjoying increased popularity over the last two decades due to its applicability to problems which are difficult to cast into a framework of scalar or vector observations. For the theory and practice on functional linear regression model and functional nonparametric regression model, one can refer to the monographs by Ramsay and Silverman (2005) and Ferraty and Vieu (2006) respectively. For the latest contributions in this field, one can refer to the book by Hsing and Eubank (2015) and also some bibliographical surveys like Goia and Vieu (2016).

Currently, the literature on spatial statistics for FDA is not extensive (see Dabo-Niang *et al.* (2010), Laksaci and Mechab (2010), Attouch *et al.* (2011), Dabo-Niang *et al.* (2012)), and for the regression model estimation we quote, (Dabo-Niang *et al.* (2011)) and (Ternynck, (2014)), which studied the kernel regression estimation for spatial functional random variables. For a review on the recent proposed methods, one can refer to Mateu and Romano (2017).

In the time series and regression, nonparametric methods were often used for prediction and characterization of nonlinear dependence, among others. The development of this type of method for spatial lattice models is weak, because the data is already on a grid, unless there is missing data, the prediction problem is less relevant, but there is still a need to explore and characterize nonlinear dependency relationships. A rather obvious reason for the lack of progress is the curse of dimensionality. There are often cases in practice where one must consider additional information.

In the nonfunctional setting, a popular way to incorporate additional covariables consists in using semiparametric modeling. It is noteworthy that semiparametric functional regression models offer a well-balanced mixture of parametric models and nonparametric models. Semiparametric functional regression models keep flexibility of parametric regression models and overcome sensitivity to dimensional effects of nonparametric approaches. An important semiparametric model is the partially linear regression model introduced by Engle *et al.* (1986) to study the effect of weather on electricity demand. Since this paper partial linear models estimation techniques have attracted much attention among statisticians and econometricians as considered by Robinson (1988), Rice (1986), Stock (1989), Chen and Shiau (1991, 1994) and Gao (1995a). Many authors have studied the estimation and application of the partially linear regression model (see, the monograph of Härdle *et al.* (2000)).

In functional regression, semiparametric models have been recently introduced by Aneiros-Pérez and Vieu (2006), who proposed a semi-functional partial linear regression (SPFLR) model. In addition, they also constructed the estimators of the parametric component and the nonparametric one, and obtained the asymptotic behaviors of the estimators respectively. Furthermore, Aneiros-Pérez and Vieu (2008) used the SFPLR model to make functional time series predictions. Besides, Aneiros-Pérez and Vieu (2011) and Shang (2014) proposed and studied the bandwidth selections by cross-validation and Bayesian approaches respectively. Meanwhile, Lian (2011) extended the (SPFLR) model to the case where the linear component is also of functional response. Ling *et al.* (2020) proposed a k-nearest-neighbours (k-NN) procedure and derived the asymptotic performances of kNN estimators, Ling *et al.* (2019) studied semi-functional partially linear regression model with responses missing at random. We can refer to Lin *et al.* (2018) and references therein for more models and discussion.

No such development has taken place for spatial lattice models for two or more dimensions, then in our model we incorporate the functional predictor directly in nonparametric component of the partial linear spatial regression model in the lattice of two dimensions. Notice that, this work extends to functional spatial case the results given in Gao *et al.* (2006). Note that the semi-functional partial linear model combines the advantages of a functional nonparametric component with the linear effect of some additional real-valued explanatory variables that may be available to the practitioner. One of the differences between this model and the single-index model (see, Ling and Vieu (2020) for more details on single-index model) is that the latter assumes that the response dependence on the covariates is given, up to the link function, through a linear projection, while the nonparametric structure of the semi-functional partial linear model allows for more general structures.

So, the aim of this work is to develop a new model that will combine the advantages of partial linearity and incorporate additional covariates, with the advantages of functional modeling with spatial dependency. Our paper is organized as follows. The semi-functional partial linear model is presented in Section 2 in the general form of regression estimation involving the spatial dependency, and the estimators of both linear and nonparametric components of the model are defined, Section 3 is devoted to the notations and assumptions, the asymptotic results are given in Section 4, while the section 5 is devoted

to the simulation and real data studies to illustrate the good behavior of the proposed estimators (parametric and nonparametric parts). The proofs of the asymptotic results are postponed in Proof section.

2.2 The Model and the estimates

Let $\mathbb{Z}^2$ be the integer lattice points in the two-dimensional Euclidean space. A point in bold $\mathbf{i} = (i_1, i_2)$ will be referred as a site. In this paper the spatial data can be seen as realizations of a measurable strictly stationary spatial process $\Gamma_{\mathbf{i}} = (Y_{\mathbf{i}}; X_{\mathbf{i}}; Z_{\mathbf{i}})$ defined on a probability space $(\Omega, \mathcal{A}, \mathbf{P})$ such that : the $\Gamma'_{\mathbf{i}}$s have the same distribution as a variable $\Gamma = (Y; X; Z)$; where Y is a real-valued and integrable variable, $X = (X^{(1)}, \cdots, X^{(p)})^T$ a $\mathbb{R}^p-$valued random variable and Z (a functional variable) valued in a separable semi-metric space ($\mathcal{E}$,$d(.,.)$) (of eventually infinite dimension). We now turn to estimation assuming that the data are available for $\mathbf{i}$ in the rectangular region $\mathcal{I}_{n,m} = \{\mathbf{i} = (i_1, i_2) \in \mathbb{N}^2, 1 \leq i_1 \leq n, 1 \leq i_2 \leq m\}$, with a sample size of nm.

We will write :$(n, m) \longrightarrow \infty$ if $\min\{n, m\} \longrightarrow \infty$, and we suppose that n, m tend to infinity at the same rate : $C_1 < |m/n| < C_2$ for some $0 < C_1 < C_2 < \infty$.

We will look at approximating the regression function $M(X_{\mathbf{i}}, Z_{\mathbf{i}}) = \mathbb{E}(Y_{\mathbf{i}}|X_{\mathbf{i}}, Z_{\mathbf{i}})$ of $Y_{\mathbf{i}}$ given $X_{\mathbf{i}}$ and $Z_{\mathbf{i}}$ by a semiparametric (partially linear) function of the form $M_0(X_{\mathbf{i}}, Z_{\mathbf{i}}) = X_{\mathbf{i}}^T\beta + g(Z_{\mathbf{i}})$ such that $\mathbb{E}\left[M(X_{\mathbf{i}}, Z_{\mathbf{i}}) - M_0(X_{\mathbf{i}}, Z_{\mathbf{i}})\right]^2$ is minimized over a class of semiparametric functions of the form $M_0(X_{\mathbf{i}}, Z_{\mathbf{i}})$, so the semi-functional partial linear spatial modeling leads us to assume that :

$$Y_{\mathbf{i}} = X_{\mathbf{i}}^T\beta + g(Z_{\mathbf{i}}) + \varepsilon_{\mathbf{i}} \quad , \ \mathbf{i} \in \mathbb{Z}^2,$$

where $\beta = (\beta_1, \cdots, \beta_p)^T$ is a vector of unknown parameters and $g(.)$ is an unknown function over a semi-metric space $\mathcal{E}$, and the noise $\varepsilon_{\mathbf{i}}$ are centered identically distributed satisfying $\mathbb{E}\left(\varepsilon_{\mathbf{i}}|X_{\mathbf{i}}^{(1)}, \cdots, X_{\mathbf{i}}^{(p)}, Z_{\mathbf{i}}\right) = 0$. To do this approximation, we propose the kernel estimate of the function $g(.)$ and the parameter β based on observations of the process $\Gamma_{\mathbf{i}}$ in some region $\mathcal{I}_{m,n}$:

Step 01 : Estimating $g(.)$ assuming β to be known. For each fixed β

$$g(z) = g(z, \beta) = \mathbb{E}\left[\left(Y_{\mathbf{i}} - X_{\mathbf{i}}^T\beta\right)|Z_{\mathbf{i}} = z\right] .$$

Can be estimated by :

$$\widehat{g}_{n,m}(z,\beta)=\sum_{i_1=1}^{n}\sum_{i_2=1}^{m}w_{n,m}(z,Z_{\mathbf{i}})\left(Y_{\mathbf{i}}-X_{\mathbf{i}}^{T}\beta\right)\ , \tag{2.1}$$

where

$$w_{n,m}(z,Z_{\mathbf{i}})=\frac{K\left(d(z,Z_{\mathbf{i}})h_{(n,m)}^{-1}\right)}{\sum\limits_{i_1=1}^{n}\sum\limits_{i_2=1}^{m}K\left(d(z,Z_{\mathbf{i}})h_{(n,m)}^{-1}\right)}, \tag{2.2}$$

and $h_{(n,m)}$being a sequence of bandwidths tending to zero as (n,m) tends to infinity, and the kernel K is a function from $\mathbb{R}^{+}$ to $\mathbb{R}^{+}$.

Step 02 : Estimating β is done by weighted least squares. We obtain

$$\begin{aligned}\widehat{\beta} &= \arg\min_{\beta\in\mathbb{R}^{p}}\sum_{i_1=1}^{n}\sum_{i_2=1}^{m}\left(Y_{\mathbf{i}}-X_{\mathbf{i}}^{T}\beta-\widehat{g}_{n,m}(z,\beta)\right)^{2}\\ &= \arg\min_{\beta\in\mathbb{R}^{p}}\sum_{i_1=1}^{n}\sum_{i_2=1}^{m}\left(\widetilde{Y}_{\mathbf{i}}-(\widetilde{X}_{\mathbf{i}})^{T}\beta\right)^{2},\end{aligned}$$

where

$$\widetilde{Y}_{\mathbf{i}}=Y_{\mathbf{i}}-\sum_{j_1=1}^{n}\sum_{j_2=1}^{m}Y_{\mathbf{i}}w_{n,m}(Z_{\mathbf{i}},Z_{\mathbf{j}})\quad\text{and}\quad\widetilde{X}_{\mathbf{i}}=X_{\mathbf{i}}-\sum_{j_1=1}^{n}\sum_{j_2=1}^{m}w_{n,m}(Z_{\mathbf{i}},Z_{\mathbf{j}})X_{\mathbf{i}}.$$

Therefore

$$\widehat{\beta}=\left(\sum_{i_1=1}^{n}\sum_{i_2=1}^{m}\widetilde{X}_{\mathbf{i}}(\widetilde{X}_{\mathbf{i}})^{T}\right)^{-1}\left(\sum_{i_1=1}^{n}\sum_{i_2=1}^{m}\widetilde{Y}_{\mathbf{i}}\widetilde{X}_{\mathbf{i}}\right). \tag{2.3}$$

We then insert $\widehat{\beta}$ in $\widehat{g}_{m,n}(t,\beta)$, to obtain

$$\widehat{g}_{m,n}(t)=\sum_{i_1=1}^{n}\sum_{i_2=1}^{m}w_{n,m}(z,Z_{\mathbf{i}})\left(Y_{\mathbf{i}}-X_{\mathbf{i}}^{T}\widehat{\beta}\right). \tag{2.4}$$

2.2.1 Hypotheses

In fact, to take into account the spatial dependency, we assume that the process $Z_{\mathbf{i}}$ is strictly stationary and satisfies a mixing condition defined in

Carbon et al (1997). As follows : there exists a function $\varphi(t) \downarrow 0$ as $t \longrightarrow \infty$, such that for S, S' subsets of $\mathbb{Z}^2$ with finite cardinals,

$$\alpha(\mathfrak{B}(S), \mathfrak{B}(S')) = \sup_{\{A \in \mathfrak{B}(S), B \in \mathfrak{B}(S')\}} \{|P(AB) - P(A)P(B)|\}$$

$$\leq \psi(Card(S), Card(S'))\varphi(dist(S, S')),$$

where Card(S) (resp. Card(S')) the cardinality of S (resp. S'), dist(S, S') the Euclidean distance between S and S' , $\mathfrak{B}(S) = \mathfrak{B}(\Gamma_{\mathbf{i}}, \mathbf{i} \in S)$ and $\mathfrak{B}(S') = \mathfrak{B}(\Gamma_{\mathbf{i}}, \mathbf{i} \in S')$ are the σ-fields generated by the random variables $\Gamma_{\mathbf{i}}$, with $\mathbf{i}$ being elements of S and S' respectively and $\psi : \mathbb{N}^2 \longrightarrow \mathbb{R}^+$ is a nondecreasing symmetric positive function in each variable. We will be assumed that ψ satisfies either :

$$\forall a, b \in \mathbb{N},\ \psi(a, b) \leq C \min(a, b)\,, \tag{2.5}$$

or

$$\forall a, b \in \mathbb{N},\ \psi(a, b) \leq C\,(a + b + 1)^{\lambda}\ , \tag{2.6}$$

for some $C > 0$ and some $\lambda \geq 1$.

Concerning the function $\varphi(.)$, we will only study the case where $\varphi(t)$ tends to zero at a polynomial rate, *i.e.*

$$\varphi(t) \leq Ct^{-\theta}, \quad \text{for some} \quad \theta > 0. \tag{2.7}$$

Now we put the following assumptions that are necessary to show our main result.

(H1) Local dependence condition

We assume that for all $\mathbf{i} \neq \mathbf{j} \in \mathbb{Z}^2$ the joint probability distribution $\nu_{\mathbf{ij}}$ of $Z_{\mathbf{i}}$ and $Z_{\mathbf{j}}$ satisfies for some constant $C > 0$ and for all $z_1, z_2 \in \mathcal{E}$

$$\exists \varepsilon_1 \in\,]0; 1]\,,\ \nu_{\mathbf{ij}}\left(B(z_1, h_{(n,m)}) \times B(z_2, h_{(n,m)})\right) = (p^{z_1}_{h_{(n,m)}} p^{z_2}_{h_{(n,m)}})^{\frac{1+\varepsilon_1}{2}},$$

where $p^{z}_{h_{(n,m)}} = P(Z \in B(z, h_{(n,m)})) = \mu\left(B(z, h_{(n,m)}\right)$, $z \in \mathcal{E}$, called small ball probability in the literature (see Ferraty and Vieu (2006)).

(H2) Conditions on the kernel :

We assume that the kernel $K : \mathbb{R}^+ \longrightarrow \mathbb{R}^+$ is a Lipschitz function of integral 1 and is such that : there exist two constants $0 < C_1 < C_2 < \infty$

$$C_1 I_{[0;1]} \leq K \leq C_2 I_{[0;1].}$$

(H3) : Let us introduce the following notation : $f^{(s)}(z) = E\left[X_{\mathbf{i}}^{(s)} | Z_{\mathbf{i}} = z\right]$, $s = 1, \cdots, p$ and $f^{(0)}(z) = E\left[Y_{\mathbf{i}} | Z_{\mathbf{i}} = z\right]$. We assume that $f^{(s)}$ for all $s = 0, 1, \cdots, p$ and g are a Lipschitz functions.

(H4) :

i) We have denoted $\gamma_{\mathbf{i}} = \left(\gamma_{\mathbf{i}}^{(1)}, \cdots, \gamma_{\mathbf{i}}^{(p)}\right)^T$, with $\gamma_{\mathbf{i}}^{(s)} = X_{\mathbf{i}}^{(s)} - \mathbb{E}[X_{\mathbf{i}}^{(s)} | Z_{\mathbf{i}}]$, $s = 1, \cdots, p$, and $\mathbf{B} = \mathbb{E}[\gamma_{\mathbf{1}}(\gamma_{\mathbf{1}})^T]$, where $\mathbf{1}$ is the site spatial $(1, 1)$. Assume that the inverse matrix of $\mathbf{B}$ exist.

ii) Let $R_{\mathbf{i}} = \gamma_{\mathbf{i}}\varepsilon_{\mathbf{i}}$ where $\varepsilon_{\mathbf{i}} = Y_{\mathbf{i}} - M_0(X_{\mathbf{i}}, Z_{\mathbf{i}})$. With assumption, if while $\varepsilon_{\mathbf{i}}$ is independent of $\gamma_{\mathbf{i}}$, the matrix $C = \sum_{i_{1,2}=-\infty}^{\infty} \mathbb{E}[R_{\mathbf{0}}(R_{\mathbf{i}})^T]$ is positive definite, where $\mathbf{0} = (0, 0)$.

(H5) : We suppose that :

i) $\mathbb{E}\left|\varepsilon_{\mathbf{1}}\right|^{\rho} + \mathbb{E}\left|\gamma_{\mathbf{1}}^{(1)}\right|^{\rho} + \cdots + \mathbb{E}\left|\gamma_{\mathbf{1}}^{(p)}\right|^{\rho} < \infty$ for some $\rho \geq 3$.

ii) For all $\mathbf{i} \neq \mathbf{j}$, $\mathbb{E}\left[Y_{\mathbf{i}} Y_{\mathbf{j}}\right| (Z_{\mathbf{i}}, Z_{\mathbf{j}})] < \infty$

iii) For all $\mathbf{i} \neq \mathbf{j}$, $\max_{1 \leq s \leq p} \mathbb{E}\left[X_{\mathbf{i}}^{(s)} X_{\mathbf{j}}^{(s)}\right| (Z_{\mathbf{i}}, Z_{\mathbf{j}})\right] < \infty$.

Comments on the assumptions

The assumptions above are common in the setting of partial linear regression models and they are quite usual in nonfunctional partial linear models, so as it is usual in functional partial linear regression models, the conditions linked with the estimation of the nonparametric component g are exactly the same as those used in pure nonparametric regression models. Therefore, we will naturally need the same set of conditions **(H1)** and **(H2)** as those proposed in Dabo-Niang *et al.* (2011), who are necessary to treat the functional nonparametric component of the model, and the second set of conditions is linked to the linear part of the model, and can be justified in detail. For example, Assumption **(H1)** can be linked with the classical local dependence condition met in the literature of real valued data when X and $(X_{\mathbf{i}}, X_{\mathbf{j}})$ admit, respectively, the densities f and $f_{\mathbf{i},\mathbf{j}}$. Such assumption can be also found in Ferraty and Vieu (2006) (Chapter 11, page 163) and in Dabo-Niang and Yao (2013), assumptions **(H2)** and **(H3)** allows the precise rate of convergence to be found, and assumption **(H4)** is necessary to apply the central limit theo-

rem to have the asymptotic normality. Finally for assumption **(H5)** we observe that, as a consequence of the expressions of our estimators in equations (2.3) and (2.4), assumptions on Y_i and g are similar to those on $X_{\mathbf{i}}^{(s)}$and $f^{(s)}$, respectively are needed as the existence of moments of higher order and to apply the result obtained in the pure non-parametric functional case concerning the convergence.

2.3 Main Results

We set $\pi(m,n) = (\log m \log n)\,(\log\log m \log\log n)^{1+\varepsilon}$, then we have

$$\sum_{(m,n)\in\mathbb{Z}^2} \frac{1}{mn}\pi(m,n) < \infty.$$

Theorem 2.3.1. *Under hypotheses **(H1)-(H5)**, if in addition* $mn(p^z_{h_{(n,m)}}/\log(mn)) \longrightarrow \infty$, *and the mixing parameters satisfies :*

For condition (2.5),

$$\left(mn\left(p^z_{h_{(n,m)}}/\log(mn)\right)^{\frac{4-\theta}{8-\theta}}(\pi((m,n))^{\frac{4}{8-\theta}}\right)^{\frac{8-\theta}{4}} \to \infty,\ \theta > 8,$$

or, for condition (2.6)

$$\left(mn\left(p^z_{h_{(n,m)}}/\log(mn)\right)^{\frac{2-\theta}{2(3+2\lambda)-\theta}}(\pi((m,n))^{\frac{4}{2(3+2\lambda)-\theta}}\right)^{\frac{2(3+2\lambda)-\theta}{4}} \to \infty,\ \theta > 2(3{+}2\lambda).$$

As (n,m) *goes to infinity, we have*

$$\sqrt{mn}\left(\widehat{\beta}-\beta\right) \xrightarrow{\mathcal{D}} N\left(0,Q\right), \tag{2.8}$$

where $Q = \mathbf{B}^{-1}C\left(\mathbf{B}^{-1}\right)^T$, *and*

$$\limsup_{(m,n)\longrightarrow\infty}\sqrt{\frac{mn}{4\log\log(mn)}}\left|\widehat{\beta}^{(s)}-\beta^{(s)}\right| \longrightarrow \sigma_s \quad a.s., \tag{2.9}$$

and $\sigma_s = (q_{ss})^{1/2}$ *with* q_{ss} *is the values of the position* (s,s) *in the matrix* Q.

The next result give the almost surely convergence with rate of the estimator $\widehat{g}_{m,n}(z)$.

Theorem 2.3.2. *Under hypotheses of Theorem (2.3.1), as (n, m) goes to infinity, we have*

$$|\widehat{g}_{n,m}(z) - g(z)| \to 0 \ \ a.s., \tag{2.10}$$

and

$$|\widehat{g}_{m,n}(z) - g(z)| = O\left(h_{(m,n)} + \sqrt{\frac{\log(mn)}{mn(p^z_{h_{(n,m)}})}}\right) \quad a.s. \tag{2.11}$$

The proofs of the Theorems (2.3.1) and (2.3.2) are postponed into the Proofs section.

2.4 Simulation and real data application

2.4.1 Simulation study

In this section, we first illustrate the finite sample behaviours of the proposed estimator $\hat{\beta}$, we have done some simulations (2 models) based on observations $(X_{\mathbf{i}}, Z_{\mathbf{i}}, Y_{\mathbf{i}}) \in (\mathbb{R}^2 \times \mathcal{E} \times \mathbb{R})$ in this case we take $p = 2$, $\mathbf{i} = (i_1, i_2)$ with $1 \leq i_1 \leq n_1$, $1 \leq i_2 \leq n_2$ and $\forall \mathbf{i} \in \mathbb{Z}^2$.

Model 1

The first model was generated as following :

$$Z_{\mathbf{i}}(t) = A_{\mathbf{i}} * (t - 0.5)^2 + B_{\mathbf{i}},$$

$$X_{\mathbf{i}} = (X^1_{\mathbf{i}}, X^2_{\mathbf{i}}),$$

and

$$Y_i = X^T_{\mathbf{i}} \ \beta + g(Z_{\mathbf{i}}) + \varepsilon_{\mathbf{i}}, \tag{2.12}$$

where $\beta = (0.5, \ 2)^T$ (^{T} mean is the transpose symbol), $g(Z) = 4.Z''$ (where Z'' denotes the second derivatives of a function Z). Thereafter, we denote by $GRF\,(m, \sigma^2, s)$ a stationary Gaussian random field with mean m and covariance function defined by $C(l) = \sigma^2 \exp\left(-\left(\frac{\|l\|}{s}\right)^2\right), l \in \mathbb{R}^2$ and $s > 0$.

Then, we have then simulated Model (3.11) with $X_{\mathbf{i}}^k \sim U(-1,2)$, $k = 1,2$. $A = D{*}\sin\left(\frac{G}{2} + .5\right), B = GRF(2.5,5,3), \varepsilon = GRF(0,.1,5), G = GRF(0,5,3)$ and $D_{\mathbf{i}} = \frac{1}{n_1 \times n_2} \sum_{\mathbf{j}} \exp\left(-\frac{\|\mathbf{i}-\mathbf{j}\|}{a}\right)$ $\left(D_{(\mathbf{i},\mathbf{j})} = \frac{1}{n_1 \times n_2} \sum_{1 \leq j_1, j_2 \leq 25} \exp\left(-\frac{\|(i_1,i_2)-(j_1,j_2)\|}{a}\right)\right)$. The function D is here to ensure and control the spatial mixing condition (even if using the Gaussian Random Fields also brings some spatial dependency).

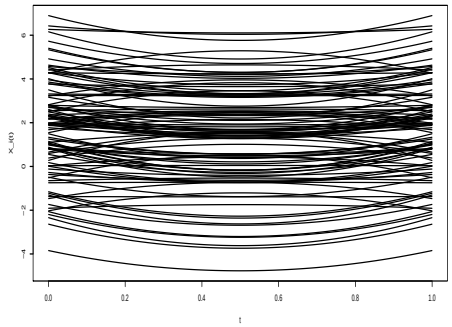

FIGURE 2.1 – The curves $Z_{\mathbf{i}}, t \in [0,1]$.

The used semi-metric is the first derivative of sample curves, given by

$$d\left(Z_{\mathbf{i}}, Z_{\mathbf{j}}\right) = \sqrt{\int_0^1 \left(Z'_{\mathbf{i}}(t) - Z'_{\mathbf{j}}(t)\right)^2 dt}, \quad \text{for } \forall Z_{\mathbf{i}}, Z_{\mathbf{j}} \in \mathcal{E}.$$

In addition, we select the usual kernel function as follows : $K(u) = \frac{3}{2}\left(1-u^2\right) 1_{[0,1]}(u)$.

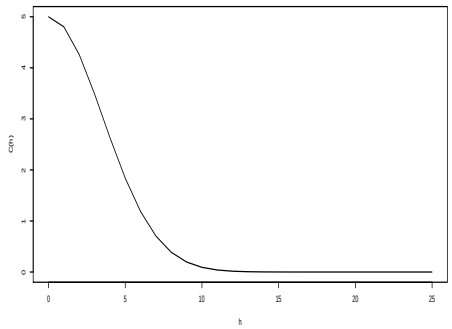

FIGURE 2.2 – Covariance function with $\sigma^2 = 5$ and $s = 5$.

