An Inversion Approach for Salt Content in Simulated Murals Based on Spectral Enhancement and Sample Partitioning

Qinghao Dong [1,2], Shuqiang Lyu [1,2], Miaole Hou [1,2,*], Yanzhu Jin [1], Xinyi Li [1]

[1] School of Geomatics and Urban Spatial Informatics, Beijing University of Civil Engineering and Architecture, No.15 Yongyuan Road, Daxing District, Beijing, China -201304020210@stu.bucea.edu.cn, lvshuqiang@bucea.edu.cn, houmiaole@bucea.edu.cn, 202306020112@stu.bucea.edu.cn, 202306020111@stu.bucea.edu.cn
[2] Beijing Key Laboratory for Architectural Heritage Fine Reconstruction & Health Monitoring, No.15 Yongyuan Road, Daxing District, Beijing, China

Keywords: Mural Disruption, Salt Content, Inversion, Sample Partitioning.

Abstract

The recurrent crystallization and dissolution of salts within murals lead to significant internal structural damage, ultimately causing paint loss and compromising mural integrity. This study explores the effectiveness of five training dataset partitioning methods in improving the accuracy of models designed to non-invasively predict salt content in murals using spectral data collected across 350 nm–2500 nm. Firstly, spectra were acquired from laboratory-simulated murals using a spectroradiometer, followed by smoothing and denoising via the Savitzky-Golay (S-G) algorithm. To further enhance salt-related spectral features and eliminate baseline drift, both first-order and second-order differentiation techniques were applied. Secondly the performance of five partitioning strategies—Random Selection (RS), Kennard-Stone (KS), Sample Set Partitioning Based on Joint X-Y Distance (SPXY), Kernel Distance-Based Sample Set Partitioning Based on Joint X-Y Distance (KSPXY), and Sample Set Partitioning Based on Joint X-Y-E Distances (SPXYE)—was evaluated. A salt content inversion model was then developed using Random Forest (RF) and Partial Least Squares Regression (PLSR). Results showed that PLSR, combined with KSPXY partitioning and first-order derivative enhancement, achieved the best predictive performance. Validation of the model with a test dataset yielded the RMSE of 0.068 and the R^2 of 0.954, indicating high accuracy. Our findings underscore the pivotal role of sample partitioning method selection in enhancing model accuracy and predictive outcomes. This study provided an effective technique for the inversion of mural salt content non-invasively, which would facilitate the preservation of these invaluable cultural artifacts.

1. Introduction

A mural is a form of painting art created on walls for decorative or other purposes. It holds significant research value, as it contains a wealth of historical information, including politics, economics, culture, science and technology, as well as production techniques(Guo et al., 2023) . However, with temperature changes, soluble salts migrate to the surface through capillary channels in the murals under humid conditions. After repeated cycles of dissolution and crystallization, the bonding force within the ground preparation layer is weakened, leading to loosening, detachment, scattering, or peeling of the pigment layer and consequently resulting in common deterioration phenomena such as plaster disruption, blistering, and salt efflorescence. Therefore, salt content testing of frescoes is very important for the prevention of salt damage and the preservation of murals.

Yu found a large amount of NaCl both inside and outside the mural walls using ion chromatography (IC). They conducted a comprehensive study on the water vapor distribution inside the mural walls by combining high-density electrical methods, microwave humidity measurement, and thermal infrared imaging with temperature and humidity monitoring results(Yuet al., 2017). Their findings confirmed that water vapor activity poses a serious threat to the safe preservation of the mural in the temple. Based on Raman spectral analysis, Madariaga discussed the formation of weathering layers on the mural of Pompeii, Italy(Madariaga et al., 2014). They explained how the chemical reaction of acidic gases with the wall materials led to the formation of salt crystals, which further exacerbated the deterioration of the mural. Sawdy and Price used Raman Inductively Coupled Plasma Atomic Emission Spectrometry (ICP-AES) combined with thermodynamic modelling software to analyses the salts in the murals. They found that the repeated crystallization and dissolution of potassium nitrate at high humidity was the main cause of damage to the mural at Cleeve Abbey, UK(Sawdy and Price, 2005). Gil et al. used optical microscopy (OM), scanning electron microscopy (SEM-EDS), μ-Raman, and Fourier Transform Infrared Spectroscopy (μ-FTIR) to identify the main salts in the walls as calcium carbonate and calcium-magnesium carbonate(Gil et al., 2015). The deposition and crystallization of these salts, along with the growth of fungal hyphae, led to cracking, flaking, and loss of adhesion of the coating. The paint layer containing malachite and staurolite was the most affected part.

In recent years, hyperspectral technology has been applied to the inversion of the material composition of mural, showing promising prospects for the inversion of salt content due to its high spectral resolution and spectral continuity. Guo experimentally simulated mural painting samples with different salt concentrations and compared the accuracy of the spectral salt index after continuum removal, first-order differentiation, and Savitzky-Golay (SG) smoothing filtering. Based on the extracted characteristic bands, Partial Least Squares Regression (PLSR), Support Vector Regression (SVR), and Random Forest (RF) models were established to estimate salt concentration(Guo et al., 2023). Ren conducted a simulation experiment to construct solutions with different concentrations of sodium dihydrogen phosphate dodecahydrate to simulate a salt-hazardous environment. They enhanced the spectral features using fractional-order differentiation and established bi- and tri-band spectral indices. Based on these indices, they developed univariate mural conductivity inversion models and constructed a conductivity-based inversion model for mural phosphate content using partial least squares(Ren and Liu, 2024).

