Computers, Materials & Continua
DOI: 10.32604/cmc.2023.033441
Article

Functional Nonparametric Predictions in Food Industry Using Near-Infrared Spectroscopy Measurement

Ibrahim M. Almanjahie[1,2,*], Omar Fetitah[3], Mohammed Kadi Attouch[3] and Tawfik Benchikh[3,4]

[1]Department of Mathematics, College of Science, King Khalid University, Abha, 62529, Saudi Arabia
[2]Statistical Research and Studies Support Unit, King Khalid University, Abha, 62529, Saudi Arabia
[3]Laboratory of Statistics and Stochastic Processes University of Djillali Liabes BP 89, Sidi Bel Abbes, 22000, Algeria
[4]Medical Faculty, Djillali Liabes University BP 89, Sidi Bel Abbes, 22000, Algeria
*Corresponding Author: Ibrahim M. Almanjahie. Email: imalmanjahi@kku.edu.sa
Received: 16 June 2022; Accepted: 08 September 2022

Abstract: Functional statistics is a new technique for dealing with data that can be viewed as curves or images. Parallel to this approach, the Near-Infrared Reflectance (NIR) spectroscopy methodology has been used in modern chemistry as a rapid, low-cost, and exact means of assessing an object's chemical properties. In this research, we investigate the quality of corn and cookie dough by analyzing the spectroscopic technique using certain cutting-edge statistical models. By analyzing spectral data and applying functional models to it, we could predict the chemical components of corn and cookie dough. Kernel Functional Classical Estimation (KFCE), Kernel Functional Quantile Estimation (KFQE), Kernel Functional Expectile Estimation (KFEE), Semi-Partial Linear Functional Classical Estimation (SPLFCE), Semi-Partial Linear Functional Quantile Estimation (SPLFQE), and Semi-Partial Linear Functional Expectile Estimation (SPLFEE) are models used to accurately estimate the different quantities present in Corn and Cookie dough. The selection of these functional models is based on their ability to construct a forecast region with a high level of confidence. We demonstrate that the considered models outperform traditional models such as the partial least-squares regression and the principal component regression in terms of prediction accuracy. Furthermore, we show that the proposed models are more robust than competing models such as SPLFQE and SPLFEE in the sense that data heterogeneity has no effect on their efficiency.

Keywords: Functional statistics; NIR; chemical component; kernel estimation

1 Introduction

As a rapid, easy, and non-destructive approach, NIR spectroscopy is commonly used for the quantitative examination of high dimension samples in a variety of disciplines, including the environment, pharmaceutical analysis, fuels, and foods [1–4]. However, because of background noise and overlapping bands, chemometrics must be used to build a suitable model to relate the intrinsic relationship between spectra and a specific component concentration [5–7]. As a result, multivariate calibrations are gaining popularity in quantitative NIR spectroscopy analysis.

On the other hand, thanks to increased storage capabilities that allow massive data to be recorded, a tremendous revolution in computer development has occurred over the last two decades. This technological advancement has aided the development of the statistical branch devoted to functional data analysis (FDA), both in terms of theoretical and methodological advancements as well as the diversification of application domains. As a result, strategies for treating observations as functional random elements have been devised. In this context, we provide some definitions that can be utilized to correct the terminology. To begin with, remember that a functional random variable is essentially a random variable with values in the space of potentially infinite dimension denoted by [8]. For example, this functional space may be a space of functions, linear operators, and so on; and the fact that these observations are intended to correspond to such a space is the domain's main source of difficulty.

Regression models the relationship between a functional (or scalar) response and a functional (or scalar) regression component. The plague of dimension, however, is a drawback to these techniques (where the speed of convergence of the estimators is a decreasing function of the dimension of the covariates). The research of the nonlinear regression model in this situation is significantly more recent than that of the linear case. Indeed, [9] obtained the initial results; the authors were interested in a nonparametric estimate of the mode of distribution of a functional variable satisfying a fractal criterion. When the observations are independent and identically distributed, [10] established the almost complete convergence of a kernel estimator of the regression function under the same conditions. However, these methods suffer from the curse of dimensionality (when analyzing a high number of covariates, where the speed of convergence of the estimators decreases as the dimension of the covariates increases). To overcome this issue, semi-parametric models, which provide a compromise between a nonparametric model and a fully parametric model, were studied. Their key advantage over nonparametric regression is faster estimator convergence.

The partial linear model is one of the most popular semi-parametric models in this literature. Reference [11] who created a semi-functional partial linear regression model, were the first to introduce this approach to the FDA. They developed estimators for each model component (parametric and nonparametric) and demonstrated their asymptotic characteristics. In addition, [12] used the model to forecast functional time series. Furthermore, using cross-validation and Bayesian techniques, [13,14] proposed and investigated bandwidth selections. Reference [15] suggested a k-nearest-neighbors (k-NN) approach and obtained the asymptotic performances of k-NN estimators, whereas [16] investigated a semi-functional partly linear regression model with random responses. Reference [17] investigated semi-functional partial linear regression for spatial data and obtained asymptotic normality of the parametric component as well as probability convergence with the rate of the nonparametric component.

The downside of classical regression, according to another section, is that the regression function estimation is sensitive to outliers and may be incorrect in some instances, such as when the distribution is multimodal or extremely asymmetric. This lack of robustness can be addressed by forecasting using conditional quantiles, as examined by [18,19]. Unfortunately, these conditional models are not always satisfactory and are difficult to calculate due to the non-continuous loss function; Its high sensitivity to extreme values, in particular.

It is possible to combine the concepts of least squares regression with conditional quantiles to create a new measure of risk known as expectile. The conditional expectile has been widely investigated, with pioneering works in [20,21].

The main goal of this study is to forecast various chemical components using a functional Near-Infrared Reflectance (NIR) spectroscopic approach and some recent statistical models based on

functional kernel estimation and semi-partial linear functional processes. Corn and cookie dough are examples of NIR spectroscopy datasets used in this investigation. In particular, for this type of data, we propose six models: KFCE, KFQE, KFEE, SPLFCE, SPLFQE, and SPLFEE.

