Journal of Hydrology 529 (2015) 1590–1600

Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol

Estimation of soil cation exchange capacity using Genetic Expression Programming (GEP) and Multivariate Adaptive Regression Splines (MARS)

S. Emamgolizadeh [a], S.M. Bateni [b,*], D. Shahsavani [c], T. Ashrafi [a], H. Ghorbani [a]

[a] Department of Water and Soil, Shahrood University, Shahrood, Iran
[b] Department of Civil and Environmental Engineering and Water Resources Research Center, University of Hawaii at Manoa, Honolulu, HI, USA
[c] Department of Mathematical Sciences, Shahrood University, Shahrood, Iran

ARTICLE INFO

Article history:
Received 13 June 2015
Received in revised form 9 August 2015
Accepted 11 August 2015
Available online 20 August 2015
This manuscript was handled by Geoff Syme, Editor-in-Chief

Keywords:
Soil cation exchange capacity
Genetic expression programming
Multivariate adaptive regression splines

SUMMARY

The soil cation exchange capacity (CEC) is one of the main soil chemical properties, which is required in various fields such as environmental and agricultural engineering as well as soil science. *In situ* measurement of CEC is time consuming and costly. Hence, numerous studies have used traditional regression-based techniques to estimate CEC from more easily measurable soil parameters (e.g., soil texture, organic matter (OM), and pH). However, these models may not be able to adequately capture the complex and highly nonlinear relationship between CEC and its influential soil variables. In this study, Genetic Expression Programming (GEP) and Multivariate Adaptive Regression Splines (MARS) were employed to estimate CEC from more readily measurable soil physical and chemical variables (e.g., OM, clay, and pH) by developing functional relations. The GEP- and MARS-based functional relations were tested at two field sites in Iran. Results showed that GEP and MARS can provide reliable estimates of CEC. Also, it was found that the MARS model (with root-mean-square-error (RMSE) of 0.318 $Cmol^+ kg^{-1}$ and correlation coefficient (R^2) of 0.864) generated slightly better results than the GEP model (with RMSE of 0.270 $Cmol^+ kg^{-1}$ and R^2 of 0.807). The performance of GEP and MARS models was compared with two existing approaches, namely artificial neural network (ANN) and multiple linear regression (MLR). The comparison indicated that MARS and GEP outperformed the MLP model, but they did not perform as good as ANN. Finally, a sensitivity analysis was conducted to determine the most and the least influential variables affecting CEC. It was found that OM and pH have the most and least significant effect on CEC, respectively.

1. Introduction

The soil cation exchange capacity (CEC) is defined as the total exchangeable cations that a soil can hold by electrostatic forces at a specific pH (Bauer and Velde, 2014). Its accurate determination is vital in soil science and environmental studies (Manrique et al., 1991; Keller et al., 2001; Belachew and Abera, 2010). CEC directly affects soil fertility by controlling the exchange of ions on the clay surfaces (Belachew and Abera, 2010). A low CEC value implies that the soil is able to hold only a small amount of nutrients that are applied through fertilization. Hence, in a soil with low CEC, the availability of nutrients to plants and microorganisms is limited (Molloy, 2007).

In situ measurement of CEC (especially in areas with large quantities of calcium carbonate and gypsum contents) is tedious, expensive and labor extensive (Carpena et al., 1972; Fernando et al., 1977; McBratney et al., 2002; Amini et al., 2005). Therefore, numerous studies have tried to empirically relate CEC to more easily measurable soil physical and chemical properties such as soil texture (more specifically clay content), soil pH, and organic matter (OM) via multiple linear regression (MLR) models. These regression-based empirical models led to the so-called pedotransfer functions (PTFs) (Drake and Motto, 1982; Breeuwsma et al., 1986; Manrique et al., 1991; Bell and Van Keulen, 1995; McBratney et al., 2002; Bortoluzzi et al., 2006; Ghorbani et al., 2015).

The main shortcoming of traditional regression-based PTFs is that they typically yield large errors because they may not be able to adequately capture the highly nonlinear and complex relationship between CEC and the relevant soil physical and chemical

* Corresponding author.
E-mail address: smbateni@hawaii.edu (S.M. Bateni).

http://dx.doi.org/10.1016/j.jhydrol.2015.08.025
0022-1694/

properties (e.g., clay, OM, pH, etc.). The inability of conventional regression-based PTFs to estimate accurate CEC values is mainly because they are obtained by assuming *a priori* a specific type of function between inputs (e.g., clay, pH, OM, bulk density, etc.) and output (CEC).

In contrast to the MLR models, several studies have used artificial neural networks (ANN) to model PFTs (Schaap et al., 1998; Amini et al., 2005; Keshavarzi and Sarmadian, 2010; Kashi et al., 2014). These studies showed that ANN performs better than the MLR models. However, ANN has its own shortcoming: It acts like a black-box and relates CEC to the relevant soil variables via a complex network that is composed of transfer functions and many coefficients.

Due to the above-mentioned shortcomings of MLR and ANNs, Genetic Expression Programming (GEP) and Multivariate Adaptive Regression Splines (MARS) are used in this study to develop PTFs. GEP is an optimization technique that was invented by Ferreira (2001). It can accurately capture the nonlinear and complex relationship between a response variable and its predictors. MARS is a non-parametric model that was developed by Friedman (1991). To the best of our knowledge, no other study has used GEP and/or MARS to establish PTFs. GEP and MARS can overcome the aforementioned deficiencies of the MLR and ANNs approaches and have distinct advantages: (1) they do not act like a black-box, (2) they provide an equation (a functional relation in which the dependent variable is stated directly in terms of the independent variables) between inputs and outputs, (3) they do not need to assume *a priori* a specific form of function to characterize the physics of the underlying problem, (4) they can capture the complex and nonlinear relationship between a response variable and its predictors and yield accurate results, and finally (5) they are more flexible than the traditional linear and nonlinear regression techniques and usually can overcome their limitations (Johari et al., 2006; Samui et al., 2011; Gandomi and Alavi, 2011; Zhang and Goh, 2013).