Currently, research on the spectral characterization of salt content in mural in China is still in its early stages, with significant differences in property estimation models for different soluble salts. As an important way into the display of China's ancient culture, mural have a long history, substantial preservation, and wide distribution, making them of great research and preservation value. Therefore, the study of spectral characteristics and attribute inversion of ancient mural is of great scientific and practical significance. To improve model accuracy, many scholars have applied various spectral transformation methods to process spectral data. However, fewer scholars have explored different sample set partitioning methods for estimating the soluble salt content of mural or combined spectral transformations with sample partitioning methods to build models and analyse their impact on estimation accuracy. To explore the effects of different sample set partitioning methods and spectral transformations on model accuracy, this study uses spectral reflectance data to develop models based on two modelling methods—partial least squares regression and random forest. The study aims to: (1) compare and analyse the effects of five sample set partitioning methods on the estimation results of mural's dissolved salt content in simulation; (2) compare the accuracy of different spectral enhancement methods combined with different sample set partitioning methods for salt content modelling; and (3) evaluate the reliability of estimating the soluble salt content of mural using reflectance spectroscopy.

2. Materials and Methods

2.1 Simulation of Murals and Acquisition of Spectral Reflectance Data

This experiment takes the temple murals as the prototype, with the simulated mural representing the structure of the interaction between the support body and the paint layer. The interaction serves as the main location for soluble salt transportation, including the coarse plaster layer and the fine plaster layer. According to the research murals production process , Dunhuang soil, loess, clarified board soil, Dunhuang sand, wheat straw and hemp rope are used as raw materials(Yao et al., 2023). Sufficient amount of deionized water is added to remove soluble salts in the materials, and desalination treatment is performed to produce simulated mural. Anhydrous sodium sulfate was chosen as the soluble salt, and the salt-to-soil ratio ranged from 0% to 1%, with each concentration set at intervals of 0.05%. A total of 60 simulated mural samples were prepared, with three samples for each of the 20 concentration levels. The salts were weighed and mixed with deionized water to fully dissolve. The mixture was then thoroughly mixed with sandy soil, hemp rope, and other raw materials. It was placed into round moulds with a height of 18mm and a diameter of 90mm. The mixture was levelled and dried in the shade at room temperature, away from light. During the experiment, the humidity and temperature of the simulated mural were continuously monitored using an MDN-6813 soil detector. This ensured that the concentration of sodium sulfate was the only variable. Several simulated murals with different concentrations of Na_2SO_4 were prepared in a laboratory setting, as shown in Figure 1.

Figure 1. Laboratory-simulated murals containing different concentrations of Na_2SO_4

The spectral reflectance of the murals surface was measured using the ASD-FieldSpec4 Hi-Res Spectroradiometer, manufactured by Analytical Spectroscopy Equipment, USA. The wavelength range of the instrument is 350–2500 nm, with sampling intervals of 1.4 nm (350–1000 nm) and 1.1 nm (1001–2500 nm). The data collection environment was a dark room after sunset, with the only light source being a 70W quartz-tungsten-halogen lamp provided with the contact probe. First, the spectral reflectance was normalized with a white standard reflectance measured on a Spectralon plate (Labsphere, Inc., North Sutton, NH, USA). Each laboratory-prepared simulated mural was measured 15 times to prevent ambient light interference and reduce operational errors. Snap the contact probe onto the surface of the mural and wait for the curve to stabilize before collecting the reflectance data. Then, rotate the probe by 90° and measure again, taking four measurements at each position.

2.2 Spectral Data Enhancement

To reduce systematic error, the spectral data were corrected for breakpoints, and the four measurements were averaged. A total of 102 spectral data points was obtained. The noisy front and rear 50 bands of each spectrum were removed, yielding the original spectra R. Savitzky-Golay smoothing and filtering were performed on R. SG smoothing is a mathematical method used for data smoothing and noise suppression, which is able to retain the original features of the data to a larger extent. A third-order polynomial is used for smoothing with a window width of 5. Spectral differentiation is a commonly used mathematical method in spectral analysis that can reveal features not easily detected in the original reflectance. It effectively removes baseline drift and noise, reduces background interference, highlights weak signals or subtle features, and improves target feature recognition. The smoothed results were processed using the First Derivative (FD) and Second Derivative (SD). The four types of datasets obtained include: original spectral reflectance (R), spectra smoothed by SG, smoothed first-order differential spectra (SG-FD), and smoothed second-order differential spectra (SG-SD). The spectra before and after processing are shown in Figure 2.

