Estimation nonparamétrique de la fonction de régression par la méthode des k-plus proches voisins pour donnés spatiales

M.S. Ahmed [1], M.K. Attouch [2], S. Dabo-Niang [1] & A. Diop [3]

[1] *Université de Lille, Laboratoire LEM, France*
E-mail: mohamed-salem.ahmed@etu.univ-lille3.fr; sophie.dabo@univ-lille3.fr
[2] *Univversité de Sidi Bel Abbées, Laboratoire de Statistique Processus Stochastiques, Algérie, E-mail: attou_kadi@yahoo.fr*
[3] *Université Gaston Berger, Saint-Louis, Sénégal*
E-mail: aliou.diop@ugb.edu.sn

Résumé. Nous proposons une généralisation de la méthode des k-plus proches voisins (Collomb (1980)) à des données spatialement dépendantes dans le but d'estimer la fonction de régression (Biau et Cadre (2004)) $r(x) = E(Y|X = x)$ à partir de réalisations d'un processus spatial strictement stationnaire $\{Z_{\mathbf{i}} = (Y_{\mathbf{i}}, X_{\mathbf{i}}) \in \mathbb{R} \times \mathbb{R}^d, \mathbf{i} \in \mathbb{Z}^N\}\,(N > 1)$ (avec $Z_{\mathbf{i}}$ de même loi que le couple de variable (X, Y)) dans une région rectangulaire $\mathcal{I}_{\mathbf{n}} = \{\mathbf{i} = (i_1, \cdots, i_N) \in \mathbb{N}^N,\, 1 \leq i_k \leq n_k,\, k = 1, \ldots, N\}$.
L'estimateur de la fonction de régression proposé est basé sur un double noyau introduit par Dabo-Niang *et al.* (2014). Ces auteurs ont proposé un estimateur à noyau de la fonction de régression dans le cas de données spatialement dépendantes en se basant sur deux noyaux : l'un contrôle la structure de dépendance spatiale et l'autre contrôle la distance entre les observations.
Nous adaptons cette méthode dans le cadre des k-plus proches voisins en utilisant sur le noyau qui contrôle les observations une fenêtre aléatoire de lissage. Cette fenêtre aléatoire de lissage est définie par la distance entre la réalisation de X au site spatial où on veut estimer la fonction de régression et la k-ième plus proche réalisation aux sites voisins. Nous établissons sous des hypothèses générales, la convergence presque complète de notre estimateur en précisant la vitesse de convergence. Sur des données simulées et réelles, nous produisons des résultats numériques tout en comparant notre estimation à celle obtenue par Dabo-Niang *et al.* (2014) avec la méthode du noyau.

Mots-clés. Régression spatiale nonparamétrique, k-NN estimateur, Vitesse de convergence, Fenêtre aléatoire de lissage.

Abstract. We propose a generalization of the method of k nearest neighbours (k-NN, Collomb (1980)) on spatially dependent data in order to estimate the regression function (Biau and Cadre, (2004)) $r(x) = E(Y|X = x)$ from the realizations of a strictly stationary spatial process $\{Z_{\mathbf{i}} = (Y_{\mathbf{i}}, X_{\mathbf{i}}) \in \mathbb{R} \times \mathbb{R}^d, \mathbf{i} \in \mathbb{Z}^N\}\,(N > 1)$ (where $Z_{\mathbf{i}}$ has same distribution as the random vector (X, Y)) observed in a rectangular region $\mathcal{I}_{\mathbf{n}} = \{\mathbf{i} = (i_1, \cdots, i_N) \in$

$\mathbb{N}^N$, $1 \leq i_k \leq n_k$, $k = 1, \ldots, N\}$.
The estimator of the regression function proposed is based on a double kernel estimation method introduced by Dabo-Niang *et al.* (2014). These authors proposed an estimator of the regression function in a case of spatially dependent data. Their estimator is based on two kernels: while one controls the spatial dependence structure the other controls the distance between the observations.
We adapt this method in the context of $k-$nearest neighbours method using on the kernel which controls the observations a random bandwidth. This random bandwidth is defined by the distance between the realization of X at the spatial site at which we want to estimate the regression function and the $k-$th nearest realization to neighbours sites. Then, we establish under general assumptions, the almost complete convergence. Finally, on simulated and real data, we compare our estimation method with that of Dabo-Niang *et al.* (2014).

