Functional Data Regression Based on KLRFSIR and Bayesian Model Averaging

Longhui Mo [1], Huajun Xi [2], Yi Ding [3,*]

[1] College of Mathematics and Physics, Chengdu University of Technology, Chengdu, 610059, China

[2] College of Science, Southern University of Science and Technology, Shenzhen, 518055, China

[3] College of Marine Sciences, Shanghai Ocean University, Shanghai, 201306, China

* Corresponding Author Email: dy296283868@sina.com

Abstract. This paper focuses on an important direction in modern data analysis: functional data regression. Functional data is typically infinite-dimensional, and this paper specifically examines the case where the predictor variable is functional data and the response variable is a continuous scalar. Based on this, the paper propose a regression method that combines kernel localized and regularized functional sliced inverse regression with Bayesian model averaging. The innovation of this method lies in two aspects: firstly, compared to traditional sliced inverse regression, it captures more detailed local information through kernel clustering, providing better estimates of the reduced subspace when the number of slices is small; secondly, model averaging effectively balances the bias and variance of the prediction model, avoiding overfitting or underfitting issues. Empirical results show that the proposed method has smaller errors in terms of mean squared error and absolute error compared to some classical methods, and it demonstrates a certain level of robustness. Finally, this paper explore the potential application of the kernel clustering concept in other dimension reduction approaches based on covariance.

Keywords: Functional Data Regression, Sliced Inverse Regression, Bayesian Model Averaging, Kernel Localized Regularization.

1. Introduction

The concept of "functional data" was first proposed by Ramsay (1982), which refers to the records obtained when the potential generation process of data is a smooth function process. Functional data regression methods have shown strong power in modern data analysis, they can capture the dynamic changes of explanatory variables and have certain predictive advantages compared to traditional regression models. [1] In recent decades, functional data analysis has been widely applied in scientific fields such as population [2], environment [3] and transportation [4].

Although the actual observation of functional data is discrete, due to the short interval between time points, there is a strong collinearity between different time points. If statistical research is conducted directly, it is inappropriate [5]. In addition, if the dimension reduction techniques such as PCA and KPCA are used, the function information of the random function, such as smoothness and volatility, is not utilized. Common tools for functional data analysis are to reconstruct random functions using basis expansion methods. Wang et al. [4] and Zhao et al. [6] use B-spline basis expansion for functional variables; Zhang et al. [7] use Fourier basis to transform discrete data with functional characteristics into functional data. However, different types of basis functions will have a significant impact on the final prediction results. In addition, the selection of the number of basis functions is also a big problem. Functional principal component analysis (FPCA) is another tool for functional data analysis. [1] Specifically, it decomposes the covariance function to obtain the decomposition form of the random function and can theoretically prove that under certain conditions, the truncated functional principal component basis function expansion converges in probability to the real random function. However, FPCA is an unsupervised dimension reduction method, unrelated to the prediction task, so after projecting the infinite-dimensional function into the finite-dimensional function space, it cannot guarantee that the obtained principal component direction is effective for

the specific task. There are also other methods, such as functional linear discriminant analysis (FLDA) and functional partial least squares (FPLS), which can better capture the function information of the original predictive variables related to the response variables [8]. However, these methods cannot guarantee to maximize the preservation of the function information of the original predictive variables. In recent years, with the rapid development of functional data analysis, functional sufficient dimension reduction (FSDR) [9] has become an important focus of researchers. This method is based on the criterion of conditional independence, theoretically can retain all effective information of the original predictive variables on the response variables, and does not involve the connection function between the predictive variables and the response variables in the dimension reduction process, so this method is also model-free. In the two-stage functional data regression method, the selection of the prediction model is another important issue. As is known to all, a single prediction model is prone to overfitting or underfitting problems. Compared with a single prediction model, the model averaging method has the following advantages. First, a single model may not be able to extract the information reflected by the data well and may even lose some useful information, while model averaging can reduce this risk by combining the predictions of multiple models. Second, model averaging can also effectively balance the bias and variance of the prediction model. [10] For example, Shi Yang et al. [11] used model averaging methods to model China's commodity housing prices and Yu Mingzhe et al. [12] constructed a global scale flood medium and long-term forecast model based on the Bayesian model averaging method, and the results obtained were better than the prediction results of a single prediction model.

