Journal of Science and Technology in Civil Engineering, HUCE, 2023, 17 (4): 122–136

APPLICATION OF THE BAYESIAN MODEL AVERAGING ALGORITHM IN EVALUATING AND SELECTING OPTIMAL SALINITY PREDICTION MODELS

Bui Duy Quynh[a,*], Ha Thi Hang[a], Duong Cong Hieu[b], Tran Xuan Truong[c], Luu Thi Dieu Chinh[d]

[a]*Faculty of Bridges and Roads, Hanoi University of Civil Engineering, 55 Giai Phong road, Hai Ba Trung district, Hanoi, Vietnam*
[b]*Institute of Geodesy Engineering Technology, Hanoi University of Civil Engineering, 55 Giai Phong road, Hai Ba Trung district, Hanoi, Vietnam*
[c]*Faculty of Geodesy – Cartography and Land Management, Hanoi University of Mining and Geology, 18 Vien street, Bac Tu Liem district, Hanoi, Vietnam*
[d]*Faculty of Hydraulic Engineering, Hanoi University of Civil Engineering, 55 Giai Phong road, Hai Ba Trung district, Hanoi, Vietnam*

Article history:
Received 14/9/2023, Revised 13/12/2023, Accepted 14/12/2023

Abstract

Salinity intrusion poses significant challenges to coastal regions worldwide. Reliable salinity prediction models can provide valuable information to mitigate the impact and influence of salinity intrusion. However, their accuracy relies mainly on the selected input variables and used optimal models. This study employs the Bayesian Model Averaging (BMA) algorithm to evaluate input variable importance and select the most reliable salinity prediction model. Based on an analysis of observed salinity data and climate data extracted from Landsat 8 OLI in the Google Earth Engine platform, the BMA algorithm identifies the significance of critical variables and optimal salinity prediction models. Various statistical metrics, including R-squared, Root Mean Square Error (RMSE), and Mean Absolute Error (MAE) from the Random Forest method, were used to verify the performance of these optimal salinity prediction models. These obtained results offer foundational knowledge and valuable insights for future studies in determining appropriate input variables and selecting the best optimal salinity prediction model.

Keywords: variable importance; salinity prediction model; Bayesian model averaging; Landsat 8; Google Earth Engine; Mekong delta.

https://doi.org/10.31814/stce.huce2023-17(4)-10

1. Introduction

Salinity intrusion is a natural phenomenon that occurs when the salinity (salt concentration) of the water in a freshwater system increases due to the intrusion of seawater or saline groundwater. Salinity intrusion in East and South Asia is a critical issue that directly and indirectly contributes to water insecurity, adversely affecting livelihoods, agricultural production, and social dynamics. The Food and Agriculture Organization (FAO) estimated that saline soils currently cover approximately 397 million hectares of land worldwide, and these lands are projected to expand by an additional 2 million hectares annually [1].

The Mekong Delta plays a crucial role in the socio-economic development of Vietnam. Its robust agricultural sector fosters economic growth and supports millions of livelihoods in rural areas.

*Corresponding author. *E-mail address:* quynhbd@huce.edu.vn (Quynh, B. D.)

In addition, it also plays an important role in facilitating trade and commerce between Vietnam and other nations. In recent years, under the impact of the salinity intrusion, the Mekong Delta has faced thousands of crops and agricultural land hectares that will be saltwater, threatening the livelihood of local communities and the food security of Vietnam directly. Therefore, it is essential to properly preserve and manage strategies in the Mekong Delta to reduce the impact of salinity intrusion and ensure sustainable development for this region. The salinity prediction models enable to capture spatial distributions in salinity, help identify vulnerable areas, and prioritize resources for mitigation efforts. The accuracy of these forecast models depends mainly on the importance of input variables and the use of optimal prediction models [2].

Salinity intrusion is often influenced by many factors that interact in complex ways. Sea-level rise due to climate change is one of the most crucial reasons for seawater penetrating further inland. Human activities, such as dam construction and excessive groundwater extraction, the topography and geomorphology of coastal regions, land-use changes, and the urbanization process can also increase the procedure of salinity intrusion [3]. Thus, capturing and understanding the interaction of these diverse factors to salinity intrusion is essential to building and developing effective strategies to mitigate the impacts of salinity intrusion [4]. Multi-spectral satellite image data can be extracted from the Google Earth Engine platform for salinity factors to serve salinity prediction models efficiently and economically. Landsat 8 OLI in the Google Earth Engine (GEE) platform is a critical satellite image source offering many advantages in exploiting salinity indicators and understanding salinity intrusion processes. Its high spatial resolution facilitates the identification of changes in coastal environments, which is crucial for monitoring salinity intrusion status over time. Many studies have utilized the salinity indicators extracted from the Landsat 8 OLI in the GEE platform to build soil salinity maps [5, 6]. The development of machine learning provides a potential approach for understanding and predicting salinity intrusion processes. By analyzing big volumes of complex data from various sources, such as satellite imagery, hydrological measurements, and climate data, these algorithms can discover the importance of input variables and hidden relationships between these variables and salinity intrusion.

Bayesian Model Averaging (BMA) algorithm is a technique that uses Bayesian models to integrate information from many different models and generate a final prediction. The main idea of BMA is to combine predictions from multiple models instead of relying on a single model [7]. In this way, the BMA is able to minimize the effects of errors and uncertainties in a single model. BMA algorithm also permits users and researchers to capture the importance of various models and their parameters, leading to more reliable and robust results. The BMA algorithm was used as a single model to manage saltwater intrusion in the "1,500-foot" sand aquifer in the Baton Rouge area, Louisiana [8], or to build top-soil salinity maps for three coastal districts of Ben Tre province in Vietnam [9]. BMA was also combined with chance-constrained (CC) programming to build a BMA-CC framework to design a hydraulic barrier to protect public supply wells of the Government St. pump station from saltwater intrusion in the "1500-foot" sand and the "1700-foot" sand of the Baton Rouge area, southeastern Louisiana [10], or it also was employed using the combination of the Boruta–artificial neural network (B-ANN) and the Boruta–support vector regression (B-SVR) models to predict long-term streamflow of the Volga River [11]. Most of these studies have indicated the number of input variables having a crucial role in assessing the salinity forecast models, and the used ML algorithms in these studies have been applied based on scientists'experiences. The number of input variables and optimal salinity prediction models need to be decided based on their importance assessments and the application of robust ML algorithms [12]. In Vietnam, some studies have applied the Landsat-8 OLI data in the GEE platform and ML algorithms in predicting the salinity intrusion in the Mekong Delta in recent years

[13–16]. However, these studies have not used the BMA algorithm, evaluated the importance of input variables, or mentioned the number of optimal salinity prediction models.

