Electrical Conductivity Inversion Method of Saline Soil based on Sentinel-2 MSI data

Yishan Sun, Xiaojie Li *, Xiaofeng Li, Tao Jiang, Xingming Zheng

Northeast Institute of Geography and Agroecology, Chinese Academy of Sciences, Changchun, 130102, China

Abstract: Electrical conductivity (EC) is not only an important index to evaluate the degree of soil salinization, but also an essential basis for judging whether saline soil can be improved and assess the effect of improvement efforts. Satellite remote sensing provides much information for large scale EC inversion of saline soil, which enables the possibility for evaluating the degree and distribution of soil salinization. Taking the salinized region of western Jilin Province as the study area, 328 salinized soil samples were collected, and the EC was measured in June 2019. The construction of the optimal spectral parameters was based on the correlation between the conductivity and the spectral reflectivity of Sentinel-2 MSI data; after satisfying the normal distribution for the Box-Cox transformation of EC, the inversion model of EC was established by using linear regression model, support vector machine (SVM), regression tree (RT), Gaussian process regression (GPR), and ensemble tree (ET). The verification results of the model on the validation set showed that the performance of GPR was optimal (R^2 = 0.66, RMSE = 0.48 mS/cm, MAE=0.52 mS/cm), which increased R^2 by 29.04% compared with the traditional linear regression model. Finally, according to the GPR model, the EC results of pixel-level resolution (10 m × 10 m) of saline soil in western Jilin Province were inversed, which provided a scientific basis for the study of the distribution characteristics and improvement scheme of saline soil.

Keywords: Sentinel-2 MSI; Spectral characteristics; Electrical Conductivity retrieval; Saline soil; Machine learning

1 Introduction

Soil salinization and secondary salinization are significant problems faced by China and the whole world. A characteristic of salinized soil is electrical conductivity (EC), wherein higher levels of salt content are strongly correlated with more excellent conductivity; therefore, EC is an important index to judge the degree of soil salinization(Lian et al., 2010).

Over the past 20 years, remote sensing has become the most common method for detecting soil EC because of its reliable real-time results and low cost (Csillag et al., 1993; Eldeiry and Garcia, 2008). In many remote sensing methods, large-scale salinized soil monitoring is based on spectral response characteristics. For the spectral characteristic response band of saline soil, many scholars have studied different remote sensing satellites and have their conclusions. The optimum band combination of saline soil monitoring was studied by Dwivedi et al. (1992), and the results show that the 1, 3, and 5 band combinations of TM data contain the most significant amount of salinization information. Wu Yunzhao et al. (2003) found that the visible (0.55-0.77 μm), near-infrared (0.9-1.03 μm, 1.27-1.52 μm), and short-wave infrared (1.94-2.15 μm, 2.15-2.31 μm, 2.33-2.4 μm) are the critical bands for identifying the saline soil. Based on the eight bands of ETM+, Shrestha(2006) established a salt prediction model of normalized vegetation index (NDVI) containing multiple spectral variables and the normalized salt index (NDSI) and salt data. It was found that band 7 (middle infrared) and band 4 (near-infrared) had the highest correlation with soil conductivity. Srivastava et al. (2015) found that the spectra between 1390 nm and 2400 nm are very sensitive to salinity changes based on the information of visible-near infrared reflectance spectra. Meti et al. (2019) found that the combination of short-wave infrared and visible bands of Sentinel-2 and Landsat-8 significantly improved the correlation of saline soil pH and EC in the arid regions of northern India. Davis et al. (2019) used Landsat OLI and Sentinel-2 MSI to reverse the conductivity of saline soil, and the result showed that MSI was superior to OLI and that the visible light band was more sensitive to soil salinity.

On this basis, many others have studied the model algorithm of the quantitative relationship between soil salinity and spectral characteristics. To sum up, the main modeling methods include linear regression, least squares, and random forest. Allbed et al. (2014a) established the correlation between the spectral index and conductivity based on IKONOS images. They used linear regression to predict the spatial change of soil salt in the Hassa oasis. Nawar et al. (2015) used multivariate adaptive regression splines to construct a soil spectrum and EC prediction model. Besides, Gorji et al. (2017) obtained the spatial distribution of saline soil around Lake Tuz in Turkey based on SI regression analysis. Zhang Suming et al. (2018) used the Kenli area of the Yellow River Delta as their research area. They combined the measured and multi-time phase remote sensing data to analyze and construct their salt inversion model. Farifthe et al. (2007) predicted the salt content of soil utilizing the partial least square regression and artificial neural network. Fan et al. (2016) carried out soil salt inversion and mapping in the Yellow River Delta region based on the PLSAR model using 30 years of multi-source Landsat data. Wang et al. (2019a) used partial least squares regression and random forest inversion to develop a salinity map of the Ebinur Lake area in northwest China, based on the extraction of conductivity and multi-band spectral indexes of saline soil from 116 sampling points. Li et al. (2019) extracted ten sensitive variables of EC from Landsat using random forest to establish a soil salinity prediction model. Wang et al. (2019b) combined soil salinity data with spectral data in order to achieve soil salinity estimation through constructing a random forest model in arid and semiarid regions. The above studies show the feasibility of quantitative analysis of soil salt. However, hyperspectral data are still obtained by data, and the application of hyperspectral data in regional soil salinization monitoring is limited by some practical factors, such as small image coverage area and others.

