BILINGUAL PUBLISHING GROUP

Journal of Environmental & Earth Sciences
https://journals.bilpubgroup.com/index.php/jees

ARTICLE

Hyperspectral Inversion and Analysis of Zinc Concentration in Urban Soil in the Urumqi City of China

***Qing Zhong*[1], *Mamattursun Eziz*[1,2*], *Mireguli Ainiwaer*[1,2], *Rukeya Sawut*[1,2]**

[1] *College of Geographical Science and Tourism, Xinjiang Normal University, Urumqi, Xinjiang, 830054, China*
[2] *Laboratory of Arid Zone Lake Environment and Resources, Xinjiang Normal University, Urumqi, Xinjiang, 830054, China*

ABSTRACT

Excessive accumulation of zinc (Zn) in urban soil can lead to environmental pollution and pose a potential threat to human health and the ecosystem. How to quickly and accurately monitor the urban soil zinc content on a large scale in real time and dynamically is crucial. Hyperspectral remote sensing technology provides a new method for rapid and nondestructive soil property detection. The main goal of this study is to find an optimal combination of spectral transformation and a hyperspectral estimation model to predict the Zn content in urban soil. A total of 88 soil samples were collected to obtain the Zn contents and related hyperspectral data, and perform 18 transformations on the original spectral data. Then, select important wavelengths by Pearson's correlation coefficient analysis (PCC) and CARS. Finally, establish a partial least squares regression model (PLSR) and random forest regression model (RFR) with soil Zn content and important wavelengths. The results indicated that the average Zn content of the collected soil samples is 60.88 mg/kg. Pearson's correlation coefficient analysis (PCC) and CARS for the original and transformed wavelengths can effectively improve the correlations between the spectral data and soil Zn content. The number of important wavelengths selected by CARS is less than the important wavelengths selected by PCC. Partial least squares regression model based on first-order differentiation of the reciprocal by CARS (CARS-RTFD-PLSR) is more stable and has the highest prediction ability (R^2 = 0.937, RMSE = 8.914, MAE = 2.735, RPD = 3.985). The CARS-RTFD-PLSR method can be used as a means of prediction of Zn content in soil in oasis cities. The results of the study can provide technical support for the hyperspectral estimation of the soil Zn content.

Keywords: Urban soil; Zinc; Hyperspectral remote sensing; Prediction; PLSR; RFR

*CORRESPONDING AUTHOR:
Mamattursun Eziz, College of Geographical Science and Tourism, Xinjiang Normal University, Urumqi, Xinjiang, 830054, China; Laboratory of Arid Zone Lake Environment and Resources, Xinjiang Normal University, Urumqi, Xinjiang, 830054, China; Email: oasiseco@126.com

ARTICLE INFO
Received: 5 September 2023 | Revised: 28 September 2023 | Accepted: 8 October 2023 | Published Online: 17 October 2023
DOI: https://doi.org/10.30564/jees.v5i2.5947

CITATION
Qing, Zh., Eziz, M., Ainiwaer, M., et al., 2023. Hyperspectral Inversion and Analysis of Zinc Concentration in Urban Soil in the Urumqi City of China. Journal of Environmental & Earth Sciences. 5(2): 76-87. DOI: https://doi.org/10.30564/jees.v5i2.5947

1. Introduction

Zinc (Zn) and its compounds are often enriched in soil by substitution reactions and adsorption and immobilization, resulting in environmental pollution [1]. Through accumulation, migration and transport in the food chain, Zn in soil eventually poses a serious threat to human health [2]. Therefore, it is vital to protect the safety of the urban soil environment by rapid and accurate monitoring of the Zn content. The traditional methods for determining heavy metal content in soil require field sampling followed by laboratory experimentation, but it is time-consuming, costly and inefficient [3,4]. Hyperspectral remote sensing technology has been applied to the prediction of heavy metal contents in soil due to the advantages of rapid, accurate, non-destructive, lower cost, and dynamic monitoring over a large area [5-7].

In recent years, hyperspectral remote sensing technology has shown good results in the prediction of soil Zn content. For example, the BPNN model has good generalization ability (R^2 = 0.74, RPIQ = 1.44) to predict the soil Zn content for Dehong Prefecture, southwest Yunnan Province, China [8], CWT-RF model has a good prediction accuracy (R^2 = 0.77, RMSE = 9.54, MAE = 7.39) to estimate the Zn content for Ordos City, Inner Mongolia Autonomous Region, China [9] and PLSR model can effectively achieve quantitative inversion (R^2 = 0.796, RMSE = 2.574) of soil Zn content in mining areas of the city of Zoucheng, Shandong Province, China [10]. El-Sayed E [11] found that the PLSR model had the optimal prediction for Bahr El-Baqar region with R^2 of 0.66, RMSE of 20.42, and RPD of 2.05. Yang et al. [12] pointed out that the PLSR model had the highest stability and accuracy (R^2 = 0.95, RMSE = 33.65) in predicting the Zn content in mining areas of the city of Tongling, Anhui Province, China.

With the acceleration of urbanization and the influence of the "Silk Road Economic Belt", eco-environmental problems in oases in the northwestern arid zones garnered more attention [13,14]. In addition, due to the impact of strong human activities, factories and high traffic volumes, the level of Zn in urban soil is higher than that in farmland and natural soil [15]. Relevant studies also reported that there is serious trace element contamination exists in soil and surface dust in Urumqi [16,17]. Therefore, it is very important to analyze the possibility of the hyperspectral inversion of Zn contents in urban soils. The main objective of this study was to find an optimal model to predict the Zn content in soil. Thus, the work of this study was to identify the important wavelengths of Zn in urban soil and evaluate the efficiency of different spectral transformations and soil Zn contents. Then, select an optimum hyperspectral prediction model for Zn content in urban soil based on the partial least squares regression (PLSR) and random forests regression (RFR). The results will solve the existing problems in the current hyperspectral inversion of Zn content in urban soil.