Salinity intrusion is a complex process influenced by various factors. In regions where salinity intrusion models may change over time, such as the Mekong Delta, the BMA algorithm can offer a powerful tool to solve the uncertainties and adapt to new data due to availability. Therefore, BMA is used in this study to select the best models that reflect the potential relationships between salinity intrusion and input variables. We first evaluate the importance of salinity variables extracted from the Landsat 8 OLI in the GEE platform and then provide optimal salinity prediction models.

In this article, the first section provides an overview of the study. Section 2 presents information on the materials used and the methods employed. Results and corresponding discussions are presented in Section 3. Finally, the conclusion is shown in Section 4.

2. Materials and Methods

2.1. Study area

The Mekong Delta is the region in the south of Vietnam where the Mekong River approaches and empties into the sea through a network of distributaries. The Mekong Delta borders Cambodia in the North, Ho Chi Minh City to the Northeast, Thailand Bay to the Southwest, and the East Sea to the East and Southeast. With a natural area of 39,712 square kilometers, the Mekong Delta includes 13 provinces and cities: An Giang, Kien Giang, Tien Giang, Hau Giang, Ben Tre, Bac Lieu, Ca Mau, Can Tho, Long An, Dong Thap, Tra Vinh, Soc Trang, and Vinh Long (Fig. 1).

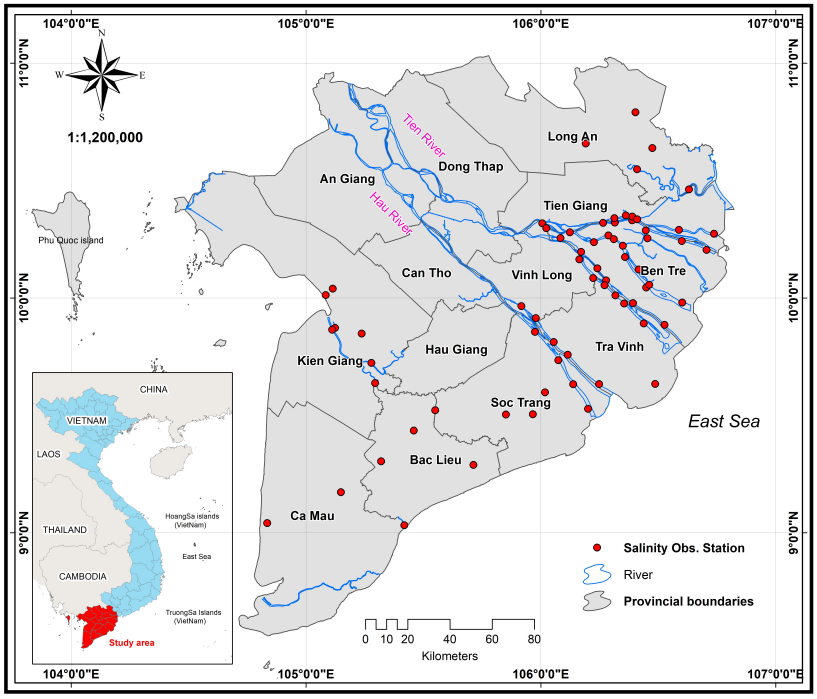

Figure 1. The study area and salinity monitoring stations

The Mekong Delta is a key economic region in Vietnam. With an average elevation of less than 1.5 m and a size of about 41,000 square kilometers, the Mekong Delta produces more than 50% of the rice and more than 65% of the seafood in Vietnam. However, the Mekong Delta has faced significant climate and environmental changes in recent decades. These challenges come from global climate

change [17] and human activities living in the Mekong Delta [18] or at the Mekong Basin level [19] leading to the Mekong Delta being threatened by storms, erosion, floods, land subsidence, and saline intrusion [20]. The salinity issue poses significant challenges to humans, agriculture, livelihood, freshwater availability, and the delicate ecosystem of the Mekong Delta.

2.2. Data used

In this study, in-situ salinity data during the day (2 hours/time) were collected from 68 salinity monitoring stations that belong to the management of the Department of Agriculture and Rural Development and the Hydrometeorological Center of Southern provinces. In addition, 67 Landsat 8 OLI satellite images with 1T levels with the 30 m spatial resolution were acquired in 6 months, from January 1^{st}, 2020, to June 30^{th}, 2020. These images were corrected by irradiance measurement, topographic correction, and map projection registration in the Google Earth Engine platform. The spatial resolution of used bands in Landsat 8 OLI satellite image is represented in Table 1.

Table 1. Description of the spatial resolution of bands in Landsat 8 imagery

Spectral bands	Wavelength (micrometers)	Spatial resolution (meters)	Repeat cycle (days)	Sensor
Band 1 (Coastal aerosol)	0.43-0.45	30	16	OLI
Band 2 (Blue)	0.45-0.51	30	16	OLI
Band 3 (Green)	0.53-0.59	30	16	OLI
Band 4 (Red)	0.64-0.67	30	16	OLI
Band 5 (Near Infrared-NIR)	0.85-0.88	30	16	OLI
Band 6 (Shortwave Infrared-SWIR1)	1.57-1.65	30	16	OLI
Band 7 (Shortwave Infrared-SWIR2)	2.11-2.29	30	16	OLI
Band 8 (Panchromatic)	0.50-0.68	15	16	OLI
Band 9 (Cirrus)	1.36-1.38	30	16	OLI
Band 10 (Thermal Infrared-TIRS1)	10.6-11.19	100	16	TIRS
Band 11 (Thermal Infrared-TIRS2)	11.50-12.51	100	16	TIRS

Table 2. Collected data in this study

No	Data	Sources	Year	Data format and spatial resolution
1	In-situ salinity data during the day (2 hours/time)	68 salinity monitoring stations - Department of Agriculture and Rural Development, Hydrometeorological Center of Southern provinces	January 1st, 2020 to June 30th, 2020	Point
2	67 Landsat 8 OLI satellite images with 1T levels	Google Earth Engine	January 1st, 2020 to June 30th, 2020	Grid, 30×30 m
3	MERIT Digital Elevation Model	MERIT DEM [21]	2018	Grid, 90×90 m

This study used 337 in-situ salinity data that were collected from 68 monitoring stations in the 6 first months of 2020, from January 1 to June 30, 2020. In this study, input data were collected from different resources and in different formats (Table 2).

To determine the relationship of salinity with spectral channels of satellite images, twenty-eight salinity factors were extracted from Landsat 8 OLI, including five bands from band 1 to band 5, five principal component analyses of band 1, 2, 3, 4, 5, and eighteen ratio bands (Table 3).

