

Predicting temperature curve based on fast k NN local linear estimation of the conditional distribution function

Ibrahim M. Almanjahie^{1,2}, Zoulikha Kaid^{1,2}, Ali Laksaci^{1,2} and Mustapha Rachdi³

¹ Department of Mathematics, College of Science, King Khalid University, Abha, South Region, Saudi Arabia

² Statistical Research and Studies Support Unit, King Khalid University, Abha, South Region, Saudi Arabia

³ AGIM Team, Laboratoire AGEIS, Université Grenoble Alpes (France), Université Grenoble Alpes, Grenoble, France

ABSTRACT

Predicting the yearly curve of the temperature, based on meteorological data, is essential for understanding the impact of climate change on humans and the environment. The standard statistical models based on the big data discretization in the finite grid suffer from certain drawbacks such as dimensionality when the size of the data is large. We consider, in this paper, the predictive region problem in functional time series analysis. We study the prediction by the shortest conditional modal interval constructed by the local linear estimation of the cumulative function of Y given functional input variable X . More precisely, we combine the k -Nearest Neighbors procedure to the local linear algorithm to construct two estimators of the conditional distribution function. The main purpose of this paper is to compare, by a simulation study, the efficiency of the two estimators concerning the level of dependence. The feasibility of these estimators in the functional times series prediction is examined at the end of this paper. More precisely, we compare the shortest conditional modal interval predictive regions of both estimators using real meteorological data.

Submitted 4 March 2021

Accepted 13 June 2021

Published 9 July 2021

Corresponding author

Ibrahim M. Almanjahie,
imalmanjahi@kku.edu.sa

Academic editor

Jianhua Xu

Additional Information and
Declarations can be found on
page 13

DOI [10.7717/peerj.11719](https://doi.org/10.7717/peerj.11719)

© Copyright

2021 Almanjahie et al.

Distributed under

Creative Commons CC-BY 4.0

OPEN ACCESS

Subjects Statistics, Environmental Impacts

Keywords Functional time series, Meteorological data, Local linear fitting, Distribution function, Kernel weighting, Conditional predictive region, k -nearest neighbors smoothing

INTRODUCTION

Over the past decades, the volume and complexity of collected data have rapidly increased the sizes and number of covariates. This increase has created significant potential and demand for scientific and technological innovations. The increased storage capacity of information, the improvement of computers and their processing capabilities, the proliferation of surveillance systems, and the improved sensors are the technological progress that has favored the emergence of this kind of data. These are now commonly used in many fields of application such as astronomy, biology, climatology, ecology, chemistry, economics, medicine, engineering sciences, etc. The standard statistical models based on the discretization of the big data in the finite grid suffer from certain drawbacks.

The first anomaly is the problem of the dimensionality curse when the size of the discretization grid is large. The second anomaly is that, with the transformation, the original data loses its characteristics such as the functionality, correlation, heteroscedasticity or the homoscedasticity of the data. In particular, the asymptotic behavior of the constructed data is related to the obtained sampling.

Recall that all the mentioned defects can substantially affect the multivariate approaches' efficiency in big data analysis. To overcome these problems, a new approach in modern statistics called functional data analysis is developed recently. Such procedure allows the use of the natural space of the data, which permits the profit from the whole information. In this modern statistics branch, the local linearity method (LLM) estimation is widely studied. It is motivated by its small bias in the estimation processing (see [Fan & Gijbels \(1996\)](#) for a uni-dimensional framework, and [Baïllo & Grané \(2009\)](#), for the Nonparametric Functional Data Analysis (NPFDM) set up in the area of functional statistics). [Barrientos-Marin, Ferraty & Vieu \(2010\)](#) studied the LLM-estimation of the nonparametric operator of the Banach explanatory variable. We cite [Berlinet, Elamine & Mas \(2011\)](#) for an alternative LLM-estimator constructed by inverting the local variance-covariance matrix of the functional variable. Concerning the LLM-estimation of the conditional cumulative distribution function (CCDF), we point out that the first result was stated by [Laksaci, Rachdi & Rahmani \(2013\)](#). Later, [Demongeot et al. \(2014\)](#) precise the least square error of the LLM-estimator of the CCDF-model. We also point out that the previous studies utilized the kernel local linearity method. However, this paper focuses on CCDF-estimation with a new weighting approach obtained by mixing the local linear fitting to the k -Nearest Neighbors (k NN) method. Indeed, the method of k NN has been received growing attention in nonparametric functional statistics. In particular, the first results on the k NN-LLM estimation were obtained by [Chikr-Elmezouar et al. \(2019\)](#). They studied the conditional density estimation using the local linear method under the k NN smoothing. They established the almost complete consistency of the obtained estimator. Recently, [Almanjahie et al. \(2021\)](#) consider the estimation of the robust regression function using the k NN. They proved the uniform consistency on the number of the neighbor of the constructed estimator. We also mention [Laksaci, Ould-Said & Rachdi \(2021\)](#) for the k NN estimation of the quantile regression. For more recent results on the k NN-LLM estimation in functional statistics, we cite [Rachdi et al. \(2021\)](#). They treated the case when the response variable is observed with missing at random and the regressor is of functional nature.

