Research on Geotechnical Data Interpolation and Prediction Techniques

Haiyong Liu[1], Yangyang Chen[2(✉)], Lu Zhao[3], and Wen Liu[3]

[1] CCCC (Guangzhou) Construction Co., Ltd, Shenzhen, Guangzhou, China
[2] School of Civil and Hydraulic Engineering, Huazhong University of Science and Technology, Wuhan, Hubei, China
824890254@qq.com
[3] CCCC Wuhan Zhixing International Engineering Consulting Co., Ltd, Wuhan, Hubei, China

Abstract. The development of underground space is vital for urbanization and infrastructure projects. Prior to construction, comprehensive geological exploration is essential to ensure stability and safety. However, acquiring complete and accurate statistical data for project management is challenging, necessitating the handling of missing data to enhance reliability. Interpolation techniques are an effective way of dealing with incomplete data. This study presents a scalable framework for geotechnical data interpolation using machine learning. The framework employs different regression models to construct estimators and accurately interpolate geotechnical data. Key considerations include model selection and parameter optimization, with complete data used as the regression target. Five regression models, Bayesian Ridge Regression (BR), Extreme Gradient Boosting Tree (XGBoost), Support Vector Machine (SVR), Random Forest (RF) and K-Nearest Neighbour (KNN), were utilised. Estimators are constructed using the regression models and iterative interpolation is used to estimate missing values for geotechnical data, with each feature treated as a result of using the different estimators. The framework is evaluated through k-fold cross-validation, demonstrating its effectiveness in imputing missing values. The interpolation results using the SVR model indicate good conformity with the original data, confirming the method's effectiveness in capturing underlying patterns. This scalable framework bridges the gap in geotechnical data interpolation research, providing a reliable solution. The proposed approach contributes to the accurate and robust interpolation of geotechnical data, facilitating informed decision-making in underground construction projects.

Keywords: underground space · geological exploration · missing data · geotechnical data interpolation · machine learning · regression models

1 Introduction

The development and utilization of underground space have gained significant importance due to the increasing demands of urbanization and infrastructure development [1]. Prior to the construction of underground works, conducting comprehensive geological exploration is crucial to ensure the stability, safety, and long-term performance of

S. H. B. D. M. Zailani et al. (Eds.): ICMSEM 2023, 259, pp. 1788–1795, 2024.
https://doi.org/10.2991/978-94-6463-256-9_182

such projects. Geological exploration data serves as a fundamental reference for various aspects of underground construction, including design, construction methodologies, operation, and maintenance [2]. However, acquiring complete and accurate statistical data for project construction and management can be challenging in practice. Consequently, dealing with missing data becomes imperative to enhance the reliability and effectiveness of the implemented program [3].

Missing data in geotechnical engineering can be classified into three categories based on their underlying mechanisms: completely random missing, random missing, and non-random missing [4]. Each category poses unique challenges in terms of data analysis and interpretation. To address these challenges, a wide range of methods for interpolating missing data have been proposed and extensively studied in the literature [5–7]. These methods encompass mean replacement techniques, regression-based approaches, expectation maximization methods, and more. However, despite the substantial body of research on data interpolation, the application of geotechnical data interpolation methods remains an area that requires further exploration and comprehensive investigation [8].

In this study, we propose a scalable framework for geotechnical data interpolation, utilizing advanced machine learning methods. Machine learning techniques have shown great promise in handling complex patterns and relationships within data, making them well-suited for addressing the challenges of geotechnical data interpolation [9]. The framework offers a systematic approach for constructing estimators using various regression models. By carefully configuring the parameters of these estimators and utilizing complete data as the regression prediction target, the proposed framework ensures accurate and reliable interpolation of geotechnical data.

The construction of accurate estimators within the framework involves several key considerations. Firstly, the selection of appropriate regression models plays a crucial role in capturing the underlying patterns and characteristics of geotechnical data. Different regression models may be more suitable for specific types of data, and their performance needs to be evaluated within the context of geotechnical engineering. Secondly, determining the optimal parameter values for the estimators is essential to achieve accurate interpolation results. The parameter selection process requires careful analysis and experimentation to strike a balance between overfitting and underfitting. Finally, leveraging the complete data as the regression prediction target enables the estimators to capture the true behavior and trends within the geotechnical data, facilitating robust interpolation.

