Theoretical and Applied Climatology (2020) 139:1111–1119
https://doi.org/10.1007/s00704-019-03048-8

ORIGINAL PAPER

Estimation of soil moisture using decision tree regression

Engin Pekel[1]

Received: 22 May 2019 / Accepted: 10 November 2019 / Published online: 26 November 2019

Abstract

Soil moisture (SM) is a significant factor in the climate system. The accurate determination of SM has high importance in food production to satisfy the increasing demand for food and the chemical processes of soil. This paper applies decision tree regression to estimate SM considering different parameters including air temperature, time, relative humidity, and soil temperature. The presented method holds a mighty advantage to determine SM since the stimulant of the decision tree regression is an algorithm that generates a decision tree from given instances. Besides, usage of decision tree regression provides an opportunity to save time. Numerical results show that the presented method offers a high coefficient of determination value (R^2), low mean squared error (MSE), and mean absolute error (MAE). The depth of the decision tree equals to five by providing higher fitness values than other depth levels. The best fitness values in the training stage are 0.00019, 0.007, and 0.842 for MSE, MAE, and R^2, respectively. In conclusion of the paper, applied decision tree regression can handle the data of SM estimation in satisfying fitness criterion.

Keywords Decision tree regression · Estimation · Learning · Soil moisture

1 Introduction

The soil is a fundamental component in satisfying the increasing demand for food and other requirements of the world population (Shukla et al., 2018). One of the most significant factors of the climate system is soil moisture (SM).SM regulates the water and energy circulation between the atmosphere and the land (Zuo and Zhang, 2008). It has a closer relationship with climate change; variety in temperature, precipitation, and other climatic factors directly result in variations in SM (Han et al., 2018). It is remarkable to analyze the SM data and achieve the distribution of SM in the regional climate system since it enables to define the threat of droughts and disasters and adjust agricultural production. It not only influences the climate change but also affects the chemical processes of soil (Badía et al., 2017; Kumar et al., 2018).

Field sampling (direct drying method), electromagnetic measurement of soil water, and remote sensing technology can measure SM. Field sampling is a direct drying method, and it calculates SM concerning the difference in soil weight before and after drying. Electromagnetic measurement utilizes the electromagnetic properties of SM. It includes time-domain reflectometry (TDR), frequency-domain reflectometry (FDR), ground-penetrating radar (GPR), and resistivity method (Huisman et al., 2001; Alamry et al., 2017). Remote sensing technology gains ground surface information in large areas and provides a basic spatial resolution (Qu et al. 2018). The algorithms in remote sensing technology consist of measurements based on visible/thermal infrared sensing (Pohn et al., 1974) and measurements based on microwave sensing (Wang et al., 1983) concerning the wavelength bands. However, a large number of factors (atmospheric correction, soil roughness, vegetation, etc.) affect the remote sensing of SM. The results of the remote sensing require a detailed validation (Rötzer et al., 2014).

This paper investigated literature considering "soil moisture" and "decision tree" keywords and found 107 papers published in total. These papers consist of 79 articles, 26 proceeding papers, and five review papers. One paper (Gorthi and Dou, 2011) deals with the estimation of SM by considering SM and decision trees in the proceeding papers. However, the parameters of the dataset used in this proceeding paper are different from the dataset used in this paper. The proceedings paper has six inputs and one output. Its inputs are air temperature, precipitation, soil temperature, vegetation indices, leaf area index, and land surface temperature. However, the inputs of this paper are time, soil temperature, air temperature, and relative humidity.

✉ Engin Pekel
enginpekel@hitit.edu.tr

[1] Department of Industrial Engineering, Faculty of Engineering, Hitit University, Çorum, Turkey

Recently, the methodology of machine learning, as a popular and significant technology of computer science, meets the demands of different academic disciplines and practical fields. Various types of machine learning methods, including artificial neural network (ANN), support vector machine (SVM), k-nearest neighbor algorithm (k-NN), and decision tree classification, are performed in pattern recognition and spatial data processing. Gill et al. (2006) implemented the SVM model to predict SM based on aerial variables and compared the results with the performed ANN model. Additionally, in their research, they found that the SVM model obtained a higher degree of accuracy in prediction compared with the ANN model. Malajner et al. (2019) applied the SVM model to estimate SM using low cost and small size commercial Ultra-Wide Band modules. Prakash et al. (2018) used machine learning techniques such as linear regression, SVM, and recurrent ANN to predict SM for several days ahead. Sarti and Mascolo (2012) also measured SM with high accuracy. They implemented the polarimetric extraction technique of AIRSAR in C and L channels. Ahmad et al. (2010) investigated SM in the western USA based on data obtained from remote sensing. They applied the SVM regression technique for ten sections and compared results with the ANN model. Their paper pointed out that the SVM model provided a better prediction for SM compared with ANN. Hajdu et al. (2018) applied the random forest method to a group of data obtained from a farm scale. They also found that their performed method made enable to catch the non-linear relationship between the ground-based and remotely sensed variables. Hajnsek et al. (2009) utilized polarimetric SAR (PolSAR) acquisitions to predict the volume of SM. Qu et al. (2018) analyzed the characteristics of SM by combining remote sensing data and implemented support vector classifier (SVC) as a machine learning model to show its characteristics.

In conclusion, the estimation of SM where the paper performs the decision tree regression with the defined parameters has not been previously addressed in the research papers.