Table 3. Factors extracted from Landsat 8 imagery

Spectral bands/Ratio bands	Formula	Sources
Coastal Aerosol, Blue, Green, Red, NIR bands	B1, B2, B3, B4, B5	[22]
PCA of bands 1, 2, 3, 4, 5 (Principal Component Analyses)	PCA1, PCA2, PCA3, PCA4, PCA5	[23]
NDVI (Normal Difference Vegetation Index)	$NDVI = \frac{NIR - R}{NIR + R}$	[24]
NDSI (Normalized Difference Salinity Index)	$NDSI = \frac{R - NIR}{R + NIR}$	[24]
NDWI (Normalized Difference Water Index)	$NDWI = \frac{G - NIR}{G + NIR}$	[25]
ND47 (Normalized Difference between TM4 and TM 7)	$ND47 = \frac{NIR - SWIR2}{NIR + SWIR2}$	[26]
NDMI (Normalized Difference Moisture Index)	$NDMI = \frac{NIR - SWIR1}{NIR + SWIR1}$	[25]
COSRI (Combined Spectral Response Index)	$COSRI = NDVI\frac{B + G}{R + NIR}$	[27]
CRSI (Canopy Response Salinity Index)	$CRSI = \sqrt{\frac{(NIR \times R) - (G \times B)}{(NIR \times R) + (G \times B)}}$	[28]
MSI (Moisture Stress Index)	$MSI = \frac{SWIR1}{NIR}$	[29]
EVI (Enhanced Vegetation Index)	$EVI = g \times \frac{NIR - R}{(NIR + c1 \times R - c2 \times B + l)}$	[30]
VSSI (Vegetation Soil Salinity Index)	$VSSI = 2 \times G - 5 \times (R + NIR)$	[31]
SI1	$SI1 = \sqrt{G^2 + R^2}$	[32]
SI2	$SI2 = \sqrt{G \times R}$	[33]
SI3	$SI3 = \sqrt{B \times R}$	[34]
SI4	$SI4 = \frac{R \times NIR}{G}$	[35]
SI5	$SI5 = \frac{B}{R}$	[36]
SI6	$SI6 = \frac{B - R}{B + R}$	[36]
SI7	$SI7 = \frac{G \times R}{B}$	[36]
SI8	$SI8 = \frac{B \times R}{G}$	[35]
SI9	$SI9 = \frac{NIR \times R}{G}$	[35]

Note: R (Red band), G (Green band), B (Blue band), NIR (Near-Infrared band), SWIR (Short-Wave infrared), SI (Salinity Index)), g = 2.5, c1 = 6.0, c2 = 7.5, l = 1.0 [37].

After calculating in the GEE platform, these salinity factors would be used to update the salinity data for each monitoring station. Because the Landsat 8 images were taken at 10 am, the salinity values at the monitoring stations would be taken at the same time as the satellite images were taken for analysis.

In this study, the location (Longitude, Latitude), altitude, and monitoring time of the salinity monitoring stations were added as input variables to analyze and select optimal salinity prediction models. On the other hand, because the Landsat 8 OLI image data are affected by clouds, scenes covered by clouds will be removed. As a result, time variables were acquired and coded for 23 days as follows: x06Th01, x13Th01, x15Th01, x22Th01, x07Th02, x14Th02, x23Th02, x01Th03, x03Th03, x10Th03, x17Th03, x19Th03, x26Th03, x04Th04, x11Th04, x18Th04, x27Th04, x04Th05, x13Th05, x29Th05, x07Th06, x14Th06, and x21Th06. Therefore, in this study, there will be a total of 54 variables as input data to evaluate the importance as well as serve for the selection of optimal salinity prediction models, including 28 factors extracted from Landsat 8 OLI imagery, 23-time variables, and 3 variables representing the coordinate location and the elevation

2.3. Methodology

In this study, the in-situ salinity data were collected from 68 salinity monitoring stations together with the Landsat 8 OLI satellite images exploited from the Google Earth Engine platform from January 1st, 2020, to June 30th, 2020. First, the correlation relationship between the in-situ salinity data was checked to ensure their positive correlation and accuracy. Second, 28 climate data were extracted from the Landsat 8 OLI satellite images in the Google Earth Engine platform. The location (Longitude, Latitude), elevation, and time variables were also utilized to assess. A total of 54 factors were selected to assess the importance and serve for selecting optimal salinity prediction models. Third, all these factors were used as input data for the Bayesian Model Averaging model to estimate the variable importance and select optimal salinity prediction models. Finally, various statistical indexes consisted of R-squared (R^2), Root Mean Square Error (RMSE), and Mean Absolute Error (MAE) exploited from the Random Forest method, which was used to verify the fit to the input data of the models as well as to confirm the performance of these optimal salinity prediction models. The research methodology is presented in Fig. 2.

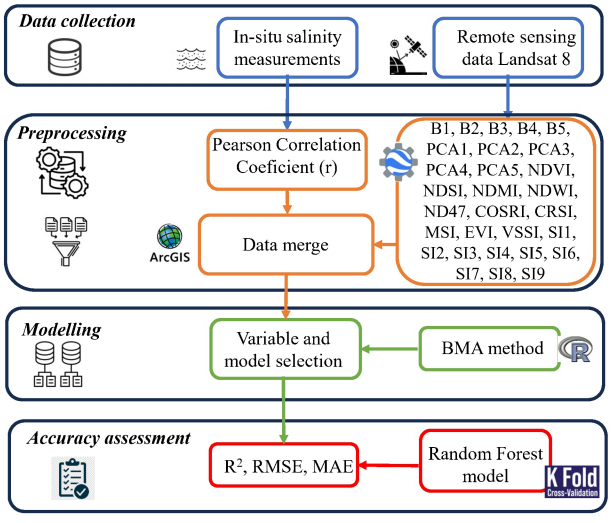

Figure 2. The methodology used in this study

2.4. Methods used

a. Pearson correlation coefficient

The Pearson correlation coefficient is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables [38]. This coefficient often ranges from −1 to +1, where −1 indicates a perfect negative correlation, +1 reflects a perfect positive correlation, and 0 denotes no linear correlation [39]. In this study, in-situ salinity data were collected from different salinity monitoring stations and at different estuaries, so the Pearson correlation coefficient was used to assess the correlation relationship between in-situ salinity data. The variables used to evaluate are in-situ salinity data and monitoring station locations. This correlation coefficient test can help understand and confirm spatial salinity variations [40].

b. Google Earth Engine platform

Google Earth Engine (GEE) is an advanced cloud computing platform that helps to preprocess and analyze satellite imagery and other geospatial data. Google Earth Engine platform has a vast and up-to-date repository of satellite imagery from various sources, allowing users to access historical and current data for comprehensive analysis. Additionally, this platform provides a wide range of built-in analysis tools and algorithms, simplifying complex geospatial computations and enabling users to focus on their research instead of coding at the beginning step [41]. In this study, Landsat 8 OLI image data were exploited from the GEE platform, and the salinity factors extracted from Landsat 8 OLI were also calculated in this platform (Fig. 3).