In conclusion, this paper proposes a functional data regression method that combines functional sufficient dimension reduction and model averaging. Specifically, first, considering that when the number of slices is relatively small, the classic functional sliced inverse regression may fail, this paper proposes a localized and regularized functional sliced inverse regression based on kernel clustering, the advantage of which is that it can capture more local function information through kernel clustering technology, and can effectively estimate the dimension reduction subspace when the number of slices is small. Second, since the connection structure between the sub-directions after dimension reduction and the response variables is unknown, this paper use the model averaging method instead of the model selection method and effectively balance the variance and bias of the prediction model through multi-model fusion. The results of actual data analysis show that: the proposed method has higher prediction accuracy and certain robustness compared to other comparison methods.

2. Theory and methods

2.1. Preliminary

Let (x, y) be random pair where $x \in L_2[0,1]$ and $y \in \mathbb{R}$. For simplicity, we assume $E[x] = \mu$, that the fourth moment X exists, $E\left[\| x \|^4\right] < \infty$. We denote the covariance operator of x by $\Gamma = E[(x-\mu) \otimes (x-\mu)]$ and the covariance operator of E[x|y] by $\Gamma_e = E\left[(E[x \mid y]-\mu) \otimes (E[x \mid y]-\mu)\right]$ where $\otimes \in \mathcal{L}(L_2[0,1], L_2[0,1])$ is defined as $(x \otimes y)z = \langle x, z \rangle y$, for any $x, y, z \in L_2[0,1]$. From the Karhunen–Loève expansion, we have

$$x = \sum_{i=1}^{\infty} \xi_i \phi_i \tag{1}$$

Where $E[\xi_i] = \lambda_i$ is the ith eigenvalue and ϕ_i is the corresponding eigenvector. Without loss of generality, we assume all eigenvalues are distinct and positive, and $\lambda_1 > \lambda_2 > \cdots > 0$. We focus on a subspace $\mathcal{S}$, spanned by K linear independent functions $\beta_1(t), \cdots, \beta_K(t)$, which is called the effective dimension reduction (EDR) space. Let $\Gamma\mathcal{S}$ be the space spanned by $\Gamma\beta_1, \cdots, \Gamma\beta_K$. Let $B_x = (\langle \beta_1, x \rangle, \cdots, \langle \beta_K, x \rangle)$. The estimation of FSIR is based on the following results with proofs omitted:

Lemma1 Suppose for all $b \in L_2[0,1]$, the conditional expectation $E[\langle b, x\rangle \mid B_x]$ is linear in $\langle \beta_1, x\rangle, \cdots, \langle \beta_K, x\rangle$. Then, $E[x \mid y] \in \Gamma\mathcal{S}$. If in addition $Var(x \mid B_x)$ is non-random, then $\Gamma - Var(x \mid y) \in \Gamma\mathcal{S}$.

In the statement of Lemma1, $Var(x \mid B_x) = E[(x - E(x \mid B_x)) \otimes (x - E[x \mid B_x]) \mid B_x]$ which is the conditional covariance operator. With abuse of notation, $\Gamma - Var(x \mid y) \in \Gamma\mathcal{S}$ means $(\Gamma - Var(x \mid y))\beta \in \Gamma\mathcal{S}$ for all $\beta \in L_2[0,1]$.

2.2. Kernel Localized and Regularized Functional Sliced Inverse Regression

We propose kernel localized and regularized functional sliced inverse regression (KLRFSIR). For any given samples $(x_1, y_1), \cdots, (x_n, y_n)$, the detailed implementation is described as follows:

(1) Estimate the operator Γ by:

$$\hat{\Gamma} = \frac{1}{n}\sum_{i=1}^{n} x_i \otimes x_i - \bar{x} \otimes \bar{x} \tag{2}$$

Where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

(2) Divide $\{y_1, \cdots, y_n\}$ into H evenly-spaced intervals $\{I_h\}_{h=1}^{H}$, where $I_h = (\tilde{y}_{s-1}, \tilde{y}_s]$ with $\min\{y_1, \cdots, y_n\} = \tilde{y}_0 < \tilde{y}_1 < \cdots < \tilde{y}_H = \max\{y_1, \cdots, y_n\}$.