To sum up, in previous studies, the quantitative estimation of the soil salinity by spectrum analysis is realized by screening the sensitive wavebands or the known spectral indexes as the modeling factors. However, this method only takes into account the relationship between the soil salinity and the sensitive waveband or the sensitive spectral index, and then construct the optimal linear and nonlinear models. However, they forgot considering whether the distribution of variables will affect the accuracy of models before modeling.

Based on Sentinel-2 MSI spectral data and measured EC of saline soil, the Box-Cox transformation of the conductivity which does not satisfy the normal distribution was performed, the relationship between different spectral parameters and transformed EC data of saline soil is explored, and the optimization of modeling variables is performed. On this basis, the nonlinear estimation model of EC is constructed by using a machine learning algorithm, and we get an inversion method that is matched with EC of carbonated (soda) saline soil in the western Jilin Province. In order to improve the inversion accuracy of EC of saline soil in the western Jilin Province, and to provide data support for accurate monitoring, evaluation, improvement, and utilization of saline soil.

2 Materials and Methods

2.1 Site descriptions and soil sampling

The western part of Jilin is part of the Songnen Plain, with the range of 121°38′-126°11′E, 43°59′-46°18′N, as shown in Figure 1, the total area is approximately 43360 square kilometers, and the terrain is flat. This area belongs to temperate continental monsoon climate; the average annual precipitation and annual evaporation are present as 400-500 mm and 1000-2000mm. (Liu et al., 2015; Xu et al., 2018). Soil evaporation is intense, which makes it easy for salt to accumulate at the surface. This severely imbalanced evaporation-precipitation ratio, coupled with the influence of local topography, hydrogeological conditions, and human activities, makes the degree of salinization in this area grave. The EC results of pixel-level resolution (10 m × 10 m) of saline soil in western Jilin Province were inversed.

Carried out the field experiment from June 20–28, 2019, and selected 328 experimental sites. In order to reduce the influence of mixed pixels, taken three points near each sampling point, and collected the soil samples by ring knife. After each soil sample was dried and sifted through 1 mm mesh, three soil samples from each sampling site were uniformly mixed into 10 g samples to prepare soil suspensions with a soil/water ratio of 1:5, the soil suspension was set aside for about 3 hours, and EC was measured using a conductivity meter (LEICI, Model DDS-307A).

In order to construct and verify the EC inversion model with pixel-level resolution (10 m × 10 m) of saline soil in western Jilin Province, 328 sample points were randomly grouped, of which randomly used two-thirds total 219 points for modeling, which was called the training dataset, and used the remaining one-third total 109 points for validation of the model, which were called validation dataset.

2.2 Sentinel-2 MSI spectral information extraction and feature construction

In order to coincide with the field sampling time, the Sentinel-2 MSI L1C multispectral data of the study area on June 23, 2019 was selected, as shown on the right side of Figure 1(false-color composite), and extracted the reflectivity of each band corresponding to the sampling points after atmospheric correction. The band parameters are shown in table 1.

Table 1 Spectral bands of Sentinel-2 MSI sensor

Acronym	Band	Band center /nm	Band width/nm	Spatial resolution/m
B1	Coastal	443	45	60
B2	Blue	492	98	10
B3	Green	560	46	10
B4	Red	665	39	10
B5	Vegetation Red Edge	703	20	20
B6	Vegetation Red Edge	739	18	20
B7	Vegetation Red Edge	779	28	20
B8	NIR	833	133	10
B8A	Vegetation Red Edge	864	32	20
B9	Water vapour	943	27	60
B10	SWIR- Cirrus	1376	76	60
B11	SWIR	1610	141	20
B12	SWIR	2186	238	20

The construction of spectral parameters includes two methods; one is generated by sensitive band combination operation (addition, multiplication), the other is to evaluate the degree of soil salinization by using the existing spectral indexes. Combined these results with the existing research results, and selected the following spectral indexes for correlation analysis with EC (formula (1)), including soil salt index SI1, SI2, SI3 (Allbed et al., 2014a; Douaoui et al., 2017; KHAN et al., 2005), SI4, SI5, NDSI, and ratio salt index (SI-T). The calculation formula for each index shown in Table 2. Calculation formula of correlation coefficient present as follows:

$$R=\frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\overline{y})^2}} \tag{1}$$

Table 2 Spectral Index Construction

Spectral Index	Formula
Soil salinity index SI1	$SI1=\sqrt{G\times R}$
Soil salinity index SI2	$SI2=\sqrt{G^2+R^2+NIR^2}$
Soil salinity index SI3	$SI2=\sqrt{G^2+R^2}$
Soil salinity index SI4	$SI4=(SWIR\times R)/G$
Soil salinity index SI5	$SI5=(B-SWIR2)/(B+SWIR2)$
Normalized salinity index (NDSI)	$NDSI=(R-NIR)/(R+NIR)$
Ratio salt index (SI-T)	$SI-T=R/NIR$

