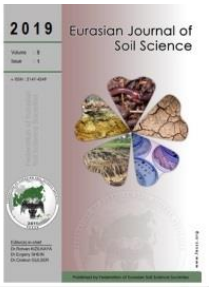
Eurasian J Soil Sci 2019, 8 (3) 196 - 207

Eurasian Journal of Soil Science

Journal homepage : http://ejss.fesss.org



Pedo-transfer functions with multiple linear regressions to predict solute-transport parameters

Md. Abdul Mojid [a,*], A.B.M. Zahid Hossain [b], Guido C.L. Wyseure [c], Md. Ali Ashraf [d]

[a] Department of Irrigation and Water Management, Bangladesh Agricultural University, Mymensingh, Bangladesh
[b] Irrigation and Water Management Division, Bangladesh Rice Research Institute, Gazipur, Bangladesh
[c] Division of Soil and Water Management, Department of Earth and Environmental Sciences, K.U. Leuven, Belgium
[d] Department of Farm Structure and Environmental Engineering, Bangladesh Agricultural University, Bangladesh

Article Info

Received : 11.10.2018
Accepted : 15.04.2019

Abstract

Transport parameters of soluble chemicals through soils are needed to assess the pollution risks of soil and groundwater resources. But, it is time consuming, laborious, expensive and, practically, impossible to experimentally measure such parameters for a wide range of solutes and soil types. So, indirect estimate of the parameters by pedo-transfer function is becoming popular. The aim of this study was to develop and evaluate pedo-transfer functions (PTFs) for solute-transport parameters by multiple linear regression (MLR) analysis. For this, transport parameters of three heavy metal /metalloid compounds ($NaAsO_2$, $Pb(NO_3)_2$, $Cd(NO_3)_2$), a pesticide (carbendazim) and an inert salt ($CaCl_2$) through 14 agricultural soils of Bangladesh were determined. The transport experiments were done in repacked soil columns under unsaturated steady-state water flow conditions. Breakthrough data of the solutes were measured with time-domain reflectometry (TDR), and velocity (V), dispersion coefficient (D) and retardation factor (R) of the solutes were determined by analyzing the data by a transfer-function method. Bulk density (γ), organic carbon (OC) content, clay (C) content, pH, median grain diameter (D_{50}) and uniformity coefficient (C_u) of the soils were determined. Regression models for V, D and R were developed with γ, OC, C, pH, D_{50} and C_u as the input variables. Bulk density and clay content were found the most sensitive input variables to the MLR models. The MLR models fairly predicted V, D and R, and thus provide a way of significantly enhancing prediction of reactive solute transport through agricultural soils.

Keywords: Soluble chemicals, soil properties, solute movement, indirect estimate.

Introduction

Pollution of agricultural soils by heavy metals and pesticide residues occurs through the application of chemical fertilizers, especially phosphate fertilizers, and pesticides. The residues of these chemicals contaminate the soil and water (both surface and groundwater), enter the food chain and cause threat to human and animal health. Industrial effluents and irrigation with wastewater further degrade soil and water quality. The characterization of soluble-chemical transport through soils is an important aspect to assess the pollution of soil and groundwater resources (Porro et al., 1993). Usually, simulation models are used to quantify solute transport through subsurface as tools to implement improved agricultural management. Solute-transport parameters, such as velocity of transport, dispersion coefficient, dispersivity and retardation factor, are among the most crucial inputs for the simulation models. Success of these models depends on our ability to properly quantify the input solute-transport parameters.

* Corresponding author.
Department of Irrigation and Water Management, Bangladesh Agricultural University, Mymensingh 2202, Bangladesh
Tel.: +880 91 67401 E-mail address: ma_mojid@yahoo.com
e-ISSN: 2147-4249 DOI: 10.18393/ejss.554942

The estimation of solute-transport parameters in soils is generally done by fitting measured breakthrough curves (BTC) of solutes to analytical solutions of the convection-dispersion equation. BTCs are constructed with the concentrations of effluent or a proxy measurement of concentration like TDR-measured electrical conductivity from leaching experiments. Such measurements are however time consuming, laborious, expensive and, practically, impossible to obtain BTCs for a wide range of solutes and soil types to sample temporal and spatial variations. So, indirect approaches are needed to predict solute-transport parameters with pedo-transfer functions (PTFs), which use basic soil properties that are often routinely available from soil survey information (Bouma, 1989). The general purpose of developing PTFs is to establish predictive models using databases of soil properties, which contain suitable predictors (basic soil properties) and desired predictands (estimated less available soil properties).

Although widely used to predict unsaturated hydraulic properties of soils (e.g., Vereecken, 1992; Gonçalves et al., 1997), PTFs are yet not well-developed and very familiar to predict solute-transport parameters. Recent developments in this field include parameterizations of solute transport, heat exchange, soil respiration, organic carbon content, root density and water uptake by vegetation (Van Looy et al., 2017). Since it is difficult to relate and compare the physical meaning of specific model parameters, there are only limited PTFs for predicting solute-transport parameters from basic soil characteristics. Several studies on PTFs for adsorption isotherm parameters mostly concern contaminants such as heavy metals (e.g., Horn et al., 2006) or excess pesticides (e.g., Kodešová et al., 2011; Moeys et al., 2011) and fertilizers (e.g., Achat et al., 2016). All these studies include soil organic carbon content as a predictor. Soil pH and clay content were reported as the other common predictors for PTFs of adsorption properties. Perfect et al. (2002) predicted dispersivity using PTFs across a range of soil textures and could explain 50% of its total variation by the parameters of soil-water retention curves using step-wise multiple regressions. Alibuyog (2007), by using PTFs from multiple linear regressions, showed great potential in predicting pore-water velocity, dispersion coefficient and dispersivity from soil physical properties than from water retention parameters. In his observations, using soil properties as predictors, the PTFs could account for more than 50% of the total variation of pore-water velocity, dispersion coefficient and dispersivity.