2. Description of the study area

The experimental field (87°28'-87°37' E and 43°48'-44°04' N) is selected in the central parts of the Urumqi, which is situated in the southern edge of the Junggar Basin, the northwest arid regions of China, and is one of the important metropolitan cities in NW China (**Figure 1**). The main soil type of this region is mainly grey desert soil [15]. The climate is regionally marked by a continental arid climate with an annual average temperature, precipitation, and evaporation of about 6.7 °C, 280 mm, and 2730 mm, respectively. Urumqi has become the capital of Xinjiang and the second-largest city in northwestern China due to its rapid economic development and expanding industrial scale. Toxic elements in the soil are accumulating and are prone to soil pollution.

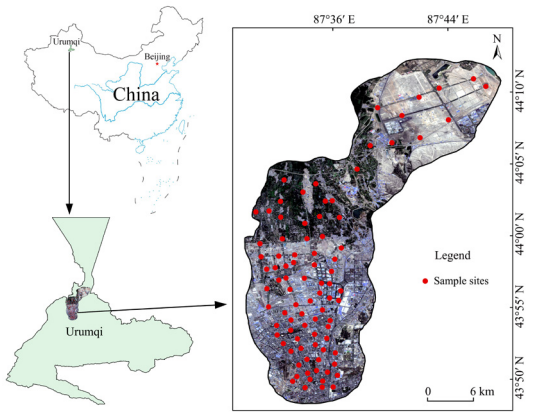

Figure 1. Location of experimental field and sample sites.

3. Materials and methods

3.1 Sample collection and analysis

A total of 88 topsoil samples (0-20 cm) were collected within the study area (**Figure 1**) in April 2021. At each sample site, five sub-samples from the topsoil (0-20 cm) layer were taken within 100 m × 100 m areas and then mixed together to form one composite soil sample, weighing more than 500 g. All the samples were returned to the laboratory and sieved for 20 meshes after naturally air dried. Each sample was divided into two groups, one for the determination of Zn content and another one for the hyperspectral measurement. The Zn content of soil samples was determined as described in "HJ 803-2016" [18], using an Inductively Coupled Plasma Mass Spectrometer (ICP-MS 7800). The analytical data quality was analyzed by the laboratory quality control methods, including the use of reagent blanks, duplicates and standard reference materials for each batch of soil samples. For the precision of the analytical procedures, a standard solution was used to compare samples to national standards (Chinese national standards samples, GSS-12). All of the soil samples were tested repeatedly, and the determined consistency of the Zn measurements was 96.5%.

3.2 Spectrometric determination and pre-processing

The spectral determination of collected soil samples was measured using a FieldSpec®3 portable object spectrometer manufactured by Analytical Spectral Devices (ASD), USA. The interval of data acquisition was 1 nm with a spectral measurement range from 350 to 2500 nm. Firstly, the instrument was preheated, and secondly, a 40 cm × 40 cm white board was placed on a 2 m × 2 m black cardboard for calibration to obtain the absolute reflectance before determining the spectral data. Finally, the soil samples were kept in a natural state on the black cardboard with the sensor probe perpendicular to 15 cm above the soil surface, and the sensor probe was optimized with a white board every 5 minutes. A total of 15 replicate measurements were taken on the same soil sample, and 15 spectral curves were collected.

The 15 spectral curves were averaged using ViewSpecPro software, and the arithmetic mean was taken as the original reflectance spectral value of the soil sample. Due to the influence of the surrounding environment and the spectral instrument itself, the spectral bands within 350-399 nm, 1350-1430 nm, 1781-1970 nm and 2401-2500 nm were excluded before constructing the hyperspectral models, which were outputted in a total of 1730 bands. The Savitzky-Golay (S-G) filter algorithm is applied for smoothing and removing noise from spectral curves. **Figure 2** illustrates the spectral reflectance curves of the original spectra and spectra processed by the S-G smoothing.

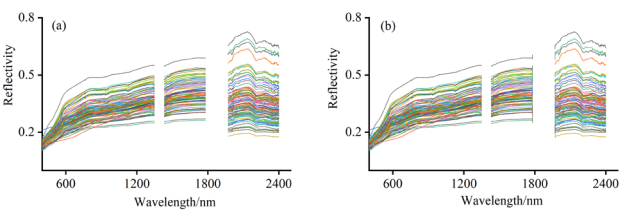

Figure2. Original (a) and Savitzky-Golay smoothing (b) spectral reflectance curve of soil.

3.3 Spectral transformation and important wavelength selection

In order to enhance the spectral information related to Zn in soil samples, the original spectral reflectance data (R) are subjected to logarithm of the

reciprocal (AT), root mean square (RMS), logarithm (LT), reciprocal of the logarithm (RL), reciprocal (RT), first-order differentiation (FD), second-order differentiation (SD), first-order differentiation of the reciprocal (RTFD), second-order differentiation of the reciprocal (RTSD), first-order differentiation of the logarithm (LTFD), second-order differentiation of the logarithm (LTSD), root mean square first-order differentiation (RMSFD), root mean square second-order differentiation (RMSSD), logarithmic first order differentiation of the reciprocal (ATFD), logarithmic second order differentiation of the reciprocal (ATSD), logarithmic first order differentiation of the reciprocal (RLFD), and logarithmic second order differentiation of the reciprocal (RLSD).

Firstly, Pearson's correlation coefficient analysis (PCC) was performed between soil Zn content and 18 forms of soil spectral data, and the bands with larger correlation coefficients were screened out as important wavelengths for hyperspectral prediction modeling. Secondly, all the original and 17 types of transformed spectral data were intelligently extracted for the important wavelengths by using Competitive Adaptive Re-weighted Sampling (CARS) to exclude further removed wavelengths with low correlation [19,20]. The CARS method is constructed in Python.