The remainder of the paper is as follows. We detail the materials and methods, including collected samples based on spectrometry and chemometric analysis in Section 2. We also introduce the modern statistical models we used for the analyses in the same section. The results and discussion of this study are presented in Section 3. Finally, we conclude our study in Section 4.

2 Materials and Methods

2.1 Data Description

This study made use of two tiny NIR spectral datasets. In addition to NIR spectra and moisture, oil, protein, and starch levels, 80 corn samples were analyzed using three spectrometers. There are two methods employed in this study: NIR spectrometer mp6 and protein content. 700 variables are captured in the wavelength range of 1100–2498 nanometers (nm) with a 2 nm digitization interval in each $(\omega_1, \ldots, \omega_{700})$ spectrum. The "Data Availability Statement" section includes a link to the data. This dataset is used in this study for the primary discussion (it was previously analyzed by [22]).

The second set of datasets examined here is the results of an experiment to assess the feasibility of using NIR spectroscopy to determine the composition of biscuit dough bits from 72 cookie dough samples (formed but unbaked biscuits) for possible online implementation. (See [23] for a detailed description of the experiment.) In a simpler term, two similar sample sets were created, with the standard recipe being changed to provide a wide range for each of the four ingredients under investigation: fat, sucrose, dry flour, and water. Each dough piece has its own NIR reflectance spectrum.

The original spectral data set consists of 700 points taken in 2 nm steps from 1100 to 2498 nm. The absorption $X_i(\omega)$ of light is measured for each wavelength ω and each sample i. $X_i(\_1), \ldots, X_i(\_700)$ represent the i th discretized spectrometric curve. The spectrometric curves are depicted in Figs. 1 and 2.

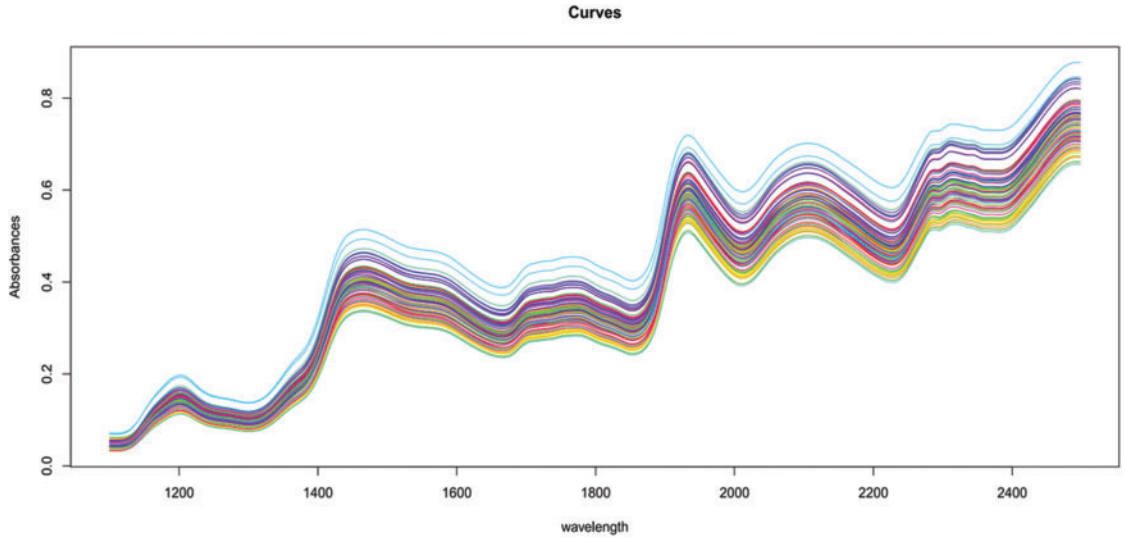

Figure 1: Corn 80 samples of near-infrared spectra

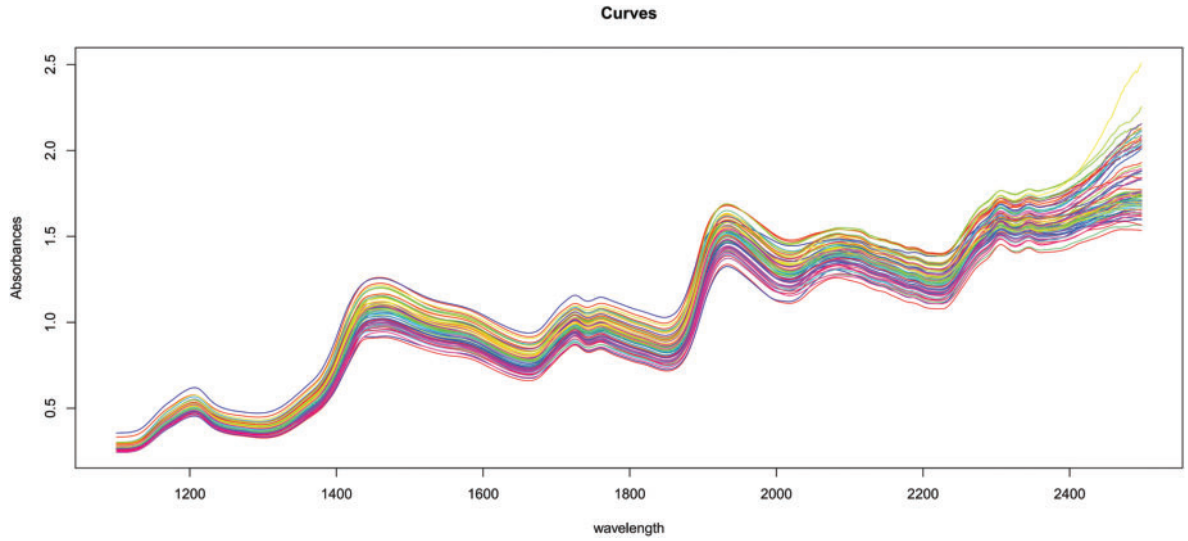

Figure 2: 72 samples of near-infrared spectra from cookie dough

It is obvious that all of these curves, even if observed at discrete places, have a continuum character. This continuous property is referred to as functional data. Figs. 1 and 2 illustrate a typical set of such data for the food business.