2.3 Modeling methods and evaluation index

Our study showed a modeling flow chart of remote sensing inversion with EC in figure 2. Firstly, a training set and single-band reflectance from Sentinel-2 MSI data were analyzed to screen out the sensitive bands. Then the spectral parameters were constructed based on the sensitive band and screened the optimal spectral parameters. We performed a pre-modeling test dataset distribution that satisfies the Gauss-Markov normality hypothesis. We found the optimal transformation for the data that does not satisfy the condition, thus improving the formality, symmetry, and homogeneity of variance of the data distribution. Finally, using the sensitive band and the optimal spectral parameters as the independent variables, the measured EC was used as the response variables to construct the inversion model and obtain more accurate modeling results, as shown in Figure 2.

2.3.1 Box-Cox Transform

In practical applications, the response variables are often not following the normal distribution, so it is not suitable for data analysis directly. Box-Cox transform was proposed by Box and Cox(1964) for the nonlinear transformation of response variables. By determining an optimal parameter λ, the non-normal data is transformed into approximately normal data, and then, the transformed data is regressed. The Box-Cox transformation of y (y >0) can be represented by formula (2).

$$y^{(\lambda)} = \begin{cases} \dfrac{y^{\lambda}-1}{\lambda}, \lambda \neq 0 \\ \ln y, \lambda = 0 \end{cases} \tag{2}$$

where y is the raw data, λ is the parameter of the change to be determined.

Box-Cox transform determines the optimal λ value by finding the maximum $L_{max}(\lambda)$ of the likelihood function. In order to calculate the pure logarithm on both sides of the likelihood function, the term A-independent constant is omitted. Formulas (3) and (4).

$$\ln ¿ \tag{3}$$

$$J(\lambda, y) = \prod_{i=1}^{n} \left| \frac{d y_i^{(\lambda)}}{d y_i} \right| \tag{4}$$

Where MSE is the mean square error, n is the data quantity.

2.3.2 Linear regression model

The traditional linear regression model is as follows:

$$y = \varepsilon + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k \tag{5}$$

Where y is the response variable of the model, $x_1 - x_k$ are independent variables, ε is a constant, and β_0, β_1, …, β_k are undetermined coefficients. In this paper, x is the spectral index of Sentinel-2 MSI data, y is the Box-Cox transform result of the measured EC.

2.3.3 Machine learning models

Because of the influence of mixed pixels and atmospheric radiation, the relationship between spectral parameters and EC of saline soil may be nonlinear, so the machine learning algorithm model is considered to invert the EC of saline soil. At present, common machine learning models include the following:

1) Support Vector Machine

V. Vapneilk and Cortes proposed a support vector machine (SVM) (Cortes and Vapnik, 1995). For regression problems that are not suitable for linear models, SVM can improve the accuracy of regression prediction by mapping the low-dimensional training dataset to the high-dimensional space construction model, and it has good generalization ability for small sample data sets.

2) Regression Tree

The regression tree(RT) is a binary decision tree for regression analysis (Mingers, 1989). The feature selection is carried out recursively, and the given input variable predicts the probability distribution of the output variable, and then, the binary regression tree is generated. The regression tree is unstable with big data sets, and the weak change of the training dataset may lead to a change in the tree structure.

3) Gaussian Process Regression

Gaussian process regression (GPR) is a new machine learning algorithm, which is a non-parametric regression probability model based on Bayesian and statistical learning theory. It is assumed that the input of the model is x, and the output is $f(x)$. A set of input sets $\left\{x_i \vee i=1,2,\ldots,n\right\}$ obtains an output set $f(x)$ through a Gaussian process regression model. Under the assumption of the mean of zero, the distribution form of $f(x)$ can be expressed as follows: $f(x) - N\left(0, K\left(\theta, x, x'\right)\right), K\left(\theta, x, x'\right)$ is a covariance matrix with super parameters (some parameters of kernel functions).

4) Ensemble Tree

The ensemble tree (ET) is a regression-lifting algorithm based on the regression tree and using the forward distribution and adding. This ensemble learning method constructs a prediction model by weighting several regression tree results when the instance is predicted. Compared with the regression tree, better results may be obtained for some datasets, thus improving prediction performance.

2.3.4 Evaluation indicators

In order to evaluate the accuracy of the inversion model, the method of V fold cross verification (VFCV)(Geisser, 1975) is used to model the data, and the determination coefficient R^2, root mean square error (RMSE) and mean absolute error (MAE) are used to evaluate the model. The calculation method is shown in the formulas (6), (7), and (8).

$$R^2=\frac{\sum_{i=1}^{N}(f_i-\bar{y})^2}{\sum_{i=1}^{N}(y_i-\bar{y})^2} \tag{6}$$

$$RMSE=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i-f_i)^2} \tag{7}$$

$$MAE=\frac{1}{N}\sum_{i=1}^{N}|y_i-f_i| \tag{8}$$

Where y_i represents a true value, f_i represents a predicted value, $\bar{y}$ represents a mean value, and N represents a sample size.