Keywords. Nonparametric spatial regression, k-NN estimator, rate of convergence, random bandwidth

Presentation of the estimator and assumptions

Let $\{Z_{\mathbf{i}} = (X_{\mathbf{i}}, Y_{\mathbf{i}}) \in \mathbb{R}^d \times \mathbb{R}\ ,\ \mathbf{i} \in \mathbb{N}^N\}$ $(d \geq 1)$ be a strictly stationary spatial process such that $Z_{\mathbf{i}}$ has same distribution as $Z = (X, Y)$ over the probability space $(\Omega, \mathcal{A}, \mathbb{P})$, $N \in \mathbb{N}^*$. We assume that Y is bounded and X has a probability density f and the process is observed in $\mathcal{I}_{\mathbf{n}} = \{\mathbf{i} \in \mathbb{N}^N\ :\ 1 \leq i_k \leq n_k\ k = 1, \ldots, N\}$, $\mathbf{n} = (n_1, \ldots, n_N) \in \mathbb{N}^N$, and $\hat{\mathbf{n}} = n_1 \times \ldots \times n_N$, we write $\mathbf{n} \to \infty$ if $\min\{n_k\} \to +\infty$ and for each $1 \leq k, l \leq N$ we have $n_l/n_k < C$. This means that the number of observations on the rectangular region expands to infinity at the same rate along all directions. Such an expansion is called isotropic divergence. Let $\|\cdot\|$ denote the Euclidian norm in $\mathbb{R}^N$ or $\mathbb{R}^d$.
In this paper, we are interested in the regression model defined by $Y_{\mathbf{i}} = r(X_{\mathbf{i}}) + \epsilon_{\mathbf{i}}$ where $r(x) = \mathbb{E}(Y|X = x)$, the noise $\epsilon_{\mathbf{i}}$ is centered, α-mixing and independent of $X_{\mathbf{s}}$. Namely, we are particularly interested in the prediction of $Y_{\mathbf{s}}$ under the condition that $X_{\mathbf{s}} = x$ (as in Dabo-Niang *et al.* (2014), Wang and Wang (2009)), that we denote in what follows $x_{\mathbf{s}}$. Using nearest neighbours sites of $\mathbf{s}$, we construct a spatial k nearest neighbours regression estimate of $r(x_{\mathbf{s}})$ with weight on the neighbours (normalized) sites as follows:

$$r_{\mathrm{kNN}}(x_{\mathbf{s}}) = \begin{cases} \dfrac{g_{\mathbf{n}}(x_{\mathbf{s}})}{f_{\mathbf{n}}(x_{\mathbf{s}})} & \text{if } f_{\mathbf{n}}(x_{\mathbf{s}}) \neq 0; \\ \overline{Y}, & \text{the empirical mean, otherwise,} \end{cases}$$

with

$$g_{\mathbf{n}}(x_{\mathbf{s}}) = \frac{1}{\hat{\mathbf{n}} h_{\mathbf{n},\mathbf{s}}^N H_{\mathbf{n},x_{\mathbf{s}}}^d} \sum_{\mathbf{i} \in \mathcal{I}_{\mathbf{n}},\, \mathbf{i} \neq \mathbf{s}} K_1\left(h_{\mathbf{n},\mathbf{s}}^{-1} \left\| \frac{\mathbf{s} - \mathbf{i}}{\mathbf{n}} \right\|\right) K_2\left(\frac{x_{\mathbf{s}} - X_{\mathbf{i}}}{H_{\mathbf{n},x_{\mathbf{s}}}}\right) Y_{\mathbf{i}}$$