2.3 Sample Set Segmentation Methods

In this experiment, the sample set partitioning method refers to dividing the full set of spectral reflectance data into a training set and a calibration set. The dataset contains 102 spectral curves and their corresponding soluble salt content each derived from a

single measurement of a simulated mural. There is one dependent variable, salt content, and 2051 independent variables, which are the reflectance values in 2051 bands (the number of bands will be reduced after differentiation). In addition to the commonly used Random Selection (RS) method, the Kennard-Stone (KS) algorithm, and the Sample Set Partitioning based on Joint X-Y Distance (SPXY) algorithm, two derivations of the SPXY method were selected: Kernel Distance-Based Sample Set Partitioning based on Joint X-Y Distance (KSPXY) and Sample Set Partitioning based on Joint X-Y-E Distance (SPXYE). In the RS method, 1000 random samples were used for regression modelling and prediction evaluation to fully verify the performance of the method. The KS algorithm is the most widely used sample partitioning algorithm. It calculates the Euclidean distances of the independent variables between two data points, selects the two data points with the greatest distance as the cluster seeds, and then clusters the data based on the minimum Euclidean distances. This process continues until the number of samples in the calibration set reaches the threshold value, with all remaining data grouped into the training set. SPXY is based on the KS algorithm and introduces the Euclidean distance between dependent variables. In this method, the distance formula used for dataset partitioning combines the Euclidean distance between the dependent variables, assuming equal weights for both variables. The joint distance is then calculated through summation for clustering. KSPXY follows the same variables and process as SPXY, but replaces the Euclidean distance formula with a kernel-based distance for calculation. Unlike the KSPXY method, the SPXYE algorithm continues to employ the traditional Euclidean distance formula, which considers both the independent and dependent variables. Moreover, it incorporates an error vector E as an additional input to enhance the partitioning process. The specific process in the initial stage is the same as SPXY, after regression based on the error vector to calculate the Euclidean distance and summed with the results of the distance value in the initial stage to obtain a new distance value as the basis for clustering again.

2.4 Modelling Methodology and Model Evaluation

The partial least squares regression is a linear method commonly used for hyperspectral attribute estimation in fields such as soil or agriculture, where the number of independent variables significantly exceeds the number of data points. It effectively addresses the issue of strong multicollinearity between independent parameters during the estimation process, as well as the dimensionality problem in multi-parameter calibration. Its core is to identify a new set of orthogonal variables that maximize the explanation of the correlation between independent and dependent variables, and to construct a regression model to predict the dependent variable. Random forest is a nonlinear ensemble learning method used for regression analysis of hyperspectral data. Random sampling with replacement is performed on the original dataset to create multiple training subsets. Each subset is used to train an independent decision tree, where the nodes are split by randomly selecting a subset of features and choosing the best features for splitting. Each decision tree is trained independently to maximize the fit to its respective subset. The final predicted values are obtained by averaging the results from all the trees.

For the evaluation metrics of the dataset, in addition to the maximum, minimum, average, and standard deviation, the Coefficient of Variation (CV) is used to supplement the evaluation of sample partitioning. The coefficient of variation is a statistical metric that measures the relative dispersion of a dataset. Unlike the standard deviation, the coefficient of variation is a dimensionless proportional value.

The Pearson Correlation Coefficient (PCC) was used to measure the correlation between the bands and salt concentration before and after the spectral enhancement process. The PCC not only indicates whether the two variables are correlated, but also quantifies the strength and direction of the correlation. The result, expressed as r, ranges from -1 to 1. A positive r value indicates a positive correlation, while a negative r value indicates a negative correlation. The absolute value of r below 0.4 indicates a weak correlation, 0.4 to 0.6 indicates a moderate correlation, and 0.6 to 1 indicates a strong correlation between the band and salt concentration index.

The accuracy of the hyperspectral prediction model was assessed using the coefficient of determination (R^2) and the root mean square error (RMSE). R^2 indicates the confidence level of the model. When R^2 is less than 0.5, it means that the model lacks predictive ability. When it is between 0.5 and 0.7, it means the model shows preliminary predictive ability. When it is greater than 0.7, it means the model indicates good predictive ability(Sawut et al., 2018) . The RMSE represents the model accuracy. A smaller value represents the higher accuracy of the model prediction.

3. Results and Discussion

3.1 Statistical Analysis of Salt Content in the Training Set

The reflectance of the simulated mural is determined by the components of sodium sulfate, Dunhuang soil, loess, clarified board soil, Dunhuang sand, and hemp rope. Fig. 3 shows the original and enhanced spectra. The reflectance curves of the simulated mural with different salt contents are parallel, showing similar morphology and trends, but there are significant differences in the reflectance values overall. The reflectance curves of the original salt-containing mural data show distinct characteristics in each wavelength band. In the visible wavelength range, the reflectance increases rapidly with the wavelength, levelling off after 800 nm. After entering the near-infrared range of 1400–2500 nm, the reflectance shows strong fluctuations and an overall decreasing trend. Asymmetric absorption valleys appeared at 1410 nm and 1940 nm, with the valley at 1940 nm being more pronounced. Its width and depth were evident, while the peak and valley positions showed no significant shift.