3.4 Modelling of hyperspectral inversion

Soil samples were randomly divided into a modeling data set (70 samples) and a validation data set (18 samples) in order to ensure the rationality of hyperspectral modeling. The modeling data set was used to build hyperspectral prediction models, while the validation data set was used to test the accuracy of prediction models. The partial least squares regression (PLSR) and random forests regression (RFR) were used to select the optimum hyperspectral prediction model. The PLSR algorithm considers both spectral information (x) and the corresponding reference values (y) of samples during modeling and transforms the original spectral data into mutually orthogonal and unrelated new variables via linear transformation, thereby eliminating multicollinearity between datasets [21].

The RFR is a relatively new data mining technique that is designed to produce accurate predictions that do not overfit the data. RFR is easy to use as it requires only three input parameters: the number of two 'random_state' and 'n_estimators'. The three input parameters are used to partition the modeling set and validation set, and determine the optimal partitioning of each tree node [22], respectively. The individual trees in the RFR ensemble are built on a bootstrapped training sample, and only a small group of predictor variables are considered at each split; this ensures that trees are de-correlated with each other. Additionally, studies have shown that the three input parameters provide accurate results [23].

3.5 Model validation

A robust model has high R^2 and RPD and low RMSE and MAE [24]. Thus, the determination coefficient (R^2), root mean square error (RMSE), mean absolute error (MAE) and residual prediction deviation (RPD) were chosen to evaluate the prediction accuracy of the hyperspectral prediction models. When $R^2 < 0.5$, the prediction model does not have prediction ability, when $0.5 \leq R^2 < 0.7$, the model has preliminary prediction ability, and when $R^2 \geq 0.7$, the model has good prediction ability [25]. When RPD $\geq$ 2.0, the prediction model has a good prediction ability, when $1.4 \leq$ RPD < 2.0, the model has the initial predictive capability, and when RPD < 1.4, the model has a poor predictive capability. In general, lower RMSE and MAE indicate better model prediction accuracy [26].

4. Results and analyses

4.1 Statistical analysis of Zn content in soil

Statistical results of Zn contents for soil samples in the Urumqi are given in **Table 1**. Standard deviation (SD) and coefficient of variation (CV) were used to measure data dispersion, with the CV used as a complement to the SD. **Table 1** shows that the Zn contents of soil samples are distributed in the range

of 34.00-200.00 mg/kg, with an average value of 60.88 mg/kg. The average Zn contents of modeling set and validation set are 60.57 mg/kg and 62.06 mg/kg, respectively. The SD of modeling set and validation set are 21.18 and 35.28 mg/kg, respectively. And, the CV values of modeling set and validation set are 0.35 and 0.57, respectively. It's clear that the average and CV values of Zn contents in the modeling set are essentially the same as those of the validation set. Influenced by two high value sample points, the SD values have a difference. This is because the main factories are located in the northern and northeastern parts of Urumqi and the roads with high traffic volumes stretch across the city center [16]. So, two high value sample points were split into modelling set and a validation set respectively. Overall, it indicates that the division of soil samples was reasonable and can be used for subsequent model construction.

4.2 Correlation between soil Zn content and reflectance data

The PCC analysis was performed between the Zn content and the spectral data after 17 types of transformations and R, which can identify the correlation between Zn content and spectral data of soil samples. The degree of correlation was expressed by the Pearson coefficient (r), and PCC was examined in the significance test at the $P < 0.01$ level (two-sided).

In **Figure 3**, the spectrum of the R, RMS, and LT showed a highly significant negative correlation with Zn content with 1730 important wavelengths selected. The spectrum of the AT, RL, and RT showed a highly significant positive correlation with Zn content with 1730 important wavelengths selected. The correlation analysis of the Zn content and spectral data processed by first-order and second-order differentiation transformed indicated that both positive and negative correlation coefficients showed extreme values, and the positive and negative correlations of the filtered important wavelengths were more uniformly distributed. Thus, 18 types of spectrum can filter out characteristic bands for data modeling, and the number of the important wavelengths is descended in the order of: R(1730) = RMS(1730) = LT(1730) = RL(1730) = RT(1730) = AT(1730) > RLFD(663) > FD(502) > RTFD(436) > RTSD(387) > RMSFD(306) > LTSD(253) = ATSD(253) > LTFD(194) = ATFD(194) > RMSSD(186) > SD(125) > RLSD(82).

The number of important wavelengths selected by CARS is descended in the order of RMSSD(25) = RMSFD(25) = RTFD(25) > RLSD(23) > RTSD(22) > ATSD(21) > LTFD(20) = ATFD(20) = FD(20) > SD(19) > R(16) > LTSD(14) > RLFD(13) = AT(13) > RMS(12) > RL(11) > LT(9) = RT(9). The number of important wavelengths selected by CARS is less than the important wavelengths selected by PCC.

4.3 The establishment and analysis of the spectral inversion prediction model

Partial least squares regression model (PLSR) and random forest regression model (RFR) were constructed to predict the Zn content of soil in this study. Based on Python, the "random_state" of three models was set as 48. Due to the randomness of the RFR model, the number of parameters ("n_estimators" and another "random_state") will disturb the predictive performance of the model. Under the consideration of model performance, model running time, sample number and other factors, the number of parameters ("n_estimators" and another "random_state") of the RFR model was set in the range from 1 to 99. The modeling set is used to construct the inversion model, whereas the validation set is used to evaluate the performance of the final model. According to the correlation coefficient between the Zn content and the spectrum, wavelengths with absolute values more than 0.272 under the processed spectral reflectance data were taken as important wavelengths. Then, the important wavelengths are selected as the independent variables (x), and the Zn contents of the soil are selected as the dependent variables (y). The hyperspectral inversion model for soil Zn content was established by the partial least squares regression (PLSR) and the random forests regression algorithms, respectively.

Table 1. Statistical values of Zn contents in soil in the Urumqi.