Predictions of solute-transport parameters, instead of direct measurements, may be accurate enough for many applications. It is therefore worthwhile to analyze databases in such a way that solute-transport parameters can be predicted from easily-measured soil properties. But, extrapolation of PTFs in different agropedoclimatic contexts limits their performance (Touil et al., 2016). So, attempts to develop solute transport PTFs have, so far, been mostly kept to small, local data sets and specific models since the local PTFs are important to properly investigate the relation between the predictors and predictands, and they could be useful in meeting the local agricultural requirements for modeling with reasonable accuracy. The prediction of transport parameters at the local scale is also a first step to simulate subsurface solute movement over larger areas (Gonçalves et al., 2001).

Regression technique is widely used to determine the relationship between predictors and predictands because of its simplicity. It can use linear regressions or nonlinear regressions depending on the expected relationship among the variables (Mojid et al., 2018). The advantage of regression analysis is that it is straightforward to carry out and easy to employ. The disadvantage is that the regression equations (e.g., linear, logarithmic or exponential) and predictors must be determined as a priori and that the relationships between soil properties and predictors may be different in different portions of the database (Van Looy et al., 2017). However, improved multiple linear regression (MLR) can be an efficient and reliable method (Touil et al., 2016) if the relationship between the dependent and independent variables is not complex.

Studies of pedo-transfer function may undertake one of two primary purposes: research or application. Investigators who intend to advance research knowledge may find it more desirable that the model is flexible and can work efficiently with various data sizes and types. Or they may intend to help mine auxiliary information (e.g., importance of input variables) given the structure or features of the model (Van Looy et al., 2017). Our study addressed the research issue. The objectives were: (i) to develop pedo-transfer functions for velocity, dispersion coefficient and retardation factor of four reactive solutes and a non-reactive solute with basic soil properties by multiple linear regression analysis and (ii) to evaluate performance of the pedo-transfer functions.

Material and Methods

Soil sampling and solute-transport measurement

Fourteen (14) soil samples of adequate quantity were collected from different locations of Bangladesh under intensive agricultural activities. The plowed upper soil layers (0–15 cm) were used in solute-transport experiments to reduce variability due to heterogeneity. The soil samples were air dried and sieved to pass through a 2-mm mesh sieve after crushing. Sub-samples from the sampled soils were analyzed for particle size distribution, gradation, pH and organic carbon (*OC*) by employing standard methods. Details of soil sampling and analysis of the samples are reported in Mojid et al. (2018).

The procedures of solute-transport experiments are described here in brief. For details, the readers are referred to Mojid et al. (2016). The experiments were done in four PVC columns (hereafter called experimental columns), each 34 cm long and 15 cm in inner diameter. These columns were filled with four of the air-dried and sieved soils under investigation in the first batch. Each column was packed to 32 cm depth. The soil columns were conditioned by leaching sufficient quantity of tap water (EC = 17 mS m^{-1}) following six wetting and drying cycles during a nine-month period. The soil columns were transferred and placed vertically and axially on four 1.2-m high supporting soil columns to simulate a thick natural soil profile. Two 3-wire TDR sensors (10 cm long and 3 cm spacing with 0.2 cm wire diameter) were inserted horizontally to each column during preparing the soil columns. One sensor was at 8 cm and the other at 28 cm below the top of the upper column; the vertical distance between the two sensors (Z) was 20 cm. A cartridge pump applied tap water through fine needles at constant rate (0.32 ± 0.02 cm h^{-1}), which was considerably lesser than the saturated hydraulic conductivities (≥0.64 cm h^{-1}) of the soils to ensure unsaturated flow. The pump distributed the applied water uniformly over the soil surface of each column through nine fine needles uniformly spaced with a PVC cap on each soil column. Water flow continued until equilibrium between the applied and drainage water was attained. A constant hanging water table, maintained at 20 cm above the base of the supporting (lower) columns, created suction in soils of the experimental (upper) columns.

First, $CaCl_2$ was used in the breakthrough experiments; it helped retaining structure of the soils during the transport experiments. At steady-state water flow condition, a pulse of 5 ml $CaCl_2$ solution (250 g l^{-1}) was introduced uniformly on each column with a syringe attached to a fine needle. The water flow (0.32 ± 0.02 cm h^{-1}) continued until the solution was completely eluted from the upper columns. A TDR100 and CR10X data logger were programmed to record water content and bulk EC of the soils at fixed interval (40, 50 or 60 minutes depending on the solute and soil types). The measurements continued until whole of the applied solute leached out from the upper columns. Measurements of water content and bulk EC were done for $CaCl_2$, $NaAsO_2$, $Pb(NO_3)_2$, $Cd(NO_3)_2$ and carbendazim. Carbendazim is a granular organic solute, which is extensively used in Bangladesh as fungicide. The molecular weight of carbendazim is 191.2 g mol^{-1} and its solubility in water at pH 7–8 is 5–7 mg l^{-1}. The pulse volume and concentration of $NaAsO_2$, $Pb(NO_3)_2$, $Cd(NO_3)_2$ and carbendazim were the same as that for $CaCl_2$ (5 ml and 250 g l^{-1}). A solute took 5–10 days to leach completely through the soil columns depending on texture of the soils and properties of the solute. After completion of the experiments, three soil samples were collected from the surface of each column by using core samplers (5 cm × 5 cm; Eijkelkamp, The Netherlands). These samples were used to determine the basic physical and hydraulic properties of the soils. Approximately, 200 g additional soil samples were also collected from each column for determining pH, EC and *OC*. Following the whole procedures, data recording on soil-water content and bulk EC, and soil sampling were done for the remaining 10 of the 14 soils in subsequent batches of experiments. In is noted that because of very time-consuming transport experiments there was no replication on each soil.