For each sample (x_i, y_i) in interval I_i, find the k nearest neighbors(KNN) $x_{i,1}, \cdots, x_{i,k}$ of x_i in I_i with distance $d_j = dis(x_i, x_{i,j})$.

(3) Compute the weighted local mean of the sample (x_i, y_i) by:

$$\hat{\mu}_{i,loc} = \frac{1}{\sum_{j=1}^{k} w_j}\sum_{j=1}^{k} w_j x_{i,j} \tag{3}$$

Where $w_j = \exp(-d_j^2 / \sigma^2)$ and σ is a given parameter of the gaussian kernel function.

(4) Estimate localized $\Gamma_{e,loc}$ by:

$$\hat{\Gamma}_{e,loc} = \frac{1}{n}\sum_{i=1}^{n} (\hat{\mu}_{i,loc} - \bar{x}) \otimes (\hat{\mu}_{i,loc} - \bar{x}) \tag{3}$$

(5) Compute the spectral decomposition of $(\hat{\Gamma}_{\rho} I)^{-1}\hat{\Gamma}_{e,loc}$ where ρ is a regularization parameter, and then obtain n eigenvectors $\phi_1, \cdots, \phi_n$ and corresponding eigenvalues $\lambda_1, \cdots, \lambda_n$. Compute the optimal target dimension by cross-validation.

2.3. Bayesian Model Averaging

For any given sample $(y, X_1, \cdots, X_k)$, consider the linear regression model:

$$M: y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \varepsilon \tag{4}$$

Where $\varepsilon \sim N(0, \sigma^2)$ and $\beta = (\beta_1, \cdots, \beta_k)$ is the parameter of model. We use Bayesian Information Criterion (BIC) to analyze this model, which is

$$m = BIC = n\log(\hat{\sigma}_e^2) + k\log(n) \tag{5}$$

Where n is the number of data samples, k is the number of predictor variables in the regression model, and $\hat{\sigma}_e^2$ is the error variance, or Mean Squared Error (MSE):

$$\hat{\sigma}_e^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{6}$$

Then, the likelihood of this model is given by

$$P\{x \mid M\} = \exp(-m/2) \tag{7}$$

Suppose we have j linear regression models $M_1, \cdots, M_j$ with BIC $m_1, \cdots, m_j$. From Bayes theorem, we know

$$P\{M_i \mid x\} = \frac{P\{x \mid M_i\}P\{M_i\}}{\sum_{l=1}^{j} P\{x \mid M_l\}P\{M_i\}} = \frac{\exp(-m_i/2)}{\sum_{l=1}^{j}\exp(-m_l/2)} \tag{8}$$

For any given sample (x, y), suppose models $M_1, \cdots, M_j$ give predicted value $\hat{y}_1, \cdots, \hat{y}_j$. Then, we predict y by

$$\hat{y} = \sum_{l=1}^{j} P\{M_i \mid x\}\hat{y}_l = \sum_{l=1}^{j} \frac{\exp(-m_l/2)}{\sum_{k=1}^{j}\exp(-m_k/2)}\hat{y}_l \tag{9}$$

2.4. Improved Method Based on KLRFSIR and Bayesian Model Averaging

This study introduces a novel functional data regression approach that amalgamates KLRFSIR and Bayesian Model Averaging. To be specific, KLRFSIR aims to address the challenge of limited slices by leveraging kernel clusters to capture more local functional information. Additionally, the issue of the unknown relationship between the reduced-dimensional subspaces and the response variables necessitates the model averaging technique, benefiting from the combination of multiple models, which results in a better tradeoff between the variance and bias of predictive models. The flow chart of the proposed method is shown in Figure 1.