On the other hand, the estimation based on the k NN method has more advantages than the Nadaraya–Watson algorithm (see [Burba, Ferraty & Vieu \(2009\)](#), for more discussion on the motivations on this approach). In this paper, we benefit from the advantages of both the k NN weighting and LLM-fitting by combining the two algorithms to provide a fast efficiency estimator for the CCDF. Specifically, we combine these approaches to construct two estimates of the CCDF and to the shortest conditional modal interval (SCMI) predictive regions using mixing functional time series. Notice also that the

smoothing parameter in the k -Nearest Neighbors method is varied randomly with respect to the observations. This feature makes the applicability of these estimators very fast and accurate because the smoothing parameter is selected with respect to the nature of the data. To highlight the smoothing parameter selection issue, we compare several cross-validation rules to select the best number of the neighborhood. The superiority of both estimators in practice is emphasized by the meteorological data; specifically, the constructed estimator to predict the yearly curve of the temperature in Europe central.

This paper is structured as follows. The two kNN-LLM estimators of the CCDF are constructed in the “Method” section. We devote the section “Results and discussions” to some discussions related to our proposed estimators. The simulation study for testing the superiority of the proposed estimators is presented in the “Simulation study” section. The performance of the constructed estimator in temperature prediction using real data is conducted in the “Real-data application” section. Our conclusion and remarks for further research are presented in the last section.

METHODS

In this section, we will construct two new estimators by combining the local linear approach to the kNN smoothing methods.

The fast kNN-LLM of the CCDF

Consider $(X_1, Y_1), (X_2, Y_2), (X_3, Y_3), \dots, (X_n, Y_n)$ be stationary sequence of random vector (X, Y) valued in $\mathcal{F} \times \mathbb{R}$, where \mathcal{F} is a separable metric and has a metric d . Let N_x be the neighborhood of fixed curve $x \in \mathcal{F}$, for which we suppose that the conditional cumulative distribution function (CCDF) $F(\cdot|x)$ has a continuous conditional density $f(\cdot|x)$.

This estimation procedure is based on the definition of CCDF, as conditional expectation:

$$F(y|x) = \mathbb{E}[\mathbb{I}_{\{Y_i \leq y\}} | X = x]$$

with \mathbb{I}_A is the indicator function of A . In the LLM technique, we approximate $F(y|x)$ locally in N_x using

$$\forall x_0 \in N_x, \quad F(y|x_0) = a_{yx} + b_{yx}d(x_0, x) + o(d(x, x_0)). \quad (1)$$

So, the kNN-LLM estimator of CCDF $F(y|x)$ is obtained by estimating a_{yx} and b_{yx} in (1), as the minimizers of the rule,

$$\text{Min}_{a,b} \sum_{i=1}^n (\mathbb{I}_{\{Y_i \leq y\}} - a - bd(X_i, x))^2 \text{Ker}\left(\frac{d(x, X_i)}{\mathbb{H}_k}\right),$$

where Ker means the kernel function and $\mathbb{H} = \min\{h \in \mathbb{R}^+, \text{ satisfies } \sum_{i=1}^n \mathbb{I}_{\text{Ba}(x,h)}(X_i) = k\}$. Similarly to [Barrientos-Marin, Ferraty, & Vieu \(2010\)](#), we obtain by derivative that the \hat{a} and \hat{b} are the solutions of

$${}^tL(Ker\Upsilon - KerL)\begin{pmatrix} \widehat{a} \\ \widehat{b} \end{pmatrix} = 0,$$

where

$${}^tL = \begin{pmatrix} 1, 1, \dots, 1 \\ d(X_1, x) \dots, d(X_n, x) \end{pmatrix}$$

and

$$KER = \text{diag}\left(Ker\left(\frac{d(x, X_1)}{\mathbb{H}_k}\right), Ker\left(\frac{d(x, X_2)}{\mathbb{H}_k}\right), \dots, Ker\left(\frac{d(x, X_n)}{\mathbb{H}_k}\right)\right)$$

with

$$\text{and } {}^t\Upsilon = (\mathbb{1}_{\{Y_1 \leq y\}}, \dots, \mathbb{1}_{\{Y_n \leq y\}}).$$

It follows that

$$\begin{pmatrix} \widehat{a} \\ \widehat{b} \end{pmatrix} = ({}^tLKERL)^{-1}({}^tLKER\Upsilon).$$

Hence,

$$\widehat{a} = (1, 0)({}^tLKERL)^{-1}({}^tLKER\Upsilon).$$

Finally, the Fast kNN-LLM of the CCDF $F(y|x)$ is

$$\widehat{F}(y|x) = \widehat{a}_{yx} = \frac{\sum_{i,j=1}^n \beta_{ij} \mathbb{1}_{\{Y_i \leq y\}}}{\sum_{i,j=1}^n \beta_{ij}},$$

where

$$\beta_{ij} = d(X_i, x)(d(X_i, x) - d(X_j, x)) \times Ker(\mathbb{H}_k^{-1}d(\mathbb{H}X_i))Ker(\mathbb{H}_k^{-1}d(x, X_j)).$$

The smooth kNN-LLM of the CCDF

An alternative estimation of CCDF is built by treating the function $F(\cdot|x)$ as a conditional expectation, i.e.,

$$\mathbb{E}[H(\ell^{-1}(y - Y_i))|X_i = x] \rightarrow F(y|x) \text{ as } \ell \rightarrow 0,$$

where H is the cumulative distribution function, $(\ell_n = \ell)$ is a positive real sequence.

