remote sensing

MDPI

Technical Note

Geographically Weighted Regression Effects on Soil Zinc Content Hyperspectral Modeling by Applying the Fractional-Order Differential

Xue Lin [1], Yung-Chih Su [1], Jiali Shang [2], Jinming Sha [1,*], Xiaomei Li [3], Yang-Yi Sun [4], Jianwan Ji [5,6] and Biao Jin [1]

[1] College of Geographical Science, Fujian Normal University, Fuzhou 350117, China; linxue9211@163.com (X.L.); surdrew@yahoo.com.tw (Y.-C.S.); jinbiao@fjnu.edu.cn (B.J.)
[2] Ottawa Research and Development Centre, Agriculture and Agri-Food Canada, 960 Carling Ave, Ottawa, ON K1A 0C6, Canada; shang.jiali@gmail.com
[3] College of Environmental Science and Engineering, Fujian Normal University, Fuzhou 350117, China; lixiaomei@fjnu.edu.cn
[4] Institute of Geophysics and Geomatics, China University of Geosciences, Wuhan 430074, China; yysun0715@gmail.com
[5] Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, Beijing 100101, China; jijw@radi.ac.cn
[6] University of Chinese Academy of Sciences, Beijing 100049, China
* Correspondence: jmsha@fjnu.edu.cn

Received: 31 December 2018; Accepted: 6 March 2019; Published: 15 March 2019

Abstract: With the development of remote sensing techniques and the increasing need for soil contamination monitoring, we estimated soil heavy metal zinc (Zn) content using hyperspectral imaging. Geographically weighted regression (GWR), an extension of the ordinary least squares (OLS) regression framework, was proposed. By estimating a set of parameters for any number of locations in a study area, GWR can probe the spatial heterogeneity in data relationships, whereas the regression parameters of an OLS model are global and aspatially-varied. The objectives of this study were: (1) To find the possible relationships between hyperspectral data and soil Zn content, and (2) to investigate the existence of their spatial heterogeneity. In this study, 67 soil samples collected from Pingtan Island, Fujian Province, China, were used to conduct laboratory hyperspectral modeling for soil Zn content estimation. Four transformations of square root, logarithm, reciprocal of logarithm, and reciprocal, as well as the fractional-order differential operations were applied to increase the amount of reflectance data in which the effective variables for modeling might be involved, and to enhance the spectral characteristics of soil Zn content. To find sensitive variables and to remove redundancy and multicollinearity in the spectra, a data sifting process was applied by selecting wavelengths with local maximum in the absolute values of the correlation coefficients with Zn content in one type of spectral data and by employing Variance Inflation Factors. Since a modeling sample size of 46 is insufficient to construct the appropriate OLS and GWR models, four methods are proposed using all 67 samples to choose explanatory variables. A random process to select 57 samples for modeling and 10 samples for validation was applied to assess model performance, in which the mean verification R^2 (R_v^2) was used as an indicator. The results show that GWR stepwise regression is the most effective method to select better variables. As the mean R_v^2 converges toward the OLS value when the bandwidth of the GWR model increases, the four variables selected by the GWR stepwise regression were used to establish the representative OLS and GWR models. The representative OLS model has the best mean verification effect among all studied models, which had a mean R_v^2 value that is 44.6% higher than the OLS model constructed using OLS stepwise regression.

Keywords: GWR; soil zinc content; hyperspectra; fractional-order differential

1. Introduction

With intensified human activity, an increasing amount of heavy metals has entered the soil and is posing serious threats to human and animal health, and the environment. Accurate and timely measurements of heavy metal content in the soil are thus urgently needed. Conventional investigations into heavy metal pollution have been mainly based on field sampling and chemical analyses in laboratories, which is time- and labor-consuming, and hence cannot satisfy the dynamic needs over large areas. In this study, we completed a modeling investigation using 67 soil samples with heavy metal Zn contents and their hyperspectral measurements. The soil samples were collected from Pingtan Island, Fujian province, China. The study was conducted to offer a reference for heavy metal content retrieval from satellite-based and aerial hyperspectral data for large-scale and highly-efficiency quantitative monitoring. Many researchers have attempted to use hyperspectral remote sensing technology to retrieve soil heavy metal content [1–3]. Not all metals have unique spectral features, and some metals only exhibit subtle spectral disparity, thereby complicating the detection of heavy metal in soil. Thus, it will be even harder to retrieve soil heavy metal content. In addition to metal type and metal concentration, other factors, such as soil organic matter, moisture content, and the state of the metals (e.g., ions or in compound) also create challenges for the retrieval processes.

Earlier studies revealed that different metals are sensitive to different wavelength regions [4,5]. For example, mercury has absorption features located at 392–455, 923–1040, and 1806–1969 nm. Comparison of three wavelength ranges (380–1100, 1000–2500, and 380–2500 nm) on the effect of prediction accuracy of mercury concentration showed that the best results were acquired using the 1000–2500 nm spectral region [5]. The degree of overlap between different heavy metal elements and organic matter at hyperspectral sensitive bands, and the positive and negative consistency of regression coefficients vary greatly; the greater the correlation with organic matter, the higher the degree of overlap, and the better the positive and negative consistency [6].

In this study, the ordinary least squares (OLS) model and the geographically weighted regression model (GWR) [7], developed in the 1990s [8,9], were employed to investigate the relationship between heavy metal Zn content and hyperspectral data in Pingtan Island. In the GWR model, the geographic locations of the spatial data are considered when calculating the regression coefficients. For each location, the regression coefficients are acquired by the local regression, in which the errors in the neighboring observations are more important, and their values are not the same over space. Thus, GWR can be used to examine the spatial heterogeneity in data relationships. Jaber and Al-Qinna [10] compared the effectiveness of the global multiple linear regression (MLR) and local spatial GWR to produce a map showing the spatial variations in soil organic carbon (SOC) in Amman-Zarqa Basin in Jordan, using Landsat Thematic Mapper (TM) data. They revealed that the GWR model produced a better modeling result. Jiang et al. [11] compared the prediction results based on GWR and OLS regression for 132 heavy metal-contaminated soil samples and transformed spectral data in Fuzhou City, Fujian province, China, and showed that the applicability of the GWR model was dependent on the spatial heterogeneity level of heavy metal influence on spectral variables. They also showed that heavy metal spectral properties were intensified by different spectral transformations, among which reciprocal transformation was most effective. Reciprocal transformation and its derivative patterns improved the performance of heavy metal prediction models. Jiang et al. [12] analyzed the influence of different spectral resolutions and transformations to establish a GWR hyperspectral prediction model of soil chromium (Cr) content in Fuzhou City. The results showed that, at a resolution of 10 nm and with the second derivative of reflectance and reflectance reciprocal as independent variables, the GWR model displayed the best prediction performance. The prediction performance of the GWR model showed a tendency to stabilize with increasing numbers of sample sites.

Given the usefulness of the spectral transformations in hyperspectral modeling, four transformations (square root, logarithm, reciprocal of logarithm, and reciprocal) were applied as pretreatments in this study. As the curves of spectral data can have larger differences with their first- and second-order derivatives, the transitional information may be omitted when only integer-order (IO) derivatives are considered. Xu et al. [13] showed that the application of fractional-order (FO) differentials enriched preprocessing of near-infrared spectral data, and could increase the accuracy of model predictions. To exploit the potential information in the spectral data, the Grünwald–Letnikov's Fractional-Order Differential Equations (FDEs) with FOs ranging from 0 to 2 and stepped by 0.2 were also applied on the raw reflectance and the transformed data in this study. Wang et al. [14] measured the visible and near-infrared (Vis-NIR) spectroscopy and soil organic matter (SOM) content of 103 soil samples collected in the Ebinur Lake basin, Xinjiang Uighur Autonomous Region, China, and employed the fractional derivative algorithm for pretreatment. Partial least squares regression (PLSR) was applied for model calibration in their study. Their results showed that the most robust model was calibrated based on 1.8-order derivative of raw spectral reflectance data. Wang et al. [15] preprocessed the hyperspectral data of 168 samples of soil obtained from an open coalmine area in Eastern Junggar Basin, China, by using FO differential algorithm. All the wavelengths among 401–2400 nm were used to calibrate the hyperspectral estimation models of soil Cr content using PLSR. Their results showed that both the fractional order differential models of the raw reflectance and the absorption rate transformation achieved the best performance at the 1.8-order derivative. The model based on the 1.8-order derivative of absorbance transformation better predicted Cr content in desert soils.

When modeling soil Zn content, it is worth discussing if collinearity exists between the selected effective variables and their collinearity. O'Brien [16] examined the measures of the degree of multicollinearity, such as the Variance Inflation Factor (VIF) and tolerance. The author found that threshold values of the VIF (and tolerance) need to be evaluated in the context of several other factors that influence the variance of regression coefficients. VIF values of 10, 20, 40, or even higher do not discount the results of regression analyses, require the elimination of one or more independent variables, suggest the use of ridge regression, or require combining independent variables into a single index. Fotheringham and Oshan [17] suggested that multicollinearity in a GWR model should be treated the same as in any regression framework, and is just one of the many factors that may cause problems with model interpretation. The issue of multicollinearity may have been overestimated in previous literature. Therefore, in this study, we examined the collinearity effect in the GWR model verification for variables of different transformations.

Since there are global and local properties in soils, both OLS and GWR models were employed to detect possible relationships between hyperspectral data and soil Zn content. Through two-step data sifting by finding the wavelengths with local maximum in absolute values of correlation coefficient with Zn content and by applying the VIFs, a total of 304 candidates were sifted from the reflectance and transformed variables, with more candidates in reciprocal transformation and FOs ≥ 0.6. It was shown that a modeling sample size of 46 might be insufficient to build the appropriate OLS and GWR models, so four methods using all 67 samples were used to select effective independent variables for GWR modeling. Finally, with the assistance of a random process in the comparison of verification effect, representative OLS and GWR models were obtained, with four variates selected through a GWR stepwise regression.

2. Materials and Methods

2.1. Study Site

Pingtan Island is situated southeast of the coast of Fuzhou City, Fujian Province, China. The coordinates of Pingtan Island are centered approximately at 119°46′ E and 25°32′ N. The area of the main land is about 274.3 km^2. Its elevation is high in the north and south, and low in the middle. The annual precipitation varies from 900 mm to 1200 mm, and the average daily temperature varies from 10.7 °C in February to 27.8 °C in July. The parent materials in the study area are weathered and alluvial materials of magmatic rock. In accordance with the Classification and Codes for Chinese Soil [18] and statistical data from a soil general survey in Pingtan Island, the six main soil types were found to be laterite, aeolian sandy soil, saline soil, paddy soil, red earth, and fluvo-aquic soil.

2.2. Data Collection and Processing

We collected 75 soil samples in July 2013 at 0–20 cm depths on Pingtan Island. To ensure that all representative land types were sampled, the data of interpreted land use, topography, and soil types were combined, and a random sampling design was used to determine the sample sites. There were 2–3 soil samples collected per site, and the samples were then mixed to make a single soil sample for subsequent analysis. The sampling mass was about 1 kg. After collection, the samples were taken to the laboratory for natural air drying. Then, the residues of gravel, plant, and animal material were removed, and the soils were ground with agate mortar and passed through a 100-mesh nylon sieve. Each sample was divided into two portions: one for chemical analysis and the other for spectral analysis. Soil sample spectra were taken using an ASD FieldSpec 3 spectroradiometer (ASD, Boulder, CO USA). When measuring the soil spectra, the soil samples were placed in sample dishes with diameter >10 cm and depth >5 cm, covered by a nearly full-absorption black material to avoid stray light interference. To reduce uncertainty and ensure data accuracy, 10 spectral curves were drawn for each sample. After the anomalous spectral curves were removed, the remaining spectra were averaged as the original spectrum of the sample. The original resampling rate was 1 nm with an average standard deviation of less than 0.01. The soil Zn content was determined using the XSERIES II Inductively Coupled Plasma Mass Spectrometer (ICP-MS) (Thermo Fisher Scientific, Waltham, MA USA) after soil microwave digestion with nitric acid, hydrofluoric acid, and hydrogen peroxide. All reagents used were guaranteed reagent (GR) and quality was controlled using the National Soil Standard Sample (GSB 04-1767-2004). According to the measured reflectance data, one sample that had unusual features relative to other measurements was removed. A total of 74 samples were retained. Then, according to the measured Zn content, a box-and-whisker plot was employed to find outliers. There were 7 outliers found, so 67 samples were used for modeling in this study. The sample size was reduced to ensure that the correlation analysis and the established models would be stable, which reflects the relationships for the majority of the Zn content in a narrower range (approximately <140 mg/kg).

After the samples were determined, hyperspectral reflectance data of the 67 soil samples were preprocessed. We found that the signal-to-noise ratios (SNRs) of measured reflectance between 350 nm and 399 nm and 2401 nm and 2500 nm were very low. Therefore, the portion of the reflectance in the 400–2400 nm range was selected for subsequent processing and analysis. The preprocessing also included a smoothing process to remove data noise and a resampling process to reduce the redundant signals in the hyperspectral reflectance data. In the smoothing process, a 2-time smoothing by using the Savitzky-Golay filter of order 2 and frame size of 25 points (25 nm) was applied, according to a trial with a simplest filter and comparisons for smaller noise vibration. After the smoothing process, the reflectance was averaged every 10 nm and spectral data size was reduced to 200 bands for each sample in the resampling process.

In this study, the 67 samples were divided into two groups: 46 for modeling and 21 for verification. The close sample sites were separated into the two groups, and then sample points were randomly divided in considerations of spatial distribution and interpolatability for verification. After the samples were randomly distributed, a few samples were adjusted. Figure 1a,b display the geographic distributions of Zn content of 46 modeling and 21 verification samples, respectively. The statistical values for the 67 measured Zn contents in Pingtan Island are shown in Table 1.

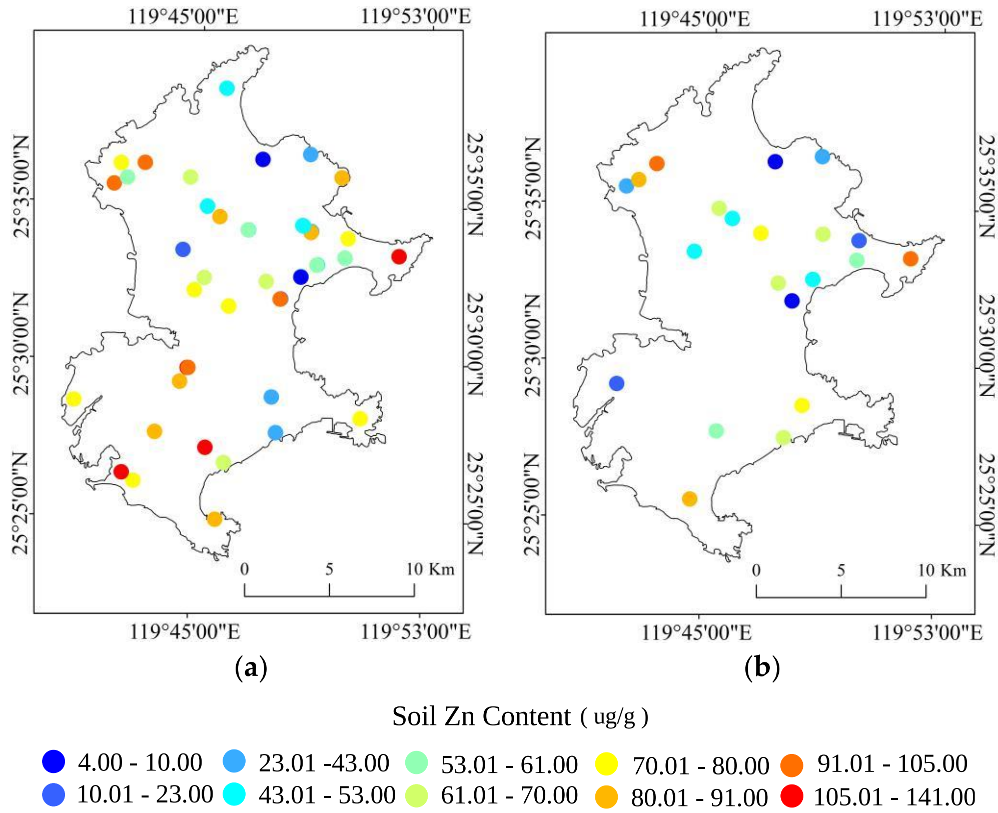


Figure 1. Spatial distribution of the 67 soil samples and their Zn contents. The colors denote the different ranges of soil Zn content, (**a**) the 46 modeling samples and (**b**) the 21 verification samples.