The analysis of TDR-measured EC was based on the relation between the concentration of a solute in soil water and EC of soil water, which is linearly related to the EC of bulk soil for constant water content (Ward et al., 1994). It is noted that the applied solutes dissociated into ions in solution (e.g., $CaCl_2$ dissociated into Ca^{2+} and Cl^-) and the positive and negative ions, especially for reactive solutes, might have different behaviors and interactions with the soil solid phase and with the existing ions on the exchange complex (Rose et al., 2006). However, for a non-reactive/inert solute like $CaCl_2$, the velocity of a solute is assumed same as the velocity of pore water in most transport experiments; although the velocity of Cl^- may differ from that of the bulk solution due to anion exclusion, the possible small discrepancy was ignored for the relatively non-reactive/inert soils used in this study. The time-series of solute concentration were determined from the TDR-measured EC. Breakthrough curves (BTCs) were drawn by plotting normalized concentrations against time. The mean travel time, τ, mass-dispersion number, N ($=D/ZV$), and retardation factor, *R*, of the solutes were fitted from the BTCs by a transfer-function method (Mojid et al., 2004; their Eqs. 5 and 7) using non-linear least-square fitting technique; the performance of the transfer-function method was reported reliable and described in detail in Mojid et al. (2004). For the physical meaning of the velocity and retardation factor of the reactive solutes ($NaAsO_2$, $Pb(NO_3)_2$, $Cd(NO_3)_2$ and carbendazim) and their differences from those of inert solutes, the readers are referred to Mojid and Vereecken (2005). For $CaCl_2$, *R* was fixed at unity assuming it to be a non-reactive solute. The transport velocity, V (= Z/τ), and dispersion coefficient, D (= $VZN = Z^2N/\tau$), of the solutes were calculated from τ, *N* and the distance, *Z*, between the input and response BTCs.

Soil property measurement

By determining the fractions of sand, silt and clay of the soils by Hydrometer method (Black, 1965) their textural classes were obtained from the Marshall's triangle (Soil Survey Staff, 1975). Soil pH was determined by a glass electrode pH meter following Jackson (1962). Twenty (20) grams of each air-dry soil was mixed with 50 ml distilled water in separate opaque plastic bottles. The suspensions were shaken with a horizontal electric shaker for 20 minutes and kept undisturbed in a control room at 25°C for five hours. The pH of the partly settled soil suspensions was measured by immersing the glass electrode. Soil organic matter, OM, was determined following the method of Walkey-Black (Jackson, 1962). Two grams of each soil were swirled in 10 ml of 1.0N $K_2Cr_2O_7$ solution. Then, 20 ml concentrated H_2SO_4 was added to it and mixed thoroughly. The mixture was kept undisturbed for 30 minutes and diluted to 200 ml with distilled water. It was titrated against $FeSO_4.7\ H_2O$ in presence of 0.5 g NaF and 30 drops of diphenylamine as indicator to dull green endpoint. The OM of the soil was calculated by

$$OM\,(\%) = 10\left(1 - \frac{K}{B}\right) \times 0.335 \tag{1}$$

where K is $FeSO_4.7\ H_2O$ (ml), which is used for titration of the sample and B is $FeSO_4.7\ H_2O$ (ml) for blank. The *OC* of the soil was calculated following Nelson and Sommers (1982) by

$$OC(\%) = \frac{\%OM}{1.724} \tag{2}$$

Gradation tests of the soils were done on the samples by a typical sieve analysis involving a nested column of sieves with wire mesh cloth. For each soil, a 500-g sample was poured into the top sieve, which had the largest screen openings. Each lower sieve in the column had smaller openings than the one above. There was a pan at the base. After shaking, the constituent materials retained on each sieve were weighed. The test was done in accordance with the British Standards, BS 1377 (1990) (Code of Practice). This exercise was repeated for the 14 soils. The median grain diameter (D_{50}) and uniformity coefficient (C_u) of the soils were calculated from the grain size distribution curves. C_u was calculated by

$$C_u = \frac{D_{60}}{D_{10}} \tag{3}$$

For determining bulk densities of the soils in the experimental columns, the soil samples in the core samplers, collected from each column after transport experiment, were dried in oven at 105°C for 24 h. The bulk densities were determined by dividing the oven dry weights of the soils by the inner volume of the core sampler. The average bulk density of each column, calculated from the three samples, was used in developing pedo-transfer functions. The textural class, bulk density, organic carbon content, relative pH (ratio of observed soil pH to the pH of a neutral soil (7) and denoted by pH′), clay content, median grain diameter and coefficient of uniformity of the soils are given in Table 1.