In chemometrics, functions–such as absorbance or emission spectra provided for food samples–are frequently employed to determine the concentration of specific components. Using spectra is often far less expensive than other types of chemical analysis.

2.2 Statistical Models

The objective is to create an equation that predicts response values Y from the spectral data X and vectorial parameter Z for future samples where Y is unknown but X and Z can be measured economically and promptly. Which we assume that those variables are linked by the following regression formula,

$$Y = r\left(Z^1, Z^2, Z^3, X\right) + \epsilon = m\left(X\right) + \sum_{k=1}^{3} Z^k \beta^k + \epsilon, \tag{1}$$

where the response variable Y is consider to be the protein content for the corn datasets and the percentage of fat content for the cookie dough datasets from the corresponding contents $Z = (Z^1, Z^2, Z^3)^t$ of moisture, oil, starch for the corn datasts and sucrose, dry flour, water for the cookie dough, as well as from the near-infrared absorbance spectra X (Figs. 1, 2 and Table 1 describe those datasets). $\beta^1, \beta^2, \beta^3$ and $m\left(\cdot\right)$ are unknown modelling the relationship between Z^1, Z^2, Z^3, X and Y; The statistical difficulty is to come up with a suitable estimator.

In this section, we will concentrate on regression models such that $\mathbb{E}(\epsilon | Z^1, Z^2, Z^3, X) = 0)$.

The main objective of this study is to show and explore various methods for modeling nonlinear relationships in datasets comprising functional data. A scalar response to functional and vectorial explanatory variables is a specific example that we focus on here. Using both traditional and robust

methods, we will propose two approaches: functional nonparametric regression and semi-functional partial linear regression.

Table 1: Inferential statistics of the datasets

Datasets	Content	$IC_{95\%}$ bounds	Mean	1st Qua	Median	3rd Qua	Sd
Corn	*Moisture*	lower	10.148%	9.817%	10.13%	10.336%	0.329%
		upper	10.318%	10.108%	10.323%	10.597%	0.45%
	Oil	lower	3.458%	3.306%	3.451%	3.557%	0.153%
		upper	3.537%	3.449%	3.556%	3.687%	0.209%
	Protein	lower	8.557%	8.23%	8.435%	8.823%	0.431%
		upper	8.779%	8.428%	8.737%	9.333%	0.59%
	Starch values	lower	64.512%	63.631%	64.574%	65.115%	0.71%
		upper	64.878%	64.544%	65.104%	65.519%	0.972%
Cookie dough	*Fat*	lower	17.847%	15.85%	17.54%	19.23%	1.691%
		upper	18.772%	17.37%	19.14%	20.66%	2.355%
	Sucrose	lower	15.676%	11.81%	15.01%	18.55%	3.354%
		upper	17.511%	14.67%	18.13%	21.42%	4.671%
	Dry flour	lower	48.339%	45.86%	48.47%	50.09%	2.347%
		upper	49.623%	48.19%	49.94%	51.95%	3.269%
	Water	lower	13.842%	12.6%	13.7%	14.94%	1.274%
		upper	14.539%	13.7%	14.89%	15.3%	1.774%

2.2.1 KFCE

Ferraty et al. [8] were the first to investigate nonparametric estimation of functional regression (see also [24]). They estimated this statistical model using the Nadaraya-Watson method. Specifically, the function $r(x)$ is written directly using the least square error criteria by

$$r(x) = \underset{t}{\operatorname{argmin}}\, \mathbb{E}\left((Y-t)^2 | X = x\right). \tag{2}$$

It follows that $r(x) = \mathbb{E}[Y|X = x]$. As a result, for any fixed curves x, we forecast the response y in relation to the criterion (2) by $\hat{r}(x)$ that is the classical kernel estimator of $r(x)$, and defined by

$$\hat{r}_{KFCE}(x) = \frac{\sum_{i=1}^{n} Y_i K\left(\frac{d(x, X_i)}{h_n}\right)}{\sum_{i=1}^{n} K\left(\frac{d(x, X_i)}{h_n}\right)}, \tag{3}$$

with K is a kernel function, h_n:h is a nonnegative real sequence and $d(.,.)$ is semi metric that measures the distance between the smoothed curves.

2.2.2 KFQE (Resp. KFEE)

In robust regressions, risk (or loss) measures such as quantile and expectile have been utilized. Quantile and expectile can characterize the entire distribution of a random variable, and there is no missing information, which decreases the sensitivity of the quantile (or expectile) to noise and outliers. Despite their variations in structure, quantiles and expectiles have similar qualities. A combination of least squares and conditional quantiles is used to create the expectile regression, which combines the advantages of both models. In order to build this regression models, one must solve the following problem:

$$r(x) = \underset{s}{\operatorname{argmin}}\, \mathbb{E}\left(\rho_p(Y-s)\,|X=x\right). \tag{4}$$

$$\rho_p(t) = \left|p - \mathbb{I}_{\{(t)\le 0\}}\right| |t| \left(\text{resp. } \rho_p(t) = \left|p - \mathbb{I}_{\{(t)\le 0\}}\right| (t)^2\right),$$

with $\rho_p(Y-s)$ is the scoring function, the $p \in [0,1]$ is called the asymmetry parameter level and $\mathbb{I}_A$ represents the event's indicator function. The model (4) has been introduced in functional statistics by [18] (rep. [25]). Its functional estimation is given as follow

$$\hat{r}(x) = \underset{s}{\operatorname{argmin}} \frac{\sum_{i=1}^{n} \rho_p(Y_i - s) K\left(\dfrac{d(x, X_i)}{h_n}\right)}{\sum_{i=1}^{n} K\left(d\dfrac{(x, X_i)}{h_n}\right)}, \tag{5}$$

for both *KFQE* and *KFEE* model.