In fact, this idea was proposed, first, by [Fan & Gijbels \(1996\)](#) in nonfunctional setup. Under this consideration, our motivation is based on the fact that the smooth kNN-LLM of the CCDF is obtained by estimating the operators a_{yx} and b_{yx} of the formula (1) as

$$\text{Min}_{(a,b) \in \mathbb{R}^2} \sum_{i=1}^n (H(\ell_l^{-1}(y - Y_i)) - a - bd(X_i, x))^2 Ker\left(\frac{d(x, X_i)}{\mathbb{H}_k}\right),$$

where $\ell_l = \min\{\ell \in \mathbb{R}^+, \text{ satisfies } \sum_{i=1}^n \mathbf{1}_{(y-\ell, y+\ell)}(Y_i) = l\}$. Then, we prove that the smooth kNN-LLM of the CCDF $F(y|x)$ is explicited by

$$\widehat{F}(y|x) = \frac{\sum_{i,j=1}^n \beta_{ij} H(\ell_l^{-1}(y - Y_i))}{\sum_{i,j=1}^n \beta_{ij}}.$$

RESULTS AND DISCUSSIONS

On the impact of this contribution

It is well known that the CCDF has pivotal role in nonparametric statistics modeling. Indeed, the nonparametric estimation of this model is an imperative step for several nonparametric model including conditional density, the conditional quantile functions and the conditional hazard. In the prediction setting, the CCDF allows constructing various predictive regions or, more specifically, predictive intervals. We mention for instance, the shortest conditional modal interval (SCMI), the conditional percentile interval and the maximum conditional density region (MCDR) (see [De Gooijer & Gannoun \(2000\)](#) for their definitions). Of course, the diversity of the applicability of CCDF highlights the importance of this conditional model, which has the power of characterizing, completely, the conditional law of the considered random variables. As mentioned in the bibliographical discussion of the introduction section, the CCDF model has been widely studied in NPFDM. However, our present work's novelty mainly estimates the CCDF model based on the combination of two fundamental approaches: the kNN and the LLM. This combination allows to construct an attractive estimator allowing to inherits the advantages of two methods. Indeed, it is well known that the LLM improves the bias property of the CKM while the weighting by the kNN-algorithm offers a sophisticated procedure for the smoothing parameter choose. It is selected locally with respect to the vicinity at the point of conditioning which is more adaptive to the data topological structure. Such adaptation is essential in nonparametric functional data analysis, where our estimators' efficiency is connected to the data structure explored through the concentration property of the probability measure of the functional variable (see [Ferraty & Vieu \(2006\)](#)). Nevertheless, the establishment of the convergence rate of the kNN-LLM estimators is more complicated than the case considered by [Laksaci, Rachdi & Rahmani \(2013\)](#). In our case, the smoothing parameter is taken to be a random variable, while it is a scalar in the classical situation. Considering the dependent case which is more general and more realistic situation this difficulty becomes more complicated. In conclusion, the principal axes of this contribution are: (1) the conditional distribution function as a pivotal model for various nonparametric conditional models, (2) the estimation method as a new proceder even in the nonfunctional case (as far as we know, there is no work in the CCDF estimation by combining the LLM to kNN) and (3) the functional time series case as a generalization for the independent case. To emphasize the usefulness of the present contribution in the prediction issue, we discuss in the following section how we can predict real future characteristic of a continuous-time process given its past.

Functional time series prediction

Recall that the nonparametric prediction is considered to be the most important application of the functional nonparametric data analysis. In particular, functional time series examples can be composed based on a continuous-time process. Indeed, consider a random variable (S_t) where $t \in [0, b)$ having real-values in a continuous-time process. So, from S_t we compose n functional random variables $(X_i)_{i=1, \dots, n}$ obtained by

$$\forall t \in [0, b), \quad X_i(t) = S_{n^{-1}((i-1)b+t)}.$$

Therefore, if our aim is to predict a future value $Y = S_{t_0}$, at fixed point $t_0 = b + s$ given $(S_t)_{t \in [0, b)}$, we then define a sequence of the interest variable Y , i.e.,

$$Y_i = S_{n^{-1}(i)b+s}, \quad i = 1, \dots, n.$$

Thereafter, we construct our predictor (conditional median, conditional quantile or the conditional mode) by using the observations $(X_i, Y_i)_{i=1, \dots, n-1}$. However, since the predictive region or, more specifically, the predictive interval is often more instructive than predicting a single-point, we focus on this kind of prediction. Formally, for all $\zeta \in (0, 1)$, the interval/region is defined as a set $I_\zeta \subset \mathbb{R}$ satisfies

$$\mathbb{P}(Y_n \in I_\zeta | X_n) = 1 - \zeta.$$

As mentioned in the above section, one of the main feature of the CCDF is the possibility to construct several predictive regions I_ζ . Of course, the efficiency of each prediction interval is assessed by the means of the length of the set I_ζ and the presence of the true value in I_ζ . It is well documented that the width of the SCMI is the smallest compared to all predictive regions with the same coverage probability (see [De Gooijer & Gannoun \(2000\)](#)). The latter is introduced by [Tong & Yao \(1995\)](#) and obtained by

$$[A_{1-\zeta}, B_{1-\zeta}] = \arg \min_{c,d} \{Leb[c, d] \mid (d|x) - F(c|x) \geq 1 - \zeta\}.$$

The $Leb(\cdot)$ refers to the Lebesgue measure. Using the CCFD estimators, we approximate the SCMI by

$$[A_{1-\zeta}(X_n), B_{1-\zeta}(X_n)] = \arg \min_{c,d} \{Leb[c, d] \mid \ddot{F}(d|X_n) - \ddot{F}(c|X_n) \geq 1 - \zeta\}.$$

where \ddot{F} means \widehat{F} or \widetilde{F} . The easy implementation of this approximation is studied and discussed in the “REAL-DATA APPLICATION” section.

SIMULATION STUDY

In this simulation study, we propose to control the behaviour of the estimators w.r.t. the dependency degree of the data. More precisely, our aim is to compare, considering finite sample, the efficiency of the estimators $\widehat{F}(y|x)$ and $\widetilde{F}(y|x)$. To do this, we use the fact that m -dependent variables are α -mixing and we generate n functional variables as follows.