The main contributions of the paper are as follows:

- The paper illustrates that the application of decision tree regression to estimate SM presents a high coefficient of determination (R^2) value and low mean squared error (MSE) and mean absolute error (MAE).
- The performed methodology provides an accurate estimation under a short computation time.
- The paper estimates of SM considering decision tree regression with the different parameters.

The remainder of this paper is organized as follows: Section 2 presents the dataset of the case study, the structure, and implementation of the performed methodology. In Section 3, the numerical results with different parameter settings are demonstrated. Section 4 provides concluding remarks and directions for future researches.

2 Material and method

2.1 Study area and experimental data

The dataset, used in this paper, is taken from the website of the University of Toronto Mississauga campus. The collected data includes three different areas. The areas of the data consist of pond, field, and forest data. Field data is only considered in this paper. Data are collected hourly at these sites using HOBO U30 data loggers equipped with sensors monitoring the inputs including time (day), soil temperature (degrees Celsius), air temperature (degrees Celsius), and relative humidity (%) and SM output ($\frac{m^3}{m^3}$). This research paper split the dataset into two parts for the training and testing stage. The initial dataset consists of 1000 instances that include four inputs and one output. Eighty percent and 20% of the dataset are kept at random for the training and testing stage, respectively.

2.2 Decision tree regression

A decision tree is a tree data structure that consists of an arbitrary number of nodes and branches at each node. A node with outgoing edges is called an internal node. Other nodes are called leaves. The instance that is used for regression or classification is split into two or more groups by an internal node concerning a specific function. The values of the input variable(s) consider a particular function in the training stage (Loh, 2011).

The stimulant of the decision tree is an algorithm that generates a decision tree from given instances. The performed algorithm aims to find the optimal decision tree by minimizing the fitness function. Since the chosen dataset in this paper does not have classes, a regression model is fit to the target variable using each of the independent variables. The dataset is split at several split points for each independent variable. At each split point, the performed algorithm calculates the error between the predicted value and the actual values concerning the pre-defined fitness function. The split point errors across the variables are compared, and the variable yielding the lowest fitness function value is chosen as the split point. This process recursively continues.

Figure 1 shows a representative decision tree that belongs to the used dataset. In Fig. 1, root, node, and leaves are the topmost decision node, a decision node, and the final decision, respectively. Decision tree regression uses a fast divide and conquer greedy algorithm that recursively splits the data into smaller parts. The efficiency of the method comes from the use of this algorithm. This greedy algorithm can cause poor decisions in lower levels of the tree because of the instability of the estimations. For instance, let the time parameter be a root node. If the value of the root node is higher than 0.5 (12 h), the decision node is the air temperature. Then, if the value of the air temperature is a present decision node and less than − 2.09, the soil temperature becomes the next decision node. Conclusively, if the value of the

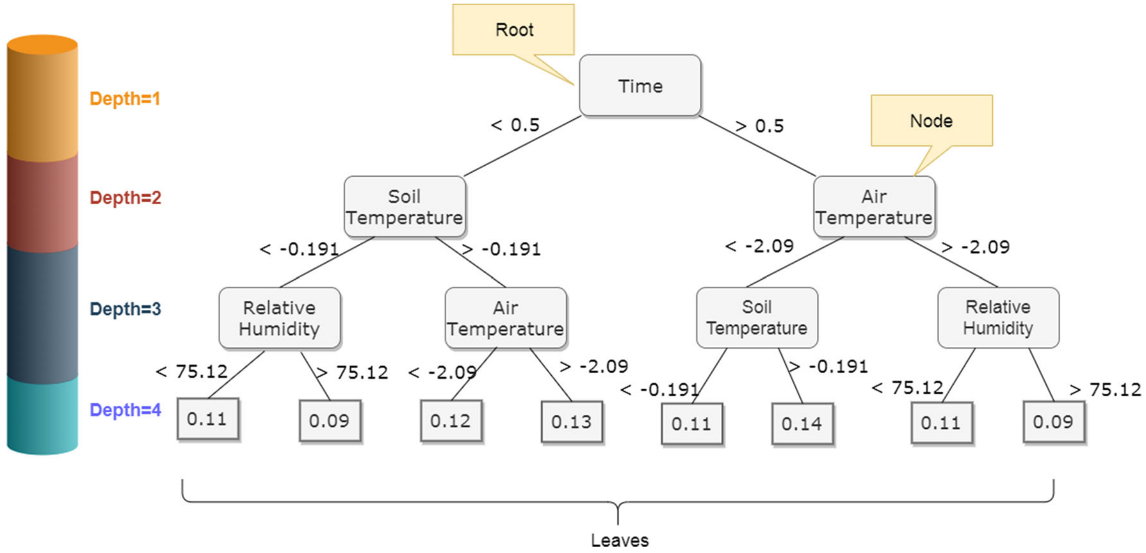


Fig. 1 A representative decision tree

soil temperature is a present decision node and higher than −0.191, the value of SM equals to 0.14 in the leaf. Otherwise, the value of SM is equaled to 0.11 in the leaf.