2.2.3 SPLFCE

This last regression is an alternative semi-parametric regression which takes into consideration the vectorial variable Z, introduced by [11]. It is defined as following

$$\hat{r}_{SPLFCE}(z,x) = \hat{m}_{SPLFCE}(x) + \sum_{k=1}^{3} z^k \hat{\beta}^k_{SPLFCE}, \tag{6}$$

where $\hat{\beta}_{SPLFCE} = \left(\hat{\beta}^1_{SPLFCE}, \hat{\beta}^2_{SPLFCE}, \hat{\beta}^3_{SPLFCE}\right)^t$ and $z = (z^1, z^2, z^3)^t$ is a fixed vector in $\mathbb{R}^3$ and $\hat{\beta}_{SPLFCE} = \left(\tilde{Z}^t\tilde{Z}\right)^{-1}\tilde{Z}^t\tilde{Y}$

with

$$\tilde{Z} = \left(\tilde{Z}_1, \ldots, \tilde{Z}_n\right)^t \text{ and } \tilde{Y} = \left(\tilde{Y}_1, \cdots, \tilde{Y}_n\right)^t$$

$$\tilde{Z}_i = Z_i - \frac{\sum_{j=1}^{n} Z_j K\left(\dfrac{d(X_i, X_j)}{h_n}\right)}{\sum_{j=1}^{n} K\left(\dfrac{d(X_i, X_j)}{h_n}\right)} \text{ and } \tilde{Y}_i = Y_i - \frac{\sum_{j=1}^{n} Y_j K\left(\dfrac{d(X_i, X_j)}{h_n}\right)}{\sum_{j=1}^{n} K\left(\dfrac{d(X_i, X_j)}{h_n}\right)},$$

and

$$\hat{m}_{SPLFCE}(x) = \frac{\sum_{i=1}^{n} \left(Y_i - Z_i^t \hat{\beta}_{SPLFCE}\right) K\left(\dfrac{d(x, X_i)}{h_n}\right)}{\sum_{i=1}^{n} K\left(\dfrac{d(x, X_i)}{h_n}\right)}.$$

2.2.4 SPLFQE (Resp. SPLFEE)

This final regression is a semi-parametric alternative to the least regression approach. It has only been considered recently in functional statistics by [19]. Our estimators are as follows:

$$\hat{r}_p(z,x) = \hat{m}_p(x) + \sum_{k=1}^{3} z^k \hat{\beta}_p^k, \tag{7}$$

for both SPLFQE and SPLFEE models, where $p \in [0,1]$ and $\hat{m}_p(x)$, $\hat{\beta}_p$ are obtained by using three-stage procedure:

Stage 1:

We obtain initial estimators of $m_p\left(X_j\right)$ and β_p ($\tilde{m}_p\left(X_j\right)$ and $\tilde{\beta}_p$) by reducing the local weighted robust loss function given below

$$\sum_{\substack{i=1 \\ i\neq j}}^{n} \rho_p\left(Y_i - m_p\left(X_j\right) - Z_i^t \beta_p\right) K\left(\frac{d\left(X_i, X_j\right)}{h_n}\right).$$

Stage 2:

$$\hat{\beta}_p = \underset{\beta p}{\operatorname{argmin}} \sum_{j=1}^{n} \rho_p\left(Y_j - \tilde{m}_p\left(X_j\right) - Z_j^t \beta_p\right).$$

Stage 3:

The final estimate of $m_p(x)$ is obtained by minimizing the following robust loss function

$$\hat{m}_p(x) = \underset{mp(x)}{\operatorname{argmin}} \sum_{i=1}^{n} \rho_p\left(Y_i - m_p(x) - Z_i^t \hat{\beta}_p\right) K\left(\frac{d(x, X_i)}{h_n}\right).$$

we conclude our study in Section 4.

3 Results and Discussions

3.1 Results

There is a strong correlation between the estimation accuracy of all the models listed above and the selection of the semi-metric $d(\cdot,\cdot)$ and bandwidth $h_n := h$ parameters. We considered (as suggested by [8]) a class of semi-metrics based on derivatives $\left\{d_q(\cdot,\cdot), q = 0, 1, \ldots, 4\right\}$, were

$$d_q\left(x_i, x_j\right) = \left(\int_0^1 \left(x_i^{(q)}(t) - x_j^{(q)}(t)\right)^2 \mathrm{dt}\right)^{1/2}.$$

In this prediction problem, the smoothing parameter h_n has the greatest impact. As a result, the optimal value h_{opt} is the data-driven bandwidth derived by a cross-validation procedure:

$$h_{opt} = \underset{h}{\operatorname{argmin}}\, CV(h) \qquad \text{where} \quad CV(h) = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_{(-i)}\right)^2,$$

with $\hat{Y}_{(-i)}$ the leave-one-out of the estimator $\hat{r}_{KFCE}(X_i), \hat{r}_{KFQE}(X_i), \hat{r}_{KFEE}(X_i), \hat{r}_{SPLFCE}(Z_i, X_i), \hat{r}_{SPLFQE}(Z_i, X_i)$ or $\hat{r}_{SPLFEE}(Z_i, X_i)$. Through this study, we take $p = 0.75$ and used the quadratic kernel K defined as $K(u) = \frac{3}{2}\left(1 - u^2\right)\mathbb{I}_{[0,1]}(u)$. Two subsamples of the original dataset, $\left\{\left(Z_i^1, Z_i^2, Z_i^3, X_i, Y_i\right); i = 1, \ldots, n\right\}$, are taken to test the semi-partial linear functional regression's prediction performance:

$\mathcal{L} = \left\{\left(Z_i^1, Z_i^2, Z_i^3, X_i, Y_i\right); i \in L\right\}$ and $\mathcal{T} = \left\{\left(Z_i^1, Z_i^2, Z_i^3, X_i, Y_i\right); i \in T\right\}$ with $L \cup T = \{1, \ldots, n\}$ and $L \cap T = \emptyset$. The learning sample $\mathcal{L}$ enables the estimation of estimators and the automatic selection of the bandwidth h_n via a cross-validation technique. The mean squared error MSE of prediction is provided by the testing sample:

$$MSE = \frac{1}{\#(\mathcal{T})} \sum_{i \in T} \left(Y_i - \hat{Y}_{(i)}\right)^2,$$

where, for all i in $\mathcal{T}$, $\hat{Y}_i$ is the value of the estimator $\hat{r}_{KFCE}(X_i)$, $\hat{r}_{KFQE}(X_i)$, $\hat{r}_{KFEE}(X_i)$, $\hat{r}_{SPLFCE}(Z_i, X_i)$, $\hat{r}_{SPLFQE}(Z_i, X_i)$ or $\hat{r}_{SPLFEE}(Z_i, X_i)$.

When the original data set is split into 100 equal parts, the Mean Square Error (MSE) values can be computed and their distribution can be seen using a beanplot. Finally, for all datasets, the testing sample $\mathcal{T}$ represents 25% of the total sample size. Figs. 3 and 4 show the accuracy of the forecasts related to one run.

Figs. 3 and 4 display the response variable prediction produced from the testing sample: the observed values *vs.* the projected values. They show that there is a significant benefit across semi-partial linear functional models when compared to kernel functional models. The semi-partial linear functional model for classical, quantile, and expectile regression outperforms the kernel functional model for classical, quantile, and expectile regression.

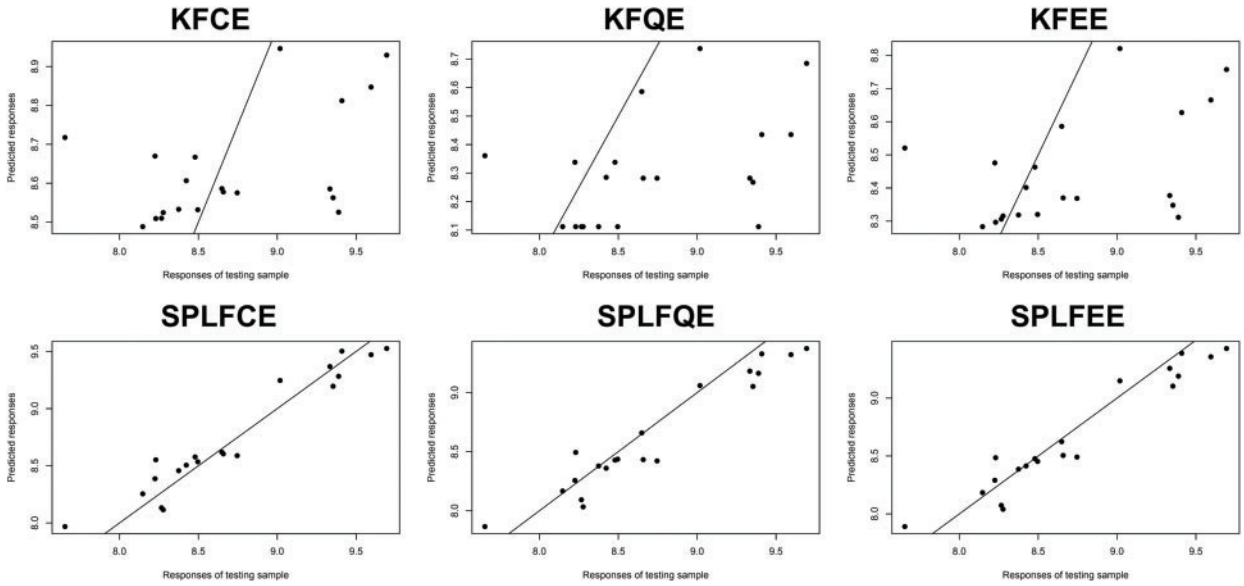

Figure 3: Predictive performance: Corn datasets

Figs. 5 and 6 show the mean squared error (MSE) of the predictions, respectively, as shown in the bean-plots. The results of a comparison between the classical (SPLFKE, SPLFQE, SPLFEE) and semi-partial (KFCE, KFQE, KFEE) methods of predicting the shape of spectrum curves show that the latter models perform better.