Pedo-transfer function development

Beyond some general conceptual understanding, there are no precise a priori relations that link predictors with the predictands. In addition, most pedo-transfer functions, PTFs, differ with the set of predictors (input variables) and predictands (output variables). PTFs were developed through multiple linear regression analyses to predict solute-transport parameters. SPSS 11.5 statistical program was used to construct multiple linear regression models for the velocity, dispersion coefficient and retardation factor of $CaCl_2$, $NaAsO_2$, $Pb(NO_3)_2$, $Cd(NO_3)_2$ and carbendazim. The input variables of the models were bulk density, organic carbon content, relative pH, clay content, median grain diameter, and uniformity coefficient of the soils (Table 1). The PTFs were developed through data validation, variable selection, and model calibration and verification. Performances of the PTFs were evaluated by sensitivity analysis and performance assessment.

Table 1. Textural class, particle size distribution, bulk density (γ, g cm^{-3}), organic carbon content (*OC*, %), relative pH (pH′), clay content (*C*, fraction), median grain diameter (D_{50}, mm) and uniformity coefficient (C_u) of nine soils used in developing and validating MLR models

Sl. No.	Texture	Particle size distribution (%)			γ (g cm^{-3})	*OC* (%)	pH′	*C* (fraction)	D_{50} (mm)	C_u
		sand	silt	clay						
1	Loamy sand	79.48	14.48	6.04	1.36	0.767	0.136	0.060	0.235	3.27
2	Silt loam	22.70	65.96	11.44	1.33	0.686	0.134	0.114	0.101	3.26
3	Sandy loam	54.48	37.02	8.50	1.33	0.452	0.132	0.085	0.200	3.58
4	Silt loam	37.32	52.00	10.68	1.29	0.753	0.136	0.107	0.112	3.30
5	Silt loam	17.24	70.08	12.68	1.40	0.523	0.143	0.127	0.095	3.86
6	Silt	8.68	77.96	13.36	1.34	0.787	0.141	0.134	0.069	2.98
7	Silt loam	18.48	66.04	15.48	1.37	0.372	0.158	0.155	0.066	2.63
8	Silt loam	4.92	75.96	19.12	1.32	0.840	0.146	0.191	0.062	2.78
9	Silty clay loam	2.68	70.16	27.16	1.41	0.554	1.09	0.272	0.037	2.43
10	Silt loam	16.96	59.84	23.20	1.26	0.987	0.96	0.232	0.045	2.61
11	Sandy loam	61.36	23.00	15.64	1.54	0.288	1.14	0.156	0.073	2.99
12	Sandy loam	74.59	15.91	9.50	1.61	0.245	1.16	0.095	0.134	3.35
13	Loamy sand	84.47	10.60	4.93	1.63	0.134	1.20	0.049	0.299	3.64
14	Silt loam	29.04	53.92	17.04	1.33	0.760	0.97	0.170	0.070	2.85

Data validation and variable selection

Data validation is a corrective measure that is taken at ‘*ab initio*’ in observations. It is crucial since existence of even a single outlier can make the whole data set to a non-linear form (Draper and Smith, 1981). The purpose of selecting the appropriate regression variables is to reduce predictors to some “optimal” subset of the available regressors. This is important since a smaller set of predictors may often provide more accurate predictions of future cases, and/or identifying only the pertinent predictor variables may significantly improve the response.

An important step in data validation process is to scale up or down the observations by focusing on their units of measurements. This is because the mean change of response-dependent variables (*Y*) with measuring units controls statistical information. For a valid data set, *Y* needs to be greater than the mean change of predictors/independent variables (*X*). Such alterations of digital magnitudes in mean due to the change in measuring units do not, however, disrupt the basic theme of analysis. The yardstick of fixing the right units of measurement is called coefficient of centrality (c), which makes the mean of *Y*s (denoted by $\overline{Y}$) compatible to the mean of *X*s (denoted by $\overline{X}$) (Rashid, 1999). A typical multiple linear regression model with the means of variables is expressed by

$$\overline{Y} = b_0\overline{X_0} + b_1\overline{X_1} + b_2\overline{X_2} + \cdots + b_k\overline{X_k} \tag{4}$$

where *b*s are regression coefficients. Dividing Eq. 4 by $\overline{Y}$ results in

$$1 = b_0c_0 + b_1c_1 + b_2c_2 + \cdots + b_kc_k = b_0c_0 + \sum_{1}^{k}(bc)_j \tag{5}$$

The coefficients of centrality, in all respects, are akin to the latent vectors, which are widely used in the latent-root regression analysis. These coefficients are always non-negative statistics and the unique value $c_o \geq 0$ in respect to the dummy variable X_o appears only at $\sum c^2_{1\to k}$ ≤1. This axiom is satisfied only if $0 \leq c_0^2 \leq 1$, and the restriction could be met by expressing the respondents, *Y*s, in higher or the predictors, *X*s, in lower measuring units. The ‘$c^2 = 0$’ implies a homogeneous model. For modeling our multiple linear regressions, the required conditions were satisfied by scaling up velocity of the solutes from ‘cm h^{-1}’ to ‘mm h^{-1}’ and dispersion coefficient of the solutes from ‘cm^2 h^{-1}’ to ‘mm^2 h^{-1}’. Soil pH was scaled down to pH′, the unit value of which implies a neutral soil. All input variables satisfied the required conditions.

The selection of variables, done at the completion of regression analysis, was accomplished with the direction of the regression coefficients (b_j) and correlation coefficients (r_j). The predictors to be consistent, b_j must be unidirectional to r_j; their anti-directional behavior results in negative correlation coefficient (−*r*), which reveals that the variables are irrelevant and must be discarded from the regression models. Based on this criterion, irrelevant variables were discarded during model development. Acceptable probability level for the MLR analysis was fixed at 5%, that is the upper level of significance for acceptance was p > 0.05. Eight soils (# 1 – 8, Table 1) were used to select variables for MLR models.