The principle of the VFCV method is to divide the data set into V parts, one from V parts as verification, the remaining V-1 as training, repeat V times, and take the mean value of each verification result as the final result to find the optimal model. VFCV can improve the generalization ability of the model to a certain extent. In this study, the V value is 10. According to research experience, it is found that tenfold cross-verification can balance deviation and variance, which is the best choice to obtain model error estimation. The closer the R^2 is to 1 in the evaluation parameter, the higher the fitting accuracy of the model; The closer the RMSE is to 0, the better the performance of the model, the smaller the difference is between the measured value and the predicted value; compared with RMSE, MAE has better robustness to outliers in the dataset and does not reduce the accuracy of the model as a whole.

3 Results

3.1 EC Measurement results

The statistical results of the measured EC for the 328 samples collected in the field are shown in Table 3. It can be seen from the table that the range of EC is 0.66 mS/cm, the standard deviation ranges from 0.06 to 5.87 mS/cm, and the coefficient of variation is significant, which indicates that the sample points have very high spatial heterogeneity.

Table 3 EC (mS/cm) Statistical Table

	Maximum	Minimum	Mean	Standard deviation	Coefficient of variation (%)
All data N=328	5.87	0.06	0.66	0.91	138
Training dataset N=219	5.87	0.06	0.72	0.98	136
Validation dataset N=109	5.79	0.08	0.53	0.73	137

3.2 Selection of sensitive bands

In this paper, the correlation between the measured EC data and the single-band spectral reflectance from the Sentinel-2 MSI was analyzed ($p<0.01$). The correlation coefficient R values were obtained, as shown in Table 4 below ($p<0.01$). The results show that the B2, B3, B4, and B8 bands were sensitive bands.

Table 4 Correlation between EC and spectral reflectance of Sentinel-2 MSI

Band	B2	B3	B4	B5	B6	B7	B8	B8A	B11	B12

R	0.42	0.43	0.41	0.34	0.29	0.28	0.42	0.24	0.15	0.20

N=219, N is number of samples

3.3 Construction of optimal spectral parameters

In order to consider the spectral characteristics synthetically, the spectral parameters of the inversion model combined B2, B3, B4 and B8 bands and spectral index SI1, SI2, SI3, SI4, SI5, NDSI, and SI-T by multiplication. Table 5 shows the correlation coefficient R ($p<0.01$) between the different spectral parameters based on the Sentinel-2 MSI data and the saline soil EC. Comparing Tables 3 and 4, the correlation between EC and spectral parameters of saline soil was higher than that of a single band reflectivity, and the band combination of R > 0.40 was selected as the spectral parameter of the estimation model. Thus B2×B3×B4, B2×B3×B8, B3×B8, B3×B4×B8, B2×B3, B2×B8, SI2, SI1and SI3 were chosen as the optimal spectral parameters.

Table 5 EC correlation analysis with spectral parameter reflectance

Spectral parameters	B2×B3×B4	B2×B3×B8	B3×B8	B3×B4×B8	B2×B3	B2×B8	SI2
R	0.52	0.51	0.48	0.48	0.47	0.43	0.42
Spectral parameters	SI1	SI3	B2×B4	NDSI	SI_T	SI5	SI4
R	0.41	0.41	0.38	0.32	0.33	0.23	0.14

N=219, N is number of samples

3.4 Box-Cox parameter λ estimation results

By testing the data of response variables and independent variables involved in modeling, we can see that the response variable EC does not conform to the normal distribution, as shown in Figure 3. The maximum likelihood estimation method proposed by Box-Cox was used to determine the parameter λ value. For different λ values ($-2 \le \lambda \le 2$), the maximum value $L_{max}(\lambda)$ of the likelihood function was calculated by the least square estimation of the linear regression model, which is expressed as Log-Likelihood $=\ln\left(L_{max}(\lambda)\right)$, with λ as the horizontal axis and

Log-Likelihood as the longitudinal axis. The results are shown in Figure 4 and Table 6. It can be seen from the results that when λ =0, Log-Likelihood was the largest. According to formula (1), the EC = ln (EC) after the Box-Cox transform was calculated as EC_bc, and the data are close to a normal distribution, as shown in Figure 3b).

Table 6 Maximum value of likelihood function and λ statistical table

λ	-2.00	1.75	-1.50	-1.25	-1.00	-0.90	-0.80
Log-Likelihood	-507.38	-449.76	-396.97	-349.42	-307.68	-292.79	-279.01
λ	-0.70	-0.60	-0.50	-0.40	-0.30	-0.20	-0.10
Log-Likelihood	-266.44	-255.15	-245.28	-236.93	-230.29	-225.53	-222.87
λ	**0.00**	0.10	0.20	0.30	0.40	0.50	0.60
Log-Likelihood	**-222.57**	-224.90	-230.16	-238.62	-250.53	-266.02	-285.13
Λ	0.70	0.80	0.90	1.00	1.25	1.50	1.75
Log-Likelihood	-307.78	-333.75	-362.77	-394.49	-483.33	-582.14	-687.60

3.5 Linear regression models for the EC retrieval

Using EC and EC_bc data as dependent variables of linear regression model in 2.3.2, respectively. The regression model and verification accuracy were obtained, as shown in Table 7 and Table 8. It can be seen from the table that the regression model with single band and spectral parameters as independent variables and EC_bc as dependent variables had the highest accuracy, and the R^2 of the verification accuracy was 0.51. Therefore, EC_bc was used to participate in the modeling. The regression model and verification accuracy were obtained, as shown in Table 7 and Table 8.