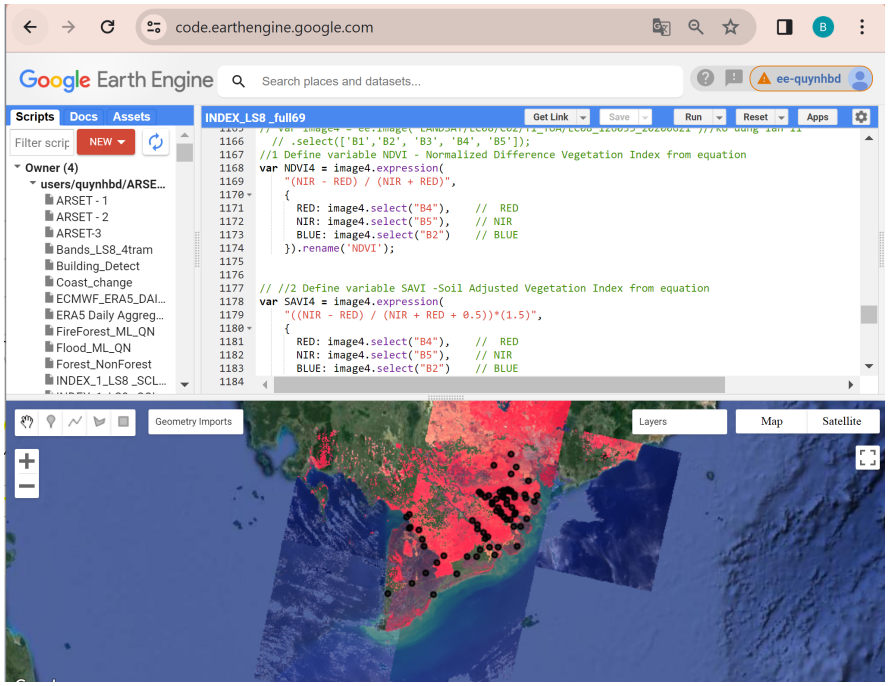

Figure 3. The salinity factors extracted from Landsat 8 were calculated in the GEE platform

After calculating in the GEE platform, these salinity factors would be used to update the salinity data for each monitoring station. Because the Landsat 8 images were taken at 10 am, the salinity values at the monitoring stations would be taken at the same time as the satellite images were taken for analysis.

c. Bayesian Model Averaging (BMA) algorithm

Bayesian Model Averaging (BMA) is a statistical algorithm used for model selection and prediction in the context of Bayesian statistics [42]. This algorithm is particularly useful when dealing with multiple competing models, each of which may have different strengths and weaknesses in explaining the observed data. In this study, the BMA technique was applied to evaluate the input data's importance and select optimal salinity prediction models. Mathematically, this algorithm can be described briefly as follows [42]:

1. Let a set of n models $M = M_1, M_2, \ldots, M_n$. Each model M_i is related to a set of parameters θ_i with $i = 1, 2, \ldots, n$;

2. Assign prior probabilities P to each model M_i, denoted as $P(M_i)$. The likelihood of observing the data D for each model M_i, denoted as $P(D|M_i, \theta_i)$. This index quantifies the fit level of each model to the observed data;

3. The posterior probability of each model M_i given the observed data D, denoted as $P(M_i|D)$ and it is calculated by using the following equation:

$$P(M_i|D) = \frac{P(D|M_i)\,P(M_i)}{\sum_{i=1}^{n}(P(D|M_i)\,P(M_i))} \tag{1}$$

4. A quantity Δ is present in all models, such as a covariate or a future observation. Then, using Bayes' theorem to calculate the posterior probability of a quantity Δ as below:

$$P(\Delta|D) = \sum_{i=1}^{n} P(\Delta|M_i, D)\,P(M_i|D) \tag{2}$$

where $P(\Delta|M_i, D)$ is the posterior probability of Δ with the known model M_i; $P(M_i|D)$ is the posterior probability of the model M_i with the observed data D.

This algorithm can build prediction models efficiently, and each prediction model has a predetermined probability. This method can identify variables that have a close relationship with the outcomes based on the actual data [42].

d. Random Forest method

Random Forest is an ensemble learning method that combines multiple decision trees [43]. This technique can provide a more robust and accurate verification of forecast models by reducing the variance and potential errors associated with individual trees. This technique also supplies feature importance measures to enhance the verification process of prediction models and ensure more accurate and reliable predictions [44]. Thus, in this study, various statistical metrics consisted of determination coefficient (R^2), Root Mean Square Error (RMSE), and Mean Absolute Error (MAE) from the Random Forest method, which was used to verify the performance of these optimal salinity prediction models. These indicators were calculated by using the following equations:

$$R^2 = \frac{\sum_{i=1}^{n}(y_i^p - \overline{y^p}) \times \left(y_i^m - \overline{y^m}\right)}{\sqrt{\sum_{i=1}^{n}\left(y_i^p - \overline{y^p}\right)^2 \times \left(y_i^m - \overline{y^m}\right)^2}} \tag{3}$$

$$RMSE = \sqrt{\frac{1}{n} * \sum_{i=1}^{n}\left(y_i^p - y_i^m\right)^2} \tag{4}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i^p - y_i^m\right| \quad (5)$$

where y_i^p and y_i^m are salinity values received from models and from monitoring stations of sample i, respectively; $\overline{y^p}$ and $\overline{y^m}$ are mean values received from prediction models and from monitoring stations; n is the total sample.

Normally, a higher R^2 value indicates a better fit of the model to the input data; meanwhile, lower RMSE and MAE values reflect better model performance.

In this study, the ML models including BMA and Random Forest algorithms were run in the R software because it is a popular selection to address the problems of statistical computing and data analysis. Moreover, the R software can supply a flexible analysis environment, a wide range of available libraries, and high accuracy.

3. Results and discussions

3.1. Correlation relationship between in-situ salinity data

In this study, the in-situ salinity data from 68 salinity monitoring stations were checked for the correlation relationship based on the Pearson correlation coefficient. Since in-situ salinity data were collected from different salinity monitoring stations and at different estuaries, the Pearson correlation coefficient was used to assess the correlation relationship between in-situ salinity data. The variables used to evaluate are in-situ salinity data and monitoring station locations. The calculated results indicated that the Pearson correlation coefficients of these data ranged from 0.756 to 0.971 (Fig. 4). These obtained values reflected the good positive correlation between these in-situ salinity data and ensured the objectivity of the input data.