2.3 Advantages and disadvantages of decision tree regression

Decision tree regression is performed to estimate the values of SM. Decision trees are a technique that can stimulate trees for predicting both categorical and real-valued targets, and hence they can be used for regression as well as classification. Decision tree regression is performed to the estimation of SM because it offers some advantages, and some of them are addressed as follows (De'ath and Fabricius, 2000):

- The method provides the flexibility to overcome a broad range of response types such as categorical, numeric data.
- It is simple to understand and to evaluate the structure of the tree.
- It enables to validate a model using statistical tests. That points out the decision tree regression as a reliable model.
- It can overcome missing values in explanatory and response variables.
- It can overcome multi-output variables.

Although decision tree regression offers many advantages, several disadvantages are existing. Some of them are as follows:

- Decision tree regression can be unsteady since small variations in the data may result in a different tree. Experiments may avoid small changes for an ideal validation.
- It may be trapped to over-fitting. The parameters of decision tree regression are required to be tuned to avoid this problem.

Decision tree regression is performed to estimate the values of SM by utilizing its advantages and avoiding its disadvantages.

2.4 Parameters of decision tree regression

vDecision tree regression has several parameters that can influence the quality of the estimation. Some of these parameters can be listed as the fitness function, depth of the tree, split sample, leaf sample, and feature number.

The fitness function aims to minimize the error between the predicted value and the observed value. A depth of the tree determines how deep the tree can grow. The deeper the tree, the more it obtains information about the data. Split sample means the minimum number of samples required to split an internal node. If the sample split equals to five and there are six samples at an internal node, then the split is allowed. Leaf sample is the minimum number of samples that are necessary to be at a leaf node. Imagine that one of the leaves has one sample, and if the leaf sample equals to two, the split is not carried out because of this leaf.

2.5 Structure of decision tree regression

In a regression problem, let $X = X_1, X_2, \ldots, X_{pn}$ be predictor variables. pn is the total number of predictor variables. Let n and $Y = Y_1, Y_2, \ldots, Y_n$ be the number of observations and a target variable that takes continuous values, respectively. vf is a feature variable and th is a threshold

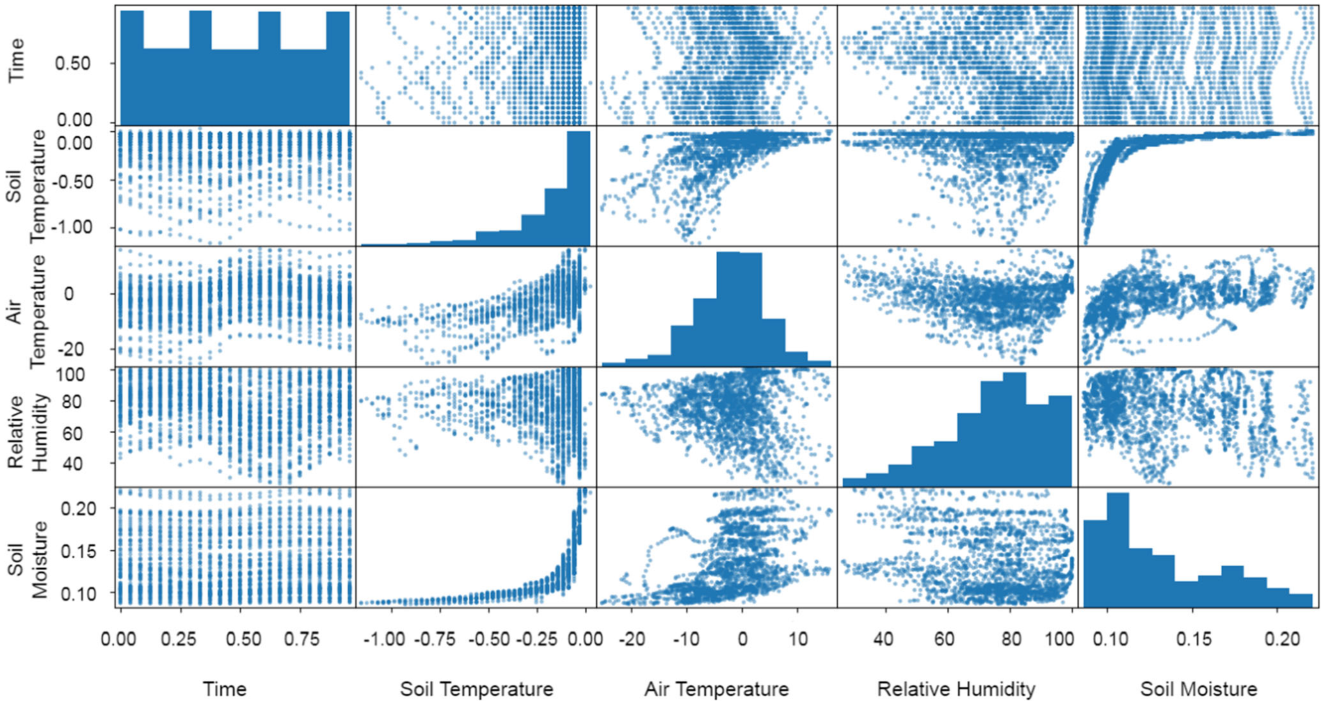


Fig. 2 Scatter matrix of the data

value. Let t and $\gamma = (vf, th_t)$ be a node, candidate split, respectively.

$$Q_l(\gamma) = (x, y)|x_{vf} \leq th_t \tag{1}$$

Equation (1) shows that Q_l that is the left side in the decision tree is found by splitting the data into γ candidate split.

$$Q_r(\gamma) = (x, y)|x_{vf} > th_t \tag{2}$$

Equation (2) shows that Q_r that is the right side in the decision tree is found by splitting the data into γ candidate split. Also, Eq. (2) can be declared as $Q_r(\gamma) = Q/Q_l(\gamma)$.