2 Methodology

This study focuses on the application of iterative interpolation as a method for imputing missing values. The core principle behind this approach is to utilize each feature as an output in a round-robin fashion, employing various estimators to perform regression [10]. To construct the estimators in this study, five regression models, BR, XGBoost, SVR, RF and KNN, were used.

In order to establish exit conditions for the estimator, we introduced two key parameters: the number of iterations and the tolerance. These parameters serve as criteria for

determining when the estimation process should be terminated. The iterative nature of the approach allows for refining the imputed values gradually, enhancing their accuracy over multiple iterations.

Given the inherent imbalanced and insufficient nature of the data, a k-fold cross-validation (CV) technique was employed in this study. By utilizing this method, we were able to assess the performance of each estimator. During the k-fold CV, the data is divided into k subsets, or folds, and the estimation process is repeated k times. This allows for a comprehensive evaluation of the estimator's performance across different subsets of the data. The scores obtained from the k-fold CV provide a robust measure of the estimators' effectiveness in imputing missing values.

To provide a visual representation of the specific steps followed in this study, we present **Fig. 1**. This figure outlines the sequential process involved in the iterative interpolation approach. It serves as a roadmap, guiding the reader through the methodology employed and facilitating a clear understanding of the experimental setup.

The evaluation indicators were calculated using the Mean Squared Error (MSE) method [11]:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_{i,actual} - y_{i,predicted}\right)^2 \tag{1}$$

where n is the number of samples, $y_{i,actual}$ is the i-th actual value and $y_{i,predicted}$ is the predicted value corresponding to the i-th actual value.

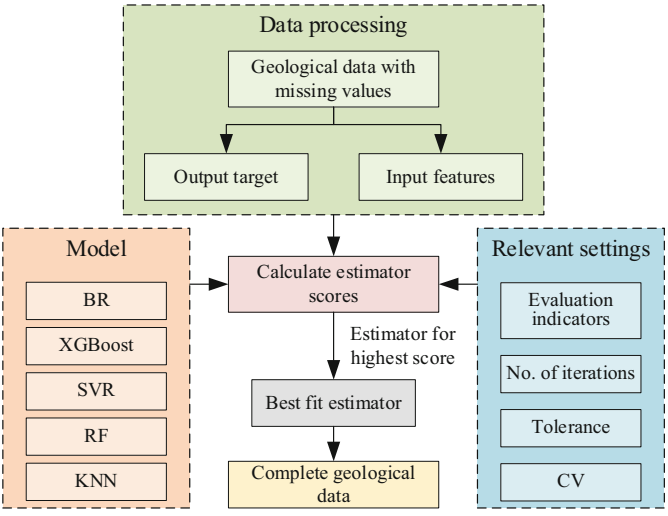

Fig. 1. Flow chart of the geotechnical data interpolation technique.

Table 1. Percentage of missing values for the characteristics of the fifteen features.

Feature	ρ	ρd	GS	ω	e	n	Sr	WL	Wp	Ip	IL	c	φ	Es	Kv
Missing percentage (%)	23	22	22	0	22	22	22	0	0	0	0	54	54	22	22

3 Case Study

3.1 Datasets

The effectiveness of the proposed method in predicting and interpolating geotechnical engineering data is verified through practical engineering cases. The dataset used in this study comprises fifteen features, namely: natural density (ρ); dry density (ρd); specific gravity of solid particles (GS); natural moisture content (ω); natural porosity (e); porosity (n); saturation (Sr); liquid limit (WL); plastic limit (Wp); plasticity index (Ip); liquidity index (IL); cohesion (c); angle of internal friction (φ); compression modulus (Es); coefficient of vertical subgrade reaction (Kv), as detailed in Table 1.