Table 1. Statistical values for measured Zn contents in Pingtan Island. $\overline{x}$, SD, and c_V denote the sample mean, sample standard deviation, and coefficient of variation of the 67 soil Zn contents, respectively.

Heavy Metal Element	Min (mg/kg)	Max (mg/kg)	$\overline{x}$ (mg/kg)	SD (mg/kg)	c_V (%)	Kurtosis	Skewness
Zn	4	141	63.180	30.529	48.321	0.161	0.080

2.3. Data Transformations and Fractional-Order Differentials

Spectral data transformations involve operations on spectral data by applying mathematical methods to obtain a new series of independent variables. Because the contents of heavy metals are relatively low in soil, the associated information reflected in the soil spectra is weak. Spectral transformations have advantages in improving the predictive efficiency of the model [11,12] and in increasing the number of modeling candidates. In this study, the pretreatments of four transformations—square root, logarithm, reciprocal of logarithm, and reciprocal—were first applied to the above preprocessed reflectance. Figure 2 displays the preprocessed reflectance and the associated transformations between 400 nm and 2400 nm. The five types of spectral curves (the original reflectance and associated transformations) shown in Figure 2 reveal notable absorption features at 1400 nm, 1900 nm, and 2200 nm.

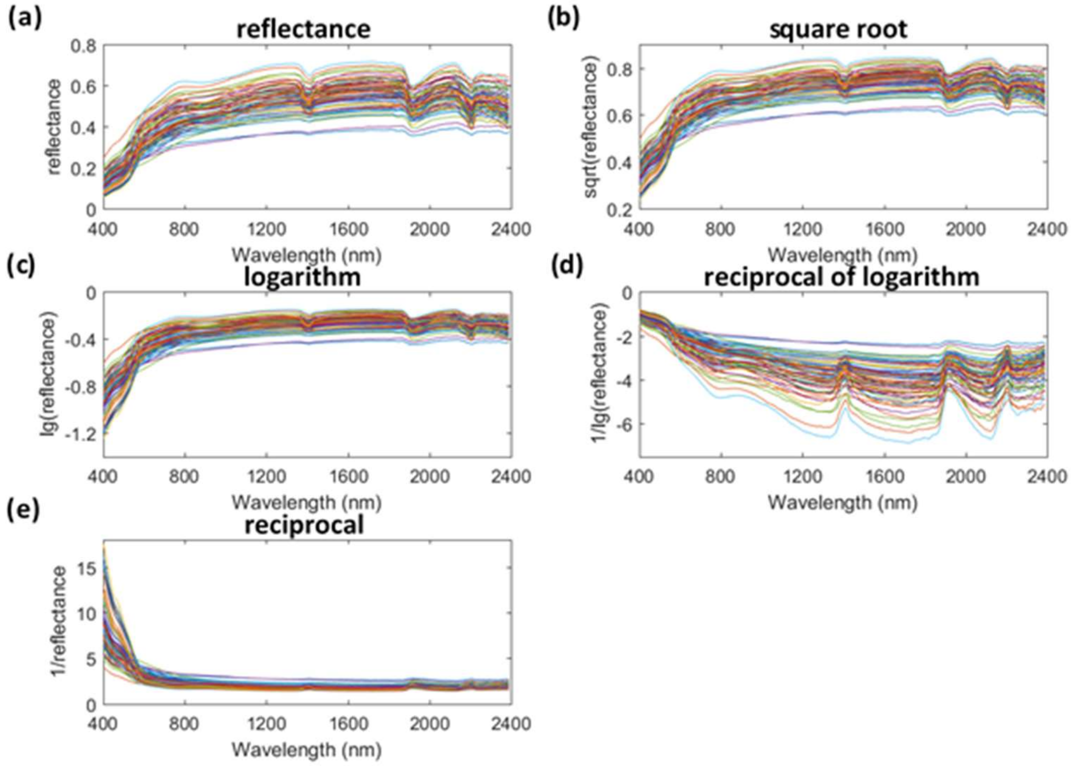


Figure 2. The original hyperspectral data and the associated four transformations: (**a**) original reflectance, (**b**) square root, (**c**) logarithm, (**d**) reciprocal of logarithm, and (**e**) reciprocal.

In addition to the four transformations, we performed differential operations on the above-mentioned five types of spectral data. The relationships and modeling between spectral data of the original, first- and second-order derivatives, and heavy metal contents were analyzed [11,12,19,20]. Because the original spectral curves can have large difference from their first- and second-order derivatives, ignoring the FO derivatives between them may result in the transitional information being omitted [13,15,21]. Fractional-order differentials were adopted in this study to exploit the potential information within the gradual changes between the curves of the original and IO derivatives. In this study, we employed the Grünwald−Letnikov's FDEs to perform the differential operations. The Grünwald−Letnikov's FDEs are given by:

$$\frac{d^v f(x)}{dx^v} \approx f(x) + (-v)f(x-1) + \frac{(-v)(-v+1)}{2}f(x-2) + \ldots + \frac{\Gamma(-v+1)}{n!\Gamma(-v+n+1)}f(x-n), \quad (1)$$

where x is the wavelength, f(x) is the reflectance or the associated transformation, v is the FO of differential, n is the difference between the upper and lower limits of the differential operation (n = x − 1), Γ is Γ-function, and

$$\Gamma(z) = \int_0^\infty e^{-\mathrm{t}} t^{z-1} dt,\ \Gamma(z+1) = z\Gamma(z) \quad (2)$$

is for complex numbers z with a positive real part (Re(z) > 0). In this study, the FDEs with FOs ranging from 0 to 2 and stepped by 0.2 were applied to the reflectance and the associated four transformations. Consequently, 11 differentials were completed and 55 types of spectral data were obtained. The spectral data of wavelengths between 420 nm and 2400 nm were retained from the differential operations.

2.4. Correlation Analyses and Data Sifting

The next step in constructing an OLS or a GWR model of Zn content involved sifting the candidate variables that could have high correlations with the Zn content from the 55 types of spectral data. Correlation analyses were completed to compute the correlation coefficient for each wavelength in the 55 types of spectral data (represented by colors in Figure 3). Figure 3a–e display the reflectance and associated four transformations (square root, logarithm, reciprocal of logarithm, and reciprocal). The x- and y-axes represent the wavelengths and FOs in differential operations, respectively. For the soil sample size of 67, the absolute values of correlation coefficients greater than 0.3125 were considered significant ($\alpha = 0.01$), which are denoted by deep blue and red at opposite sides of the color bar. The five bands having significant correlation with the Zn content are around 420–540 nm, 1400 nm, 1900–1940 nm, 2100–2200 nm, and 2250–2400 nm, especially at the lower FOs. Note that the bands of 1400 nm, 1900–1940 nm, and 2100–2200 nm are the corresponding bands with notable characteristics of absorption, as shown in Figure 2. We still found significantly correlated wavelengths between these five bands in the spectral data. In some transformations, for example, the reciprocal (Figure 3e), the extensions of significant correlation coefficients from lower to higher FOs are remarkable. The reciprocal transformed spectra have the largest number of data with significant correlations with the Zn content. The polarities of correlation of the reflectance and the transformations of square root and logarithm (Figure 3a–c) are roughly consistent, but are opposite to those of the reciprocal of logarithm and reciprocal (Figure 3d,e).

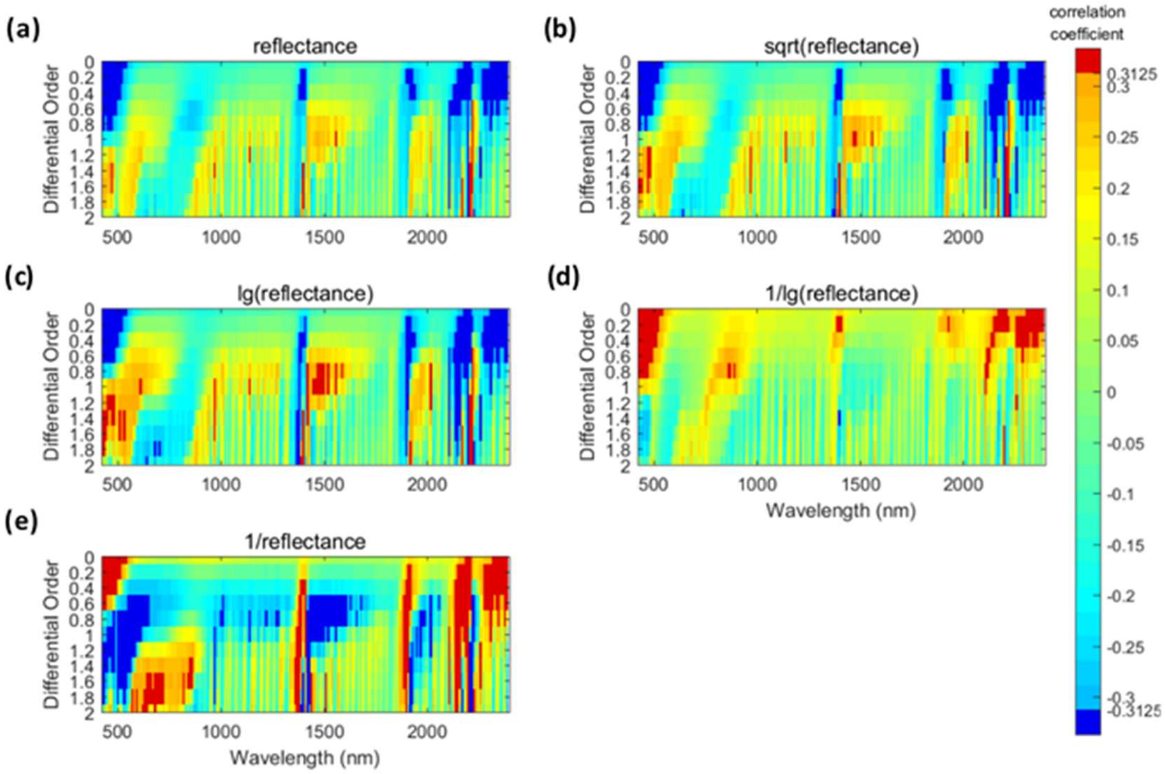


Figure 3. The correlation coefficients of hyperspectral reflectance data and the four transformations and different orders of differential operations applied: (**a**) reflectance, (**b**) square root, (**c**) logarithm, (**d**) reciprocal of logarithm, and (**e**) reciprocal. The x- and y-axes represent the wavelengths and fractional-orders (FOs) in differential operations from 0 to 2, respectively. The colors denote the correlation coefficients. For the soil sample size of 67, the absolute values of correlation coefficients greater than 0.3125 are considered to be significant ($\alpha = 0.01$), which are denoted by deep blue and red in the color bar.

Figure 3 reveals that there were high redundancies in the data that were significantly correlated with Zn content. In this study, the redundancy and multicollinearity in each of the 55 types of spectra were removed. To remove the redundancy and multicollinearity, the wavelengths were first sifted by selecting those with local maxima in the absolute values of the correlation coefficients in one type of the spectral data. After this sifting, the VIFs were employed. The wavelengths were filtered sequentially until the VIFs of all the sifted wavelengths in one type of spectral data were less than 10. Figure 4 displays the numbers of sifted candidates for modeling in the 55 types of data. In total, 304 candidates were sifted from the 55 types of spectral data. Among the 55 types of data, the FOs ≥ 0.6 and the reciprocal transformation yielded relatively more candidates than others. For example, there were 14 candidates in FO = 0.8 and FO = 1.8 of reciprocal. In contrast, there was only one candidate in FOs = 1.2 to 2 of the reciprocal of the logarithm. A flowchart showing the methods in this study is provided in Figure 5.

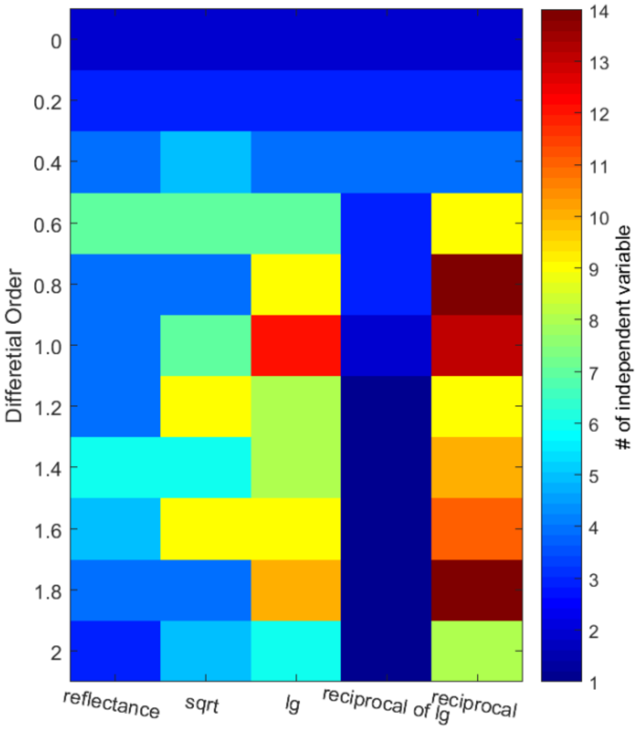


Figure 4. Numbers of sifted candidates for geographically weighted regression (GWR) and ordinary least squares (OLS) modeling in 55 types of spectral data. The x-axis denotes reflectance and the four transformations. We sifted 304 candidates from the 55 types of spectral data.

