Comparative Analysis of Predictive Performance in Nonparametric Functional Regression: A Case Study of Spectrometric Fat Content Prediction

Kurdistan M. Taher Omar[1,*] and Sameera Abdulsalam Othman[2]

[1]*Department of Mathematics, Faculty of Science, University of Zakho, Kurdistan Region, Iraq*

[2]*Department of Mathematics, College of Basic Education, University of Dohuk, Kurdistan Region, Iraq*

Abstract: *Objective*: This research aims to compare two nonparametric functional regression models, the Kernel Model and the K-Nearest Neighbor (KNN) Model, with a focus on predicting scalar responses from functional covariates. Two semi-metrics, one based on second derivatives and the other on Functional Principle Component Analysis, are employed for prediction. The study assesses the accuracy of these models by computing Mean Square Errors (MSE) and provides practical applications for illustration.

Method: The study delves into the realm of nonparametric functional regression, where the response variable (Y) is scalar, and the covariate variable (x) is a function. The Kernel Model, known as funopare.kernel.cv, and the KNN Model, termed funopare.knn.gcv, are used for prediction. The Kernel Model employs automatic bandwidth selection via Cross-Validation, while the KNN Model employs a global smoothing parameter. The performance of both models is evaluated using MSE, considering two different semi-metrics.

Results: The results indicate that the KNN Model outperforms the Kernel Model in terms of prediction accuracy, as supported by the computed MSE. The choice of semi-metric, whether based on second derivatives or Functional Principle Component Analysis, impacts the model's performance. Two real-world applications, Spectrometric Data for predicting fat content and Canadian Weather Station data for predicting precipitation, demonstrate the practicality and utility of the models.

Conclusion: This research provides valuable insights into nonparametric functional regression methods for predicting scalar responses from functional covariates. The KNN Model, when compared to the Kernel Model, offers superior predictive performance. The selection of an appropriate semi-metric is essential for model accuracy. Future research may explore the extension of these models to cases involving multivariate responses and consider interactions between response components.

Keywords: Nonparametric regression, Functional data, Kernel function, Functional covariates, KNN estimator, Semi-metrics.

1. NTRODUCTION

The proposed article deals with the functional nonparametric regression model

$$\mathcal{Y} = m(x) + e,$$

where $\mathcal{Y}$ is a scalar response, x is a function. Recently, functional data analysis has been developed sharply because of the rising number of states coming from different fields of applied science (biometric, chemometric, medicine, environmetrics, etc.), where data are collected as curves. Measuring devices are now more stronger and the quality of collected data is more precise. Therefore, collecting data in a short time that lets us deal with a great data set of variables leads to expanding the finite-dimensional statistical structures to an infinite-dimensional data technique. Therefore, this type of data needs special statistical structures or procedures. Ramsay and Silverman (1997) [1] pointed to statistics of functional data, after that this field of statistics is becoming more popular because there are several applications in nonparametric regression of functional data analysis using curves and images as data. So, several different issues include functional data analysis (for instance, prediction, and classification).

Ramsay and Silverman (2005) and Cao *et al*. (2020) [2, 3] also define the functional data in detail as well as mention several different instances and utilize the linear regression model and multiple regression method (parametric model) for predicting scalar response from the functional covariate. Cardot *et al*. (2003) [4] discussed also the linear method for regression with functional data analysis. While consecrating status projects and applied studies decreasing from collaborative study to explicate how functional data plans work out in application are discussed by Ramsay and Silverman (2002) [5].

In the 1960s, freemodelling was the idea of Mahalandobis (1961) [6], who studied a regression analysis. Then, nonparametric regression statistics have been expanded in multivariate model and functional data methods. Therefore, initially, statistical studies, and nonparametric functional prediction issues have been studied both in real data and multivariate data. The work of Nadaraya (1964) and Watson (1964) [7, 8] fixed these ideas.

*Address correspondence to this author at the Department of Mathematics, Faculty of Science, University of Zakho, Kurdistan Region, Iraq;
E-mail: kurdistan.taher@uoz.edu.krd

There is a lot of a good manner about nonparametric statistics, so that [9] studies the applied nonparametric regression. However, only classical framework, real and multivariate, is of concern. The local weighting models in the finite-dimensional data are extremely popular in the society of non-parameterizations, particularly the kernel, particularly the kernel. Ferraty and Vieu (2006) [10], Omar and Wang (2019) [11], and Midi *et al*. (2021) [12] mention some fundamental statistical concepts on local weighting techniques and expansion to the functional data analysis as well kernel weighting in special regard. There are different types of kernel methods with a lot of automatic choices of bandwidth (smoothing parameter).