Y_t is the mean predicted value at terminal nodes. Let n be the number of a sample at the current node.

$$\overline{Y}_t = \frac{1}{n} \sum_{i \in n} Y_i \tag{3}$$

Equation (3) shows the calculation of the mean predicted value at terminal nodes. The value of this calculation is used in Eqs. (4)–(6).The fitness function is calculated concerning MSE, MAE, and R^2 in this paper.

$$S(X_t) = \frac{1}{n} \sum_{i \in n} \left(Y_i - \overline{Y}_t\right)^2 \tag{4}$$

$$S(X_t) = \frac{1}{n} \sum_{i \in n} |Y_i - \overline{Y}_t| \tag{5}$$

$$S(X_t) = 1 - \frac{\sum_{i \in n} \left(Y_i - \overline{Y}_t\right)^2}{\sum_{i \in n} \left(Y_i - \widetilde{Y}\right)^2} \tag{6}$$

Let $Y^{\check{}}$ be the mean value of the observed output. Equations (4), (5), and (6) show the calculation of MSE, MAE, and R^2, respectively.

Equations (4), (5), and (6) are used as fitness functions to reach satisfying fitness criteria in estimation. Let n_l and n_r be the number of a sample left and right child (side), respectively.

$$I(Q, \gamma) = \frac{n_l}{n} S(Q_l(\gamma)) + \frac{n_r}{n} S(Q_r(\gamma)) \tag{7}$$

I is an impurity function. Equation (7) shows how the impurity function is calculated. The impurity function is minimized by considering Q and γ parameters.

Algorithm 1 presents the pseudo-code for the construction of decision tree regression.

Algorithm 1

Pseudocode for decision tree regression

```
(1) Start with a single node
(2) For each $X$, find the fitness function value ($S$) and choose the split that
    offers the minimum value of the fitness function
(3) In each new node, go back to step (2). If a stopping criterion is
    reached, exit.
```