Table 7 Linear inversion model before and after soil conductivity transformation

Variable	Regression model	Model accuracy R^2	Verification accuracy R^2
Single band	$EC=-1.266+0.001\times B2+0.002\times B3-0.001\times B4+0.003\times B8$	0.21	0.21
	$EC\_bc=-3.078+0.001\times B2+0.001\times B3-0.001\times B4+7.90\times 10^{-5}\times B8$	0.32	0.39
Spectral parameters	$EC=1.962+1.673\times 10^{-8}\times(B2\times B3)-1.160\times 10^{-7}\times(B2\times B8)+2.008\times 10^{-7}\times(B3\times B8)+2.109\times 10^{-11}\times(B2\times B3\times B4)+5.073\times 10^{-11}\times(B2\times B3\times B8)+1.178\times 10^{-}$	0.37	0.37

	$^{11}\times(B3\times B4\times B8)+0.001\times SI2-0.001\times SI3$		
	$EC\_bc=-0.434+4.586\times10^{-8}\times(B2\times B3)-1.862\times10^{-7}\times(B2\times B8)+1.711\times10^{-7}\times(B3\times B8)+2.839\times10^{-11}\times(B2\times B3\times B4)+3.724\times10^{-11}\times(B2\times B3\times B8)+1.658\times10^{-11}\times(B3\times B4\times B8)+0.001\times SI2-0.011\times SI3$	0.46	0.46
Single band and Spectral parameters	$EC=0.543+0.001\times B2+0.002\times B3+0.001\times B4+0.003\times B8+3.790\times10^{-8}\times(B2\times B3)-4.196\times10^{-7}\times(B2\times B8)+2.246\times10^{-7}\times(B3\times B8)+1.448\times10^{-11}\times(B2\times B3\times B4)+4.789\times10^{-11}\times(B2\times B3\times B8)+3.191\times10^{-11}\times(B3\times B4\times B8)-0.004\times SI2$	0.42	0.45
	$EC\_bc=1.489+0.001\times B2+0.001\times B3+0.001\times B4+0.003\times B8+7.403\times10^{-8}\times(B2\times B3)-4.354\times10^{-7}\times(B2\times B8)+1.962\times10^{-7}\times(B3\times B8)+1.731\times10^{-11}\times(B2\times B3\times B4)+3.464\times10^{-11}\times(B2\times B3\times B8)+3.900\times10^{-11}\times(B3\times B4\times B8)-0.004\times SI2$	0.49	0.51

N=219, N is number of samples

Table 8 Evaluation Indexes of linear inversion models of EC_bc

Variable	Training dataset			Validation dataset		
	RMSE /(mS/cm)	R^2	MAE /(mS/cm)	RMSE /(mS/cm)	R^2	MAE /(mS/cm)
Single band	0.55	0.32	0.50	0.56	0.39	0.51
Spectral parameters	0.52	0.46	0.48	0.55	0.46	0.49
Single-band and Spectral parameters	**0.53**	**0.49**	**0.44**	**0.56**	**0.51**	**0.44**

3.6 Machine learning models for the EC retrieval

The optimal spectral parameters selected from 3.3 were used as the input, and the EC_bc was used as the output to build the model with five algorithms of SVM, RT, GPR, and ET, respectively. The inversion results of each model to the validation dataset are shown in Figure 5. In order to quantitatively describe the inversion accuracy of the model, the evaluation index results of the five models are shown in Table 9.

Table 9 Evaluation Indexes of five models

Model	Training dataset			Validation dataset		
	RMSE/(mS/cm)	R^2	MAE/(mS/cm)	RMSE/(mS/cm)	R^2	MAE/(mS/cm)
LINEAR	0.53	0.49	0.44	0.56	0.51	0.44
SVM	0.43	0.58	0.48	0.44	0.65	0.53
RT	0.50	0.58	0.57	0.52	0.57	0.53
GPR	**0.42**	**0.61**	**0.58**	**0.48**	**0.66**	**0.52**

ET	0.51	0.61	0.53	0.49	0.62	0.54

As can be seen from Table 9, the traditional linear regression model had the worst results among the evaluation indexes of the five models. Among the four machine learning models, the R^2 of the GPR model was 0.66, the RMSE was 0.48, and the MAE was 0.52. The prediction performance of the GPR model was the best, SVM was the second, and RT was the lowest. In the comprehensive view, the accuracy of the five models for the inversion of the saline soil EC was GPR> SVM> ET> RT> LINEAR. Figure 6 shows the comparison between the measured values and the predicted values of 109 points in the data set verified by the GPR model.

