EXTREME GRADIENT BOOSTING REGRESSION MODEL FOR SOIL THERMAL CONDUCTIVITY

by

Ahmet Hasim YURTTAKAL[*]

Department of Computer Technologies, Yozgat Bozok University, Yozgat, Turkey

Original scientific paper
https://doi.org/10.2298/TSCI200612001Y

The thermal conductivity estimation for the soil is an important step for many geothermal applications. But it is a difficult and complicated process since it involves a variety of factors that have significant effects on the thermal conductivity of soils such as soil moisture and granular structure. In this study, regression was performed with the extreme gradient boosting algorithm to develop a model for estimating thermal conductivity value. The performance of the model was measured on the unseen test data. As a result, the proposed algorithm reached 0.18 RMSE, 0.99 R^2, and 3.18% MAE values which state that the algorithm is encouraging.

Key words: *gradient boosting, thermal conductivity, regression*

Introduction

Conduction of heat transfer is the transfer of heat through matter by molecular excitement within the material without bulk motion of the matter. The fundamental formula for the conduction heat transfer is expressed in terms of heat transfer rate by heat conduction law, aka Fourier's law:

$$q = -k\nabla T \tag{1}$$

where k is material's thermal conductivity and ∇T – the local temperature gradient. In other words, the rate of heat transfer through a material is proportional to the temperature gradient [1]. The thermal conductivity value k at the proportion is constant and unique for each material. It determines the heat transfer in matters [2]. Thus, it is crucial to compute its value as accurately as possible, especially in geothermal applications. There has been a vast number of studies on the prediction of k in various fields in the literature [3, 4].

Soil is a material and can be in solid, liquid, and gaseous stage. Its thermal conductivity varies depending on its stage [5]. In other words, soil's thermal connectivity is determined by the contributions of the stages in which materials' properties change, thus, its thermal conductivity depends on the properties that change in time and space. Thus, the thermal conductivity itself is a feature of the soil, and it varies depending on the mineral structure, texture, moisture content, amount of organic matter, grain shape, thermal conductivity, and shape of the aggregates [6].

On the other hand, it is well known that large amounts of organic matter show low thermal conductivity, so, sand soils' thermal conductivity is higher than clay soils. Moreover, heat transfer increases as the volume weight of the soil increases, and the porosity decreases.

[*] Author's e-mail: ahmet.yurttakal@bozok.edu.tr

Thus, the thermal conductivity of dry soil increases with the addition of water to the soil, because liquids transmit heat easily according to air [7].

Knowing the thermal properties of the soil is of great importance for the researchers in soil science, and microclimate as well as in many areas of agricultural engineering. Furthermore, the early growth and development of a crop can be largely determined by the microclimate [8]. It is also important to estimate the thermal conductivity of the soil for geothermal applications such as ground source heat pumps and borehole thermal energy storage [9]. Thus, many researchers spent a considerable amount of time predicting soils' thermal conductivity. He *et al*. [10] developed a model for the prediction of soil conductivity from matric potential. Usowicz *et. al* [11] analyzed soil water content and aggregation status in soil thermal conductivity. Go *et al*. [12] proposed a new empirical model to estimate the thermal conductivity that can be applied to unsaturated granite soils. In another study, He *et al*. [13] presented a comparative analysis in which 20 models were evaluated to estimate solid thermal conductivity. It is worth noting that, in general, but especially for soil, thermal conductivity prediction is a difficult problem since many parameters affect the thermal conductivity. Therefore, the majority of the prediction models have been developed for specific soil types [14].

In this study, the gradient boosting regression method was employed to develop a model for predicting the thermal conductivity of soils. The model uses soils' dry density, porosity, saturation degree, quartz content, sand content, and clay content as input and returns an empirical coefficient as the output, an estimate to the thermal conductivity constant. Training the model was repeated so that the optimum hyperparameters were determined, then regression was performed. Error squared value R^2 for the model is 0.9943 in training data and 0.8067 in test data.

Material and methods

Regression analysis is used to describe possible relationship between two or more variables. Regression shows the functional form of the linear relationship between two or more variables, but provides estimation about the other when the value of one of the variables is known [15]. The data set used and the recommended regression model are explained in this section.

Dataset

In training the gradient boosting regression method based predictive model, the dataset created by Zhang *et al*. [14] was used. The dataset contains 257 thermal conductivity measurements collected from the studies by Chen [16], Zhang *et al*. [17], Tarnawski *et al*. [18], McCombie *et al*. [19], Tarnawski *et al*. [20, 21], and Tokoro *et al*. [22]. Table 1 presents some statistical outputs such as mean, standard deviation, minimum and maximum of each input values of dry density, porosity, saturation degree, quartz content, sand content, clay content, and thermal conductivity measurements in the dataset. These values belong to different type of soils.