Model calibration and verification

The MLR models were calibrated with experimental data of five soils (#9 – 13, Table 1); the purpose was to determine appropriate values of the regression coefficients (Eq. 4). Unlike usually employed procedure of calibrating a model by utilizing only one known data set, we utilized five known data sets and determined the overall average regression coefficients for the data sets. So, the calibration was done by comparing the model-estimated solute-transport parameters (*V*, *D* and *R*) with their measured values, while ensuring least errors between the two parameter sets. The obtained models were verified with the measured soil properties in evaluating solute-transport parameters. Data of a silt loam soil (#14, Table 1), not used in variable selection and model calibration, was used to verify accuracy of the models.

Parameter sensitivity analysis

Sensitivity analysis was done to evaluate relative importance of each input variable in the performance of the MLR models. At first, the model was run by using the measured input variables and the observed error was recorded for the solutes. Afterwards, the model was run by changing the input variables by ± 5% and ± 10%, and the error was recorded in each run. The sensitivity of an input variable was estimated by the ratio of error obtained with ± 5% or ± 10% changing of the variable to the error obtained with original (measured) value of the variable. An error ratio of less than unity implies that there is no effect of the input variable in generating output of the model. A larger error ratio, on the other hand, indicates more sensitivity of the input variable on the output of the model. The input variables were ranked in order of their degree of influence on the model output based on the error ratio.

Model performance assessment

Improving the accuracy of pedo-transfer functions, PTFs, requires studying how prediction uncertainty can be apportioned to different sources of uncertainty in inputs. The performance of the MLR models in simulating transport parameters of the five solutes in homogenous soil columns against their measured values was evaluated by using goodness-of-fit parameters following Piegorsch and Bailer (2005), Sarmah et al. (2005) and Phillips (2006). The most common metrics used to evaluate performance of the PTFs are root-mean-square errors (*RMSE*s), mean errors (*ME*s) and coefficient of determination (r^2). The *RMSE* quantifies the root of the average bivariate variance between estimated and measured quantities. It was calculated by

$$RMSE = \left[\sum_{i=1}^{n} (P_i - O_i)^2 / n \right]^{1/2} \tag{6}$$

where P_i is predicted and O_i is measured solute-transport parameters, and *n* is the number of observations. An *RMSE* of zero indicates no difference between the measured and simulated solute-transport parameters; the smaller an *RMSE* the better is the performance of the model. Modeling efficiency (*EF*) is a measure of accuracy of simulation and is an indicator of overall agreement between the measured and predicted results. *EF* of the model was calculated by

$$EF = \frac{\sum_{i=1}^{n} (O_i - O_m)^2 - \sum_{i=1}^{n} (P_i - O_i)^2}{\sum_{i=1}^{n} (O_i - O_m)^2} \tag{7}$$

where O_m is the average of measured values, and O_i and P_i represent the same meaning as in Eq. 6. An *EF* of unity implies a perfect match between the predicted and measured results. A negative *EF* implies that the predicted values are worse than simply using the observed mean as the best estimate of the data.

Databases used in the development of PTFs usually do not reflect the true population of soils in a region, and, as a result, PTFs tend to be biased to the database on which they are calibrated (Schaap and Leij, 1998). Mean error (*ME*) provides the size and sign of such systematic errors or bias of the prediction error. This error is computed by

$$ME = \sum_{i=1}^{n} (O_i - P_i) / n \tag{8}$$

Negative *ME* values indicate an average underestimation of the quantity being evaluated, while its positive values indicate an overestimation of the target variables. For a truly well-performing PTF, both *RMSE* and *ME* should be as low as possible.

Bias is a persistent positive or negative deviation of the measured value from the true value that arises from erroneous assumptions in the learning algorithm. This error, expressed as a percentage of overall error and denoted by *BOE*, is calculated by (Geman et al., 1992)

$$BOE = \frac{ME^2}{MSE} \times 100 \tag{9}$$

A mean square error (*MSE*), which measures the average of the square of error with the error being the amount by which the estimator differs from the quantity to be estimated, is calculated by

$$MSE = (RMSE)^2 \tag{10}$$

Results and Discussion

MLR models

Pedo-transfer functions, in the form of multiple linear regression, MLR, for predicting transport parameters of $CaCl_2$, $NaAsO_2$, $Pb(NO_3)_2$, $Cd(NO_3)_2$ and carbendazim are compared in Table 2 along with the inconsistent input variables, coefficient of determination (r^2) and probability level (p). The MLR models for predicting velocity, *V*, of the solutes were significant over probability level, p = 0.014 to 0.032. The coefficients of determination of the models, $r^2 \geq 0.99$, revealed that over 99% variation in *V* was justified due to the contributions of organic carbon, *OC*; bulk density, γ; clay content, *C*; median grain diameter, D_{50}; and uniformity coefficient, C_u, of the soils. When relative soil reaction, pH′, was included, in addition to these input variables, the models became insignificant in predicting *V*, with probability level exceeding 0.05. It thus revealed that pH′ exerted inconsistent effects on the output of the models. The opposite signs (not shown) obtained between the regression coefficients (*b*s in Eq. 4) and correlation coefficient (*r*) of pH′ also confirmed this inconsistency. So, pH′ was ignored for modeling velocity of the solutes.