3.7 The inversion results of saline soil EC in the west of Jilin Province

In order to reflect the EC of the large-area of saline soil in the west of Jilin, according to the most accurate GPR model in 3.6, based on the Sentinel-2 MSI data of June 23, 2019, EC of the pixel-level resolution of the saline soil in the western part of Jilin Province was obtained by inversion in 2019. The results are shown in Figure 7.

In order to quantify the degree of soil salinization in this study area, according to the classification criterion of Kissell and Sonon (2008), the degree of salinization of inversion EC was graded and mapped. The results are shown in Figure 8. It can be seen from the results that the soil salinization in the study area tends to increase gradually from east to west. Mild saline soil was mainly distributed in Qianguo County, Changling County, and Fuyu City. Moderate and severe saline soil was mainly distributed in Zhenlai County, the junction of Da'an City, Qianan County, and Tongyu County, and a small area of extremely saline soil was distributed in Da'an City, Qianan County, and Zhenlai County.

In order to quantitatively describe the area of the soil with different degrees of salinization, the areas of several salinized soils in Figure 8 were counted, and the results are shown in Table 10.

According to the statistical data, after many years of improvement, the degree of soil salinization in the western part of Jilin Province in 2019 was mainly mild, accounting for 54.48% of the total area, moderate and severe salinization covered 33.29% of the area, and the extremely heavy salinization was 2.26% of the study area.

Table 10 Statistics of soil salinity grades in western Jilin Province in 2019

Soil Salinity Level(mS/cm)	Non-Saline Soil (0-0.15)	Low Salinity (0.16-0.50)	Medium Salinity (0.51-1.25)	Strongly Salinity (1.26-1.75)	Very High Salinity (1.76-2.0)	Excessively High Salinity (>2.0)
Area (km^2)	653.72	3572.07	1975.91	206.71	79.68	68.06
Percent (%)	9.97	54.48	30.14	3.15	1.22	1.04

4. Discussion

When the spectral index is selected, it is an important prerequisite that invalid information generated by the superimposed spectrum can be compressed, and the practical information of saline soil characteristics can be highlighted in order to improve the accuracy of the model. At present, the commonly used spectral indices are the NDVI, the NDSI, and the others mentioned above. We believe that on the one hand, these indices did not use the sensitive band to superimpose useful spectral information to delve into the spectral characteristics of saline soil; on the other hand, the presence of alkali-resistant crops such as soda can lead to the error of using NDVI to retrieve soil salinization. Allbed et al. (2014b) expressed a similar view that salt recognition based on vegetation index would not work in bare land. Therefore, the index of NDVI was avoided in this paper.

We performed a Box-Cox transformation on the EC data of the original saline soil to determine an optimal λ, thereby transforming the non-normal data into approximately normal data. Subsequently, the single-band and spectral parameters were used as independent variables, and the regression model was obtained after the Box-Cox transformation. After verification, the accuracy was $R^2 = 0.51$, which is a particular improvement over the accuracy of 0.45 without conversion (Section 3.5 Table 6). Besides, the spectral parameters were constructed by multiplying the sensitive band by Box-Cox transforming the EC data of the original saline soil and combining the single band as the modeling factor, the selectivity of the modeling was increased, and the synergy between the spectral segments was enhanced.

In existing studies, researchers (Atman et al.,2018; Bannari et al., 2018) have found that the short-wave infrared band of Sentinel-2 MSI, which can distinguish different grades of saline soils by combined with visible light bands, is more sensitive to saline soils in arid regions. Meti et al. (2019) once again demonstrated that the combination of visible light bands of Sentinel-2 MSI and short-wave infrared could significantly improve the correlation with soil EC (R=0.60-0.70). Also, several studies have demonstrated the potential of the short-wave infrared band of Sentinel-2 MSI in distinguishing saline soils (Bannari et al., 2016; Bannari et al., 2008; FARIFTEH et al., 2007). Researchers (Bannari et al., 2018) have found that light with short-wavelength infrared wavelengths can easily detect soils that are predominantly rich in sulfate minerals, chlorides, and small amounts of bicarbonate. According to this, we constructed a spectral index composed of short-wave infrared and visible light bands (Section 2.2 Table 1 SI4, SI5). $R_{SI4}=0.14$, $R_{SI5}=0.23$. However, the results show a poor correlation, indicating that the above conclusions do not apply to saline soils in western Jilin. We speculate that the western part of Jilin belongs to the Songnen plain, and the type of saline soil is inland soda saline soil, which main salt composition is $NaHCO_3$ and Na_2CO_3 with containing a small amount of sulfate and chloride, thus has present strong alkalinity. We know that saline soils in the arid area, which mainly contain chloride-sulfate saline and sulfated soils, belong to slightly alkaline soils. Due to the differences in the chemical composition, the characteristic bands of different types of saline soils are different.