Since (in these conditions) model is based on Gaussian random fields with covariance function C and scale $s = 5$ (see Figure 2.2), observations of sites $\mathbf{i}$

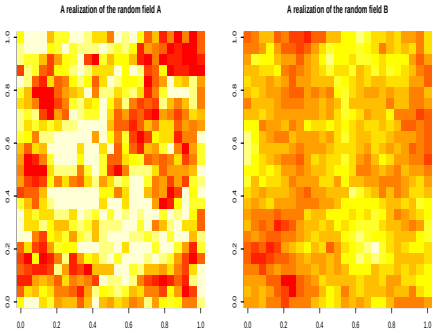

FIGURE 2.3 – Random field simulation.

and $\mathbf{j}$ with $\|\mathbf{i}-\mathbf{j}\| < 15$ are spatial dependent and nearly independent from $\|\mathbf{i}-\mathbf{j}\| \geq 15$. So, our observations are a mixture of i.i.d. and dependent observations (see Figure 2.3). Thus, to move away from independence, it suffices to lower the value of a.

In order to check the performance of the proposed estimator, denoted by $\widehat{r}(x,z) = x^T\widehat{\beta} + \widehat{g}_{m,n}(z)$ withe $(z,x) \in (\mathcal{E} \times \mathbb{R}^2)$, we randomly split our data $(X_\mathbf{i}, Z_\mathbf{i}, Y_\mathbf{i})_\mathbf{i}$ into two subsets : Learning sample $(X_\mathbf{i}, Z_\mathbf{i}, Y_\mathbf{i})_{\mathbf{i}\in I}$ and test sample $(X_\mathbf{i}, Z_\mathbf{i}, Y_\mathbf{i}) \in I'$. The training sample was used to choose the smoothing parameters ($h_{k_{opt}}$ and h_{opt} for the $k-$Nearest Neighbors $k-$NN and cross-validation CV procedures, respectively).

- The h_{opt} is the data-driven bandwidth obtained by a cross-validation procedure : $h_{opt} = \arg\min_h CV(h)$ where $\quad CV(h) = \sum_{\mathbf{i}\in I}\left(Y_\mathbf{i} - \widehat{r}^{CV}_{(-\mathbf{i})}(X_\mathbf{i}, Z_\mathbf{i})\right)^2$
- The $h_{k_{opt}}$ is the bandwidth corresponding to the optimal number of neighbours obtained by a cross-validation procedure :

$$h_k = \min\left\{h \in \mathbb{R}^+ \text{such that } \sum_{\mathbf{i}\in I}\mathbb{I}_{B(x,h)}(Z_\mathbf{i}) = k\right\}$$

$$\text{with } k_{opt} = \arg\min_k CV(k) \text{ where } \quad CV(k) = \sum_{\mathbf{i}\in I}\left(Y_i - \widehat{r}^{kNN}_{(-\mathbf{i})}(X_\mathbf{i}, Z_\mathbf{i})\right)^2$$

with $\widehat{r}^{CV}_{(-\mathbf{i})}$ and $\widehat{r}^{kNN}_{(-\mathbf{i})}$ are the values of the estimator $\widehat{r}_{(-\mathbf{i})}$ calculate at $(X_\mathbf{i}, Z_\mathbf{i})$ (see Ferraty and Vieu (2006) for more details). On the one hand, the accurate of estimates, $\widehat{\beta}$ and $\widehat{g}(\cdot)$, of β and $g(\cdot)$ was measured through the Square and

Mean Square Errors (SE and SE_g) :

$$\mathrm{SE} = \frac{1}{2}\sum_{j=1}^{2} |\widehat{\beta}_j - \beta_j|^2 \text{ and } \mathrm{SE_g} = \frac{1}{\#\left(I'\right)} \sum_{\mathbf{i}\in I'} \left(\widehat{g}\left(Z_{\mathbf{i}}\right) - g\left(Z_{\mathbf{i}}\right)\right)^2.$$

where $\#(I')$ is the size of testing sample I'. Figures 2.4 display boxplots of the corresponding $M = 100$ replications of SE and SE_g, respectively.

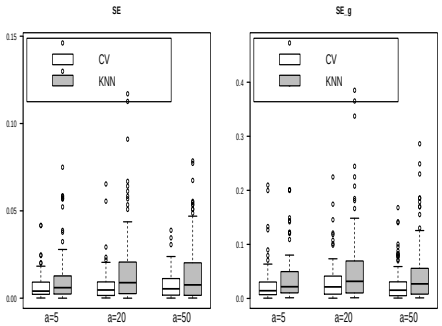

FIGURE 2.4 – Left panel : Mean square errors SE of $\widehat{\beta}$. Right panel : Mean square errors SE_g of $\widehat{g}(.)$, with difference values of a.

To evaluate the efficiency of the proposed model in this prediction issue, we use the Mean Square Error (MSE) for the three models : the vectorial nonparametric regression VNR (see Hädle (2000)) defined as $\widehat{m}_V\left(z\right) = \frac{\sum_{\mathbf{i}\in I} Y_{\mathbf{i}} K\left(\frac{\|z-Z_{\mathbf{i}}\|}{h}\right)}{\sum_{\mathbf{i}\in I} K\left(\frac{\|z-Z_{\mathbf{i}}\|}{h}\right)}$; The FNR functional nonparametric regression introduced by Dabo-Niang *et al.* (2011) defined as $\widehat{m}_F\left(z\right) = \frac{\sum_{\mathbf{i}\in I} Y_{\mathbf{i}} K\left(\frac{d(z,Z_{\mathbf{i}})}{h}\right)}{\sum_{\mathbf{i}\in I} K\left(\frac{d(z,Z_{\mathbf{i}})}{h}\right)}$, and our proposed model SFPLR ($\widehat{r}(x,z)$).

We compute the MSE of the three methods as accuracy measure and are defined as follows :

$$MSE_{VNR} = \frac{1}{\#(I')} \sum_{\mathrm{i}\in I'} \left(Y_{\mathbf{i}} - \widehat{m}_V(Z_{\mathbf{i}})\right)^2, MSE_{FNR} = \frac{1}{\#(I')} \sum_{\mathrm{i}\in I'} \left(Y_{\mathbf{i}} - \widehat{m}_F(Z_{\mathbf{i}})\right)^2$$

$$\text{and } MSE_{SFPLR} = \frac{1}{\#(I')} \sum_{\mathrm{i}\in I'} \left[Y_{\mathbf{i}} - \left(X_{\mathbf{i}}^T\widehat{\beta} + \widehat{g}_{m,n}(Z_{\mathbf{i}})\right)\right]^2,$$

The results obtained for difference values of a are presented in Figure 2.5, where the predicted values are plotted versus the true values.

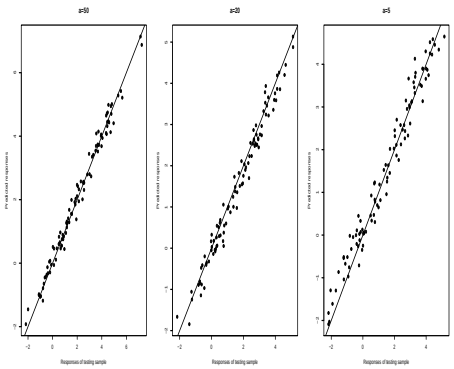

FIGURE 2.5 – Predictions of the first model (3.11) with difference values of a.

TABLE 2.1 – The values of the estimator $\hat{\beta}$ for the combinations of sample sizes (n_1, n_2).

$n_2 \rightarrow$ n_1 $\downarrow$	10	25	50
10	$(0.52831, 1.96795)$	$(0.49505, 2.03652)$	$(0.53218, 2.03006)$
25	$(0.47158, 1.97875)$	$(0.47977, 2.02017)$	$(0.48580, 1.99346)$
50	$(0.49979, 2.02081)$	$(0.50650, 2.00135)$	$(0.50580, 2.01516)$

By Tables 2.1, we can observe that the values of estimators $\widehat{\beta}$ converge to the true values β as n_1 and n_2 increases.

Hence, the MSE under the vectorial nonparametric regression (VNR), functional nonparametric regression (FNR) model and the SFPLR model respectively. Then, we report it in Table 2.2 below.

TABLE 2.2 – Mean squared error (MSE) for VNR, FNR and SFPLR model respectively.

n_1 ↓	Model → n_2 ↓	VNR MSE_{VNR}	FNR MSE_{FNR}	SFPLR MSE_{SFPLR}
10	10	3.53577	3.35116	0.12159
	25	3.39926	3.29978	0.10579
	50	3.35303	3.34031	0.11164
25	10	3.59592	3.40146	0.08223
	25	3.36307	3.33534	0.10814
	50	3.29238	3.27908	0.11695
50	10	3.42431	3.35671	0.10403
	25	3.31251	3.32350	0.10064
	50	3.28555	3.28448	0.10239

By Table 2.2, we observe that the SFPLR model 3.11 shows the better prediction effects in comparison to VNR and FNR models.

Model 2

For the second example we will take :

$$Z_{\mathbf{i}}(t) = A_{\mathbf{i}} \cos(2\pi t), t \in [0, 1].$$

In this case the function $g(.)$ could be $g(T) = \frac{A}{\pi^2} T''$. for the other parameters we leave them as they were, except for the $A = D \times (2\cos(2G) + \exp(-4G^2))$ The curve of $Z(t)$, is drawing on Figure 2.6

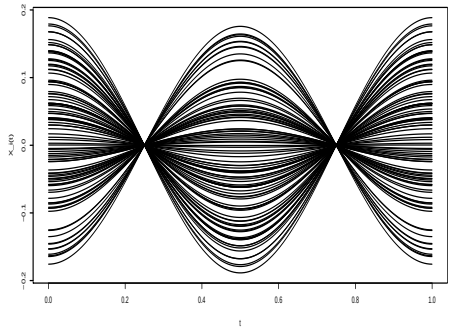

FIGURE 2.6 – The curves $Z_{\mathbf{i}}, t \in [0, 1]$ for the second case.

Figures 2.7 display boxplots of the corresponding $M = 100$ replications of SE and SE_g, respectively.

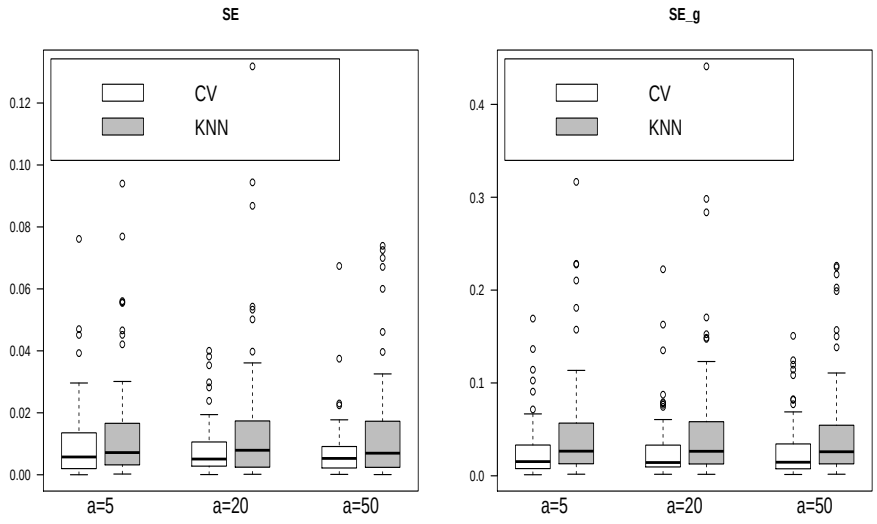

FIGURE 2.7 – Left panel : Mean square errors SE of $\widehat{\beta}$. Right panel : Mean square errors SE_g of $\widehat{g}(.)$, with difference values of a for the second case.

Figure 2.8 illustrate the predicted values versus the true values obtained for difference values of a.

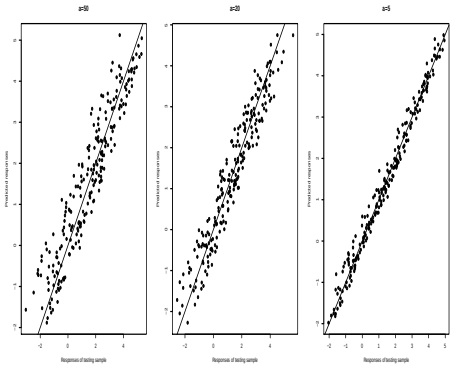

FIGURE 2.8 – Predictions of the second model with difference values of a.

TABLE 2.3 – The values of the estimator $\hat{\beta}$ for the second model according to the sizes (n_1, n_2).

$n_2 \rightarrow$ / $n_1 \downarrow$	10	25	50
10	(0.4756084 , 1.9477183)	(0.4877611 , 2.0516753)	(0.488783 , 2.012454)
25	(0.4810403 , 1.9945778)	(0.5018481 , 2.0165716)	(0.5149852 , 1.9861143)
50	(0.5173907 , 1.9979463)	(0.5052932 , 2.0240304)	(0.5008047 , 2.0157739)

By Tables 2.3, we can observe the convergence of the estimators $\widehat{\beta}$ to the true values β as n_1 and n_2 increases.

TABLE 2.4 – The second case : MSE for VNR, FNR and SFPLR model respectively.

$n_1 \downarrow$	Model $\rightarrow$ / $n_2 \downarrow$	VNR MSE_{VNR}	FNR MSE_{FNR}	SFPLR MSE_{SFPLR}
10	10	2.02512	0.57763	0.10323
	25	2.13186	0.22909	0.09246
	50	1.97262	0.20480	0.08423
25	10	2.20037	0.25174	0.08932
	25	1.38338	0.15638	0.07207
	50	1.01466	0.12872	0.08034
50	10	1.59915	0.22025	0.09377
	25	0.91456	0.12871	0.07917
	50	0.83417	0.12104	0.06077

We notice in Table 2.4 that the SFPLR model performs better than the FNR and VNR models when the data become large.

2.4.2 Real data analysis

Air pollution is one of the most influential factors in human health. Many different chemical substances contribute to the quality of this last. These chemicals come from a variety of sources. On the one hand, there are natural sources such as forest fires, volcanic eruptions, wind erosion, pollen dispersal, evaporation of organic compounds, and natural radioactivity. Furthermore, on

the other hand, human industrial activity represents the artificial air pollution sources.

The main objective of this section is to apply the theoretical results obtained in the previous section to real data. More specifically, in spatial functional prediction context, we examine the performance of the proposed estimator by the semi-functional partial linear regression (SFLPR) approach through some applications which point out the importance of taking into account the spatial locations of the data.

In this real data example, we are interested in the prediction of the future Ozone O_3 concentration given the curve of the previous days of the O_3 and Sulfur dioxide SO_2. For this purpose application, we considered daily O_3 concentration between January 01,2018 and October 30,2018 and the SO_2 for the day October 30,2018. The data controlled by 246 stations in the United States. Figure 2.9 presents the locations of 246 stations in the USA. These observations are available on the following site : https://www.epa.gov/outdoor-air-quality-data.

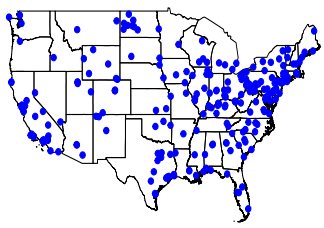

FIGURE 2.9 – The locations of 246 stations in the USA.

Where we assume that the observations are linked by the following regression formula,

$$Y = g(Z) + \beta X + \epsilon.$$

where

the response variable can be taken as : $Y = O_3(244)$(for the day October 30,2018), for the functional variable we take : $Z = O_3(t); t = 1, ..., 243$ (from January 01,2018 to October 29,2018), the additional variable is taken : $X = SO_2$ (for the day October 30,2018).

Specifically, according to the notations of the previous section, the functional predictor $Z_{\mathbf{i}}$ is the curve of the daily ozone in the **i**th station (defined by

its geographic coordinates $\mathbf{i} = (Latitude; Longitude)$), the parametric part is the concentration of the SO_2 ($X_\mathbf{i}$) and $Y_\mathbf{i}$ is future ozone concentration in the same station.

The functional covariates $Z_\mathbf{i}$ are given in the following Figure 2.10.

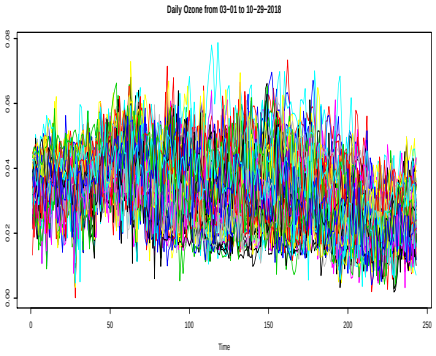

FIGURE 2.10 – Daily ozone concentration in 246 stations in USA.

Our main goal is to compare our estimator for the SFLPR with the functional nonparametric regression (FNR) and the vectorial nonparametric regression (VNR). The kernel K is chosen to be quadratic defined as $K(u) = \frac{3}{4}\left(\frac{12}{11} - u^2\right)\mathbb{I}_{[0,1]}(u)$. The choice of bandwidth parameter h is a crucial question in nonparametric estimation, we propose to choose the optimal bandwidth by using cross-validation procedure. We adopt the selection rule, proposed by Ferraty and Vieu (2006).

We suggests to use standard PCA semi-metrics as following :

$$d_q^{PCA}\left(Z_\mathbf{i}, Z_\mathbf{j}\right) = \sqrt{\sum_{k=1}^{q}\left(\int\left[Z_\mathbf{i}(t) - Z_\mathbf{j}(t)\right]v_k(t)dt\right)^2}.$$

Here, we take $q = 4$, and the v_k is selected among the eigenfunctions of the empirical covariance operator

$$\Gamma_Z^{\hat{n}}(s,t) = \frac{1}{\hat{n}}\sum_{\mathbf{i}\in I} Z_\mathbf{i}(s)Z_\mathbf{i}(t).$$

We randomly split our data $(Z_\mathbf{i}, X_\mathbf{i}, Y_\mathbf{i})_\mathbf{i}$ (we note that $X_\mathbf{i} \in \mathbb{R}$ i.e. $p = 1$) into two subsets : Learning sample $(Z_\mathbf{i}, X_\mathbf{i}, Y_\mathbf{i})_{\mathbf{i}\in I}$ (123 stations), and test sample

$(Z_{\mathbf{i}}, X_{\mathbf{i}}, Y_{\mathbf{i}})_{\mathbf{i}\in I'}$, (123 stations). We use the Mean Square Error (MSE) as accuracy measure defined as follows :

$$MSE = \frac{1}{123}\sum_{\mathbf{i}\in I'}\left(Y_{\mathbf{i}} - \widetilde{Y}_{\mathbf{i}}\right)^2,$$

where $\widetilde{Y}_{\mathbf{i}}$ is the estimator for the three methods VNR, FNR and SFLPR.

Figure 2.11, show the results obtained for the ozone prediction derived from the testing sample by the three models (the SFLPR in the left panel, the FNR in the middle panel and the VNR in the right panel). This is illustrated by the $MSE_{VNR} = 0.02384$, $MSE_{FNR} = 0.02351$ and $MSE_{SFLPR} = 0.02019$.

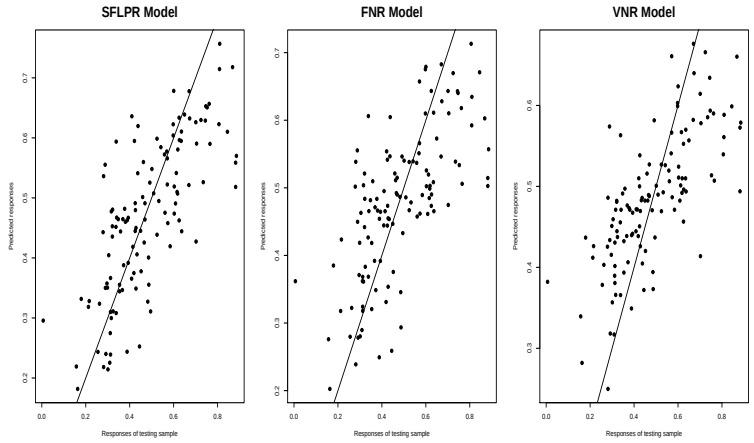

FIGURE 2.11 – Comparison of the prediction results between the three methods.

Figure 2.12 give a comparison of the MSE for the three Models and it shows the good performance of our approach.

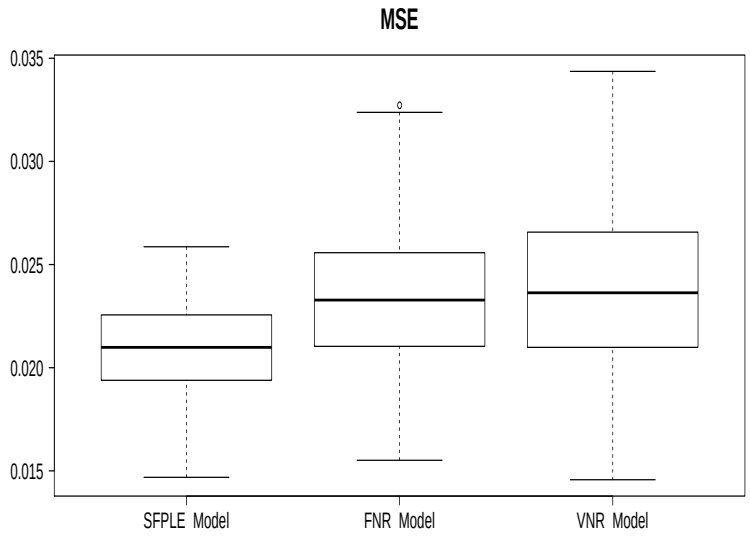

FIGURE 2.12 – Boxplot of the prediction MSE for the three methods.

The estimation of the k-NN kernel offers, in certain situations, a significant advantage over the estimation of the classical kernel. The specificity of the k-NN estimator is that it is flexible to all kinds of covariate heterogeneity, and this by using the local data structure. This method consists of using an appropriate number of neighbors and adapted a random smoothing window allowing us to know more about the dependence of the local data. However, there are recognized problems with this method. It tends to be heavily biased at the edge of the data cloud because the prediction sites will likely be associated with a more central sample value due to the asymmetric neighborhood. Extremely small values and extremely large values will be overestimated and underestimated, respectively, if the sample data does not cover the full range of variability. Bias can also be a problem within the data cloud if the covariates are not evenly distributed. In k-NN methods, the match is found using the covariates available for each site. As the number of samples increases, the chances of getting an exact match in the covariates increase. It would be wise to use the method proposed by Ternynck (2014) to choose sites that will be included in the estimator and thus reduce the bias. This confirms the result obtained in the graph 2.13 which shows the superiority of the cross-validation method to that of the k-NN method.

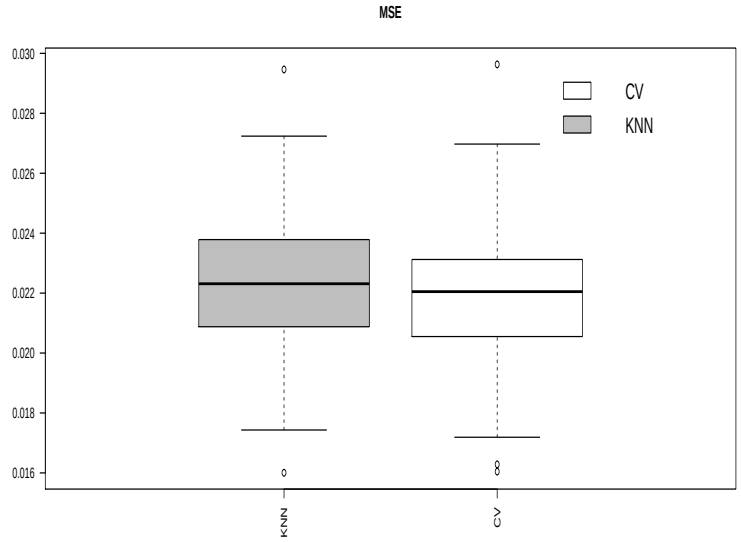

FIGURE 2.13 – Boxplots of the MSE's obtained when one predicts from the SFLPR model, using the $k-$NN and the CV procedures.

2.5 Proofs

Firstly, we state the following Lemmas essential to establish the Theorems 2.3.1 and 2.3.2.

Throughoutthissection,C denotes a generic positive constant which may take different values from one formula to another.

Lemma 2.5.1. *(Carbon et al. (1997)). Let the sets $E_1, \cdots, E_r$ containing each m sites and such that, for all $i \neq j$ and for $1 \leq i, j \leq r$, $dist(E_i, E_j) \geq \delta_0$. Let $V_1, \cdots V_r$ a sequence of random variables with real values and measurable respectively with respect to $B(E_1), \cdots, B(E_r)$. Let be W_l with values in $[a; b]$. There exists a sequence of independent random variables $V_1^*, \cdots V_r^*$ such that W_l^* has the same distribution as W_l and satisfies :*

$$\sum_{l=1}^{r} \mathbb{E}\,|W_l - W_l^*| \leq 2r(b-a)\psi((r-1)m, m)\varphi(\delta_0)\,.$$

Lemma 2.5.2. ***Theorem 4.1, Bulinski, A., & Shashkin, A. (2006)).*** *Suppose that Υ is a random field satisfying all the conditions below :*

1) $\sup_{\mathbf{j}\in\mathbb{Z}^d} \mathbb{E}\,|\Upsilon_{\mathbf{j}}|^p < \infty$ *for somme* $p > 2$

2) $\sigma^2 := \sum_{\mathbf{j}\in\mathbb{Z}^d} cov(\Upsilon_{\mathbf{0}}, \Upsilon_{\mathbf{j}}) \neq 0$, *with* $\mathbf{0}$ *is the site* $(0,\ldots,0)$.