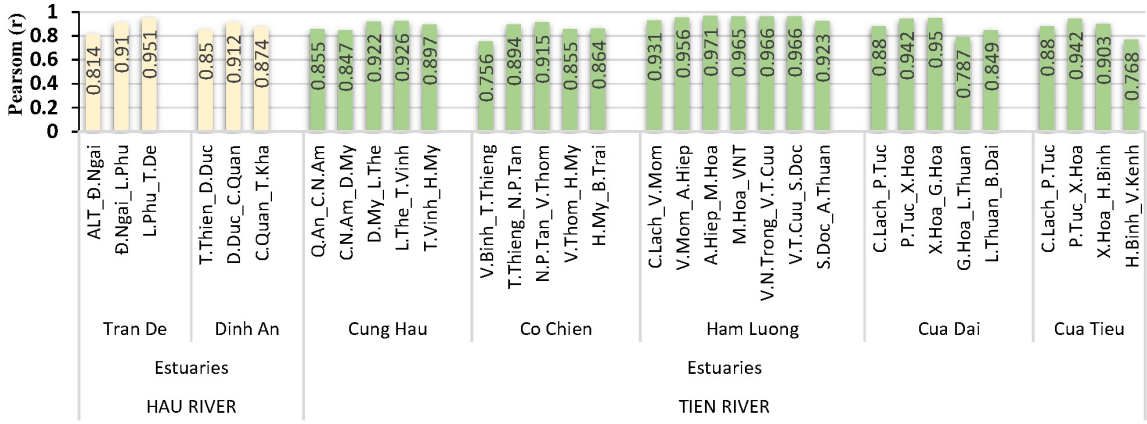

Figure 4. Testing of correlation relationship between in-situ salinity data

3.2. Measuring variable importance and selecting optimal salinity prediction models

There are a total of 54 input variables, including 28 factors extracted from Landsat 8 OLI imagery, 23-time variables, and 3 variables representing the coordinate location and the elevation. We ran the BMA package in R for selecting the optimal salinity prediction models.

The obtained results are presented in Fig. 5, in which the vertical axis denotes the number of used variables, and the horizontal axis describes the number of received models. The results in Fig. 5 show that the variables in the blue band mean that the regression coefficient is negative, the red band is the positive regression coefficient, and the yellow color band does not participate

in any model. Variables such as time variables (x06Th01, x22Th01, x07Th02, x23Th02, x10Th03, x26Th03, x11Th04, x13Th05, x29Th05, x14Th06), coordinate variables (Latitude, Longitude) of salinity monitoring sations, NDMI, MSI, and PCA2 appear 100% in all selected models. These results prove the highest importance of these input factors in the proposed models. Followed by variables such as SI5, COSRI, NDVI, DEM, VSSI, and B5 with appear frequencies of 61.2%, 49.8%, 49.6%, 48.6%, 45.7%, and 42.8%, respectively. Variables with the appear frequencies less than 25% are EVI, NDWI, ND47, SI3, SI4, SI6, SI7, B2, and B3. Remaining variables such as time variables (x13Th01, x15Th01, x14Th02, x01Th03, x03Th03, x17Th03, x19Th03, x04Th04, x18Th04, x27Th04, x04Th05, x07Th06, x21Th06), B1, B4, PCA1, PCA3, PCA4, PCA5, S1, S2, S8, S9, and NDSI that are not statistically significant in the proposed models.

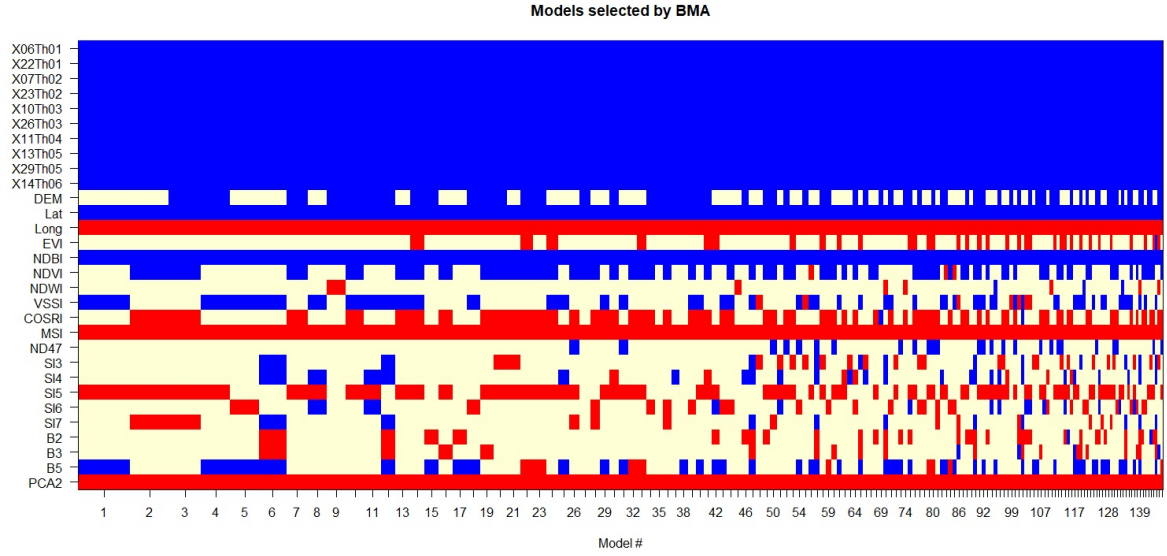

Figure 5. Appear frequency of input factors based on BMA analysis result

There were 147 established salinity prediction models (Fig. 5), but only five best salinity prediction models based on the above-analysis results of the BMA algorithm. The result showed the five best salinity prediction models out of 147 selected models (Fig. 6).