Table 2. Pedo-transfer functions of MLR models for $CaCl_2$, NaAsO2, $Pb(NO_3)_2$, $Cd(NO_3)_2$ and carbendazim with the values of the regression coefficients, inconsistent input parameters, coefficients of determination (r^2) and probability of uncertainty of the models

Solutes	MLR models	Inconsistent parameters	r^2	p-value
$CaCl_2$	$V = 11.99 + 0.561(\%OC) - 2.31(\gamma) - 21.80(C) + 4.39(D_{50}) + 0.13(C_u)$	pH′	0.990	0.025
	$D = 11.10 - 1.02(\%OC) + 2.39(\gamma) + 3.60(pH') - 14.94(C) + 33.78(D_{50})$	C_u	0.999	0.001
$NaAsO_2$	$V = 10.38 + 0.513(\%OC) - 1.31(\gamma) - 21.57(C) + 4.68(D_{50}) + 0.19(C_u)$	pH′	0.997	0.016
	$D = 23.35 + 0.31(\gamma) + 0.97(pH') - 30.63(C) + 0.54C_u$	%*OC*, D_{50}	0.927	0.047
	$R = 1.14 + 0.021(\%OC) + 0.173(\gamma) - 0.056(pH') + 0.631(C) - 0.34(D_{50})$	C_u	0.998	0.005
Pb $(NO_3)_2$	$V = 12.29 + 0.489(\%OC) - 2.86(\gamma) - 21.22(C) + 4.39(D_{50}) + 0.254(C_u)$	pH′	0.998	0.032
	$D = 16.16 - 0.025(\%OC) - 19.43(C) + 3.97(D_{50}) + 0.55(C_u)$	γ, pH′	0.963	0.017
	$R = 1.16 + 0.011(\%OC) + 0.092(\gamma) - 0.098(pH') + 0.855(C)$	D_{50}, C_u	0.984	0.005
Cd $(NO_3)_2$	$V = 12.19 + 0.35(\%OC) - 3.18(\gamma) - 19.07(C) + 5.62(D_{50}) + 0.31(C_u)$	pH′	0.990	0.026
	$D = 9.90 + 3.50(\gamma) - 9.92(C) + 4.89(D_{50})$	pH′, %OC, C_u	0.939	0.007
	$R = 1.21 + 0.090(\gamma) + 0.273(C) - 0.694(D_{50})$	pH′, %OC, C_u	0.994	0.001
Carbendazim	$V = 12.82 + 0.433(\%OC) - 3.09(\gamma) - 21.65(C) + 4.05(D_{50}) + 0.221(C_u)$	pH′	0.994	0.014
	$D = 13.28 - 0.178(\%OC) + 0.371(\gamma) - 14.86(C) + 0.046(D_{50})$	pH′, C_u	0.983	0.005
	$R = 1.17 + 0.018(\%OC) + 0.112(\gamma) - 0.053(pH') + 0.566(C) - 0.563(D_{50})$	C_u	0.999	0.003

In modeling dispersion coefficient, *D*, all the five solutes encountered one or more inconsistent input variables. Although the MLR models for *D*, with inclusion of all input variables, were significant at p > 0.05, the uniformity coefficient, C_u, was inconsistent. Elimination of C_u from the models improved significant level of the models from 0.03 to 0.001. Modeling *D* for $NaAsO_2$ was insignificant with probability level exceeding 0.05 when all input variables were considered. Organic carbon, *OC*, and median grain diameter exerted inconsistent effects on model outputs. When these variables were discarded, the significant level of the models improved to 0.047. For $Pb(NO_3)_2$, γ and pH′ exerted inconsistent effects on the model outputs, and their exclusion from the models reduced probability level to 0.017. Organic carbon, pH′ and C_u exhibited inconsistency in modeling *D* for $Cd(NO_3)_2$. When these variables were excluded, the models became significant with probability level improved from 0.188 to 0.007. For carbendazim, when all input variables were considered in the model, γ and pH′ appeared inconsistent, and their elimination significantly improved the probability level although C_u then became inconsistent. Consequently, C_u was also discarded. The accuracy of the models however improved surprisingly when γ was re-introduced; γ became consistent with a model probability of 0.005. The coefficients of determination of the models for the five solutes ranged from 0.927 to 0.999. The uniformity coefficient always put inconsistent influence for modeling retardation factor, *R*, of the solutes. In addition to this, pH′ and *OC* were also inconsistent in case of $Cd(NO_3)_2$, and D_{50} was inconsistent in case of $Pb(NO_3)_2$. All inconsistent input variables were eliminated from the models to obtain improved levels of model probability.

Model performance

The simulated velocity of the solutes agreed well with the measured velocity as illustrated in Figure 1. In predicting solute velocity, *V*, by the MLR models, the coefficients of determination were large (r^2 = 0.955–0.996). The RMSEs were 0.084–0.126 (Table 3). The efficiency, *EF*, of the models was 99% for all the solutes under investigation. The mean errors, *ME*s, of the models were –0.006 to –0.008 for $CaCl_2$, $Cd(NO_3)_2$ and carbendazim; the negative *ME*s imply that the models slightly overestimated *V* during validation. For $NaAsO_2$ and $Pb(NO_3)_2$, *ME*s of the models were 0.0027–0.0028, implying that the models slightly underestimated *V*. The bias components of overall error, *BOE*, were considerably small (0.051–0.99%).

