Spatial Variable Co-estimation Method Based on Kriging Interpolation and Random Forest

Siyao Chu[a,*], Qunli Geng[b], Zhanbo Liang[c]

School of Computer Science and Technology, Zhoukou Normal University, Zhoukou, Henan, China

[a]2123432698@qq.com, [b]18538103475@163.com, [c]2085596390@qq.com

*Corresponding author

Keywords: Spatial variable co-estimation; Kriging interpolation; Kriging random forest; Estimation error

Abstract: Spatial variable co-estimation is crucial in engineering and fields like geology/mining. Traditional methods like Kriging have limitations with complex correlations. This paper studies a method for spatial attribute data co-estimation. It preprocesses data, divides it into grids, and interpolates. By comparing BN-NN and KRF, it finds KRF is more accurate and robust, and it effectively estimates variables with insufficient sampling. This research provides a more reliable tool for related fields, enabling more accurate spatial data analysis and prediction, which is of great significance for optimizing resource exploration, environmental monitoring, and urban planning, and promoting the development of these fields.

1. Introduction

With the development of technology, spatial statistics has become crucial in fields such as geological exploration, environmental monitoring, and agricultural research. However, the problem of limited sample data for spatial variable estimation persists. The co-estimation method, especially the co-Kriging approach, has been developed but faces challenges like high computational complexity. Fortunately, artificial intelligence and machine learning offer new ways to solve these issues[1].

Previous studies in this domain have numerous shortcomings. Ahn et al. focused only on single variables, disregarding valuable co-variable information[2]. Fuentes et al.'s use of linear models was ineffective in capturing nonlinear relationships in spatial data, leading to inaccurate estimations[3]. Hossain and Timmer lacked a comprehensive model and parameter comparison and optimization process, resulting in suboptimal model selection[4]. Chen and Zhuang failed to consider data heterogeneity, introducing biases[5]. Additionally, Tzoumas et al. couldn't handle outliers properly, further reducing the accuracy and reliability of the estimations[6].

This paper makes the following improvements and innovations. Firstly, a comprehensive data preprocessing method is proposed to ensure the quality and integrity of the data, which provides a solid foundation for subsequent analysis. Secondly, multiple methods are used to analyze the correlation between the target variable and the co-variables, and the most relevant co-variables are selected more accurately. Finally, by comparing different estimation methods, the Kriging random forest (KRF) method with higher accuracy and robustness is determined, which provides a more effective solution for spatial variable co-estimation.

2. Methodology

2.1. Sampling and Interpolation

In this spatial variable analysis, the research area is a square with X from 51250.0000m to 64500.0000m and Y from 78750.0000m to 92000.0000m, divided into 50m×50m grids for sampling. Data preprocessing involves collecting and loading the target and co-variable data, reshaping them into a matrix for consistency, and merging them. A resampling model (10% - 90%) creates datasets with unselected points as NaN. The "natural neighborhood interpolation" estimates unsampled values

  DOI: 10.25236/icmmct.2025.006

using the formula[7]:

$$f(x) = \sum_{i=1}^{n} w_i \cdot f(x_i) \tag{1}$$

where $f(x)$ is the estimate at x, w_i is the weight of x_i, and $f(x_i)$ is the known value at x_i.

2.2. Variable Correlation Analysis

To analyze the correlation between the target variable and potential co-variables, the random forest method is utilized. Before applying this method, the data is preprocessed by cleaning, feature selection, and standardization to ensure its quality and suitability for analysis[8].

The random forest model is constructed by recursively building decision trees. In this process, the best split point for each feature is determined to minimize the impurity of the child nodes after splitting. Gini impurity is calculated using the formula:

$$Gini(D) = 1 - \sum_{i=1}^{c} p_i^2 \tag{2}$$

where D is the data set, c is the number of categories, and p_i is the proportion of the i-th category samples in the data set. Information gain is calculated using the formula:

$$IG(D, A) = H(D) - \sum_{v \in A} \frac{|D^v|}{|D|} H(D^v) \tag{3}$$

where D is the data set, A is the feature, v is all possible values of the feature, D^v is the subset of the data set when the feature takes the value v, and $H(D)$ is the entropy of the data set:

$$H(D) = -\sum_{i=1}^{c} p_i \log_2 p_i \tag{4}$$

The feature importance is determined by calculating the contribution of each feature in the random forest model. For each decision tree, the out-of-bag samples (i.e., the unselected samples) are used for prediction, and the prediction error is calculated. The average of the out-of-bag errors of all decision trees is calculated. For each feature, its value is randomly shuffled, and the out-of-bag error is recalculated. The difference between the out-of-bag errors before and after shuffling is calculated. The greater the difference, the higher the importance of the feature. The formula is:

$$\text{Importance}(A) = \text{OOB\_Error}(A_{\text{shuffed}}) - \text{OOB\_Error}\left(A_{\text{original}}\right)$$

where OOB_Error is the out-of-bag error, A_{shuffed} is the feature with its value randomly shuffled, and A_{original} is the original feature.