So, It is well known that the bandwidth chosen is a decisive point in nonparametric prediction kernel regression function according to (Ferraty and Vieu (2003, 2006), Ferraty *et al*. (2007), Rachdi and Vieu (2007) and Ismaeel *et al*. (2022)) [10, 13, 14, 15, 16] who propose an automatic data-driven operation for selecting this parameter as well as the provision of theoretical support to the Functional Cross-Validation study of bandwidth chosen.

According to Ferraty and Vieu (2006) and Doori (2019) [10 17], the functional characteristic of free-parameter comes from the nature of the subject to be predicted (for example functional density, and functional regression), which is not supposed to be parametrizable by a finite number of real qualities.

Consequently, Ramsay and Silverman (2005) [2] have performed studies in functional data analysis that extend to nonparametric functional data analysis, one of that studies is when the output is real value and the explanatory is a function by (Ferraty and Vieu (2003, 2004, 2006), Ferraty *et al*. (2007), Burba *et al*. (2009), and Midi and Ismaeel (2018)) [10, 13, 14, 18, 19, 20]. Nevertheless, the study of prediction function and scalar response developed by Baíllo and Grané (2009) [21] introduced a new nonparametric regression method in the context of functional data, proposing a local linear regression estimator that can be compared with a Nadaraya-Watson type Kernel Regression by Monto Carlo study.

The main purpose of this article is to compare two methods Kernel function and the K-Nearest Neighbor Method in nonparametric functional data and use two types of semi-metrics for both of them the first built on the second derivatives and the second type on the functional principle component analysis. Then, Spectrometric data and the Canadian Weather Station are two examples for applying models and then comparing results.

The rest of the article is arranged as follows. Section 2 points to the nonparametric functional regression (Kernel method and K-Nearest Neighbor Model). Two instances of real data are presented in Section 3. Section 4, includes the conclusion.

2. METHODOLOGY

The general shape of the nonparametric functional regression is:

$$\mathcal{Y} = m(x_i) + \varepsilon_i, \quad i = 1, \ldots, n,$$

where $\mathcal{Y}$ is a response variable, x is a covariate function and ε_i is independent random error.

This paper considers on two models to predict the different regression factors: the first one is built on the kernel model and is called *funopare.kernel.cv,* and the second one utilizes the KNN model and is called *funopare.knn.gcv*. The first one utilizes an automatic bandwidth selection by Cross-Validation structure. The main purpose is to compute the formula:

$$m_{CV}^{kernel} = \frac{\sum_{i=1}^{n} y_i \, K(d_q(\mathbf{x}_i, \mathbf{x})/h_{opt})}{\sum_{i=1}^{n} K\,(d_q(\mathbf{x}_i, \mathbf{x})/h_{opt})},$$

where k is the functional kernel and h_{opt} is a bandwidth of the kernel method is gotten by a Cross-Validation structure:

$$h_{opt} = arg\min_{h} CV(h),$$

where

$$CV(h) = \sum_{i=1}^{n} \left(y_i - m_{(-i)}^{kernel}(\mathbf{x}_i)\right)^2,$$

with

$$m_{(-i)}^{kernel}(\mathbf{x}) = \frac{\sum_{j=1, j\neq i}^{n} y_j \, K\left(d_q(\mathbf{x}_j, \mathbf{x})/h\right)}{\sum_{j=1, j\neq i}^{n} K\,(d_q(\mathbf{x}_j, \mathbf{x})/h}).$$

The semi-metric $d_q(.,.)$ and the functional kernel $K(.)$ have been fixed by the partitioner by (Ferraty and Vieu (2004, 2006), and Burba *et al*. (2009)) [10, 18, 19].

The KNN model uses a global smoothing parameter, and the main aim is to calculate the quantity:

$$m_{GCV}^{KNN}(\mathbf{x}) = \frac{\sum_{i=1}^{n} y_i \, K\left(\mathbf{d}_q(\mathbf{x}_i, \mathbf{x})/h_{kopt}(x)\right)}{\sum_{i=1}^{n} K\left(\mathbf{d}_q(\mathbf{x}_i, \mathbf{x})/h_{kopt}(x)\right)},$$