Table 1. Some statistics for the dataset

	ρ_d [gcm^{-3}]	η	S_r	m_q	m_s	m_c	k [Wm^{-1}K^{-1}]	k_{sat} [Wm^{-1}K^{-1}]	k_{dry} [Wm^{-1}K^{-1}]	κ
Mean	1.44	0.46	0.33	0.61	0.62	0.09	1.17	1.95	0.24	3.24
Std	0.18	0.07	0.22	0.31	0.39	0.11	0.63	0.68	0.07	2.39
Min	0.98	0.35	0.04	0.00	0.00	0.00	0.10	0.39	0.09	0.20
Max	1.83	0.63	0.70	1.00	1.00	0.42	2.96	3.29	0.56	14.2

Extreme gradient boosting (XGBoost)

Recall that boosting is an ensemble-based learning algorithm. It returns different weights for training data distribution after each iteration. Every boosting iteration adds weight to the miss classified error sample but subtracts from the correctly classified sample, so it effectively changes the training data distribution [23]. On the other side, extreme gradient boosting is a combination of gradient descent and boosting, named gradient boosting machine (GBM). The GBM uses second order gradient statistics to minimize the optimization problem:

$$\delta(\phi) = \sum_i l(y_{\text{true}}, y_{\text{pred}}) + \sum_k \Omega(f_k)$$

with

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \|w\|^2 \tag{2}$$

where l is the loss function and Ω – the regularization function. The loss function l is a differentiable convex funaction and measures the difference between the prediction y_{pred} and the target y_{true}. Besides, the regularization function Ω penalizes the complexity of the model. As a tree-based algorithm, GBM is purposed to find the best candidate split points, which is non-trivial for large dataset. Chen and Guestrin [24] purpose a novel distributed weighted quantile sketch algorithm that can handle weighted data with a provable theoretical guarantee, resulting a new scalable and efficient algorithm called extreme gradient boosting (XGBoost). The XGBoost is also provided in many programming languages such as R and Python.

Model evaluation

In the implementation of the model, ρ_{d} [gcm^{-3}], η, S_{r}, m_{q}, m_{s}, m_{c} variables are accepted as input parameters, while κ empirical coefficient parameter is the output. In calculating empirical coefficient κ, it was used two normalization formulas for thermal conductivity. One of these, suggested by Johansen [25], states that the normalized thermal conductivity value k_r:

$$k_r = \frac{k - k_{\text{dry}}}{k_{\text{sat}} - k_{\text{dry}}} \tag{3}$$

where k is thermal conductivity value, and k_{sat} and k_{dry} are soils' saturated and dry values, respectively.

On the other hand, the second formula is given by Cote and Konrad [26] and says:

$$k_r = \frac{\kappa S_r}{1 + (\kappa - 1)S_r} \tag{4}$$

where S_r is saturation degree.

Combining these two formulas yields the formula for κ:

$$\kappa = \frac{(1 - S_r)(k - k_{\text{dry}})}{S_r(k_{\text{sat}} - k)} \tag{5}$$

Model performance is measured with root mean square error (RMSE), mean absolute percentage error (MAPE), R^2, Akaika information criterion corrected (AICc) and bayesian information criterion (BIC) metrics. Now, we briefly recall each of these metrics.

The MAPE is the average relative error and used to measure prediction accuracy in various predictive methods, as in regression problems. The mathematical definition is given:

$$MAPE = \frac{1}{n}\sum_{t=1}^{n} \left| \frac{y_{\text{true}} - y_{\text{pred}}}{y_{\text{true}}} \right| \tag{6}$$

where y_{true} is the true value, y_{pred} – the predicted value, and n – the number of predictions.

The RMSE is the standard deviation of prediction errors and measures how spread out these prediction errors are:

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}\left(y_{\text{true}} - y_{\text{pred}}\right)^{2}} \tag{7}$$

The R^2 is the proportion of the variance in the dependent variable that is predictable from the independent variables:

$$R^{2} = 1 - \frac{\sum_{i}(y_{\text{true}} - y_{\text{pred}})^{2}}{\sum_{i}\left(y_{\text{true}} - \frac{1}{n}\sum_{i=1}^{n} y_{\text{true}}\right)^{2}} \tag{8}$$