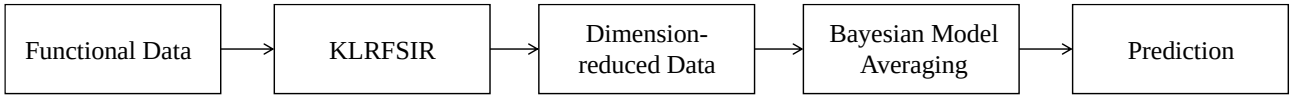

Figure 1. Flow chart of the proposed method

3. Data analysis

3.1. Data Sources and Experimental Settings

In this experiment, we used two datasets, one is the corn dataset, and the other is the meat dataset. The corn dataset was collected by Cargill, which includes 80 corn samples, using three instruments (m5, mp5, and mp6), with a wavelength range of 1100-2498nm. Each sample contains 700 absorbance data and includes the moisture content of each corn sample. The meat dataset is provided by Tecator company, downloaded from http://lib.stat.cmu.edu/datasets/tecator, with a total of 240 samples. The data is generated by Tecator Infratec Food and Feed Near Infrared Spectrometer in the wavelength range of 850-1050nm. Each sample includes 100 absorbance spectral data, as well as the moisture, fat, and protein content in the meat based on analytical chemistry determination.

By processing the spectral data of corn and meat using B-spline basis functions, the spectral curves of each sample were obtained as shown in Figures 2-3.

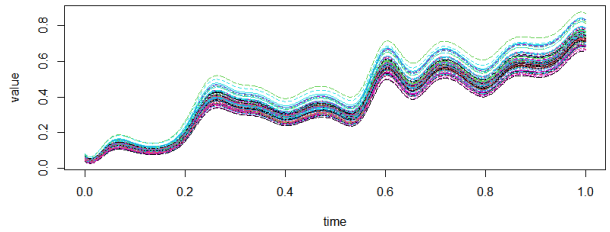

Figure 2. Spectral curves of corn data

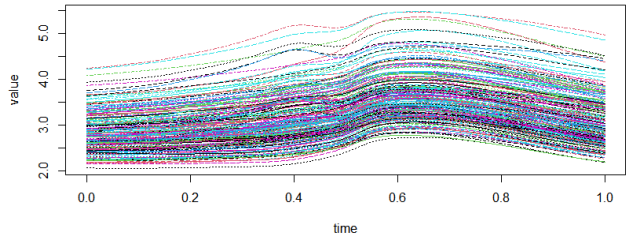

Figure 3. Spectral curves of meat data

The goal of this experiment is to reduce the dimensionality of the high-dimensional absorbance data in the corn and meat datasets, then use the moisture, fat, and protein content of corn and meat as independent variables, construct feature vectors using different dimension reduction methods ("BSPLINE", "PCA", "KLFSIR", "FSIR", "FSAVE", etc.), and then use these feature vectors as independent variables for analysis and prediction through the Generalized Linear Model (GLM). We evaluate the effectiveness and stability of various dimension reduction methods by comparing the mean squared error between the predicted values and the values determined by analytical chemistry. The specific formula for Mean Squared Error (MSE) is as follows:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 \tag{10}$$

The experiment was conducted on R version 4.3.2. The main packages used were FDA, FNN, BMA, MASS, glmnet, etc. The dimension of the feature vector is determined by the cumulative explained variance ratio, which is 98% in this experiment. The number of slices is determined by the method of cross-validation, initially set to 4. Regarding the corn dataset, 50 samples were randomly selected as the training set in the experiment, and the remaining samples were used as test data. Regarding the meat dataset, 140 samples were randomly selected as the training set in the experiment, and the remaining samples were used as test data. We repeated the above partitioning program 100 times to obtain the mean squared error in different simulation processes. The final calculated mean represents the accuracy of the algorithm, and the standard deviation represents the robustness of the algorithm. The smaller both are, the better the performance of the algorithm.

3.2. Analysis of experimental results

The comparison results are shown in Tables 1-6. Since the experiment was simulated 100 times, we compare the error distribution through boxplots, as shown in Figures 4-9.