When the node's sample number is below a threshold or its impurity is low enough, the decision tree construction stops. Pruning is done by setting parameters like max depth and min sample number to prevent overfitting.

For regression, the final prediction is the average of multiple decision trees' results; for classification, it's the majority vote. Feature importance is calculated by using out-of-bag samples for prediction, getting the average error, shuffling each feature's values, recalculating the error, and taking the difference between the original and shuffled errors.

2.3. Model Comparison and Selection

In this section, Kriging - random forest (KRF) and Bayesian neural network (BN - NN) are compared. For KRF, data loading and preprocessing are done first. Then, different sampling rates (0.1 - 0.9) are set for resampling the target and co-variables using rand sample function. The fitcensemble function trains the random forest model with selected parameters for stability. In prediction, an appropriate batch size (e.g., 1000) is set and results are combined[9],[10].

For BN - NN, after similar data handling, the fitcnb function trains the Bayesian network first. Then, its output and co-variables are used to train the neural network with fitnet function and selected hidden nodes. Batch prediction and error calculation steps are like KRF's.

Error metrics such as mean absolute error (MAE) is calculated using the formula:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{5}$$

where y_i is the true value, $\hat{y}_i$ is the predicted value, and n is the sample size.

Mean square error (MSE) is calculated using the formula:

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

Root mean square error (RMSE) is calculated using the formula:

$$RMSE = \sqrt{MSE} \tag{7}$$

The coefficient of determination (R^2) is calculated using the formula:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{8}$$

where $\bar{y}$ is the average value of the target variable.

3. Results and Discussion

3.1. Sampling and Interpolation Results

The distribution of the target variable across the research area is clearly depicted in the contour maps generated from the sampling and interpolation process , as shown in Figure 1. By analyzing the relationship between the sampling density and the mean absolute error (MAE), it is found that the error tends to decrease as the sampling density increases. For instance, at a 90% sampling density, the MAE is relatively low, suggesting a more accurate estimation. This is because a higher sampling density provides more data points, enabling a better representation of the variable's distribution. In contrast, at a 10% sampling rate, significant errors occur due to the scarcity of data points, as there is insufficient information to accurately capture the variable's behavior.

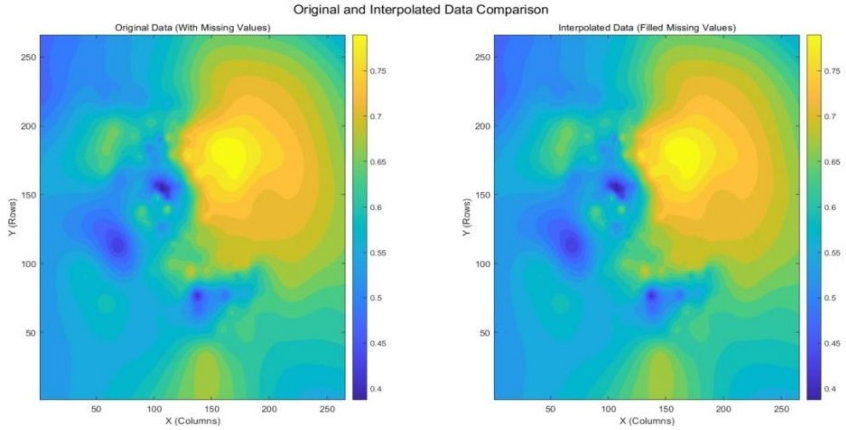

Figure 1 Interpolation Result Graph.

3.2. Variable Correlation Analysis Results

Using the random forest model, the correlations between the target variable and each co-variable are calculated. It is determined that co-variable 1 and co-variable 3 have the strongest correlations with the target variable , as shown in Figure 2. This finding is crucial as it allows for the selection of the most relevant co-variables for subsequent co-estimation. By focusing on these highly correlated

co-variables, the estimation accuracy can be improved since they are more likely to contain valuable information related to the target variable.