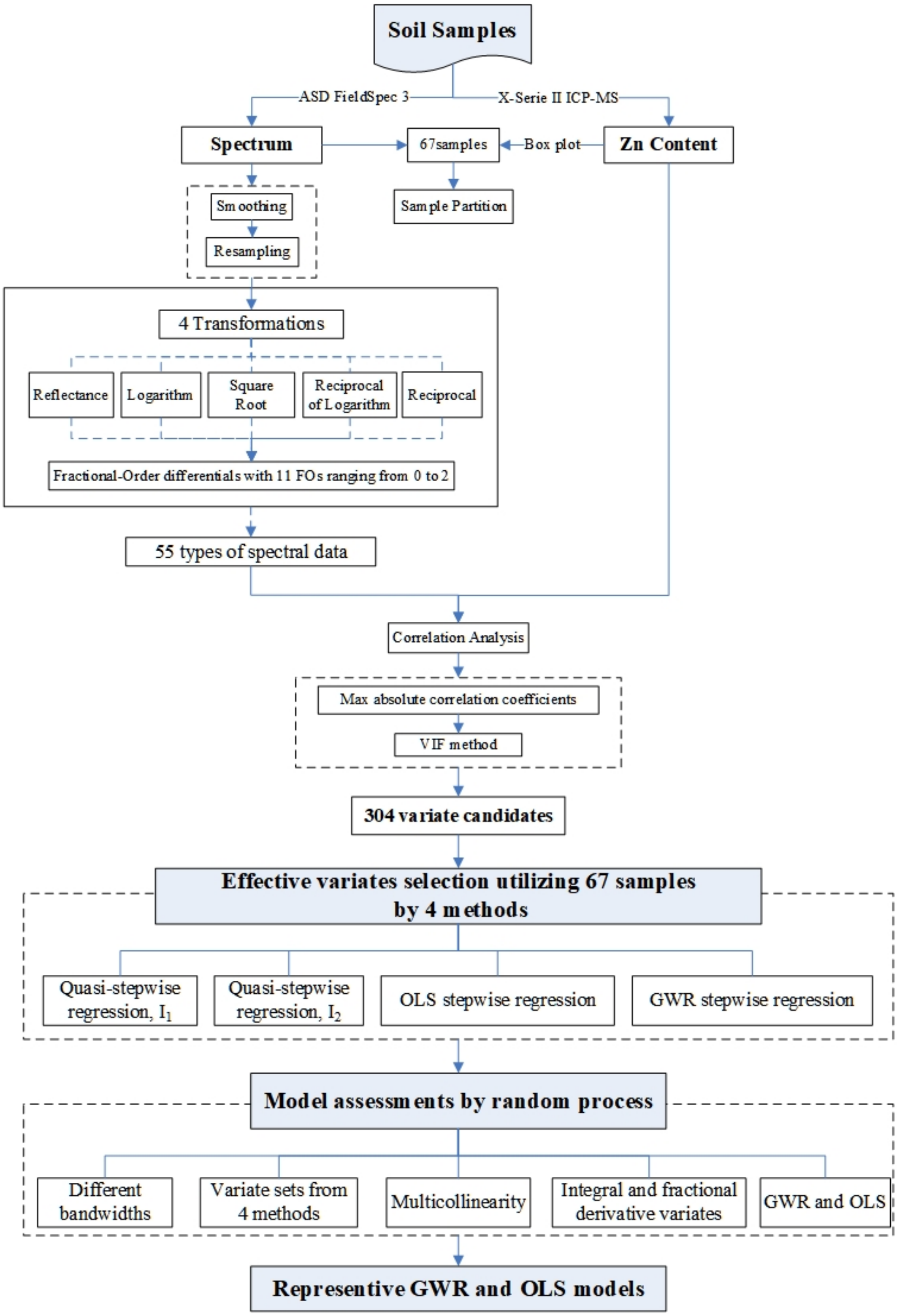

Soil Samples
ASD FieldSpec 3
X-Serie II ICP-MS
Spectrum
67samples
Box plot
Zn Content
Smoothing
Resampling
Sample Partition
4 Transformations
Reflectance
Logarithm
Square Root
Reciprocal of Logarithm
Reciprocal
Fractional-Order differentials with 11 FOs ranging from 0 to 2
55 types of spectral data
Correlation Analysis
Max absolute correlation coefficients
VIF method
304 variate candidates
Effective variates selection utilizing 67 samples by 4 methods
Quasi-stepwise regression, I1
Quasi-stepwise regression, I2
OLS stepwise regression
GWR stepwise regression
Model assessments by random process
Different bandwidths
Variate sets from 4 methods
Multicollinearity
Integral and fractional derivative variates
GWR and OLS
Representive GWR and OLS models

Figure 5. Flowchart showing the methods of this study.

2.5. Geographically Weighted Regression

In this study, the GWR [7] modeling for soil Zn content was conducted using samples from Pingtan Island. GWR is a spatial extension from the OLS model in Equation (3). The geographic location of each sample is considered in the GWR model, and the regression parameters could be estimated regionally. The extended model is expressed by Equation (4):

$$y_i = \beta_0 + \sum_{k=1}^{p} \beta_k x_{ik} + \varepsilon_i, \tag{3}$$

$$y_i = \beta_{i0}(u_i, v_i) + \sum_{k=1}^{p} \beta_{ik}(u_i, v_i) x_{ik} + \varepsilon_i, \tag{4}$$

where y_i is the dependent variable of the ith sample, which denotes the Zn content in this study; x_{ik} is the kth explanatory variable of the ith sample; p is the total number of explanatory variables; β_k is the coefficient of the kth explanatory variable; β_0 denotes the intercept; ε_i is the error term; and (u_i, v_i) represents the geographic coordinate of the ith sample. The estimated coefficients β_{ik} for each sample i vary depend on the associated location. In the calculation of the GWR coefficients β_{jk} (k = 0, 1, ..., p) for the jth sample, the coefficients can be obtained from the whole samples of size n so that element 5 is the lowest.

$$\sum_{i=1}^{n} w_{ji}[y_i - \beta_{j0}(u_j, v_j) - \sum_{k=1}^{p} \beta_{jk}(u_j, v_j) x_{ik}]^2, \tag{5}$$

where w_{ji} is the weight value of observation at location i for estimating the coefficients at location j, which is related to the distance between i and j. In this study, the Gaussian-distributed weight values with fixed bandwidth θ were employed:

$$w_{\mathrm{ij}} = \exp(-{d_{ij}}^2/\theta^2), \tag{6}$$

where d_{ij} denotes the distance between i and j. The set of regression coefficients at location i is estimated by weighted least squares. The matrix expression for the estimation is:

$$\hat{\boldsymbol{\beta}}_i = (\boldsymbol{X}^T \boldsymbol{W}_i \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{W}_i \boldsymbol{Y}, \tag{7}$$

where $\mathbf{X} = [\mathbf{X}_0, \mathbf{X}_1, ..., \mathbf{X}_p]$, $\mathbf{X}_k$ is a column vector of the kth explanatory variable, $\mathbf{X}_0$ is a column vector of ones for the intercept, $\mathbf{Y}$ is a column vector of the dependent variable, $\hat{\boldsymbol{\beta}}_i = (\beta_{i0}, \ldots, \beta_{ip})^T$ is the vector of local regression coefficients, and $\mathbf{W}_i$ is the diagonal matrix denoting the geographical weighting of each observation for the ith regression point. In this study, a distance D was used as a reference standard for bandwidth, which is given by:

$$\mathrm{D} = \sqrt{(u_{max} - u_{min})^2 + (v_{max} - v_{min})^2}, \tag{8}$$

where u_{max}, u_{min}, v_{max}, and v_{min} are the extrema of the geographic coordinates of the 67 samples, and D ≈ 31.5 km was obtained. To avoid large variance (too small bandwidth) and large bias (too large bandwidth) in the local estimates, the bandwidths were set to the range of D/4 to 3D/4 in this study.

While spatial analyses estimated the significance of regression coefficients or associated explanatory variables across space, the t-values of the regression coefficients provide information about the significance of the regression coefficients. The t-values of the estimated GWR regression coefficients were computed by approximating the variance. The covariance of the estimated regression coefficients at location i was approximated as [7] (p.55):

$$\mathrm{var}\left[\hat{\boldsymbol{\beta}}_i\right] = \boldsymbol{A}_i \boldsymbol{A}_i^T \hat{\sigma}^2, \tag{9}$$

where $\mathbf{A}_i$ is defined as

$$\boldsymbol{A}_i = (\boldsymbol{X}^T \boldsymbol{W}_i \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{W}_i, \tag{10}$$

and the estimated error variance, $\hat{\sigma}^2$, is given as

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\{n - [2trace(\boldsymbol{H}) - trace(\boldsymbol{H}^T\boldsymbol{H})]\}}, \tag{11}$$

where the i$^{\mathrm{th}}$ row of the hat matrix is defined as

$$\boldsymbol{H}_i = \boldsymbol{X}_i\boldsymbol{A}_i \tag{12}$$

with $\mathbf{X}_\mathrm{i}$ being the i$^{\mathrm{th}}$ row of **X**, and the trace of a matrix is the sum of the matrix diagonal elements. The t-values were computed by dividing each of the local estimates by the corresponding local standard error of the estimate. The t-values are useful in a purely exploratory sense as they can be used to detect the significance of the regression coefficients in a map [22].

3. Results and Discussion

3.1. Modeling Results of the 46 Samples

An OLS model and two GWR models were constructed from the 46 modeling samples shown in Figure 1a. The 21 verification samples (Figure 1b) were used to validate the results. Firstly, the bandwidth of D/4 was chosen for the two GWR models. Stepwise regression was used to select the explanatory variables for the OLS model. The t-value of a regression coefficient is an indicator to determine if a variable should be included or removed from an OLS model. A new variable with the largest t-value is added into the model, and the t-values of all selected variables have to satisfy $|t| > t_{\alpha = 0.05}$. The two obtained variables x_1 and x_2 for the OLS model (denoted by M_1) are reciprocal, FO = 0.8, 1560 nm; and square root, FO = 1.2, and 1140 nm, respectively, which are listed in Table 2. These two variables were also used to construct the GWR model (denoted by M_2). In addition to the OLS model, stepwise regression was employed to select the explanatory variables for GWR model (denoted by M_3) construction. For the GWR model, the stepwise regression method used was the same as that used for the OLS model except the indicator, for which the average $|t|$ of one explanatory variable was used because the t-values vary with location. In this study, the averages of $|t|$ of all the selected variables in the GWR stepwise regression are greater than 2. There are two variables, x_1 and x_2, obtained in this GWR stepwise regression model, which are logarithm, FO = 1, and 1140 nm; and reciprocal, FO = 1, and 2020 nm, respectively (Table 2).

Table 2. The modeling R^2 and adjusted R^2, and the verification R^2 and adjusted R^2 (denoted by $R_m{}^2$, $R_{adj,m}{}^2$, $R_v{}^2$ and $R_{adj,v}{}^2$, respectively), as well as the variates used to construct the models. Three bandwidths of D/4, D/2 and 3D/4 for GWR model (M_2) with the same variates as the OLS model (M_1) were examined. In M_3, the same bandwidths were investigated to build the GWR models using GWR stepwise regression.

	M_1 (OLS)	**M_2 (GWR)**			**M_3 (GWR)**		
		D/4	**D/2**	**3D/4**	**D/4**	**D/2**	**3D/4**
$R_m{}^2$	0.288	0.545	0.396	0.342	0.610	0.430	0.403
$R_{adj,m}{}^2$	0.255	0.524	0.368	0.311	0.591	0.389	0.360
$R_v{}^2$	−0.074	−0.186	−0.075	−0.067	−0.204	−0.181	−0.217
$R_{adj,v}{}^2$	−0.194	−0.318	−0.194	−0.186	−0.337	−0.390	−0.432
x_1	reciprocal, FO = 0.8, 1560 nm				logarithm, FO = 1, 1140 nm	reciprocal, FO = 0.8, 1560 nm	
x_2	sqrt, FO = 1.2, 1140 nm				reciprocal, FO = 1, 2020 nm	reciprocal, FO = 0.8, 1010 nm	
x_3						reciprocal, FO = 0.6, 1010 nm	

The modeling results of M_1, M_2, and M_3 (Figure 6) show the R^2 and $R_{adj}{}^2$ obtained from the 46 modeling samples (denoted by $R_m{}^2$ and $R_{adj,m}{}^2$ in Table 2, respectively) in the two GWR models are better than those in M_1, and those of M_3 are the largest. The predicted Zn contents for 21 verification samples were directly calculated from the regression coefficients of the OLS model and the associated variates. The regression coefficients in the GWR model for the 21 verification samples were interpolated from those of the 46 samples. The interpolated coefficients and the associated variates were used to calculate the predicted Zn content in the GWR model. The verification results of the 21 samples show that R^2 and $R_{adj}{}^2$ (denoted by $R_v{}^2$ and $R_{adj,v}{}^2$, respectively) of the two GWR and the OLS models are all negative. The $R_v{}^2$ and $R_{adj,v}{}^2$ of M_1 are largest whereas those of M_3 are smallest, which is inverse to the results predicted by the 46 samples.

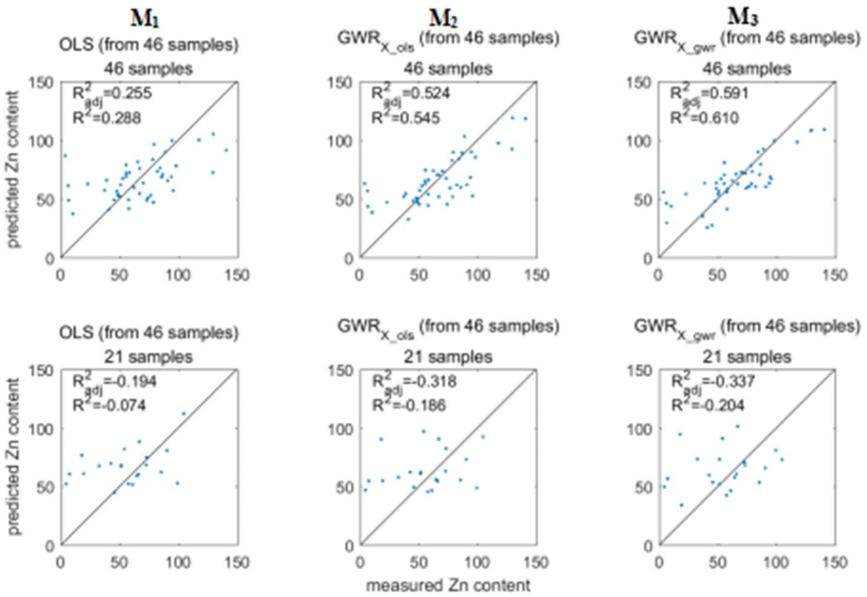


Figure 6. Modeling and verification results for 46 modeling and 21 verification samples for an OLS model (M_1) and two GWR models (M_2 and M_3). The left, middle, and right columns denote M_1, M_2, and M_3, respectively. The top and bottom rows represent the associated modeling and verification results, respectively. The associated R^2 and $R_{adj}{}^2$ are displayed on the top-left of each panel.

We examined the models with the two variates in M_2 for the two larger bandwidths of D/2 and 3D/4, as well as models built by the GWR stepwise regression (M_3) for changing the bandwidths to D/2 and 3D/4 to find the explanatory variables (Table 2). For the two bandwidths in M_3, the same three variates were selected. Although the $R_v{}^2$ and $R_{adj,v}{}^2$ in M_2 improved, these values were still negative using bandwidths D/2 and 3D/4 in M_2 and M_3. Thus, the results in Figure 6 and Table 2 might suggest that the 46 samples were not sufficiently large for building an appropriate model for the 67 samples. Although higher $R_m{}^2$ and $R_{adj,m}{}^2$ could be obtained for the 46 samples in M_2 and M_3, the associated $R_v{}^2$ and $R_{adj,v}{}^2$ for the 21 samples were negative. This may imply that the optimal variates suitable for the modeling samples are not applicable to the verification samples, or the interpolated coefficients of the GWR model are not applicable to them. To achieve more accurate predictions for new samples, $R_v{}^2$ and $R_{adj,v}{}^2$ are considered more important than the $R_m{}^2$ and $R_{adj,m}{}^2$ in assessing the GWR and OLS models.