Table 1. Comparison of MSE for corn m5 dataset

Index	BSPLINE	PCA	KLFSIR	FSIR	FSAVE
mean	0.109720073	0.089823177	0.053154008	0.094474595	0.091894234
sd	0.022657749	0.018364994	0.020981079	0.017543065	0.01695933

Table 2. Comparison of MSE for corn mp5 dataset

Index	BSPLINE	PCA	KLFSIR	FSIR	FSAVE
mean	0.106010179	0.097211605	0.057389065	0.10065846	0.097037736
sd	0.02094802	0.017234212	0.018712791	0.016951088	0.016736997

Table 3. Comparison of MSE for corn mp6 dataset

Index	BSPLINE	PCA	KLFSIR	FSIR	FSAVE
mean	0.107651248	0.098812996	0.053732258	0.102524777	0.099213403
sd	0.024087932	0.020455291	0.014490116	0.019736184	0.018647686

Table 4. Comparison of MSE for moisture prediction in meat dataset

Index	BSPLINE	PCA	KLFSIR	FSIR	FSAVE
mean	93.56118303	78.66850442	16.25126135	82.18606097	81.55127633
sd	10.90281036	8.368181928	3.879793141	8.462272336	8.357824453

Table 5. Comparison of MSE error for fat prediction in meat dataset

Index	BSPLINE	PCA	KLFSIR	FSIR	FSAVE
mean	159.3645572	159.36538023	27.82305092	137.1111964	135.7568493
sd	20.43513663	14.76033049	6.776371544	15.20394061	14.84438069

Table 6. Comparison of MSE for protein prediction in meat dataset

Index	BSPLINE	PCA	KLFSIR	FSIR	FSAVE
mean	10.08590838	8.026270643	3.91980429	8.650597104	8.511855755
sd	0.88562264	0.672660406	0.679046862	0.711641974	0.715648103

Comparing the prediction model constructed based on the feature vectors extracted by the KLFSIR method with the prediction results obtained by the feature vectors extracted by other methods, it can be seen that after dimension reduction of absorbance data of three different instruments by using the KLFSIR method, the prediction of moisture content in corn dataset and moisture content, fat content, and protein content in meat dataset is better than other dimension reduction methods. At the same time, the KLFSIR method also shows better robustness. The sufficient dimension reduction of the KLFSIR method is based on the independence criterion, which may make the data dimension reduction effect better than the BSPLINE model and PCA model. Secondly, by embedding the kernel clustering method, the KLFSIR method retains more local information of high-dimensional data than other methods, so the results are better than the FSIR and FSAVE methods.