Data set	Samples/n	Minimum	Maximum	Average	SD	CV
Modeling set (mg/kg)	70	37.00	164.00	60.57	21.18	0.35
Validation set (mg/kg)	18	34.00	200.00	62.06	35.52	0.57
Total (mg/kg)	88	34.00	200.00	60.88	24.81	0.41

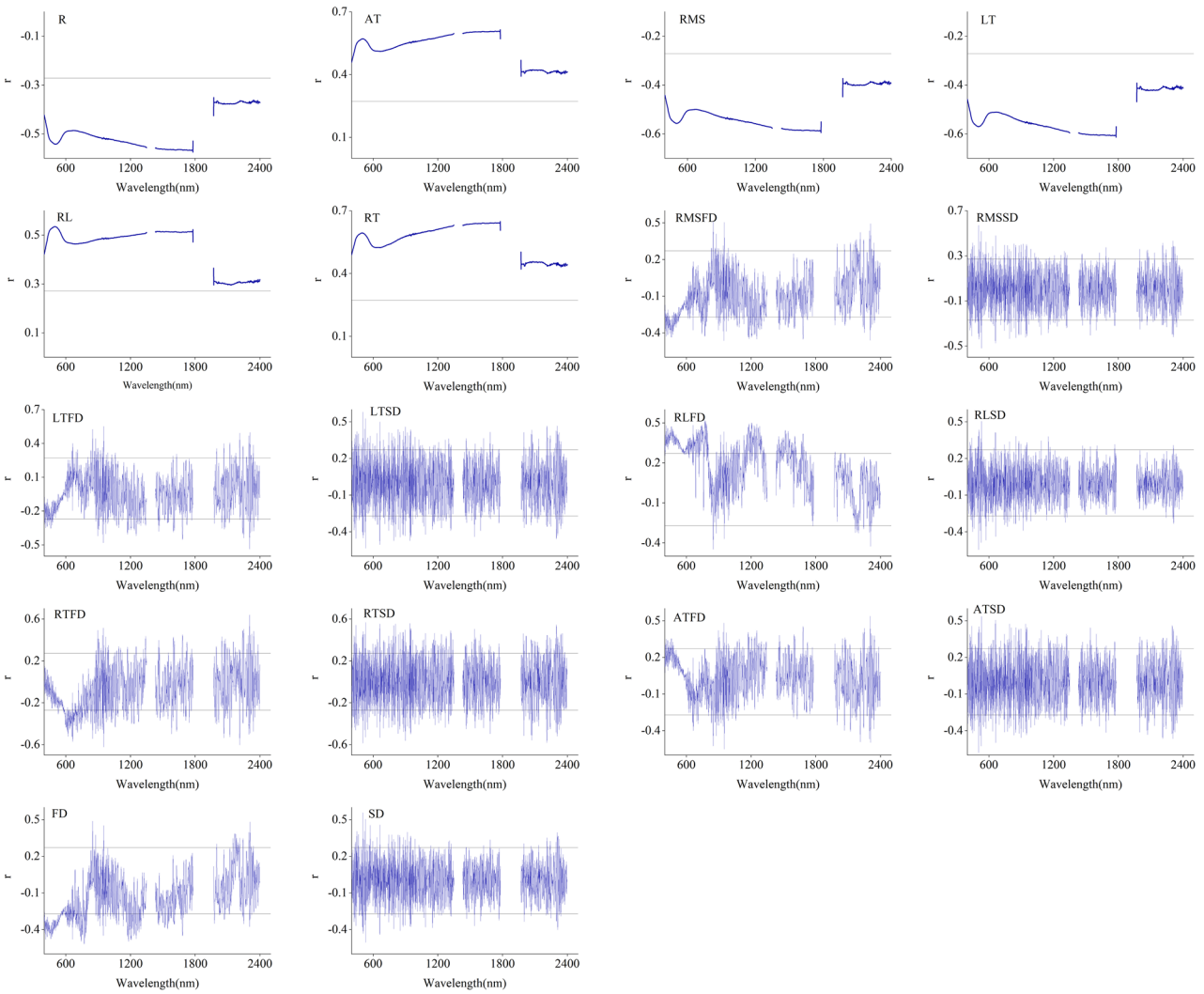

Figure 3. Correlations of PCC between soil Zn content and spectral reflectance data.

The analysis of the PLSR model

The basic statistics related to the stability and accuracy of the PLSR model are given in **Table 2**.

As shown in **Table 2**, the R^2 inversed by the PLSR model based on the important wavelengths selected by PCC range from 0.208 to 0.540, RMSE values range from 24.089 to 31.621, MAE values range from 3.981 to 4.095, and RPD values range from 1.123 to 1.475. For the RTSD-PLSR model (R^2 = 0.540, RMSE = 24.089, MAE = 3.981, and RPD = 1.475), the estimation capability of the remaining models is poor and the prediction accuracy is low.

The ranges of R^2, RMSE, MAE, and RPD values inversed by PLSR based on the important wavelengths selected by CARS are 0.135-0.937, 8.914-33.039, 2.735-4.305, and 3.985-1.075, respectively. The prediction accuracy of CARS-RTFD-PLSR model is highest (R^2 = 0.937, RMSE = 8.914, MAE = 2.735, RPD = 3.985). PLSR model based on AT, RMS, RT, RMSFD, LTFD, ATFD, and FD has the better estimation capability.

Table 2. Statistics of accuracy parameters of PLSR model for soil Zn content in Urumqi.