3) For any pair of disjoint finite sets $\mathbf{I},\mathbf{J} \subset \mathbb{Z}^{\mathbf{d}}$, *and any pair of bounded Lipschitz functions* $f : \mathbb{R}^{|\mathbf{I}|} \to \mathbb{R}$ *and* $g : \mathbb{R}^{|\mathbf{J}|} \to \mathbb{R}$ *one has*

$$|cov(f(\Upsilon_{\mathbf{i}}, \mathbf{i}\in\mathbf{I}), (f(X_{\mathbf{j}}, \mathbf{j}\in\mathbf{J}))| \leq Lip(f)Lip(g)(|\mathbf{I}| \wedge |\mathbf{J}|)\theta_r.$$

$|V|$ *is the cardinality of a finite set* V, $r = dist(\mathbf{I},\mathbf{J}) = \min\{\|\mathbf{i}-\mathbf{j}\| : \mathbf{i}\in\mathbf{I}, \mathbf{j}\in\mathbf{J}\}$, *with the norm* $\|z\| = \max_{i=1,\cdots,d} |z_i|$, $z = (z_1,\cdots,z_d) \in \mathbb{Z}^d$, *and, for* $F : \mathbb{R}^n \to \mathbb{R}$,

$$Lip(F) = \sup_{x\neq y} \frac{|F(x)-F(y)|}{|x_1-y_1|+\cdots+|x_n-y_n|}$$

and $\theta_r \leq c_0 r^{-\lambda}$ *or* $\theta_r \leq c_0 e^{-\lambda r}$, $r\in\mathbb{N}$, *for somme* $c_0 > 1$ *and* $\lambda > 0$.

Then we have :

$$\limsup_{\mathbf{n}\to\infty} \frac{S_{\mathbf{n}}}{\sqrt{2d\ Var(S_{\mathbf{n}})\log\log\hat{\mathbf{n}}}} = 1 \quad a.s.,$$

where $S_{\mathbf{n}} = \sum_{\mathbf{i}\in G_\tau} \Upsilon_{\mathbf{i}}$, $\mathbf{n} = (n_1\ldots,n_d) \in G_\tau$ *and* $\hat{\mathbf{n}} = n_1\times\ldots\times n_d$, *with* $G_\tau = \bigcap_{s=1}^{d} \left\{ j\in\mathbb{N}^d : j_s \geq \left(\prod_{s'\neq s} j_{s'}\right)^{\tau} \right\}$ *for any* $\tau \in (0, 1/(d-1))$.

Lemma 2.5.3. *Let us introduce the following notations :*

$$\chi_{(n,m)}(z) = \frac{1}{nm\mathbb{E}\left(K\left(d(z,Z)h_{(m,n)}^{-1}\right)\right)} \sum_{i_1=1}^{m}\sum_{i_2=1}^{n} K\left(d(z,Z_{\mathbf{i}})h_{(m,n)}^{-1}\right),$$

$$\zeta_{(n,m)}(z) = \frac{1}{nm\mathbb{E}\left(K\left(d(z,Z)h_{(m,n)}^{-1}\right)\right)} \sum_{i_1=1}^{m}\sum_{i_2=1}^{n} Y_{\mathbf{i}} K\left(d(z,Z_{\mathbf{i}})h_{(m,n)}^{-1}\right),$$

$$f_{(n,m)}^{(s)}(z) = \frac{1}{nm\mathbb{E}\left(K\left(d(z,Z)h_{(m,n)}^{-1}\right)\right)} \sum_{i_1=1}^{m}\sum_{i_2=1}^{m} X_{\mathbf{i}}^{(s)} K\left(d(z,Z_{\mathbf{i}})h_{(m,n)}^{-1}\right) \text{ for } s = 1,\cdots,p.$$

Under assumptions (H1), and (H2), if in addition the mixing satisfies (3.2) or (3.3) with $\sum_{t=1}^{\infty} t\,(\varphi(t))^a < \infty$, *for some* $0 < a < 1/2$.

$$\forall \Phi_{(n,m)} \in \left\{ \chi_{(n,m)}, \zeta_{(n,m)}, f_{(n,m)}^{(1)}, \cdots, f_{(n,m)}^{(p)} \right\},$$

$$\lim_{(n,m)\to\infty} \left(nmp_{h_{(n,m)}}^{z}\right) var\left(\Phi_{(n,m)}(z)\right) < \infty, \in E.$$

Proof :

By using Lemme 2.5.1, the proof is the same as in Niang et al. (2011) for the case of $\mathbb{Z}^2$, and changing $Y_{\mathbf{i}}$ by $X_{\mathbf{i}}^{(s)}$ if $\Phi_{_{(n,m)}} = f_{_{(n,m)}}^{(s)}$ for each $s = 1, \cdots, p$.

Lemma 2.5.4. *Under assumptions* ***(H1)-(H3)*** *,* ***(H4)(i)*** *and* ***(H5),*** *(g not included in* ***(H3))*** *, and if in addition*

$mn(p_{h_{(n,m)}}^z / \log(mn)) \longrightarrow \infty$ *and the mixing satisfies : condition* (2.5) *with,*

$$\left(mn \left(p_{h_{(n,m)}}^z / \log(mn) \right)^{\frac{4-\theta}{8-\theta}} (\pi((m,n))^{\frac{4}{8-\theta}} \right)^{\frac{8-\theta}{4}} \to \infty,\ \theta > 8,$$

or, condition (2.6) *with*

$$\left(mn \left(p_{h_{(n,m)}}^z / \log(mn) \right)^{\frac{2-\theta}{2(3+2\lambda)-\theta}} (\pi((m,n))^{\frac{4}{2(3+2\lambda)-\theta}} \right)^{\frac{2(3+2\lambda)-\theta}{4}} \to \infty,\ \theta > 2(3+2\lambda).$$

As (n, m) *goes to infinity, we have*

$$\frac{1}{mn} \sum_{i_1=1}^{m} \sum_{i_2=1}^{n} \widetilde{X}_{\mathbf{i}} (\widetilde{X}_{\mathbf{i}})^T \longrightarrow \mathbf{B} \quad a.s. \tag{2.13}$$

Proof :

The (r, s)th element of $\frac{1}{mn} \sum_{i_1=1}^{m} \sum_{i_2=1}^{n} \widetilde{X}_{\mathbf{i}} (\widetilde{X}_{\mathbf{i}})^T$ can be written as

$$\begin{aligned} \left(\frac{1}{mn} \sum_{i_1=1}^{m} \sum_{i_2=1}^{n} \widetilde{X}_{\mathbf{i}} (\widetilde{X}_{\mathbf{i}})^T \right)_{rs} &= \frac{1}{mn} \sum_{i_1=1}^{m} \sum_{i_2=1}^{n} \gamma_{\mathbf{i}}^{(r)} \gamma_{\mathbf{i}}^{(s)} + \frac{1}{mn} \sum_{i_1=1}^{m} \sum_{i_2=1}^{n} \Delta_{\mathbf{i}}^{(r)} \gamma_{\mathbf{i}}^{(s)} \\ &+ \frac{1}{mn} \sum_{i_1=1}^{m} \sum_{i_2=1}^{n} \gamma_{\mathbf{i}}^{(r)} \Delta_{\mathbf{i}}^{(s)} + \frac{1}{mn} \sum_{i_1=1}^{m} \sum_{i_2=1}^{n} \Delta_{\mathbf{i}}^{(r)} \Delta_{\mathbf{i}}^{(s)}, \end{aligned} \tag{2.14}$$

with

$$\Delta_{\mathbf{i}} = (\Delta_{\mathbf{i}}^{(1)}, \cdots, \Delta_{\mathbf{i}}^{(p)})^T = E[X_{\mathbf{i}} | Z_{\mathbf{i}}] - \sum_{j_1=1}^{m} \sum_{j_2=1}^{n} w_{m,n}(Z_{\mathbf{i}}, Z_{\mathbf{j}}) X_{\mathbf{i}},$$

and

$$\Delta_{\mathbf{i}}^{(s)} = E[X_{\mathbf{i}}^{(s)} | Z_{\mathbf{i}}] - \sum_{j_1=1}^{m} \sum_{j_2=1}^{n} w_{m,n}(Z_{\mathbf{i}}, Z_{\mathbf{j}}) X_{\mathbf{i}}^{(s)} \text{ , for } s = 1, \cdots, p.$$

Now, the Strong Law of Large Numbers give

$$\frac{1}{mn}\sum_{i_1=1}^{m}\sum_{i_2=1}^{n}\gamma_{\mathbf{i}}^{(r)}\gamma_{\mathbf{i}}^{(s)} \longrightarrow \mathbb{E}\left[\gamma_{\mathbf{1}}^{(r)}\gamma_{\mathbf{1}}^{(s)}\right] = \mathbf{B}_{rs} \quad a.s. \tag{2.15}$$

In particular,

$$\frac{1}{mn}\sum_{i_1=1}^{m}\sum_{i_2=1}^{n}\gamma_{\mathbf{i}}^{(s)}\gamma_{\mathbf{i}}^{(s)} = O(1)\,. \tag{2.16}$$

Furthermore, by applying directly Theorem 6 in Niang *et al.* (2011), we can see that :

$$\max_{1\leq k\leq p}\max_{\mathbf{i}\in\mathcal{I}_{m,n}}\left|\Delta_{\mathbf{i}}^{(k)}\right| = \Lambda \;\to\; 0 \quad a.s. \tag{2.17}$$

By (2.16) and (2.17), and by using the Cauchy–Schwartz inequality, it is easy to check that

$$\begin{aligned}\frac{1}{mn}\sum_{i_1=1}^{m}\sum_{i_2=1}^{n}\Delta_{\mathbf{i}}^{(r)}\gamma_{\mathbf{i}}^{(s)} &\leq \frac{1}{mn}\left(\sum_{i_1=1}^{m}\sum_{i_2=1}^{n}\left(\Delta_{\mathbf{i}}^{(r)}\right)^2\right)^{1/2}\left(\sum_{i_1=1}^{m}\sum_{i_2=1}^{n}\left(\gamma_{\mathbf{i}}^{(s)}\right)^2\right)^{1/2}\\ &\leq \Lambda\left(\frac{1}{mn}\sum_{i_1=1}^{m}\sum_{i_2=1}^{n}\gamma_{\mathbf{i}}^{(s)}\gamma_{\mathbf{i}}^{(s)}\right)^{1/2} \;\to\; 0 \quad a.s.\end{aligned}$$

So,

$$\left(\frac{1}{mn}\sum_{i_1=1}^{m}\sum_{i_2=1}^{n}\Delta_{\mathbf{i}}^{(r)}\gamma_{\mathbf{i}}^{(s)} + \frac{1}{mn}\sum_{i_1=1}^{m}\sum_{i_2=1}^{n}\gamma_{\mathbf{i}}^{(r)}\Delta_{\mathbf{i}}^{(s)} + \frac{1}{mn}\sum_{i_1=1}^{m}\sum_{i_2=1}^{n}\Delta_{\mathbf{i}}^{(r)}\Delta_{\mathbf{i}}^{(s)}\right) \;\to\; 0 \quad a.s. \tag{2.18}$$

We conclude this proof by using (2.14),(2.15) and (2.18).

Proof of Theorem 2.3.1

The proof of first result (2.8) is based on the following decomposition and the Lemma 2.5.4 :

$$\begin{aligned}\sqrt{mn}\left(\widehat{\beta}-\beta\right) &= \left(\tfrac{1}{mn}\sum_{i_1=1}^{m}\sum_{i_2=1}^{n}\widetilde{X}_{\mathbf{i}}(\widetilde{X}_{\mathbf{i}})^T\right)^{-1}\left(\tfrac{1}{\sqrt{mn}}\sum_{i_1=1}^{m}\sum_{i_2=1}^{n}R_{\mathbf{i}} + \tfrac{1}{\sqrt{mn}}\sum_{i_1=1}^{m}\sum_{i_2=1}^{n}\gamma_{\mathbf{i}}\left(\Delta_{\mathbf{i}}^{(0)}-\Delta_{\mathbf{i}}^{T}\beta\right)\right)\\ &+\left(\tfrac{1}{mn}\sum_{i_1=1}^{m}\sum_{i_2=1}^{n}\widetilde{X}_{\mathbf{i}}(\widetilde{X}_{\mathbf{i}})^T\right)^{-1}\left(\tfrac{1}{\sqrt{mn}}\sum_{i_1=1}^{m}\sum_{i_2=1}^{n}\Delta_{\mathbf{i}}\varepsilon_{\mathbf{i}} + \tfrac{1}{\sqrt{mn}}\sum_{i_1=1}^{m}\sum_{i_2=1}^{n}\Delta_{\mathbf{i}}\left(\Delta_{\mathbf{i}}^{(0)}-\Delta_{\mathbf{i}}^{T}\beta\right)\right)\,,\end{aligned} \tag{2.19}$$

where

$$\Delta_{\mathbf{i}}^{(0)} = E[Y_{\mathbf{i}}|T_{\mathbf{i}}] - \sum_{j_1=1}^{n}\sum_{j_2=1}^{m} w_{n,m}(T_{\mathbf{i}}, T_{\mathbf{j}})Y_{\mathbf{i}}\,.$$

So, we have

$$\widehat{\beta} - \beta = \left(\mathbf{B}^{-1} + o(1)\right)\left(\frac{1}{mn}\sum_{i_1=1}^{m}\sum_{i_2=1}^{n} R_{\mathbf{i}} + o(1)\right) \quad a.s. \tag{2.20}$$

On the other side by applying directly Theorem 6.1.1 in Lin *et al.* (1996), on $(R_{\mathbf{i}}, \mathbf{i} \in \mathbb{Z}^2)$ which are a strictly stationary $\alpha-$mixing random field with $\mathbb{E}(R_{\mathbf{i}}) = 0$, $\mathbb{E}\,|R_{\mathbf{i}}|^3 < \infty$, and according to the condition **(H4)(ii)**, we have :

$$\frac{1}{\sqrt{mn}}\sum_{i_1=1}^{m}\sum_{i_2=1}^{n} R_{\mathbf{i}} \xrightarrow{\mathcal{D}} N\left(0, \mathbf{C}\right). \tag{2.21}$$

We conclude the proof of the result (2.8) by using (2.20) and (2.21).

According to (2.20), we have

$$\widehat{\beta}^{(s)} - \beta^{(s)} = \frac{1}{mn}\sum_{i_1=1}^{m}\sum_{i_2=1}^{n} \mathbf{b}_s^T R_{\mathbf{i}} + o(1)\ a.s.,$$

where $\mathbf{b}_s^T = (b_s^1, \cdots, b_s^p)$, with $b_s^r = \left(\mathbf{B}^{-1}\right)_{sr}$.

Therefore, it suffices to show that,

$$\limsup_{(m,n)\longrightarrow\infty} \left|\sqrt{\frac{1}{2mn\log\log(mn)}}\left(\sum_{i_1=1}^{m}\sum_{i_2=1}^{n} \mathbf{b}_s^T R_{\mathbf{i}}\right)\right| \longrightarrow \sqrt{2}\sigma_{_s}.$$

Considering, $\Upsilon_{\mathbf{i}} = \mathbf{b}_s^T R_{\mathbf{i}}$, who satisfies all the conditions, we have

$$Var\left(\sum_{i_1=1}^{m}\sum_{i_2=1}^{n} \mathbf{b}_s^T R_{\mathbf{i}}\right) = \mathbf{b}_s^T Var\left(\sum_{i_1=1}^{m}\sum_{i_2=1}^{n} R_{\mathbf{i}}\right)\mathbf{b}_s = mnq_{ss}(1 + o(1)).$$

So

$$\limsup_{(m,n)\longrightarrow\infty} \left|\sqrt{\frac{1}{4d_{ss}mn\log\log(mn)}}\left(\sum_{i_1=1}^{m}\sum_{i_2=1}^{n} \mathbf{b}_s^T R_{\mathbf{i}}\right)\right| \longrightarrow 1 \quad a.s.$$

Thus, (2.9) holds.

Proof of Theorem 2.3.2

From (2.4), we can write

$$\widehat{g}_{n,m}(t) = \sum_{i_1=1}^{n}\sum_{i_2=1}^{m} w_{n,m}(z, Z_{\mathbf{i}}) \left(g\left(Z_{\mathbf{i}}\right) + \varepsilon_{\mathbf{i}}\right) - \sum_{i_1=1}^{m}\sum_{i_2=1}^{n} w_{n,m}(z, Z_{\mathbf{i}}) X_{\mathbf{i}}^T \left(\widehat{\beta} - \beta\right) .$$

So we have

$$\begin{aligned} \widehat{g}_{m,n}(z) - g(z) = & \left(\sum_{i_1=1}^{n}\sum_{i_2=1}^{m} w_{n,m}(z, Z_{\mathbf{i}}) \left(g\left(Z_{\mathbf{i}}\right) + \varepsilon_{\mathbf{i}}\right) - g(z)\right) - \left(\sum_{i_1=1}^{n}\sum_{i_2=1}^{m} w_{n,m}(z, Z_{\mathbf{i}}) X_{\mathbf{i}}^T \left(\widehat{\beta} - \beta\right)\right) \\ = & \; S_1 - S_2, \end{aligned}$$

and

$$|\widehat{g}_{m,n}(z) - g(z)| \leq \; |S_1| + |S_2| \; . \tag{2.22}$$

The Theorem 6 in Niang *et al.* (2011), give

$$|S_1| \to \; 0 \; a.s. \tag{2.23}$$

On the other hand, we have

$$|S_2| \leq \; \left\| \sum_{i_1=1}^{n}\sum_{i_2=1}^{m} w_{n,m}(z, Z_{\mathbf{i}}) X_{\mathbf{i}} \right\| \left\| \widehat{\beta} - \beta \right\| .$$

So

$$|S_2| \leq \; \left\| \sum_{i_1=1}^{n}\sum_{i_2=1}^{m} w_{n,m}(z, Z_{\mathbf{i}}) X_{\mathbf{i}} - \mathbb{E}\left(X_{\mathbf{i}} | Z_{\mathbf{i}} = z\right) \right\| \left\| \widehat{\beta} - \beta \right\| + \; \left\| \mathbb{E}\left(X_{\mathbf{i}} | Z_{\mathbf{i}} = z\right) \right\| \left\| \widehat{\beta} - \beta \right\| .$$

Under the Theorem 2.3.1, we have $\left\| \widehat{\beta} - \beta \right\| \to 0$ *a.s.*, and according to the conditions **(H4) and (H5)(iii)**, we have $\|\mathbb{E}(X_{\mathbf{i}} | Z_{\mathbf{i}} = z)\| < \infty$.

That implies

$$|S_2| \to \; 0 \;\; a.s. \tag{2.24}$$

From (2.22) -(2.24) the proof of (2.10) is holds.

For the second result (2.11) of the Theorem 2.3.2, we use the same decomposition of $\widehat{g}_{m,n}(t) - g(t)$, and by fixed z in $\mathcal{E}$, from the Theorem 11 in Niang *et al.* (2011), we can get

$$|S_1| = O\left(h_{(m,n)} + \sqrt{\frac{\log(mn)}{mnp^z_{h_{(n,m)}}}}\right) \ a.s. \tag{2.25}$$

So by using (2.22), (2.24) and (2.25), we have

$$|\widehat{g}_{m,n}(z) - g(z)| = O\left(h_{(m,n)} + \sqrt{\frac{\log(mn)}{mnp^z_{h_{(n,m)}}}}\right) \quad a.s.$$

Bibliographie

[1] Aneiros-Pérez G. and Vieu P. (2006). Semi-functional partial linear regression. Stat. Probab. Lett., 76, (11), 1102-1110.

[2] Aneiros-Pérez G. and Vieu P. (2008). Nonparametric time series prediction. A semi-functional partial linear modeling. Journal of Multivariate Analysis, 99, 834-857.

[3] Aneiros-Pérez, G. and Vieu, P. (2011). Automatic estimation procedure in partial linear model with functional data. Stat. Pap., 52, (4), 751-771.

[4] Attouch, M. K., Gheriballah, A., and Laksaci, A. (2011) Robust nonparametric estimation for functional spatial regression. In Ferraty, F., editor, Recent Advances in Functional Data Analysis and Related Topics, Contributions to Statistics, 27-31. Physica-Verlag HD.

[5] Bulinski, A., and Shashkin, A., (2006). Strong invariance principle for dependent random fields. Dynamics and Stochastics, 128-143.

[6] Carbon, M., Tran, L. T., and Wu, B., (1997). Kernel density estimation for random fields (density estimation for random fields). Stat Probab Lett, 36, (2), 115-125.

[7] Carbon, M., Francq, C., and Tran, L. T. (2007). Kernel regression estimation for random fields. Journal of Statistical Planning and Inference, 137 ; (3) ;778-798.

[8] Dabo-Niang, S., Yao, A.-F., Pischedda, L., Cuny, P., and Gilbert, F. (2010). Spatial mode estimation for functional random fields with application to bioturbation problem. Stochastic Environmental Research and Risk Assessment, 24, (4), 487-497.

[9] Dabo-Niang, S., Rachdi, M., and Yao, A.F., (2011). Kernel regression estimation for spatial functional random variables. Far East Journal of Theoretical Statistics, 37, (2), 77-113.

[10] Dabo-Niang, S., Kaid, Z., and Laksaci, A., (2012). On spatial conditional mode estimation for a functional regressor. Stat Probab Lett, 82, (7), 1413-1421.

[11] Dabo-Niang, S. and Yao, A.-F. (2013). Kernel spatial density estimation in infinite dimension space. Metrika, 76, (1), 19-52.

[12] Engle R., Granger C., Rice J. and Weiss A., (1986). Nonparametric estimates of the relation between weather and electricity sales, J. Amer. Statist. Assoc., 81, 310-320.

[13] Ferraty, F. and Vieu, P. (2006). Nonparametric Functional Data Analysis : Theory and Practice. Springer Series in Statistics. Springer.

[14] Gao, J. T. (1995a). Asymptotic theory for partially linear models. Commun. Statist. Theory Method, 22, 3327-3354.

[15] Gao, J. T., Lu, Z. and Tjϕstheim D., (2006). Estimation in semiparametric spatial regression. The Annals of Statistics, 34, (3), 1395-1435.

[16] Goia, A., Vieu, P., (2016). An introduction to recent advances in high/infinite dimensional statistics. J. Multivariate Anal., 146, 1-6.

[17] Härdle W, Liang H. and Gao J (2000). Partially linear models. Physica-Verlag, Heidelberg.

[18] Hsing, T. and Eubank, R.L., (2015). Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators. JohnWiley & Sons.

[19] Laksaci, A. and Mechab, B. (2010). Estimation non paramétrique de la fonction de hasard avec variable explicative fonctionnelle : cas des données spatiales. Revue Roumaine de Mathématiques Pures et Appliquées, 55, (1), 35-51.

[20] Lian H., (2011). Functional partial linear model. J. Nonparametr. Stat., 23, (1), 115-128.

[21] Lin, Z. and Lu, C. (1996). Limit Theory for Mixing Dependent Random Variables. Kluwer, Dordrecht.

[22] Ling, N., Aneiros-Pérez, G. and Vieu, P., (2020). knn estimation in functional partial linear modeling. Statistical Papers, 61, (1), 423-444.

[23] Ling, N. and Vieu P., (2018). Nonparametric modelling for functional data : selected survey and tracks for future. Statistics, 52, (4), 934-949.

[24] Ling, N. and Vieu P., (2020). On semiparametric regression in functional data analysis. WIRES Computational Statistics, 12, (6), 20-30.

[25] Ling N., Kan R., Vieu P., and Meng S., (2019). Semi-functional partially linear regression model with responses missing at random. Metrika, 82, 39-70.

[26] Mateu, J. and Romano, E., (2017). Advances in spatial functional statistics. Stoch Environ Res Risk Assess, 31, 1-6.

[27] Ramsay J. and Silverman B., (2005). Functional data analysis. Springer series in statistics. Springer, New York.

[28] Rice, J. (1986). Convergence rates for partially splined models. Statist. Probab. Lett., 4, 203-208.

[29] Robinson, P. M. (1988). Kernel smoothing in partial linear models. J. Roy. Statist. Soc. Ser. B, 50, 413-436.

[30] Shang, H., (2014). Bayesian bandwidth estimation for a semi-functional partial linear regression model with unknown error density. Comput. Stat., 29, 829-848.

[31] Ternynck, C., (2014). Spatial regression estimation for Functional data with spatial dependency. Journal de la Société Française de Statistique, Vol. 155, No. 2.