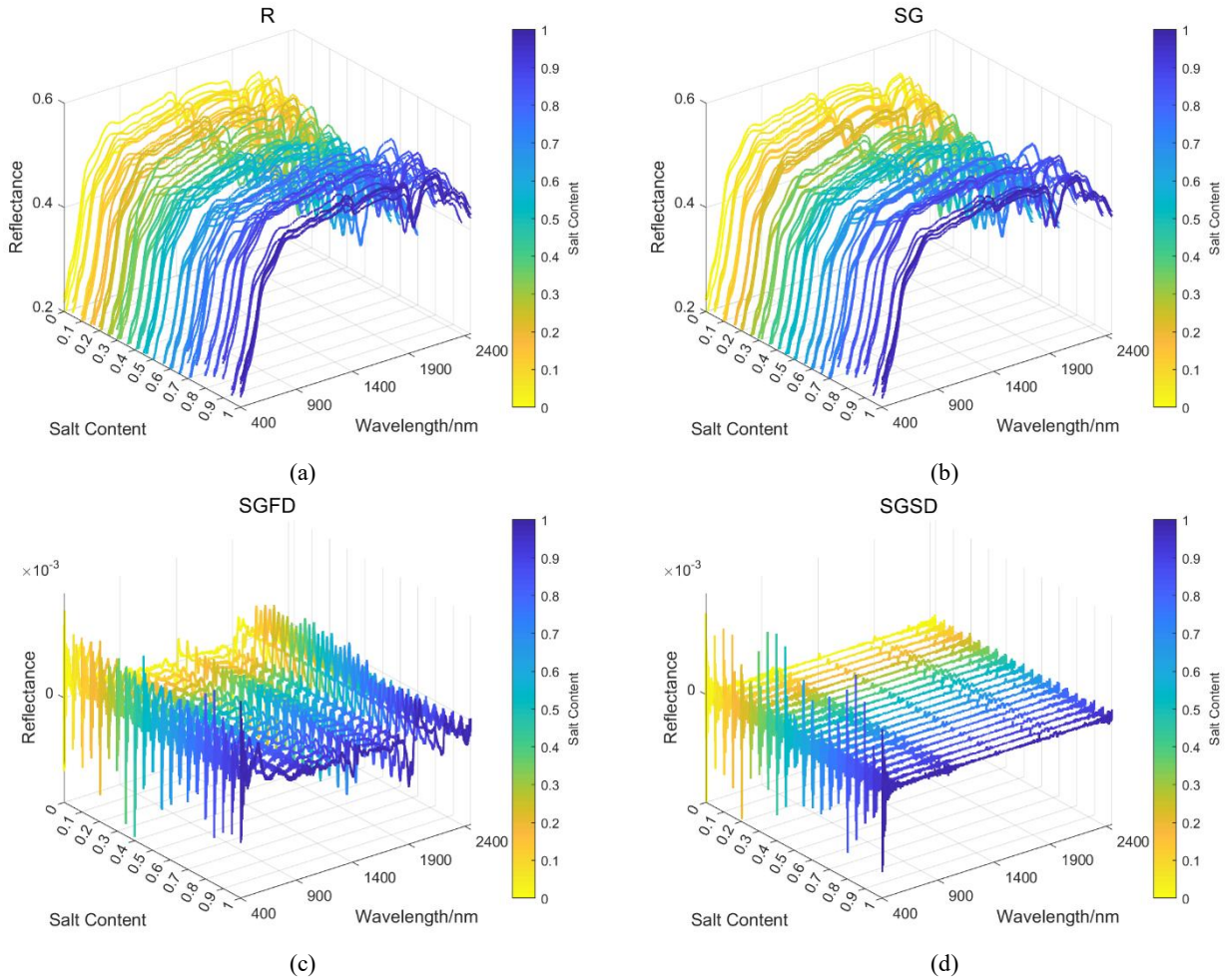

Figure 3. Spectral Curves of the Original and Enhanced Spectra(a) Spectral Curve of the Original Data; (b) Spectral Curve of the SG Data; (c) Spectral Curve of the SG-FD Data; (d) Spectral Curve of the SG-SD Data.

Sample partitioning methods should effectively capture the characteristics of small datasets to rationally divide the training and calibration sets. In this study, the range (maximum-minimum), mean, standard deviation, and coefficient of variation were used to describe the characteristics of the simulated mural salt concentration dataset. In Table 1, the statistics of the full set without spectral enhancement and the training set with different sample partitioning methods are presented. All sample partitioning is made into training and calibration sets at a 7:3 ratio, with the RS method generating 1000 sets through random sampling. However, only the dataset corresponding to the optimal inversion results is selected for analysis in the table. In the comparison analysis, the range data (Range) is consistent across all methods. The values closest to the original dataset in the remaining metrics are underlined. The results show that, in terms of coefficient of variation and mean, SPXY, KSPXY, and SPXYE perform better. The RS method results are more different from the original dataset and the KS methods are generally closer to the original dataset. On the standard deviation metric, the methods were ranked as KSPXY < SPXYE < KS = RS < SPXY. In contrast, on the coefficient of variation metric, the ranking was SPXY < SPXYE < KS < KSPXY < RS.

Taken together, the five sample partitioning methods perform reasonably well on the raw data, with SPXYE showing an overall advantage. SPXY and KSPXY also demonstrate potential worth noting. The ultimate goal of the study is to select the data inversion model that provides high accuracy. Therefore, both a linear model (PLSR) and a nonlinear model (RF) should be used to perform salt concentration inversion on the five training sets. The results should then be further analysed by calculating the accuracy of the calibration set.