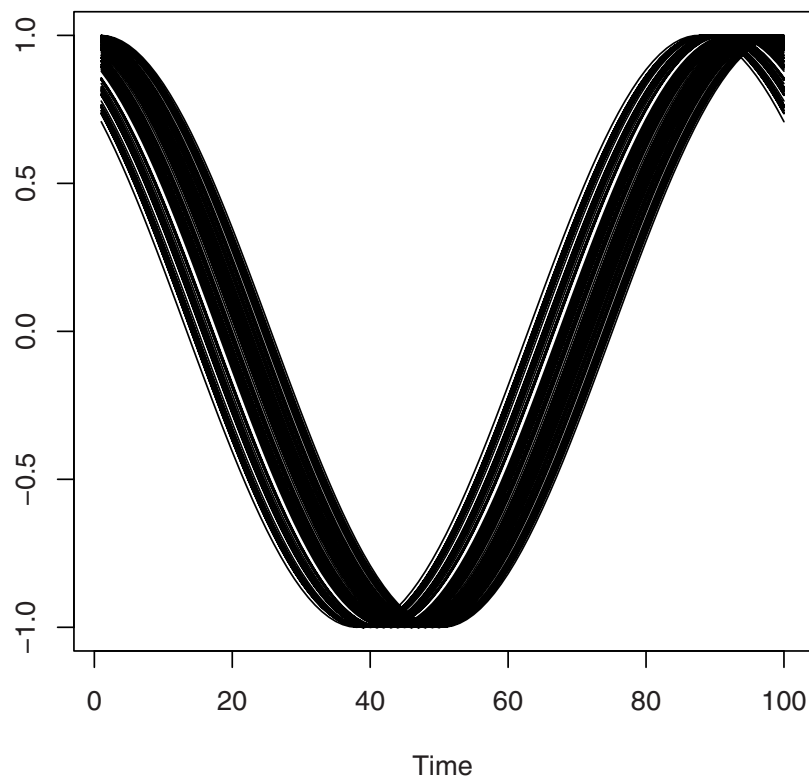



Figure 1 A sample of 100 curves.

Full-size  DOI: [10.7717/peerj.11719/fig-1](https://doi.org/10.7717/peerj.11719/fig-1)

In the first, we draw $n + m - 1$ independent functional variables by

$$S_j(t) = 2 \cos(tW_{1j}) + 0.2W_{2j}, \text{ for } j = 1, \dots, n + m - 1,$$

where W_1 and W_2 are uniformly distributed on $[0, \pi/4]$. Next, we simulate the m -dependent functional variables defined by:

$$X_i(t) = \sum_{j=i}^{i+m-1} S_j(t).$$

Discretizing the curves X_i 's on the same grid leads to the construction of 100 equispaced measurements in $(0, 2\pi)$. In Fig. 1, we plotted the functional variables associated to the strong case where $m = 4$.

For the response variable we consider four regression models:

$$\text{Model M1} : Y_i = 5 \int_0^{2\pi} \log((4 - X_i(t))^2 + 2) dt + \varepsilon_i,$$

$$\text{Model M2} : Y_i = \int_0^{2\pi} \exp(-X_i(t)) dt + 1.5 \int_0^{2\pi} \exp(X_i(t)) dt + \varepsilon_i,$$

$$\text{Model M3 : } Y_i = \int_0^{2\pi} (5 \log((4 - X_i(t))^2 + 2)) dt + \int_0^{2\pi} \exp(-X_i(t)) dt \\ + 1.5 \delta_i \int_0^{2\pi} \exp(X_i(t)) dt,$$

$$\text{Model M4 : } Y_i = \int_0^{2\pi} \exp(-X_i(t)) dt + 1.5 \int_0^{2\pi} \exp(X_i(t)) dt \\ + 5 \delta_i \int_0^{2\pi} \log((4 - X_i(t))^2 + 2) dt,$$

where $\varepsilon_i \sim N(0, 0.25)$ (resp. $\delta_i \sim \text{Exp}(2)$). Note that, based on these models and with given $X = x$, the CCDF of Y is explicitly determined according to the distributions of ε_i and δ_i , which permit the determination of the theoretical CCDF, $F(y|x)$.

Now, we specify quickly the different parameters involved in both estimators. Note that the parameters of the two estimators are the kernel K , the semi-metric d , the number k and/or l of neighbors. So, for this numerical study, we point out that we have taken a quadratic kernel supported within $(0, 1)$ and used the L_2 metric and the numbers of neighbors k, l are chosen using the following cross-validation criterion, defined as

$$\sum_{j=1}^n (F(Y_j|X_j) - \bar{F}^{-j}(Y_j|X_j)),$$

where \bar{F}^{-j} denotes the leave-one-out-curve estimate of the \tilde{F} and \hat{F} .

The performances of the two estimates is examined by comparing their average absolute errors:

$$AE(\tilde{F}) = \frac{1}{n} \sum_{i=1}^n |F(Y_i|X_i) - \tilde{F}(Y_i|X_i)| \text{ and } AE(\hat{F}) = \frac{1}{n} \sum_{i=1}^n |F(Y_i|X_i) - \hat{F}(Y_i|X_i)|.$$

Thus, in order to control behaviors of both estimates w.r.t. the level of dependency, we plotted in Fig. 2, the curves of $AE(\tilde{F})$ and $AE(\hat{F})$ w.r.t. the values of m .