[32] Stock, C. J. (1989). Nonparametric policy analysis. J. Amer. Statist. Assoc., 89, 567-575.

[33] Tran, L. T., (1990). Kernel density estimation on random fields. Journal of Multivariate Analysis, 34, 1, 37-53.

Chapitre 3

Kernel estimator of the regression function for spatial data with responses missing at random

Dans ce chapitre, on s'intéresse à l'estimation par la méthode du noyau de la fonction de régression spatiale lorsque le régresseur fonctionnel est complètement observé et qu'une partie des réponses est manquante au hasard (MAR). On commence par donner la version spatiale de l'estimateur de la fonction de régression. Ensuite, Les propriétés asymptotiques de cet estimateur telles que le taux de convergence en probabilité est obtenue dans certaines conditions faibles. Enfin, une étude avec des données simulées est menée pour étudier les propriétés de l'échantillon fini de la méthode proposée.

Les résultats de ce chapitre font l'objet d'un article soumis pour expertise.

3.1 Introduction

Current research in many fields such as environmental sciences, meteorology, geography, economics, epidemiology, geophysics, image processing, and many others are often confronted with the analysis of large amounts of information with a spatial component (geographical position). Therefore, modeling such activities requires finding correlations between random variables in one place with others in neighboring locations ; this is an essential feature of spatial data analysis.

On the other hand, in recent decay, the new branch of statistics devoted to analyzing functional data (FDA) offered a new dynamic in theoretical and methodological developments and the diversification of application fields. This

has become possible thanks to the computer tool's progress in storage capacities, making it possible to record and analyze increasingly large data. Among the reference works on the subject, we can cite the monographs of Ramsay and Silverman (2005) for the applied aspects, Bosq (2000) for the theoretical aspects, Ferraty and Vieu (2006) for a nonparametric study. We can refer to the book of Hsing and Eubank (2015) and some bibliographic surveys such as Goia and Vieu (2016) and Ling and Vieu (2018) for the latest contributions in this field.

Functional data combined with spatial dependency offers the possibility to combining knowledge from spatial statistics and functional data analysis. and the spatial functional statistics extend this approach to process samples of functions recorded at different locations in a region (the so-called spatially correlated functional data). This combination has a promising future both on the applied and on the theoretical sides of statistics (see Mateu and Romano (2017) for a new contribution and a recent bibliography on the subject).

As an essential tool of the functional data analysis, the functional regression aims to model the relationship between the scalar response variable Y and the functional regressor X. The first results on estimating the regression function (in semi-metric space) were developed by Ferraty and Vieu (2000). Then, the theory and methods in this research field are well developed, see, for instance, Ferraty and Vieu (2002,2004), Ferraty *et al.* (2007, 2012), Dabo-Niang *et al.*(2009), the monograph by Ferraty and Vieu (2006, 2011) and the references therein. The regression function estimation in this spatial functional context was considered by Dabo-Niang *et al.* (2011). The authors establish the almost sure convergence rates of the spatial kernel estimator of the functional nonparametric regression. Simultaneously, the robust regression's asymptotic normality was obtained by Attouch *et al.* (2012).

All the work cited is devoted to full data analysis. However, this is not the case in many applications, including, for example, the analysis of survival data. This topic has been widely studied and widely covered in the multivariate case, especially on how to impute missing data and the precision of this imputation according to the types of missing data (see, for example, Little and Rubin, 2020 ; Trivellore, 2015, Graham, 2020). Commonly used imputation methods for missing responses include linear regression imputation, kernel regression imputation, etc. When the explanatory variables are finite-dimensional, much work on the regression function with missing data and its statistical inferences

can be found in the statistical literature We can cite the work of Ali and Abu-Salih (1988) and Siepman and Yang (1994) for parametric regression, Chen (1994), Chu and Cheng (1995) for nonparametric regression with a kernel. Nittner (2003) and Efromovich (2011 a, b) consider the case where some observations on the covariates are MAR, but the variable response observations are fully observed. While Boente *et al.* (2009) studied the robust regression model with data missing. Regression with the response variable and/or predictors (covariates) are MAR was considered by Efromovich (2014). For the latest contributions in this field, one can refer to the book by Little and Rubin (2020).When the explanatory variables are functional, very little literature has been reported to investigate the functional nonparametric regression model's statistical properties for the missing data. Ferraty et al. (2013) proposed the first time to estimate the mean of a scalar response based on an i.i.d. functional sample in which explanatory variables are observed for each subject, while response variables are MAR. They generalize the results of Cheng (1994) and they established the regression operator estimator's asymptotic properties when the functional regressor is completely observed and part of the responses is missing at random (MAR). Ling *et al.* (2015) considered the regression function's estimation and proved their asymptotic properties when the predictor variables are functional and stationary ergodic with MAR response. While, Rachdi *et al.* (2020) combined the k-nearest neighbor (k-NN) and local linear estimation methods to estimate the regression function, when the regressor variable is functional, and the response variable is a scalar but observed with a few random missing observations (MAR). Ling et al. (2020) established nonparametric quantile regression estimation for functional data with responses missing at random

To our knowledge, no work has been devoted to nonparametric regression for functional spatial data with response MAR.

For this, our objective is to study the estimation of the regression function by kernel approach when the observations are spatially dependant, and the variable response is MAR.

The organization of the paper is as follows. In Section 2, we describe some assumptions on the model **(1.1)** and construct precisely the estimator of $r(.)$ with MAR, while in Section 3, we present the main results of our work. In Section 4, we illustrate our methodology by a simulation study to compare the classical spatial nonparametric functional model incomplete and the model

with MAR. Finally, the proofs of the main results are postponed to Section 5.

3.2 The Model and the estimates

We investigate a nonparametric estimate of the conditional expectation of the real random variable with missing at random (MAR) $Y_{\mathbf{i}}$ given the functional random field $X_{\mathbf{i}}$ valued in a separable semi-metric space ($\mathcal{E}$,$d(.,.)$) (of eventually infinite dimension). The week consistencies of the estimate are shown.

Let $\mathbb{Z}^2$ be the integer lattice points in the two-dimensional Euclidean space. A point in bold $\mathbf{i} = (i_1, i_2)$ will be referred as a site. We now turn to estimation assuming that the data are available for $\mathbf{i}$ in the rectangular region $\mathcal{I}_{\mathbf{n}} = \{\mathbf{i} = (i_1, i_2) \in \mathbb{N}^2, 1 \leq i_1 \leq n_1, 1 \leq i_2 \leq n_2\}$, with a sample size of $\widehat{\mathbf{n}}$=$n_1 n_2$. We will write $\mathbf{n} \to \infty$, if $\min(n_1, n_2)$. The noparametric spatial modeling leads us to assume that :

$$Y_{\mathbf{i}} = r(X_{\mathbf{i}}) + \varepsilon_{\mathbf{i}},\ \mathbf{i} \in \mathbb{Z}^2.$$

where $r(.)$ is an unknown function over a semi-metric space $\mathcal{E}$, and the noise $\varepsilon_{\mathbf{i}}$ are centered identically distributed satisfying $\mathbb{E}\left(\varepsilon_{\mathbf{i}}|X_{\mathbf{i}}\right) = 0$, and $\sigma^2 = var(\varepsilon_{\mathbf{i}}) < \infty$.

It is clear that in the case of complete data, a well-known N-W kernel-type estimator of $r(.)$ is given by

$$\widetilde{r}_{\mathbf{n}}(x) = \frac{\sum\limits_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} Y_{\mathbf{i}} K\left(d(x, X_{\mathbf{i}}) b_{\mathbf{n}}^{-1}\right)}{\sum\limits_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} K\left(d(x, X_{\mathbf{i}}) b_{\mathbf{n}}^{-1}\right)},\ x \in \mathcal{E}, \tag{3.1}$$

where $b_{\mathbf{n}}$ being a sequence of bandwidths tending to zero as $\mathbf{n}$ tends to infinity, and the kernel K is a function from $\mathbb{R}^+$ to $\mathbb{R}^+$.

However, in missing mechanism with MAR for the response variable, and in the spatial case of dimention 2, we present the Bernoulli random variable δ, so an available incomplete sample of size $\widehat{\mathbf{n}}$ from (X, Y, δ) is $\{(X_{\mathbf{i}}, Y_{\mathbf{i}}, \delta_{\mathbf{i}}), \mathbf{i} \in \mathcal{I}_{\mathbf{n}}\}$, where $X_{\mathbf{i}}$ is observed completely, $\delta_{\mathbf{i}} = 1$ if $Y_{\mathbf{i}}$ is observed, and $\delta_{\mathbf{i}} = 0$ otherwise. Meanwhile the Bernoulli random variable δ is satisfied with $\mathbb{P}(\delta = 1/X = x, Y = y) = \mathbb{P}(\delta = 1/X = x) = \pi(x)$, where $\pi(x)$ is a function operator, which is called the conditional probability of the observing response given

the predictor and is often unknown. This mechanism shows that δ and Y are conditionally independent given X, we present the estimator of $r(.)$, which is given by :

$$\widetilde{r}_{\mathbf{n}}(x) = \frac{\sum\limits_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} \delta_{\mathbf{i}} Y_{\mathbf{i}} K\left(d(x, X_{\mathbf{i}}) h_{\mathbf{n}}^{-1}\right)}{\sum\limits_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} \delta_{\mathbf{i}} K\left(d(x, X_{\mathbf{i}}) h_{\mathbf{n}}^{-1}\right)} = \frac{\widetilde{f}_{\mathbf{n}}(x)}{\widetilde{g}_{\mathbf{n}}(x)} \tag{3.2}$$

whith

$$\widetilde{f}_{\mathbf{n}}(x) = \frac{1}{\widehat{\mathbf{n}}\mathbb{E}\left(K\left(d(x,X)h_{\mathbf{n}}^{-1}\right)\right)} \sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} \delta_{\mathbf{i}} Y_{\mathbf{i}} K\left(d(x, X_{\mathbf{i}}) h_{\mathbf{n}}^{-1}\right), \quad x \in \mathcal{E}$$

$$\widetilde{g}_{\mathbf{n}}(x) = \frac{1}{\widehat{\mathbf{n}}\mathbb{E}\left(K\left(d(x,X)h_{\mathbf{n}}^{-1}\right)\right)} \sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} \delta_{\mathbf{i}} K\left(d(x, X_{\mathbf{i}}) h_{\mathbf{n}}^{-1}\right), \quad x \in \mathcal{E}$$

and $h_{\mathbf{n}}$ being a sequence of bandwidths tending to zero as $\mathbf{n}$ tends to infinity, and the kernel K is a function from $\mathbb{R}^+$ to $\mathbb{R}^+$.

Note that $\widetilde{r}_{\mathbf{n}}(x)$ is constructed only using $\delta_{\mathbf{i}} = 1$ observations or the subjects with missing responses being discarded. In what follows, we will define the new estimators of $r(.)$ based on $\widetilde{r}_{\mathbf{n}}(.)$ by kernel regression imputing every missing $Y_{\mathbf{i}}$, we denote $Z_{\mathbf{i}} = Y_{\mathbf{i}}$ if $\delta_{\mathbf{i}} = 1$and $Z_{\mathbf{i}} = r(X_{\mathbf{i}})$ othewise. Under MAR assumption, we have

$$Z_{\mathbf{i}} = \delta_{\mathbf{i}} Y_{\mathbf{i}} + (1- \delta_{\mathbf{i}}) r(X_{\mathbf{i}}) = r(X_{\mathbf{i}}) + \delta_{\mathbf{i}}\, \varepsilon_{\mathbf{i}},$$

and the missing responses $Y_{\mathbf{i}}$ can be imputed by estimating $Z_{\mathbf{i}}$ which is defined as

$$\widetilde{Z}_{\mathbf{i}} = \delta_{\mathbf{i}} Y_{\mathbf{i}} + (1- \delta_{\mathbf{i}}) \widetilde{r}_{\mathbf{n}}(X_{\mathbf{i}})$$

Finally, based on $\left\{\left(X_{\mathbf{i}}, \widetilde{Z}_{\mathbf{i}}\right), \mathbf{i} \in \mathcal{I}_{\mathbf{n}}\right\}$, we obtain the estimator of $r(.)$ as follows :

$$\widehat{r}_{\mathbf{n}}(x) = \frac{\sum\limits_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} \widetilde{Z}_{\mathbf{i}} K\left(d(x,X_{\mathbf{i}}) b_{\mathbf{n}}^{-1}\right)}{\sum\limits_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} K\left(d(x,X_{\mathbf{i}}) b_{\mathbf{n}}^{-1}\right)} = \frac{\widehat{f}_{\mathbf{n}}(x)}{\widehat{g}_{\mathbf{n}}(x)},$$

whith

$$\widehat{f}_{\mathbf{n}}(x) = \frac{1}{\widehat{\mathbf{n}}\mathbb{E}\left(K\left(d(x,X)b_{\mathbf{n}}^{-1}\right)\right)} \sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} \widetilde{Z}_{\mathbf{i}} K\left(d(x, X_{\mathbf{i}}) b_{\mathbf{n}}^{-1}\right)$$

$$\widehat{g}_{\mathbf{n}}(x) = \frac{1}{\widehat{\mathbf{n}}\mathbb{E}\left(K\left(d(x,X)b_{\mathbf{n}}^{-1}\right)\right)} \sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} K\left(d(x, X_{\mathbf{i}}) b_{\mathbf{n}}^{-1}\right)$$

and $b_{\mathbf{n}}$being a sequence of bandwidths tending to zero as $\mathbf{n}$ tends to infinity

3.3 Some notations and assumptions

In fact, to take into account the spatial dependency, we assume that the process $\Gamma_{\mathbf{i}}$ is strictly stationary and satisfies a mixing condition defined in Carbon et al (1997). As follows : there exists a function $\varphi(t) \downarrow 0$ as $t \longrightarrow \infty$, such that for S, S' subsets of $\mathbb{Z}^2$ with finite cardinals,

$$\alpha(\mathfrak{B}(S), \mathfrak{B}(S')) = \sup_{\{A\in\mathfrak{B}(S), B\in\mathfrak{B}(S')\}} \{|P(AB) - P(A)P(B)|\}$$

$$\leq \psi(Card(S), Card(S'))\varphi(dist(S, S')), \tag{3.3}$$

where Card(S) (resp. Card(S')) the cardinality of S (resp. S'), dist(S, S') the Euclidean distance between S and S' , $\mathfrak{B}(S) = \mathfrak{B}(\Gamma_{\mathbf{i}}, \mathbf{i} \in S)$ and $\mathfrak{B}(S') = \mathfrak{B}(\Gamma_{\mathbf{i}}, \mathbf{i} \in S')$ are the σ-fields generated by the random variables $\Gamma_{\mathbf{i}}$, with $\mathbf{i}$ being elements of S and S' respectively and $\psi : \mathbb{N}^2 \longrightarrow \mathbb{R}^+$ is a nondecreasing symmetric positive function in each variable. We will be assumed that ψ satisfies either :

$$\forall a, b \in \mathbb{N},\ \psi(a, b) \leq C \min(a, b)\,, \tag{3.4}$$

or

$$\forall a, b \in \mathbb{N},\ \psi(a, b) \leq C\,(a + b + 1)^{\lambda}\,, \tag{3.5}$$

for some $C > 0$ and some $\lambda \geq 1$.

Concerning the function $\varphi(.)$, we will only study the case where $\varphi(t)$ tends to zero at a polynomial rate, $i.e.$

$$\varphi(t) \leq Ct^{-\theta}, \quad \text{for some} \quad \theta > 0. \tag{3.6}$$

3.3.1 Some notations

Let

$$\varsigma_{\mathbf{n}}(x) := \frac{\mathbb{E}\big(\widetilde{f}_{\mathbf{n}}(x)\big)}{\mathbb{E}(\widetilde{g}_{\mathbf{n}}(x))}$$

Now, define the bias term by

$$\begin{aligned} B_{\mathbf{n}}(x) &:= \varsigma_{\mathbf{n}}(x) - r(x) \\ &:= \frac{\mathbb{E}\big(\widetilde{f}_{\mathbf{n}}(x)\big) - r(x)\mathbb{E}(\widetilde{g}_{\mathbf{n}}(x))}{\mathbb{E}(\widetilde{g}_{\mathbf{n}}(x))} \end{aligned}$$

and the centered variate

$$Q_{\mathbf{n}}(x) := \left(\widetilde{f}_{\mathbf{n}}(x) - \mathbb{E}\left(\widetilde{f}_{\mathbf{n}}(x)\right)\right) - r(x)\left(\widetilde{g}_{\mathbf{n}}(x) - \mathbb{E}\left(\widetilde{g}_{\mathbf{n}}(x)\right)\right)$$

Then, it can be seen that :

$$\widetilde{r}_{\mathbf{n}}(x) - \varsigma_{\mathbf{n}}(x) = \frac{Q_{\mathbf{n}}(x) - B_{\mathbf{n}}(x)\left(\widetilde{g}_{\mathbf{n}}(x) - \mathbb{E}\left(\widetilde{g}_{\mathbf{n}}(x)\right)\right)}{\widetilde{g}_{\mathbf{n}}(x)} \tag{3.7}$$

3.3.2 Assumptions

We put the following assumptions that are necessary to show our main result.

(H1) Local dependence condition

We assume that for all $\mathbf{i} \neq \mathbf{j} \in \mathbb{Z}^2$ the joint probability distribution $\nu_{\mathbf{ij}}$ of $Z_{\mathbf{i}}$ and $Z_{\mathbf{j}}$ satisfies, for some constant $C > 0$ and for all $z_1, z_2 \in \mathcal{E}$:

$$\exists \varepsilon \in]0;1]\,,\ \nu_{\mathbf{ij}}\left(B(z_1, h_{\mathbf{n}}) \times B(z_2, h_{\mathbf{n}})\right) = (\phi_{h_{\mathbf{n}}}^{z_1}\phi_{h_{\mathbf{n}}}^{z_2})^{\frac{1+\varepsilon}{2}},$$

where $\phi_{h_{\mathbf{n}}}^{z} = P(Z \in B(z, h_{\mathbf{n}})) = \mu\left(B(z, h_{\mathbf{n}}),\ z \in \mathcal{E}\right.$, called small ball probability in the literature (see Ferraty and Vieu (2006)).

(H2) Conditions on the kernel :

We assume that the kernel $K : \mathbb{R}^+ \longrightarrow \mathbb{R}^+$ is a Lipschitz function of integral 1 and is such that : there exist two constants $0 < C_1 < C_2 < \infty$

$$C_1 I_{[0;1]} \leq K \leq C_2 I_{[0;1].}$$

(H3) Conditions on the conditional moment :

We suppose that :

i) for all $m \geq 2$, $\mathbb{E}[Y^m | X = x] < M_m(x) < \infty$, where M_m is a continuous function, for all x.

ii) For all $\mathbf{i} \neq \mathbf{j}$, $\mathbb{E}\left[Y_{\mathbf{i}} Y_{\mathbf{j}} | (X_{\mathbf{i}}, X_{\mathbf{j}})\right] < \infty$.

(H4) Local smoothness and continuous conditions :

i) $\exists \alpha > 0$ and a constant $C > 0$ such that $|r(u) - r(v)| \leq C d(u, v)^{\alpha}$, for all u, v in $\mathcal{E}$.

ii) We suppose that the functions $\pi(.)$ is continuous in a neighborhood of x, that is

$$\sup_{\{u : d(x,u) \leq h\}} |\pi(u) - \pi(x)| = \circ(1)\,, \quad as\, h \to 0\,,$$

iii) $\exists \varepsilon > 0$, such that $\mathbb{E}|Y_{\mathbf{1}}|^{2+\varepsilon} < \infty$.

Comments on the assumptions

The assumptions above are quite usual in nonfunctional partial linear models, so as it is usual in functional regression models, the conditions **(H1)** and **(H2)** as those proposed in Dabo-Niang *et al.* (2011), who are necessary to treat the functional nonparametric component of the model, and assuption **(H3)** is necessary to capture the convergence of the estimator, finally the assumption **(H4)** is necessary to have the rate of convergence of $Q_{\mathbf{n}}(x)$.

It is worth being noted that the results in our work extend the complete data in Dabo-Niang et al. (2011) to MAR case.

3.4 Main Results

Theorem 3.4.1. *Under hypotheses **(H1)-(H4)**, if in addition* $\widehat{\mathbf{n}}(\phi^x_{h_{\mathbf{n}}}/\log(\widehat{\mathbf{n}})) \longrightarrow \infty$*, as* $\mathbf{n}$ *goes to infinity, we have :*

$$\sqrt{\widehat{\mathbf{n}}(\phi^x_{h_{\mathbf{n}}}/\log(\widehat{\mathbf{n}}))}\,|\widetilde{r}_{\mathbf{n}}(x) - r(x) - B_{\mathbf{n}}| \to 0_p \ . \tag{3.8}$$

In addition, if $\widehat{\mathbf{n}}h^{2\alpha}_{\mathbf{n}}(\phi^x_{h_{\mathbf{n}}}/\log(\widehat{\mathbf{n}})) \longrightarrow 0$*, as* $\mathbf{n}$ *goes to infinity, we have :*

$$\sqrt{\widehat{\mathbf{n}}(\phi^x_{h_{\mathbf{n}}}/\log(\widehat{\mathbf{n}}))}\,|\widetilde{r}_{\mathbf{n}}(x) - r(x)| \to 0\,_p, \tag{3.9}$$

Corollary 3.4.2. *Under hypotheses theorem 3.4.1,*
and if $\widehat{\mathbf{n}}h^{2\alpha}_{\mathbf{n}}(\phi^x_{h_{\mathbf{n}}}/\log(\widehat{\mathbf{n}})) \longrightarrow 0$*, as* $\mathbf{n}$ *goes to infinity, we have :*

$$\sqrt{\widehat{\mathbf{n}}(\phi^x_{h_{\mathbf{n}}}/\log(\widehat{\mathbf{n}}))}\,|\widehat{r}_{\mathbf{n}}(x) - r(x)| \to 0\,_p, \tag{3.10}$$

3.5 Simulation

In this section, we first illustrate the finite sample behaviours of the proposed estimator $\widehat{r}$, we have done some simulations based on observations $(X_{\mathbf{i}}, Y_{\mathbf{i}}, \delta_{\mathbf{i}}) \in (\mathcal{E} \times \mathbb{R} \times \{0,1\})$ in this case we take $\mathbf{i} = (i_1, i_2)$ with $1 \leq i_1 \leq n_1$, $1 \leq i_2 \leq n_2$ and $\forall \mathbf{i} \in \mathbb{Z}^2$.

The model was generated as following :

$$X_{\mathbf{i}}(t) = \cos(2\pi A_{\mathbf{i}} t) + B_{\mathbf{i}} t, \quad t \in [0,1],$$

and

$$Y_i = r(X_{\mathbf{i}}) + \varepsilon_{\mathbf{i}}, \tag{3.11}$$

where $r(X) = 5.\frac{1}{\int_0^1 |X(t)|dt}$. Thereafter, we denote by $GRF\,(m, \sigma^2, s)$ a stationary Gaussian random field with mean m and covariance function defined by $C(l) = \sigma^2 \exp\left(-\left(\frac{\|l\|}{s}\right)^2\right), l \in \mathbb{R}^2$ and $s > 0$. Then, we have then simulated Model (3.11) with $A = D{*}\sin\left(\frac{G}{2} + .5\right), B = GRF(2.5, 5, 3), \varepsilon = GRF(0, .1, 5)$, $G = GRF(0, 5, 3)$ and $D_{\mathbf{i}} = \frac{1}{n_1 \times n_2} \sum_{\mathbf{j}} \exp\left(-\frac{\|\mathbf{i}-\mathbf{j}\|}{a}\right)$ $\left(D_{(\mathbf{i},\mathbf{j})} = \frac{1}{n_1 \times n_2} \sum_{1 \le j_1, j_2 \le 25} \exp\left(-\frac{\|(i_1,i_2)-(j_1,j_2)\|}{a}\right)\right)$. The function D is here to ensure and control the spatial mixing condition (even if using the Gaussian Random Fields also brings some spatial dependency).

For the missing mechanism, we adopted it as in Ferraty et al. (2013) :

$$p(x) = \mathbb{P}(\delta = 1 | X = x) = \text{expit}\left(2\alpha \int_0^1 x^2(t)\mathrm{d}t\right),$$

where $\text{expit}(u) = e^u/(1+e^u)$ for $\forall u \in \mathbb{R}$. The parameter α controls the degree of dependency between the functional curve X and the variable δ. To keep control the quantity $p(x)$, we compute $\overline{\delta} = 1 - \frac{1}{n_1 \times n_2} \sum_{i_1=1}^{n_1} \sum_{i_2=1}^{n_2} \delta_{(i_1,i_2)}$.