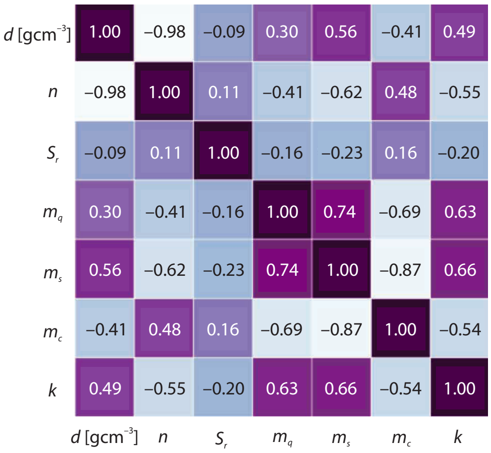
The AICc is a statistical metric to compare the quality of statistical models to each other. It takes each model under consideration, and rank them best to worst. In other words, it relatively measures models trained on a dataset:

$$AICc = n\log\left[\sum(y_{\text{true}} - y_{\text{pred}})^{2}\right] + 2n_{p} + \frac{2n_{p}(n_{p}+1)}{n_{y} - n_{p} - 1} \tag{9}$$

In the formula n_p, n_y is the number of parameters and sample size, respectively.

Finally, BIC is also used for model selection among a finite number of models. It partially depends on the likelihood function. It is closely related to AICc. Its mathematical definition is given:

$$BIC = n\log\left[\sum(y_{true} - y_{pred})^{2}\right] + n_{p}\log(n) \tag{10}$$



	d [gcm^{-3}]	n	S_r	m_q	m_s	m_c	k
d [gcm^{-3}]	1.00	−0.98	−0.09	0.30	0.56	−0.41	0.49
n	−0.98	1.00	0.11	−0.41	−0.62	0.48	−0.55
S_r	−0.09	0.11	1.00	−0.16	−0.23	0.16	−0.20
m_q	0.30	−0.41	−0.16	1.00	0.74	−0.69	0.63
m_s	0.56	−0.62	−0.23	0.74	1.00	−0.87	0.66
m_c	−0.41	0.48	0.16	−0.69	−0.87	1.00	−0.54
k	0.49	−0.55	−0.20	0.63	0.66	−0.54	1.00

Figure 1. Variables correlation

Experimental results

The proposed method was implemented in an open source Python environment. Figure 1, the correlation levels of the variables are given. While there is a positive high correlation between quartz content and sand content, there is a negative high correlation between dry density and porosity.

While 10% of the data set selected randomly is reserved for testing without being included in the training, 80% of the remaining part is reserved for training and 20% for validation. After that, hyper parameter tuning was done. The optimum parameters of XGBoost algorithms with LightGBM are given in tab. 2.

Table 2. Optimum hyper-parameters

	Hyper-parameters
LightGBM	n_estimators = 580, learning_rate = 0.03, num_iterations = 1000, boosting_type = 'dart', min_data_in_leaf = 1, max_depth = 6, num_leaves = 32
XGBoost	learning_rate = 0.21, n_jobs = 4, n_estimators = 640, max_depth = 5, min_child_weight = 5, subsample = 0.8, reg_alpha = 0.01, colsample_bytree = 0.8, gamma = 0

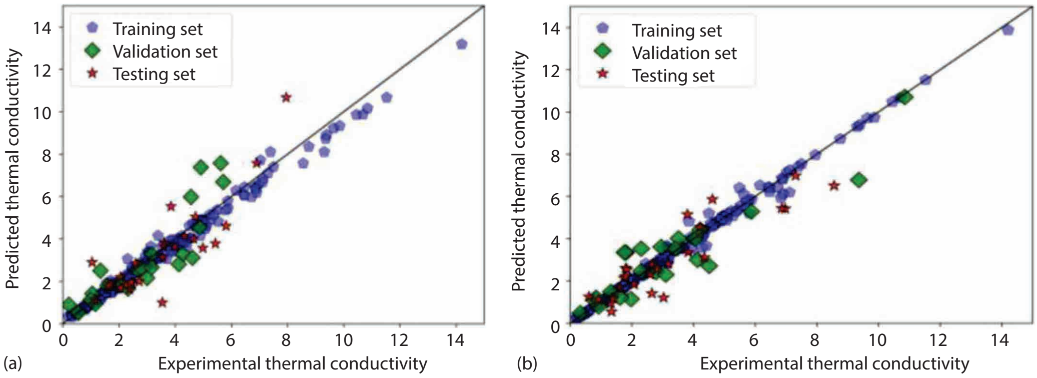


Figure 2. Regression-line; (a) LightGBM and (b) XGBoost

According to the optimum hyper parameters obtained, fig. 2(a) shows the regression-line obtained with the LightGBM algorithm, while fig. 2(b) shows the regression-line obtained with the XGBoost. It appears that the XGBoost algorithm performs better in regression.