Where $h_{kopt}(x)$ is the bandwidth corresponding to the optimal number of neighbours obtained by a Cross-Validation method:

$$h_{kopt} = arg\min_{h} GCV(k),$$

where

$$GCV(k) = \sum_{i=1}^{n} (y_i - m_{-i}^{KNN}(\mathbf{x}_i))^2$$

with

$$m_{-i}^{KNN}(\mathbf{x}) = \frac{\sum_{j=1, j\neq i}^{n} y_i \, K\left(\mathbf{d}_q(\mathbf{x}_j, \mathbf{x})/h_k(x)\right)}{\sum_{j=1, j\neq i}^{n} K\left(\mathbf{d}_q(\mathbf{x}_j, \mathbf{x})/h_k(x)\right)}.$$

The same number of neighbors used at any curve supplies a global selection. setting the semi-metric $(\mathbf{d}_q(.,.))$ and the functional kernel $K(.)$ (see for example, Ferrety and Vieu (2004, 2006), and Burba *et al*. (2009)) [10, 18, 19].

3. APPLICATIONS

This section is devoted to the applications of the nonparametric functional prediction model based on regression. The section contains two different examples of prediction scalar response values from functional covariates. The first is the pre-diction of a percentage of fat content in the pieces of meat from the spectrometric curves. The second is the prediction of precipitation over one year from the mean monthly temperatures for 35 Canadian Weather Stations.

Example 1

Spectrometric Data: Predicting Fat Content from Spectrometric Data by Two Different Methods (KNN-Model and Kernel Model):

This instance is practical among the nonparametricians's community because Tecator data was a starting point for improving the nonparametric functional data, and then several implementations have been done on this kind of data by different methods (see for example, Ferraty and Vieu (2003, 2006), Ferraty *et al*. (2007), and Burba *et al*. (2009)) [10, 13, 14, 19]. A brief characterization of Tecator data comes from the quality control issue and can be found at http://lib.stat.cmu.edu/datasets/tecator.

For every meat sample, the data contain a 100 channel Spectrum of absorbance, where the absorbance is the − log10 of the transmittance measured by the spectrometer and the plot is the absorbance versus wavelength. The wavelength ranges between 850-1050 nm, and more detail about this shape of data can be found in the book published by Ferraty and Vieu (2006) [10].

Figure **1** also shows that each unit is clearly as a discretized curve, which can see each unit as a persistent curve.

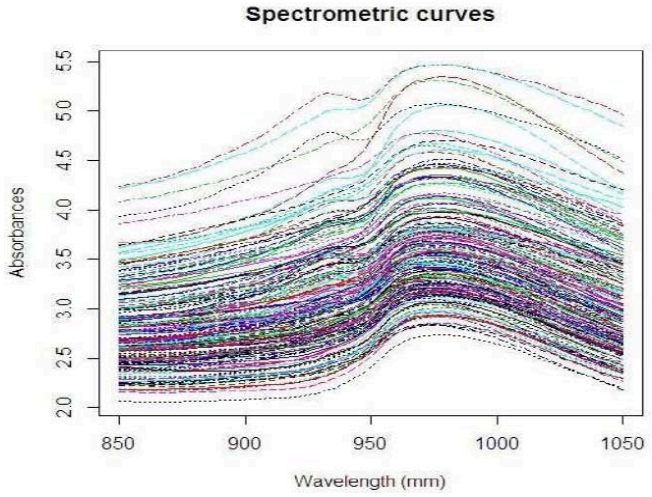

Figure 1: The Spectrometric curves.

The goal of Tecator data is to allow for the finding of the ratio of specific chemical tenor because chemistry method analysis would take more time and it will be more costly. For that reason, a regression functional problem with a response is a good way for this type for example Y is the rate of fat variable in the part of the meat (the response variable), and utilize the spectrometric curves (functional covariate) to estimate y. This instance is also fixed (see for instance, Ferraty and Vieu (2006) and Shang (2014)) [10, 22], so we can use the same case study but different methods to address this problem.

The descriptive of the data in this way, since for each curve I among 215 pieces of meat, which notice one spectrometric discretized object (x_i), which correlates with the absorbance measured on a grid of 100 wavelengths. Let y_i be represent fat part for each unit i.

The goal of this study is to estimate fat purport yˆ from a new curve x.

Then, the rendering of nonparametric functional regression, splitting the datainto two groups. The first sample is learning which contains 160 curves (x_i, y_i), i=1,2,...,160, for this sample utilizes both the covariates function and the responding output variables, which allows us to structure the functional kernel estimators with optimal bandwidth. The rest of group called the testing sample includes the last 55 objects, which is useful for obtaining predictions of the model and computing their achievement.