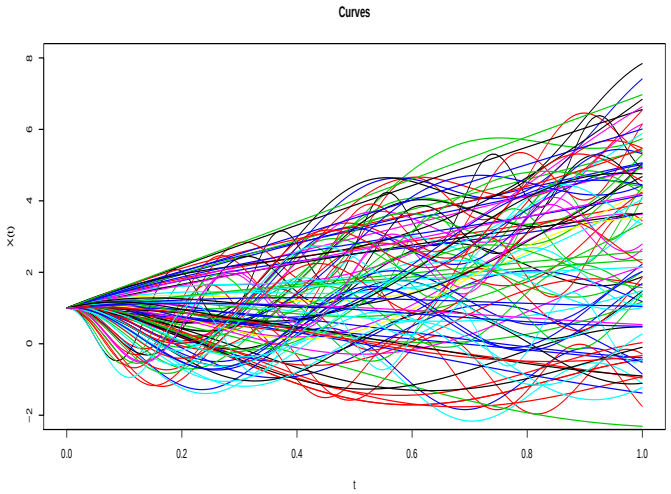

FIGURE 3.1 – The curves $X_{\mathbf{i}}, t \in [0, 1]$.

The used semi-metric is the first derivative of sample curves, given by

$$d\,(X_{\mathbf{i}}, X_{\mathbf{j}}) = \sqrt{\int_0^1 \left(X'_{\mathbf{i}}(t) - X'_{\mathbf{j}}(t)\right)^2 dt}, \qquad \text{for } \forall X_{\mathbf{i}}, X_{\mathbf{j}} \in \mathcal{E}.$$

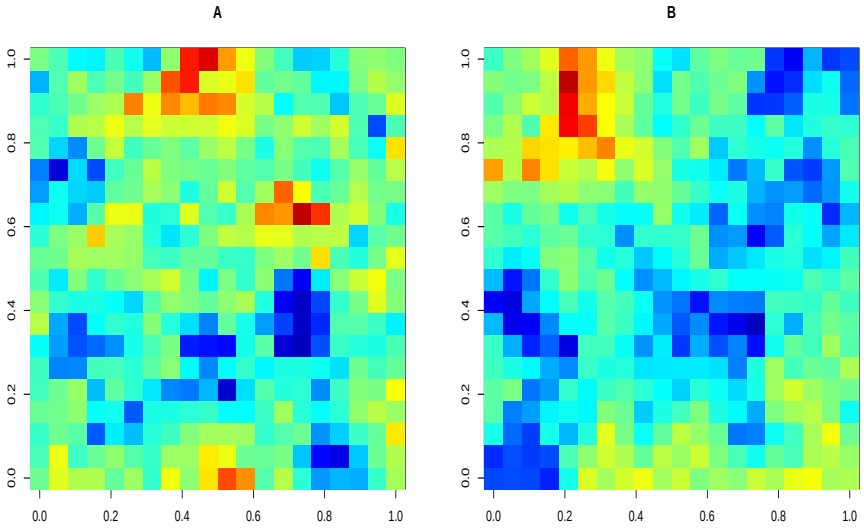

FIGURE 3.2 – Random field simulation.

In addition, we select the usual kernel function as follows : $K(u) = \frac{3}{2}(1-u^2)\,\mathbb{1}_{[0,1]}(u)$.

Since (in these conditions) model is based on Gaussian random fields with covariance function C and scale $s = 5$? observations of sites $\mathbf{i}$ and $\mathbf{j}$ with $\|\mathbf{i}-\mathbf{j}\| < 15$ are spatial dependent and nearly independent from $\|\mathbf{i}-\mathbf{j}\| \geq 15$. So, our observations are a mixture of i.i.d. and dependent observations (see Figure 3.1). Thus, to move away from independence, it suffices to lower the value of a (our results are based on $a = 0.5$).

Our main goal is to compare our estimator (MAR, $\widehat{r}_{\mathbf{n}}(x)$) with naive (MARV, $\widetilde{r}_{\mathbf{n}}(x)$) and the estimator in the complete data case (ECD, $\widetilde{r}_{n,m}(x)$) introduced by Dabo-Niang (2011).

In order to check the performance of the proposed estimator, we randomly split our data $(X_{\mathbf{i}}, Y_{\mathbf{i}}, \delta_{\mathbf{i}})_{\mathbf{i}}$ into two subsets : Learning sample $(X_{\mathbf{i}}, Y_{\mathbf{i}}, \delta_{\mathbf{i}})_{\mathbf{i}\in I}$ and test sample $(X_{\mathbf{i}}, Y_{\mathbf{i}}) \in I'$ (without missing data). The training sample was used to choose the smoothing parameters $h_{k_{opt}}$ the $k-$Nearest Neighbors $k-$NN cross-validation procedures :
The $h_{k_{opt}}$ is the bandwidth corresponding to the optimal number of neighbours obtained by a cross-validation procedure :

$$h_k = \min\left\{h \in \mathbb{R}^+ \text{such that } \sum_{\mathbf{i}\in I} \mathbb{1}_{B(x,h)}(Z_{\mathbf{i}}) = k\right\}$$

with $k_{opt} = \arg\min_k CV(k)$ where $CV(k) = \sum_{\mathbf{i}\in I}\left(Y_i - \widehat{r}_{\mathbf{n}}^{(-\mathbf{i})}(X_{\mathbf{i}})\right)^2$ with $\widehat{r}_{\mathbf{n}}^{(-\mathbf{i})}$ are the the leave one out of $\widehat{r}_{\mathbf{n}}$ (see Ferraty and Vieu (2006) for more details). On the one hand, the accurate of estimate, $\widehat{r}_{\mathbf{n}}(\cdot)$ of $r(\cdot)$ was measured through the Square and Mean Square Errors MSE :

$$\text{MSE} = \frac{1}{\#(I')}\sum_{\mathbf{i}\in I'}\left(\widehat{r}_{\mathbf{n}}(X_{\mathbf{i}}) - r(X_{\mathbf{i}})\right)^2.$$

where $\#(I')$ is the size of testing sample I'.
The results obtained for the three models are presented in Figure 3.3, where the predicted values are plotted versus the true values.

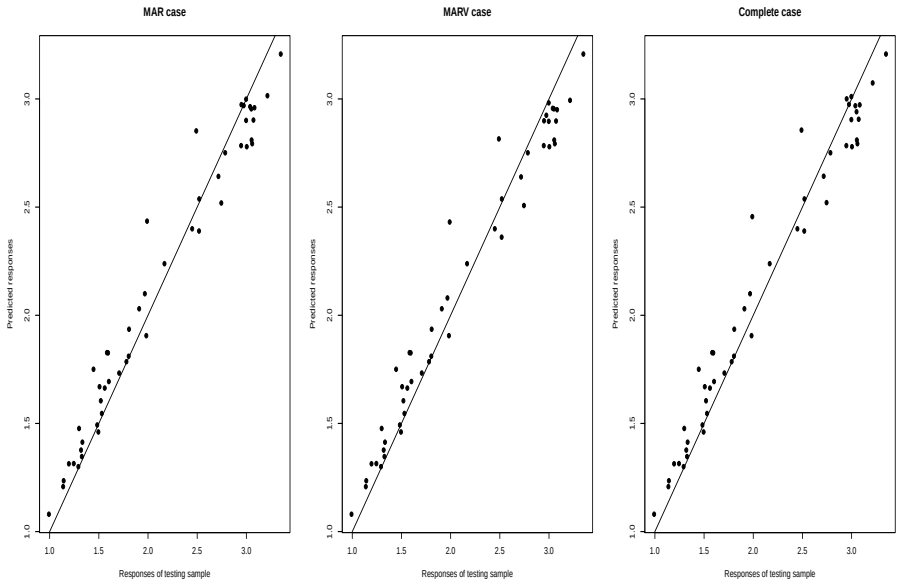

FIGURE 3.3 – Predictions of the 3 model .

Hence, the MSE under the MAR, MARV model and the complet model respectively. Then, we report it in Table 3.1 below.

TABLE 3.1 – Mean squared error (MSE) for MAR, MARV and Complet data.

n_1	n_2	α	$\overline{\delta}$	MSE (MAR)	MSE (MARV)	MSE (Complet)
10	10	0.5	0.1386	0.2234	0.2211	0.2097
		2.0	0.0393	0.2439	0.2433	0.2356
	20	0.5	0.1353	0.1463	0.1483	0.1285
		2.0	0.0280	0.1340	0.1361	0.1318
	30	0.5	0.1340	0.1267	0.1257	0.1058
		2.0	0.0217	0.1117	0.1128	0.1100
20	10	0.5	0.1431	0.1581	0.1552	0.1334
		2.0	0.0251	0.1656	0.1659	0.1608
	20	0.5	0.1344	0.1108	0.1106	0.0954
		2.0	0.0228	0.0956	0.0975	0.0937
	30	0.5	0.1302	0.0920	0.0894	0.0778
		2.0	0.0216	0.0794	0.0793	0.0781
30	10	0.5	0.1426	0.1346	0.1336	0.1233
		2.0	0.0262	0.1178	0.1187	0.1156
	20	0.5	0.1330	0.0824	0.0801	0.0720
		2.0	0.0253	0.0764	0.0767	0.0748
	30	0.5	0.1303	0.0552	0.0543	0.0467
		2.0	0.0215	0.0579	0.0584	0.0567

We can observe in the table 3.1 that the naive version offers a better MSE when the rate of missing data is small ; conversely, when the rate increases, the estimator MAR offers a better estimate. Note also that the MSE decreases significantly when the number of observations $\mathbf{n}$ grows. Such a numerical result is consistent with the theoretical ones of Theorem 3.4.1.

3.6 Proofs

Firstly, we state the following Lemmas essential to establish the Theorem.

Lemma 3.6.1. *Let $\widetilde{H}_{\mathbf{n}}(x) = \widetilde{f}_{\mathbf{n}}(x)$ or $\widetilde{g}_{\mathbf{n}}(x)$. If assumption **(H2)** hold, with 3.3 and the condition* (3.4) *or 3.5 is satisfied with*

$$\sum_{t=1}^{\infty} t(\varphi(t))^a < \infty, \text{ for some } 0 < a < 1/2 \ ,$$

then, we have :

$$\lim_{\mathbf{n}\to\infty} \left(\widehat{\mathbf{n}}\phi^x_{h_{\mathbf{n}}}\right) var(\widetilde{H}_{\mathbf{n}}(x)) < \infty, \ x \in \mathcal{E} \, .$$

The proof is the same as the lemma 4 in Dabo-Niang et al. (2011). Note that $\sum_{t=1}^{\infty} t(\varphi(t))^a < \infty$ hold if $\theta > 4$ (see lemma 4.2 in Carbon et *al*) (1997).

Lemma 3.6.2. *Under the assumptions* ($\boldsymbol{H_1}$)-($\boldsymbol{H_2}$) *and if* $\widehat{\mathbf{n}}\phi^x_{h_\mathbf{n}} \to \infty$ *as* $\mathbf{n} \to \infty$, *we have for any* $x \in \mathcal{E}$

$$\widetilde{g}_\mathbf{n}(x) \xrightarrow{p} \pi(x)\,. \tag{3.12}$$

Proof :
For the demonstration, we start by writing :

$$\begin{aligned} \widetilde{g}_\mathbf{n}(x) &= \frac{1}{\widehat{\mathbf{n}}\mathbb{E}\left(K\left(d(x,X)h_\mathbf{n}^{-1}\right)\right)} \sum_{i_1=1}^{n_1}\sum_{i_2=1}^{n_2} \delta_\mathbf{i} K\left(d(x,X_\mathbf{i})h_\mathbf{n}^{-1}\right) \\ &= \mathbb{E}\left(\widetilde{g}_\mathbf{n}(x)\right) + R_\mathbf{n}(x) \end{aligned}$$

where

$$\mathbb{E}\left(\widetilde{g}_\mathbf{n}(x)\right) = \frac{1}{\widehat{\mathbf{n}}\mathbb{E}\left(Kd(x,X)h_\mathbf{n}^{-1}\right)} \sum_{i_1=1}^{n_1}\sum_{i_2=1}^{n_2} \mathbb{E}\left[\delta_\mathbf{i} K\left(d(x,X_\mathbf{i})h_\mathbf{n}^{-1}\right)\right]$$

and

$$R_\mathbf{n}(x) = \frac{1}{\widehat{\mathbf{n}}\mathbb{E}\left(K\left(d(x,X)h_\mathbf{n}^{-1}\right)\right)} \sum_{i_1=1}^{n_1}\sum_{i_2=1}^{n_2} \delta_\mathbf{i} K\left(d(x,X_\mathbf{i})h_\mathbf{n}^{-1}\right) - \mathbb{E}\left[\delta_\mathbf{i} K\left(d(x,X_\mathbf{i})h_\mathbf{n}^{-1}\right)\right]$$

First, we need to establish

$$\mathbb{E}\left(\widetilde{g}_\mathbf{n}(x)\right) \xrightarrow{p} \pi(x), as\ \mathbf{n} \to \infty \tag{3.13}$$

By the properties of conditional expectation and the mechanism of MAR, combining the assumptions ($\mathbf{H_1}$), ($\mathbf{H_2}$) and the continuous property of $\pi(x)$, we have, for a given $\mathbf{k} \in \mathbb{Z}^2$:

$$\begin{aligned} \mathbb{E}\left(\widetilde{g}_\mathbf{n}(x)\right) &= \frac{1}{\widehat{\mathbf{n}}\mathbb{E}\left(K\left(d(x,X)h_\mathbf{n}^{-1}\right)\right)} \sum_{i_1=1}^{n_1}\sum_{i_2=1}^{n_2} \mathbb{E}\left[\delta_\mathbf{i} K\left(d(x,X_\mathbf{i})h_\mathbf{n}^{-1}\right)\right] \\ &= \frac{1}{\mathbb{E}\left(K\left(d(x,X)h_\mathbf{n}^{-1}\right)\right)} \mathbb{E}\left[\delta K\left(d(x,X)h_\mathbf{n}^{-1}\right)\right] \\ &= \frac{1}{\mathbb{E}\left(K\left(d(x,X)h_\mathbf{n}^{-1}\right)\right)} \mathbb{E}\left(\mathbb{E}(\delta_\mathbf{k} K d(x,X_\mathbf{k})h_\mathbf{n}^{-1} | X_\mathbf{k} = x)\right) \\ &= \mathbb{E}\left(\pi(x)\frac{K\left(d(x,X_\mathbf{k})h_\mathbf{n}^{-1}\right)}{\mathbb{E}\left(K\left(d(x,X)h_\mathbf{n}^{-1}\right)\right)}\right) \end{aligned}$$

converge to $\pi(x)$ as $\mathbf{n} \to \infty$

Second, we will prove that

$$R_{\mathbf{n}}(x) \xrightarrow{p} 0 as \mathbf{n} \to \infty \tag{3.14}$$

Let

$$Z_{\mathbf{n},\mathbf{i}} = \frac{\delta_{\mathbf{i}} K\left(d(x,X_{\mathbf{i}})h_{\mathbf{n}}^{-1}\right) - \mathbb{E}\left[\delta_{\mathbf{i}} K\left(d(x,X_{\mathbf{i}})h_{\mathbf{n}}^{-1}\right)\right]}{\widehat{\mathbf{n}}\mathbb{E}\left(K\left(d(x,X)h_{\mathbf{n}}^{-1}\right)\right)}$$

Then, we have

$$R_{\mathbf{n}}(x) = \sum_{i_1=1}^{n_1}\sum_{i_2=1}^{n_2} Z_{\mathbf{n},\mathbf{i}}.$$

Now, following the proof of theorem 5 in Dabo-Niang et al. (2011), according to lemma 3.6.1, we have

$$R_{\mathbf{n}}(x) \xrightarrow{p} 0, \text{ as } \mathbf{n} \to \infty.$$

hence the result of lemma 3.6.2

Lemma 3.6.3. *Under the assumptions (**H1**)-(**H4**) and if* $\widehat{\mathbf{n}}h_{\mathbf{n}}^{2\alpha}\phi_{h_{\mathbf{n}}}^{x} \to 0$ *as* $\mathbf{n} \to \infty$*, we have :*

$$B_{\mathbf{n}}(x) = O_p(h_{\boldsymbol{n}}^{\alpha}), \tag{3.15}$$

and

$$\sqrt{\widehat{\mathbf{n}}\phi_{h_{\mathbf{n}}}^{x}}B_{\mathbf{n}}(x)\left(\widetilde{g}_{\mathbf{n}}(x) - \mathbb{E}\left(\widetilde{g}_{\mathbf{n}}(x)\right)\right) \xrightarrow{p} 0, \ \ as \ \mathbf{n} \to \infty \tag{3.16}$$

Proof : We have

$$B_{\mathbf{n}}(x) = \frac{\mathbb{E}\left(\widetilde{f}_{\mathbf{n}}(x)\right) - r(x)\mathbb{E}(\widetilde{g}_{\mathbf{n}}(x))}{\mathbb{E}(\widetilde{g}_{\mathbf{n}}(x))}$$

Then by 3.13, we need to show that

$$\mathbb{E}\left(\widetilde{f}_{\mathbf{n}}(x)\right) - r(x)\mathbb{E}\left(\widetilde{g}_{\mathbf{n}}(x)\right) = o_{a.s}(h^{\alpha})$$

Similar to the proof of lemma 3.6.2, according to **(H2)** and **(H4i)**, it follows that :

$$\begin{aligned}
\mathbb{E}\left(\widetilde{f}_{\mathbf{n}}(x)\right) - r(x)\mathbb{E}\left(\widetilde{g}_{\mathbf{n}}(x)\right) &= \frac{1}{\mathbb{E}\left(K\left(d(x,X)h_{\mathbf{n}}^{-1}\right)\right)} \sum_{i_1=1}^{n_1}\sum_{i_2=1}^{n_2} \mathbb{E}\left[\left(Y_{\mathbf{i}} - r(x)\right)\delta_{\mathbf{i}} K\left(d(x,X_{\mathbf{i}})h_{\mathbf{n}}^{-1}\right)\right] \\
&= \frac{1}{\mathbb{E}\left(K\left(d(x,X)h_{\mathbf{n}}^{-1}\right)\right)} \sum_{i_1=1}^{n_1}\sum_{i_2=1}^{n_2} \mathbb{E}\{\mathbb{E}\left[\left(Y_{\mathbf{i}} - r(x)\right)\delta_{\mathbf{i}} K\left(d(x,X_{\mathbf{i}})h_{\mathbf{n}}^{-1}\right)|X_{\mathbf{i}}\right] \\
&= \frac{1}{\mathbb{E}\left(K\left(d(x,X)h_{\mathbf{n}}^{-1}\right)\right)} \sum_{i_1=1}^{n_1}\sum_{i_2=1}^{n_2} \mathbb{E}\left[\left(r(X_{\mathbf{i}}) - r(x)\right)\pi(X_{\mathbf{i}}) K\left(d(x,X_{\mathbf{i}})h_{\mathbf{n}}^{-1}\right)\right] \\
&\leq \sup_{u\in B(x,h)}|r(x) - r(u)| \left|\frac{1}{\mathbb{E}\left(K\left(d(x,X)h_{\mathbf{n}}^{-1}\right)\right)} \sum_{i_1=1}^{n_1}\sum_{i_2=1}^{n_2} \mathbb{E}\left[\pi(X_{\mathbf{i}}) K\left(d(x,X_{\mathbf{i}})h_{\mathbf{n}}^{-1}\right)\right]\right| \\
&= O_{a.s}(h_{\mathbf{n}}^{\alpha})
\end{aligned}$$

And from 3.13 the result follow.

Now, from 3.16, we observe that, $\sqrt{\widehat{\mathbf{n}}\phi_{h_{\mathbf{n}}}^{x}}B_{\mathbf{n}}(x)\left(\widetilde{g}_{\mathbf{n}}(x) - \mathbb{E}\widetilde{g}_{\mathbf{n}}(x)\right) = \sqrt{\widehat{\mathbf{n}}\phi_{h_{\mathbf{n}}}^{x}}B_{\mathbf{n}}(x)R_{\mathbf{n}}(x)$.

By 3.14, 3.15 and the condition $\widehat{\mathbf{n}}h_{\mathbf{n}}^{2\alpha}\phi_{h_{\mathbf{n}}}^{x} \to 0$ as $\mathbf{n}\to\infty$, the result is over.

Proof of Theorem 3.4.1

First, we present the proof of 3.8.

According to 3.12 and 3.16 , and the fact that

$$\widetilde{r}_{\mathbf{n}}(x) - \varsigma_{\mathbf{n}}(x) = \frac{Q_{\mathbf{n}}(x) - B_{\mathbf{n}}(x)\left(\widetilde{g}_{\mathbf{n}}(x) - \mathbb{E}\left(\widetilde{g}_{\mathbf{n}}(x)\right)\right)}{\widetilde{g}_{\mathbf{n}}(x)},$$

we have

$$\sqrt{\widehat{\mathbf{n}}\phi_{h_{\mathbf{n}}}^{x}}\frac{Q_{\mathbf{n}}(x) - B_{\mathbf{n}}(x)(\widetilde{g}_{\mathbf{n}}(x) - \mathbb{E}(\widetilde{g}_{\mathbf{n}}(x)))}{\widetilde{g}_{\mathbf{n}}(x)} = O_p(1),$$

That implies

$$\sqrt{\widehat{\mathbf{n}}(\phi_{h_{\mathbf{n}}}^{x}/\log(\widehat{\mathbf{n}}))}\,|\widetilde{r}_{\mathbf{n}}(x) - \varsigma_{\mathbf{n}}(x)| \to 0_p \ .$$

Secondly, concerning the proof of 3.9, we observe that

$$\widetilde{r}_{\mathbf{n}}(x) - r(x) = \widetilde{r}_{\mathbf{n}}(x) - \varsigma_{\mathbf{n}}(x) + B_{\mathbf{n}}(x)\,.$$

According to 3.8 and 3.15, and by the condition $\widehat{\mathbf{n}}h_{\mathbf{n}}^{2\alpha}(\phi_{h_{\mathbf{n}}}^{x}/\log(\widehat{\mathbf{n}})) \longrightarrow 0$, we have the result.

Proof of corollary 3.4.2

To demonstrate the corollary, it suffices to see that

$$\begin{aligned} |\widehat{r}_{\mathbf{n}}(x) - r(x)| &= \left| \frac{\sum\limits_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} \widetilde{Z_{\mathbf{i}}} K\left(d(x,X_{\mathbf{i}})b_{\mathbf{n}}^{-1}\right)}{\sum\limits_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} K\left(d(x,X_{\mathbf{i}})b_{\mathbf{n}}^{-1}\right)} - r(x) \right| \\ &= \left| \frac{\sum\limits_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} (\delta_{\mathbf{i}} Y_{\mathbf{i}} + (1-\delta_{\mathbf{i}})\widetilde{r}_{\mathbf{n}}(X_{\mathbf{i}})) K\left(d(x,X_{\mathbf{i}})b_{\mathbf{n}}^{-1}\right)}{\sum\limits_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} K\left(d(x,X_{\mathbf{i}})b_{\mathbf{n}}^{-1}\right)} - r(x) \right| . \end{aligned}$$

From where, we have :

$$\begin{aligned} \delta_{\mathbf{i}} Y_{\mathbf{i}} + (1- \delta_{\mathbf{i}})\widetilde{r}_{\mathbf{n}}(X_{\mathbf{i}}) &\leq \left| \frac{\sum\limits_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} Z_{\mathbf{i}} K\left(d(x,X_{\mathbf{i}})b_{\mathbf{n}}^{-1}\right)}{\sum\limits_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} K\left(d(x,X_{\mathbf{i}})b_{\mathbf{n}}^{-1}\right)} - r(x) \right| + \left| \frac{\sum\limits_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} (1-\delta_{\mathbf{i}})(\widetilde{r}_{\mathbf{n}}(X_{\mathbf{i}}) - r(X_{\mathbf{i}})) K\left(d(x,X_{\mathbf{i}})b_{\mathbf{n}}^{-1}\right)}{\sum\limits_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} K\left(d(x,X_{\mathbf{i}})b_{\mathbf{n}}^{-1}\right)} \right| \\ &= I_1 + I_2. \end{aligned}$$

Then, of the fact that $\mathbb{E}\left(Z_{\mathbf{i}}|X_{\mathbf{i}} = x\right) = r(x)$, so the result in Dabo *et al* (2011), given $\sqrt{\widehat{\mathbf{n}}(\phi^x_{h_{\mathbf{n}}}/\log(\widehat{\mathbf{n}}))}I_1$ converges in probability to 0.

Secondly, according to assumption **H2**, and the result of theorem above, we have $\sqrt{\widehat{\mathbf{n}}(\phi^x_{h_{\mathbf{n}}}/\log(\widehat{\mathbf{n}}))}I_2$ converges in probability to 0.

Bibliographie

[1] Ali, M. A., Abu-Salih, M. S. (1988). On estimation of missing observations in linear regression models, Sankhya. Indian J. Statist. 50 (B),404-411.

[2] Attouch, M., Chouaf, B. and Laksaci, A. (2012) Nonparametric M-estimation for functional spatial data , Communication of the Korean Statistical society, 19, 193-211.

[3] Boente, G., Gonzalez-Manteiga, W. and Perez-Gonzalez, A. (2009). Robust nonparametric estimation with missing data. J. Statist. Plann. Inference, 139, 571-592.

[4] Bosq, D. Linear Processes in Function Spaces : Theory and applications. Lecture Notes in Statistics, 149, Springer. (2000).

[5] Carbon, M., Tran, L. T., and Wu, B., (1997). Kernel density estimation for random fields (density estimation for random fields). Stat Probab Lett, 36, (2), 115-125.

[6] Cheng, P.E., (1994). Nonparametric estimation of mean functionals with data missing at random. J. Amer. Statist. Assoc. 89, 81-87.