At the same time, because of the different driving factors and formation mechanism of saline soil, there are many factors affecting salt, which lead to the complex nonlinear relationship between salt and spectrum. Therefore, the linear regression model is not a good reflection of this relationship; the machine learning algorithm solves the nonlinear problem of the model, which can effectively improve the accuracy of saline soil conductivity inversion. In the machine learning algorithm, the GPR model performs better (Boedecker et al., 2014; Rasmussen et al., 2005) in calculating the probability of the super-parameter acquisition and the variable output compared with the common SVM, the neural network, and RT. The model uses a Gaussian process to deduce the function distribution of the training dataset, obtains the super optimal parameters based on the kernel function, and uses the training dataset to train the super parameters to realize the prediction output; the model works better for high-dimensional small samples and non-linear regression.

5 Conclusion

In this study, according to the correlation of electrical conductivity characteristics and spectral reflectance of each band of Sentinel-2 MSI, the sensitive band was screened, and the optimal spectral parameters were constructed by mathematical operations such as multiplying the sensitive band. The EC_bc was obtained by the Box-Cox transformation of EC data, which did not satisfy the normal distribution, and we constructed the linear regression models of EC with spectral parameters and a model of EC_bc with spectral parameters, respectively. The verification results showed that the accuracy of the model R^2 after EC transformation was improved from 0.45 to 0.51. Therefore, we established the nonlinear inversion models of GPR, ET, SVM, and RT of

EC_bc. Then using validation set, the inversion accuracy of salt soil EC_bc was as follows: GPR > ET > SVM > RT > LINEAR. The most accurate GPR model for the validation dataset inversion R^2 was 0.66, proving the validity of the model. Finally, according to the model, the pixel resolution results of saline soil EC were inversed in western Jilin Province in 2019, which provides necessary data support for evaluating the salinization degree of soil and the effectiveness of the improvement scheme.

Acknowledgments

This research was funded by the National Natural Science Foundation of China (No.41671350).

References

Allbed, A., Kumar, L., Aldakheel, Y.Y., 2014a. Assessing soil salinity using soil salinity and vegetation indices derived from IKONOS high-spatial resolution imageries: Applications in a date palm dominated region. Geoderma, 230-231, 1-8. doi: 10.1016/j.geoderma.2014.03.025

Allbed, A., Kumar, L., Sinha, P., 2014b. Mapping and Modelling Spatial Variation in Soil Salinity in the Al Hassa Oasis Based on Remote Sensing Indicators and Regression Techniques. Remote Sensing, 6, 1137-1157. doi: 10.3390/rs6021137

Bannari, A., El-Battay, A., Bannari, R., Rhinane, H., 2018. Sentinel-MSI VNIR and SWIR Bands Sensitivity Analysis for Soil Salinity Discrimination in an Arid Landscape. Remote Sensing, 10, 20. doi: 10.3390/rs10060855

Bannari, A., Guedon, A.M., El-Ghmari, A., 2016. Mapping Slight and Moderate Saline Soils in Irrigated Agricultural Land Using Advanced Land Imager Sensor (EO-1) Data and Semi-

Empirical Models. Communications in Soil Science and Plant Analysis, 47, 1883-1906. doi: 10.1080/00103624.2016.1206919

Bannari, A., Guedon, A.M., El-Harti, A., Cherkaoui, F.Z., El-Ghmari, A., 2008. Characterization of Slightly and Moderately Saline and Sodic Soils in Irrigated Agricultural Land using Simulated Data of Advanced Land Imaging (EO-1) Sensor. Communications in Soil Science and Plant Analysis, 39, 2795-2811. doi: 10.1080/00103620802432717

Boedecker, J., Springenberg, J.T., Wulfing, J., Riedmiller, M., 2014. Approximate real-time optimal control based on sparse Gaussian process models, 2014 IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning (ADPRL). doi: 10.1109/ADPRL.2014.7010608

Box, G.E.P., Cox, D.R., 1964. AN ANALYSIS OF TRANSFORMATIONS. Journal of the Royal Statistical Society Series B-Statistical Methodology, 26, 211-252.

Cortes, C., Vapnik, V.N., 1995. Support Vector Networks. Machine Learning, 20, 273-297. doi: 10.1023/A:1022627411411

Csillag, F., Pasztor, L., Biehl, L.L., 1993. Spectral Band Selection for the Characterization of Salinity Status of Soils. Remote Sensing of Environment, 43, 231-242. doi: 10.1016/0034-4257(93)90068-9

Davis, E., Wang, C., Dow, K., 2019. Comparing Sentinel-2 MSI and Landsat 8 OLI in soil salinity detection: a case study of agricultural lands in coastal North Carolina. International Journal of Remote Sensing, 40, 6134-6153. doi: 10.1080/01431161.2019.1587205

Douaoui, E.K. et al., 2017. Detecting salinity hazards within a semiarid context by means of combining soil and remote-sensing data. Geoderma, 134, 217-230. doi: 10.1016/j.geoderma.2005.10.009

Dwivedi, R.S., Rao, B.R.M., 1992. The Selection of the Best Possible Landsat Tm Band Combination for Delineating Salt-Affected Soils. International Journal of Remote Sensing, 13, 2051-2058. doi: 10.1080/01431169208904252