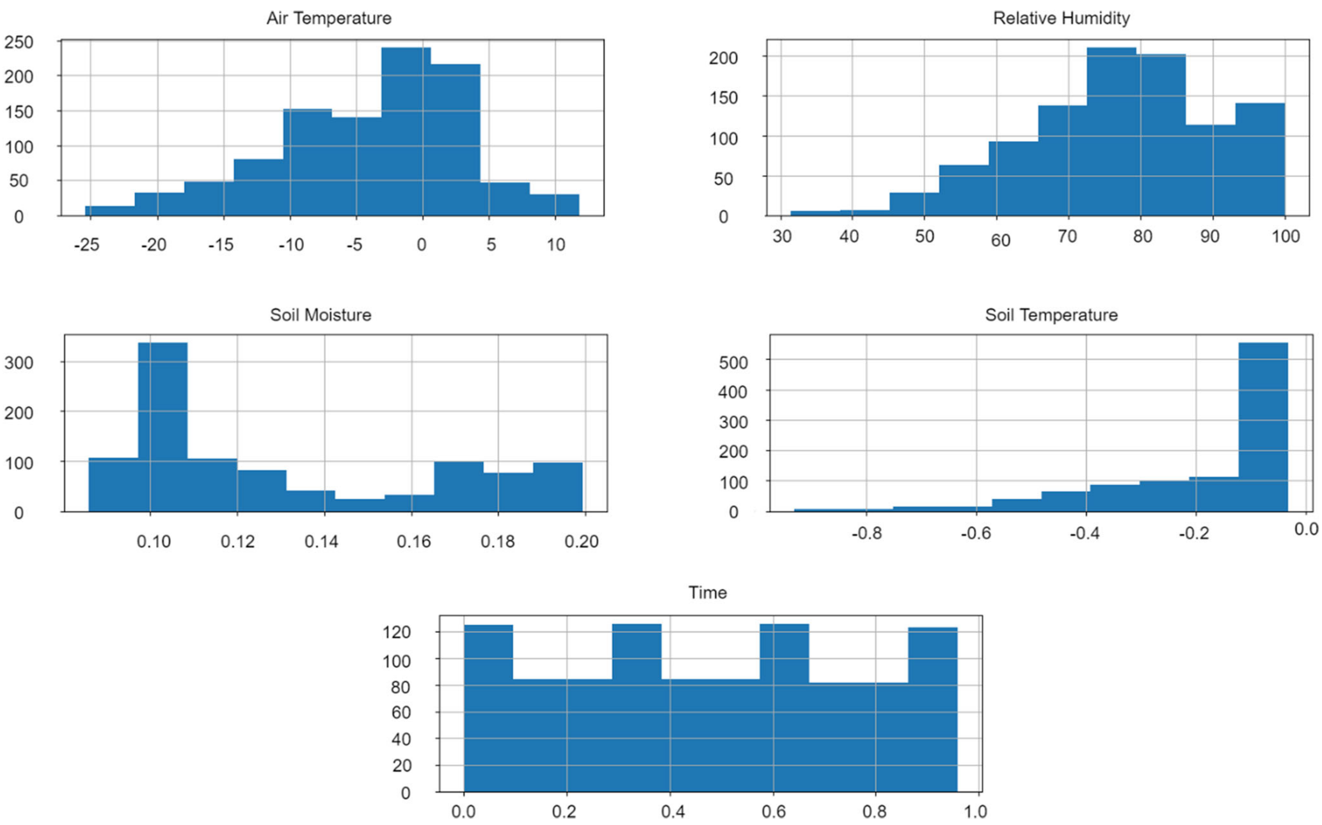


Fig. 3 Histogram of the data

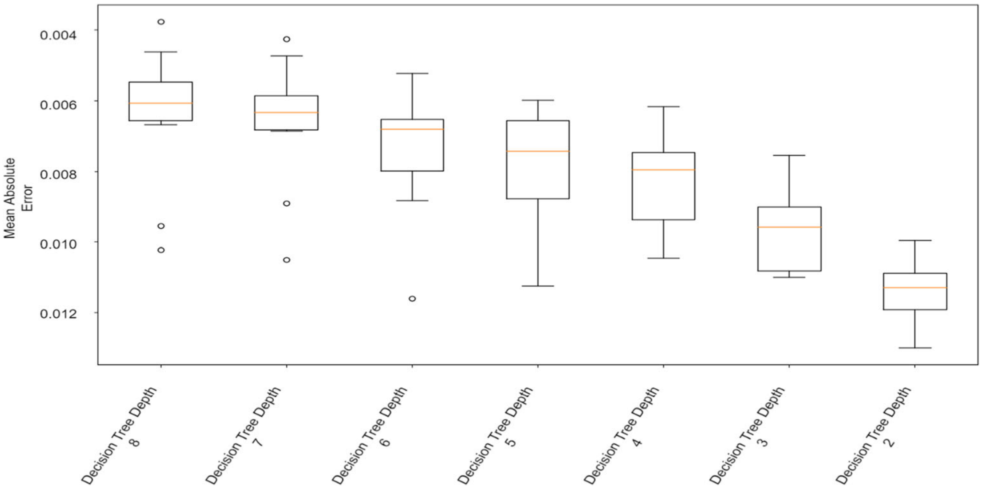
3 Numerical results

This section provides all characteristics of all features and numerical results of the estimation experiment. The method of decision tree regression has been run on a computer that has a 32-bit Windows 7 operating system, 2.9 GHz processor, and 4-GB memory. The performed method has been implemented in the Python programming language. The total time that consists of the training, the testing, and the estimation stages takes approximately 3 s.

3.1 Data structure

Firstly, the structure of data is presented to illustrate the complexity of the chosen data in Figs. 2 and 3. Later, Figs. 4, 5, and 6 provide the training results obtained from

Fig. 4 MAE results under different depths



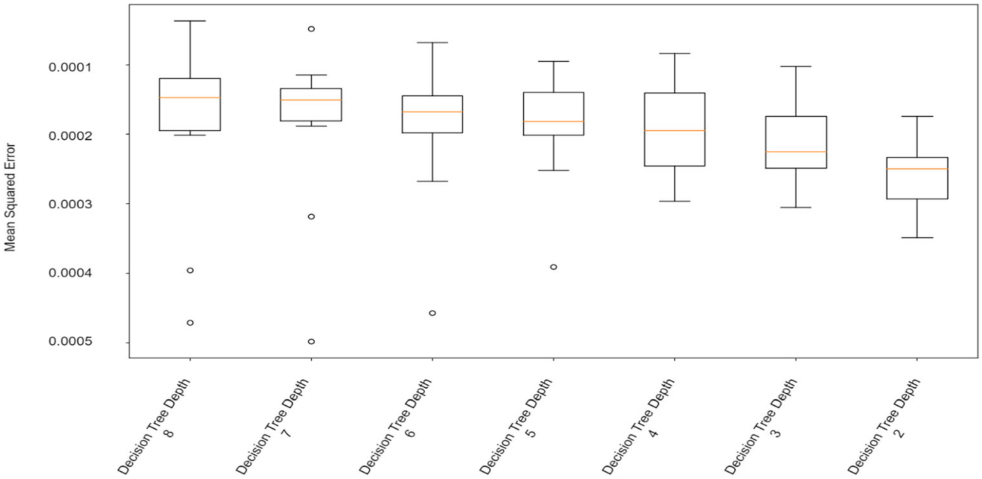


Fig. 5 MSE results under different depths

different depths of the decision tree concerning the different fitness criteria that include MSE, MAE, and R^2. Finally, Figs. 7 and 8 provide the estimation results of all and the best decision tree depth, respectively.

Figure 2 shows the scatter matrix of all features in the data. A scatter matrix is a pair-wise scatter plot of all features, including input and output variables. The scatter matrix in Fig. 2 illustrates any correlation between each feature. This correlation can be either positive or negative. For example, in Fig. 2, the correlations between soil temperature and SM pairs, air temperature, and soil temperature pairs do not show a linear structure. In light of the non-linear structure, Fig. 2 shows that the data is quite complex to estimate SM.