[7] Chu, C. K., Cheng, P. E. (1995). Nonparametric regression estimation with missing data. J. Statist. Planning Inference. 48,85-99.

[8] Dabo-Niang, S., Rachdi, M., and Yao, A.-F. (2011). Kernel regression estimation for spatial functional random variables. Far East Journal of Theoretical Statistics, 37(2) :77–113.

[9] Dabo-Niang, S. and Rhomari, N. (2009). Kernel regression estimation in a Banach space. J. Statistical Planning and Inference, 139, 1421-1434.

[10] Efromovich, S., (2011b) Nonparametric Regression with Predictors Missing at Random, J. Amer. Statist. Assoc., 106 : 306-319.

[11] Efromovich, S., (2011a). Nonparametric regression with responses missing at random. J. Statist. Plann. Inference 141, 3744-3752.

[12] Efromovich, S., (2014). Nonparametric regression with missing data, Wiley Interdisciplinary Reviews Computational Statistics, 6 (4), 265-275.

[13] Ferraty, A. Laksaci, A. Tadj, P. Vieu.(2012). Estimation de la fonction de régression pour variable explicative et réponses fonctionnelles dépendante.C. R. Acad. Sci.Paris, Ser.I 350, 717-720.

[14] Ferraty, F., Mas, A. and Vieu, P. (2007). Nonparametric regression on functional data : inference and practical aspects. Aust. N. Z. J. Stat., 49 (3), 267-286

[15] Ferraty, F., Sued, F., Vieu, P., (2013). Mean estimation with data missing at random for functional covariables. Statistics 47 (4), 688-706.

[16] Ferraty, F. and Vieu, P. (2000) Dimension fractale et estimation de la régression dans des espaces vectoriels semi-normés. C.R. Math. Acad. Sci. Paris., 330 (2), 139-142.

[17] Ferraty, F. et Vieu, P. (2002) The functional nonparametric model and application to spectrometric data. Comput. Statist. 17 (4), 545-564.

[18] Ferraty, F., Vieu, P. (2004). Nonparametric models for functional data, with application in regression times series prediction and curves discrimination. J. Nonparametric Statist., 16, 111-127.

[19] Ferraty, F. and Vieu, P. (2006). Nonparametric functional data analysis. Theory and Practice. Springer Series in Statistics. New York.

[20] Ferraty, F., Vieu, P. (2011). Kernel regression estimation for functional data. In the Oxford Handbook of Functional Data Analysis (Ed. F. Ferraty and Y. Romain). Oxford University Press.

[21] Hsing, T. and Eubank, R.L., (2015). Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators. John Wiley and Sons.

[22] Goia, A., Vieu, P., (2016). An introduction to recent advances in high/infinite dimensional statistics. J. Multivariate Anal., 146, 1-6.

[23] Graham, J. W., (2020). Missing data analysis and design. NY : Springer, New York.

[24] Lin Z., Lu C. Limit Theory for Mixing Dependent Random Variables. Mathematics and Its Applications, 378, Springer.(1996).

[25] Ling ; N., Liang L., Vieu P (2015) Nonparametric regression estimation for functional stationary ergodic data with missing at random. J Stat Plan Inference 162 :75-87.

[26] Ling ; N., Liu Y, Vieu P (2016) Conditional mode estimation for functional stationary ergodic data with responses missing at random. Statistics 50,1-23.

[27] Ling ; N. and Vieu P., (2018). Nonparametric modelling for functional data. selected survey and tracks for future. Statistics, 52, (4), 934-949.

[28] Little, R. J. A. and Rubin, D. B. (2020) ; Statistical Analysis with Missing Data, 3rd Edition ; Wiley Series in Probability and Statistics

[29] Mateu, J. and Romano, E., (2017). Advances in spatial functional statistics. Stoch Environ Res Risk Assess, 31, 1-6.

[30] Nittner, T., (2003). Missing at random (MAR) in nonparametric regression, a simulation experiment. Stat. Methodol. Appl. 12, 195-210.

[31] Rachdi M.,, Laksaci A., Kaid Z., Benchiha A., Fahimah A. Al Awadhi, 2021. k-Nearest neighbors local linear regression for functional and missing data at random, Statistica Neerlandica, Netherlands Society for Statistics and Operations Research, 75(1), 42-65

[32] Ramsay, J. and Silverman, B. (2005) Functional Data Analysis (Second Edition) Spinger-Verlag, New York.

[33] Siepman, H. R., Yang, S.-S. (1994). Generalized least squares estimation of multivariate nonlinear models with missing data. Commun. Statist., 23(6), 1565-1579.

[34] Trivellore R. ; (2015). Missing Data Analysis in Practice. Chapman and Hall, Taylor and Francis Group.

Chapitre 4

SFPLR for spatial data with responses missing at random

Dans ce chapitre, on s'intéresse à l'estimation par la méthode du noyau de la fonction de régression semi-fonctionnelle partiellement linéaire pour les données spatiale lorsque le régresseur fonctionnel est complètement observé et qu'une partie des réponses est manquante au hasard (MAR). On généralise les résultats obtenus dans le premier chapitre, en montrant la convergence presque sûr et la normalité asymptotique de la version fonctionnelle.

4.1 Introduction

In recent decades, the problem of missing data has garnered a lot of attention within the statistical community. Responses may be missing for a number of common reasons. These include equipment malfunction, contamination of samples, manufacturing defects, drop out in clinical trials, weather conditions, incorrect data entry, etc.

In this paper, we consider missing responses in the context of functional regression analysis.

In other hand, in regression, the semi-parametric models were introduced in order to balance the low flexibility of linear regression and the high sensitivity to dimensional effects of nonparametric approaches. Their main interest is to take into account a priori, that there is a certain linear relationship in order to reduce the cost of the estimation that a nonparametric model would have, while keeping the effect underlying the nonparametric model for explain other relationships. These models have demonstrated their utility in many fields of applied sciences, such as economics, environmental studies, medicine, ect ...

(see Härdle it et al. (2000) for a large survey). In this context, partially linear estimation remains the most used method. These are models where the response variable Y is generally explained by the sum of an unknown linear combination of the components of a multivariate random vector X and an unknown transformation of another multivariate explanatory variable Z , Historically, and since the pioneering work of Engle et al. (1986), estimation techniques on partially linear models (PLR) has become very popular and several articles have been published for i.i.d. data (see, for example, Robinson (1988), Stock (1989), Linton (1995), or Liang (2000)) as well as for dependent data (see, for example, Gao (1995) or Aneiros-Pérez *et al.* (2004)).

In 2006, Aneiros and Vieu extended this model to the case where Z is valued in a semi-metric space possibly of infinite dimension with i.i.d. data. They call it semi-functional partially linear regression (SFPLR). The estimators of *beta* and $g(.)$ are obtained successively by using the least squares method and Nadaraya-Watson estimator.

The model is presented as

$$Y = X^T\beta + g(T) + \varepsilon, \tag{4.1}$$

where Y is a scalar response variable, $X = \left(X^{(1)}, \cdots, X^{(p)}\right)$ is random vector $\in \mathbb{R}^p$, $\beta = (\beta_1, \beta_2, \ldots, \beta_p)^T$ is a vector of unknown real parameters, $g(.)$ is an unknown smooth functional operator of another functional explanatory variable T with valued in some abstract infinite dimensional spaceH with associated semi-metric denoted by $d(.,.)$ and ε is an error such that $\mathbb{E}\left(\varepsilon|X,T\right) = 0$.

They proved the asymptotic normality of the estimator of β and obtained the convergence rate of the estimator of $g(.)$. The same authors in 2008 extended the model to the domain of time series and established the asymptotic properties of these estimators. At the same time, Lian (2011) extended the SFPLR model to the case where the linear component is also a functional characteristic. Benallou *et al.* (2021) we studied the semi-functional partial linear regression for spatial data with a both parametric and nonparametric modeling. In this case we obtain the asymptotic normality of the parametric component, and probability convergence with rate of the nonparametric component under spatial dependency.

The results above on SFPLR model (4.1) are concerned with the sample being observed completely. However, if the data are observed incompletely there is little work. When all the explanatory variables are of finite dimensio-

nality, we can cite Wang et al. (2004), Wang and Sun (2007). Nevertheless,when the explanatory variables X is in the case of infinite dimensionality or it is of functional feature, there are only limited papers devoted to investigate the functional regression model under the case of responses with MAR. Ling *et al.* (2018) being the first to study the estimation of the SFPLR from missing and identical independent data, he generalized the results in Wang et al. (2004).

The development of this type of method for spatial data is low.

In the multivariate case, we can quote , Haworth J. and Cheng T. (2012), where uses a kernel nonparametric regression model for spatiotemporal data to developed a predict future unit travel time values of road links in central London, UK, under the assumption of sensor malfunction. The model's performance is compared to another form of nonparametric regression, K-nearest neighbors, which is also useful for forecasting in case of missing data. Puranik *et al.* (2021) take the information from neighboring regions. The spatial autocorrelation inherent in the data presents a study that aims to estimate the missing aggregate-level spatial public health data. Zhang *et al.* (2014) propose two extensions of the regression based on the concurrent functional linear model to solve missing data problems in a series of co-located spatial images.

Semiparametric regression analysis with missing response at random.

Already exist, no work have been devoted the spatial SFPLR model with response MAR, so our cotrubution is to include the missing data in the spatial SFPLR model definite in chapter 1

4.2 Estimation and assumptions

4.2.1 Estimation

Let $\mathbb{Z}^2$ be the integer lattice points in the two-dimensional Euclidean space. A point in bold $\mathbf{i} = (i_1, i_2)$ will be referred as a site. In this chapter the spatial data can be seen as realizations of a measurable strictly stationary spatial process $\Lambda_{\mathbf{i}} = (Y_{\mathbf{i}}, X_{\mathbf{i}}, T_{\mathbf{i}})$ defined on a probability space $(\Omega, \mathfrak{A}, P)$ such that the $\Lambda_{\mathbf{i}}'$s have the same distribution as a variable $\Lambda = (Y; X; T)$; where Y is a real-valued and integrable variable, $X = \left(X^{(1)}, \cdots, X^{(p)}\right)$ a $\mathbb{R}^p-$valued random variable and T (a functional variable) valued in a separable semi-metric space $(\mathcal{E}; d(.;.))$. We now turn to estimation assuming that the data are available for $\mathbf{i}$ in the rectangular region $\mathcal{I}_{\mathbf{n}} = \{\mathbf{i} = (i_1, i_2) \in \mathbb{N}^2, 1 \leq i_1 \leq n_1, 1 \leq i_2 \leq n_2\}$,

with a sample size of $\widehat{\mathbf{n}}$=$n_1 n_2$. We will write $\mathbf{n} \to \infty$, if $\min(n_1, n_2)$.

the semi-functional partial linear spatial modeling leads us to assume that :

$$Y_{\mathbf{i}} = X_{\mathbf{i}}^T \beta + g(T_{\mathbf{i}}) + \varepsilon_{\mathbf{i}}, \;\; \mathbf{i} \in \mathbb{Z}^2,$$

where $\beta = (\beta_1, \dots \beta_p)^T$ is a vector of unknown parameters and $g(.)$ is an unknown function over a semi-metric space $\mathcal{E}$, and $\varepsilon_{\mathbf{i}}$ are centered identically distributed satisfying $\mathbb{E}\left(\varepsilon_{\mathbf{i}} | X^{(1)}, \cdots, X^{(p)}, T_{\mathbf{i}}\right) = 0$.

It is clear that in the case of complete data, according to chapter1 a well-known N-W kernel-type estimator of β is given by :

$$\widehat{\beta} = \left(\sum_{\mathbf{i} \in \mathcal{I}_{\mathbf{n}}} \widetilde{X}_{\mathbf{i}} (\widetilde{X}_{\mathbf{i}})^T \right)^{-1} \left(\sum_{\mathbf{i} \in \mathcal{I}_{\mathbf{n}}} \widetilde{Y}_{\mathbf{i}} \widetilde{X}_{\mathbf{i}} \right) \tag{4.2}$$

where

$$\widetilde{Y}_{\mathbf{i}} = Y_{\mathbf{i}} - \sum_{\mathbf{i} \in \mathcal{I}_{\mathbf{n}}} Y_{\mathbf{i}} w_{\mathbf{n}}(T_{\mathbf{i}}, T_{\mathbf{j}}) \text{ and } \widetilde{X}_{\mathbf{i}} = X_{\mathbf{i}} - \sum_{\mathbf{i} \in \mathcal{I}_{\mathbf{n}}} w_{\mathbf{n}}(T_{\mathbf{i}}, T_{\mathbf{j}}) X_{\mathbf{i}}$$

and

$$w_{\mathbf{n}}(T_{\mathbf{i}}, T_{\mathbf{j}}) = \frac{K\left(d(T_{\mathbf{i}}, T_{\mathbf{j}}) h_{\mathbf{n}}^{-1}\right)}{\sum_{\mathbf{i} \in \mathcal{I}_{\mathbf{n}}} K\left(d(T_{\mathbf{i}}, T_{\mathbf{j}}) h_{\mathbf{n}}^{-1}\right)} \tag{4.3}$$

with $h_{\mathbf{n}}$being a sequence of bandwidths tending to zero as $\mathbf{n}$ tends to infinity, and the kernel K is a function from $\mathbb{R}^+$ to $\mathbb{R}^+$.

and by insert $\widehat{\beta}$ the estimator of $g(.)$ is given by :

$$\widehat{g}_{\mathbf{n}}(t) = \sum_{\mathbf{i} \in \mathcal{I}_{\mathbf{n}}} w_{\mathbf{n}}(t, T_{\mathbf{i}}) \left(Y_{\mathbf{i}} - X_{\mathbf{i}}^T \widehat{\beta} \right) \tag{4.4}$$

So in missing mechanism with MAR for the response variable, an available incomplete sample of size $\widehat{\mathbf{n}}$ from $(Y; X; T, \delta)$ is $\{(Y_{\mathbf{i}}, X_{\mathbf{i}}, T_{\mathbf{i}}, \delta_{\mathbf{i}}), \mathbf{i} \in \mathcal{I}_{\mathbf{n}}\}$, where $X_{\mathbf{i}}$ and $T_{\mathbf{i}}$ are observed completely, $\delta_{\mathbf{i}} = 1$ if $Y_{\mathbf{i}}$ is observed, and $\delta_{\mathbf{i}} = 0$ otherwise. Meanwhile the Bernoulli random variable δ is satisfied with $\mathbb{P}(\delta = 1/X = x, T = t, Y = y) = \mathbb{P}(\delta = 1/X = x, T = t) = P(x, t)$, where $P(x, t)$ is a function operator, which is called the conditional probability of the observing

response given the predictor and is often unknown. This mechanism shows that δ and Y are conditionally independent given X and T.

Now, similar to Ling *and al.* (2018) when the simple are **i.i.d**, we obtein that

$$\widehat{\beta} = \left(\sum_{\mathbf{i} \in \mathcal{I}_\mathbf{n}} \delta_\mathbf{i} \widetilde{X}_\mathbf{i} (\widetilde{X}_\mathbf{i})^T \right)^{-1} \left(\sum_{\mathbf{i} \in \mathcal{I}_\mathbf{n}} \delta_\mathbf{i} \widetilde{Y}_\mathbf{i} \widetilde{X}_\mathbf{i} \right) \tag{4.5}$$

where

$$\widetilde{Y}_\mathbf{i} = Y_\mathbf{i} - \sum_{\mathbf{i} \in \mathcal{I}_\mathbf{n}} \delta_\mathbf{i} w_\mathbf{n}(T_\mathbf{i}, T_\mathbf{j}) Y_\mathbf{i} \text{ and } \widetilde{X}_\mathbf{i} = X_\mathbf{i} - \sum_{\mathbf{i} \in \mathcal{I}_\mathbf{n}} \delta_\mathbf{i} w_\mathbf{n}(T_\mathbf{i}, T_\mathbf{j}) X_\mathbf{i}$$

and

$$w_\mathbf{n}(T_\mathbf{i}, T_\mathbf{j}) = \frac{K\left(d(T_\mathbf{i}, T_\mathbf{j}) h_\mathbf{n}^{-1}\right)}{\sum\limits_{\mathbf{i} \in \mathcal{I}_\mathbf{n}} \delta_\mathbf{i} K\left(d(T_\mathbf{i}, T_\mathbf{j}) h_\mathbf{n}^{-1}\right)} \tag{4.6}$$

and $g(.)$ is estimated by

$$\widehat{g}_\mathbf{n}(t) = \sum_{\mathbf{i} \in \mathcal{I}_\mathbf{n}} \delta_\mathbf{i} w_\mathbf{n}(t, T_\mathbf{i}) \left(Y_\mathbf{i} - X_\mathbf{i}^T \widehat{\beta} \right) \tag{4.7}$$

4.2.2 Some notations and assumptions

The spatial dependency, we assume that the process $\Lambda_\mathbf{i}$ is strictly stationary and satisfies a mixing condition defined in Carbon et al (1997). As follows :

There exists a function $\varphi(t) \downarrow 0$ as $t \longrightarrow \infty$, such that for S, S' subsets of $\mathbb{Z}^2$ with finite cardinals,

$$\alpha(\mathfrak{B}(S), \mathfrak{B}(S')) = \sup_{\{A \in \mathfrak{B}(S), B \in \mathfrak{B}(S')\}} \{|P(AB) - P(A)P(B)|\}$$

$$\leq \psi(Card(S), Card(S'))\varphi(dist(S, S')), \tag{4.8}$$

where Card(S) (resp. Card(S')) the cardinality of S (resp. S'), dist(S, S') the Euclidean distance between S and S' , $\mathfrak{B}(S) = \mathfrak{B}(\Lambda_{\mathbf{i}}, \mathbf{i} \in S)$ and $\mathfrak{B}(S') = \mathfrak{B}(\Lambda_{\mathbf{i}}, \mathbf{i} \in S')$ are the σ-fields generated by the random variables $\Lambda_{\mathbf{i}}$, with $\mathbf{i}$ being elements of S and S' respectively and $\psi : \mathbb{N}^2 \longrightarrow \mathbb{R}^+$ is a nondecreasing symmetric positive function in each variable. We will be assumed that ψ satisfies either :

$$\forall a, b \in \mathbb{N},\ \psi(a, b) \leq C \min(a, b)\,, \tag{4.9}$$

or

$$\forall a, b \in \mathbb{N},\ \psi(a, b) \leq C\,(a + b + 1)^{\lambda}\,, \tag{4.10}$$

for some $C > 0$ and some $\lambda \geq 1$.

Concerning the function $\varphi(.)$, we will only study the case where $\varphi(t)$ tends to zero at a polynomial rate, *i.e.*

$$\varphi(t) \leq Ct^{-\theta}, \quad \text{for some} \quad \theta > 0. \tag{4.11}$$

For a fixed $t \in \mathcal{E}$, we denote by $\mathfrak{B}(t, h) = \{t' \in \mathcal{E};\ d(t, t'\} < h\}$, the ball with center t and radius h.

Now we put the following assumptions that are necessary to show our main result.

(H1) Local dependence condition

We assume that for all $\mathbf{i} \neq \mathbf{j} \in \mathbb{Z}^2$ the joint probability distribution $\nu_{\mathbf{ij}}$ of $T_{\mathbf{i}}$ and $T_{\mathbf{j}}$ satisfies for some constant $C > 0$ and for all $x, y \in \mathcal{E}$

$$\exists \varepsilon_1 \in\,]0; 1]\,,\ \nu_{\mathbf{ij}}\left(B(x, h_{\mathbf{n}}) \times B(y, h_{\mathbf{n}})\right) = (\phi^x_{h_{\mathbf{n}}} \phi^y_{h_{\mathbf{n}}})^{\frac{1+\varepsilon_1}{2}},$$

where $\phi^x_{h_{\mathbf{n}}} = P(T \in B(x, h_{\mathbf{n}})) = \mu\left(B(x, h_{\mathbf{n}})\right.$ called small ball probability in the literature (see Ferraty and Vieu (2006)).

(H2) Conditions on the kernel :

We assume that the kernel $K : \mathbb{R}^+ \longrightarrow \mathbb{R}^+$ is a Lipschitz function of integral 1 and is such that : there exist two constants $0 < C_1 < C_2 < \infty$

$$C_1 I_{[0;1]} \leq K \leq C_2 I_{[0;1]}$$

(H3) : Let us introduce the following notation : $f^{(s)}(t) = E\left[X^{(s)}_{\mathbf{i}} | T_{\mathbf{i}} = t\right]$, $s = 1, \cdots, p$ and $f^{(0)}(t) = E\left[Y_{\mathbf{i}} | T_{\mathbf{i}} = t\right]$.

We assume that $f^{(s)}(.)$ for all $s = 0, 1, \cdots, p$ and $g(.)$ are smooth in the sense that for somme $c > 0$ and $\alpha > 0$, we have :

$$|\xi(u) - \xi(v)| \leq cd(u,v)^{\alpha},$$

$\forall u, v \in \mathcal{E}$ and $\forall \xi \in \left\{g, f^{(1)}, \cdots, f^{(p)}\right\}$.

(H4) :

i) We have denoted $\theta_{\mathbf{i}} = \left(\theta_{\mathbf{i}}^{(1)}, \cdots, \theta_{\mathbf{i}}^{(p)}\right)^T$, with $\theta_{\mathbf{i}}^{(s)} = X_{\mathbf{i}}^{(s)} - \mathbb{E}[\delta_{\mathbf{i}} X_{\mathbf{i}}^{(s)} | T_{\mathbf{i}}]$, $s = 1, \cdots, p$. and $\mathbf{\Sigma} = \mathbb{E}[P(X_{\mathbf{1}}, T_{\mathbf{1}}) \left(\theta_{\mathbf{1}} (\theta_{\mathbf{1}})^T\right)]$, where $\mathbf{1}$ is the site spatial $(1,1)$. Assume that the inverse matrix of $\mathbf{\Sigma}$ exist.

ii) Let $R_{\mathbf{i}} = \theta_{\mathbf{i}} \varepsilon_{\mathbf{i}}$. While $\varepsilon_{\mathbf{i}}$ is independent of $\theta_{\mathbf{i}}$, assume that the matrix $C = \sum_{i_1=-\infty}^{\infty} \sum_{i_2=-\infty}^{\infty} \mathbb{E}[P(X_{\mathbf{i}}, T_{\mathbf{i}}) \left(R_{\mathbf{0}} (R_{\mathbf{i}})^T\right)]$ is positive definite.

(H5) : We suppose that :

i) $\mathbb{E}\left|\varepsilon_{\mathbf{1}}\right|^{\rho} + \mathbb{E}\left|\theta_{\mathbf{1}}^{(1)}\right|^{\rho} + \cdots + \mathbb{E}\left|\theta_{\mathbf{1}}^{(p)}\right|^{\rho} < \infty$ for some $\rho \geq 3$.

ii) For all $\mathbf{i} \neq \mathbf{j}$, $\mathbb{E}\left[Y_{\mathbf{i}} Y_{\mathbf{j}}\right| (T_{\mathbf{i}}, T_{\mathbf{j}})] < \infty$

iii) For all $\mathbf{i} \neq \mathbf{j}$, $\max\limits_{1 \leq s \leq p} \mathbb{E}\left[X_{\mathbf{i}}^{(s)} X_{\mathbf{j}}^{(s)} | (T_{\mathbf{i}}, T_{\mathbf{j}})\right] < \infty$.

Comments on the assumptions

The assumptions above in general is already used in chapter1 with a little defference in the condition **(H4).**

4.3 Main Results

In this Section, we present the asymptotic normality of the estimator $\widehat{\beta}$ and the rate of convergence of $\widehat{g}_{\mathbf{n}}(t)$ defined in 4.5 and 4.6, respectively.

Theorem 4.3.1. *Under hypotheses **(H1)-(H5)**, as* $\mathbf{n}$ *goes to infinity, we have :*

$$\sqrt{\widehat{\mathbf{n}}} \left(\widehat{\beta} - \beta\right) \xrightarrow{D} N\left(0, \mathbf{\Sigma}^{-1} \mathbf{C} \left(\mathbf{\Sigma}^{-1}\right)^T\right) \tag{4.12}$$

Theorem 4.3.2. *Under hypotheses of 4.3.1, if in addition* $\mathbf{n}\phi_{h_{\mathbf{n}}}^{x} / \log(\widehat{\mathbf{n}}) \longrightarrow \infty$, *and the mixing satisfies :*

– *condition (4.9) with* $\theta > 4$ *in (4.11)* ,

or

– *condition (4.10) with* $\left(\mathbf{n}\left(\phi_{h_{\mathbf{n}}}^{x} / \log(\widehat{\mathbf{n}})\right)^{\frac{2-\theta}{2(1+2\lambda)-\theta}}\right) \to \infty$, $\theta > 2(1+2\lambda)$ *in* (4.11). *We have :*

$$|\widehat{g}_{\mathbf{n}}(t) - g(t)| \xrightarrow{p} 0. \tag{4.13}$$

4.4 Proof

Throughout this section, C denotes a generic positive constant which may take different values from one formula to another. We first need to state some preliminary results.