3.2. Variate Selections from 67 Samples

For a GWR model, the distributions of the regression coefficients and the significance of the associated variates at different locations are often compared. Thus, the accuracy of the interpolated regression coefficients is important. As mentioned in Section 3.1., the $R_v{}^2$ and $R_{adj,v}{}^2$ for the OLS and GWR models were all negative, and the number of samples to construct appropriate OLS and GWR models should be bigger. To improve the accuracy of the interpolation of the regression coefficients and hence produce a better verification effect with the GWR model, four methods are proposed to construct the GWR models. In Method 1, a quasi-stepwise regression was employed to select independent variables. In order to emphasize $R_{adj,v}{}^2$, in which the number of independent variables is considered, an indicator I_1 calculated by $\bar{t}_{new} \times R_{adj,m}{}^2 \times R_{adj,v}{}^2$ was created for the quasi-stepwise regression, where $\bar{t}_{new}$ is the averaged absolute t-values of the 46 modeling samples of a new variate to be included, and $R_{adj,m}{}^2$ and $R_{adj,v}{}^2$ are the associated adjusted R^2 of the modeling and 21 verification samples as mentioned before, respectively. A new variate with the largest I_1 was included in the quasi-stepwise regression of Method 1. Since $R_{adj,v}{}^2$ was emphasized, the $\bar{t}_{new}$ of the newly added variate decreased. The standard method used to remove the variates in the GWR model involves setting the average |t| of variates in the model to greater than 1. The results of Method 1 are shown in Table 3 with different bandwidths of D/4, D/2, and 3D/4. Focus on I_1, $\bar{t}_{new}$, $R_{adj,m}{}^2$, and $R_{adj,v}{}^2$ rather than the variates. Considering the values of I_1, $R_{adj,m}{}^2$, and $R_{adj,v}{}^2$, the variates obtained at the third step of bandwidth of D/4 with the largest I_1 of 0.163 were selected to construct the GWR model in Method 1. The selected two variates were reciprocal, FO = 1, 1560 nm; and reciprocal, FO = 0.8, and 2310 nm. The $R_{adj,v}{}^2$ from the two variates was the largest compared with those from other variate sets listed in Table 3. The $R_{adj,m}{}^2$ increased with increasing number of variates. The $R_{adj,m}{}^2$ of the two selected variates is quite large compared with the values in bandwidths D/2 and 3D/4. Therefore, using I_1 as an indicator for a new variate seems to be reasonable in Method 1. The only shortcoming in Method 1 is that $\bar{t}_{new}$ is not the largest among the candidates. The lower mean value of |t| may lead to the insignificance of the regression coefficients in some areas. Table 5 shows that the minimum t-values of the two variates of Method 1 are insignificant. If the areas of insignificant regression coefficients overlap, the selected independent variables may have no effect on the predicted Zn content.

Table 3. Records of variates, I_1, $\bar{t}_{new}$, $R_{adj,m}{}^2$, and $R_{adj,v}{}^2$ in the quasi-stepwise regression of Method 1 with different bandwidths of D/4, D/2, and 3D/4.

Bandwidth = D/4				
Variates	I_1	$\bar{t}_{new}$	$R_{adj,m}{}^2$	$R_{adj,v}{}^2$
reciprocal, FO = 1.6, 1510 nm	0.0458	1.0826	0.2822	0.1498
reciprocal, FO = 1.6, 1510 nm*; reciprocal, FO = 1, 1560 nm	0.1421	1.8641	0.3916	0.1947
reciprocal, FO = 1, 1560 nm; reciprocal, FO = 0.8, 2310 nm	0.1634	1.743	0.4477	0.2094
reciprocal, FO = 1, 1560 nm; reciprocal, FO = 0.8, 2310 nm*; reciprocal, FO = 0.8, 2020 nm	0.1147	1.4718	0.4997	0.1559
reciprocal, FO = 1, 1560 nm; reciprocal, FO = 0.8, 2020 nm; reciprocal, FO = 1.2, 970 nm	0.0817	1.0291	0.5489	0.1446
Bandwidth = D/2				
Variates	I_1	$\bar{t}_{new}$	$R_{adj,m}{}^2$	$R_{adj,v}{}^2$
logarithm, FO = 1, 1560 nm	0.0236	2.6336	0.1794	0.0500
logarithm, FO = 1, 1560 nm; reciprocal, FO = 2, 2200 nm	0.0786	2.0766	0.2387	0.1587
logarithm, FO = 1, 1560 nm; reciprocal, FO = 2, 2200 nm; sqrt, FO = 0.6, 1380 nm;	0.0673	1.4093	0.2783	0.1717
logarithm, FO = 1, 1560 nm; reciprocal, FO = 2, 2200 nm; sqrt, FO = 0.6, 1380 nm; reciprocal, FO = 1, 2200 nm	0.1106	2.0686	0.3555	0.1504
logarithm, FO = 1, 1560 nm; reciprocal, FO = 2, 2200 nm; sqrt, FO = 0.6, 1380 nm; reciprocal, FO = 1, 2200 nm; logarithm, FO = 1, 2200 nm	0.0754	1.6947	0.3979	0.1119
logarithm, FO = 1, 1560 nm; reciprocal, FO = 2, 2200 nm; sqrt, FO = 0.6, 1380 nm*; reciprocal, FO = 1, 2200 nm; logarithm, FO = 1, 2200 nm; reciprocal of logarithm, FO = 0.2, 1930 nm	0.0659	1.0264	0.4062	0.1580
logarithm, FO = 1, 1560 nm; reciprocal, FO = 2, 2200 nm; reciprocal, FO = 1, 2200 nm; logarithm, FO = 1, 2200 nm; reciprocal of logarithm, FO = 0.2, 1930 nm; logarithm, FO = 1.6, 2170 nm	0.0885	1.1712	0.4224	0.1790
Bandwidth = 3D/4				
Variates	I_1	$\bar{t}_{new}$	$R_{adj,m}{}^2$	$R_{adj,v}{}^2$
reciprocal, FO = 0.8, 2220 nm	0.0160	2.6938	0.1473	0.0404
reciprocal, FO = 0.8, 2220 nm; reciprocal, FO = 1, 1560 nm	0.0525	2.1102	0.2116	0.1176
reciprocal, FO = 0.8, 2220 nm; reciprocal, FO = 1, 1560 nm; sqrt, FO = 0.6, 1380 nm	0.0399	1.2801	0.2269	0.1374
reciprocal, FO = 0.8, 2220 nm; reciprocal, FO = 1, 1560 nm; sqrt, FO = 0.6, 1380 nm; reciprocal, FO = 1.2, 2020 nm	0.0669	1.8941	0.2791	0.1265
reciprocal, FO = 0.8, 2220 nm; reciprocal, FO = 1, 1560 nm; sqrt, FO = 0.6, 1380 nm; reciprocal, FO = 1.2, 2020 nm; logarithm, FO = 1.2, 2020 nm	0.1016	1.8254	0.3257	0.1709
reciprocal, FO = 0.8, 2220 nm*; reciprocal, FO = 1, 1560 nm; sqrt, FO = 0.6, 1380 nm *; reciprocal, FO = 1.2, 2020 nm; logarithm, FO = 1.2, 2020 nm; reciprocal of logarithm, FO = 1.8, 2200 nm	0.0918	2.1128	0.3804	0.1143
reciprocal, FO = 1, 1560 nm; reciprocal, FO = 1.2, 2020 nm; logarithm, FO = 1.2, 2020 nm; reciprocal of logarithm, FO = 1.8, 2200 nm; logarithm, FO = 1.8, 700 nm	0.0827	1.4927	0.4124	0.1343

* Denotes variates with average |t| less than 1.

To emphasize the t-value, I_2 calculated by $\bar{t}_{new} \times t_{min} \times R_{adj,m}{}^2 \times R_{adj,v}{}^2$ was used as an indicator in the quasi-stepwise regression GWR models of Method 2, where t_{min} is the minimum of the absolute t-values in the 46 modeling samples of a new variate. Similarly, a new variate with the largest I_2 was included in the Method 2 model, and the standard that average $|t|$ of variates greater than 1 remained the same. Table 4 displays the results of Method 2 with different bandwidths of D/4, D/2, and 3D/4. Four variates, obtained in the fifth step of bandwidth 3D/4 with maximum I_2 of 0.267 among values at the three bandwidths, were determined as the explanatory variables for Method 2. The four explanatory variables are reciprocal, FO = 1, 1560 nm; reciprocal, FO = 1.2, and 2020 nm; reciprocal of logarithm, FO = 2, and 2200 nm; and logarithm, FO = 1.2, and 2020 nm. The associated $R_{adj,v}{}^2$ of 0.192 is the second largest, which is comparable to the 0.209 obtained in Method 1, and the $R_{adj,m}{}^2$ of 0.359 is acceptable. Table 5 shows that the mean $|t|$ of the four variates were all greater than 2 in the final model calibrated by the 46 samples in Method 2. When all the 67 samples were used, the associated $\bar{t}$ and t_{min} increased. The t_{min} of the four variates have smaller differences from the associated $\bar{t}$ compared with those in Method 1.

Table 4. A record of variates, I_2, $\bar{t}_{new}$, t_{min}, $R_{adj,m}{}^2$, and $R_{adj,v}{}^2$ in the quasi-stepwise regression of Method 2 with different bandwidths of D/4, D/2, and 3D/4.

Bandwidth = D/4					
Variates	I_2	$\bar{t}_{new}$	t_{min}	$R_{adj,m}{}^2$	$R_{adj,v}{}^2$
reciprocal, FO = 1, 1560 nm	0.0174	1.9250	0.6434	0.3303	0.0424
reciprocal, FO = 1, 1560 nm; reciprocal, FO = 2, 640 nm	0.0415	1.8741	0.3729	0.4261	0.1395
reciprocal, FO = 1, 1560 nm; reciprocal, FO = 2, 640 nm; reciprocal of logarithm, FO = 2, 2200 nm	0.0179	1.3117	0.2945	0.4441	0.1041
reciprocal, FO = 1, 1560 nm; reciprocal, FO = 2, 640 nm; reciprocal of logarithm, FO = 2, 2200 nm; logarithm, FO = 0.6, 2380 nm	0.0076	1.1330	0.1056	0.5363	0.1182
reciprocal, FO = 1, 1560 nm; reciprocal, FO = 2, 640 nm; reciprocal of logarithm, FO = 2, 2200 nm; logarithm, FO = 0.6, 2380 nm; reciprocal, FO = 1.8, 820 nm	0.0125	1.3054	0.0673	0.6454	0.2198
reciprocal, FO = 1, 1560 nm; reciprocal, FO = 2, 640 nm; reciprocal of logarithm, FO = 2, 2200 nm; logarithm, FO = 0.6, 2380 nm; reciprocal, FO = 1.8, 820 nm; reciprocal, FO = 1.2, 1970 nm	0.1091	1.6182	0.7625	0.6897	0.1281
Bandwidth = D/2					
Variates	I_2	$\bar{t}_{new}$	t_{min}	$R_{adj,m}{}^2$	$R_{adj,v}{}^2$
logarithm, FO = 1, 1560 nm	0.0505	2.6336	2.1384	0.1794	0.05
logarithm, FO = 1, 1560 nm; logarithm, FO = 2, 2200 nm;	0.1249	2.1617	1.6863	0.2399	0.1428
logarithm, FO = 1, 1560 nm; logarithm, FO = 2, 2200 nm; reflectance, FO = 1, 1560 nm	0.064	1.9158	1.5799	0.2943	0.0719
logarithm, FO = 1, 1560 nm; logarithm, FO = 2, 2200 nm; reflectance, FO = 1, 1560 nm; reciprocal, FO = 1.2, 2020 nm	0.0467	1.7515	1.2311	0.3474	0.0623
logarithm, FO = 1, 1560 nm; logarithm, FO = 2, 2200 nm; reflectance, FO = 1, 1560 nm; reciprocal, FO = 1.2, 2020 nm; logarithm, FO = 1.2, 2020 nm	0.0410	1.1916	0.8381	0.3508	0.1170
Bandwidth = 3D/4					
Variates	I_2	$\bar{t}_{new}$	t_{min}	$R_{adj,m}{}^2$	$R_{adj,v}{}^2$
reciprocal, FO = 0.8, 2220 nm	0.0405	2.6938	2.5271	0.1473	0.0404
reciprocal, FO = 0.8, 2220 nm; reciprocal, FO = 1, 1560 nm	0.1010	2.1102	1.9241	0.2116	0.1176
reciprocal, FO = 0.8, 2220 nm*; reciprocal, FO = 1, 1560 nm; reciprocal, FO = 1.2, 2020 nm	0.0578	1.8003	1.5415	0.2590	0.0804
reciprocal, FO = 1, 1560 nm; reciprocal, FO = 1.2, 2020 nm; reciprocal of logarithm, FO = 2, 2200 nm	0.1100	2.0887	1.9006	0.3049	0.0909
reciprocal, FO = 1, 1560 nm; reciprocal, FO = 1.2, 2020 nm; reciprocal of logarithm, FO = 2, 2200 nm; logarithm, FO = 1.2, 2020 nm	0.2670	2.1087	1.8366	0.3591	0.1920
reciprocal, FO = 1, 1560 nm; reciprocal, FO = 1.2, 2020 nm; reciprocal of logarithm, FO = 2, 2200 nm; logarithm, FO = 1.2, 2020 nm; logarithm, FO = 1.8, 680 nm	0.1119	2.1072	1.7310	0.4132	0.0742
reciprocal, FO = 1, 1560 nm; reciprocal, FO = 1.2, 2020 nm; reciprocal of logarithm, FO = 2, 2200 nm; logarithm, FO = 1.2, 2020 nm; logarithm, FO = 1.8, 680 nm; reciprocal of logarithm, FO = 0.4, 2140 nm	0.0141	1.1922	1.0272	0.4263	0.0269
reciprocal, FO = 1, 1560 nm; reciprocal, FO = 1.2, 2020 nm; reciprocal of logarithm, FO = 2, 2200 nm; logarithm, FO = 1.2, 2020 nm; logarithm, FO = 1.8, 680 nm; reciprocal of logarithm, FO = 0.4, 2140 nm; reciprocal, FO = 0.8, 1010 nm	0.0204	1.3752	1.2141	0.4493	0.0271

* Denotes the variates with average $|t|$ less than 1.

Table 5. The mean and minimum of |t| ($\bar{t}$ and t_{min}, respectively) of the regression coefficients of the explanatory variables in the GWR models built using Method 1 (D/4), Method 2 (3D/4), and Method 3 (3D/4), and the t-values of regression coefficients in the OLS model built in Method 3. For Methods 1 and 2, $\bar{t}$ and t_{min} in the GWR models calibrated by the 46 and 67 samples are both displayed. In Method 3, the values in a GWR model with bandwidth of 3D/4 are displayed. Here, $\bar{t}$, t_{min} and t-values are the values for the final models instead of the values standing for new variate.