$$f_{\mathbf{n}}(x_{\mathbf{s}}) = \frac{1}{\hat{\mathbf{n}} h_{\mathbf{n},\mathbf{s}}^N H_{\mathbf{n},x_{\mathbf{s}}}^d} \sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}},\, \mathbf{i}\neq\mathbf{s}} K_1\left(h_{\mathbf{n},\mathbf{s}}^{-1}\left\|\frac{\mathbf{s}-\mathbf{i}}{\mathbf{n}}\right\|\right) K_2\left(\frac{x_{\mathbf{s}}-X_{\mathbf{i}}}{H_{\mathbf{n},x_{\mathbf{s}}}}\right)$$

where K_1 and K_2 are two kernels from $\mathbb{R}$ to $\mathbb{R}_+$ and $\mathbb{R}^d$ to $\mathbb{R}_+$ respectively, $\frac{\mathbf{i}}{\mathbf{n}} = \left(\frac{i_1}{n_1}, \cdots, \frac{i_N}{n_N}\right)$, and $H_{\mathbf{n},x_{\mathbf{s}}} = \min\left\{h \in \mathbb{R}_+^* \,|\, \sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}},\, \mathbf{i}\neq\mathbf{s}} \mathbb{I}_{\{\|X_{\mathbf{i}}-x_{\mathbf{s}}\|<h\}} = k(\mathbf{n})\right\}$ and

$h_{\mathbf{n},\mathbf{s}} = \min\left\{h \in \mathbb{R}_+^* \,|\, \sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}},\, \mathbf{i}\neq\mathbf{s}} \mathbb{I}_{\left\{\left\|\frac{\mathbf{i}-\mathbf{s}}{\mathbf{n}}\right\|<h\right\}} = k_{\mathbf{n}}^1\right\}$ where $k_{\mathbf{n}}^1$, $k(\mathbf{n})$ are positive integers sequences and $H_{\mathbf{n},x_{\mathbf{s}}}$ is a positive random variable which depends on $\{X_{\mathbf{i}},\, \mathbf{i} \in I_{\mathbf{n}}\}$. $\mathbb{I}_A(\cdot)$ is the indicator function on a subset A.

In parallel, in order to show the differences between the k-NN method and the traditional kernel approach, we recall the kernel regression function given in Dabo-Niang *et al.* (2014):

$$r_{NW}(x_{\mathbf{s}}) = \begin{cases} \dfrac{g_{\mathbf{n}}(x_{\mathbf{s}})}{f_{\mathbf{n}}(x_{\mathbf{s}})} & \text{if } f_{\mathbf{n}}(x_{\mathbf{s}}) \neq 0; \\ \dfrac{1}{\hat{\mathbf{n}}} \displaystyle\sum_{\mathbf{i}\in I_{\mathbf{n}}} Y_{\mathbf{i}} & \text{otherwise,} \end{cases}$$

with

$$g_{\mathbf{n}}(x_{\mathbf{s}}) = \frac{1}{\hat{\mathbf{n}} h_{\mathbf{n}}^d \rho_{\mathbf{n}}^N} \sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}},\, \mathbf{i}\neq\mathbf{s}} K_1\left(\rho_{\mathbf{n}}^{-1}\left\|\frac{\mathbf{s}-\mathbf{i}}{\mathbf{n}}\right\|\right) K_2\left(\frac{x_{\mathbf{s}}-X_{\mathbf{i}}}{h_{\mathbf{n}}}\right) Y_{\mathbf{i}}$$

$$f_{\mathbf{n}}(x_{\mathbf{s}}) = \frac{1}{\hat{\mathbf{n}} h_{\mathbf{n}}^d \rho_{\mathbf{n}}^N} \sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}},\, \mathbf{i}\neq\mathbf{s}} K_1\left(\rho_{\mathbf{n}}^{-1}\left\|\frac{\mathbf{s}-\mathbf{i}}{\mathbf{n}}\right\|\right) K_2\left(\frac{x_{\mathbf{s}}-X_{\mathbf{i}}}{h_{\mathbf{n}}}\right)$$

where $\rho_{\mathbf{n}}$ and $h_{\mathbf{n}}$ are bandwidths tending to zero such that $\hat{\mathbf{n}} h_{\mathbf{n}}^d \rho_{\mathbf{n}}^N \longrightarrow \infty$.