Table 3. Statistical indices for performance assessment of MLR models for $CaCl_2$, $NaAsO_2$, $Pb(NO_3)_2$, $Cd(NO_3)_2$ and carbendazim

Solute solutions	Parameters	RMSE	EF	ME	BOE (%)
$CaCl_2$	*V*	0.110	0.990	–0.006	0.30
	D	0.046	0.999	0.000	0.00
	R	0.019	0.598	–0.003	3.23
$NaAsO_2$	*V*	0.091	0.993	0.003	0.09
	D	0.373	0.927	–0.004	0.01
	R	0.004	0.992	0.004	74.41
$Pb(NO_3)_2$	*V*	0.126	0.987	0.003	0.05
	D	0.213	0.963	0.012	0.34
	R	0.007	0.965	–0.005	54.82
$Cd(NO_3)_2$	*V*	0.114	0.990	–0.007	0.33
	D	0.169	0.939	0.009	0.28
	R	0.006	0.987	0.004	51.40
Carbendazim	*V*	0.084	0.994	–0.008	1.00
	D	0.074	0.983	–0.002	0.08
	R	0.003	0.997	–0.002	59.60

There were good agreements between the measured and simulated dispersion coefficients, *D*, for the solutes (Figure 2) with large coefficients of determination (r^2 = 0.982–0.997). *RMSE* of the models was 0.213 for $Pb(NO_3)_2$ and 0.373 for $NaAsO_2$ (Table 3). Models' efficiency varied from 0.927 to 0.999. *ME*s of the models for $CaCl_2$, $Pb(NO_3)_2$ and $Cd(NO_3)_2$ were positive (0.000–0.0124), but those for the other solutes were negative (–0.0021 to –0.0041). These results revealed that the MLR models for $CaCl_2$, $Pb(NO_3)_2$ and $Cd(NO_3)_2$ slightly underestimated *D*. For the other solutes, the models slightly overestimated *D*. The *BOE*s were perceptibly small (0–0.338%) for all the solutes.

The measured retardation factors, *R*, agreed well with the estimated *R* of all the solutes (Figure 3) with large coefficients of determination (r^2 = 0.971–0.993). Small RMSEs (0.003–0.019, Table 3) revealed good match between the measured and simulated *R*s of the solutes. Modeling efficiencies for *R* in case of $NaAsO_2$, $Pb(NO_3)_2$, $Cd(NO_3)_2$ and carbendazim were 0.992, 0.965, 0.987 and 0.997, respectively. For $CaCl_2$, *EF* was significantly low (0.598). The mean errors of the models, 0.0035 for $NaAsO_2$ and 0.0042 for $Cd(NO_3)_2$, imply slightly underestimation of *R* by the models. For the other solutes, *ME*s were –0.0023 to –0.0051; these small negative *ME*s indicate minimal overestimation of *R*. *BOE*s of the models were considerably large (3.23 to 74.41%), implying that the MLR models for *R* might miss the appropriate relations between the input parameters and target output (*R*).

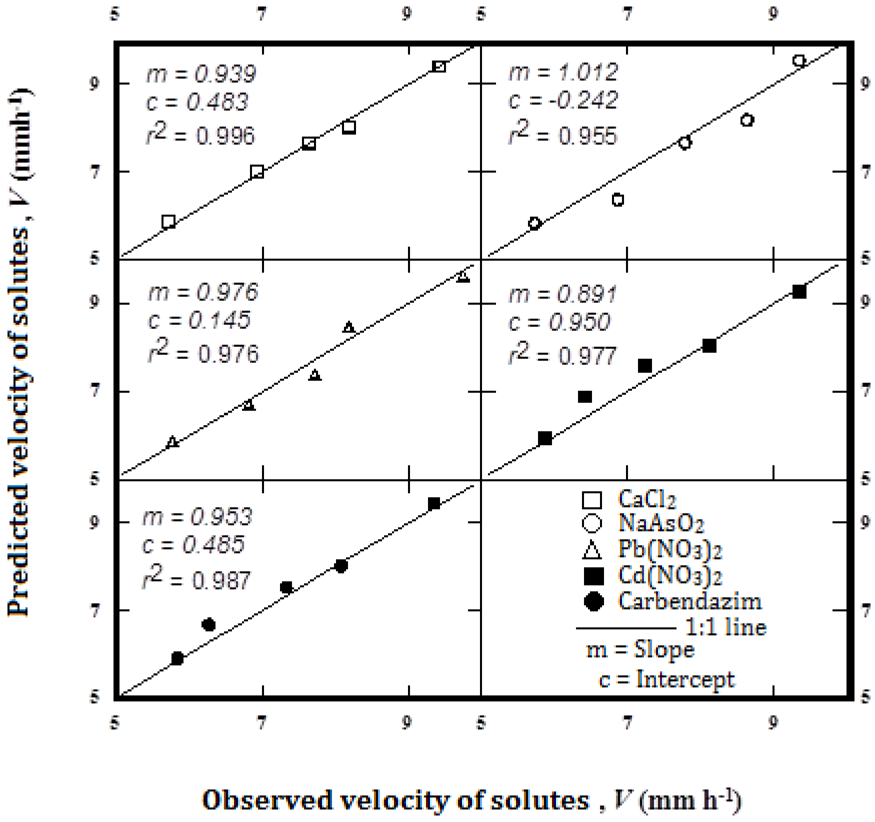


Figure 1. Predicted velocities of $CaCl_2$, $NaAsO_2$, $Pb(NO_3)_2$, $Cd(NO_3)_2$ and carbendazim by MLR models versus their observed velocities.