Method 1 (46 samples)	**variates**	**reciprocal, FO = 1, 1560 nm;**	**reciprocal, FO = 0.8, 2310 nm**		
Method 1 (67 samples)	$\bar{t}$	1.657	1.743		
	t_{min}	0.119	0.001		
	$\bar{t}$	2.305	2.312		
	t_{min}	0.142	0.060		
Method 2 (46 samples)	**variates**	**reciprocal, FO = 1, 1560 nm**	**reciprocal, FO = 1.2, 2020 nm**	**reciprocal of logarithm, FO = 2, 2200 nm**	**logarithm, FO = 1.2, 2020 nm**
Method 2 (67 samples)	$\bar{t}$	2.458	2.396	2.607	2.109
	t_{min}	2.234	2.101	2.449	1.837
	$\bar{t}$	3.682	3.345	3.364	3.002
	t_{min}	3.507	2.927	3.212	2.589
Method 3 (67 samples)	**variates**	**reciprocal, FO = 0.8, 1560 nm**	**reciprocal, FO = 1.2, 2020 nm**	**logarithm, FO = 1.2, 2020 nm**	**reciprocal of logarithm, FO = 1.6, 2200 nm**
OLS (67 samples)	$\bar{t}$	3.479	4.166	3.889	2.467
	t_{min}	3.219	3.938	3.667	2.405
	t	−3.558	−4.272	−4.001	2.607

A more general method of constructing the GWR model was used in Method 3. In Method 3, stepwise regression for an OLS model, as described in Section 3.1. for M_1, was adopted for choosing the independent variables. Here, all 67 samples were used in the stepwise regression. The four selected variates in the OLS and for GWR models are shown in Table 5, which are reciprocal, FO = 0.8, 1560 nm; reciprocal, FO = 1.2, 2020 nm; logarithm, FO = 1.2, 2020 nm; reciprocal of logarithm, FO = 1.6, 2200 nm. In Method 4, all 67 samples were used to construct the GWR stepwise regression model using the method for M_3 in Section 3.1. In Method 4, five bandwidths of D/4, 3D/8, D/2, 5D/8, and 3D/4 were examined. The selected variates and modeling results for the five bandwidths are shown in Table 6, with only one variate selected for bandwidth D/4, and four variates chosen for bandwidths 3D/8, D/2, 5D/8, and 3D/4. The four variates of bandwidth 3D/8 are the same as those obtained by Method 2, and only FO in the second variate of bandwidths D/2, 5D/8, and 3D/4 was changed from 2 to 1.6 compared with bandwidth 3D/8. Compared with the variates obtained by the OLS stepwise regression in Method 3, only the FO in the first variate of the last 3 bandwidths in Table 6 was changed from 0.8 to 1.

Table 6. The obtained variates and the associated R_m^2 and $R_{adj,m}^2$ for the GWR models by applying the GWR stepwise regression in Method 4 with different bandwidths of D/4, 3D/8, D/2, 5D/8, and 3D/4.

Bandwidth	**Variates**	$\mathbf{R_m}^2$	$\mathbf{R_{adj,m}}^2$
D/4	reciprocal, FO = 0.8, 2020 nm	0.3673	0.3576
3D/8	reciprocal, FO = 1, 1560 nm; reciprocal of logarithm, FO = 2, 2200 nm; reciprocal, FO = 1.2, 2020 nm; logarithm, FO = 1.2, 2020 nm	0.4644	0.4299
D/2	reciprocal, FO = 1, 1560 nm; reciprocal of logarithm, FO = 1.6, 2200 nm; reciprocal, FO = 1.2, 2020 nm; logarithm, FO = 1.2, 2020 nm	0.4634	0.4288
5D/8	reciprocal, FO = 1, 1560 nm; reciprocal of logarithm, FO = 1.6, 2200 nm; reciprocal, FO = 1.2, 2020 nm; logarithm, FO = 1.2, 2020 nm	0.4572	0.4221
3D/4	reciprocal, FO = 1, 1560 nm; reciprocal of logarithm, FO = 1.6, 2200 nm; reciprocal, FO = 1.2, 2020 nm; logarithm, FO = 1.2, 2020 nm	0.4542	0.4190

Notably, two variates of reciprocal, FO = 1.2, 2020 nm; and logarithm, FO = 1.2, 2020 nm are present in the models built by Methods 2, 3, and 4. By comparing the R_m^2 and $R_{adj,m}^2$ listed in Table 6, the values of bandwidth D/4 are smallest, and those of bandwidth 3D/8 are largest. However, to assess a GWR model, the verification results have to be considered. The assessment of GWR models in Methods 1–4 and the OLS model in Method 3 are presented in Section 3.3. Larger $\bar{t}$ and t_{min} of the regression coefficients in the Method 3 GWR model using bandwidth 3D/4 and the OLS model in Method 3 are shown in Table 5. The t-values in the GWR model using Method 4 with bandwidth 3D/4 are shown in Figure 8.

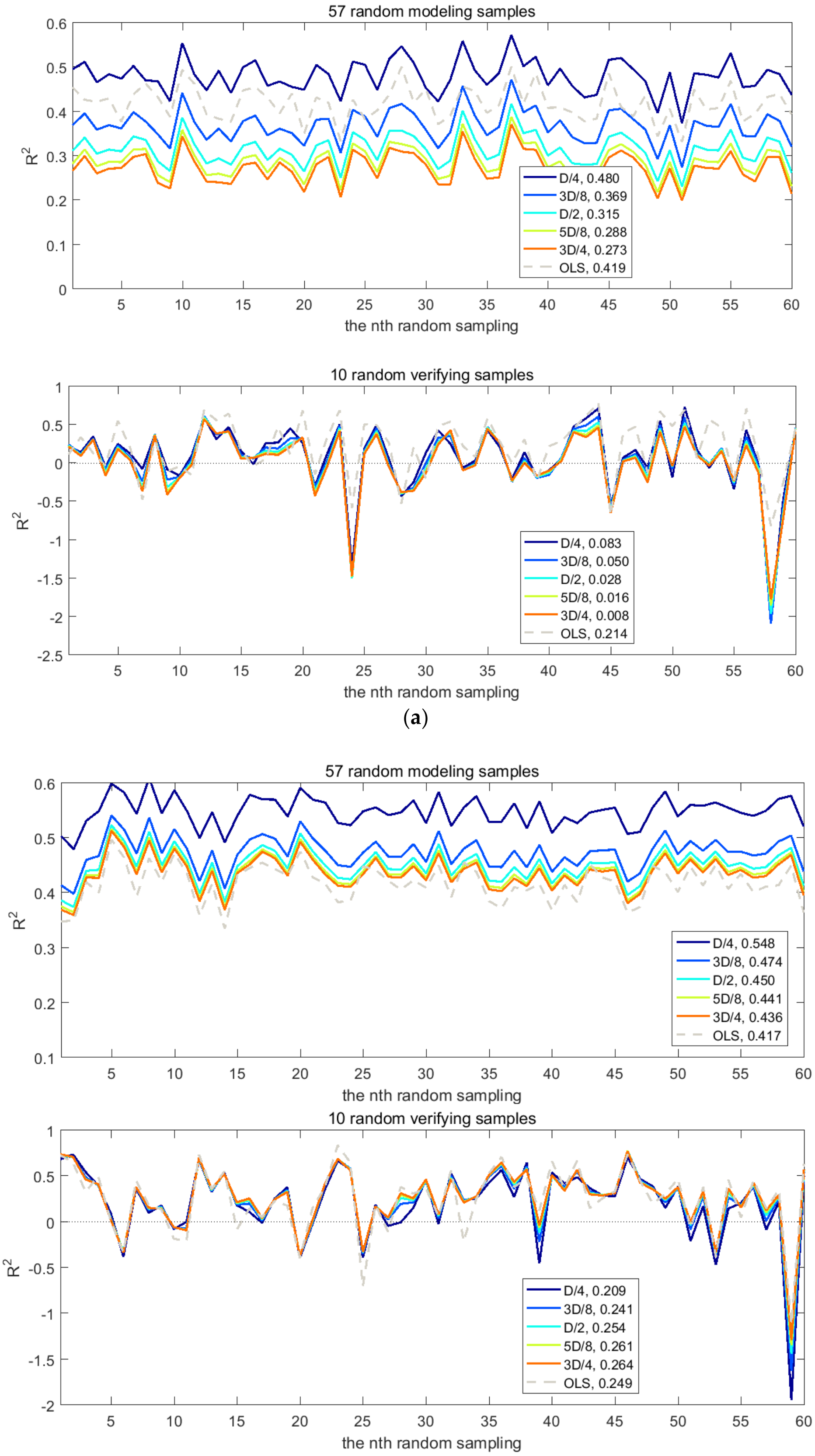
3.3. Model Assessments

In order to assess the GWR models built using Methods 1–4 and the OLS model, we conducted a random sampling process to select 57 samples for modeling and 10 samples for validation. The selection of 57 samples for modeling is comparable with the total sample size of 67. Despite the fact that the difference in modeling samples and total sample size can yield differences in retrieved coefficients in both GWR and OLS models, we examined the averaged effects of accuracy of the interpolation in the regression coefficients of GWR model and the error of the predicted Zn content in both models for 10 verification samples. Figure 7 displays the modeling and verification results of 60 repeated random processes. The OLS model in Figure 7 was built via stepwise regression from 67 samples using Method 3. In Figure 7a, the GWR modeling and its associated verification effects of the two variates in Method 1 are displayed, with bandwidths being changed from D/4 to 3D/4. The bandwidth D/4, which was used to find the two variates in Method 1, has the best effects on both modeling and verification when two variates were used. As bandwidth increases, the averaged values of modeling R_m^2 and verification R_v^2 decrease. The mean modeling result of D/4 is better than that of the OLS model; however, the mean verification effects in these models are much worse than that of the OLS model.

Table 7. The obtained variates and the associated R_m^2 and $R_{adj,m}^2$ for the GWR models obtained by applying GWR stepwise regression in Method 4 with different bandwidths of D/4, 3D/8, D/2, 5D/8, and 3D/4. Only IO derivatives are the candidates in the GWR stepwise regressions.

Bandwidth	Variates	R_m^2	$R_{adj,m}^2$
D/4	reciprocal, FO = 1, 2020 nm	0.337	0.326
3D/8	reciprocal, FO = 1, 2020 nm; reciprocal, FO = 1, 1560 nm; reciprocal of logarithm, FO = 2, 2200 nm; logarithm, FO = 1, 2020 nm	0.457	0.422
D/2	reciprocal, FO = 1, 2020 nm; reciprocal, FO = 1, 1560 nm; reciprocal of logarithm, FO = 2, 2200 nm; logarithm, FO = 1, 2020 nm; reciprocal of logarithm, FO = 0, 2200 nm; sqrt, FO = 2, 640 nm	0.518	0.469
5D/8	reciprocal, FO = 1, 2020 nm; reciprocal, FO = 1, 1560 nm; reciprocal of logarithm, FO = 2, 2200 nm; logarithm, FO = 1, 2020 nm; reciprocal of logarithm, FO = 0, 2200 nm; sqrt, FO = 2, 640 nm	0.506	0.456
3D/4	reciprocal, FO = 2, 2200 nm; reciprocal, FO = 1, 1560 nm;	0.294	0.272

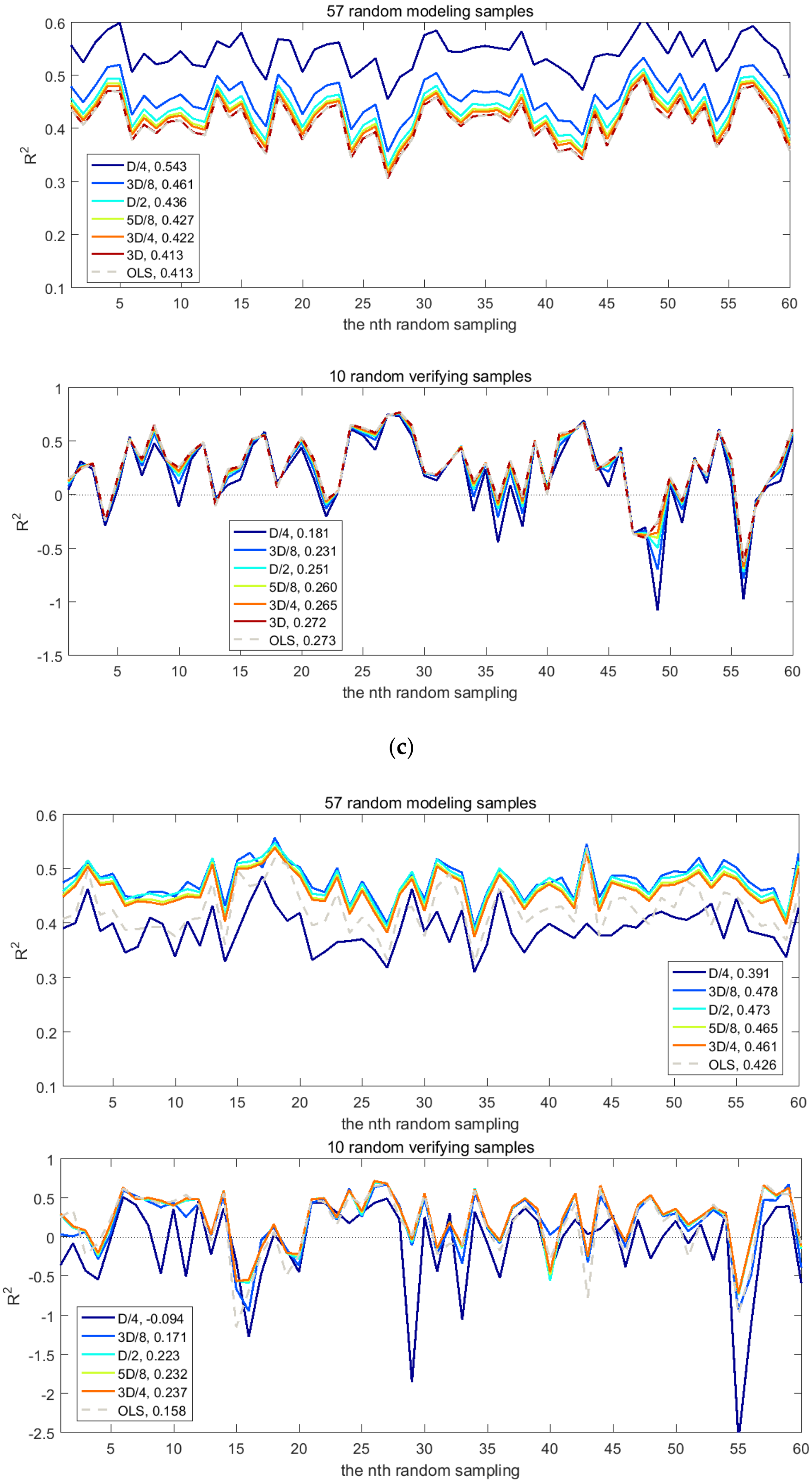
In Method 2, four variates were selected for modeling in which bandwidth 3D/4 was used, as shown in Table 4. The effects of bandwidth on the four variates were investigated by changing the bandwidth from D/4 to 3D/4 (Figure 7b). The smaller the bandwidth, the larger the mean R_m^2, and the OLS model has the smallest mean R_m^2 in the 60 repetitions of the random process. In contrast to the 10 random verification samples, the mean R_v^2 increases with increasing bandwidth. The 3D/4 used in finding the four variates in Method 2 produced the best verification effect, in which the associated mean R_v^2 is greater than that of the OLS model. R_v^2 is considered to be more important in assessing the variation of the regression coefficients, and a larger bandwidth should be adopted for the four variates.