Fig. 6 indicates the five best salinity prediction models with the number of corresponding input variables, of which the vertical axis describes *intercept*, the order of input variables (from X06Th01 to PCA2), the number of variables (*nVar*), R-squared or coefficient of determination (*r*2), the Bayesian Information Criterion (*BIC*) value, and the post-probability (*post prob*) values; the horizontal axis reflects p factorial (*p*!), Expected Value (*EV*), Standard Deviation (*SD*), and the name of the best prediction models. These models could determine 66% salinity values (r^2 = 0.66) based on selected input variables. The best salinity prediction models were selected from the calculated results of the post-probability values and the Bayesian Information Criterion (BIC) value, so there were five selected optimal models. The first optimal model consists of 18 input variables, r^2 = 0.662, BIC value is −259.6748, and the post-probability value is 0.048; followed by the second optimal model with 19 input variables, r^2 = 0.667, BIC value is −259.0558, and the post-probability value is 0.036; the third optimal model with 20 input variables, r^2 = 0.672, BIC value is −258.7030, and the post-probability value is 0.030; the fourth optimal model with 19 input variables, r^2 = 0.667, BIC value is −258.4910, and the post-probability value is 0.027; and the last optimal model with 18 input variables, r^2 = 0.661, BIC value is −258.4546, and the post-probability value is 0.026 (Fig. 6 and Table 4).

```
  147  models were selected
 Best  5  models (cumulative posterior probability =  0.1669 ):

           p!=0    EV         SD         model 1     model 2     model 3     model 4     model 5
Intercept  100.0  -773.7944   115.8940  -769.7800   -845.5726   -773.7928   -707.0637   -766.3536
X06Th01    100.0   -15.4758     1.6595   -16.0107    -15.8034    -14.9879    -15.4048    -16.0913
X22Th01    100.0   -10.1153     1.4859   -10.3963    -10.4172     -9.7530     -9.8344    -10.3538
X07Th02    100.0   -11.6479     1.5748   -11.3709    -11.8615    -11.6005    -11.0963    -11.4471
X23Th02    100.0   -10.3600     1.5706   -10.4340    -10.2937     -9.8906    -10.0171    -10.2682
X10Th03    100.0    -5.5153     1.4085    -5.6648     -5.9978     -5.3888     -5.1062     -5.8230
X26Th03    100.0    -6.8534     1.5353    -7.0778     -7.2770     -6.7354     -6.6096     -7.1775
X11Th04    100.0    -8.8141     1.8339    -9.2723     -9.2484     -8.7105     -8.7670     -9.4257
X13Th05    100.0    -7.9943     1.3786    -8.2617     -8.6564     -8.1981     -7.8211     -8.0711
X29Th05    100.0   -12.2841     2.1699   -12.3127    -12.5686    -12.2274    -12.0491    -12.5611
X14Th06    100.0   -10.0555     2.1937   -10.9067    -11.0091    -10.3073    -10.2417    -10.7856
DEM         48.6    -0.2253     0.2695        .           .       -0.4361     -0.4019           .
Lat        100.0   -14.4786     1.2571   -14.1974    -15.0225    -14.5848    -13.8842    -14.4935
Long       100.0     7.9900     1.1147     7.8879      8.5519      7.9103      7.3325      8.0109
EVI         16.1     3.1159     8.9906          .           .           .           .           .
NDMI       100.0    76.4308    14.5826    75.1060     81.8273     77.1387     71.2128     76.1525
NDVI        49.6   -40.1759    47.9645          .    -78.8348    -75.7160           .           .
NDWI         5.1     0.6120     5.3073          .           .           .           .           .
VSSI        45.7  -130.6680   416.2907   -55.6760           .           .    -47.5363    -58.2601
COSRI       49.8    19.1093    22.3552          .     38.9959     38.3687           .           .
MSI        100.0    77.5981    12.0040    77.9651     82.2761     79.2786     75.3702     78.2042
ND47        11.2    -1.5418     5.1321          .           .           .           .           .
SI3         19.0  -723.1829  2656.2764          .           .           .           .           .
SI4         20.3   -66.3451   203.4759          .           .           .           .           .
SI5         61.2    13.3735    17.8494    15.8568     19.8492     18.1203     14.1122           .
SI6         22.9    -6.2806    53.7870          .           .           .           .     46.2626
SI7         18.8  -132.4127   557.2526          .     94.3979     77.5912           .           .
B2          18.5   256.2569   856.3809          .           .           .           .           .
B3          13.3   303.6919  1138.1721          .           .           .           .           .
B5          42.8  -627.8731  1934.4019  -383.4450           .           .    -329.1321   -394.9936
PCA2       100.0     3.3555     0.7302     3.2604      3.3127      3.3181      3.2427      3.2732

nVar                                        18          19          20          19          18
r2                                       0.662       0.667       0.672       0.667       0.661
BIC                                   -259.6748   -259.0558   -258.7030   -258.4910   -258.4546
post prob                                0.048       0.036       0.030       0.027       0.026
```

Figure 6. Selection of salinity factors and 5 optimal salinity prediction models using the BMA package in R software

Table 4. Five optimal salinity prediction models and the number of selected variables

Variables	Optimal salinity prediction models				
	Model 1 (18 variables)	Model 2 (19 variables)	Model 3 (20 variables)	Model 4 (19 variables)	Model 5 (18 variables)
Time variables (x06Th01, x22Th01, x07Th02, x23Th02, x10Th03, x26Th03, x11Th04, x13Th05, x29Th05, x14Th06)	X	X	X	X	X
Elevation (m)	-	-	X	X	-
Latitude	X	X	X	X	X
Longitude	X	X	X	X	X
NDMI	X	X	X	X	X
NDVI	-	X	X	-	-
VSSI	X	-	-	X	X
COSRI	-	X	X	-	-
MSI	X	X	X	X	X
SI5	X	X	X	X	-
SI6	-	-	-	-	X
SI7	-	X	X	-	-
B5	X	-	-	X	X
PCA2	X	X	X	X	X

3.3. Verification of 5 optimal salinity prediction models

In this study, the Random Forest model only validated the obtained optimal models from the BMA technique, including model 1 (18 variables), model 2 (19 variables), model 3 (20 variables), model 4 (19 variables), and model 5 (18 variables) based on determination coefficient (R^2), Root Mean Square Error (RMSE), and Mean Absolute Error (MAE). Since 337 in-situ salinity data are continuing input variables, so several statistical indexes, including determination coefficient (R^2), Root Mean Square Error (RMSE), and Mean Absolute Error (MAE) exploited from the Random Forest method were utilized to validate the performance of 5 optimal salinity prediction models, including model 1 (18 variables), model 2 (19 variables), model 3 (20 variables), model 4 (19 variables), and model 5 (18 variables).

