Water Depth Inversion based on Landsat-8 Date and Random Forest Algorithm

Jin Zhang¹,²,³*, Shujun Li²,³ and Mo Wang²,³

¹32023 Troop, Dalian, Liaoning, 116021, China
²Department of Military Oceanography and Hydrography & Cartography, Dalian Naval Academy, Dalian, Liaoning, 116018, China
³Key Laboratory of Hydrographic Surveying and Mapping of PLA, Dalian Naval Academy, Dalian, Liaoning, 116018, China

*Corresponding author’s e-mail: zjin_vic@163.com

Abstract. In view of the fact that the traditional water depth inversion model is susceptible to water quality and environmental factors, the water depth inversion model is constructed using the random forest algorithm using the high resolution remote sensing image of the Huangyan Island area and the corresponding measured water depth data. The influence of the number of training sets on the inversion of remote sensing water depth is explored. The results show that the inversion accuracy of random forest is higher, and for this method, the more features in the training set, the better the effect of remote sensing inversion.

1. Introduction.
Water depth is an important marine element in shallow sea areas such as coastal zones and islands, as well as one of the main contents of marine surveying and charting. Whether it is the construction of ports, docks, waterways, anchorages, or the navigation safety of ships, Marine scientific research, Marine engineering construction and Marine environment evaluation; water depth data is required as a basic guarantee. Compared with traditional side-scan sonar and multi-beam measurement methods, water depth remote sensing methods have the advantages of fast, large-scale, quasi-synchronous and high spatial resolution measurements. It can compensate for the insufficiency of measuring the water depth data of the vessel that cannot reach the sea area directly [1-3]. It is an indirect means of water depth measurement and a new technical method, which creates a new path for water depth exploration.

Machine learning is one of the most popular and active research branches in the field of computer science and technology [4]. In this paper, random forest is used for remote sensing water depth inversion [5]. Huangyan island is taken as the research area, using measured water depth data and Landsat8 OLI sensor L1GT level image data. Using machine learning methods model features and labels of the training set. The trained model is applied to the features of the test set to obtain the predicted tag value. Finally, the accuracy of the predicted label value and the label value in the test set are evaluated, and it was verified by the measured water depth value.
2. Data preparation.

2.1. Research area and data.

In this paper, the image is Landsat8 OLI sensor L1GT level data. The data obtained from the Geospatial Data Cloud Platform of the Computer Network Information Center of the Chinese Academy of Sciences. The strip number is 118, the line number is 49, and the spatial resolution is 30m. The imaging time is 2:29:31 on September 17, 2014 (GMT time), and the imaging area is Huangyan Island, as shown in Figure 1. The water depth data used in this paper is the measured water depth data, with a total of 11601 water depth points. After removing the water depth point which depth value is less than or equal to 0, the remaining 11416 water depth points have a water depth range of 0-30m. In this study, remote sensing image data needs to be pre-processed by radiometric calibration, Image cropping, atmospheric correction and noise removal. Since the measured water depth data is based on the depth reference plane, it is the data after the water level correction. Depth datum Since the measured water depth data is based on the Depth datum, it is the data after the water level correction. The water depth of the remote sensing image is the instantaneous water depth. Therefore, before the remote sensing water depth inversion, we need to perform the tide correction on the measured water depth data. Figure 2 is Preprocessed data.

![Image 1](image1.png)

Figure 1. The imaging is Huangyan Island

![Image 2](image2.png)

Figure 2. The preprocessed data

2.2. Stratified sampling.

Usually we do not directly establish a remote sensing water depth inversion model for the preprocessed data, but divide it into a training set and a test set. The training set participates in the establishment of the model, and the test set is used to verify the accuracy of the model. The amount of data in a typical training set is about 80% of the total number of data [6]. The commonly used partitioning methods are simple random sampling and stratified sampling. The simple random sampling method is simple and easy to implement, but the sampling results are not representative when the total amount of data is small. The stratified sampling method overcomes the above disadvantages, and it divides the population into several layers according to one or several characteristics (attributes), and then randomly samples from each layer. Although this sampling method is complicated, it can ensure that the extracted samples are distributed in each layer of the whole, and the statistical effect is remarkable. Therefore, this paper obtains 9132 training sets (shown in Table 1) and 2284 test sets (shown in Table 2) by stratified sampling from the matching of 11416 data at a ratio of 4:1.