Methods	Average	Standard Deviation	CV	Range
Total	0.437	0.295	0.676	1
RS	0.392	0.292	0.745	1
KS	0.443	0.292	0.659	1
SPXY	0.423	0.290	<u>0.687</u>	1
KSPXY	0.449	<u>0.296</u>	0.658	1
SPXYE	<u>0.442</u>	0.293	0.664	1

Table 1. Training set design results for salt concentration of simulated murals

3.2 Inversion Results for the Original Dataset

The inversion results of the original dataset using the PLSR and RF models are presented in Tables 2 and 3, respectively. Based on the calibration set performance, the RF model generally demonstrates superior predictive accuracy compared to the PLSR model. The results under the RF model, as shown in Table 2, indicate that the R_C^2 values are all above 0.926, and $RMSE_C$ is below 0.084. However, when evaluating the calibration set performance, it is observed that the coefficients of determination for the remaining methods are all below 0.376 except for the R-KSPXY-RS model, whose R_V^2 exceeds 0.6. This suggests poor

predictive performance and indicates potential overfitting. It is possibly caused by the use of full-band data containing excessive noise and indistinct features. Spectral feature enhancement may help mitigate this overfitting issue.

Techniques	RF			
	$RMSE_C$	R_C^2	$RMSE_V$	R_V^2
RS	0.084	0.926	0.129	0.706
KS	0.079	0.927	0.246	0.341
SPXY	0.073	0.936	0.241	0.376
KSPXY	0.073	0.934	0.197	0.615
SPXYE	0.072	0.939	0.271	0.186

Table 2. Inversion results of raw spectral data in RF model

As shown in Table 3, under the PLSR model, the RS-PLSR model yields the best performance. Most calibration set models exhibit more stable performance except for the KS method whose R_C^2 is 0.518(below the 0.6 threshold), indicating a slightly weaker fitting effect. Similarly, the KS model also shows the highest $RMSE_C$ value at 0.20, further confirming its relatively poor fitting. In the calibration set, the R_V^2 for SPXY is the smallest at 0.825, still greater than 0.6, suggesting that the PLSR model has some generalization. The maximum and minimum $RMSE_V$ of 0.124 and 0.082 were observed in the KS and RS methods, respectively. KSPXY-PLSR yields the best results, except for the RS method.

Techniques	PLSR			
	$RMSE_C$	R_C^2	$RMSE_V$	R_V^2
RS	0.156	0.681	0.082	0.896
KS	0.201	0.518	0.124	0.832
SPXY	0.126	0.797	0.123	0.825
KSPXY	0.167	0.658	0.095	0.878
SPXYE	0.140	0.758	0.111	0.855

Table 3. Inversion results of raw spectral data in PLSR model

Therefore, using different sample partitioning methods for the same full dataset yields slightly different results, even when the inversion model is the same. Some sample partitioning methods can better enhance model stability, suggesting that the choice of sample partitioning method does impact the modelling prediction results.

3.3 Enhanced Dataset Modelling and Predictive Validation Results Analysis

The correlation coefficient r between salt content and reflectance was calculated for each wavelength using PCC for the four datasets (R, SG, SG-FD, SG-SD). A two-tailed significance test was then performed to determine whether the correlation coefficient was statistically significant. A heat map was generated based on the results (Fig. 3), with the horizontal coordinates representing the corresponding wavebands and enhancement methods. The colours on the left indicate the correlation strength. Redder colours represent stronger positive correlations, while bluer colours indicate stronger negative correlations. From the figure, it can be seen that the raw reflectance dataset is predominantly yellow, indicating a weak positive correlation with salt content overall. The calculated results show that the highest correlation coefficient occurs at 425 nm, with a value of 0.244, all of which are below 0.4. The correlation of the SG dataset slightly improves, with the highest correlation coefficient of 0.253 occurring at 499 nm, still indicating a weak correlation. From the figure, it can be seen that after first-order differentiation and second-order differentiation, red and blue colours appear in different bands, significantly enhancing both positive and negative correlations. The strongest negative correlation in the first-order differentiation occurs at 2260 nm, with a correlation coefficient of -0.812, while the strongest positive correlation occurs at 1040 nm, with a coefficient of 0.653. In the second-order differentiation, the strongest positive correlation is at 2273 nm, with a correlation coefficient of 0.761, and the strongest negative correlation is at 1857 nm, with a coefficient of -0.731.

Overall, the correlations of the results for the enhancement methods are ranked as SG-FD > SG-SD > SG > R. The pre-processing significantly enhances the correlations, with more bands showing positive and negative polar distributions. The spectral differentiation method was effective in identifying key bands highly correlated with salt content, and the effects of different pre-treatment methods on correlation enhancement varied significantly.