(**a**)

(**b**)

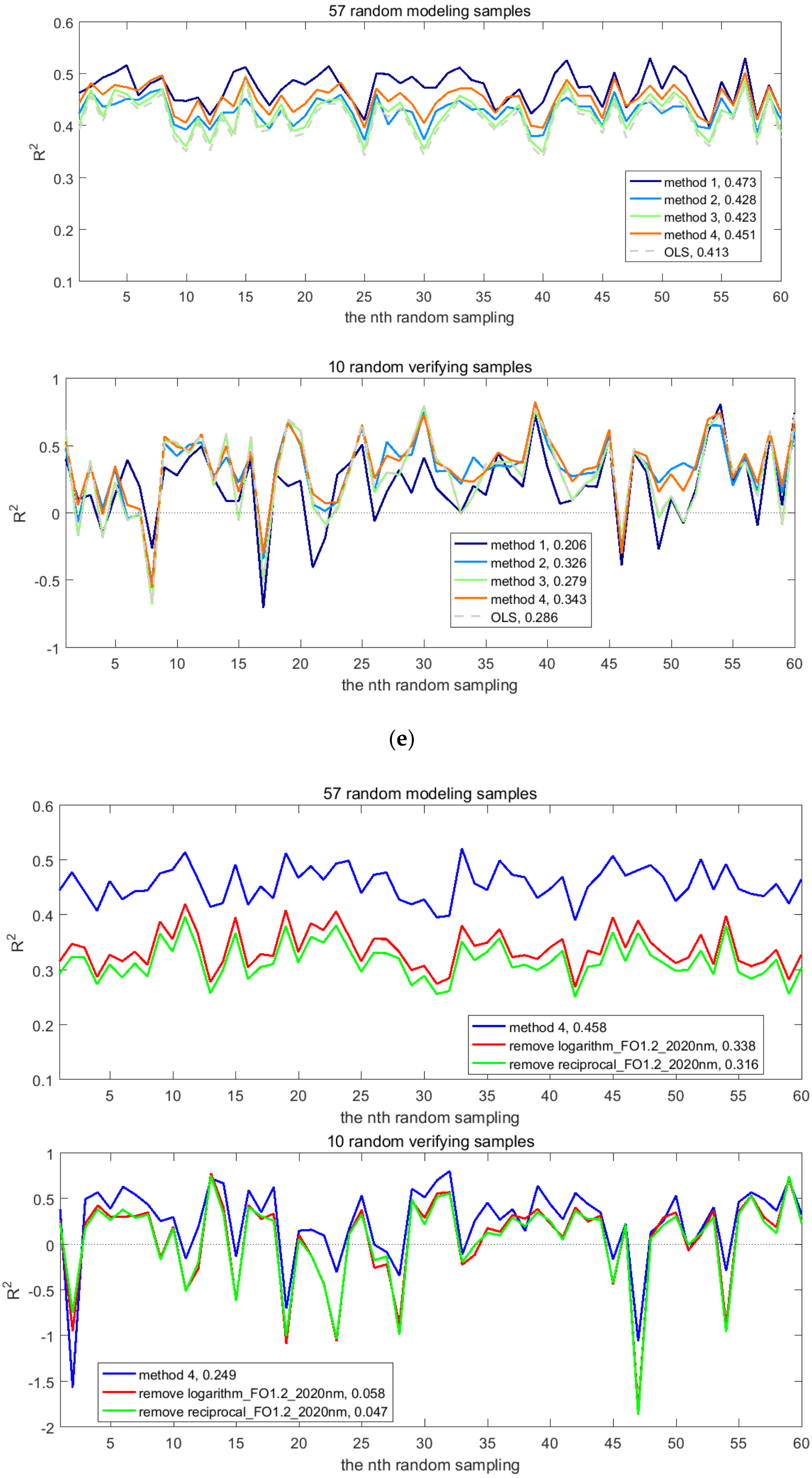
Figure 7. *Cont.*



(**c**)

(**d**)

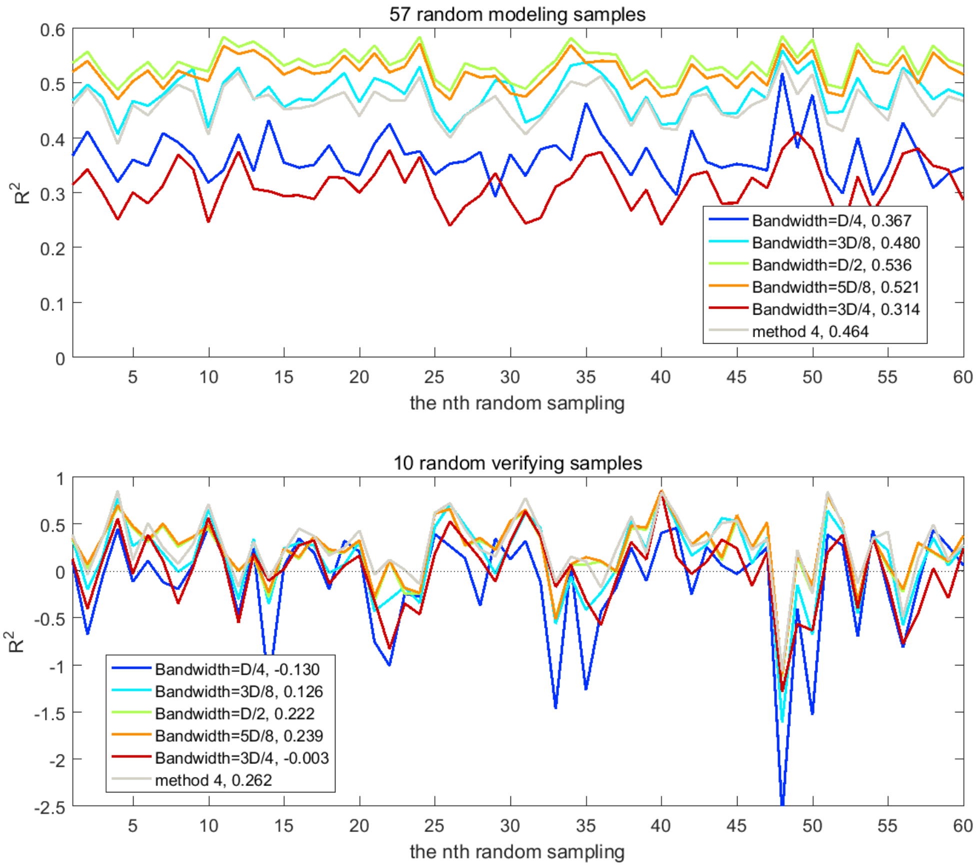
Figure 7. *Cont.*



(**e**)

(**f**)

Figure 7. *Cont.*



(**g**)

Figure 7. The results of the 60-repetition random processes by repeatedly selecting 57 samples for modeling and 10 samples for validation. The top and bottom panels display the 60 repetitions of the modeling results from the 57 randomly-picked samples and verification results from the left 10 samples, respectively. Values in the legend represent the associated mean R^2 of the 60-repetition random process. (**a**) The GWR models with two variates (D/4) in Method 1 and with bandwidths changing from D/4 to 3D/4; (**b**) the GWR models with four variates (3D/4) in Method 2 and with bandwidths changing from D/4 to 3D/4; (**c**) the GWR models with four variates obtained by OLS stepwise regression and with bandwidths changing from D/4 to 3D; (**d**) the GWR models with variates obtained by GWR stepwise regression for different bandwidths from D/4 to 3D/4 as shown in Table 6; (**e**) the GWR models with the superior mean R_v^2 in Methods 1–4; (**f**) the GWR model with the superior mean R_v^2 in Method 4, and those with one of the two variates x_3 and x_4 removed; and (**g**) the GWR models with variates of IO derivatives obtained by GWR stepwise regression for different bandwidths from D/4 to 3D/4 as shown in Table 7. In (**a**)–(**e**), the OLS models in Method 3 are compared with the GWR models, whereas in (**f**) and (**g**), the GWR models with the superior mean R_v^2 in Method 4 are compared to the other GWR models.

The OLS model and GWR models of different bandwidths built from the 57 random modeling samples with the four selected variates (shown in Table 5) using Method 3 are shown in Figure 7c. Similar to Figure 7b, tendencies in the mean values of R_m^2 and R_v^2 changing with bandwidths can be found in Figure 7c, in which the mean R_m^2 increases with decreasing bandwidth. Conversely, R_v^2 increases with increasing bandwidth. In comparison, the modeling and verification effects of bandwidth 3D approached those of the OLS model, where the mean R_m^2 is the lowest and the mean R_v^2 is the highest, and the mean of R_m^2 and R_v^2 converge toward the OLS values as the bandwidth increases for the four variates. This implies that with increasing bandwidth, local estimates approach global estimates. This bandwidth effect can also be seen in Figure 7a,b. Based on the mean $R_v{}^2$, the OLS and GWR models with larger bandwidth are more favorable for the four variates.

In Method 4, one variate was selected for bandwidth D/4, and four variates were chosen for bandwidths 3D/8, D/2, 5D/8 and 3D/4, as shown in Table 6. Note that the set of four variates for bandwidth 3D/8 is the same as that obtained with Method 2, and only FO in its second variate in Table 6 is different from that for bandwidths D/2, 5D/8, and 3D/4, which changes from FO = 2 to FO = 1.6. The 60-repetition random process shows that both the modeling and verification effects of bandwidth D/4 are the worst among all models in Figure 7d. The variations in the $R_m{}^2$ curves are similar for bandwidths D/2, 5D/8, and 3D/4, and there is only a small difference between the curves of bandwidth 3D/8 and the other three bandwidths. Comparing the modeling effects of the 57 random samples for the four bandwidths, the mean $R_m{}^2$ of 3D/8 is the largest, and that of 3D/4 is the smallest. The variations in $R_v{}^2$ are also similar for bandwidths D/2, 5D/8, and 3D/4, while the mean $R_v{}^2$ of 3D/4 is largest, and that of 3D/8 is the smallest. The model constructed using bandwidth 3D/4 with the largest $R_v{}^2$ among the five GWR stepwise regression models (Figure 7d) also has better modeling and verification effects than the OLS model.

After analyzing the models using Methods 1–4, the models with superior mean $R_v{}^2$ values were compared with the OLS model in which there are two variates in Method 1 with bandwidth D/4 and four variates in Methods 2–4 with bandwidth 3D/4 (Figure 7e). Method 4 has the best mean verification effect, and its mean modeling effect is only lower than that of Method 1 in the 60-repetition random process. The mean $R_v{}^2$ of Method 2 is also better among the models in Figure 7e, where one of the variates (reciprocal of logarithm, 2200 nm) has a discrepancy in FO = 2 from that in Method 4 (FO = 1.6). The model in Method 4 with the best mean $R_v{}^2$ from the random process was chosen as the designated representative model for soil Zn content estimation on Pingtan Island.

The VIFs set for four variates shown in the last row of Table 6 in Method 4 (bandwidth 3D/4), which are defined in sequence as x_1, x_2, x_3, and x_4, shows that the VIF values are 1.023, 1.502, 52.990, and 56.334, respectively. This suggests that there is high collinearity between variates x_3 and x_4. The collinearity effects on modeling and verification were investigated by the 60-repetiton random process (Figure 7f). It is not surprising that the mean $R_m{}^2$ obtained by using all four variates in Method 4 is greater than those obtained with one of x_3 and x_4 removed. The verification results show that, for most models established from the 57 random samples, using any of the three variates is apparently worse than using all four variates. From this investigation, all four variates should be adopted together because the 10 random samples were more accurately predicted.

In this study, we investigated the effects of applying the fractional-order differential operation to spectral data in building GWR models. The effectiveness of the differential operation of the FO relative to the operation of IO was analyzed using a 60-repetition random process (Figure 7). Table 7 displays the results of the GWR stepwise regression for candidates of the IO derivatives using M_3 in Section 3.1. with bandwidths D/4, 3D/8, D/2, 5D/8, and 3D/4. There are one, four, six, six, and two variates selected for the five bandwidths, respectively, as shown in Table 7. In Figure 7g, the modeling results of the 57 random samples show that the models with the same six variates of the IO derivatives (D/2 and 5D/8) have better modeling effects, in which bandwidth D/2 is the best. For the 10 random verification samples, the four variates in the representative GWR model with Method 4 have the best mean verification effect, whereas for the effect in the IO derivative models, the bandwidth 5D/8 is the best. Table 8 shows that for the IO derivative model with bandwidth 5D/8, the mean $R_v{}^2$ of the 10,000-repetition random process of 0.220 is large compared with those in Methods 1–3, and the representative GWR model of Method 4 improves the $R_v{}^2$ by 10% (mean $R_v{}^2$ = 0.242). The mean $R_v{}^2$ of the 10,000-repetition random process for the model constructed by the same six variates of IO derivatives with bandwidth 3D/4 was also computed. The $R_v{}^2$ value of 0.228, which is still less than that of the representative GWR model, suggests that the FO differential operation could slightly increase the accuracy of modeling soil Zn content on Pingtan Island with fewer explanatory variables required when the GWR stepwise regression is used. This might have resulted from the similarities in the four variates between the models, and that the FOs of some of the variates are close to or equal to IOs.

Table 8. The mean of R_m^2, $R_{adj,m}^2$, R_v^2, and $R_{adj,v}^2$ of the 10,000-repetition random process by repeatedly selecting 57 samples for modeling and 10 samples for validation.

	Method 1 (D/4)	**Method 2 (3D/4)**	**Method 3 (3D/4)**	**Method 4 (3D/4)**	**OLS**	**IO (5D/8)**	**IO (3D/4)**	**OLS-2**	**OLS-3**
Mean R_m^2	0.480	0.438	0.432	0.459	0.423	0.519	0.510	0.452	0.488
Mean $R_{adj,m}^2$	0.461	0.395	0.389	0.418	0.378	0.461	0.451	0.410	0.427
Mean R_v^2	0.082	0.205	0.163	0.242	0.175	0.220	0.228	0.253	0.244
Mean $R_{adj,v}^2$	−0.180	−0.431	−0.508	−0.364	−0.486	−1.340	−1.316	−0.346	−1.270

The results shown in Figure 7c, in which the four variates used were obtained via OLS stepwise regression, display a convergence effect as the bandwidth increases. It is expected that the OLS models established using the four variates (bandwidth 3D/4) in Method 4 and six variates (bandwidth 5D/8) of IO derivatives may have better average verification results than the associated GWR models. Table 8 depicts the associated coefficients of determination of the two OLS models, denoted as OLS-2 and OLS-3, respectively. For the four variates in Method 4, the R_v^2 of the OLS-2 is 0.253, surpassing the value of 0.242 of the GWR model. The R_v^2 of the OLS-3 is 0.244, improving the IO derivatives GWR model (5D/8), and is also greater than that of the model with Method 4. Here, the OLS-2 with the highest R_v^2 is also designated to be the representative model for the OLS model. Nevertheless, a GWR model describes the possible spatial variations in the parameter estimates, and the mean verification effect of OLS-2 in R_v^2 is 4.5% higher than that of the representative GWR model. In the following section, we outlined our analyses of parameter estimates for both representative GWR and OLS models.