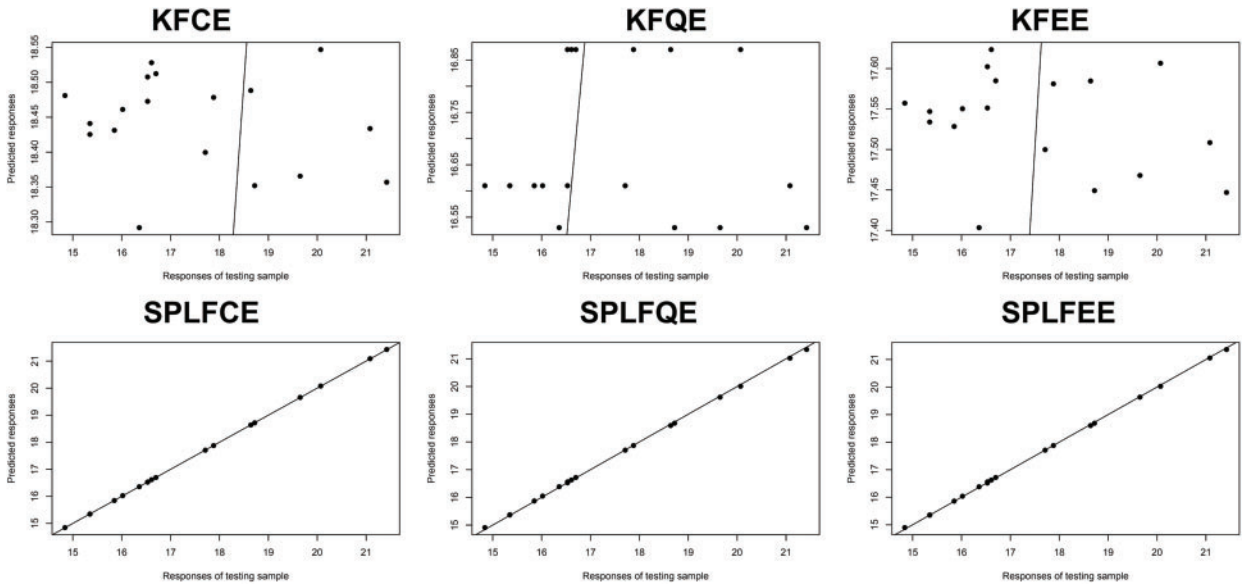

Figure 4: Predictive performance: Cookie dough datasets

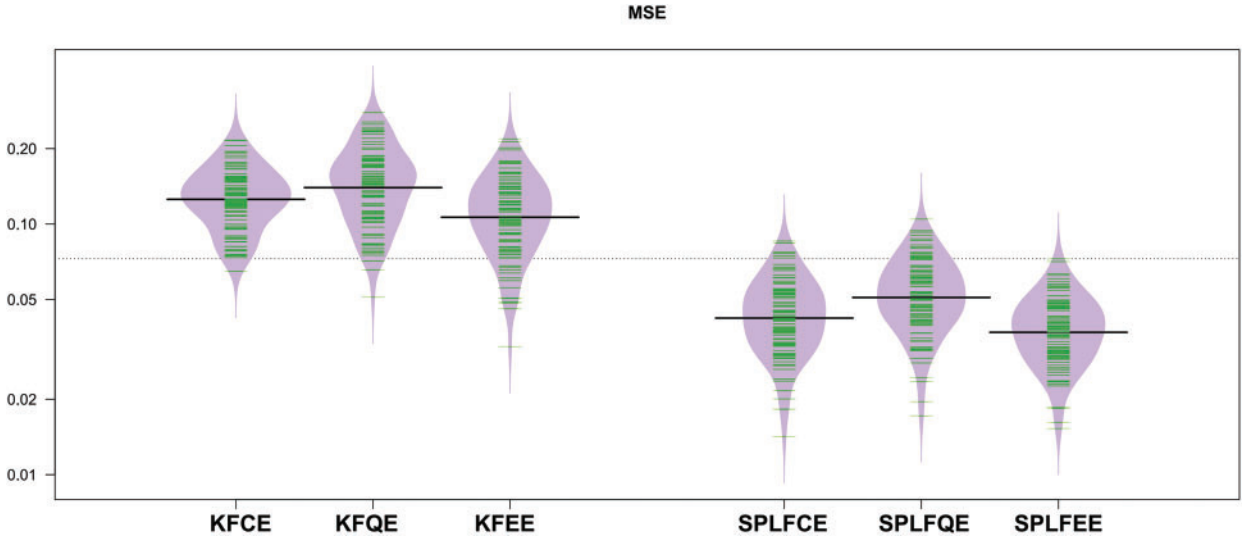

Figure 5: Bean-plots of the *MSE* using four models: Corn datasets

The MSE values for the two semi-partial linear functional models, SPLFCE, SPLFQE, and SPLFEE, are reasonably stable and less when compared to the kernel functional models, KFCE, KFQE, and KFEE. Although the performance of the studied models varies, the variability of the MSE for the three proposed models is relatively stable for the semi-partial linear functional for classical, quantile, and expectile regression models as compared to the kernel functional for classical, quantile, and expectile regression models.

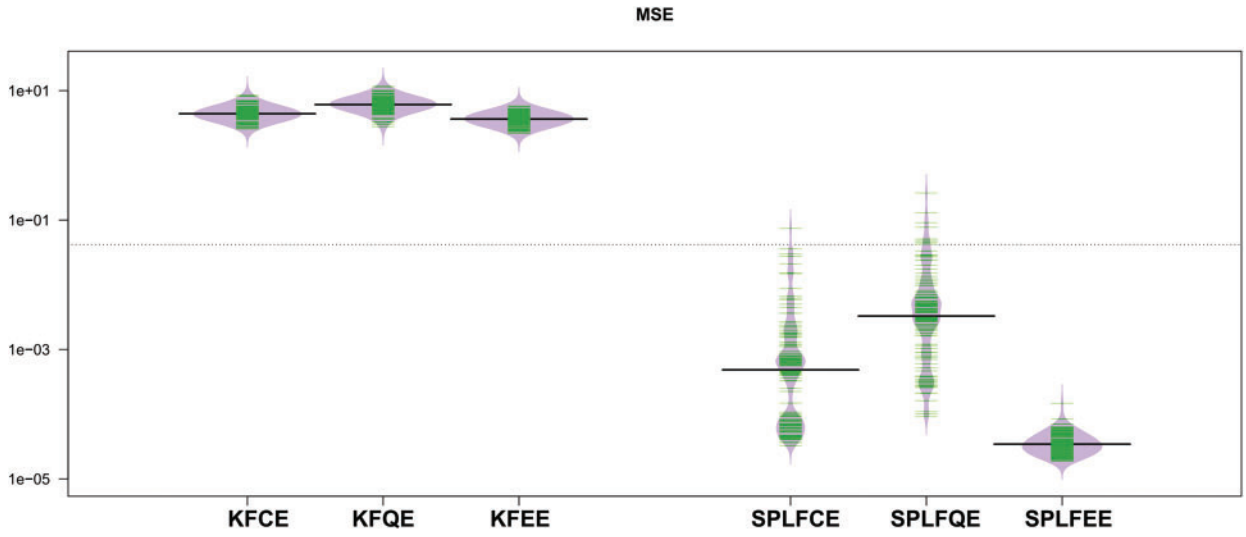

Figure 6: Bean-plots of the *MSE* using four models: Cookie dough datasets

It is worth mentioning that the functional semi-parametric regression approach includes a distance measure $d(.,.)$. The most informative proximity measure $d(.,.)$ may change depending on the nature of what we need to predict from spectrometric curves, as discussed in the preceding section, but we never know which one is the most relevant in advance. For example, if one wishes to concentrate on our spectrometric datasets, one can specify $\mathcal{D} = \{d_q; q = 0, 1, \ldots, 4\}$. In this case, the family $\mathcal{D}$ is simply indexed by the four leading integers. Table 2 indicates the amount of predictive power that can be expected to improve.