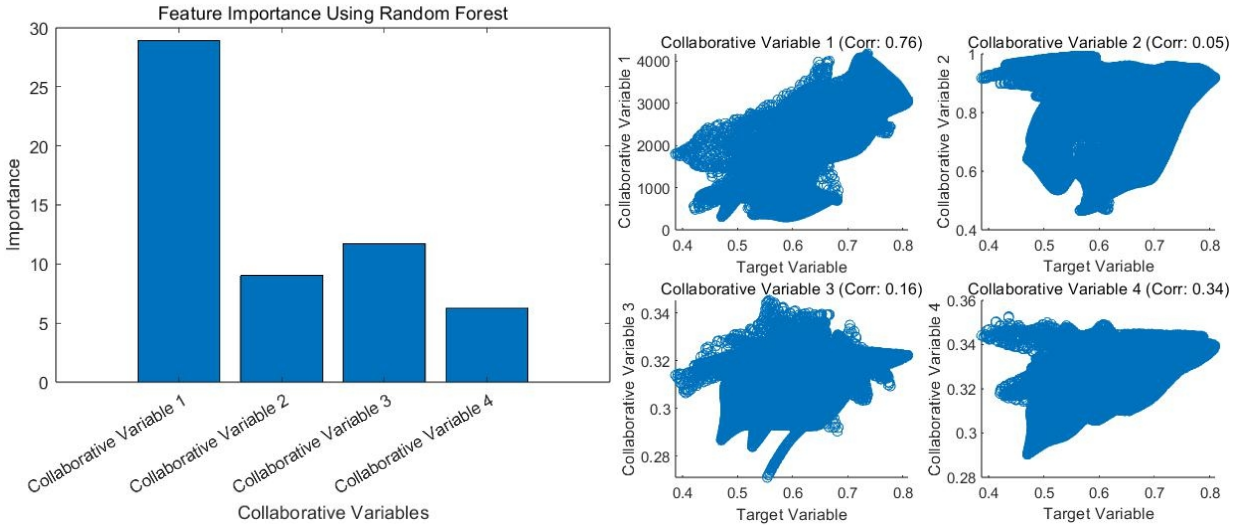

Figure 2 Collaborative Variable Correlation Graph

3.3. Model Comparison and Selection Results

The interpolation results of both the KRF and BN - NN methods at different sampling rates are presented. It is observed that with the increase of the sampling rate, the smoothness and accuracy of the interpolation results of both methods improve, as shown in Figure 3 and Figure 4. However, in terms of error metrics such as MAE, MSE, RMSE, and the coefficient of determination (R^2), the KRF method demonstrates better performance. Specifically, at low sampling rates, the KRF method maintains relatively low errors and better stability compared to the BN - NN method. The sensitivity analysis also indicates that the KRF method is less sensitive to the change of the sampling rate, which shows its robustness. In the robustness test, when the noise level increases, the KRF method shows smaller increases in the estimation error, further confirming its superiority in dealing with complex data conditions.

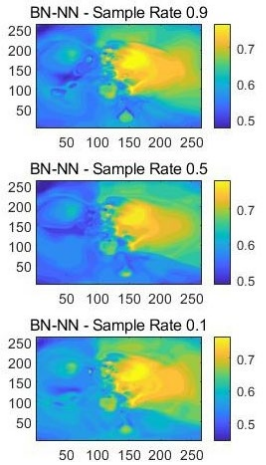

Figure 3 Bayesian Result Graph

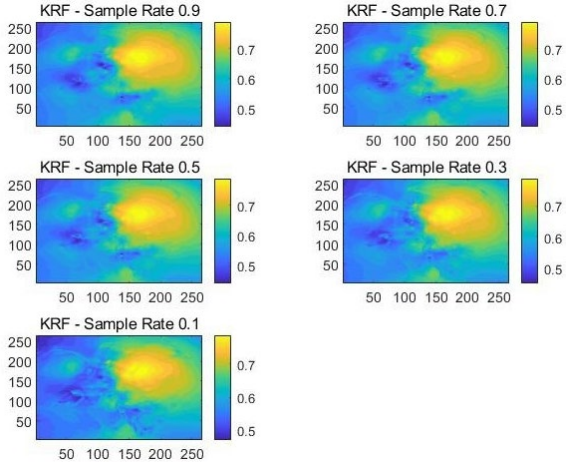

Figure 4 Interpolation Result Graph

3.4. Robustness Test Results

The average absolute errors (MAE) of BN - NN and KRF at different noise levels and sampling rates are shown in Figure 5. With the increase of the sampling rate, the MAE of both methods is significantly reduced, indicating that a higher sampling rate can significantly improve the accuracy of the interpolation results. When the noise level increases, the KRF method shows smaller increases in the estimation error compared to BN - NN, demonstrating its better robustness in dealing with complex data conditions.

3.5. Sensitivity Analysis Results

The estimation errors at different sampling rates are shown in Figure 6. With the increase of the sampling rate, the estimation error gradually decreases, indicating that a higher sampling rate can significantly improve the prediction accuracy of the model. The KRF method is less sensitive to the change of the sampling rate compared to BN - NN, further highlighting its stability and reliability in the estimation process.