To account for spatial dependency, we assume that the process $(Z_{\mathbf{i}})$ satisfies a mixing condition defined as follows: there exists a function $\varphi(x) \searrow 0$ as $x \to \infty$, such that

$$\begin{aligned} \alpha\left(\sigma\left(S\right), \sigma\left(S'\right)\right) &= \sup\left\{\left|\mathbb{P}(A\cap B) - \mathbb{P}\left(A\right)\mathbb{P}\left(B\right)\right|, \quad A \in \sigma\left(S\right), \quad B \in \sigma\left(S'\right)\right\} \\ &\leq \psi\left(\mathrm{Card}(S), \mathrm{Card}(S')\right) \varphi\left(\mathrm{dist}(S, S')\right) \end{aligned} \tag{1}$$

where S and S' are two finite sets of sites, $\mathrm{Card}(S)$ denotes the cardinality of the set S, $\sigma\left(S\right) = \{Z_{\mathbf{i}}, \mathbf{i} \in S\}$ and $\sigma\left(S'\right) = \{Z_{\mathbf{i}}, \mathbf{i} \in S'\}$ are σ-fields generated by $Z_{\mathbf{i}}$, $\mathrm{dist}(S, S')$ is the Euclidean distance between S and S', and $\psi(\cdot)$ is a positive symmetric function nondecreasing in each variable. We recall that the process $(Z_{\mathbf{i}})$ is said to be strongly mixing if $\psi \equiv 1$. As usual, we will assume that one of both conditions on $\varphi(i)$ is verified:

$$\varphi\left(i\right) \leq Ci^{-\theta}, \qquad \text{for some } \theta > 0$$

i.e. that $\varphi(i)$ tends to zero at a polynomial rate, or

$$\varphi(i) \leq C \exp(-si), \qquad \text{for some } s > 0$$

i.e. that $\varphi(i)$ tends to zero at an exponential rate. Then, to address the mixing condition, we impose that for u large, $K_1(u)$ is asymptotically exponential or polynomial. These conditions are satisfied, for instance, by several kernels with compact support such as triangular (Bartlett), Epanechnikov, Parzen kernels. For all compact support kernels, both conditions are verified at least asymptotically. That is sufficient to ensure the control of the mixing condition. Concerning the function $\varphi(\cdot)$, for the sake of simplicity, we will only study the case where $\varphi(\cdot)$ tends to zero at a polynomial rate, that is:

$$\varphi(t) \leq Ct^{-\theta} \qquad , \ \theta > 0 \, , \, t \in \mathbb{R}_+^*. \tag{2}$$

Before stating the main results, the following set of assumptions are listed and all along the paper, when no confusion is possible, we will denote by C strictly positive generic constant.

(H1) Functions f and $r(\cdot)$ are continuous at $x_{\mathbf{s}}$, f is lipschitzian in a neighbor of $x_{\mathbf{s}}$: $|f(x_{\mathbf{s}}) - f(y)| \leq C\|x_{\mathbf{s}} - y\|$, for all y in a neighbor of $x_{\mathbf{s}}$. In addition, $f(x_{\mathbf{s}}) > 0$.

(H2) The density $f_{X_{\mathbf{i}}X_{\mathbf{j}}}$ of $(X_{\mathbf{i}}, X_{\mathbf{j}})$ is bounded and $\left|f_{X_{\mathbf{i}}X_{\mathbf{j}}}(u,v) - f(u)f(v)\right| \leq C$ for all $\mathbf{i} \neq \mathbf{j}$ and (u,v) in a neighbor of $(x_{\mathbf{s}}, x_{\mathbf{s}})$.

(H3) $k(\mathbf{n}) \sim \hat{\mathbf{n}}^{\gamma}$ and $k_{\mathbf{n}}^{1} \sim \hat{\mathbf{n}}^{\tilde{\gamma}}$, where γ, $\tilde{\gamma} \in]0,1[$, $\gamma < \tilde{\gamma}$, $\gamma + \tilde{\gamma} > 1$ and $\tilde{\gamma} - \gamma < 1/(\gamma + \tilde{\gamma})$.