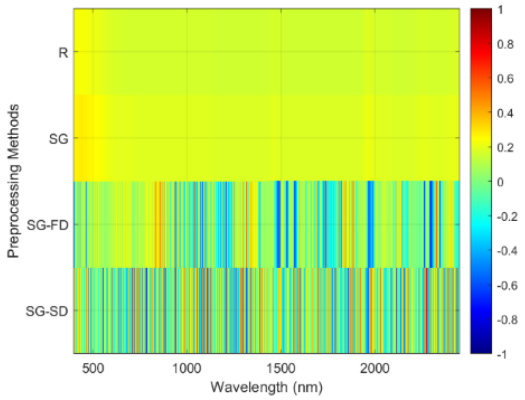

Figure 3. Correlations between salt content and spectral reflectance data

Combined with a comprehensive analysis of the regression models, Tables 4 and 5 present the statistical results of the PLSR and RF prediction models built under the five data partitioning methods for the spectrally enhanced dataset. Under the PLSR model for the differentiated dataset, the minimum training set coefficient of determination is 0.675, which occurs in KS-SG-FD. The results range from 0.5 to 0.7, indicating a certain level of predictive ability. The maximum training set root mean square error (0.158) was achieved with the same method. The PLSR model fit was stable, demonstrating good model correction capability. The minimum R_V^2 (0.726) appeared in the KS-SG-SD dataset. The $RMSE_V$ ranged from 0.049 to 0.133, indicating strong overall generalization ability and model stability. The R_V^2 and $RMSE_V$ of the SG-FD and the SG-SD under the five sample partitioning methods are better than the statistical results of R. Therefore, it can be concluded that the PLSR model built using the spectrally enhanced dataset performs well in terms of accuracy and model stability. The model ranking based on $RMSE_V$ is: RS<KSPXY<SPXY<SPXYE<KS.

The results under PLSR show that RS-SG-SD has the highest evaluation, with the coefficient of determination values for the calibration set and training set at a high level (0.985 and 0.939). This indicates that random sampling combined with the SG-SD method has good potential for fitting and generalization of the data. The SG-FD method of the KSPXY technique shows outstanding performance with the highest R_V^2 (0.954), indicating that this method has significant advantages in data selection and preprocessing. SPXY-SG-SD shows more balanced results for

both the training and calibration sets. From the statistical results, the KS method has the worst model calibration results, the second-worst test results, and the poorest generalization ability.

Methods		$RMSE_C$	R_C^2	$RMSE_V$	R_V^2
RS	-	0.156	0.681	0.082	0.896
	SG	0.113	0.852	0.065	0.934
	SG-FD	0.095	0.904	0.049	0.945
	SG-SD	0.038	0.985	0.058	0.939
KS	-	0.201	0.518	0.124	0.832
	SG	0.124	0.789	0.171	0.726
	SG-FD	0.158	0.675	0.114	0.876
	SG-SD	0.137	0.765	0.117	0.852
SPXY	-	0.126	0.797	0.123	0.825
	SG	0.114	0.826	0.133	0.814
	SG-FD	0.091	0.891	0.077	0.930
	SG-SD	0.086	0.908	0.077	0.923
KSPXY	-	0.167	0.658	0.095	0.878
	SG	0.099	0.861	0.102	0.895
	SG-FD	0.082	0.912	0.068	0.954
	SG-SD	0.095	0.882	0.075	0.936
SPXYE	-	0.140	0.758	0.111	0.855
	SG	0.098	0.873	0.121	0.838
	SG-FD	0.115	0.819	0.101	0.906
	SG-SD	0.092	0.882	0.103	0.897

Table 4. Statistics of accuracy parameters of PLSR model

The R_C^2 values of the RF models built for the differentiated dataset are higher than 0.96, while the R_V^2 are all above 0.735. The overfitting issue has been improved, and the coefficients are also higher than the maximum R_V^2 of the original dataset (0.706). This indicates that improving the data relevance contributes to the stability of the RF models. Meanwhile, the model's $RMSE_V$ has decreased, demonstrating that spectral differentiation has an excellent effect under the RF model. Based on $RMSE_V$, the ranking is as follows: RS<KS<KSPXY<SPXY<SPXYE.

Taken together, the RS-SG-FD model has the best prediction ability, with an $RMSE_V$ of 0.081 and an R_V^2 of 0.921. The KS method improves its performance for the post-differential calibration set, while the KSPXY results are more stable when dealing with differentiated data, closely matching the performance of the RS optimal results. The SPXYE method shows balanced performance among all the partitioning methods, demonstrating better applicability. Overall, RS still yields the best results, followed by KS, while KSPXY exhibits high stability in partitioning complex data. For cases that require further optimization of data distribution, the KSPXY combined with the SG-FD method achieves the best results, and this combination is recommended for data preprocessing.

Methods		$RMSE_C$	R_C^2	$RMSE_V$	R_V^2
RS	-	0.084	0.926	0.129	0.706
	SG	0.107	0.875	0.165	0.615
	SG-FD	0.053	0.968	0.081	0.921
	SG-SD	0.055	0.967	0.136	0.735
KS	-	0.079	0.927	0.246	0.341
	SG	0.096	0.873	0.253	0.404
	SG-FD	0.054	0.963	0.099	0.907
	SG-SD	0.054	0.963	0.132	0.813
SPXY	-	0.073	0.936	0.241	0.376
	SG	0.090	0.894	0.281	0.267
	SG-FD	0.052	0.966	0.125	0.847
	SG-SD	0.058	0.961	0.127	0.821
KSPXY	-	0.073	0.934	0.197	0.615
	SG	0.091	0.900	0.247	0.327
	SG-FD	0.054	0.965	0.114	0.861
	SG-SD	0.057	0.960	0.115	0.858
SPXYE	-	0.072	0.939	0.271	0.186
	SG	0.084	0.910	0.303	0.114
	SG-FD	0.046	0.973	0.164	0.758
	SG-SD	0.054	0.961	0.159	0.775