3.4. Analyses of Parameter Estimates

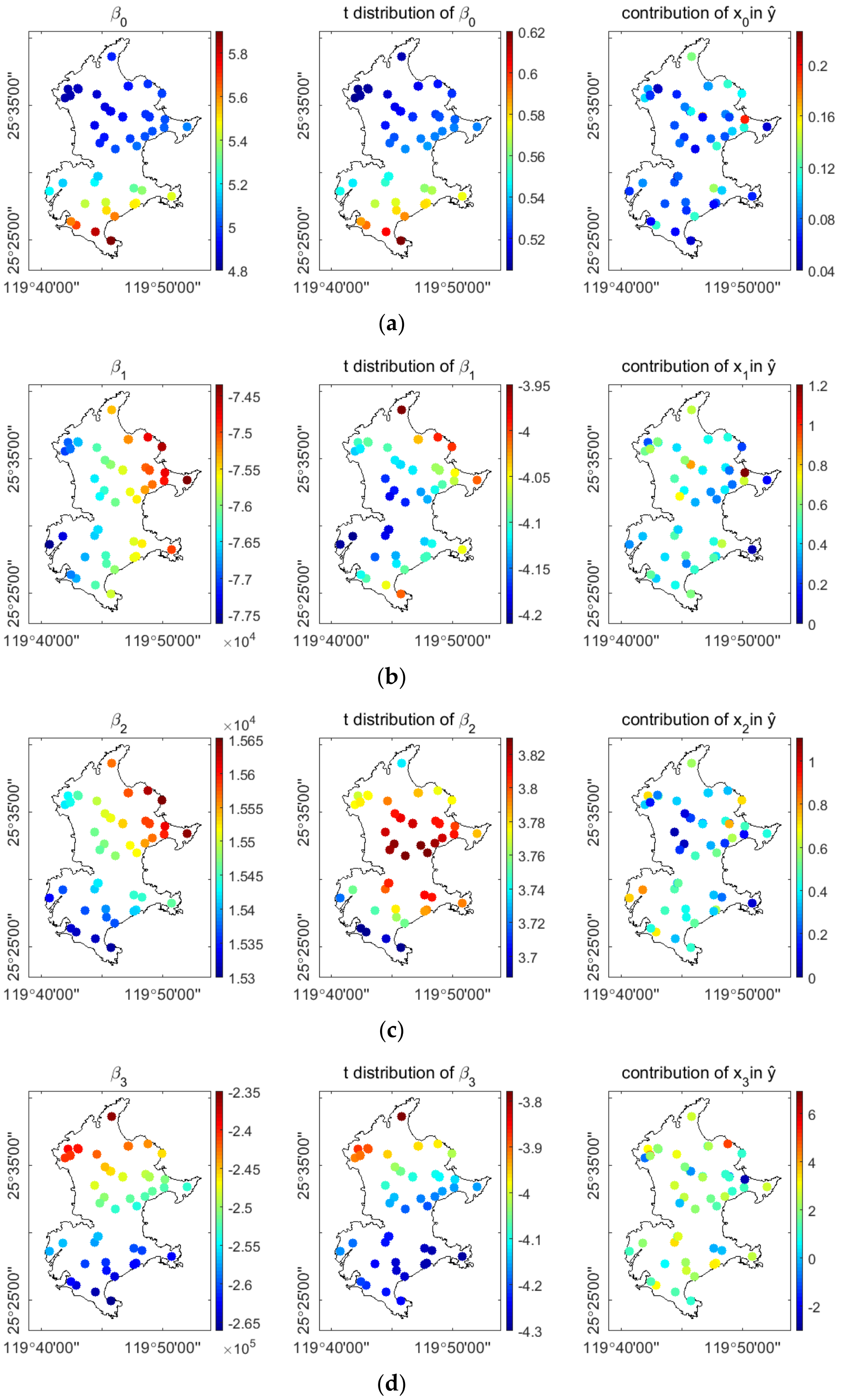
From Section 3.3., an OLS model and a GWR model (bandwidth 3D/4) with four variates using Method 4 were chosen to be the representative models for soil Zn content estimation on Pingtan Island, as they produced superior mean verification results compared to the other studied models. To compare the performance of the two models, the parameter estimates, standard errors (SE), and t-values for the OLS model are outlined in Table 9. Figure 8 displays the spatial distributions of the parameter estimates, associated t-values, and contributions of C_{ik} of each variate in the predicted Zn content $\hat{y}$ for the GWR model. The contribution C_{ik} is:

$$C_{ik} = \frac{\beta_{ik}(u_i, v_i)x_{ik}}{\hat{y}_i}, \tag{13}$$

where i = 1, 2, ..., 67 represents each of the 67 samples; k = 0, 1, ..., 4 represents the intercept and the four explanatory variables of the two models in this study; and $x_{i0} = 1$ for all i for the intercept. Note that the kth estimate β_{ik} is a single value for all i of the OLS model, and it varies spatially for the GWR model. As a whole, the five parameter estimates and the associated t-values in the OLS model are comparable to those in the GWR model, as shown in Figure 8.

Table 9. The parameter estimates (β), standard errors (SE), and t-values for the representative OLS model.

Parameter	β_0	β_1	β_2	β_3	β_4
β	5.922	-7.460×10^4	1.551×10^4	-2.515×10^5	-1.018×10^6
SE	9.264	1.799×10^4	3.989×10^3	5.961×10^4	2.604×10^5
t-value	0.639	−4.148	3.888	−4.219	−3.909


β_0
t distribution of β_0
contribution of x_0 in ŷ
β_1
t distribution of β_1
contribution of x_1 in ŷ
β_2
t distribution of β_2
contribution of x_2 in ŷ
β_3
t distribution of β_3
contribution of x_3 in ŷ
25°35'00"
25°25'00"
119°40'00"
119°50'00"
5.8 5.6 5.4 5.2 5 4.8
0.62 0.6 0.58 0.56 0.54 0.52
0.2 0.16 0.12 0.08 0.04
-7.45 -7.5 -7.55 -7.6 -7.65 -7.7 -7.75 $\times 10^4$
-3.95 -4 -4.05 -4.1 -4.15 -4.2
1.2 1 0.8 0.6 0.4 0.2 0
$\times 10^4$ 1.565 1.56 1.555 1.55 1.545 1.54 1.535 1.53
3.82 3.8 3.78 3.76 3.74 3.72 3.7
1 0.8 0.6 0.4 0.2 0
-2.35 -2.4 -2.45 -2.5 -2.55 -2.6 -2.65 $\times 10^5$
-3.8 -3.9 -4 -4.1 -4.2 -4.3
6 4 2 0 -2

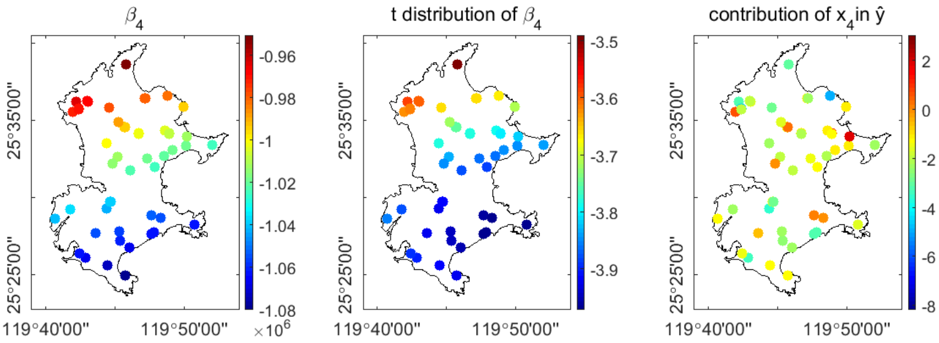
(**a**)

(**b**)

(**c**)

(**d**)

Figure 8. *Cont.*



(**e**)

Figure 8. Spatial distributions of parameter estimates for β$_0$, β$_1$, β$_2$, β$_3$, and β$_4$; the associated t-values; and the contributions of C$_{ik}$ in predicted Zn content for intercept and the four variates of the representative GWR model. The left, middle, and right columns denote the parameter estimates, t-values, and the contributions C$_{ik}$, respectively. (**a**) The intercept and (**b**)–(**e**) for the four variates x$_1$, x$_2$, x$_3$, and x$_4$, respectively.

The t-values show that β$_0$ is the most insignificant in both models. Comparing the ranges of β$_0$, β$_1$, β$_2$, β$_3$, and β$_4$ in Figure 8 and the associated SE in the OLS model (Table 9), the spatial non-stationarity of the five parameter estimates is insignificant [11,12]. In β$_0$, the parameter estimates present a tendency where the larger values are found in the south, and the smaller values are found in the north. The values of β$_0$ fall approximately in a range of 4.8 to 6.0 mg/kg, the contributions of x$_0$ are positive, and most are less than 0.15. The magnitudes of parameter estimates may not represent the same as their significance. In β$_1$, the parameter estimates show that the negative value in the east is greater than that in the west. Detailed comparisons of the significance of β$_1$ showed that the central west has the largest significance, and the significance level decreases gradually from the center. For β$_2$, the values are positive and higher in the northeast and lower in the southwest. The most significant area of β$_2$ is around the central-north section. The contributions of x$_1$ and x$_2$ are less than 0.75. For the negative values of β$_3$ and β$_4$, the spatial variations in their values and the associated t-values between the two estimates are similar. This may be due to the high collinearity between the two variates.

In the parameter estimates, the negative degree is larger in the south and smaller in the north. As x$_3$ and x$_4$ are different only in the reciprocal and logarithm transformations, and both β$_3$ and β$_4$ are negative, there are opposite polarities in the contributions of x$_3$ and x$_4$ in the predicted Zn content. Many samples have contributions greater than 1 or less than –1 in x$_3$ and x$_4$, and more than 73% range from 0 to 3 and from –3 to 0, respectively. Figure 7f shows that the model with four variates is better than those with one of x$_3$ or x$_4$ removed during verification. Figure 9 displays the spatial distributions of C$_{ik}$ for the intercept and the four variates in the OLS model. The trends in C$_{ik}$ of the five terms in both models are not as clear as in the parameter estimates and t-values, and the five distributions of C$_{ik}$ in the OLS model agree with those in the GWR model in general.

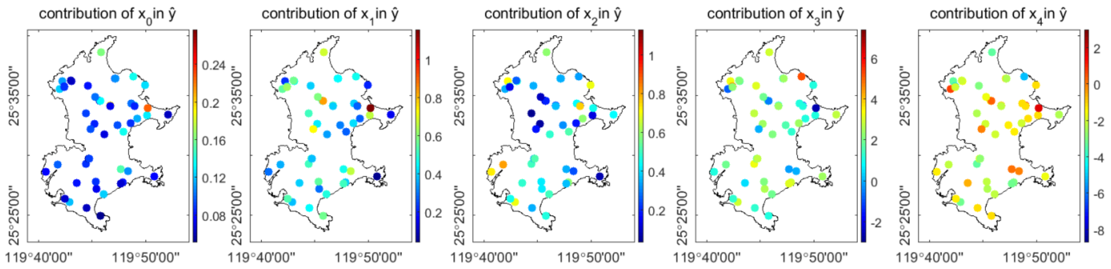

contribution of $x_0$in ŷ
contribution of $x_1$in ŷ
contribution of $x_2$in ŷ
contribution of $x_3$in ŷ
contribution of $x_4$in ŷ
0.24
0.2
0.16
0.12
0.08
1
0.8
0.6
0.4
0.2
6
4
2
0
-2
2
0
-2
-4
-6
-8
25°35'00"
25°25'00"
119°40'00"
119°50'00"

Figure 9. Spatial distributions of contributions C$_{ik}$ in predicted Zn content for intercept and the four variates in the representative OLS model. The columns from left to right denote the contributions C$_{ik}$ of the intercept and the four variates x$_1$, x$_2$, x$_3$, and x$_4$, respectively.

3.5. Performance of the Representative Models

The performance of the representative GWR and OLS models constructed by the 67 samples is shown in Figure 10. The R^2 of both models are approximately 0.45, which could be further improved by finding locally effective variates or applying chemical analyses of soil constituents and investigating the hyperspectral retrieval mechanisms of Zn content. To investigate the model performance at different Zn concentration levels, soil samples with Zn content <25 mg/kg were separated from those with Zn content >25 mg/kg. The $R_{adj}{}^2$, R^2, RMSE, and average relative error δ, which is defined as mean$\left(|\hat{y}_i - y_i|/y_i\right)$, for Zn content below 25 mg/kg of the representative GWR model are –73.868, –36.434, 40.977, and 537.2%, and those for Zn content above 25 mg/kg are 0.396, 0.438, 17.773, and 19.7%, respectively (Table 10). For Zn contents below 25 mg/kg, the RMSE is greater than the value 22.338 of the 67 samples, and δ is much greater than the value 89.2% of the 67 samples. For Zn contents above 25 mg/kg, the RMSE decreased by 20.4%, and δ decreased drastically by 77.9% relative to the value of the 67 samples. These results suggest that the nine samples with Zn content below 25 mg/kg have larger error and lower prediction accuracy in the representative GWR model. Similar results were found in the representative OLS model. The $R_{adj}{}^2$ and R^2 in the representative OLS model for Zn contents above 25 mg/kg are greater than and the associated RMSE and δ are less than the representative GWR model in this study.

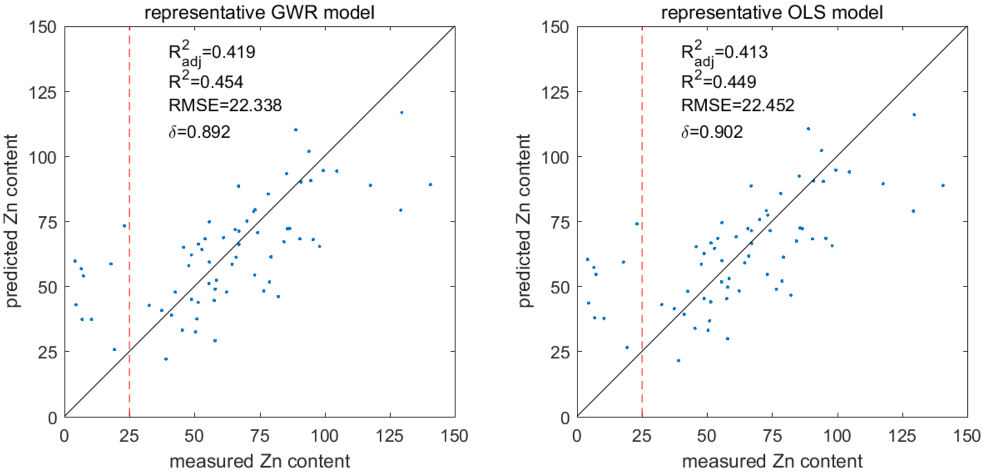


Figure 10. Scatter plots of the representative GWR (left panel) and OLS (right panel) models constructed using the 67 samples. The $R_{adj}{}^2$, R^2, RMSE, and average relative error δ calculated using the 67 samples are shown in the top of two panels. The red dashed vertical line indicates the measured Zn content of 25 mg/kg.