(H4) (i) K_1 is bounded nonnegative function satisfying

$$C_1 \mathbb{I}_{[0,1]}(t) \leq K_1(t) \leq C_2 \mathbb{I}_{[0,1]}(t) \, , \qquad for \ t \in \mathbb{R}. \tag{3}$$

where C_1 and C_2 are positive numbers.

(ii) The kernel K_2 is bounded, of compact support and:

$$\forall u \in \mathbb{R}^d, \ K_2(u) \leq K_2(tu) \quad \forall t \in]0,1[. \tag{4}$$

(H5) $\forall n, m \in \mathbb{N} \quad \psi(n,m) \leq C \min(n,m)$ and $\theta > 2N(\gamma + \tilde{\gamma})/(\gamma + \tilde{\gamma} - 1)$.

(H6) $\forall n, m \in \mathbb{N} \quad \psi(n,m) \leq C(n+m+1)^{\tilde{\beta}}$, $\tilde{\beta} \geq 1$ and $\theta > N(\gamma + \tilde{\gamma} + 1 + 2\tilde{\beta})/(\gamma + \tilde{\gamma} - 1)$.

(H7) $r(\cdot)$ is lipschitzian in a neighbor of $x_{\mathbf{s}}$.

Asymptotic results

In this section, we establish in Theorem 1 the almost complete (a.co.) convergence of the estimator r_{kNN} if $r(\cdot)$ is continuous. When $r(\cdot)$ is Lipschtiz function, we obtain the rate of this convergence (see Theorem 2).

Theorem 1 *Under assumptions (H1)-(H4) and (H5) or (H6), we have*

$$|r_{\mathrm{kNN}}(x_{\mathbf{s}}) - r(x_{\mathbf{s}})| \underset{\mathbf{n}\to\infty}{\longrightarrow} 0 \quad a.co. \tag{5}$$

Theorem 2 *Under assumptions (H1)-(H4), (H7) and (H5) or (H6), as* $\mathbf{n} \to \infty$*, we have*

$$r_{\mathrm{kNN}}(x_{\mathbf{s}}) - r(x_{\mathbf{s}}) = \mathcal{O}\left(\left(\frac{k(\mathbf{n})}{\hat{\mathbf{n}}}\right)^{1/d} + \left(\frac{\hat{\mathbf{n}}\log(\hat{\mathbf{n}})}{k_{\mathbf{n}}^{1}k(\mathbf{n})}\right)^{1/2}\right) \quad a.co. \tag{6}$$

For the proofs of Theorems 1 and 2, the difficulty comes from randomness of the window $H_{\mathbf{n},x_{\mathbf{s}}}$. So, we do not have in the numerator and in denominator of $r_{\mathrm{kNN}}(x_{\mathbf{s}})$ sums of independents variables. The idea is to frame $H_{\mathbf{n},x_{\mathbf{s}}}$ by two non-random windows. More generally, these technical tools could be useful as long as one has to deal with random bandwidths.
We also provide some numerical results comparing the performance of our estimator and the kernel one, towards some simulated but also real data.

Bibliographie

[1] Biau, G. and Cadre, B. (2004). *Nonparametric spatial prediction.* Statistical Inference for Stochastic Processes, **7**, 327–349.
[2] Burba, F., Ferraty, F. and Vieu, P. (2009). *k-Nearest Neighbour method in functional nonparametric regression.* Journal of Nonparametric Statistics , **21** (4), 453–469.
[3] Collomb, G. (1980). *Estimation de la régression par la méthode des k points les plus proches avec noyau: quelques propriétés de convergence ponctuelle*, Lecture Notes in Mathematics. Springer, **821**, 159-175.
[4] Dabo-Niang, S. ,Ternynck, C. and Yao, A.-F.(2014). *Spatial regression in the multivariate context : a new kernel estimator considering spatial dependency*, Submitted.
[5] Wang, H. and Wang, J. (2009). *Estimation of the trend function for spatio-temporal models*, Journal of Nonparametric Statistics, **21** (5), 567–588.