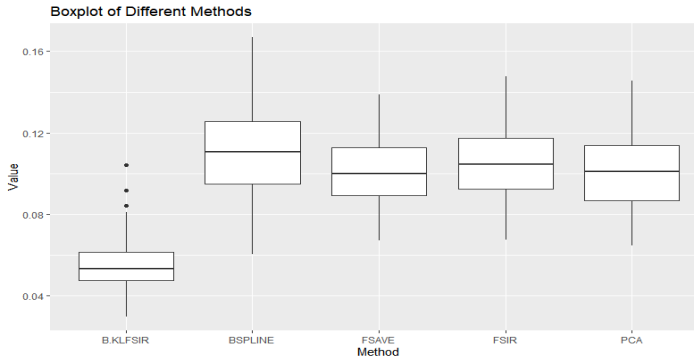

Figure 4. Comparison of MSE distribution for corn m5 dataset

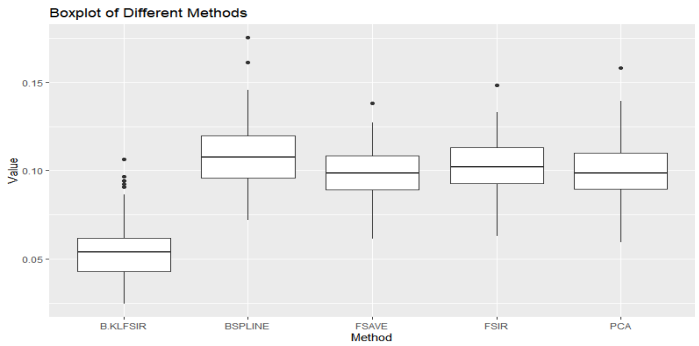

Figure 5. Comparison of MSE distribution for corn mp5 dataset

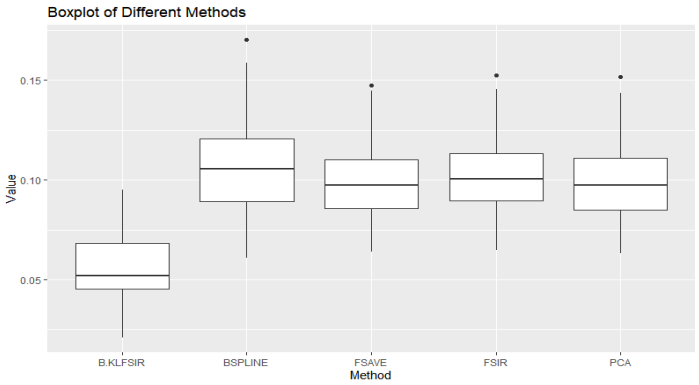

Figure 6. Comparison of MSE distribution for corn mp6 dataset

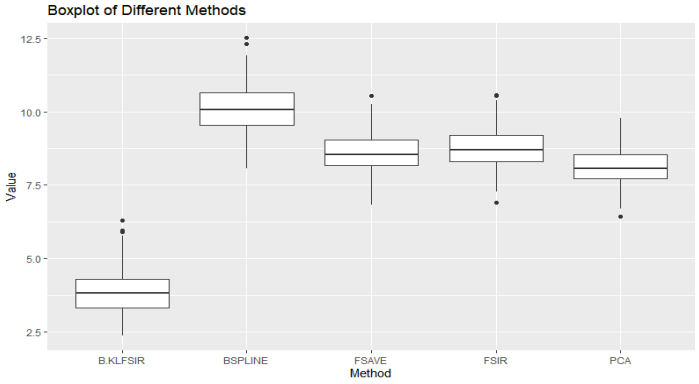

Figure 7. Comparison of MSE distribution for moisture prediction in meat dataset

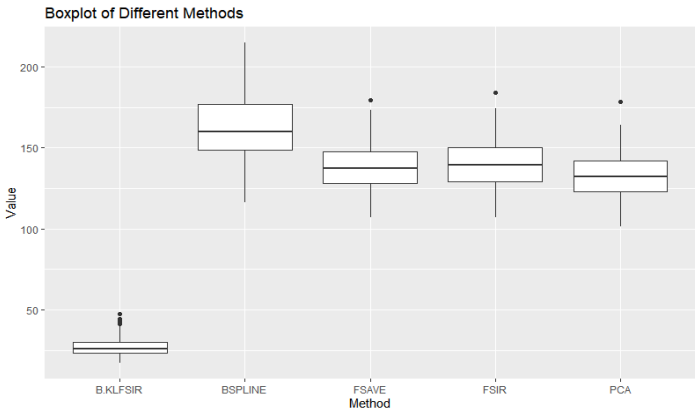

Figure 8. Comparison of MSE distribution for fat prediction in meat dataset

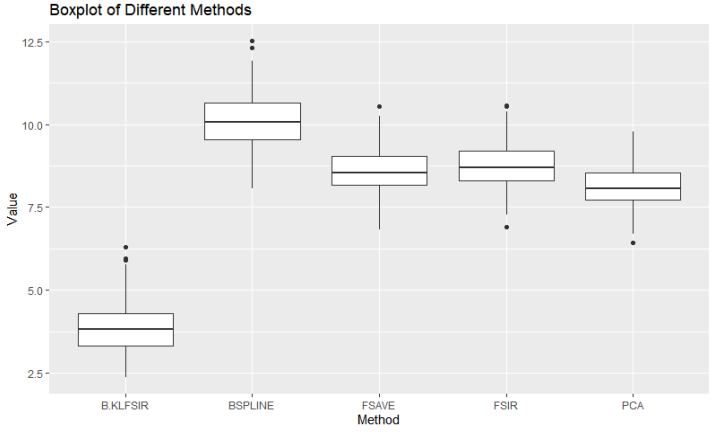

Figure 9. Comparison of MSE distribution for protein prediction in meat dataset

Finally, by introducing the Bayesian model averaging method, the variance and bias of the GLM model are well balanced, and the prediction results are further improved. In general, the KLFSIR method shows better dimension reduction effect and prediction performance in this experiment, which may be due to the combined effects of its independence criterion, kernel clustering embedding, and model averaging characteristics. This analysis helps to gain a deeper understanding of why the KLFSIR method is superior to other methods in this task and also provides valuable references for further research.