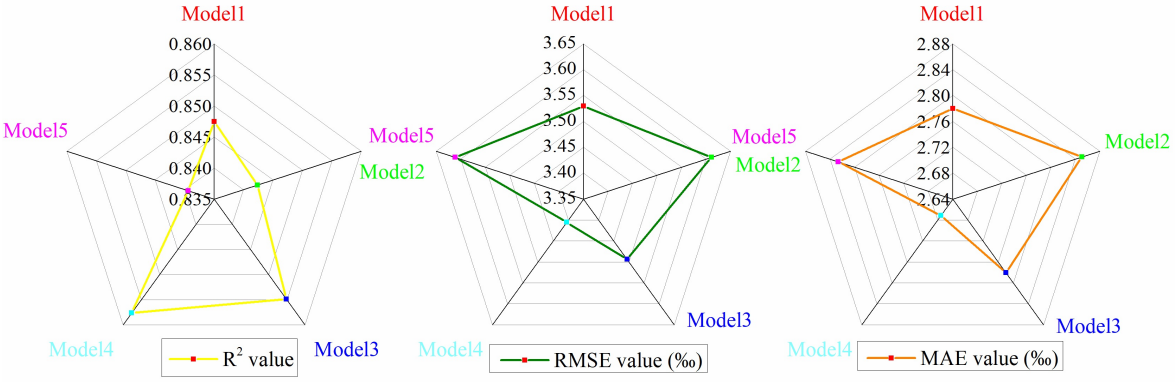

Figure 7. Statistical indexes of 5 optimal salinity prediction models

The calculated results demonstrated that the R-squared values ranged from 0.839 to 0.858, the RMSE values ranged from 3.405 (‰) to 3.612 (‰), and the MAE values ranged from 2.672 (‰) to 2.851 (‰) (Fig. 7). The obtained values generally denoted five optimal salinity prediction models that fit the input data and have good performance.

4. Conclusions

By analyzing observed salinity data and climate data extracted from Landsat 8 OLI in the Google Earth Engine platform, this study uses the BMA algorithm to identify the significance of relevant input variables and filter optimal salinity prediction models. The analysis results from the BMA algorithm show that only about 23 input variables play an important role in the total 54 input variables and only five optimal salinity prediction models out of 147 synthesized models. In addition, various statistical metrics consisted of determination coefficient (R^2), Root Mean Square Error (RMSE), and Mean Absolute Error (MAE) from the Random Forest method, which was used to verify the fit to the input data of the models and to confirm the performance of these optimal salinity prediction models. The results indicated these five best salinity prediction models that fit the input variables and had a good performance. These findings of this study can supply foundational knowledge and a basis for future studies, determining appropriate input variables and selecting the most suitable salinity prediction model to ensure the efficiency of salinity mitigation strategies. In this study, we add the coordinate and elevation as input variables; these variables have not been used in any of the published studies before. The results of this study can provide valuable information and basic background for further studies in shaping the selection of suitable input variables and optimal salinity prediction models to save time and effort. This research only focuses on studying the application of machine learning models in evaluating the importance of input variables and selecting the optimal salinity prediction model. The construction of salinity prediction maps will be conducted in further studies in the future.

Acknowledgment

This research was funded by the Ministry of Education and Training under the project "Assessing saltwater intrusion risk under climate change impact" grant number CT.2022.01.XDA.04.

References

[1] Negacz, K., Malek, Z., de Vos, A., Vellinga, P. (2022). Saline soils worldwide: Identifying the most promising areas for saline agriculture. *Journal of Arid Environments*, 203:104775.

[2] Lal, A., Datta, B. (2020). Application of the group method of data handling and variable importance analysis for prediction and modelling of saltwater intrusion processes in coastal aquifers. *Neural Computing and Applications*, 33(9):4179–4190.

[3] Sudha Rani, N. N. V., Satyanarayana, A. N. V., Bhaskaran, P. K. (2015). Coastal vulnerability assessment studies over India: a review. *Natural Hazards*, 77(1):405–428.

[4] Smajgl, A., Toan, T. Q., Nhan, D. K., Ward, J., Trung, N. H., Tri, L. Q., Tri, V. P. D., Vu, P. T. (2015). Responding to rising sea levels in the Mekong Delta. *Nature Climate Change*, 5(2):167–174.

[5] Aksoy, S., Yildirim, A., Gorji, T., Hamzehpour, N., Tanik, A., Sertel, E. (2022). Assessing the performance of machine learning algorithms for soil salinity mapping in Google Earth Engine platform using Sentinel-2A and Landsat-8 OLI data. *Advances in Space Research*, 69(2):1072–1086.

[6] Kabiraj, S., Jayanthi, M., Vijayakumar, S., Duraisamy, M. (2022). Comparative assessment of satellite images spectral characteristics in identifying the different levels of soil salinization using machine learning techniques in Google Earth Engine. *Earth Science Informatics*, 15(4):2275–2288.

[7] Rojas, R., Feyen, L., Dassargues, A. (2008). Conceptual model uncertainty in groundwater modeling: Combining generalized likelihood uncertainty estimation and Bayesian model averaging. *Water Resources Research*, 44(12).

[8] Tsai, F. T.-C. (2010). Bayesian model averaging assessment on groundwater management under model structure uncertainty. *Stochastic Environmental Research and Risk Assessment*, 24(6):845–861.

[9] Nguyen, T. T. M., Nguyen, N. T., Du Xuan Nguyen, C. T. T., Tri, D. Q. (2023). Mapping of Top-Soil Salinity Zoning in the Coastal Area of Ben Tre Province, Vietnam. *International Journal of Intelligent Systems and Applications in Engineering*, 11(4s):473–490.

[10] Chitsazan, N., Pham, H. V., Tsai, F. T. (2014). Bayesian Chance-Constrained Hydraulic Barrier Design under Geological Structure Uncertainty. *Groundwater*, 53(6):908–919.

[11] Abbasi, M., Dehban, H., Farokhnia, A., Roozbahani, R., Bahreinimotlagh, M. (2022). Long-Term Streamflow Prediction Using Hybrid SVR-ANN Based on Bayesian Model Averaging. *Journal of Hydrologic Engineering*, 27(11).

[12] Ruidas, D., Pal, S. C., Saha, A., Chowdhuri, I., Shit, M. (2022). Hydrogeochemical characterization based water resources vulnerability assessment in India's first Ramsar site of Chilka lake. *Marine Pollution Bulletin*, 184:114107.

[13] Nguyen, T. G., Tran, N. A., Vu, P. L., Nguyen, Q.-H., Nguyen, H. D., Bui, Q.-T. (2021). Salinity intrusion prediction using remote sensing and machine learning in data-limited regions: A case study in Vietnam's Mekong Delta. *Geoderma Regional*, 27:e00424.

[14] Nguyen, P. T. B., Koedsin, W., McNeil, D., Van, T. P. D. (2018). Remote sensing techniques to predict salinity intrusion: application for a data-poor area of the coastal Mekong Delta, Vietnam. *International Journal of Remote Sensing*, 39(20):6676–6691.

[15] Tran, T. T., Pham, N. H., Pham, Q. B., Pham, T. L., Ngo, X. Q., Nguyen, D. L., Nguyen, P. N., Veettil, B. K. (2022). Performances of Different Machine Learning Algorithms for Predicting Saltwater Intrusion in the Vietnamese Mekong Delta Using Limited Input Data: A Study from Ham Luong River. *Water Resources*, 49(3):391–401.