Figure 3 presents a view of the data concerning the histogram graph. Figure 3 shows the fundamental frequency distribution of all inputs and output. The area of the bar indicates the frequency of occurrences for each instance. A careful examination of Fig. 3 reveals that air temperature and relative humidity show a distribution that approximates the normal distribution. The weighted values of air temperature change between − 3 and 4 °C. The weighted values of relative humidity change between 72 and 86%.The weighted values of soil temperature change between − 0.1 and 0 °C. Time input shows a different distribution than the others because the values of time have a homogeneous distribution. The values of time change between 0 and 1. Zero and 1 represent the range in 24 h. SM is an output that changes between 0.02 and 0.20.

3.2 Training the performed algorithm

A cross-validation is applied in this paper for the training data to avoid overfitting. Cross-validation carries out splitting the training dataset into ten subsets. The validation procedure holds one of the ten subsets, and the rest of the subsets are trained concerning the fitness function. Experiments are carried out ten times for each kept subset in the training stage.

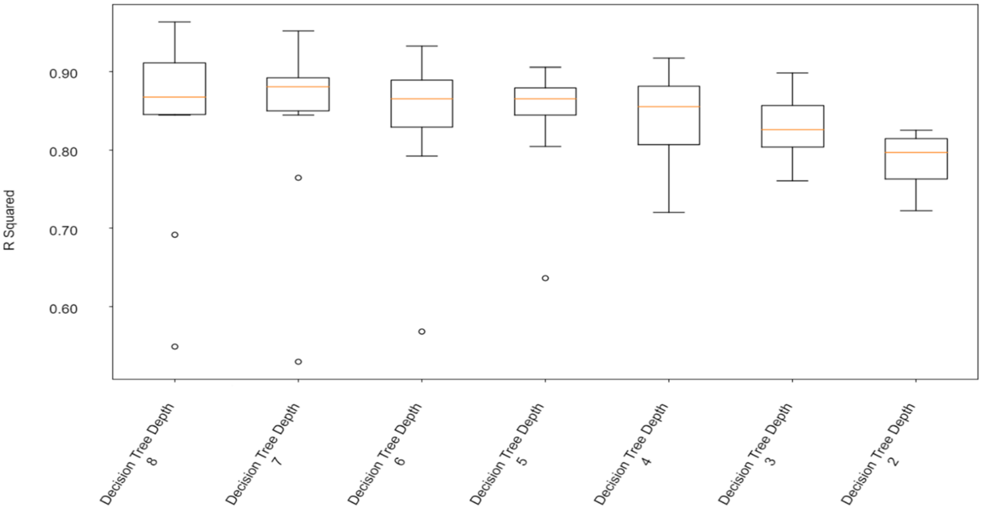


Fig. 6 R^2results under different depths

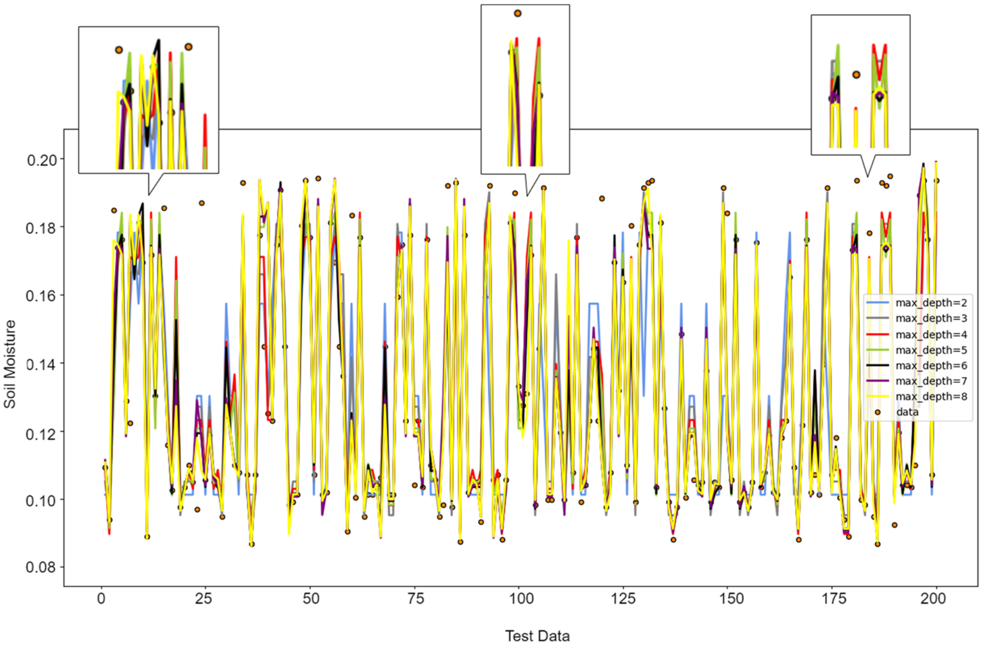


Fig. 7 Estimation results of SM under different depths

Figure 4 presents the inconsistent points, quartile 1, median value, and quartile 3 of MAE results concerning the different decision tree depths in the training stage. Figure 4 shows that the median value (orange line) of MAE lowers as the depth of the decision tree increases. However, the deviation of MAE results increases, the more the depth of the decision tree is. Besides, an increased deviation of MAE may result in overfitting. To avoid overfitting, the value that the depth of decision tree equals to 5 (0.007, 0.0016) is more reasonable because it has the least MAE value compared with 2 (0.011, 0.0009), 3 (0.009, 0.0011), and 4 (0.008, 0.0014) decision tree depths and the least deviation compared with 6 (0.007, 0.0017), 7 (0.007, 0.0018), and 8 (0.007, 0.0021) decision tree depths. The values in the parenthesis show the mean value and standard deviation of MAE in the training stage, respectively.