It can be seen that, both errors increase substantially relatively to the values of m . Furthermore, it is clear that in the models M3 and M4 (heteroscedastic case), the estimate \hat{F} outperforms \tilde{F} , but in the models M1 and M2 (homoscedastic case), the estimate \tilde{F} is significantly better than \hat{F} .

REAL-DATA APPLICATION

In this section, we show the applicability of the proposed estimators to a real data example. To do that, we consider the problem of predicting the monthly average temperature one year ahead. For this purpose, we consider the same data set used by *Laksaci, Lemdani & Said (2011)* which are available at the website https://www.met.hu/en/eghajlat/magyarorszag_eghajlata/eghajlati_adatsorok/Debrecen/adatok/napi_adatok/index.php. This data were collected by Debrecen's station, Hungary (northern latitude $47^\circ 35' 44''$ and eastern longitude $21^\circ 38' 43''$). They are monthly measurements (1,200 months = 100 years) from 1901 to 2000. The latter can be viewed as a continuous process denoted by S_t .

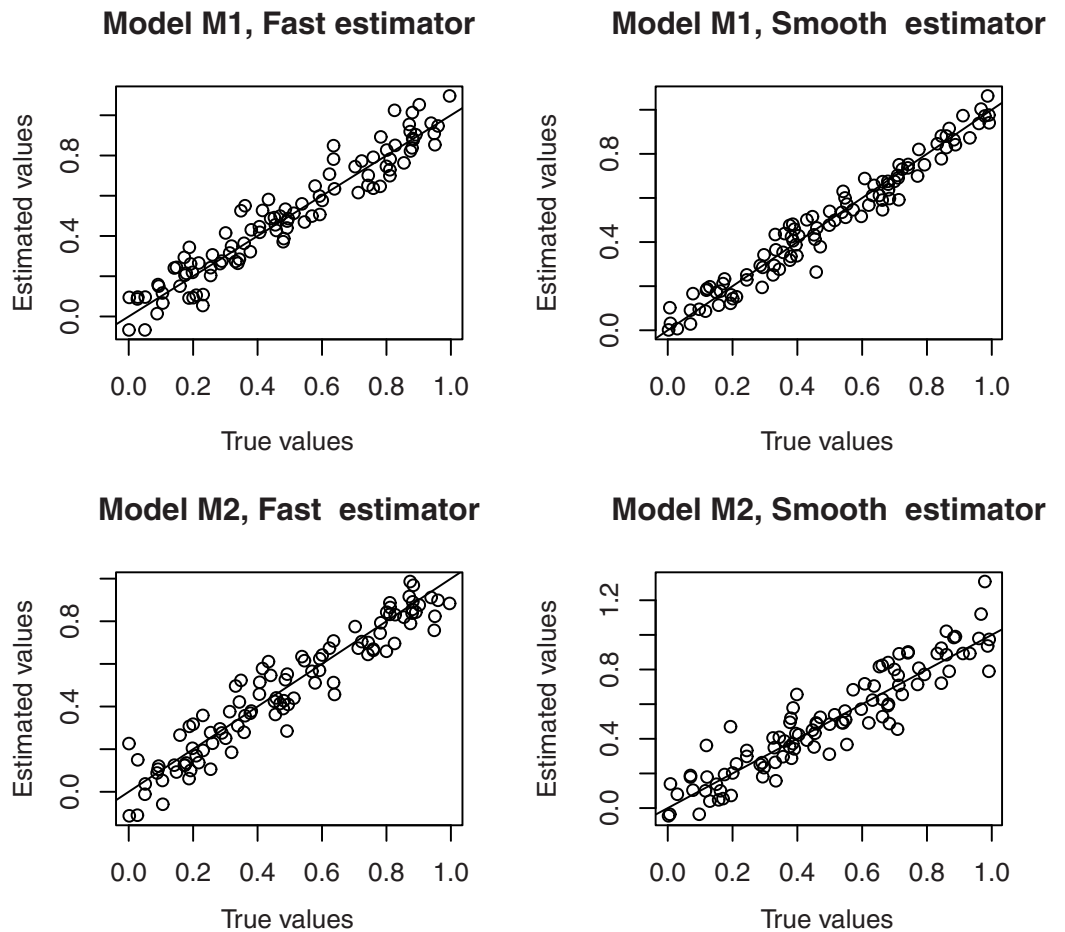


Figure 2 $AE(\tilde{F})$ (dotted line) and $AE(\hat{F})$ (continuous line).

Full-size  DOI: [10.7717/peerj.11719/fig-2](https://doi.org/10.7717/peerj.11719/fig-2)

As noticed in the previous section, from S_b , we construct $n + 1 = 100$ curves $(X_i(t))$, $i = 1, \dots, n + 1$, where X_i denotes the average temperature curve observed during the (1 year) 12 months of the i^{th} year. The process (S_t) and the curves $(X_i)_i$ are plotted in Figs. 3 and 4.