Lemma 4.4.1. *(**Carbon et al. (2007)**) Let the sets $E_1, \cdots, E_r$ containing each m sites and such that, for all $i \neq j$ and for $1 \leq i, j \leq r$, $dist(E_i, E_j) \geq \vartheta_0$. Let $V_1, \cdots V_r$ a sequence of random variables with real values and measurable respectively with respect to $\mathfrak{B}(E_1), \cdots, \mathfrak{B}(E_r)$. Let be W_l with values in $[a; b]$. There exists a sequence of independent random variables $V_1^*, \cdots V_r^*$ such that W_l^* has the same distribution as W_l and satisfies :*

$$\sum_{l=1}^{r} \mathbb{E}\left|W_l - W_l^*\right| \leq 2r(b-a)\psi((r-1)m, m)\varphi(\vartheta_0)$$

Lemma 4.4.2. *(**Carbon et al. (1997)**) Denote by $\mathcal{L}_r(\mathcal{F})$ the class of $\mathcal{F}$-measurable random variables X which satisfy : $\|X\|_r = \left(E\left|X\right|^r\right)^{1/r} < \infty$. Suppose that $X \in \mathcal{L}_r\left(\mathfrak{B}(E)\right), Y \in \mathcal{L}_s\left(\mathfrak{B}(E')\right), 1 < r, s, t < \infty$ and $\frac{1}{r} + \frac{1}{s} + \frac{1}{t} = 1$ then,*

$$\left|\mathbb{E}\left[XY\right] - \mathbb{E}\left[X\right]\mathbb{E}\left[Y\right]\right| \leq$$
$$C\left\|X\right\|_r \left\|Y\right\|_s \left\{\psi(Card(S), Card(S'))\varphi(dist(S, S'))\right\}^{1/t}.$$

For bounded random variables with probability 1, we have : $|\mathbb{E}XY - \mathbb{E}X\mathbb{E}Y| \leq C\{\psi(\mathrm{Card}(S), \mathrm{Card}(S'))\varphi(\mathrm{dist}(S, S'))\}$.

Lemma 4.4.3. *Let us introduce the following notations :*

$$M_{\mathbf{n}}(t) = \frac{1}{\widehat{\mathbf{n}}E\left(K\left(d(t,T)h_{\mathbf{n}}^{-1}\right)\right)} \sum_{i_1=1}^{n_1}\sum_{i_2=1}^{n_2} K\left(d(t, T_{\mathbf{i}})h_{\mathbf{n}}^{-1}\right),$$

$$N_{\mathbf{n}}(t) = \frac{1}{\widehat{\mathbf{n}}E\left(K\left(d(t,T)h_{\mathbf{n}}^{-1}\right)\right)} \sum_{i_1=1}^{n_1}\sum_{i_2=1}^{n_2} \delta_{\mathbf{i}} Y_{\mathbf{i}} K\left(d(t, T_{\mathbf{i}})h_{\mathbf{n}}^{-1}\right),$$

$$L_{\mathbf{n}}^{(s)}(t) = \frac{1}{\widehat{\mathbf{n}}E\left(K\left(d(t,T)h_{\mathbf{n}}^{-1}\right)\right)} \sum_{i_1=1}^{n_1}\sum_{i_2=1}^{n_2} \delta_{\mathbf{i}} X_{\mathbf{i}}^{(s)} K\left(d(t, T_{\mathbf{i}})h_{\mathbf{n}}^{-1}\right) \text{ for } s = 1, \cdots, p.$$

*Under assumptions **(H1)-(H4)** and **(H6)**, if in addition the mixing satisfies (4.9) or (4.10) with* $\sum_{i=1}^{\infty} i\,(\varphi(i))^a < \infty$, *for some* $0 < a < 1/2$.

Then $\forall G_{\mathbf{n}} \in \left\{M_{\mathbf{n}}, N_{\mathbf{n}}, L_{\mathbf{n}}^{(1)}, \cdots, L_{\mathbf{n})}^{(p)}\right\}$, *we have*

$$\lim_{\mathbf{n}\to\infty} \left(\widehat{\mathbf{n}}\phi_{h_{(n,m)}}^{t}\right) \mathbf{var}\left(G_{\mathbf{n}}(t)\right) < \infty, \quad t \in \mathcal{E}.$$

Proof :

The proof is the same as in (Dabo & al.2011) for the case of $N = 2$.

Lemma 4.4.4. *Under assumptions **(H1)-(H4)** and **(H5)(iii)** (g not included in **(H4)**) and by using the result of Lemme3 with* $G_{(n,m)} \neq \zeta_{(n,m)}$, *as* $\mathbf{n}$ *goes to infinity, we have :*

$$\frac{1}{\widehat{\mathbf{n}}}\sum_{i_1=1}^{n_1}\sum_{i_2=1}^{n_2}\widetilde{X}_{\mathbf{i}}(\widetilde{X}_{\mathbf{i}})^T \longrightarrow \mathbf{\Sigma} \quad \textit{in probability.} \tag{4.14}$$

Proof :

The (r,s)th element of $\frac{1}{\widehat{\mathbf{n}}}\sum_{i_1=1}^{n_1}\sum_{i_2=1}^{n_2}\widetilde{X}_{\mathbf{i}}(\widetilde{X}_{\mathbf{i}})^T$ can be written as

$$\left(\frac{1}{\widehat{\mathbf{n}}}\sum_{i_1=1}^{n_1}\sum_{i_2=1}^{n_2}\widetilde{X}_{\mathbf{i}}(\widetilde{X}_{\mathbf{i}})^T\right)_{rs} = \frac{1}{\widehat{\mathbf{n}}}\sum_{i_1=1}^{n_1}\sum_{i_2=1}^{n_2}\theta_{\mathbf{i}}^{(r)}\theta_{\mathbf{i}}^{(s)} + \frac{1}{\widehat{\mathbf{n}}}\sum_{i_1=1}^{n_1}\sum_{i_2=1}^{n_2}\Delta_{\mathbf{i}}^{(r)}\theta_{\mathbf{i}}^{(s)}$$

$$+ \frac{1}{\widehat{\mathbf{n}}}\sum_{i_1=1}^{n_1}\sum_{i_2=1}^{n_2}\Delta_{\mathbf{i}}^{(s)}\theta_{\mathbf{i}}^{(r)} + \frac{1}{\widehat{\mathbf{n}}}\sum_{i_1=1}^{n_1}\sum_{i_2=1}^{n_2}\Delta_{\mathbf{i}}^{(r)}\Delta_{\mathbf{i}}^{(s)}, \tag{4.15}$$

with

$$\Delta_{\mathbf{i}} = (\Delta_{\mathbf{i}}^{(1)}, \cdots, \Delta_{\mathbf{i}}^{(p)})^T = \mathbb{E}[\delta_{\mathbf{i}}X_{\mathbf{i}}|T_{\mathbf{i}}] - \sum_{i_1=1}^{n_1}\sum_{i_2=1}^{n_2}\delta_{\mathbf{i}}w_{\mathbf{n}}(T_{\mathbf{i}}, T_{\mathbf{j}})X_{\mathbf{i}}$$

and

$$\Delta_{\mathbf{i}}^{(s)} = \mathbb{E}[\delta_{\mathbf{i}}X_{\mathbf{i}}^{(s)}|T_{\mathbf{i}}] - \sum_{i_1=1}^{n_1}\sum_{i_2=1}^{n_2}\delta_{\mathbf{i}}w_{\mathbf{n}}(T_{\mathbf{i}}, T_{\mathbf{j}})X_{\mathbf{i}}^{(s)} \text{ for } s = 1, \cdots, p.$$

So

$$\Delta_{\mathbf{i}}^{(s)} = f^{(s)}(T_{\mathbf{i}}) - \widehat{f^{(s)}}(T_{\mathbf{i}}),$$

where $\widehat{f^{(s)}}$ is the kernel estimator of $f^{(s)}$ defined in **(H4)**, and by the result of chapter 2, we can see that :

$$\max_{1\leq k\leq p}\max_{\mathbf{i}\in\mathcal{I}_{n,m}}\left|\Delta_{\mathbf{i}}^{(k)}\right| \xrightarrow{p} 0\,. \tag{4.16}$$

Now since

$$\frac{1}{\widehat{\mathbf{n}}}\sum_{i_1=1}^{n_1}\sum_{i_2=1}^{n_2}\theta_{\mathbf{i}}^{(r)}\theta_{\mathbf{i}}^{(s)} \longrightarrow \mathbb{E}\left[P(X_{\mathbf{1}},T_{\mathbf{1}})\theta_{\mathbf{1}}^{(r)}\theta_{\mathbf{1}}^{(s)}\right] = \mathbf{\Sigma}_{rs} \textit{ in probability}, \tag{4.17}$$

we conclude this proof by using (4.15)-(4.17), and the Cauchy-Schwarz inequality.

Proof of Theorem 4.3.1

The proof of Theorem 4.3.1 is based on the following decomposition and the lemmas above :

$$\widehat{\beta}-\beta = \left(\tfrac{1}{\widehat{\mathbf{n}}}\sum_{i_1=1}^{n_1}\sum_{i_2=1}^{n_2}\widetilde{X}_{\mathbf{i}}(\widetilde{X}_{\mathbf{i}})^T\right)^{-1}\left(\tfrac{1}{\widehat{\mathbf{n}}}\sum_{i_1=1}^{n_1}\sum_{i_2=1}^{n_2}\widetilde{X}_{\mathbf{i}}\left(\widetilde{Y}_{\mathbf{i}}-\left(\widetilde{X}_{\mathbf{i}}\right)^T\beta\right)\right)$$

We can write

$$\begin{aligned}\sqrt{\widehat{\mathbf{n}}}\left(\widehat{\beta}-\beta\right) &= \left(\tfrac{1}{\widehat{\mathbf{n}}}\sum_{i_1=1}^{n_1}\sum_{i_2=1}^{n_2}\widetilde{X}_{\mathbf{i}}(\widetilde{X}_{\mathbf{i}})^T\right)^{-1}\left(\tfrac{1}{\sqrt{\widehat{\mathbf{n}}}}\sum_{i_1=1}^{n_1}\sum_{i_2=1}^{n_2}R_{\mathbf{i}}+\tfrac{1}{\sqrt{\widehat{\mathbf{n}}}}\sum_{i_1=1}^{n_1}\sum_{i_2=1}^{n_2}\gamma_{\mathbf{i}}\left(\Delta_{\mathbf{i}}^{(0)}-\Delta_{\mathbf{i}}^{T}\beta\right)\right)\\ &+\left(\tfrac{1}{\widehat{\mathbf{n}}}\sum_{i_1=1}^{n_1}\sum_{i_2=1}^{n_2}\widetilde{X}_{\mathbf{i}}(\widetilde{X}_{\mathbf{i}})^T\right)^{-1}\left(\tfrac{1}{\sqrt{\widehat{\mathbf{n}}}}\sum_{i_1=1}^{n_1}\sum_{i_2=1}^{n_2}\Delta_{\mathbf{i}}\varepsilon_{\mathbf{i}}+\tfrac{1}{\sqrt{\widehat{\mathbf{n}}}}\sum_{i_1=1}^{n_1}\sum_{i_2=1}^{n_2}\Delta_{\mathbf{i}}\left(\Delta_{\mathbf{i}}^{(0)}-\Delta_{\mathbf{i}}^{T}\beta\right)\right)\end{aligned} \tag{4.18}$$

where

$$\Delta_{\mathbf{i}}^{(0)} = E[\delta_{\mathbf{i}}Y_{\mathbf{i}}|T_{\mathbf{i}}] - \sum_{i_1=1}^{n_1}\sum_{i_2=1}^{n_2}\delta_{\mathbf{i}}w_{\mathbf{n}}(T_{\mathbf{i}},T_{\mathbf{j}})Y_{\mathbf{i}}$$

So

$$\Delta_{\mathbf{i}}^{(0)} = f^{(0)}(T_{\mathbf{i}}) - \widehat{f^{(0)}}(T_{\mathbf{i}})$$

and by applying the result of chapter 2

$$\max_{\mathbf{i}\in\mathcal{I}_{n,m}}\left|\Delta_{\mathbf{i}}^{(0)}\right| \xrightarrow{p} 0 \tag{4.19}$$

So by using (4.17), (4.19) and the Cauchy-Schwarz inequality we have

$$\frac{1}{\widehat{\mathbf{n}}}\sum_{i_1=1}^{n_1}\sum_{i_2=1}^{n_2}\gamma_{\mathbf{i}}\left(\Delta_{\mathbf{i}}^{(0)}-\Delta_{\mathbf{i}}^{T}\beta\right)+\frac{1}{\widehat{\mathbf{n}}}\sum_{i_1=1}^{n_1}\sum_{i_2=1}^{n_2}\Delta_{\mathbf{i}}\varepsilon_{\mathbf{i}}+\frac{1}{\widehat{\mathbf{n}}}\sum_{i_1=1}^{n_1}\sum_{i_2=1}^{n_2}\Delta_{\mathbf{i}}\left(\Delta_{\mathbf{i}}^{(0)}-\Delta_{\mathbf{i}}^{T}\beta\right)=o_P(1)\,. \tag{4.20}$$

On the other side by applying directly **theorem 6.1.1** in **[20]** on $\{R_{\mathbf{i}},\mathbf{i}\in\mathbb{Z}^2\}$ which are a strictly stationary α-mixing random field with $\mathbb{E}(R_{\mathbf{i}})=0$, and $\mathbb{E}\left|R_{\mathbf{i}}\right|^3<\infty$, and according to the condition **(H5)(ii)** we have :

$$\frac{1}{\sqrt{\widehat{\mathbf{n}}}}\sum_{i_1=1}^{n_1}\sum_{i_2=1}^{n_2}R_{\mathbf{i}}\xrightarrow{D}N\left(0,\mathbf{C}\right)\,. \tag{4.21}$$

We conclude this proof by using (4.18), (4.20), (4.21), and Lemma 4.4.4.

Proof of Theorem 4.3.2.
We can write

$$\widehat{g}_{\mathbf{n}}(t)=\sum_{i_1=1}^{n_1}\sum_{i_2=1}^{n_2}w_{\mathbf{n}}(t,T_{\mathbf{i}})\left(g\left(T_{\mathbf{i}}\right)+\varepsilon_{\mathbf{i}}\right)-\sum_{i_1=1}^{n_1}\sum_{i_2=1}^{n_2}w_{\mathbf{n}}(t,T_{\mathbf{i}})X_{\mathbf{i}}^{T}\left(\widehat{\beta}-\beta\right)$$

So we have

$$\begin{aligned}\widehat{g}_{\mathbf{n}}(t)-g(t) &= \left(\sum_{i_1=1}^{n_1}\sum_{i_2=1}^{n_2}w_{\mathbf{n}}(t,T_{\mathbf{i}})\left(g\left(T_{\mathbf{i}}\right)+\varepsilon_{\mathbf{i}}\right)-g(t)\right)-\left(\sum_{i_1=1}^{n_1}\sum_{i_2=1}^{n_2}w_{\mathbf{n}}(t,T_{\mathbf{i}})X_{\mathbf{i}}^{T}\left(\widehat{\beta}-\beta\right)\right)\\ &= S_1-S_2\,,\end{aligned}$$

and

$$|\widehat{g}_{\mathbf{n}}(t)-g(t)|\leq\ |S_1|+|S_2| \tag{4.22}$$

Firstly $S_1=\widehat{g}^{*}_{\mathbf{n}}(t)-g(t)$ with $\widehat{g}^{*}_{\mathbf{n}}(t)=\sum_{i_1=1}^{n_1}\sum_{i_2=1}^{n_2}w_{\mathbf{n}}(t,T_{\mathbf{i}})\left(g\left(T_{\mathbf{i}}\right)+\varepsilon_{\mathbf{i}}\right)$

and the result of chapter 2 give :

$$|S_1|\xrightarrow{p}\ 0 \tag{4.23}$$

On the other hand we have :

$$|S_2| \leq \left\| \sum_{i_1=1}^{n_1} \sum_{i_2=1}^{n_2} w_{\mathbf{n}}(t, T_{\mathbf{i}}) X_{\mathbf{i}} \right\| \left\| \widehat{\beta} - \beta \right\|$$

So

$$|S_2| \leq \left\| \sum_{i_1=1}^{n_1} \sum_{i_2=1}^{n_2} \delta_{\mathbf{i}} w_{\mathbf{n}}(t, T_{\mathbf{i}}) X_{\mathbf{i}} - \mathbb{E}\left(\delta_{\mathbf{i}} X_{\mathbf{i}} | T_{\mathbf{i}} = t\right) \right\| \left\| \widehat{\beta} - \beta \right\| + \left\| \mathbb{E}\left(\delta_{\mathbf{i}} X_{\mathbf{i}} | T_{\mathbf{i}} = t\right) \right\| \left\| \widehat{\beta} - \beta \right\|$$

Under the result of Theorem1we have $\left\| \widehat{\beta} - \beta \right\| = o_P(1)$, and the result in chapter1 give : $\left\| \sum_{i_1=1}^{n_1} \sum_{i_2=1}^{n_2} \delta_{\mathbf{i}} w_{\mathbf{n}}(t, T_{\mathbf{i}}) X_{\mathbf{i}} - \mathbb{E}\left(\delta_{\mathbf{i}} X_{\mathbf{i}} | T_{\mathbf{i}} = t\right) \right\| = o_P(1)$, at the same time under the condition **(H4)** we have $\left\| E\left(\delta_{\mathbf{i}} X_{\mathbf{i}} | T_{\mathbf{i}} = t\right) \right\| < \infty$.

That implies

$$|S_2| = o_P(1) \tag{4.24}$$

From (4.22)-(4.24) the proof is over.

Bibliographie

[1] Aneiros-Pérez G, González-Manteiga W., Vieu. P. (2004). Estimation and testing in a partial regression model under long-memory dependence, Bernoulli 10 :49–78.

[2] Aneiros-Pérez G, Vieu P (2006) Semi-functional partial linear regression. Stat Probab Lett 76 :1102-1110.

[3] Aneiros-Pérez G, Vieu P (2008) Nonparametric time series prediction : a semi-functional partial linear modeling. J Multivar Anal. 99,834-857.

[4] Benallou M., Attouch M.K., Benchikh T., Fetitah O., (2021). Asymptotic results of semi-functional partial linear regression estimate under functional spatial dependency. Communications in Statistics - Theory and Methods. https ://www.tandfonline.com/doi/abs/10.1080/03610926.2020.1871021

[5] Haworth J.,, Cheng T., (2012). Non-parametric regression for space-time forecasting under missing data. Computers, Environment and Urban Systems 36 (6), 538-550.

[6] Engle R., Granger C., Rice J. and Weiss A., (1986). Nonparametric estimates of the relation between weather and electricity sales, J. Amer. Statist. Assoc., 81, 310-320.

[7] Gao, J.T. (1995). The laws of the iterated logarithm of some estimates in partly linear models. Statist. Probab. Lett., 25, 153-162.

[8] Härdle W, Liang H. and Gao J (2000). Partially linear models. Physica-Verlag, Heidelberg.

[9] Lian H (2011) Functional partial linear model. J Nonparametr. Stat. 23,115-128.

[10] Liang ; H. (2000). Asymptotic normality of parametric part in partially linear models with measurement error in the nonparametric part. J. Statist. Plann. Inference, 86(1),51-62.

[11] Linton, O. (1995). Second order approximation in the partially linear regression model, Econometrica, 63, 1079-1112.

[12] N. Ling, R. Kan, P.Vieu, S.Meng Semi-functional partially linear regressionmodel with responses missing at random.

[13] Puranik A., Binub V.S. and Seena B., (2021). Estimation of missing values in aggregate level spatial data. Clinical Epidemiology and Global Health journal. 9, 304-309.

[14] Robinson, P. M. (1988). Root-N-consistent semiparametric regression. Econometrica, 56, 931-954.

[15] Wang QH, Linton O., and Wolfgang H., (2004). Semiparametric regression analysis with missing response at random. J. Amer. Statist. Assoc. 99,334-345.

[16] Wang Q.H., Sun Z., (2007). Estimation in partially linear models with missing responses at random, J. Multivariate Anal. 98, 1470-1493.

[17] Zhang J., Clayton M.K. and Townsend P.A., (2014). Missing Data and Regression Models. for Spatial Images. IEEE Transactions on geoscience and remote sensing. 53 (3). 1574-1582.

Conclusion et Perspectives

Conclusion

Dans cette thèse, nous avons considéré la modélisation régression semi fonctionnelle partiellement linéaire en utilisant l'estimation par noyau quand les données sont spatialement . Plus précisément, nous avons considéré deux cas, le cas où la variable réponse est complètement observé et le cas où la variable réponse présente des données manquantes au hasard (MAR)

Dans le premier chapitre, nous avons estimé les paramètres du modèle semi-fonctionnelle partiellement linéaire, à savoir les paramètres de la partie paramétrique par la méthode de moindre carré et la partie non paramétrique par la méthode de noyau. Différents types des convergences ont été abordé à savoir la convergence en probabilité, la convergence presque sûr.

Les propriétés asymptotiques de ces deux estimateurs sont établies. Il démontre la convergence presque sûr et la normalité asymptotique de l'estimateur de la partie paramétrique et la convergence presque sûr de l'estimateur de la partie non-paramétrique. La performance des estimateurs est illustré par des données simulées et réelles.

La deuxième partie de la thèse est consacré au cas où la variable réponse est " Missing at random ". Nous avons considéré la régression non paramétrique pure et le modèle semi fonctionnelle partiellement linéaire. Il généralise les résultats obtenus du première partie, en montrant la convergence presque sûr et la normalité asymptotique de la version fonctionnelle. Une étude comparative avec des données simulées entre ce modèle et les modèles déjà existants est proposée.

Perspectives

L'importance de ce sujet peut être aussi exprimée en fonction des nombreuses perspectives qu'il engendre, parmi lesquels on cite :

- On peut envisager l'estimation robuste dans les modèles semi fonctionnelle partiellement linéaire pour les données spatiale.
- Il est important de considérer l'estimation locale linéaire ou l'estimation par la méthode des k plus proches voisins comme une méthode d'estimation alternative qui reste à être développer pour notre modèle.
- Une autre méthode d'estimation peut-être envisager en utilisant deux noyaux afin de contrôler à la fois la distance entre les observations et les emplacements spatiaux.
- On peut s'intéresser également aux méthodes dites récursives qui permettent une mise à jour des estimations séquentielles des données spatiales.
- Une autre alternative est d'utiliser les modèles de régression partiellement lineaires et additifs. En effet, l'estimation non-parametrique dans le cas fonctionnelle est soumise à la même contrainte de détérioration de sa vitesse de convergence. Pour pallier ce problème, la réponse adéquate est l'introduction d'une structure additive de la partie non-paramétrique par des méthodes appropriées.
- Des procédures bootstrap naïves et sauvages peuvent être proposées pour approximer la distribution des estimateurs.
- D'autres formes de régressions basées sur d'autres caractéristiques conditionnelles (quantile et mode) qui utilisent la méthode de noyau peuvent être considérées.

Bibliographie générale

Bibliographie

[1] Aerts, M., Claeskens, G., Hens, N., Molenberghs, G. (2002). Local multiple imputation. Biometrika 89(2) :375-388.

[2] Ali, M. A., Abu-Salih, M. S. (1988). On estimation of missing observations in linear regression models, Sankhya. Indian J. Statist. 50 (B),404-411.

[3] Anselin, L. and Florax, R. J. G. M. (1995). New Directions in Spatial Econometrics. Advances in Spatial Science. Springer.

[4] Aneiros-Pérez G, González-Manteiga W., Vieu. P. (2004). Estimation and testing in a partial regression model under long-memory dependence, Bernoulli 10 :49–78.

[5] Aneiros-Pérez, G., Ferraty, F. and Vieu, P. (2015a). Variable selection in partial linear regression with functional covariate, Statistics, Vol. 49, No. 6, pp. 1322–1347.

[6] Aneiros-Pérez, G., Ling, N. and Vieu P., (2015b). Error variance estimation in semi-functional partially linear regression models, J Nonparametr Stat, Vol 27, No. 3, pp. 316–330.

[7] Aneiros-Pérez G. and Vilar-Fernández J.M (2008). Local polynomial estimation in partial linear regression models under dependence. Computational statistics et data analysis, 52 (5), 2757-2777.

[8] Aneiros-Pérez G. and Vieu P. (2006). Semi-functional partial linear regression. Stat. Probab. Lett., 76, (11), 1102-1110.

[9] Aneiros-Pérez G. and Vieu P. (2008). Nonparametric time series prediction. A semi-functional partial linear modeling. Journal of Multivariate Analysis, 99, 834-857.

[10] Aneiros-Pérez, G. and Vieu, P. (2011). Automatic estimation procedure in partial linear model with functional data. Stat.Pap., 52, (4), 751-771.

[11] Aneiros-Pérez, G. and Vieu, P. (2013). Testing linearity in semi-parametric functional data analysis, Computational Statistics, Vol 28,No. 2, pp. 413–434.

[12] Araujo A.„ Gine E., (1980). The Central Limit Theorem for Real and Banach Valued Random Variables, Wiley, New York, 1980.

[13] Attouch, M., Chouaf, B. and Laksaci, A. (2012) Nonparametric M-estimation for functional spatial data , Communication of the Korean Statistical society, 19, 193-211.