Table 5 Statistics of accuracy parameters of RF model

3.4 Analysis of RS Results

Although the RS method demonstrated the best results under both the PLSR and RF models, which only indicates the potential of the RS method to achieve optimal results, it does not prove that RS is the best method for mural salinity inversion. Therefore, 1,000 sets of data were collected under the RS method, as well as 1,000 sets for each of the enhanced treatment datasets, with results evaluated in both the PLSR and RF models. These results include the mean (Average, Ave), median (Median, Med), minimum (Min), maximum (Max), and overall Confidence Interval (CI), Confidence Interval Upper bound r (CIU) and Confidence Interval Lower bound (CIL). The results are displayed in Table 6.

			Ave	Med	Min	Max	CIL	CIU
PLSR	R_C^2	R	0.778	0.782	0.511	0.914	0.775	0.782
		SG	0.856	0.866	0.603	0.953	0.853	0.859
		SG-FD	0.936	0.928	0.770	0.999	0.933	0.938
		SG-SD	0.918	0.915	0.745	0.997	0.915	0.920
	R_V^2	R	0.746	0.755	0.333	0.904	0.741	0.751
		SG	0.830	0.837	0.490	0.938	0.827	0.833
		SG-FD	0.891	0.896	0.694	0.961	0.889	0.893
		SG-SD	0.859	0.864	0.571	0.955	0.856	0.862
	$RMSE_C$	R	0.133	0.132	0.080	0.201	0.132	0.134
		SG	0.107	0.105	0.060	0.180	0.106	0.108
		SG-FD	0.067	0.077	0.010	0.138	0.065	0.069
		SG-SD	0.079	0.083	0.015	0.150	0.078	0.081
	$RMSE_V$	R	0.136	0.135	0.082	0.198	0.135	0.137
		SG	0.110	0.109	0.065	0.160	0.109	0.111
		SG-FD	0.089	0.087	0.049	0.138	0.088	0.090
		SG-SD	0.101	0.101	0.058	0.152	0.100	0.102

RF	R_C^2	R	0.928	0.928	0.901	0.957	0.928	0.929
		SG	0.893	0.893	0.860	0.920	0.892	0.893
		SG-FD	0.971	0.971	0.957	0.985	0.971	0.971
		SG-SD	0.966	0.966	0.953	0.979	0.966	0.966
	R_V^2	R	0.482	0.507	0.001	0.794	0.472	0.493
		SG	0.251	0.254	0.000	0.638	0.243	0.260
		SG-FD	0.783	0.797	0.237	0.921	0.778	0.788
		SG-SD	0.742	0.752	0.419	0.896	0.738	0.747
	$RMSE_C$	R	0.078	0.079	0.058	0.089	0.078	0.079
		SG	0.096	0.096	0.081	0.107	0.096	0.096
		SG-FD	0.050	0.050	0.037	0.057	0.049	0.050
		SG-SD	0.054	0.054	0.042	0.061	0.054	0.054
	$RMSE_V$	R	0.207	0.205	0.129	0.329	0.204	0.209
		SG	0.249	0.248	0.165	0.348	0.247	0.251
		SG-FD	0.133	0.130	0.081	0.229	0.131	0.134
		SG-SD	0.146	0.144	0.082	0.214	0.144	0.147

Table 6. Statistics of accuracy parameters of PLSR and RF model in RS

In the PLSR model, the performance difference of the RS methods can be observed based on the mean, median, and confidence interval results. Overall, the models under various data processing methods show some accuracy and stability. From the results of the training set, the values of the mean and median are very close to each other, indicating that the distribution of the model's prediction results is more consistent, and the range of the confidence intervals is smaller (e.g., 0.775 to 0.782 for the unprocessed R data), which suggests that the model has good stability. Under different treatments, the mean value gradually increases from 0.778 to a maximum of 0.936, demonstrating the effect of spectral enhancement on model performance and the good suitability of the PLSR model. In the calibration set results, the decrease in the mean value is limited, and the ranges of the minimum and maximum values are relatively narrow, indicating that the model has good generalization ability. In addition, the narrow confidence intervals further indicate that the model has a certain degree of reliability in processing RS subset data. However, when comparing the same spectral enhancement methods, the RS sample partitioning method shows a R_V^2 higher than 0.6 and a smaller $RMSE_V$ than the other four methods, with improvements of 0.15%, 29.4%, 4.8%, and 4.9%, respectively. This suggests that the sample partitioning results of RS have some stability under the PLSR model, though most of the inversion results are not as good as those of the other four methods.

Similarly, in the RF model, the minimum values of the training set are generally high, with all values above 0.859. This indicates that the model has a strong ability to fit different data processing methods on the training set. However, in the calibration set, the minimum value drops significantly. For example, the minimum value of the unprocessed R data is only 0.001, demonstrating the instability of the model in practical applications. This significant difference in performance between the training and calibration sets suggests that the RF model is overfitting on the RS subset data. The combination of the mean, minimum, and confidence interval metrics shows that the RF method has insufficient generalization ability. Statistics for the four calibration sets with a coefficient of determination higher than 0.6 and root-mean-square error smaller than the other four methods are 25.9%, 0.03%, 7%, and 5.5%, respectively. However, most of the sample results are still inferior to those of the other four methods.