Of course, the proposed predictive interval's efficiency is closely connected to the parameters' choices in the estimator of the conditional distribution function. For the real data example, we compare estimators F and \hat{F} in the SCMI estimation (with $\zeta = 0.1$). For the computational study, we use the same kernel K , that is, the quadratic kernel on $(0,1)$. The latter is adequate with this type of nonparametric approach. It is usually used in nonparametric functional statistics and incorporates the technical assumptions of the theoretical development of this kind of model. Concerning the choice of the metric d , we point out that it is closely related to the nature of the functional variable and its smoothing property. The Principal Component Analysis (PCA) metric is more suitable for this type of discontinuous functional regressors. For the choice of k (or h), we utilize the same

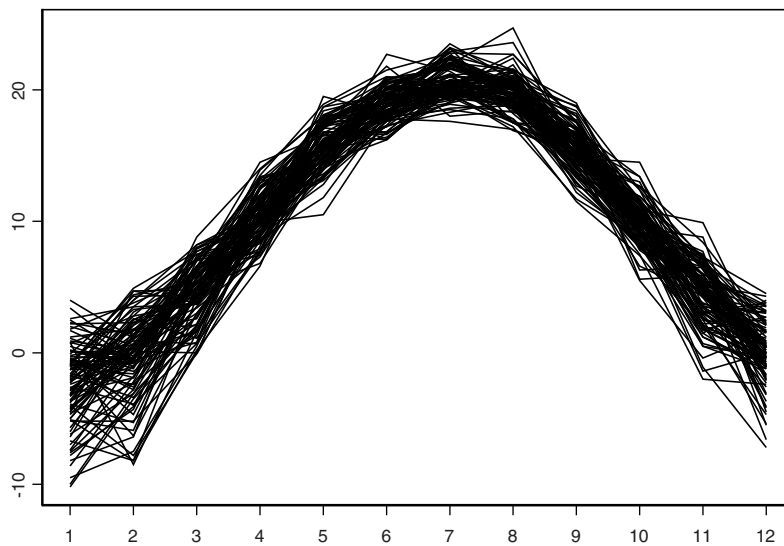


Figure 3 Mean temperature by year.

Full-size  DOI: 10.7717/peerj.11719/fig-3

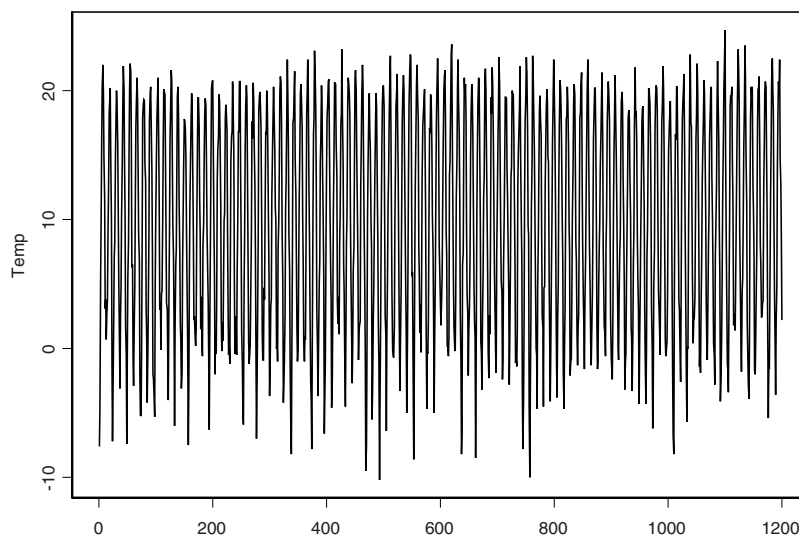


Figure 4 Monthly mean temperature.

Full-size  DOI: 10.7717/peerj.11719/fig-4

cross-validation method used by *De Gooijer & Gannoun (2000)*, which is based on the criterion

$$CV = \frac{1}{n} \sum_{j=1}^n \sum_{i=1}^n (1_{\{Y_i \leq Y_j\}} - \hat{F}^{-j}(Y_i | x_j))^2.$$

This criterion is optimised over the same subsets of k (or h) proposed by *Rachdi & Vieu (2007)*, that is, $\{5, 10, 20, \dots, 45\}$.

To determine the SCMI predictive interval of the whole curve of the last year ($i = 100$) of this sample knowing the functional covariates X_{99} , we use the first 98 curves as a training sample. Then, we predict the CCDF knowing X_{99} by $\hat{F}(\cdot | X_{99}^*)$ and $\tilde{F}(\cdot | X_{99}^*)$ where X_{99}^*

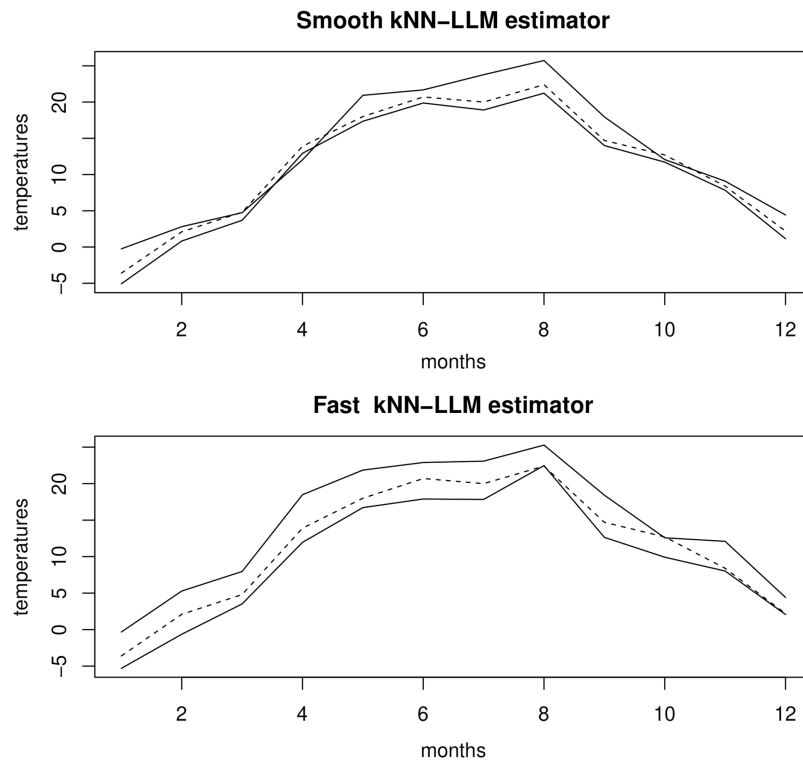


Figure 5 Results obtained by smooth kNN-LLM and fast kNN-LLM estimators.