Transformation	PCC				CARS			
	R^2	RMSE	MAE	RPD	R^2	RMSE	MAE	RPD
R	0.376	25.051	4.037	1.418	0.496	24.209	3.919	1.467
AT	0.329	29.095	4.215	1.221	0.550	23.833	3.935	1.490
RMS	0.303	29.650	4.232	1.198	0.532	24.297	4.050	1.462
LT	0.329	29.095	4.215	1.221	0.480	25.624	3.805	1.386
RL	0.271	30.334	4.048	1.171	0.200	31.770	4.651	1.118
RT	0.346	28.735	3.951	1.236	0.506	24.977	3.653	1.422
RMSFD	0.263	30.502	3.894	1.165	0.591	22.722	4.128	1.563
RMSSD	0.389	27.776	4.079	1.279	0.332	29.023	4.307	1.224
LTFD	0.481	25.579	3.752	1.389	0.762	17.323	3.536	2.050
LTSD	0.427	26.899	4.032	1.320	0.135	33.039	4.305	1.075
RLFD	0.208	31.621	4.095	1.123	0.168	32.402	4.294	1.096
RLSD	0.372	28.158	4.066	1.261	0.299	29.732	4.038	1.195
RTFD	0.358	28.451	3.916	1.248	0.937	8.914	2.735	3.985
RTSD	0.540	24.089	3.981	1.475	0.337	28.932	4.045	1.228
ATFD	0.474	25.765	3.926	1.379	0.510	24.856	4.114	1.429
ATSD	0.426	26.905	4.036	1.320	0.333	29.017	4.367	1.224
FD	0.384	27.889	4.036	1.274	0.687	19.833	3.871	1.791
SD	0.367	28.263	4.095	1.257	0.397	27.594	4.062	1.287

R (original spectral reflectance data); AT (logarithm of the reciprocal); RMS (root mean square); LT (logarithm); RL (reciprocal of the logarithm); RT (reciprocal); RMSFD (root mean square first-order differentiation); RMSSD (root mean square second-order differentiation); LTFD (first-order differentiation of the logarithm); LTSD (second-order differentiation of the logarithm); RLFD (logarithmic first order differentiation of the reciprocal); RLSD (logarithmic second order differentiation of the reciprocal); RTFD (first-order differentiation of the reciprocal); RTSD (second-order differentiation of the reciprocal); ATFD (logarithmic first order differentiation of the reciprocal); ATSD (logarithmic second order differentiation of the reciprocal); FD (first-order differentiation), SD (second-order differentiation).

In general, CARS is superior to PCC, and the CARS-RTFD-PLSR model is better than the RT-FD-PLSR model. A map of the spatial distribution (**Figure 4**) illustrates the relationship between the predicted contents of Zn and the measured contents of Zn in the study area.

The analysis of the RFR model

In **Table 3**, the ranges of R^2, RMSE, MAE and RPD values inversed by the RFR model based on the important wavelengths selected by PCC are 0.477-0.799, 1.414-4.714, 12.417-21.481, and 7.535-25.120, respectively. The R^2 is higher than 0.5 except for the RLSD-RFR model, so the RFR model has good prediction ability. The best inverse prediction model is the FD-RFR model (R^2 = 0.799, RMSE = 2.711, MAE = 12.417, and RPD = 13.102).

The R^2 inversed by the RFR model based on the important wavelengths selected by CARS ranges from 0.316 to 0.856, the ranges of RMSE and MAE values are 1.100-7.377 and 10.074-20.343, and the RPD values are 4.815-30.127. The prediction accuracy of CARS-LTFD-RFR model is highest (R^2 = 0.856, RMSE = 2.514, MAE = 10.074, and RPD = 14.129).

CARS is superior to PCC, and the CARS-LTFD-RFR model is better than the FD-RFR model. However, all the estimation capability of the RFR model is good because the values of RPD are higher than 2.0. A map of the spatial distribution illustrates the relationship between the predicted contents of Zn and the measured contents of Zn in the study area (**Figure 5**).

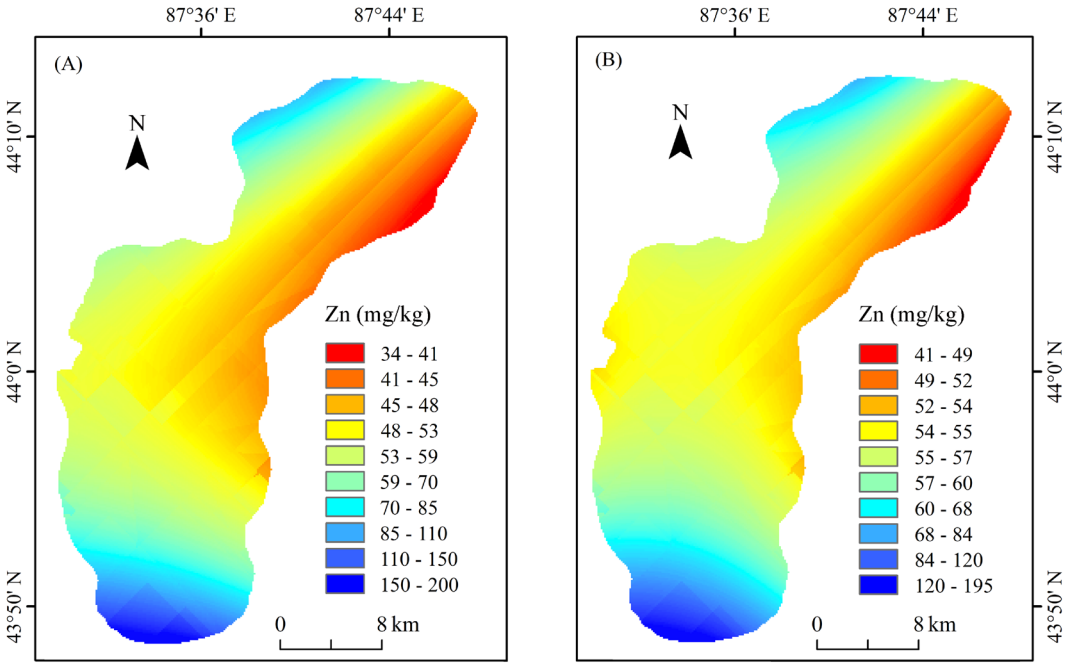

Figure 4. Distribution of Zn content based on measured values (A) and PLSR predicted values (B).

Table 3. Statistics of accuracy parameters of RFR model for Zn content of soils in Urumqi.