Eldeiry, A.A., Garcia, L.A., 2008. Detecting soil salinity in alfalfa fields using spatial modeling and remote sensing. Soil Science Society of America Journal, 72, 201-211. doi: 10.2136/sssaj2007.0013

Fan, X.W., Weng, Y.L., Tao, J.M., 2016. Towards decadal soil salinity mapping using Landsat time series data. International Journal of Applied Earth Observation and Geoinformation, 52, 32-41. doi: 10.1016/j.jag.2016.05.009

FARIFTEH, Meer, V.D., ATZBERGER, CARRANZA, E., J.M., 2007. Quantitative analysis of salt-affected soil reflectance spectra: A comparison of two adaptive methods (PLSR and ANN). Remote Sensing of Environment, 110, 59-78. doi: 10.1016/j.rse.2007.02.005

Geisser, S., 1975. PREDICTIVE SAMPLE REUSE METHOD WITH APPLICATIONS. Journal of the American Statistical Association, 70, 320-328. doi: 10.1080/01621459.1975.10479865

Gorji, T., Sertel, E., Tanik, A., 2017. Recent Satellite Technologies for Soil Salinity Assessment with Special Focus on Mediterranean Countries. Fresenius Environmental Bulletin, 26, 196-203.

KHAN et al., 2005. Assessment of hydrosaline land degradation by using a simple approach of remote sensing indicators. Agricultural Water Management, 77, 96-109. doi: 10.1016/j.agwat.2004.09.038

Li, Z. et al., 2019. Spatial Prediction of Soil Salinity in a Semiarid Oasis: Environmental Sensitive Variable Selection and Model Comparison. Chinese Geographical Science, 29, 784-797.

doi: CNKI:SUN:ZDKX.0.2019-05-005

Lian, Y. et al., 2010. Quantitative Assessment of Impacts of Regional Climate and Human Activities on Saline-alkali Land Changes: A Case Study of Qian'an County, Jilin Province. Chinese Geographical Science, 20, 91-97. doi: 10.1007/s11769-010-0091-3

Liu, X., Xiao, C., Zhang, Y., Qiao, Y., 2015. Analysis on the Evolution Characteristics and Trend of the 52-year Precipitation Distribution in the West of Jilin Province. Water Resources and Power, 33, 11-14.

Meti, S., Lakshmi, P., Nagaraja, M., Shreepad, V., 2019. SENTINEL 2 AND LANDSAT-8 BANDS SENSITIVITY ANALYSIS FOR MAPPING OF ALKALINE SOIL IN NORTHERN DRY ZONE OF KARNATAKA, INDIA. ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XLII-3/W6, 307-313. doi: 10.5194/isprs-archives-XLII-3-W6-307-2019

Mingers, J., 1989. An Empirical Comparison of Pruning Methods for Decision Tree Induction. Machine Learning, 4, 227-243. doi: 10.1023/A:1022604100933

Nawar, S., Buddenbaum, H., Hill, J., 2015. Estimation of soil salinity using three quantitative methods based on visible and near-infrared reflectance spectroscopy: a case study from Egypt. Arabian Journal of Geosciences, 8, 5127-5140. doi: 10.1007/s12517-014-1580-y

Rasmussen, C.E., Williams, C. K. I., 2005. Gaussian Processes for Machine Learning. MIT Press. doi: 10.1142/S0129065704001899

Shrestha, R.P., 2006. Relating soil electrical conductivity to remote sensing and other soil properties for assessing soil salinity in northeast Thailand. Land Degradation & Development, 17, 677-689. doi: 10.1002/ldr.752

Srivastava, R. et al., 2015. Development of hyperspectral model for rapid monitoring of soil organic carbon under precision farming in the Indo-Gangetic Plains of Punjab, India. Journal of the Indian Society of Remote Sensing, 43, 751-759. doi: 10.1007/s12524-015-0458-0

Wang, J.Z. et al., 2019a. Capability of Sentinel-2 MSI data for monitoring and mapping of soil salinity in dry and wet seasons in the Ebinur Lake region, Xinjiang, China. Geoderma, 353, 172-187.

Wang, S.J., Chen, Y.H., Wang, M.G., Li, J., 2019b. Performance Comparison of Machine Learning Algorithms for Estimating the Soil Salinity of Salt-Affected Soil Using Field Spectral Data. Remote Sensing, 11, 26.

Wu, J. Z., Tian, Q. J., Ji, J.F., Chen, J., Hui, F. M., 2003. Theory, Method and Application of Soil Optical Remote Sensing. Remote Sensing Information, 01, 40-47+52(in Chinese).

Xu, S., Liang, H., Fu, S., Hu, Y., 2018. Variation characteristics of evaporation in Jilin province from 1951 to 2015. Journal of Meteorology and Environment, 34, 71-77(in Chinese).

Zhang, S., Zhao, G., 2019. A Harmonious Satellite-Unmanned Aerial Vehicle-Ground Measurement Inversion Method for Monitoring Salinity in Coastal Saline Soil. Remote Sensing, 11, 1700. doi: 10.3390/rs11141700