Table 2: MSE of the six models according to derivation degree q

Datasets	q	KFCE	KFQE	KFEE	SPLFCE	SPLFQE	SPLFEE
Corn	0	0.140046	0.180314	0.141321	0.047479	0.060149	0.038829
	1	0.132705	0.153369	0.117859	0.044554	0.052448	0.039988
	2	0.143072	0.172831	0.134176	0.044524	0.053287	0.038016
	3	0.142950	0.160119	0.126371	0.048978	0.058533	0.041505
	4	0.129085	0.161383	0.119972	0.042774	0.049528	0.037037
Cookie dough	0	3.011332	4.161048	2.533001	0.001462	0.008604	0.000018
	1	4.512168	6.097559	3.917729	0.005593	0.026370	0.000011
	2	4.431158	6.300766	3.366549	0.000966	0.006512	0.000011
	3	2.869686	3.779619	2.458226	0.007894	0.065315	0.000014
	4	3.435075	4.627528	2.891879	0.005159	0.021796	0.000013

The previous three regression models appear to have done well and have a reasonable level of accuracy for this prediction task. On the other hand, the performance of the analyzed models varies according to the input vectorial variable Z. The MSE errors are summarized in Table 3.

Table 3: MSE of the six models according to vectorial input variable

Datasets	Vectorial covariate Z	KFCE	KFQE	KFEE	SPLFCE	SPLFQE	SPLFEE
Corn	Oil (Z^2), starch values (Z^3)	0.1322	0.1496	0.1161	0.0530	0.0633	0.0452
	Moisture (Z^1), starch values (Z^3)	0.1290	0.1417	0.1087	0.0464	0.0547	0.0452
	Moisture (Z^1), oil (Z^2)	0.1301	0.1422	0.1095	0.1059	0.1351	0.1288
Cookie dough	Dry flour (Z^2), water (Z^3)	4.6307	6.2975	3.8824	0.9471	1.0400	0.9300
	Sucrose (Z^1), water (Z^3)	4.8707	6.5970	4.0496	0.1536	0.1606	0.1542
	Sucrose (Z^1), dry flour (Z^2)	4.9172	6.6873	4.0675	0.3365	0.3399	0.3085

Table 3 show that the values of MSE are relatively more significant for the six proposed models, namely KFCE, KFQE, KFEE, SPLFCE, SPLFQE, and SPLFEE, in the absence of moisture Z^1 for the corn dataset. While for the Cookie dough dataset, the *MSE* of KFCE, KFQE, and KFEE (resp. SPLFCE, SPLFQE, and SPLFEE) is more significant in the absence of water Z^3 (resp. sucrose Z^1).

3.2 Discussion

KFCE, KFQE, KFEE, SPLFCE, SPLFQE, and SPLFEE were examined and used to multivariate calibration of two small NIR spectral datasets in nonparametric functional and semi-partial linear functional statistics (80 observations of corn dataset and 72 observations of Cookie dough dataset). Protein and fat levels were measured and ranged between $7.654\% - 9.711\%$ and $14.84\% - 21.67\%$, respectively. The primary purpose of this research is to forecast those examined parameters using near-infrared spectra curves and the related moisture, oil, starch, and sucrose contents for corn datasets and sucrose, dry flour, and water for cookie dough.

The comparison of both prediction plots in Figs. 3 and 4 show that the SPLFCE, SPLFQE, and SPLFEE procedures outperform the KFCE, KFQE, and KFEE approaches. As illustrated in Figs. 5–6 and Table 3, the SPLFCE, SPLFQE, and SPLFEE models, respectively, are the best models (with a small MSE). In particular, semi-partial linear functional models with moisture quantities for corn datasets and water for cookie dough produce the best results. To the contrary, the results in Table 2 show that using semi-metrics significantly increases predictive performance with the same sample size.

4 Conclusion

Modern statistics approaches for predicting scalar responses given functional variables are examined in this paper. Corn and cookie dough quality can be predicted using cutting-edge statistical models, such as the spectroscopic approach. By contrast to traditional methods which result in data loss following transformation, the presented models do not suffer from this problem.

This work extends the multidimensional framework by setting the response variable in a dimensional space. The comparison of the generated estimator when there are outlier's outputs is an important natural issue that we would be treating and that has not been considered in this work (see, for example, [17] and [26]). This research combines contemporary chemistry techniques with recent developments in mathematical statistics to produce effective prediction processes that use the entire curves as regressors. This research is conclusive. We may say that such a combination is economically advantageous, as well as more adaptable, resilient, and speedy. New statistical methods have been compared to traditional analytical methods as a quick and reliable alternative. For food analysis, those methods have become increasingly common in the last decade. A list of references can be found here (e.g., [27]).

Acknowledgement: The authors extend their appreciation to the Deanship of Scientific Research at King Khalid University for funding this work.

Data Availability: The corn and cookie dough data sets that used in this paper are, respectively, available at http://software.eigenvector.com/Data/Corn/index.html and https://rdrr.io/cran/ppls/man/cookie.html.

Funding Statement: This work is funded by the Deanship of Scientific Research at King Khalid University for funding this work through Large Groups Project under grant number RGP. 2/132/43.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] J. H. Qu, D. Liu, J. H. Cheng, D. W. Sun, J. Ma *et al.,* "Applications of near-infrared spectroscopy in food safety evaluation and control: A review of recent research advances," *Critical Reviews in Food Science and Nutrition*, vol. 55, pp. 1939–1954, 2015.