Among these features, ω, WL, Wp, Ipand IL were recorded as complete, with no missing values. However, both c and φ exhibited a significant amount of missing data, with 52.43% of their values unavailable. These two parameters represent the most severely affected features in the dataset. This study refines the missing values in the geotechnical data using data interpolation techniques. To verify the validity of the scheme, two complete data were selected as prediction targets, namely WL and IL. The remaining data, ω, Wp, Ip and other data after interpolation were selected as inputs. The predictive ability for WL and IL reflects the performance of the constructed interpolated estimator for geotechnical data.

To visualize the distribution of missing values across the 100 samples, **Fig. 2** displays the dataset with missing value markers. In the figure, white areas indicate the regions where data is missing or recorded as incomplete. This visualization highlights the extent and pattern of missing values within the geotechnical dataset.

3.2 Missing Value Interpolation and Prediction Results Analysis

To validate the approach proposed in this study, two complete data were selected as prediction targets, i.e. WL and IL. The remaining data, i.e. the complete ω, Wp, Ip and other data after interpolation, were selected as inputs.The predictive performance of the interpolation estimator is shown in the following table. The 10-fold cross-validated prediction results for WL and IL are shown in **Figs. 3** and **4**. Looking at **Figs. 3** and **4**, it is evident that there are differences in the mean and error (standard deviation) of the MSE scoring indices across the prediction results produced by the five models.

Specifically, as shown in **Fig. 3**, the SVR model showed the smallest mean MSE of 2.57, while the BR model showed the highest mean MSE of 3.54. The SVR model showed the smallest error of 1.09, while the BR model showed the highest error of 3.54. Overall, the interpolated estimator of geotechnical parameters constructed using SVR predicted WL better and showed better robustness.

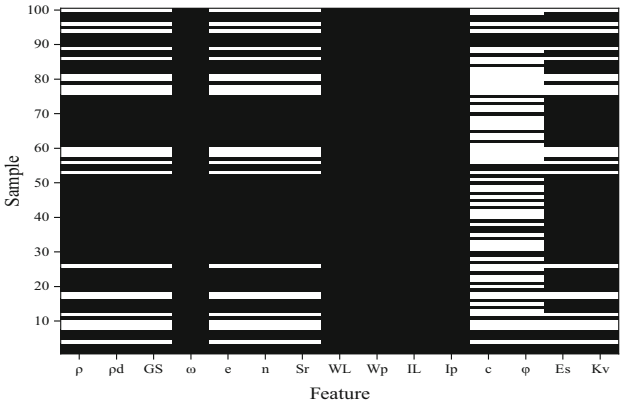

Fig. 2. Missing value markers from 1 to 100 samples.

Furthermore, **Fig. 4** shows that for the IL model predictions, the RF model shows the smallest mean MSE of 0.1457 and XGBoost shows the highest mean MSE of 0.3005. SVR shows the smallest error of 0.0123 and XGBoost shows the highest error of 0.028. This indicates that the geotechnical data estimator constructed using SVR is able to accurately predict IL data and exhibits excellent robustness.

When considering a specific target output, it is essential to select an appropriate model to achieve better interpolation of geotechnical data. Each model possesses distinct strengths and limitations, making careful consideration necessary to ensure optimal performance for a given target variable.

Given the satisfactory stability of the SVR model, it was selected to construct an estimator for the interpolation of geotechnical data. The results of this interpolation process are presented in **Fig. 5**, where the red points represent the original data, and the blue points represent the interpolated data. The conformity between the interpolated and original data points in **Fig. 5** serves as evidence supporting the effectiveness of the proposed geotechnical data interpolation method. The absence of pronounced anomalies or irregularities in the interpolated data suggests that the SVR-based estimator successfully captures and reproduces the underlying patterns of the geotechnical data.

4 Conclusions

This study proposes a framework for interpolation of geotechnical data and validates the interpolation using complete data.The main results of which are as follows:

(1) The proposed scalable framework based on machine learning methods successfully addresses the challenge of geotechnical data interpolation. Quantitatively, the SVR model demonstrates superior performance with the smallest MSE, while the KNN model exhibits low error. This indicates the reliability and accuracy of the interpolation results.