Several studies have recently used GEP and MARS to identify complex relationships between inputs and outputs in numerous engineering problems (Yang et al., 2003; Zakaria et al., 2010; Gutiérrez et al., 2011; Kayadelen, 2011; Samadianfard et al., 2012; Landeras et al., 2012; Sattar, 2013). Yang et al. (2004) applied MARS and ANN models to estimate soil temperature at different depths and found that MARS outperformed ANN. Quirós et al. (2009) showed that MARS can classify land covers in a test zone in the south western of Spain more accurately that the parallelepiped and maximum likelihood (ML) methods. Kisi and Shiri (2012) utilized GEP and ANN to assess the daily suspended sediment concentration in the Eel River (California, USA). Their findings indicated that GEP outperformed ANN. Sattar (2013) showed that GEP can reliably estimate the critical shear stress of cohesive soils from their mineral contents. As mentioned above, GEP and MARS have unique characteristics and their performance is comparable or even better than the commonly used methods in engineering problems. The primary objective of this study is to develop and test GEP- and MARS-based PTFs. It is worth noting that this is the first study which explores the potential ability of GEP and MARS to model PTFs and estimate CEC. The secondary goal is to compare performance of the two developed PTFs with each other and the existing PTFs in literature.

2. Data, methods and models

2.1. Study sites and data

This study uses soil physical and chemical data collected from two field sites in Iran to develop GEP- and MARS-based PTFs. The sites are located in the city of Semnan (35°34′ N to 35°34.9′ N and 53°28′ E to 53°28.9′ E) and Taybad (34°41.9′ N to 34°42′ N and 60°46′ E to 60°46.9′ E). Each site is about 400 hectares (Fig. 1). The temperature and soil moisture regimes in both sites are thermic and aridic, respectively. Most soils of the study area are Entisols and Aridisols (USDA, soil taxonomy 2010).

Five hundreds soil samples (250 samples in each site) were collected from the top 30 cm of soil profile in the two field sites. In the samples, soil CEC, texture (percent of sand, silt, and clay), pH, and OM percentage were measured by Bower's method (Sparks et al., 1996), hydrometer technique (Gee and Bauder, 1986), pH-meter and the Walkley–Black approach (Walkley and Black, 1934; Nelson and Sommers, 1982), respectively. To have an overview of the measured variables (i.e., percent of sand, silt, clay, OM, pH, and CEC), their statistical indices are shown in Table 1. As indicated in Table 1, clay fraction and OM content of soil samples vary from 1.2% to 3.72% and 10% to 27%, respectively. The measured pH values show that all soil samples fall in the alkalinity range. This is

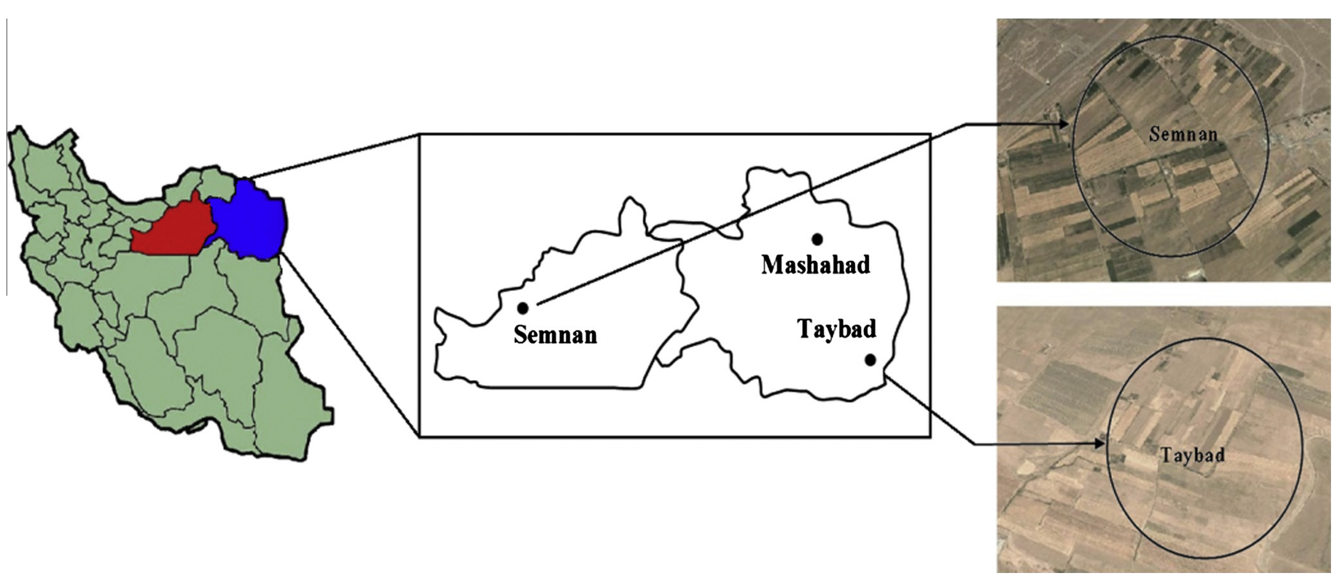

Fig. 1. Graphical location of the sites in the Province of Razavi Khorasan (filled by blue color) and Semnan (filled by red color) in Iran. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)