[2] D. Pan, G. Crull, S. Yin and J. Grosso, "Low level drug product API form analysis–avalide tablet NIR quantitative method development and robustness challenges," *Journal of Pharmaceutical and Biomedical Analysis*, vol. 89, pp. 268–275, 2014.

[3] A. O'Reilly, R. Coffey, A. Gowen and E. Cummins, "Evaluation of near-infrared chemical imaging for the prediction of surface water quality parameters," *International Journal of Environmental Analytical Chemistry*, vol. 95, pp. 403–418, 2015.

[4] J. M. Harnly, P. B. Harrington, L. L. Botros, J. Jablonski, C. Chang *et al.,* "Characterization of near-infrared spectral variance in the authentication of skim and nonfat dry milk powder collection using ANOVA-PCA, pooled-ANOVA, and partial least-squares regression," *Journal of Agricultural and Food Chemistry*, vol. 62, pp. 8060–8067, 2014.

[5] P. Tong, Y. Du, K. Zheng, T. Wu and J. Wang, "Improvement of NIR model by fractional order Savitzky–Golay derivation (FOSGD) coupled with wavelength selection," *Chemometrics and Intelligent Laboratory Systems*, vol. 143, pp. 40–48, 2015.

[6] X. Shao, X. Bian, J. Liu, M. Zhang and W. Cai, "Multivariate calibration methods in near infrared spectroscopic analysis," *Analytical Methods*, vol. 2, pp. 1662–1666, 2010.

[7] U. Thissen, B. Üstün, W. J. Melssen and L. M. Buydens, "Multivariate calibration with least-squares support vector machines," *Analytical Chemistry*, vol. 76, pp. 3099–3105, 2004.

[8] F. Ferraty and P. Vieu, "Nonparametric functional data analysis: Theory and practice," in *Springer Series in Statistics*, 2006.

[9] T. Gasser, P. Hall and B. Presnell, "Nonparametric estimation of the mode of a distribution of random curves," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 60, pp. 681–691, 1998.
[10] F. Ferraty and P. Vieu, "Dimension fractale et estimation de la régression dans des espaces vectoriels semi-normés," *Comptes Rendus de L'Académie des Sciences-Series I-Mathematics*, vol. 330, pp. 139–142, 2000.
[11] G. Aneiros-Pérez and P. Vieu, "Semi-functional partial linear regression," *Statistics & Probability Letters*, vol. 76, pp. 1102–1110, 2006.
[12] G. Aneiros-Perez and P. Vieu, "Nonparametric time series prediction: A semi-functional partial linear modeling," *Journal of Multivariate Analysis*, vol. 99, pp. 834–857, 2008.
[13] G. Aneiros-Pérez and P. Vieu, "Automatic estimation procedure in partial linear model with functional data," *Statistical Papers*, vol. 52, pp. 751–771, 2011.
[14] H. L. Shang, "Bayesian bandwidth estimation for a semi-functional partial linear regression model with unknown error density," *Computational Statistics*, vol. 29, pp. 829–848, 2014.
[15] N. Ling and P. Vieu, "On semiparametric regression in functional data analysis," in *Wiley Interdisciplinary Reviews: Computational Statistics*, Wiley Online Library, pp. e1538, 2021.
[16] N. Ling, R. Kan, P. Vieu and S. Meng, "Semi-functional partially linear regression model with responses missing at random," *Metrika*, vol. 82, pp. 39–70, 2019.
[17] 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*, vol. 51, pp. 1–21, 2021.
[18] A. Laksaci, M. Lemdani and E. Ould-Saïd, "A generalized L1-approach for a kernel estimator of conditional quantile with functional regressors: Consistency and asymptotic normality," *Statistics & Probability Letters*, vol. 79, pp. 1065–1073, 2009.
[19] H. Ding, Z. Lu, J. Zhang and R. Zhang, "Semi-functional partial linear quantile regression," *Statistics & Probability Letters*, vol. 142, pp. 92–101, 2018.
[20] C. Kuan, J. Yeh and Y. Hsu, "Assessing value at risk with CARE, the conditional autoregressive expectile models," *Journal of Econometrics*, vol. 150, pp. 261–270, 2009.
[21] W. Ehm, T. Gneiting, A. Jordan and F. Krüger, "Of quantiles and expectiles: Consistent scoring functions, choquet representations and forecast rankings," *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, vol. 78, pp. 505–562, 2016.
[22] X. Bian, S. Li, X. Shao and P. Liu, "Variable space boosting partial least squares for multivariate calibration of near-infrared spectroscopy," *Chemometrics and Intelligent Laboratory Systems*, vol. 158, pp. 174–179, 2016.
[23] P. Brown, T. Fearn and M. Vannucci, "Bayesian wavelet regression on curves with application to a spectroscopic calibration problem," *Journal of the American Statistical Association*, vol. 96, pp. 398–408, 2001.
[24] F. Ferraty, A. Laksaci, A. Tadj and P. Vieu, "Rate of uniform consistency for nonparametric estimates with functional variables," *Journal of Statistical Planning and Inference*, vol. 140, pp. 335–352, 2010.
[25] M. Mohammedi, S. Bouzebda and A. Laksaci, "The consistency and asymptotic normality of the kernel type expectile regression estimator for functional data," *Journal of Multivariate Analysis*, vol. 181, pp. 104673, 2021.
[26] G. Boente and A. Vahnovan, "Robust estimators in semi-functional partial linear regression models," *Journal of Multivariate Analysis*, vol. 154, pp. 59–84, 2017.
[27] I. Almanjahi, Z. Chiker-El-Mezouar, A. Laksaci and M. Rachdi, "kNN local linear estimation of the conditional cumulative distribution function: Dependent functional data case," *C. R. Math. Acad. Sci. Paris*, vol. 356, pp. 1036–1039, 2018.