Figure 5 presents the inconsistent points, quartile 1, median value, and quartile 3 of MSE results concerning the different decision tree depths in the training stage. Additionally, it shows that the median value (orange line) of MSE lowers as the depth of the decision tree increases. However, the deviation of MSE increases, the more the depth of the decision tree is. Besides, an increased deviation of MSE may result in overfitting. To avoid overfitting, the value that the depth of decision tree equals to 5 (0.00019, 0.00008) should be chosen because it has the least MSE value compared with 2 (0.00026,

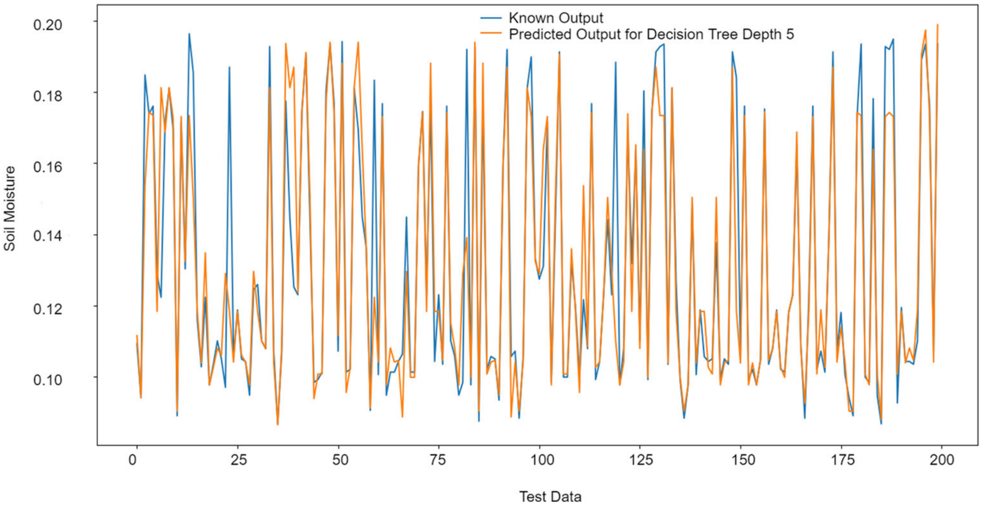


Fig. 8 Estimation result of SM

0.00005), 3 (0.00021, 0.00006), and 4 (0.00020, 0.00007) decision tree depths and the least deviation compared with 6 (0.00020, 0.00010), 7 (0.00019, 0.00012), and 8 (0.00020, 0.00013) decision tree depths. The values in the parenthesis show the mean value and standard deviation of MSE in the training stage, respectively.

Figure 6 presents the inconsistent points, quartile 1, median value, and quartile 3 of R^2 results concerning the different decision tree depths in the training stage. Figure 6 shows that the median value (orange line) of R^2 increases as the depth of the decision tree increases. However, the deviation of R^2 increases, the more the depth of the decision tree is. Besides, an increased deviation of R^2 may result in overfitting. To avoid overfitting, the value that the depth of decision tree equals to 5 (0.842, 0.075) is more reasonable because it has the highest R^2 value compared with 2 (0.785, 0.036), 3 (0.825, 0.042), and 4 (0.842, 0.056) decision tree depths and the least deviation compared with 6 (0.837, 0.097), 7 (0.838, 0.111), and 8 (0.830, 0.121) decision tree depths. Furthermore, R^2 value begins to decrease, and the standard deviation begins to increase after the depth of the decision tree exceeds 5. The values in the parenthesis show the mean value of R^2 and the standard deviation of R^2 in the training stage, respectively.

3.3 Testing the performed algorithm

This paper splits the dataset into two parts for training (80%) and testing (20%) stage. The initial dataset consists of 1000 instances that include four inputs and one output. The testing stage includes 200 instances, and Fig. 7 shows these instances concerning different decision tree depths.

Figure 7 illustrates the estimation results of SM concerning the different decision tree depths in the testing stage. Blue, grey, red, green, black, purple, and yellow colors are denoted two (0.0003, 0.0122, 0.736), three (0.0002, 0.0093, 0.827), four (0.0002, 0.0081, 0.840), five (0.0002, 0.0076, 0.843), six (0.0002, 0.0074, 0.826), seven (0.0002, 0.0069, 0.838), and eight (0.0002, 0.0067, 0.834) decision tree depths, respectively. Estimation results of SM in the testing stage are denoted by these colors. The values in the parenthesis show MSE, MAE, and R^2 in the testing stage, respectively. The circle denotes the observed data that belongs to SM. The depth of the decision tree that equals to five shows the more suitable pattern than the others.

Figure 8 illustrates the estimation result of SM by comparing the predicted values to the known values of SM in the testing data. The performed method of decision tree regression presents decent estimated SM values that cover the known SM values in Fig. 8. A careful examination in Fig. 8 shows that sharp spikes are present in the up and downsides. Almost the sharp spikes in the downside and most of the sharp spikes in the upside are estimated by satisfying fitness criteria.