Ferraty and Vieu (2003, 2006), Burba *et al*. (2009), [10, 13, 19] prove that the functional prediction regression model measures the performance as:

i) Calculate the Square Errors: $S_{ei} = (y_i - \hat{y}_i)^2$, i = 161, ..., 215.

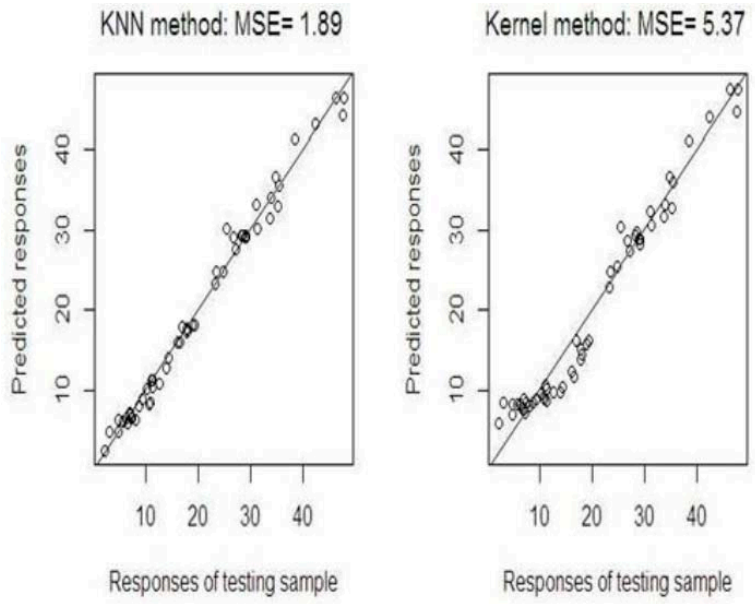

Figure 2: The prediction of fat content from Spectrometric curves.

ii) Compute the Mean Square Errors: $MSE = \frac{1}{55}\sum_{i=161}^{215} Se_i$

Running the *funopare.kernel.cv* and *funopare.knn.gcv* following the prediction based on the conditional expectation model by R procedure (see website of Nonparametric Functional Data Analysis (NFDA)). In this analysis, the semi-metric builds on the second derivatives because the curves are smooth, and the second derivative is use for both the kernel method and the KNN method. Figure **2** displays the predicted responses by these two models on the testing sample (responses of testing sample) when the semi-metric based on the second derivatives, so that the result by the KNN method is more accurate than the kernel method because of the fineness of the grid of the curves.

Example 2

Canadian Weather Stations: Prediction of Precipitation Over whole year from the Mean Monthly Temperatures for 35 Stations by Two Different Methods (KNN-Method and Kernel Method):

Data were obtained from Canadian Weather Stations, and set in the R package. In this instance, only 35 stations for each of the mean monthly precipitation and the mean monthly temperature.

Figure **3** displays 35 objects of the mean monthly temperature which contains 12 observations, and one station is places as a discretized curve. Figure **4** shows the mean monthly precipitation of the 35 Stations, with each station representing one curve. Then, for each station, I among 35 curves have one discretized curve x_i which corresponds to the average monthly temperature measured at 12-time points. The prediction of each station over the whole year contains y_i from Canadian Weather Stations.

The mean monthly temperatures represent functional covariates, and the mean monthly precipitation corresponds the output. The goal is to predict the mean monthly precipitation over the whole year $\hat{y}$, from a new mean monthly.

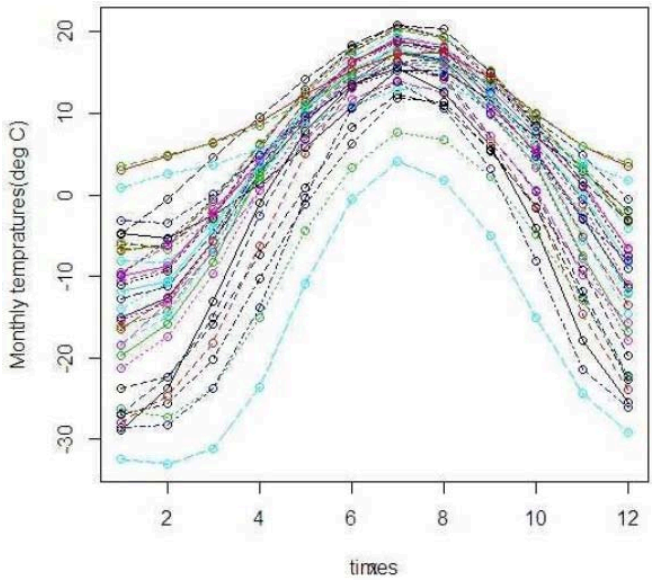

Figure 3: The mean monthly temperature at thirty-five Canadian Weather Stations.