| XR | YR | CA | B | G | R | NIR | SW1 | SW2 | XS | YS | DC |
|----|----|----|---|---|---|-----|-----|-----|----|----|----|

Table 1. The text file of Training set.
For the 9132 training sets, the tidal corrected water depth points cover the various water depths. Converting the UTM coordinates of the training set water depth to latitude and longitude, roughly obtaining the latitude and longitude distribution map of the following water depth points (as shown in Figure 3). For the 2284 test sets, the water depth points are superimposed on the remote sensing image. Figure 4 shows the overall distribution of the test water depth in the Huangyan Island area.

This section preprocesses the image data and measured water depth data of the Huangyan Island area. The stratified sampling method divides 11416 matching results into 9132 training sets and 2284 test sets. Whether it is a training set or a verification set, it has a certain representativeness to the sample. The stratified sampling method has a positive impact on the establishment of the subsequent remote sensing water depth inversion model.

### 2.3. Feature number processing.

Most machine learning methods are sensitive to the number of features, in order to explore the influence of the number of features of data sets on Machine Learning-Based Remote Sensing Water Depth Inversion, seeking the best machine learning method under hyperparameters. In this paper, 9132 training sets and 2284 test set data are divided into three training sets and corresponding test sets according to the number of factors, and then the water depth inversion experiment is carried out. Taking the training set as an example, the first training set (shown as Table 3) contains only the single factor characteristics of the band (CA, B, G, NIR, SW1, SW2), the second training set (shown as Table 4) adds the coordinate single factor feature (XR, YR) on the first one. The third training set (shown as Table 5) has the largest number of factors, and on the second basis, it adds a well-correlated two-factor feature (B*G, CA/B, G/CA, XR*G, YR/B, G/YB).

#### Table 3. Training set one.

| CA | B | G | R | NIR | SW1 | SW2 | DC |
|----|---|---|---|-----|-----|-----|----|
|    |   |   |   |     |     |     |    |

For the 9132 training sets, the tidal corrected water depth points cover the various water depths. Converting the UTM coordinates of the training set water depth to latitude and longitude, roughly obtaining the latitude and longitude distribution map of the following water depth points (as shown in Figure 3). For the 2284 test sets, the water depth points are superimposed on the remote sensing image. Figure 4 shows the overall distribution of the test water depth in the Huangyan Island area.

This section preprocesses the image data and measured water depth data of the Huangyan Island area. The stratified sampling method divides 11416 matching results into 9132 training sets and 2284 test sets. Whether it is a training set or a verification set, it has a certain representativeness to the sample. The stratified sampling method has a positive impact on the establishment of the subsequent remote sensing water depth inversion model.

### 2.3. Feature number processing.

Most machine learning methods are sensitive to the number of features, in order to explore the influence of the number of features of data sets on Machine Learning-Based Remote Sensing Water Depth Inversion, seeking the best machine learning method under hyperparameters. In this paper, 9132 training sets and 2284 test set data are divided into three training sets and corresponding test sets according to the number of factors, and then the water depth inversion experiment is carried out. Taking the training set as an example, the first training set (shown as Table 3) contains only the single factor characteristics of the band (CA, B, G, NIR, SW1, SW2), the second training set (shown as Table 4) adds the coordinate single factor feature (XR, YR) on the first one. The third training set (shown as Table 5) has the largest number of factors, and on the second basis, it adds a well-correlated two-factor feature (B*G, CA/B, G/CA, XR*G, YR/B, G/YB).