Transformation	PCC				CARS			
	R^2	RMSE	MAE	RPD	R^2	RMSE	MAE	RPD
R	0.509	3.012	14.228	11.793	0.437	1.179	14.926	30.127
AT	0.584	2.593	14.722	13.698	0.470	3.477	17.097	10.216
RMS	0.509	3.012	14.228	11.793	0.437	3.693	16.259	9.618
LT	0.508	3.012	14.253	11.793	0.488	2.357	14.426	15.070
RL	0.584	2.593	14.806	13.698	0.457	1.100	14.148	32.291
RT	0.584	2.593	14.722	13.698	0.518	5.215	15.646	6.811
RMSFD	0.578	2.671	21.481	13.298	0.598	3.435	12.992	10.341
RMSSD	0.640	2.027	14.822	17.523	0.316	2.678	18.884	13.264
LTFD	0.524	1.500	14.126	23.680	0.856	2.514	10.074	14.129
LTSD	0.714	2.216	13.622	16.029	0.388	3.359	18.903	10.575
RLFD	0.744	3.614	12.963	9.828	0.463	7.377	15.206	4.815
RLSD	0.477	2.269	15.729	15.654	0.318	2.721	20.343	13.054
RTFD	0.692	3.300	16.944	10.764	0.747	2.095	11.648	16.955
RTSD	0.683	2.711	14.694	13.102	0.536	1.886	13.846	18.834
ATFD	0.628	1.414	17.556	25.120	0.709	2.887	12.352	12.303
ATSD	0.655	2.828	15.556	12.560	0.454	3.435	18.111	10.341
FD	0.799	2.711	12.417	13.102	0.777	1.886	11.778	18.834
SD	0.593	4.714	16.889	7.535	0.465	2.528	15.949	14.051

R (original spectral reflectance data); AT (logarithm of the reciprocal); RMS (root mean square); LT (logarithm); RL (reciprocal of the logarithm); RT (reciprocal); RMSFD (root mean square first-order differentiation); RMSSD (root mean square second-order differentiation); LTFD (first-order differentiation of the logarithm); LTSD (second-order differentiation of the logarithm); RLFD (logarithmic first order differentiation of the reciprocal); RLSD (logarithmic second order differentiation of the reciprocal); RTFD (first-order differentiation of the reciprocal); RTSD (second-order differentiation of the reciprocal); ATFD (logarithmic first order differentiation of the reciprocal); ATSD (logarithmic second order differentiation of the reciprocal); FD (first-order differentiation), SD (second-order differentiation).

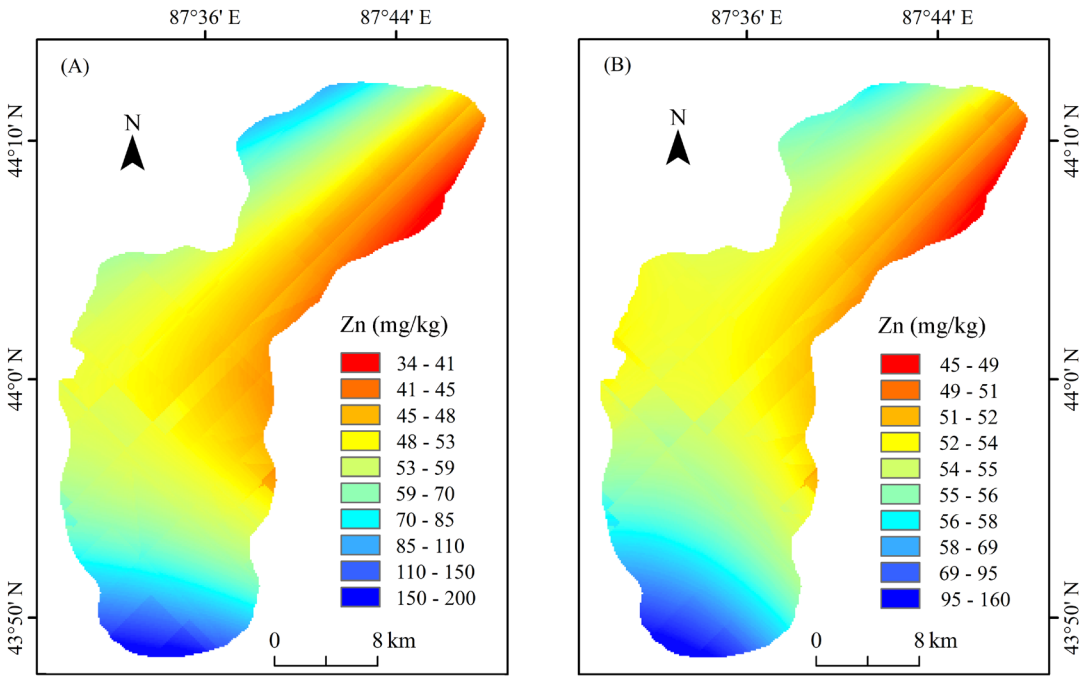

Figure 5. Distribution of Zn content based on field measured values (A) and RFR predicted values (B).

4.4 Discussion of optimal prediction models

On the one hand, estimating Zn content in soils using hyperspectral remote sensing is a cost-efficient method but challenging due to the effects of natural environmental conditions and soil properties [27]. On the other hand, high-data dimensionality is a common problem in hyperspectral data processing, so the inversion accuracy of the constructed model is biased by redundant spectra and noise [23,28].

In this study, the predicted accuracy of the soil Zn content is $R^2_{CARS\text{-}RTFD\text{-}PLSR} > R^2_{RLSD\text{-}RFR} > R^2_{CARS\text{-}LTFD\text{-}RFR} > R^2_{RTSD\text{-}PLSR}$. Therefore, combined with the performance of the prediction accuracy of soil Zn content, the prediction accuracy of PLSR among the modeling methods is significantly better than that of RFR. As shown in **Tables 2 and 3**, the fitness, stability and accuracy of the prediction model are changed to different degrees after processing methods of the original spectral data. The best predict prediction model is the CARS-RTFD-PLSR (partial least squares regression model based on first-order differentiation of the reciprocal by CARS) model (R^2 = 0.937, RMSE = 8.914, MAE = 2.735, and RPD = 3.985), which has the better ability to invert the soil heavy metal content in the study area. The scatter plot of the measured and predicted values of Zn content modeling by CARS-RTFD- PLSR and R-PLSR model was exhibited in **Figure 6**.