4. Conclusion

In conclusion, this study introduces a novel functional data regression approach that combines functional dimensionality reduction and model averaging. The proposed approach addresses the challenges associated with limited slices in classical functional sliced inverse regression by introducing a localized regularization method grounded in kernel clustering. Furthermore, the study opts for model averaging to extract more information reflected in the data and thus strikes a better balance between the variance and bias of predictive models.

Looking ahead, several promising directions for research and application emerge. Firstly, the proposed methodology could be extended to a wide range of real-world datasets, including biomedical research, environmental modeling, and so on. Moreover, various kernel functions, clustering algorithms, and different weighting schemes to optimize the ensemble's predictive performance can be incorporated into our method to enhance its adaptability to diverse data patterns. Additionally, evaluating the performance of this approach across diverse datasets and addressing the computational intricacies warrants continued investigation.

5. References

[1] Yan Mingyi. Statistical analysis of functional data: ideas, methods and applications [J]. Statistical Research, 2007 (02): 87-94. DOI:10.19343/j.cnki.11-1302/c.2007.02.016.

[2] Wei Yanhua, Ma Liping, Wang Bingcan. Research on population change trends and regional differences in China based on functional data [J]. Statistics and Decision, 2022, 38 (08): 82-86. DOI: 10.13546/j.cnki.tjyjc.2022.08.016.

[3] Li Ke, Liu Ya, Meng Zhenju et al. Spatial and temporal characteristics of air quality in Hunan Province-empirical analysis based on functional data [J / OL]. Journal of Natural Sciences of Hunan Normal University: 1-23 [2023-08-30]. http://kns.cnki.net/kcms/detail/43.1542.N.20230808.1222.002.html.

[4] Wang Desheng. Traffic flow characteristic analysis based on functional data [C] // China Association for Science and Technology, Ministry of Transport, Chinese Academy of Engineering, Hubei Provincial People's Government. Proceedings of the 2022 World Transport Conference (WTC2022) (Transportation Engineering and Air Transport Section). People's Communications Press Co., Ltd., 2022: 5. DOI:10.26914/c.cnkihy.2022.019779.

[5] Zhu Rong, Zou Guohua, Zhang Xinyu. Model averaging method for partial functional linear model [J]. System Science and Mathematics, 2018, 38 (07): 777-800.

[6] Zhao Xiaoya, Qiao Zhiying, Zhou Xiaomin et al. Processing and vibration prediction of heterogeneous time series data from multiple sources in cold rolling process based on FDA-LSTM [J]. Vibration and Shock, 2022, 41 (22): 202-210. DOI:10.13465/j.cnki.jvs.2022.22.024.

[7] Zhang Lin, Liu Jichun, Ma Jingyu et al. Short-term power forecasting method based on functional characteristic data of photovoltaic [J]. Electric Drive, 2021, 51 (12): 66-73. DOI:10.19457/j.1001-2095.dqcd21358.

[8] Guochang W, Xinyuan S. Functional Sufficient Dimension Reduction for Functional Data Classification [J]. Journal of Classification, 2018, 35 (2).

[9] LI K C. Sliced inverse regression for dimension Reduction [J]. Publications of the American Statistical Association, 1991, 86(414): 316-327.

[10] Zhang Xinyu, Zou Guohua. Model averaging method and its application in prediction [J]. Statistical Research, 2011, 28 (06): 97-102. DOI: 10.19343/j.cnki.11-1302/c.2011.06.018.

[11] Shi Yang, Zhao Shangwei. Prediction of China's housing prices under sudden public health events - based on model averaging method [J]. Journal of Minzu University of China (Natural Science Edition), 2022, 31 (01): 71-77.

[12] Yu Mingzhe, Liu Jianyu, You Yuanyuan et al. Evaluation of long-term forecasting of global floods based on Bayesian model averaging [J]. Geographical Sciences, 2022, 42 (09): 1646-1653. DOI:10.13249/j.cnki.sgs.2022.09.014.