#### Table 3. Training set one.

| CA | B | G | R | NIR | SW1 | SW2 | DC |
|----|---|---|---|-----|-----|-----|----|
|    |   |   |   |     |     |     |    |
Table 4. Training set two.

| XR   | YR   | CA | B   | G   | R   | NIR | SW1 | SW2 | DC   |
|------|------|----|-----|-----|-----|-----|-----|-----|------|
| 586620 | 1671120 | 806 | 738 | 352 | 62  | 37  | 31  | 28  | 10.20 |
| 581430 | 1678110 | 894 | 837 | 591 | 177 | 139 | 135 | 99  | 9.30  |
| 589200 | 1671390 | 1007 | 972 | 445 | 31  | 14  | 22  | 9   | 11.00 |

Table 5. Training set three.

| XR   | YR   | CA | B   | G   | R   | NIR | SW1 | SW2 | B*   | G   | CA/B | G/CA | XR*G | YR/B | G/YR | DC   |
|------|------|----|-----|-----|-----|-----|-----|-----|------|-----|------|------|------|------|------|------|
| 5866  | 16711 | 806 | 738 | 352 | 62  | 37  | 31  | 28  | 5976  | 1.09 | 0.44 | 206490 | 240  | 2264.39 | 0.0002 | 10.20 |
| 5814  | 16781 | 894 | 837 | 591 | 177 | 139 | 135 | 99  | 4946  | 1.04 | 0.66 | 343625 | 130  | 2004.91 | 0.0003 | 9.30  |
| 5892  | 16713 | 972 | 445 | 31  | 14  | 22  | 9   | 4325 | 1.04  | 0.44 | 262194 | 000  | 1719.54 | 0.0003 | 11.00 |

3. Modeling and Experiment.

Machine learning is one of the most popular and active research branches in the field of computer science and technology. Random Forest is a kind of Ensemble learning [7]. It uses decision tree as the base learner, uses Bagging method to homogenize multiple base learners, and further introduces random feature selection in decision tree training process [8]. This way of ensemble learning is often better than data for a single best learner. Therefore, based on the decision tree, this paper uses the method of random forest to carry out remote sensing water depth inversion.

3.1. Decision tree.

The decision tree is a powerful machine learning algorithm that can handle classification, regression, and even multiple output problems, and can fit well to complex data sets. The decision tree algorithm mainly includes three steps: feature selection, decision tree generation and decision tree pruning. Commonly used decision tree algorithms are ID3, C4.5, and CART. Among them, the CART is the most widely used decision tree algorithm. For the classification problem, it uses the Gini Index to select features.

The Gini index can be expressed as:

\[ G_i = 1 - \sum_{k=1}^{n} P_{i,k}^2 \]  \tag{1}

\( P_{i,k} \) is the ratio of samples, \( k \) is the training samples, \( i \) is the node category.

For the regression problem, the loss function of CART decision tree algorithm is defined as follows:

\[ J(k, t_k) = \frac{m_{\text{right}}}{m} \text{MSE}_{\text{right}} + \frac{m_{\text{left}}}{m} \text{MSE}_{\text{right}} \]  \tag{2}

\( t \) is the threshold of the characteristic \( k \).

\[ \text{MSE}_{\text{node}} = \sum_{i=\text{node}}^{\hat{y}_i - y^{(i)}}^2, \quad \hat{y}_i = \frac{1}{m_{\text{node}}} \sum_{i=\text{node}} y^{(i)} \]  \tag{3}

In this paper, the CART decision tree algorithm is used to perform remote sensing water depth inversion. There is no upper limit on the depth of the tree and the number of leaf nodes. The minimum value of divided samples number in the internal node is set to 2, the minimum value of divided samples number in the leaf node is set to 1, the maximum number of features is set to the total number of features.
in the corresponding training set, and the threshold of the impurity reduction of the node sample is set to 0. The weighted Impurity Decrease expression is as follows:

\[
WID = \frac{N_t}{N} \left( \text{impurity} - \frac{N_{\text{right}}}{N_t} \cdot \text{impurity}_{\text{right}} - \frac{N_{\text{left}}}{N_t} \cdot \text{impurity}_{\text{left}} \right)
\]

(4)

\(N\) is the total number of samples, \(N_t\) is the number of samples in the current node, and \(N_{\text{left}}\) and \(N_{\text{right}}\) are the number of samples in the left and right child nodes.