The R^2 calculated by the PLSR model constructed based on CARS-RTFD of the important wavelengths is significantly higher than that modeled from the original spectral data, and both the RMSE and MAE are significantly decreased. From **Figure 6**, it can be intuitively seen that the prediction accuracy of the PLSR model based on original spectral data was not high, and the R^2 between the measured and predicted values was 0.496. The prediction accuracy of the CARS-RTFD was improved significantly, and the predicted and measured values presented a good agreement with each other, with R^2 of 0.937, which was improved by 0.441 compared with the R-PLSR model. Overall, a faster and more convenient method for estimating Zn content in soil is described in this work. This method provides an effective way for predicting soil Zn contents in oasis cities.

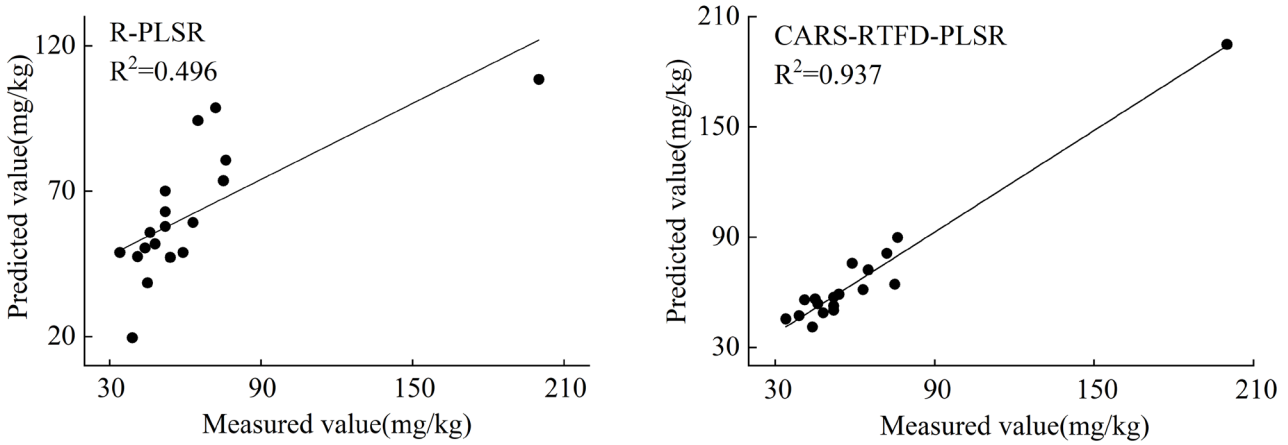

Figure 6. Measured and CARS-RTFD-PLSR predicted values of Zn content in soil.

5. Conclusions

To find an optimal model to predict the soil Zn content for the study area, the PLSR model and the RFR model were constructed based on the important wavelengths and Zn content from soil samples. The results of this study lead to the following conclusions:

1) Transformed spectral data with Pearson's correlation coefficient analysis and CARS can obviously reduce the interference of the environmental background and improve the correlations between soil spectral reflectance data and Zn contents of soil. The first-order differentiation of the reciprocal (RTFD) has the most significant enhancement of spectral features.

2) The results showed that the CARS-RTFD-PLSR model is more stable with the highest prediction ability (R^2 = 0.937, RMSE = 8.914, MAE = 2.735, and RPD = 3.985) for soil Zn content in the research region. The CARS-RTFD-PLSR method can provide a reference method and technical support for the prediction of soil Zn content in oasis cities.

Overall, the results of this study demonstrate the possibility of directly applying hyperspectral remote sensing approaches to estimating soil Zn contents in oasis cities. This method can provide technical support for the hyperspectral estimation of the soil Zn content and can require rapid detection of Zn a contamination of soil. However, the limitation of this study is the lack of combination of hyperspectral and remote sensing imagery, which needs to be further verified in subsequent studies.

Author Contributions

Qing Zhong completed all the experiments to obtain data, processed the data, and wrote the main manuscript text. Mamattursun Eziz provided support and gave guidance for this study. Rukeya Sawut and Mireguli Ainiwaer taught the methods. All authors reviewed the manuscript.

Conflict of Interest

The authors declare no conflicts of interest.

Funding

This research was funded by the National Natural Science Foundation of China (No. U2003301) and the Tianshan Talent Training Project of Xinjiang.

Acknowledgement

The original version of this paper was substantially improved thanks to the constructive comments by anonymous reviewers.

References

[1] Zhao, R.X., 2004. Huan jing wu ran hua xue

(Chinese) [Environmental pollution chemistry]. China Industry Press: Beijing.

[2] Chen, Y.Z., Wang, F., Wang, G., et al., 2012. Research advances on zinc pollution and remediation of soil system. Fujian Journal of Agricultural Sciences. 27(8), 901-908.

[3] Patel, A.K., Ghosh, J.K., Sayyad, S.U., 2022. Fractional abundances study of macronutrients in soil using hyperspectral remote sensing. Geocarto International. 37(2), 474-493.

[4] Yang, Y., Cui, Q.F., Jia, P., et al., 2021. Estimating the heavy metal concentrations in topsoil in the Daxigou mining area, China, using multispectral satellite imagery. Scientific Reports. 11, 11718.

[5] Wei, L.F., Pu, H.C., Wang, Z.X., et al., 2020. Estimation of soil arsenic content with hyperspectral remote sensing. Sensor. 20, 4056-4071.

