

Investigation of the Optimal Method of Data Processing to Increase Accuracy of Simulation of Surface Soil Salinity (Case study: MARVAST)

A. Habibipoor^{1*} – A. Talebi² – A. A. Karimian³ – F. Dehghani⁴ – M. H. Mokhtari⁵

Received: 26-4-2016

Accepted: 31-5-2017

Introduction: Salinity is one of the problems of arid and semi-arid soils. Identification and classification of saline/alkaline soils is necessity for dealing with difficult situations and correct management. Considering the nature of salinity data and selection of befitting methods to process data before use artificial neural network, can result in better simulations. The aim of this study was to investigate the optimal method for data processing to enhance the accuracy of surface soil salinity simulation and improve the efficiency of decision tree algorithm.

Materials and Methods: The study area was 88940.4 hectares of *Marvast* plain located in central Iran ($54^{\circ} 5' \text{ to } 54^{\circ} 18'$ east longitude and $30^{\circ} 10' \text{ to } 30^{\circ} 35'$ north latitude). This region faces with problems of soil and water resources salinity. In this study, the effect of data processing on increasing accuracy of simulation of soil surface salinity was assessed in *Marvast* region using decision tree algorithm. For this purpose, the decision tree algorithm was applied and simulation was performed using three approaches i.e. original data, logarithmic data and standardized data. Finally, five statistics including R, Rmse, %Rmse, MAE and Bias were calculated to evaluate the performance of used simulation methods.

Results and Discussion: In this study, when the logarithmic data was used, the composition of band 7 – elevation was identified as the most appropriate condition. The created tree can estimate the soil salinity by five laws:

If elevation is less than 1519, then the average of surface soil salinity will be 147.9 ds/m.

If elevation is between 1519 to 1569.9, then the average of surface soil salinity will be 43.6 ds/m.

If elevation is between 1569.9 to 1609.8, then the average of surface soil salinity will be 17.5 ds/m.

If elevation is more or equal to 1609.8 and pixel value of band 7 (ETM+ sensor) in selected point is less than 0.295, then the average of surface soil salinity will be 4.7 ds/m.

If elevation is higher or equal to 1609.8 and pixel value of band 7 (ETM+ sensor) in selected point is more than or equal to 0.295, then the average of surface soil salinity will be 1.4 ds/m.

For the approach of using the logarithmic data, decision tree algorithm used two parameters out of 46 independent variables introduced into the model. R, Rmse, %Rmse, MAE and Bias for this method was computed to be 0.76, 0.49, 38.57, 0.37 and -0.14, respectively. The application of logarithmic data was recognized as the best method considering the lower calculated error and its less input requirement. Using Easy fit software, the distribution of salinity data was found to be Log Pearson 3. Thus, the use of logarithmic data improved model performance. Our findings were in agreement with those of Afkhami et al (2015) who increased the simulation accuracy of suspended sediment with artificial intelligence methods (Artificial neural networks and ANFIS) using logarithmic data.

Conclusions: As effective factors for soil salinity simulation vary in different regions, application of a unique method and indicator to estimate soil salinity in deferent region may not be possible. The application of semi intelligent algorithm which limits user intervention and selects effective parameters for simulation would increase the simulation accuracy. Furthermore, considering the nature of salinity data and selection of befitting methods to process before using decision tree algorithm can effectively improve model performance. The current study was conducted to select an appropriate approach to enhance the simulation accuracy of surface soil salinity. The results demonstrate that the performance of decision tree algorithm as one of the artificial intelligence models can be affected by input data. In this study, Log-Pearson3 distribution was defined as the distribution of salinity data. Moreover, despite existence of significant correlation coefficients for three simulation methods, the error was lower when logarithmic data was used. Since the probability distribution of salinity data in the studied area was logarithmic (Log-Pearson 3), the reduction in error rate can be attributed to

1, 2, 3, 5- PhD Student, Associate Professor and Assistant Professors Yazd University, Iran

(*-Corresponding Author Email: Email: AzamHabibipoor@yahoo.com)

4-Faculty of national salinity research center, Yazd, Iran

the probability distribution of salinity data.

Keywords: Normalization, Probability distribution, Regression decision tree, Electrical conductivity

Archive of SID