Table 10. The performance of the representative GWR and OLS models for the 67 samples: 9 samples with Zn content <25 mg/kg and 58 samples with Zn content >25 mg/kg.

	Representative GWR model			**Representative OLS model**		
	67 samples	**9 samples (<25 mg/kg)**	**58 samples (>25 mg/kg)**	**67 samples**	**9 samples (<25 mg/kg)**	**58 samples (>25 mg/kg)**
$R_{adj}{}^2$	0.419	–73.868	0.396	0.413	–76.057	0.399
R^2	0.454	–36.434	0.438	0.449	–37.529	0.441
RMSE	22.338	40.977	17.773	22.452	41.571	17.725
δ	89.2%	537.2%	19.7%	90.2%	545.2%	19.6%

3.6. Summary

In this study, four transformations (square root, logarithm, reciprocal of logarithm, and reciprocal) and 11 differential operations were applied to the preprocessed hyperspectral reflectance data of soil samples to increase the number of candidates for OLS and GWR modeling. By finding the wavelength regions in the data of a certain transformation and of a certain differential operation with local maxima in the absolute values of correlation coefficient, and by employing the VIF, 304 candidates were identified. Firstly, 46 modeling samples were used to construct one OLS (M_1) and two GWR models (M_2 and M_3). The results showed that the R_v^2 of the three models were all negative, suggesting that 46 samples might be too few to build an appropriate model for 67 samples. High R_m^2 and $R_{adj,m}^2$ were achieved for 46 samples using M_2 and M_3, whereas the rest of the 21 samples had low R_v^2 and $R_{adj,v}^2$. This likely implies that the optimal variates suitable for the modeling samples or the interpolated coefficients of GWR model are not applicable to the verification samples. To construct an appropriate model for Zn content estimation, four methods were proposed, in which $R_{adj,v}^2$ was considered in the first two methods, and OLS and GWR stepwise regressions were used in the other two methods using all 67 samples. A 60-repetition random process, by repeatedly selecting 57 samples for modeling and 10 samples for validation, was used to assess the model performance. The results revealed that the model with four variates (bandwidth 3D/4) in Method 4 produced the best verification effect compared with all GWR models and the OLS model in Method 3. There is a convergence effect in both R_m^2 and R_v^2 for GWR models in Method 3. Subsequently, the four variates in Method 4 were incorporated into the OLS model, and the best verification effect was achieved. The OLS and the associated GWR models with four variates from Method 4 were designated as the representative OLS and GWR models for soil Zn content estimation on Pingtan Island.

The spectral characteristics of heavy metal can be enhanced by different spectral transformations, and the reciprocal transformation and its derivative forms could improve the performance of prediction models [11]. As depicted in Figure 4, more candidates are in the data of FOs ≥ 0.6 with the reciprocal transformation. We found two similar variates in the reciprocal transformed data among the four selected variates from Methods 2 to 4, which were used for comparison in Figure 7e, and among the six variates obtained by the GWR stepwise regression of IO derivatives (bandwidth 5D/8). The two variates selected for Method 1 were reciprocal data. The reciprocal transformation was effective, which is in agreement with previously published results [11,12]. There was only one variate of reciprocal of the logarithm selected in Methods 2–4 and two of those were selected in the IO derivative model, despite the fact that the number of candidates was less and only one candidate was sifted for FOs = 1.6 and 2 in the transformation of reciprocal of logarithm, as shown in Figure 4. One logarithm variate was also selected for models of Methods 2–4 and the IO derivative model, and one square root variate was selected for the IO derivative model. Results from this study showed that the applied spectral transformations are useful in extracting information from hyperspectral data, and the reciprocal transformations of the original reflectance and the logarithm of the reflectance data are the most efficient.

Overall, the variates selected using Methods 2–4 and GWR stepwise regression of IO derivatives are similar. The similarities of the FOs of the reciprocal data at 1560 nm are 0.8 and 1, the FOs of the reciprocal of logarithm at 2200 nm are 0, 1.6 and 2, and the FOs of reciprocal and logarithm at 2020 nm are 1 and 1.2. Tables 5 and 9 and Figure 8 show that the regression coefficients of the four selected variates using Methods 2–4 are significant. Except for 2200 nm, the two wavelengths of 1560 nm and 2020 nm are not the wavelengths among the five bands with a significant correlation shown in Figure 3. These two specific wavelengths are needed for modeling the soil Zn content, and not the other four bands, given their obvious variations in spectral data shown in Figure 2.

Our results demonstrate a clear convergence effect, as both R_m^2 and R_v^2 approach the associated OLS values when bandwidth increases (Figure 7c). Based on these results, the mean R_v^2 was computed from the 10,000-repetition random process for the OLS model established using the four variates (bandwidth 3D/4) in Method 4. We found that the averaged verification result was the best among all studied models. Comparing the mean R_v^2 of the OLS model in Method 3 with that of OLS-2 (Table 8), the mean R_v^2 of OLS-2 improved by about 44.6%. Comparing the five parameter estimates of the representative GWR model (Figure 8) and the corresponding SEs of the representative OLS model (Table 9), the spatial non-stationarity of the five parameter estimates is insignificant, and the global relationships between the four selected hyperspectral variables and soil Zn content are more prominent. In Method 2, the t-value is emphasized, and the criterion to retain the variates in the GWR stepwise regression model in Method 4 is the average of $|t| > 2$. This may result in the global relationships of selected variables being prominent in Methods 2 and 4. Jiang et al. [11] showed that there was significant spatial non-stationarity for each parameter in the GWR prediction model of Zn content sampled from Fuzhou. The discrepancy between the two studies may have resulted from the differences in the study procedures and the fact that the area of Fuzhou is much larger than that of Pingtan Island (the area of Fuzhou is 11,968 km^2 in contrast to Pingtan's 274.3 km^2).

4. Conclusions and Recommendations

In this study, we applied spectral transformations and FO differential operations to spectral data to build appropriate GWR and OLS models for soil Zn content estimation on Pingtan Island, Fujian, China. There were similarities in four selected variates between the different methods we used to select them: reciprocal, FO = 0.8 and 1, and 1560 nm; reciprocal of logarithm, FO = 0, 1.6, and 2, and 2200 nm; reciprocal and logarithm, FO = 1 and 1.2, and 2020 nm. To better predict new samples and obtain accurate variations in regression coefficients, the mean R_v^2 of a random process was used as an indicator to assess the models. The results showed that the GWR stepwise regression is the most effective method to select better variables. A GWR model constructed using GWR stepwise regression with the bandwidth of 3D/4 was designated as the representative GWR model according to the mean R_v^2. Due to a convergence effect in the mean R_v^2 when bandwidths increased, a representative OLS model constructed using the four variables, selected through the GWR stepwise regression with the best verification effect among all studied models, was obtained. The GWR technique allows each data point to be weighted by its distance from the regression point. The closer the data point to the regression point, the more weight it receives. This enables the creation of a local model of relationship (as opposed to a global) to be measured. Therefore, GWR models tend to produce better results. The results presented in this study provide a reference for space-based or aerial hyperspectral soil Zn content modeling.

Despite the encouraging results obtained from this study, limitations exist in the proposed methodology that warrant improvement in future studies. For instance, the 67 soil samples were too few for testing the robustness of the proposed method. Pingtan is a small island, where spatial correlation can affect the accuracy and precision of the results. When designing future studies, a much larger number of soil samples should be collected across a wider range of geographical locations. Simultaneous comparison of the modeling results from several study sites in Fujian Province, China, should be considered, as well as conducting a new sampling on Pingtan Island. To improve prediction accuracy, finding locally effective variates and investigating the hyperspectral retrieval mechanisms are suggested. In addition to using hyperspectral signals directly, other indirect methods should be explored to provide an additional source of verification. For instance, vegetation suffering from heavy metal stress can exhibit certain symptoms through leaf coloration, shape, and structure changes. Recent advances in nanoscale science and technology have offered new tools to detect and quantify metal ions for environmental monitoring and should be explored for retrieving soil heavy metal content.

Author Contributions: Conceptualization, J.S. (Jinming Sha), X.L. (Xue Lin), J.J. and B.J.; methodology, Y.-C.S. and X.L. (Xue Lin); software, X.L. (Xue Lin) and Y.-C.S.; validation, X.L. (Xue Lin) and Y.-C.S.; formal analysis, X.L. (Xue Lin) and Y.-C.S.; investigation, X.L. (Xue Lin), J.S. (Jinming Sha), X.L. (Xiaomei Li) and Y.-C.S.; resources, X.L. (Xiaomei Li) and J.S. (Jinming Sha); data curation, X.L. (Xue Lin), X.L. (Xiaomei Li), J.S. (Jinming Sha) and J.J.; writing—original draft preparation, X.L. (Xue Lin) and Y.-C.S.; writing—review and editing, J.S. (Jiali Shang), X.L. (Xue Lin), Y.-C.S. and Y.-Y.S.; visualization, X.L. (Xue Lin), J.S. (Jiali Shang) and Y.-C.S.; supervision, J.S. (Jinming Sha); project administration, J.S. (Jinming Sha); funding acquisition, J.S. (Jinming Sha).

Funding: This research was funded by (1) International University Cooperation on Land Protection in European-Asiatic Countries (IUCLAND) (561841-EPP-1-2015-1-IT-EPPKA2-CBHE-JP), supported by EU Erasmus+ Project, 2015.10–2019.01; (2) integrated geospatial information technology and its application to Resource and environmental management towards the GEOSS (247608), supported by Chinese Ministry of Science and Technology (MOST), 2012–2018; (3) soil heavy metal monitoring based on hyperspectral data by use of GWR model (2018I0005), supported by the Fujian Provincial Department of the Science & Technology, 2018–2020; and (4) monitoring land use/cover change by use of the high resolution SAR data (NO.[2016]105), supported by China-Slovenia governmental cooperation Project, 2016.0–2018.10.

Acknowledgments: We are grateful to Prof. Wenhui Chen of College of Geographical Science, Fujian Normal University for lending us a laboratory, and instructing us in using ASD FieldSpec 3 spectroradiometer. We would like to thank our laboratory members Zhenglong Zhou and Shang Gao for their help in soil sample preparation.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. He, J.; Zhang, S.; Zha, Y.; Jiang, J. Review of retrieving soil heavy metal content by hyperspectral remote sensing. *Remote Sens. Technol. Appl.* **2015**, *30*, 407–412. [CrossRef]
2. Chi, G.; Guo, N.; Chen, X. Hyperspectral remote sensing monitoring on heavy metal contaminated farmland. *Soils Crops* **2017**, *6*, 243–250. [CrossRef]
3. Wang, F.; Gao, J.; Zha, Y. Hyperspectral sensing of heavy metals in soil and vegetation: Feasibility and challenges. *ISPRS J. Photogramm. Remote Sens.* **2018**, *136*, 73–84. [CrossRef]
4. Liu, J.; Dong, Z.; Sun, Z.; Ma, H.; Shi, L. Study on hyperspectral characteristics and estimation model of soil mercury content. *IOP Conf. Ser. Mater. Sci. Eng.* **2017**, *274*, 012030. [CrossRef]
5. Wu, Y.Z.; Chen, J.; Ji, J.F.; Tian, Q.J.; Wu, X.M. Feasibility of reflectance spectroscopy for the assessment of soil mercury contamination. *Environ. Sci. Technol.* **2005**, *39*, 873–878. [CrossRef] [PubMed]
6. Xia, F.; Peng, J.; Wang, Q.L.; Zhou, L.Q.; Shi, Z. Prediction of heavy metal content in soil of cultivated land: Hyperspectral technology at provincial scale. *J. Infrared Millim. Waves* **2015**, *34*, 593–598, 605. [CrossRef]
7. Fotheringham, A.S.; Brunsdon, C.; Charlton, M. *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*; Wiley: Chichester, UK, 2002; ISBN 978-0-470-85525-6.
8. Fotheringham, A.S.; Charlton, M.; Brunsdon, C. The geography of parameter space: An investigation of spatial non-stationarity. *Int. J. Geogr. Inf. Syst.* **1996**, *10*, 605–627. [CrossRef]
9. Brunsdon, C.; Fotheringham, A.S.; Charlton, M.E. Geographically weighted regression: A method for exploring spatial nonstationarity. *Geogr. Anal.* **1996**, *28*, 281–298. [CrossRef]
10. Jaber, S.M.; Al-Qinna, M.I. Global and local modeling of soil organic carbon using Thematic Mapper data in a semi-arid environment. *Arab. J. Geosci.* **2015**, *8*, 3159–3169. [CrossRef]
11. Jiang, Z.; Yang, Y.; Sha, J. Application of GWR model in hyperspectral prediction of soil heavy metals. *Acta Geogr. Sin.* **2017**, *72*, 533–544. [CrossRef]
12. Jiang, Z.L.; Yang, Y.S.; Sha, J.M. Study on GWR model applied for hyperspectral prediction of soil chromium in Fuzhou City. *Acta Ecol. Sin.* **2017**, *37*, 8117–8127. [CrossRef]
13. Xu, J.; Feng, X.; Guan, L.; Wang, S.; Hu, Q. Fractional differential application in reprocessing infrared spectral data. *Control Instrum. Chem. Ind.* **2012**, *39*, 347–351. [CrossRef]
14. Wang, J.; Tiyip, T.; Ding, J.; Zhang, D.; Liu, W.; Wang, F. Quantitative estimation of organic matter content in arid soil using Vis-NIR spectroscopy preprocessed by fractional derivative. *J. Spectrosc.* **2017**, *2017*, 1375158. [CrossRef]
15. Wang, J.; Tiyip, T.; Zhang, D. Spectral detection of chromium content in desert soil based on fractional differential. *Trans. Chin. Soc. Agric. Mach.* **2017**, *48*, 152–158. [CrossRef]
16. O'brien, R.M. A caution regarding rules of thumb for variance inflation factors. *Qual. Quant.* **2007**, *41*, 673–690. [CrossRef]
17. Fotheringham, A.S.; Oshan, T.M. Geographically weighted regression and multicollinearity: Dispelling the myth. *J. Geogr. Syst.* **2016**, *18*, 303–329. [CrossRef]
18. Standards Press of China. *Classification and Codes for Chinese Soil*; Standards Press of China: Beijing, China, 2009.
19. Li, S.; Li, H.; Sun, D.; Zhou, L.; Bao, J. Characteristic and diagnostic bands of heavy metals in Beijing agricultural soils based on spectroscopy. *Chin. J. Soil Sci.* **2011**, *42*, 730–735. [CrossRef]
20. Xie, X.; Sun, B.; Hao, H. Relationship between visible-near infrared reflectance spectroscopy and heavy metal of soil concentration. *Acta Pedol. Sin.* **2007**, *44*, 982–993. [CrossRef]

21. Zhang, D.; Tiyip, T.; Zhang, F.; Kelimu, A.; Xia, N. Effect of fractional differential algorithm on hyperspectral data of saline soil. *Acta Opt. Sin.* **2016**, *36*, 0330002. [CrossRef]
22. Chasco Yrigoyen, C.; García Rodríguez, I.; Vicéns Otero, J. Modeling spatial variations in household disposable income with geographically weighted regression(1). *Estadística Española* **2008**, *50*, 321–360.