[6] Tan, K., Wang, H.M., Chen, L.H., et al., 2021. Estimating the distribution trend of soil heavy metals in mining area from HyMap airborne hyperspectral imagery based on ensemble learning. Journal of Hazardous Materials. 401, 1-17.

[7] Ye, M., Zhu, L., Li, X.J., et al., 2022. Estimation of the soil arsenic concentration using a geographically weighted XGBoost model based on hyperspectral data. Science of the Total Environment. 858, 159798-159798.

[8] Bian, Z.J., Sun, L.N., Tian, K., et al., 2021. Estimation of heavy metals in Tailings and soils using hyperspectral technology: A case study in a Tin Polymetallic mining area. Bulletin of Environmental Contamination and Toxicology. 107, 1022-1031.

[9] Zhang, B., Guo, B., Zou, B., et al., 2022. Retrieving soil heavy metals concentrations based on GaoFen-5 hyperspectral satellite image at an opencast coal mine, Inner Mongolia, China. Environmental Pollution. 300, 118981-118992.

[10] Hou, L., Li, X.J., Li, F., 2018. Hyperspectral-based inversion of heavy metal content in the soil of coal mining areas. Journal of Environmental Quality. 48, 57-63.

[11] Omran, E.S.E., 2016. Inference model to predict heavy metals of Bahr El Baqar soils, Egypt using spectroscopy and chemometrics technique. Modeling Earth Systems and Environment. 2, 1-17.

[12] Yang, H.F., Xu, H., Zhong, X.N., 2022. Prediction of soil heavy metal concentrations in copper tailings area using hyperspectral reflectance. Environmental Earth Sciences. 81, 183-193.

[13] Wei, B., Jiang, F., Li, X., et al., 2010. Heavy metal induced ecological risk in the city of Urumqi, NW China. Environmental Monitoring and Assessment. 160, 33-45.

[14] Li, J.M., Zhang, Y.T., 2019. Wu lu mu qi bu tong gong neng qu lin dai tu rang zhong jin shu wu ran te zheng fen xi (Chinese) [Characteristics of heavy metal pollution in forest belt soil of different functional zones in Urumqi, Xinjiang]. Journal of Environmental Sciences. 28, 1859-1866.

[15] Sidikjan, N., Eziz, M., Li, X., et al., 2022. Spatial distribution, contamination levels, and health risks of trace elements in topsoil along an urbanization gradient in the City of Urumqi, China. Sustainability. 14(19), 12646.

[16] Hini, G., Eziz, M., Wang, W., et al., 2020. Spatial distribution, contamination levels, sources, and potential health risk assessment of trace elements in street dusts of Urumqi city, NW China. Human and Ecological Risk Assessment: An International Journal. 26(8), 2112-2128.

[17] Yao, X.D., Wang, J., Wang, Y.M., et al., 2022. Wu lu mu qi mou gong ye yuan qu tu rang zhong jin shu qian zai sheng tai feng xian ping jia (Chinese) [Potential ecological risk assessment on heavy metals in the soil of an industrial park in Urumqi, China]. Transactions of Nonferrous Metals Society of China. 12, 160-166.

[18] Soil and Sediment-Determination of Aqua Regia Extracts of 12 Metal Elements-Inductively Coupled Plasma Mass Spectrometry [Internet]. Ministry of Environmental Protection of the People's Republic of China; 2016. Available from: https://english.mee.gov.cn/Resources/standards/Soil/Method_Standard4/201607/t20160704_357088.shtml

[19] Yuan, Z.R., Wei, L.F., Zhang, Y.X., et al., 2020.

Hyperspectral inversion and analysis of heavy metal arsenic content in farmland soil based on optimizing CARS combined with PSO-SVM algorithm. Spectroscopy and Spectral Analysis. 40(2), 567-573.

[20] Zhong, X.J., Yang, L., Zhang, D.X., et al., 2022. Effect of different particle sizes on the prediction of soil organic matter content by visible-near infrared spectroscopy. Spectroscopy and Spectral Analysis. 42(8), 2542-2550.

[21] Ma, X.M., Zhou, K.F., Wand, J.L., et al., 2022. Optimal bandwidth selection for retrieving Cu content in rock based on hyperspectral remote sensing. Journal of Arid Land. 14(1), 102-114.

[22] Samuel, N.A., Anna, F.H., Andreas, A., et al., 2021. Advances in soil moisture retrieval from multispectral remote sensing using unoccupied aircraft systems and machine learning techniques. Hydrology and Earth System Sciences. 25, 2739-2758.

[23] Michelle, D., Onisimo, M., Riyad, I., 2011. Examining the utility of random forest and AISA Eagle hyperspectral image data to predict Pinus patula age in KwaZulu-Natal, South Africa. Geocarto International. 26(4), 275-289.

[24] Rukeya, S., Nijat, K., Abdugheni, A., et al., 2018. Possibility of optimized indices for the assessment of heavy metal contents in soil around an open pit coal mine area. International Journal of Applied Earth Observation and Geoinformation. 73, 14-25.

[25] Vohland, M., Joachim, B., Joachim, H., et al., 2011. Comparing different multivariate calibration methods for the determination of soil organic carbon pools with visible to near infrared spectroscopy. Geoderma. 166, 198-205.

[26] Wang, Y.Y., Niu, R.Q., Lin, G., et al., 2023. Estimate of soil heavy metal in a mining region using PCC-SVM-RFECV-AdaBoost combined with reflectance spectroscopy. Environmental Geochemistry and Health. Ahead of print. DOI: https://doi.org/10.1007/s10653-023-01488-W

[27] Liu, W.W., Li, M.J., Zhang, M.Y., et al., 2020. Hyperspectral inversion of mercury in reed leaves under different levels of soil mercury contamination. Environmental Science and Pollution Research. 27, 22935-22945.

[28] Zhou, M., Zou, B., Tu, Y.L., et al., 2022. Spectral response feature bands extracted from near standard soil samples for estimating soil Pb in a mining area. Geocarto International. 37(26), 13248-13267.