[16] Nguyen, H. D., Van, C. P., Nguyen, T. G., Dang, D. K., Pham, T. T. N., Nguyen, Q.-H., Bui, Q.-T. (2023). Soil salinity prediction using hybrid machine learning and remote sensing in Ben Tre province on Vietnam's Mekong River Delta. *Environmental Science and Pollution Research*, 30(29):74340–74357.

[17] Minderhoud, P. S. J., Coumou, L., Erkens, G., Middelkoop, H., Stouthamer, E. (2019). Mekong delta much lower than previously assumed in sea-level rise impact assessments. *Nature Communications*, 10 (1).

[18] Anthony, E. J., Brunier, G., Besset, M., Goichot, M., Dussouillez, P., Nguyen, V. L. (2015). Linking rapid erosion of the Mekong River delta to human activities. *Scientific Reports*, 5(1).

[19] Eslami, S., Hoekstra, P., Nguyen Trung, N., Ahmed Kantoush, S., Van Binh, D., Duc Dung, D., Tran Quang, T., van der Vegt, M. (2019). Tidal amplification and salt intrusion in the Mekong Delta driven by anthropogenic sediment starvation. *Scientific Reports*, 9(1).

[20] Minderhoud, P. S. J., Middelkoop, H., Erkens, G., Stouthamer, E. (2020). Groundwater extraction may drown mega-delta: projections of extraction-induced subsidence and elevation of the Mekong delta for the 21st century. *Environmental Research Communications*, 2(1):011005.

[21] MERIT DEM. *Multi-Error-Removed Improved-Terrain DEM (Last update: 15 Oct, 2018)*. 12 May, 2019.

[22] Forkuor, G., Dimobe, K., Serme, I., Tondoh, J. E. (2017). Landsat-8 vs. Sentinel-2: examining the added value of sentinel-2's red-edge bands to land-use and land-cover mapping in Burkina Faso. *GIScience & Remote Sensing*, 55(3):331–354.

[23] Baig, M. H. A., Zhang, L., Shuai, T., Tong, Q. (2014). Derivation of a tasselled cap transformation based on Landsat 8 at-satellite reflectance. *Remote Sensing Letters*, 5(5):423–431.

[24] Rouse, J. W., Haas, R. H., Schell, J. A., Deering, D. W. (1974). *Monitoring vegetation systems in the Great Plains with ERTS*. NASA Spec. Publ, 1, p. 309.

[25] Gao, B.-c. (1996). NDWI—A normalized difference water index for remote sensing of vegetation liquid water from space. *Remote Sensing of Environment*, 58(3):257–266.

[26] Wu, W. (2019). *A Brief Review on Soil Salinity Mapping by Optical and Radar Remote Sensing*. Springer Singapore, 53–65.

[27] Fernández-Buces, N., Siebe, C., Cram, S., Palacio, J. L. (2006). Mapping soil salinity using a combined spectral response index for bare soil and vegetation: A case study in the former lake Texcoco, Mexico. *Journal of Arid Environments*, 65(4):644–667.

[28] Scudiero, E., Skaggs, T. H., Corwin, D. L. (2015). Regional-scale soil salinity assessment using Landsat ETM + canopy reflectance. *Remote Sensing of Environment*, 169:335–343.

[29] Huntjr, E., Rock, B. (1989). Detection of changes in leaf water content using Near- and Middle-Infrared reflectances. *Remote Sensing of Environment*, 30(1):43–54.

[30] Huete, A., Didan, K., Miura, T., Rodriguez, E. P., Gao, X., Ferreira, L. G. (2002). Overview of the radiometric and biophysical performance of the MODIS vegetation indices. *Remote Sensing of Environment*, 83(1–2):195–213.

[31] Dehni, A., Lounis, M. (2012). Remote Sensing Techniques for Salt Affected Soil Mapping: Application to the Oran Region of Algeria. *Procedia Engineering*, 33:188–198.

[32] Yahiaoui, I., Douaoui, A., Zhang, Q., Ziane, A. (2015). Soil salinity prediction in the Lower Cheliff plain (Algeria) based on remote sensing and topographic feature analysis. *Journal of Arid Land*, 7(6):794–805.

[33] Douaoui, A. E. K., Nicolas, H., Walter, C. (2006). Detecting salinity hazards within a semiarid context by means of combining soil and remote-sensing data. *Geoderma*, 134(1–2):217–230.

[34] Khan, N. M., Rastoskuev, V. V., Sato, Y., Shiozawa, S. (2005). Assessment of hydrosaline land degradation by using a simple approach of remote sensing indicators. *Agricultural Water Management*, 77(1–3): 96–109.

[35] Abbas, A., Khan, S. (2007). Using remote sensing techniques for appraisal of irrigated soil salinity. In *International Congress on Modelling and Simulation (MODSIM)*, Modelling and Simulation Society of Australia and New Zealand, 2632–2638.

[36] 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(19–20): 2795–2811.

[37] Fakhri, S. A., Sayadi, S., Latifi, H., Khare, S. (2019). An optimized Enhanced Vegetation Index for Sparse Tree Cover Mapping across a Mountainous Region. In *2019 IEEE International Workshop on Metrology for Agriculture and Forestry (MetroAgriFor)*, IEEE, 146–151.

[38] Mukaka, M. M. (2012). A guide to appropriate use of correlation coefficient in medical research. *Malawi medical journal*, 24(3):69–71.

[39] Benesty, J., Chen, J., Huang, Y., Cohen, I. (2009). *Pearson Correlation Coefficient*. Springer Berlin Heidelberg, 1–4.

[40] Wu, W., Mhaimeed, A. S., Al-Shafie, W. M., Ziadat, F., Dhehibi, B., Nangia, V., De Pauw, E. (2014). Mapping soil salinity changes using remote sensing in Central Iraq. *Geoderma Regional*, 2–3:21–31.

[41] Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., Moore, R. (2017). Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*, 202:18–27.

[42] Fragoso, T. M., Bertoli, W., Louzada, F. (2017). Bayesian Model Averaging: A Systematic Review and Conceptual Classification. *International Statistical Review*, 86(1):1–28.

[43] Chinh, L. T. D., Hang, H. T. (2023). An analysis of the relative variable importance to flood fatality using a machine learning approach. *Journal of Science and Technology in Civil Engineering (STCE) - HUCE*, 17(1):125–136.

[44] Rodriguez-Galiano, V., Sanchez-Castillo, M., Chica-Olmo, M., Chica-Rivas, M. (2015). Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. *Ore Geology Reviews*, 71:804–818.