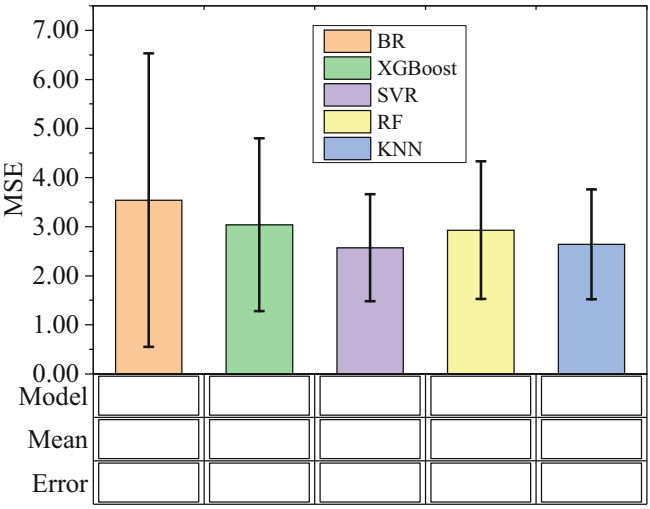

Fig. 3. WL's predicted results.

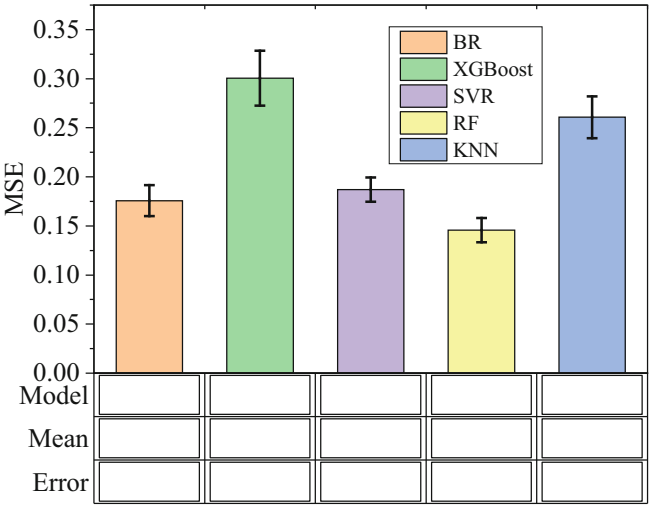

Fig. 4. IL's predicted results.

(2) The qualitative analysis reveals that the interpolated data aligns well with the distribution of the original data, indicating the ability of the framework to capture the underlying patterns of geotechnical data. The absence of pronounced anomalies or irregularities further supports the effectiveness of the proposed geotechnical data interpolation method.