Then, the execution of nonparametric functional regression, using the cross-validation procedure which splits the sample into two subsamples, the first 25 curves used as a learning sample and the last 10 stations used as a testing sample.

Ferraty and Vieu (2003, 2006), Burba *et al*. (2009), and Shang (2014) [10, 13, 19, 22] prove that the

functional prediction regression model is measured by the performance as below:

i) Calculate the Square Errors: $Se_i = (y_i - \hat{y}_i)^2, i = 26,27,\ldots,35.$

ii) Compute the Mean Square Errors: $MSE = \frac{1}{10}\sum_{i=26}^{35} Se_i$.

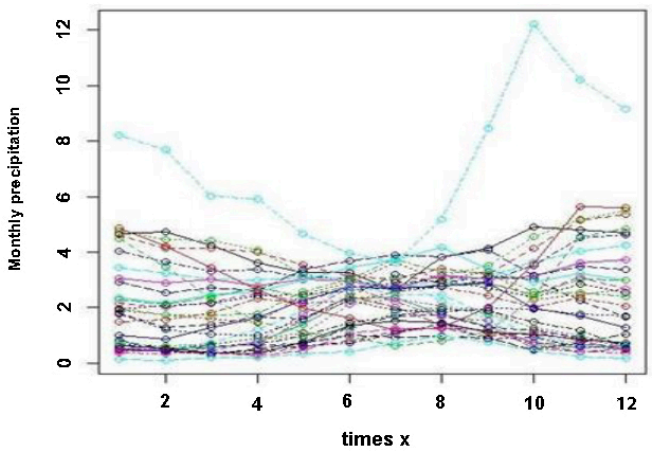

Figure 4: The mean monthly precipitation at thirty-five Canadian Weather Stations temperature station **x**.

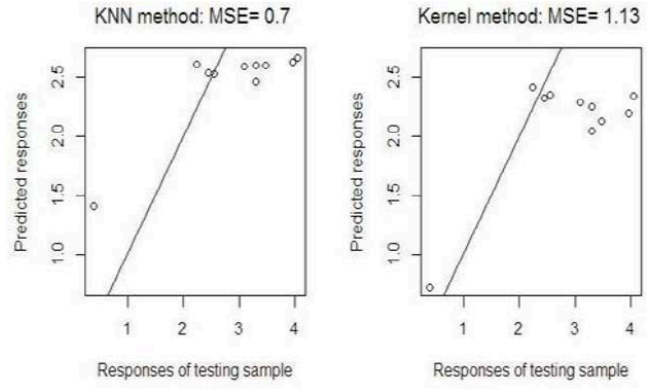

Figure 5: Prediction of precipitation over one year from mean monthly temperature curves by KNN method and kernel method with a semi-metric set up on the second derivative.

We run the funopare.kernel.cv and funopare.knn.gcv corresponding to the pre-diction based on the conditional expectation (i.e. regression) model by R procedure (see the NFDA website). Also, this analysis utilizes the semi-metric build on second derivatives (q=2) because the curve is smooth, for both the kernel method and KNN method. Figure **5** shows the predicted response values versus the responses of testing the sample for both methods based on the second derivatives. As found that, both methods give poor prediction but the result produced by the KNN method is more accurate than the kernel method. The reason for that result leads to the collection of data from different stations in Canada. Thus, there is a distance between curves as well as a difference in weather (Atlantic, Continental, Pacific, and Arctic).

4. CONCLUSION

The main point of this article is to present a comparison between two methods in functional nonparametric regression where the output is a scalar value and the covariates are functions. The presented model, notes that the K-Nearest Neighbor predictor supplies perfect estimation when compared with the results obtained from the kernel estimator model. Using the semi-metrics for the measure of closeness between the covariate function (semi-metric build on second derivatives and functional principle component analysis), as noted by Ferraty and Vieu (2006) [10]. The use of the K-NN method and kernel model are clarified through some numerical examples. In future studies, we will try to research functional nonparametric regression when the function is covariate and the response is multivariate and then take the engagement between different components of the responses into consideration.