Full-size  DOI: 10.7717/peerj.11719/fig-5

is the nearest curve to X_{99} in the training sample $(Y_i^j, X_i)_{i=1, \dots, 99}$ with $Y_i^j = X_{i+1}(j)$, for each fixed month $j = 1, \dots, 12$. Figure 5 displays the results. The dashed curve represents the observed data and the solid curves represent the estimated values for the two extremities of the SCMI predictive interval. We observe that the result of the \tilde{F} is significantly better than the \hat{F} one. In the sense that it has an average mean length ($M.L = 1.23$) versus $M.L = 1.97$ for \hat{F} . Of course, this gain is not surprising.

In the second illustration, we emphasize the importance of the kNN-LLM estimation of CCDF on the construction of the SCMI predictive interval by comparing the two kNN-LLM estimators, f and \hat{F} , to their competitive estimators constructed by the kNN-Nadraya-Watson (kNN-NW) method. More precisely, we compare the estimators \tilde{F} and \hat{F} to the smooth ($\tilde{F}_1(y|x)$) and fast ($\hat{F}_1(y|x)$) kNN-NW estimators, where

$$\hat{F}_1(y|x) = \frac{\sum_{i=1}^n \text{Ker}\left(\frac{d(x, X_i)}{H_k}\right) H(\ell_1^{-1}(y - Y_i))}{\sum_i \text{Ker}\left(\frac{d(x, X_i)}{H_k}\right)}$$

and

$$\tilde{F}_1(y|x) = \frac{\sum_{i=1}^n \text{Ker}\left(\frac{d(x, X_i)}{H_k}\right) 1_{\{Y_i \leq y\}}}{\sum_{i=1}^n \text{Ker}\left(\frac{d(x, X_i)}{H_k}\right)}.$$

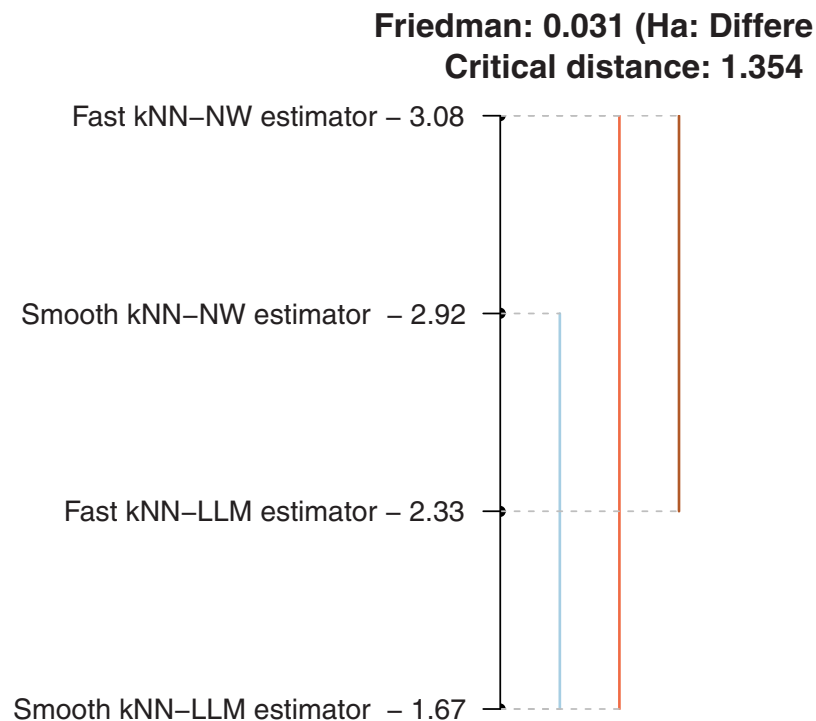


Figure 6 Comparison of the average means of length intervals.

Full-size DOI: [10.7717/peerj.11719/fig-6](https://doi.org/10.7717/peerj.11719/fig-6)

Of course, in order to conduct a comprehensive comparison, we must treat the four models. For this reason, we have used the same kernel, the metric as well as the same selection method of the optimal number of neighbor k . Similar to the first illustration, this comparative analysis is performed over the 12 functional time series $(Y_i^j, X_i)_{i=1\dots 100}$ with $Y_i^j = X_{i+1}(j)$, for each fixed month $j = 1, \dots, 12$. For each fixed j , we split (randomly) the functional time series $(Y_i^j, X_i)_{i=1\dots 100}$ into two parts: the learning sample (70 observations) and the test sample (30 observations). Next, we compute the SCMI predictive intervals for the curves of the testing sample using the estimators \tilde{F} , \hat{F} , \tilde{F}_1 and \hat{F}_1 . The efficiency of the four models is evaluated using the Nemenyi test plots for the average mean of the interval-length (ML). The comparison results are displayed in Fig. 6.

Once again, the conclusion is without surprise. It confirms the statement mentioned in the first illustration. More precisely, the local linear approach is more accurate than the Nadaraya-Watson method. This conclusion emphasizes the superiority of the local linear over the classical kernel method, which has been shown in the multivariate statistics through the bias term. On the other hand, it should be noted that all these four functional approaches have a stratification performance in this context of predictive issues.