3.2. Random Forest.
Random Forest is a kind of Ensemble Learning. It uses decision tree as the base learner and uses Bagging method to homogenize multiple base learners, and further introduces Random feature selection in the process of decision tree training. This way of ensemble learning is often better than data for a single best learner [9].

3.2.1. Bagging.
The Bagging method is an integrated method based on Bootstrap Sampling [10]. The basic idea is to use the self-sampling method to sample \(k\) samples containing \(m\) training samples, train a base learner based on each sample set, and finally combine these base learners. This method of feature selection leads to the diversity of spanning trees, but it also gets smaller deviations, which leads to a better model. In general, the random forest method is very convenient and fast for selecting the features in the data set. It is simple to implement and has small computational time complexity. It has been applied in many practical problems and shows strong performance. It is a very Excellent machine learning algorithm.

3.3. Experiment.
When using the random forest method for water depth inversion, the hyperparameters need to be set more than the decision tree. Therefore, this paper uses Grid Search to find the optimal hyperparameters. Grid search is actually exhausting all the possible values of each parameter, and then evaluating all the parameter combinations with certain evaluation criteria, and finally selecting the process with the best performance set of hyperparameters.

3.3.1. The first Training set.
For the first training set, the grid search takes 2 minutes and 16 seconds. The results are shown in Table 6. As you can see, the best hyperparameter combination is \{max_features: 4, n_estimators: 30, bootstrap: True\}, which has an RMSE of 1.96402m on the training set.

| Max_features | N_estimators | Bootstrap | RMSE  |
|--------------|--------------|-----------|-------|
| 2            | 3            | True      | 2.27708 |
| 2            | 10           | True      | 2.05065 |
| 2            | 30           | True      | 1.98243 |
| 4            | 3            | True      | 2.20234 |
| 4            | 10           | True      | 2.02812 |
| 4            | 30           | True      | 1.96402 |
| 6            | 3            | True      | 2.19394 |
| 6            | 10           | True      | 2.01457 |
| 6            | 30           | True      | 1.96779 |
| 2            | 3            | False     | 2.28340 |
| 2            | 10           | False     | 2.07854 |
3.3.2. The second Training set.
For the second training set, the grid search takes 2 minutes and 20 seconds. The results are shown in Table 7. The best hyperparameter combination is \{max\_features: 8, n\_estimators: 30, bootstrap: True\} which has an RMSE of 1.89519m, and the RMSE error is reduced compared to the first training set.

Table 7. Root mean square error of different parameter combination on training set two.

| Max\_features | N\_estimators | Bootstrap | RMSE   |
|---------------|----------------|-----------|--------|
| 2             | 3              | True      | 2.19875|
| 2             | 10             | True      | 1.98258|
| 2             | 30             | True      | 1.91174|
| 4             | 3              | True      | 2.15930|
| 4             | 10             | True      | 1.96036|
| 4             | 30             | True      | 1.91209|
| 6             | 3              | True      | 2.12933|
| 6             | 10             | True      | 1.94526|
| 6             | 30             | True      | 1.89917|
| 8             | 3              | True      | 2.15104|
| 8             | 10             | True      | 1.95677|
| 8             | 30             | True      | **1.89519**|
| 2             | 3              | False     | 2.15900|
| 2             | 10             | False     | 2.00583|
| 3             | 3              | False     | 2.15243|
| 3             | 10             | False     | 2.01526|
| 4             | 3              | False     | 2.16042|
| 4             | 10             | False     | 1.98881|

3.3.3. The third Training set.
For the third training set, the grid search takes 2 minutes and 21 seconds. The results are shown in Table 8. The best hyperparameter combination is \{max\_features: 8, n\_estimators: 30, bootstrap: True\} which has an RMSE of 1.86675m, and the RMSE error is the smallest among the three training set data.