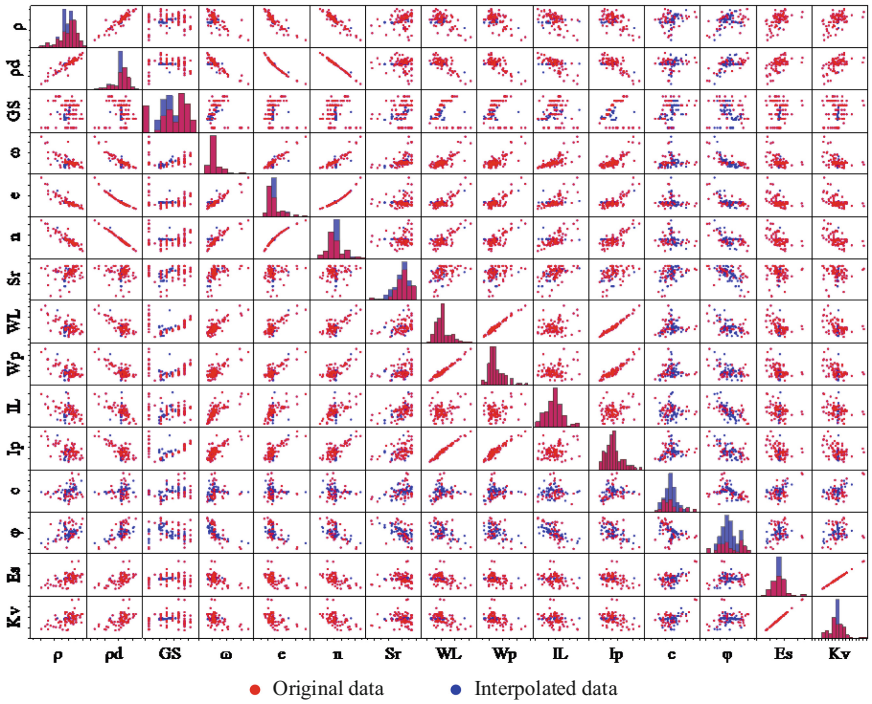

Fig. 5. Original geological data and interpolated geological data.

(3) The framework provides a systematic approach for constructing estimators using different regression models and setting appropriate parameters. Leveraging complete data as the regression prediction target ensures accurate and robust interpolation of geotechnical data. This framework can enhance the reliability and effectiveness of geotechnical projects by providing accurate estimations for missing data.

References

1. Liu, W.L., Chen, E.J., Yao, E.L., Wang, Y.Y., & Chen, Y.Y. (2021). Reliability analysis of face stability for tunnel excavation in a dependent system. Reliability Engineering & System Safety, 206, 107306. https://doi.org/10.1016/j.ress.2020.107306.
2. Tao, Y. (2022) Data on prediction of geological characteristics during shield tunnelling in mixed soil and rock ground. Data in Brief, 45: 108726. https://doi.org/10.1016/j.dib.2022.108726.
3. Zona, A., Kammouh, O., & Cimellaro, G.P. (2020). Resourcefulness quantification approach for resilient communities and countries. International Journal of Disaster Risk Reduction, 46, 101509. https://doi.org/10.1016/j.ijdrr.2020.101509.
4. Jafrasteh, B., Hernández-Lobato, D., Lubián-López, S.P., & Benavente-Fernández, I. (2023). Gaussian processes for missing value imputation. Knowledge-Based Systems, 273, 110603. https://doi.org/10.1016/j.knosys.2023.110603.

5. Cao, J.J., Yao, G., & Da Silva, N.V. (2022). Interpolation of irregularly sampled noisy seismic data with the nonconvex regularization and proximal method. Pure and Applied Geophysics, 179(2), 663–678. https://doi.org/10.1016/j.jappgeo.2023.105073.
6. Xu, C., Wang, J., Hu, M., & Wang, W. (2022). A new method for interpolation of missing air quality data at monitor stations. Environment International, 169, 107538. https://doi.org/10.1016/j.envint.2022.107538.
7. Tang, J., Xia, H., Aljerf, L., Wang, D., & Ukaogo, P.O. (2022). Prediction of dioxin emission from municipal solid waste incineration based on expansion, interpolation, and selection for small samples. Journal of Environmental Chemical Engineering, 10(5), 108314. https://doi.org/10.1016/j.jece.2022.108314.
8. Kim, H.S., & Ji, Y. (2022). Three-dimensional geotechnical-layer mapping in Seoul using borehole database and deep neural network-based model. Engineering Geology, 297, 106489. https://doi.org/10.1016/j.enggeo.2021.106489.
9. Liu, W.L., Li, A., Fang, W.L., Love, P.E., Hartmann, T., & Luo, H.B. (2023). A hybrid data-driven model for geotechnical reliability analysis. Reliability Engineering & System Safety, 231, 108985. https://doi.org/10.1016/j.ress.2022.108985.
10. Swami, A., & Jain, R. (2013). Scikit-learn: machine learning in python. Journal of Machine Learning Research, 12(10), 2825–2830. https://doi.org/10.48550/arXiv.1201.0490.
11. Liu, W. L., Li, A., & Liu, C.J.(2022). Multi-objective optimization control for tunnel boring machine performance improvement under uncertainty. Automation in construction, 139. https://doi.org/10.1016/j.autcon.2022.104310.