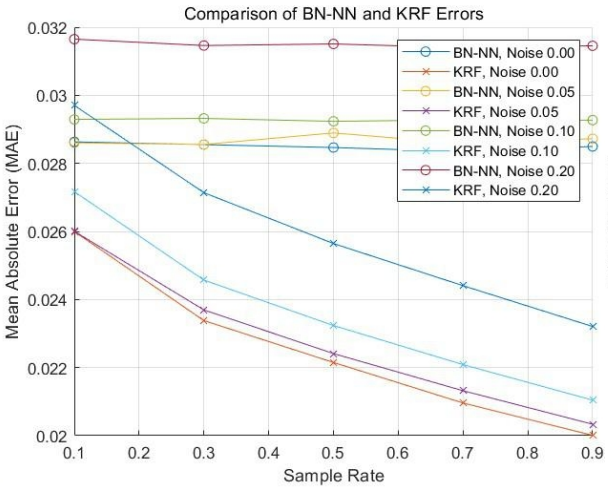

Figure 5 Robustness Test

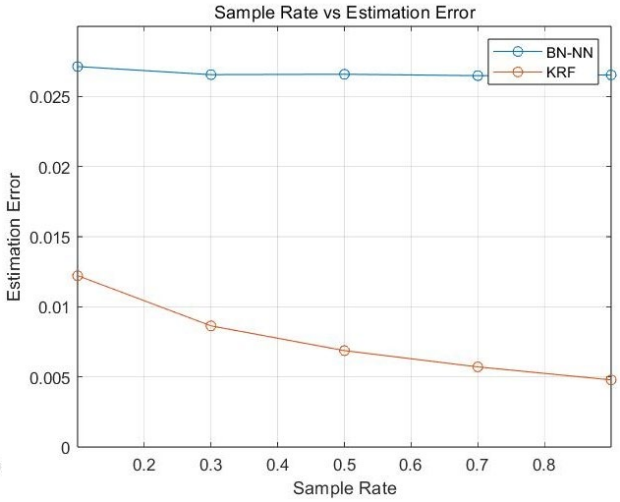

Figure 6 Sensitivity Analysis

In conclusion, the results of this study suggest that the KRF method is a more effective and reliable approach for spatial variable co-estimation, providing valuable insights and tools for related research and applications.

4. Conclusion

This paper studies the spatial variable co-estimation method based on Kriging interpolation and random forest. Through data preprocessing, model establishment and solution, and result analysis, it is found that the Kriging - random forest (KRF) method is an effective and optimal method for spatial variable co-estimation. It can effectively deal with the problems of limited sample data and complex spatial correlations, and has higher accuracy and robustness. This research provides a new method and idea for the field of spatial variable estimation, and has certain guiding significance for practical applications such as geological exploration and environmental monitoring. Future research can further explore the application of this method in different fields and improve the performance of the model.

References

[1] Zhang Z, Shi Y, Wang F, et al. Co-Kriging-guided interpolation for mapping forest aboveground biomass by integrating global ecosystem dynamics investigation and Sentinel-2 data. Remote Sens. 2022;14(23):5860.

[2] Ahn S, Ryu D-W, Lee S. A machine learning-based approach for spatial estimation using the spatial features of coordinate information. ISPRS Int J Geo-Inf. 2020;9(10):587.

[3] Fuentes M, Heaton MJ, Gneiting T. Nonlinear spatial time series modeling: New perspectives. Stat Methods Spat. Anal. 2024;45(2):58-71.

[4] Hossain MR, Timmer D. Machine learning model optimization with hyperparameter tuning approach. CORE. 2023;8(1):34-46.

[5] Chen X, Zhuang Y. Modeling Spatially Stratified Heterogeneous Data: Applications in Geospatial Analysis. Biometrics. 2023;68(2):331-340.

[6] Tzoumas V, Antonante P, Carlone L. Outlier-robust spatial perception: hardness, general-purpose algorithms, and guarantees. IEEE Int Conf Intell Robots Syst. 2020:5145-5152.

[7] Cotta B. A fast algorithm for natural neighbor interpolation. Geophys J Int. 1995;122(3):837-857.

[8] Strobl C, Boulesteix AL, Kneib T, Augustin T, Ziegler A. Conditional variable importance for random forests. BMC Bioinformatics. 2008;9:307.

[9] Alizadeh R, Imani F, Tsujii H, et al. A comparison study of surrogate models: Random forest and Kriging. J Model Simul Comput. 2023;14(3):214-228.

[10] Wu Z, Yao F, Zhang J, Liu H. Estimating forest aboveground biomass using a combination of geographical random forest and empirical Bayesian Kriging models. MDPIs. 2024;16(11):1859.