In fig. 3, the order of importance of features in regression is given. In the regression process, clay content, quarz content and saturation degree parameters have more impact on performance, respectively.

The metrics obtained according to the training and test data are given in tab. 3. It showed higher performance in XGBoost regression process than LightGBM algorithm.

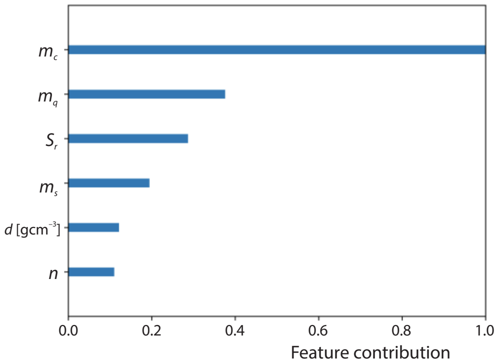


Figure 3. Feature importance

Table 3. Regression performance metrics

	LightGBM		XGBoost	
	Train Set	Test Set	Train Set	Test Set
RMSE	0.3073	1.0701	0.1811	0.9180
MAE [%]	7.40	24.94	3.18	26.90
R^2	0.9842	0.6052	0.9943	0.8067
AICc	885.078	146.796	743.069	154.334
BIC	864.048	134.826	722.039	142.365

Lu *et al.* [27], Barry *et al.* [28], and Cote and Konrad [24] made estimation analysis on six different scenarios according to the value ranges of quarz content, sand content saturation degree parameters. The proposed method made a regression considering all parameters without considering the value ranges of the parameters. Zhang *et al.* [14] reached 941 AICc value in its regression process considering all parameters, while the recommended method reached 743.069 AICc.

The study made important contributions to the literature. Regression was performed with the XGBoost algorithm. The results were compared with the LightGBM algorithm. In the regression process, optimum parameters were determined by hyper tuning. The researchers were given an idea about which parameters could be used. Input parameters affecting the most regression were determined. Model performance was analyzed on unseen test data. Regression was performed considering all input parameters. High performance metrics are achieved.

Conclusion

Throughout the study, we developed a predictive model for the thermal connectivity constants for different types of soils. The model was built on the gradient boosting regression method. The prediction is made by the features of dry density, porosity, saturation degree, quartz content, sand content, and clay content. The model training was repeated by fine-tuning hyperparameters until almost optimum hyperparameters were determined, then the regression was applied. The XGBoost algorithm achieved 0.18 RMSE, 0.99 R^2, 3.18% MAE on the training set with these calculated almost optimum parameters. Moreover, error squared value R^2 for the model was 0.9943 in training data and 0.8067 in test data. The model received the best AICc value of 743.069 in the literature.

Nomenclature

k – thermal conductivity, [Wm^{-1}K^{-1}]
k_{dry} – thermal conductivity of dry soil, [Wm^{-1}K^{-1}]
k_{sat} – thermal conductivity of saturated soil, [Wm^{-1}K^{-1}]
m_{c} – clay content
m_{q} – quarz content
m_{s} – sand content
S_r – saturation degree
R^2 – R squared error

Greek symbol

η – porosity
κ – empirical coefficient
ρ_{d} – dry density, [gcm^{-3}]

Acronyms

AICc – akaika information criterion corrected
BIC – bayesian information criterion
MAE – mean absolute error
RMSE – root mean square error

References

[1] Bergman, T. L., *et al.*, *Fundamentals of Heat and Mass Transfer*, John Wiley and Sons, New York, USA, 2011

[2] Grigull, U., Sandner, H., *Heat Conduction*, Springer-Verlag, Berlin, 1984

[3] Li, X., *et al.*, Thermal Properties of Evaporitic Rocks and their Geothermal Effects on the Kuqa Foreland Basin, Northwest China, *Geothermics*, *88* (2020), 101898

[4] Do, T. M., *et al.*, Thermal Conductivity of Controlled Low Strength Material (CLSM) under Various Degrees of Saturation Using a Modified Pressure Plate Extractor Apparatus – A Case Study for Geothermal Systems, *Applied Thermal Engineering*, *143* (2018), Oct., pp. 607-613

[5] Yun, T. S., Santamarina, J. C., Fundamental Study of Thermal Conduction in Dry Soils, *Granular Matter*, *10* (2008), 3, 197

[6] Nassar, I., *et al.*, Simultaneous Transfer of Heat, Water, and Solute in Porous Media: II. Experiment and Analysis, *Soil Science Society of America Journal*, *56* (1992), 5, pp. 1357-1365