4 Conclusion

This paper addresses the estimation of SM by performing decision tree regression which is not discussed concerning given parameters before in a research paper. This paper contributes to the literature by the following aspects. First, the application of decision tree regression to estimate SM presents a high R^2 value and low MSE and MAE values. It means that the performed method satisfies three different fitness criteria with higher values. Second contribution is that the proposed methodology provides an accurate estimation under a short computation time. The final contribution is regarding the estimation of SM considering decision tree regression performed with different parameters.

The numerical results show that the performed decision tree regression can handle the data of SM by estimating with satisfying fitness criteria. The obtained results of MSE, MAE, and R^2in the training and the testing stage illustrate that the depth of the decision tree that equals to 5 provides higher fitness values than others.

Further research needs to be conducted to improve the fitness criteria of the estimation of SM. The researchers may apply design of experiment methodology (DOE) to the estimation of SM since the DOE may reduce the required instance number of data. The reduced instance number of data may provide better fitness values. Additionally, other supervised learning methods, including kernel ridge regression, deep learning, SVM, and neural network, may be employed to the estimation parameters of SM in this paper.

References

Ahmad S, Kalra A, Stephen H (2010) Estimating soil moisture using remote sensing data: a machine learning approach. Adv Water Resour 33(1):69–80

Alamry AS, van der Meijde M, Noomen M, Addink EA, van Benthem R, de Jong SM (2017) Spatial and temporal monitoring of soil moisture using surface electrical resistivity tomography in Mediterranean soils. Catena 157:388–396

Badía D, López-García S, Martí C, Ortíz-Perpiñá O, Girona-García A, Casanova-Gascón J (2017) Burn effects on soil properties associated to heat transfer under contrasting moisture content. Sci Total Environ 601:1119–1128

De'ath G, Fabricius KE (2000) Classification and regression trees: a powerful yet simple technique for ecological data analysis. Ecology 81(11):3178–3192

Gill MK, Asefa T, Kemblowski MW, McKee M (2006) Soil moisture prediction using support vector machines 1. J Am Water Resour Assoc 42(4):1033–1046

Gorthi S, Dou H (2011) Prediction models for the estimation of soil moisture content. In: International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, Washington, DC, pp 945–953

Hajdu I, Yule I, Dehghan-Shear MH (2018) Modelling of near-surface soil moisture using machine learning and multi-temporal sentinel 1 images in New Zealand. In: International Geoscience and Remote Sensing Symposium, Valencia, pp 1422–1425

Hajnsek I, Jagdhuber T, Schon H, Papathanassiou KP (2009) Potential of estimating soil moisture under vegetation cover by means of PolSAR. Trans Geosci Remote Sens 47(2):442–454

Han J, Mao K, Xu T, Guo J, Zuo Z, Gao C (2018) A soil moisture estimation framework based on the cart algorithm and its application in china. J Hydrol 563:65–75

https://www.utm.utoronto.ca/geography/resources/environmental-datasets, 24.01.2019

Huisman JA, Sperl C, Bouten W, Verstraten JM (2001) Soil water content measurements at different scales: accuracy of time domain reflectometry and ground-penetrating radar. J Hydrol 245(1-4):48–58

Kumar SV, Dirmeyer PA, Peters-Lidard CD, Bindlish R, Bolten J (2018) Information theoretic evaluation of satellite soil moisture retrievals. Remote Sens Environ 204:392–400

Loh WY (2011) Classification and regression trees. Wiley Interdiscip Rev: Data Min Knowl Discovery 1(1):14–23

Malajner M, Gleich D, Planinsic P (2019) Soil type characterization for moisture estimation using machine learning and UWB-Time of Flight measurements. Measurement 146:537–543

Qu Y, Qian X, Song H, Xing Y, Li Z, Tan J (2018) Soil moisture investigation utilizing machine learning approach based experimental data and Landsat5-TM images: a case study in the Mega City Beijing. Water 10(4):423

Pohn HA, Offield TW, Watson K (1974) Thermal inertia mapping from satellite-discrimination of geologic units in Oman. J Res US Geol Surv 2(2):147–158

Prakash S, Sharma A, Sahu SS (April) Soil moisture prediction using machine learning. In: Second International Conference on Inventive Communication and Computational Technologies, Coimbatore, pp 1–6

Rötzer K, Montzka C, Bogena H, Wagner W, Kerr YH, Kidd R, Vereecken H (2014) Catchment scale validation of SMOS and ASCAT soil moisture products using hydrological modeling and temporal stability analysis. J Hydrol 519:934–946

Sarti M, Mascolo L (2012) An investigation of different polarimetric decomposition techniques for soil moisture estimation. In: Tyrrhenian Workshop on Advances in Radar and Remote Sensing, Naples, pp 209–213

Shukla G, Garg RD, Srivastava HS, Garg PK (2018) An effective implementation and assessment of a random forest classifier as a soil spatial predictive model. Int J Remote Sens 39(8):2637–2669

Wang JR, O'Neill PE, Jackson TJ, Engman ET (1983) Multifrequency measurements of the effects of soil moisture, soil texture, and surface roughness. IEEE Trans Geosci Remote Sens 1:44–51

Zuo ZY, Zhang RH (2008) Spatial and temporal variations of soil moisture in spring in East China. Sci China Earth Sci 38(11):1428–1437

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.