[14] Attouch, M. K., Gheriballah, A., and Laksaci, A. (2011) Robust nonparametric estimation for functional spatial regression. In Ferraty, F., editor, Recent Advances in Functional Data Analysis and Related Topics, Contributions to Statistics, 27-31. Physica-Verlag HD.

[15] Attouch, M., Laksaci, A. and Ould-Said, E. Asymptotic distribution of robust estimator for functional nonparametric models. Comm. Statist. Theory Methods., 38 (2009), no. 8-10, 1317-1335.

[16] Azzedine, N., Laksaci, A. and Ould-Said, E. On robust nonparametric, regression estimation for a functional regressor. Statist. Probab. Lett., 78 (2008), no. 18, 3216-3221

[17] Barrientos-Marin, J., Ferraty, F. and Vieu, P. Locally modelled regression and functional data. J. Nonparametr. Stat., 22, (2010), no. 5-6,617-632.

[18] Benallou M., Attouch M.K., Benchikh T., Fetitah O., (2021). Asymptotic results of semi-functional partial linear regression estimate under functional spatial dependency. Communications in Statistics - Theory and Methods. https ://www.tandfonline.com/doi/abs/10.1080/03610926.2020.1871021.

[19] Benhenni, K., Ferraty, F., Rachdi, M. and Vieu, P. Local smoothing regression with functional data. Comput. Statist., 22 (2007), no. 3, 353-369.

[20] Besse, P., Cardot, H. et Stephenson D. (2000) Autoregressive Forecasting of Some Functional Climatic Variations. Scandinavian Journal of Statistics 27 ; 673-687.

[21] Biau, G. and Cadre, B. (2004). Nonparametric spatial prediction. Statistical Inference for Stochastic Processes, 7(3) :327-349.

[22] Boente, G., Gonzalez-Manteiga, W. and Perez-Gonzalez, A. (2009). Robust nonparametric estimation with missing data. J. Statist. Plann. Inference, 139, 571-592.

[23] Boente, G. and Vahnovan, A. (2017). Robust estimators in semi-functional partial linear regression models, Journal of Multivariate Analysis, Vol 154, No. C, pp. 59–84.

[24] Bosq, D. Linear Processes in Function Spaces : Theory and applications. Lecture Notes in Statistics, 149, Springer. (2000).

[25] Brown, L. D., Levine, M., and Wang, L. (2016). A semiparametric multivariate partially linearmodel : A difference approach.Journal of Statistical Planning and Inference, 178 :99-111.

[26] Bulinski, A., and Shashkin, A., (2006). Strong invariance principle for dependent random fields. Dynamics and Stochastics, 128-143.

[27] Burba, F., Ferraty, F. and Vieu, P. k-nearest neighbour method in functional nonparametric regression. J. Nonparametr. Stat., 21 (2009), no. 4, 453-469.

[28] Carbon, M., Tran, L. T., and Wu, B., (1997). Kernel density estimation for random fields (density estimation for random fields). Stat Probab Lett, 36, (2), 115-125.

[29] Carbon, M., Francq, C., and Tran, L. T. (2007). Kernel regression estimation for random fields. Journal of Statistical Planning and Inference, 137 (3),778-798.

[30] Chang, X.W. et Qu, L. (2004). Wavelet estimation of partially linear models. Computationnal Statistics and Data Analysis. 47(1), 31-48.

[31] Chen, J. et Shao, J. (2000). Nearest Neighbor Imputation for Survey Data. Journal of Official Statistics, 16, 113-131.

[32] Cheng, P. (1994). Nonparametric estimation of mean functionals with data missing at random, J. Amer. Statist. Assoc., 89, 81-87.

[33] Chouaf, A. and Laksaci, A. On the functional local linear estimate for spatial regression. Stat. Risk Model., 29 (2012), no. 3, 189-214.

[34] Chu, C. K., Cheng, P. E. (1995). Nonparametric regression estimation with missing data. J. Statist. Planning Inference. 48,85-99.

[35] Crambes, C., Delsol, L. and Laksaci, A. (2008). Robust nonparametric estimation for functional data. J. Nonparametr. Stat., 20 (7), 573-598.

[36] Cressie, N. A. C. (1993). Statistics for Spatial Data. Wiley, New York.

[37] Cressie, N. and Wikle, C. K. (2011). Statistics for Spatio-Temporal Data. Wiley Series in Probability and Statistics. John Wiley and Sons.

[38] Dabo-Niang, S., Kaid, Z., and Laksaci, A. (2011b). Sur la régression quantile pour variable explicative fonctionnelle : Cas des données spatiales. CRAS, 349(23) :1287-1291.

[39] Dabo-Niang, S., Kaid, Z., and Laksaci, A., (2012a). On spatial conditional mode estimation for a functional regressor. Stat Probab Lett, 82, (7), 1413-1421.

[40] Dabo-Niang, S., Kaid, Z., and Laksaci, A., (2012b). Spatial conditional quantile regression : Weak consistency of a kernel estimate, Rev. Roumaine Math. Pures Appl 57, 311-339.

[41] Dabo-Niang, S., Rachdi, M., and Yao, A.F., (2011). Kernel regression estimation for spatial functional random variables. Far East Journal of Theoretical Statistics, 37, (2), 77-113.

[42] Dabo-Niang, S. and Rhomari, N. (2003). Estimation non paramétrique de la régression avec variable explicative dans un espace métrique. Comptes Rendus Mathematique, 336(1) :75-80.

[43] Dabo-Niang, S. and Rhomari, N. (2009). Kernel regression estimation in a Banach space. J. Statistical Planning and Inference, 139, 1421-1434.

[44] Dabo-Niang, S., Ternynck, C. and Yao, A.-F. (2016). Nonparametric prediction of spatial multivariate data. Journal of Nonparametric Statistics, 28 , 428-458.

[45] Dabo-Niang, S. and Thiam, B. (2010). Robust quantile estimation and prediction for spatial processes. Stat. Probab. Letters, 80(17) :1447-1458.

[46] Dabo-Niang, S. and Yao, A.-F. (2007). Kernel regression estimation for continuous spatial processes. Mathematical Methods of Statistics, 16(4) :298-317.

[47] Dabo-Niang, S., Yao, A.-F., Pischedda, L., Cuny, P., and Gilbert, F. (2010). Spatial mode estimation for functional random fields with application to bioturbation problem. Stochastic Environmental Research and Risk Assessment, 24, (4), 487-497.

[48] Efromovich, S., (2011b) Nonparametric Regression with Predictors Missing at Random, J. Amer. Statist. Assoc., 106 : 306-319.

[49] Efromovich, S., (2011a). Nonparametric regression with responses missing at random. J. Statist. Plann. Inference 141, 3744-3752.

[50] Efromovich, S., (2014). Nonparametric regression with missing data, Wiley Interdisciplinary Reviews Computational Statistics, 6 (4), 265-275.

[51] El Machkouri, M. and Stoica, R. (2010). Asymptotic normality of kernel estimates in a regression model for random fields. Journal of Nonparametric Statistics, 22(8) :955-971.

[52] Engle R., Granger C., Rice J. and Weiss A., (1986). Nonparametric estimates of the relation between weather and electricity sales, J. Amer. Statist. Assoc., 81, 310-320.

[53] Ferraty, F., Goia, A. and Vieu, P. (2002). Functional nonparametric model for time series : a fractal approach for dimension reduction. Test, 11 (2), 317-344.

[54] Ferraty, I. Van Keilegom, P. Vieu.(2012b). Regression when both response and predictor are functions, J. Multivar.Anal.109, 10-28.

[55] Ferraty, A. Laksaci, A. Tadj, P. Vieu.(2011). Kernel regression with functional response.Electronic. J. Stat.5 159-171.

[56] Ferraty, A. Laksaci, A. Tadj, P. Vieu.(2012a). Estimation de la fonction de régression pour variable explicative et réponses fonctionnelles dépendante.C. R. Acad. Sci.Paris, Ser.I 350, 717-720.

[57] Ferraty, F., Mas, A. and Vieu, P. (2007). Nonparametric regression on functional data : inference and practical aspects. Aust. N. Z. J. Stat., 49 (3), 267-286.

[58] Ferraty, F., Sued, M, Vieu, P. (2013). Mean estimation with data missing at random for functional covariables, Statistics, 47, 688-706.

[59] Ferraty, F. and Vieu, P. (2000) Dimension fractale et estimation de la régression dans des espaces vectoriels semi-normés. C.R. Math. Acad. Sci. Paris., 330 (2), 139-142.

[60] Ferraty, F. et Vieu, P. (2002) The functional nonparametric model and application to spectrometric data. Comput. Statist. 17 (4) 545-564.

[61] Ferraty, F. and Vieu, P. (2003) Curves discrimination : a nonparametric functional approach Computational Statistics and Data Analysis 44 (1-2) 161-173.

[62] Ferraty, F., Vieu, P. (2004). Nonparametric models for functional data, with application in regression times series prediction and curves discrimination. J. Nonparametric Statist., 16, 111-127.

[63] Ferraty, F. and Vieu, P. (2006). Nonparametric functional data analysis. Theory and Practice. Springer Series in Statistics. New York.

[64] Ferraty, F., Vieu, P. (2011). Kernel regression estimation for functional data.Inthe Oxford Handbook of Functional Data Analysis (Ed. F. Ferraty and Y. Romain).Oxford University Press

[65] Gao, J.T. (1995). The laws of the iterated logarithm of some estimates in partly linear models. Statist. Probab. Lett., 25, 153-162.

[66] Gao, J. T., Lu, Z. and Tjϕstheim D., (2006). Estimation in semiparametric spatial regression. The Annals of Statistics, 34, (3), 1395-1435.

[67] Gasser, T., Hall, P. et Presnell, B. (1998) Nonparametric estimation of the mode of a distribution of random curves. J. R. Statomptes. Soc. Ser. B Stat. Methodol. 60 (4) 681-691.

[68] Gheriballah, A., Laksaci, A., and Rouane, R. (2010). Robust nonparametric estimation for spatial regression. Journal of Statistical Planning and Inference, 140(7) :1656 - 1670.

[69] Germán A., Raña P., Vieu p., et Vilar J. (2017). Bootstrap in semi-functional partial linear regression under dependence. TEST. DOI 10.1007/s11749-017-0566-y.

[70] Giraldo, R., Delicado, P. and Mateu, J. (2011). Geostatistics with infinite dimensional data : a generalization of cokriging and multivariable spatial prediction. Matematica : ICM-ESPOL, 9 (1), 16-21.

[71] Giraldo, R., Dabo-Niang, S. and Martinez, S. (2018). Statistical modeling of spatial big data : An approach from a functional data analysis perspective. Statistics and Probability Letters, 136, 126-129.

[72] Goia, A., Vieu, P., (2016). An introduction to recent advances in high/infinite dimensional statistics. J. Multivariate Anal., 146, 1-6.

[73] Graham, J. W., Missing data analysis and design. NY : Springer, New York, 2012.

[74] Guyon, X. (1995).Random Fields on a Network - Modeling, Statistics, and Applications, Springer, New-York

[75] Hallin, M., Lu, Z., and Tran, L. T. (2004). Local linear spatial regression. The Annals of Statistics, 32(6), 2469-2500.

[76] Hallin, M., Lu, Z. and Yu, K. (2009). Local linear spatial quantile regression, Bernouilli. 15, 659-686.

[77] Hamilton, S. A. and Truong, Y. K. (1997). Local linear estimation in partly linear models. J. Multivariate Anal., 60(1), 1-19.

[78] Härdle, W. (1990), Applied Nonparametric Regression, Econometric Society Monographs, Cambridge : Cambridge University Press.

[79] Härdle W, Liang H. and Gao J (2000). Partially linear models. Physica-Verlag, Heidelberg.

[80] Haworth J.,, Cheng T., (2012). Non-parametric regression for space-time forecasting under missing data. Computers, Environment and Urban Systems 36 (6), 538-550.

[81] Healy, M.J.R. and Westmacott,M.(1956). Missing values in experiments analyzed on automatic computers. Appl. Statist

[82] Hsing, T. and Eubank, R.L., (2015). Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators. John Wiley and Sons.

[83] Ibrahim F, Ali-Hajj H, Demongeot J, et al. Regression model for surrogate data in high dimensional statistics. Commun Stat Theory Methods. Pages 3206-3227 49, 2020 - Issue 13

[84] Kara-Zaitri, L., Laksaci A., Rachdi M. et Vieu P. (2016). Uniform in bandwidth consistency for various kernel estimators involving functional data,Journal of Nonparametric Statistics, DOI : 10.1080/10485252.2016.1254780

[85] Karácsony, Z. et Filzmoser, P. (2010). Asymptotic normality of kernel type regression estimators for random fields. Journal of Statistical Planning and Inference, 140 : 872-886.

[86] Kraus D., (2015). Components and completion of partially observed functional data. J R Stat Soc B 77 :777-801

[87] Laksaci, A. and Maref, F. (2009). Estimation non paramétrique de quantiles conditionnels pour des variables fonctionnelles spatialement dépendantes. Comptes Rendus Mathématique, 347(17-18) :1075-1080.

[88] Laksaci, A. and Mechab, B. (2010). Estimation non paramétrique de la fonction de hasard avec variable explicative fonctionnelle : cas des données spatiales. Revue Roumaine deMathématiques Pures et Appliquées, 55(1) :35-51.

[89] Li, J. and Tran, L. T. (2009). Nonparametric estimation of conditional expectation. J. Statist. Plann. Inference. 139, 164-175.

[90] Lian H., (2011). Functional partial linear model. J Nonparametr Stat, 23(1),115-128.

[91] Liang ; H. (2000). Asymptotic normality of parametric part in partially linear models with measurement error in the nonparametric part. J. Statist. Plann. Inference, 86(1),51-62.

[92] Liang ; H, Wang S, Carroll R (2007) Partially linear models with missing response variables and error-prone covariates. Biometrika 94 :185-198.

[93] Ling ; N., Liang LL, Vieu P (2015) Nonparametric regression estimation for functional stationary ergodic data with missing at random. J Stat Plan Inference 162 :75-87.

[94] Ling ; N., Liu Y, Vieu P (2016) Conditional mode estimation for functional stationary ergodic data with responses missing at random. Statistics 50 :1-23.

[95] Ling ; N. and Vieu P., (2018). Nonparametric modelling for functional data : selected survey and tracks for future. Statistics, 52, (4), 934-949.

[96] Ling, N., Aneiros-Pérez, G. and Vieu, P., (2017), knn estimation in functional partial linear modeling, Statist. Papers, Vol 61, No. 1, pp. 423–444.

[97] Ling ; N. and Vieu P., (2020). On semiparametric regression in functional data analysis. WIRES Computational Statistics, 12, (6), 20-30.

[98] Linton, O. (1995). Second order approximation in the partially linear regression model, Econometrica, 63, 1079-1112.

[99] Little, R, Rubin, D. : Statistical Analysis with Missing Data, Second Edition, Wiley, New York, (2002)

[100] Little, R. J. A. and Rubin, D. B. (2020) ; Statistical Analysis with Missing Data, 3rd Edition ; Wiley Series in Probability and Statistics

[101] Lu, Z. and Chen, X. (2004). Spatial kernel regression estimation : weak concistency, Stat. and Probab. Lett., 68, pp. 125-136.

[102] Mason, A., Richardson, S., Plewis, I. and Best, N. (2012) Strategy for modelling nonrandom missing data mechanisms in observational studies using Bayesian methods. Journal of Official Statistics, 28, 279-302

[103] Masry, E. (2005). Nonparametric regression estimation for dependent functional data : Asymptotic normality. Stoch. Proc. and their Appl., 115, 155-177.

[104] Mateu, J. and Romano, E., (2017). Advances in spatial functional statistics. Stoch Environ Res Risk Assess, 31, 1-6.

[105] Menezes, R., García-Soidán, P., and Ferreira, C. (2010). Nonparametric spatial prediction under stochastic sampling design. Journal of Nonparametric Statistics, 22(3), 363-377.

[106] Molenberghs G., Fitzmaurice G., Kenward M.K., Tsiatis A., Verbeke G., (2015). Handbook of Missing Data Methodology.

[107] Müller U.U., (2009). Estimating linear functionals in nonparametrique regression with responses missing at random. The Annals of Statistics. 37 (5A), 2245-2277.

[108] Nadaraya, E. A. (1964). On estimating regression. Theory of Probability and Its Applications, 9(1) :141-142.

[109] Nerini, D., Monestiez, P., Manté, C., 2010. Cokriging for spatial functional data. J. Multivariate Anal. 101 (2), 409-418.

[110] Nittner T (2003) Missing at random (MAR) in nonparametric regression-a simulation experiment. Stat Methods Appl 12 :195-210.

[111] Ould Abdi A., Diop A., Dabo-Niang S. et Ould Abdi S.A. (2010a), Estimation non paramétrique du mode conditionnel dans le cas spatial Nonparametric estimation of conditional mode in the spatial case CRAS ; 348 (13), 815-819.

[112] Ould Abdi A., Diop A., Dabo-Niang S. et Ould Abdi S.A., (2010b), Consistency of a Nonparametric Conditional Quantile Estimator for Random Fields, Mathematical Methods of Statistics, 19 (1) :1-21.

[113] Perez-Gonzalez, A., Vilar-Fernandez and J. M. Gonzalez-Manteiga, W. (2009). Asymptotic properties of local polynomial regression with missing data and correlated errors. Ann. Inst. Statist. Math., 61, 85-109.

[114] Rachdi, M., Laksaci, A., Almanjahie, I.M. and Chikr-Elmezouar, Z. (2020). FDA :theoretical and practical efficiency of the local linear esti-

mation based on the kNN smoothing of the conditional distribution when there are missing data. J. Stat. Comput.Simul., 90, 1479-1495.

[115] Puranik A., Binub V.S. and Seena B., (2021). Estimation of missing values in aggregate level spatial data. Clinical Epidemiology and Global Health journal. 9, 304-309.

[116] Rachdi M.„ Laksaci A., Kaid Z., Benchiha A., Fahimah A. Al Awadhi, 2021. k-Nearest neighbors local linear regression for functional and missing data at random, Statistica Neerlandica, Netherlands Society for Statistics and Operations Research, 75(1), 42-65.

[117] Rachdi, M., Vieu, P. (2007). Nonparametric regression functional data : Automatic smoothing parameter selection. J. Statist. Plan. Inf., 137, 2784-2801.

[118] Ramsay, J. (2008). Fda problems that i like to talk about. Personal communication.

[119] Ramsay, J. and Silverman, B. (1997) Functional Data Analysis Springer-Verlag, New York.

[120] Ramsay, J. and Silverman, B. (2002) Applied functional data analysis : Methods and case studies Spinger-Verlag, New York.

[121] Ramsay, J. and Silverman, B. (2005) Functional Data Analysis (Second Edition) Spinger-Verlag, New York.

[122] Rao, J.N.K. (1996). On variance estimation with imputed survey data. Journal of the American Statistical Association, 91, 499-506.

[123] Rice, J. (1986). Convergence rates for partially splined models. Statist. Probab. Lett., 4, 203-208.

[124] Ripley, B. D. (1981). Spatial Statistics. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc.

[125] Robinson, P. M. (1988). Root-N-consistent semiparametric regression. Econometrica, 56, 931–954.

[126] Robinson, P. M. (2011). Asymptotic theory for nonparametric regression with spatial data. Journal of Econometrics, 165(1) :5-19.

[127] Schafer J.L. and Graham J.W. (2002). Missing Data : Our View of the State of the Art. Psychological Methods, 7 (2), 147-177.

[128] Shang H., (2014). Bayesian bandwidth estimation for a semi-functional partial linear regression model with unknown error density. Comput. Stat., 29 (3-4), 829-848

[129] Siepman, H. R., Yang, S.-S. (1994). Generalized least squares estimation of multivariate nonlinear models with missing data. Commun. Statist., 23(6), 1565-1579.

[130] Silverman, B.W. (1986), Density Estimation for Statistics and Data Analysis, in Monographs on Statistical Subjects, London : Chapman and Hall.

[131] Speckman P. (1988). Kernel smoothing in partial linear models. J. Roy. Statist. Soc. Ser. B, 50, 413-436.

[132] Stock, C. J. (1989). Nonparametric policy analysis. J. Amer. Statist. Assoc., 89, 567-575.

[133] Ternynck, C., (2014). Spatial regression estimation for Functional data with spatial dependency. Journal de la Société Française de Statistique, Vol. 155, No. 2.

[134] Tran, L. T. (1990). Kernel density estimation on random fields. Journal of Multivariate Analysis, 34(1) :37-53.

[135] Tran, L. T. and Yakowitz, S. (1993). Nearest neighbor estimators for random fields. Journal of Multivariate Analysis, 44(1) :23-46.

[136] Trivellore R. ; (2015). Missing Data Analysis in Practice. Chapman and Hall, Taylor and Francis Group.

[137] Tsiatis A (2006) Semiparametric theory and missing data. Springer, New York.

[138] Trivellore R. ; (2015). Missing Data Analysis in Practice. Chapman and Hall, Taylor and Francis Group.

[139] V.A. Volkonskii, Yu.A. Rozanov, (1959). Some limit theorems for random functions, Theory Probab. Appl. 4, 178–197.

[140] Wand, M.P., and Jones, M.C. (1995), Kernel Smoothing, in Monographs on Statistics and Applied Probability, London : Chapman and Hall.

[141] Wang QH, Linton O, Wolfgang H (2004) Semiparametric regression analysis with missing response at random. J Am Stat Assoc 99 :334-345.

[142] Wang Q.H., Linton O., and Wolfgang H., (2004). Semiparametric regression analysis with missing response at random. J. Amer. Statist. Assoc. 99,334-345.

[143] Wang Q.H., Sun Z., (2007). Estimation in partially linear models with missing responses at random, J. Multivariate Anal. 98, 1470-1493.

[144] Wang, H.,Wang, J., (2009). Estimation of the trend function for spatio-temporal models. Journal of Nonparametric Statistics, 21 : 567-588.

[145] Watson, G. S. (1964). Smooth regression analysis. Sankhya : The Indian Journal of Statistics, Series A, pages 359-372.

[146] Xu, R., Wang, J., 2008.L1-estimation for spatial nonparametric regression. Nonparametric Statist. 20, 523-537

[147] Yates, F. (1933). The analysis of replicated experiments when the field results are incomplete. Emporium J. Exp. Agriculture, 1, 129-142.

[148] Zhang J., Clayton M.K. and Townsend P.A., (2014). Missing Data and Regression Models. for Spatial Images. IEEE Transactions on geoscience and remote sensing. 53 (3). 1574-1582.

ملخص

في هذا العمل ، نأخذ في الاعتبار النماذج الخطية الجزئية عندما تكون المتغيرات التوضيحية وظيفية وتعتمد على المكاني. أولاً ، نبدأ بإعطاء النسخة المكانية لمقدري النواة للمكونين (الخطي واللامعلمي) عند اكتمال البيانات. تم تحديد الخصائص المقاربة لهذين المقدرين. لقد أثبتنا التقارب المؤكد تقريبًا والحالة الطبيعية المقاربة لمقدر الجزء البارامترى والتقارب شبه المؤكد لمقدر الجزء غير البارامترى. يتم توضيح أداء المقدرين من خلال بيانات محاكاة وحقيقية. في الحالة الثانية ، نهتم بالحالة التي يكون فيها متغير الاستجابة "مفقود عشوائيًا". نقوم بتعميم النتائج التي تم الحصول عليها في الجزء الأول ، من خلال إظهار التقارب في الاحتمالية والحالة الطبيعية المقاربة للإصدار الوظيفي. تم اقتراح دراسة مقارنة مع بيانات محاكاة بين هذا النموذج والنماذج الحالية

Résumé

Dans ce travail, nous considérons les modèles partielles linéaires lorsque variables explicatives sont fonctionnelles et spatialement dépendantes. Dans un premier temps, on commence par donner la version spatiale des deux estimateurs à noyau pour les deux composantes (linéaire et non paramétrique) quand les données sont complètes. Les propriétés asymptotiques de ces deux estimateurs sont établies. On démontre la convergence presque sûr et la normalité asymptotique de l'estimateur de la partieparamétrique et la convergence presque sûr de l'estimateur de la partie non-paramétrique. La performance des estimateurs sont illustrés par des données simulées et réelles. Dans la deuxième, on s'intéresse au cas où la variable réponse est « Missing at random ». On généralise les résultats obtenus dans la première partie, en montrant la convergence en probabilité et la normalité asymptotique de la version fonctionnelle. Une étude comparative avec des données simulées entre ce modèle et les modèles déjà existants est proposée.

Abstract

In this work, we consider partial linear models when explanatory variables are functional and spatially dependent. First, we start by giving the spatial version of the two kernel estimators for the two components (linear and nonparametric) when the data are complete. The asymptotic properties of these two estimators are established. We prove the almost sure convergence and asymptotic normality of the estimator of the parametric part and the almost sure convergence of the estimator of the non-parametric part. The performance of the estimators are illustrated by simulated and real data. In the second, we are interested in the case where the response variable is "Missing at random". We generalize the results obtained in the first part, by showing the convergence in probability and the asymptotic normality of the functional version. A comparative study with simulated data between this model and existing models is proposed.