[7] Abu-Hamdeh, N. H., Reeder, R. C., Soil Thermal Conductivity Effects of Density, Moisture, Salt Concentration, and Organic Matter, *Soil Science Society of America Journal*, *64* (2000), 4, pp. 1285-1290

[8] Oladunjoye, M., *et al.*, Variability of Soil Thermal Properties of a Seasonally Cultivated Agricultural Teaching and Research Farm, University of Ibadan, South-Western Nigeria, *Global Journal of Science Frontier Research Agriculture and Veterinary*, *13* (2013), 8, pp. 41-64

[9] Zhang, N., Wang, Z., Review of Soil Thermal Conductivity and Predictive Models, *International Journal of Thermal Sciences*, *117* (2017), July, pp. 172-183

[10] He, H., *et al.*, A New Model for Predicting Soil Thermal Conductivity from Matric Potential, *Journal of Hydrology*, *589* (2020), Oct., 125167

[11] Usowicz, B., *et al.*, Effects of Aggregate Size on Soil Thermal Conductivity: Comparison of Measured and Model-Predicted Data, *International Journal of Heat and Mass Transfer*, *57* (2013), 2, pp. 536-541

[12] Go, G.-H., *et al*., A Reliable Model to Predict Thermal Conductivity of Unsaturated Weathered Granite Soils, *International Communications in Heat and Mass Transfer*, *74* (2016), May, pp. 82-90
[13] He, H., *et al.*, Modelling of Soil Solid Thermal Conductivity, *International Communications in Heat and Mass Transfer*, *116* (2020), 104602
[14] Zhang, N., *et al*., A Unified Soil Thermal Conductivity Model Based on Artificial Neural Network, *International Journal of Thermal Sciences*, *155* (2020), 106414
[15] Draper, N. R., Smith, H., *Applied Regression Analysis*, John Wiley and Sons, New York, USA, 1998
[16] Chen, S. X., Thermal Conductivity of Sands, *Heat and Mass Transfer*, *44* (2008), 10, 1241
[17] Zhang, N., *et al*., Thermal Conductivity of Quartz Sands by Thermo-Time Domain Reflectometry Probe and Model Prediction, *Journal of Materials in Civil Engineering*, *27* (2015), 12, 04015059
[18] Tarnawski, V., *et al*., Canadian Field Soils III. Thermal-Conductivity Data and Modelling, *International Journal of Thermophysics*, *36* (2015), 1, pp. 119-156
[19] McCombie, M., *et al*., Thermal Conductivity of Pyroclastic Soil (Pozzolana) from the Environs of Rome, *International Journal of Thermophysics*, *38* (2017), 2, 21
[20] Tarnawski, V., *et al*., Volcanic Soils: Inverse Modelling of Thermal Conductivity Data, *International Journal of Thermophysics*, *40* (2019), 2, 14
[21] Tarnawski, V., *et al*., Thermal Conductivity of Standard Sands: Part III – Full Range of Saturation, *International Journal of Thermophysics*, *34* (2013), 6, pp. 1130-1147
[22] Tokoro, T., *et al*., Estimation Methods for Thermal Conductivity of Sandy Soil with Electrical Characteristics, *Soils and Foundations*, *56* (2016), 5, pp. 927-936
[23] Bisri, A., Wahono, R. S., Penerapan Adaboost untuk penyelesaian ketidakseimbangan kelas pada Penentuan kelulusan mahasiswa dengan metode Decision Tree, *Journal of Intelligent Systems*, *1* (2015), 1, pp. 27-32
[24] Chen, T.,Guestrin, C., The Xgboost: A Scalable Tree Boosting System, *Proceedings*, 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, Cal., USA, 2016, pp. 785-794
[25] Johansen, O., Thermal Conductivity of Soils, Ph. D. thesis, Trondheim, Norway (CRREL Draft Translation 637, 1977) ADA, 44002, 1975
[26] Cote, J., Konrad, J.-M., A Generalized Thermal Conductivity Model for Soils and Construction Materials, *Canadian Geotechnical Journal*, *42* (2005), 2, pp. 443-458
[27] Lu, S., *et al*., An Improved Model for Predicting Soil Thermal Conductivity from Water Content at Room Temperature, *Soil Science Society of America Journal*, *71* (2007), 1, pp. 8-14
[28] Barry-Macaulay, D., *et al*., Evaluation of Soil Thermal Conductivity Models, *Canadian Geotechnical Journal*, *52* (2015), 11, pp. 1892-1900

Paper submitted: June 12, 2020
Paper revised: November 11, 2020
Paper accepted: November 15, 2020