REFERENCES

[1] Ramsay J, Silverman BW. Functional Data Analysis, Springer, New York 1997. https://doi.org/10.1007/978-1-4757-7107-7

[2] Ramsay J, Silverman BW. Functional Data Analysis, Second ed., Springer, New York 2005. https://doi.org/10.1007/b98888

[3] Cao G, Wang S, Wang L. Estimation and inference for functional linear regression models with partially varying regression coefficients. Stat 9(1) (2020): p.e286Montgomery AA, Peters TJ, Little P. Design, analysis and presentation of factorial randomised controlled trials. BMC Medical Research Methodology 2003; 3(1): 1-5. https://doi.org/10.1002/sta4.286

[4] Cardot H, Ferraty F, Sarda P. Spline estimators for the functional linear model. Statistica Sinica 2003; 13(3): 571-592.

[5] Ramsay J, Silverman BW. Applied Functional Data Analysis: Methods and case studies, Springer, New York 2002. https://doi.org/10.1007/b98886

[6] Mahalanobis PC. A method of fractile graphical analysis. Sankhy: The Indian Journal of Statistics Series A 1961; 23(1): 41-64.

[7] Nadaraya EA. On estimating regression. Theory of Probability and Its Applications 1964; 9(1): 141-142. https://doi.org/10.1137/1109020

[8] Watson GS. Smooth regression analysis. Sankhyā: The Indian Journal of Statistics, Series A 1964; pp. 359-372.

[9] Härdle W. Applied nonparametric regression, Cambridge university press, UK 1990. https://doi.org/10.1017/CCOL0521382483

[10] Ferraty F, Vieu P. Nonparametric functional data analysis: theory and practice. Springer Science, New York 2006.

[11] Omar KMT, Wang B. Nonparametric regression method with functional covariates and multivariate response. Communications in Statistics-Theory and Methods 2019; 48(2): 368-380. https://doi.org/10.1080/03610926.2017.1410716

[12] Midi H, Sani M, Ismaeel SS, Arasan J. Fast Improvised Influential Distance for the Identification of Influential Observations in Multiple Linear Regression. Sains Malaysiana 2021; 50(7): 2085-2094. https://doi.org/10.17576/jsm-2021-5007-22

[13] Ferraty F, Vieu P. Curves discrimination: a nonparametric functional approach. Computational Statistics and Data Analysis 2003; 44(1): 161-173. https://doi.org/10.1016/S0167-9473(03)00032-X

[14] Ferraty F, Mas A, Vieu P. Nonparametric regression on functional data: inference and practical aspects. Australian and New Zealand Journal of Statistics 2007; 49(3): 267-286. https://doi.org/10.1111/j.1467-842X.2007.00480.x

[15] Rachdi M, Vieu P. Nonparametric regression for functional data: automatic smoothing parameter selection. Journal of Statistical Planning and Inference 2007; 137(9): 2784-2801. https://doi.org/10.1016/j.jspi.2006.10.001

[16] Ismaeel SS, Omar KMT, Wang B. K-nearest neighbor method with principal component analysis for functional nonparametric regression. Baghdad Science Journal 2022; 19(6 (Suppl.)): 1612. https://doi.org/10.21123/bsj.2022.6476

[17] Doori A. Hazard Rate Estimation Using Varying Kernel Function for Censored Data Type I. Baghdad Science Journal 2019; 16(3 (Suppl.)): 0793-0793.

[18] Ferraty F, Vieu P. Nonparametric models for functional data, with application in regression, time series prediction and curve discrimination. Nonparametric Statistics 2004; 16(1-2): 111-125. https://doi.org/10.1080/10485250310001622686

[19] Burba F, Ferraty F, Vieu P. k-Nearest Neighbor method in functional nonparametric regression. Journal of Nonparametric Statistics 2009; 21(4): 453-469. https://doi.org/10.1080/10485250802668909

[20] Midi H, Ismaeel SS. Fast improvised diagnostic robust measure for the identification of high leverage points in multiple linear regression. Journal of Statistics and Management Systems 2018; 21(6): 1003-1019. https://doi.org/10.1080/09720510.2018.1466443

[21] Báıllo A, Gran´e A. Local linear regression for functional predictor and scalar response. Journal of Multivariate Analysis 2009; 100(1): 102-111. https://doi.org/10.1016/j.jmva.2008.03.008

[22] Shang HL. Bayesian bandwidth estimation for a nonparametric functional regression model with mixed types of regresses and unknown error density. arXiv preprint arXiv:1403.1913, 2014. https://doi.org/10.1080/10485252.2014.916806

Received on 27-09-2023 Accepted on 20-10-2023 Published on 10-11-2023

https://doi.org/10.6000/1929-6029.2023.12.22