In summary, the RS method demonstrated high fitting accuracy and stability in the PLSR model, with the mean, median, and confidence intervals reflecting the robustness and usability of the model in the inversion task. However, in the RF model, overfitting was observed with the unenhanced data, and the model lacked sufficient generalization, leading to poor results. When comparing the other methods under the same conditions, fewer better results were achieved, especially for the enhanced data, where the number of successful inversions was less than 7%.

4. Conclusion

This paper explores the impact of sample partitioning and spectral enhancement methods on the accuracy of reflectance data in predicting the salt content of mural surfaces. The dataset is somewhat discrete and difficult to predict. A comparative study was conducted using five sample partitioning methods (RS, KS, SPXY, KSPXY, and SPXYE) along with spectral differentiation, and the following findings were made:

(1) In order to obtain a representative data subset, the KSPXY method offers some advantages, while the SPXYE algorithm is more stable in its performance. RS, on the other hand, is the least stable one.

(2) The sample partitioning method has a significant impact on model prediction. When combined with different prediction algorithms, the KSPXY prediction model shows higher stability and accuracy than the other three methods, except for the RS method. While the RS method, when combined with two prediction algorithms, achieves the best inversion results and generalization ability in 1000 experiments. It is not more than 7% better than the other four methods. This indicates that the RS method has the potential for optimal results, but it requires significant computational power and extensive data screening.

(3) The differentiation significantly improves the correlation between spectra and salt content, while highlighting data features and enhancing the inversion of full-band or multi-source data. Among all methods, SG-FD and SG-SD further improve fitting accuracy and generalization ability. Particularly in the RF model, although this method has some advantages with small samples, it is generally effective in noise reduction, and the overfitting problem with the enhanced dataset is mitigated.

(4) Different modelling methods exhibit some variation in their ability to handle the simulated mural data, particularly in the inversion of full-band data. Comparing the PLSR and RF model results, the models show some differences, especially with overfitting in the RF model. The least squares PLSR, however,

provides better results due to its integration with principal component analysis, which helps remove redundant noise and select key features more effectively.

The KS method is ineffective for discrete data. The SPXY and SPXYE methods offer slight advantages in dataset partitioning, but their predictive results are average. The optimal method in this experiment is KSPXY combined with SG-FD data, which performs best under the PLSR model, with an $RMSE_C$ of 0.082 and an R_C^2 of 0.912. The $RMSE_V$ is 0.068, with an R_V^2 of 0.954, demonstrating effective data selection and preprocessing. This experiment primarily investigates the impact of sample partitioning methods on salt content prediction for temple murals, with less focus on the regression modelling method, which requires further research.

Acknowledgements

This research was supported by the National Natural Science Foundation of China (No.42171356， No.42171444).

References

Guo, Z.Q., Lyu, S.Q., Hou, M.L., 2023: Estimation of the soluble salt concentration in murals based on spectral transformation and feature extraction modelling. *J. Appl. Spectrosc.*. 90, 1123–1132.

Gil, M., Martins, M.R., Carvalho, M.L., Souto, C., Longelin, S., Cardoso, A., Mirão, J., Candeias, A.E., 2015: Microscopy and Microanalysis of an Extreme Case of Salt and Biodegradation in 17th Century Wall Paintings. *Microsc. Microanal.* 21, 606–616.

Madariaga, J.M., Maguregui, M., De Vallejuelo, S.F., Knuutinen, U., Castro, K., Martinez-Arkarazo, I., Giakoumaki, A., Pitarch, A., 2014: In situ analysis with portable Raman and ED-XRF spectrometers for the diagnosis of the formation of efflorescence on walls and wall paintings of the Insula IX 3 (Pompeii, Italy). *J. Raman Spectrosc.* 45, 1059–1067.

Ren, Y., Liu, F., 2024: The spectral inversion model for electrical conductivity in mural plaster following phosphate erosion based on fractional order differentiation and novel spectral indices. *Herit. Sci.* 12, 1–27.

Sawut, R., Kasim, N., Abliz, A., Hu, L., Yalkun, A., Maihemuti, B., Qingdong, S., 2018: Possibility of optimized indices for the assessment of heavy metal contents in soil around an open pit coal mine area. *Int. J. Appl. Earth Obs. Geoinf.* 73, 14–25.

Sawdy, A., Price, C., 2005: Salt damage at Cleeve Abbey, England. *J. Cult. Herit.* 6, 125–135.

Yu Z.R, Wang Y.W, Wang X.W., Zhao L.Y., Guo Q.L., Wang X.D. 2017: Research on the Water Vapor Source Induced Diseases of Wall Paintings in Longxing Temple. *Advances in Earth Sci.*, 32(6): 668-676.

Yao Y.X., Huang Y.Z., Ma Y., Qi Y.M., Wei S.Y. 2023: The Identification and Analysis of the Materials and Workmanship for the Water-and-Land-Murals of Daxiong Dian (Hall) of Princess Temple， Fanshi. *Spectrosc. Spectr. Anal.* 43(04): 1155-1161.