CONCLUSION

In this paper, we have studied the nonparametric estimation of the cumulative distribution function of the scalar response variable given a functional explanatory variable. Two new estimators are constructed by combining the local linear approach to the kNN smoothing

methods. The first one is built by a fast algorithm based on the conditional cumulative distribution as classical regression of the indicator function. The second estimator is obtained by integration of the double kernel conditional density estimator. The latter gives a smooth estimator of the conditional cumulative distribution function. An empirical analysis is conducted to compare both estimators and their easy implementation in practice. Both artificial and real data carry out the finite sample performance of the two estimators. Undoubtedly, the present contribution highlights the conditional distribution function's potential impact as a pivotal model in functional data analysis. It is involved in various conational nonparametric models, and its estimation is indispensable as preliminary steps to estimate numerous nonparametric functional models. For instance, we have focused on the prediction problem in this paper, and we have constructed two predictors (single prediction and region predictor). The artificial data analysis shows that both estimators have a satisfactory efficiency in different sinarios of regression data analysis, including homoscedasticity case, heteroscedasticity case, mixture models case, light-tailed or heavy-tailed conditional distribution cases. In conclusion, we can say that the functional data analysis through the conditional distribution has a significant impact in practice, and the proposed estimators of this contribution are fast, very easy to implement, and have a good performance in the prediction issues. Finally, let us note that the present contribution opens several prospects for future research. For example, it will be very interesting to compare the efficiency of our approach to other functional models such as the robust regression and the relative regression. Such models allow to reduce the sensitivity of the kNN approach to the noisy data, missing values, and the presence of outliers.

ACKNOWLEDGEMENTS

The authors would like to thank the reviewers for their valuable comments and suggestions which improved the quality of this article substantially.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This work was supported by the Deanship of Scientific Research at King Khalid University through the Research Groups Program under the grant number R.G.P. 1/64/42.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
King Khalid University: R.G.P. 1/64/42.

Competing Interests

The authors declare that they have no competing interests.

Author Contributions

- Ibrahim M. Almanjahie conceived and analyzed the data, authored or reviewed drafts of the paper, and approved the final draft.
- Zoulikha Kaid conceived and prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.
- Ali Laksaci conceived and analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.
- Mustapha Rachdi conceived and authored or reviewed drafts of the paper, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

The data used in our paper is available at the Hungarian Meteorological Service: https://www.met.hu/en/eghajlat/magyarorszag_eghajlata/eghajlati_adatsorok/Debrecen/adatok/napi_adatok/index.php

The code is available as a [Supplemental File](#).

Supplemental Information

Supplemental information for this article can be found online at <http://dx.doi.org/10.7717/peerj.11719#supplemental-information>.

REFERENCES

- Almanjahie IM, Alahmari WM, Laksaci A, Rachdi M. 2021.** Computational aspects of the k nn local linear smoothing for some conditional models in high dimensional statistics. In: *Communications in Statistics-Simulation and Computation*. Oxfordshire: Taylor & Francis.
- Baillo A, Grané A. 2009.** Local linear regression for functional predictor and scalar response. *Journal of Multivariate Analysis* **100**:102–111.
- Barrientos-Marin J, Ferraty F, Vieu P. 2010.** Locally modelled regression and functional data. *Journal of Nonparametric Statistics* **22**(5):617–632.
- Berlinet A, Elamine A, Mas A. 2011.** Local linear regression for functional data. *Journal of Multivariate Analysis* **63**(5):1047–1075.
- Burba F, Ferraty F, Vieu P. 2009.** k -nearest neighbor method in functional non-parametric regression. *Journal of Nonparametric Statistics* **21**:453–469.
- Chikr-Elmezouar Z, Almanjahie IM, Laksaci A, Rachdi M. 2019.** FDA: strong consistency of the k nn local linear estimation of the functional conditional density and mode. *Journal of Nonparametric Statistics* **31**:175–195.
- De Gooijer A, Gannoun J. 2000.** Nonparametric conditional predictive regions for time series. *Computational Statistics & Data Analysis* **3**(3):259–275.
- Demongeot J, Laksaci A, Rachdi M, Rahmani S. 2014.** On the local linear modelization of the conditional distribution for functional data. *Sankhya A* **76**:328–355.
- Fan J, Gijbels I. 1996.** *Local polynomial modelling and its applications*. London: Chapman & Hall.
- Ferraty F, Vieu P. 2006.** Nonparametric functional data analysis. In: *Theory and Practice*. New York: Springer Series in Statistics.

- Laksaci A, Lemdani M, Said EO. 2011.** Asymptotic results for an l_1 -norm kernel estimator of the conditional quantile for functional dependent data with application to climatology. *Sankhya A* **73(1)**:125–141.
- Laksaci A, Ould-Said E, Rachdi M. 2021.** Uniform consistency in number of neighbours of the knn estimator of the conditional quantile model. In: *Metrika*. Berlin, Germany: Springer.
- Laksaci A, Rachdi M, Rahmani S. 2013.** Spatial modelization: local linear estimation of the conditional distribution for functional data. *Spatial Statistics* **6**:1–23.
- Rachdi M, Vieu P. 2007.** Nonparametric regression for functional data: automatic smoothing parameter selection. *Journal of Statistical Planning and Inference* **137(9)**:2784–2801.
- Rachdi M, Laksaci A, Kaid Z, Benchiha A, Al-Awadhi F. 2021.** k-nearest neighbors local linear regression for functional and missing data at random. *Statistica Neerlandica* **75**:42–65.
- Tong H, Yao Q. 1995.** On initial-condition sensitivity and prediction in nonlinear stochastic systems. *Bulletin of the International Statistical Institute* **IP10.3**:395–412.