Table 8. Root mean square error of different parameter combination on training set three.

| Max\_features | N\_estimators | Bootstrap | RMSE   |
|---------------|----------------|-----------|--------|
| 2             | 3              | True      | 2.15558|
| 2             | 10             | True      | 1.96517|
| 2             | 30             | True      | 1.92504|
| 4             | 3              | True      | 2.15073|
| 4             | 10             | True      | 1.94526|
| 4             | 30             | True      | 1.87169|
| 6             | 3              | True      | 2.09133|
| 6             | 10             | True      | 1.95971|
| 6             | 30             | True      | **1.86675**|
|   |   |   |   |
|---|---|---|---|
| 3 | 8 | True | 2.09826 |
| 3 | 8 | True | 1.92218 |
| 30 | 8 | True | 1.87022 |
| 3 | 10 | True | 2.09197 |
| 10 | 10 | True | 1.93242 |
| 30 | 10 | True | 1.86993 |
| 3 | 12 | True | 2.14376 |
| 10 | 12 | True | 1.93974 |
| 30 | 12 | True | 1.87957 |
| 3 | 2 | False | 2.19032 |
| 3 | 2 | False | 1.9973 |
| 10 | 3 | False | 2.15824 |
| 3 | 10 | False | 1.99105 |
| 3 | 4 | False | 2.14642 |
| 10 | 4 | False | 1.99321 |

3.4. Analysis.
Comparing the three tables, the other parameters are the same, we found that the RMSE on the training set not obtained by the self-sampling method is higher than the self-sampling method. For the first training set and the third training set, it is not that the more features are used, the smaller the error will be. The error of the water depth inversion of the random forest method, whether MAE or RMSE, is significantly reduced compared to the previous model or method. Among them, the accuracy of the random forest method trained by the third training set is the highest in the three training sets, and the MAE is 1.19937m. The water depth deducted by this method has certain practical value.

4. Conclusion
Based on remote sensing images and measured water depth data, this paper uses the random forest model to invert the water depth in Huangyan Island and compare it with the traditional model. The results show that the accuracy of the random forest water depth inversion model is better than that of the traditional water depth model, and the water depth deducted by this method has certain practical value. Compared with traditional method, it can effectively and quickly obtain large-scale underwater terrain information, and provides an efficient and scientific reference for the remote sensing bathymetry.

References

[1] Cheng Peng. Bathymetry Retrieval from Hyperion Data in Shallow water[D]. University of Chinese Academy of Science(The Institute of Remote Sensing and Digital Earth), 2017.

[2] Yong Kai. Research on Feature Selection and Model Optimization of Random Forest[D]. Harbin Institute of Technology, 2008.

[3] Lei Zhen. Random Forest and Its Application in Remote Sensing[D]. Shanghai Jiao Tong University, 2012.

[4] Kogut, Weistock. Classifying airborne bathymetry data using the Random Forest algorithm[J]. Remote Sensing Letters,2019,10(9).

[5] Qiu Yaowei, Shen Wei, Hui Xiao. Water Depth Inversion Based on WorldView-2 Date and Random Forest Algorithm [j], Remote Sensing Information, 2019, 34(02): 75-79.

[6] Cao Zhengfeng. Study on Optimization of Random Forest algorithm[D]. Capital University of Economics and Business, 2014.

[7] Wang Cheng, Wang Kai. An improved Random Forest Algorithm based on Decision Trees Clustering Reduction[J]. Journal of Nanjing University of Posts and Telecommunications (Natural Science Edition),2019, 39(03):91-97.

[8] Science - Applied Geoscience; New Applied Geoscience Findings Has Been Reported by Investigators at University of Canterbury (Effects of Pre-processing Methods On Landsat Oli-
8 Land Cover Classification Using Obia and Random Forests Classifier) [J]. Science Letter, 2019.

[9] Andrei DORNIK, Lucian DRAGUT, Petru URDEA. Classification of Soil Types Using Geographic Object-Based Image Analysis and Random Forests[J]. Pedosphere, 2018, 28(06):913-925.

[10] Ma Jingyi, Xie Bangchang. A Comparison on Random Forest and Bagging Classification Tree in Classification[J]. Statistics and Information Forum, 2010, 25(10):18-22.