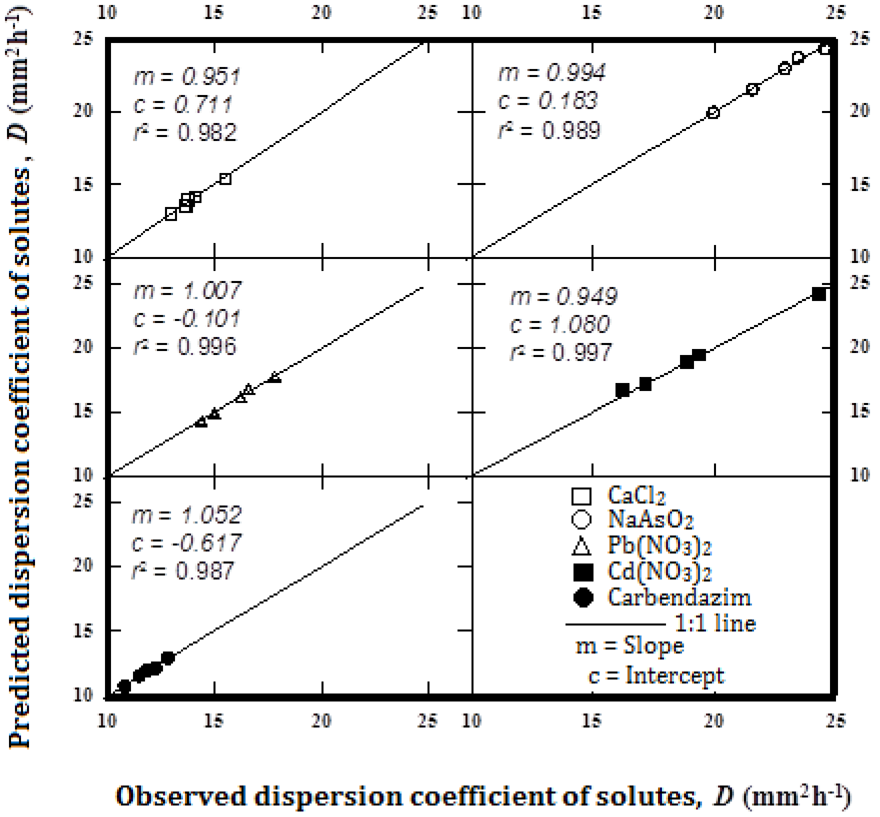


Figure 2. Predicted dispersion coefficients $CaCl_2$, $NaAsO_2$, $Pb(NO_3)_2$, $Cd(NO_3)_2$ and carbendazim by MLR models versus their observed dispersion coefficients.

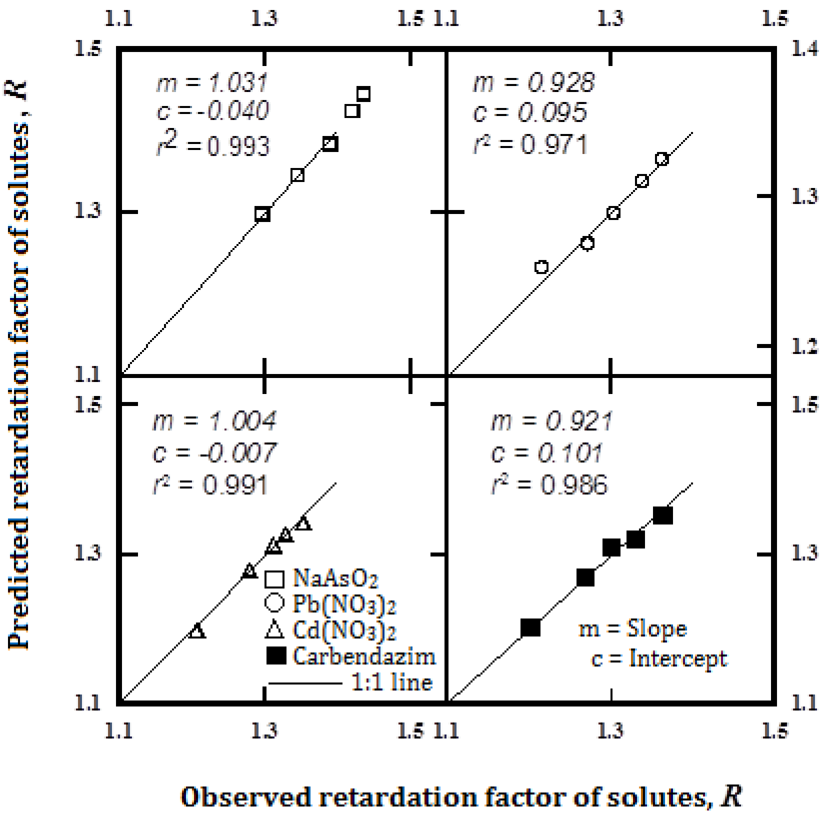


Figure 3. Predicted retardation factors of NaAsO2, $Pb(NO_3)_2$, $Cd(NO_3)_2$ and carbendazim by MLR models versus their retardation factors.

Sensitivity of input parameters

Bulk density, γ, and clay content, *C*, of the soils predominantly governed velocity of the solutes, *V*, in the MLR models, while organic carbon, *OC*, influenced them to a lesser extent. Bulk density ranked as the most impact-generating variable and clay content ranked as the second most influential variable for predicting *V* except for $NaAsO_2$. These two input variables (γ and *C*) were inversely related to *V* as was also observed by Dian-qing et al. (2010). For $NaAsO_2$, γ and *C* followed reversed ranking than for the other solutes. Median grain diameter, D_{50}, ranked as third to influence *V* except for $Cd(NO_3)_2$ for which coefficient of uniformity, C_u, was the third level and D_{50} was the fourth level influential regressors. Organic carbon ranked fifth in controlling the model output except for $CaCl_2$. Uniformity coefficient ranked as the fifth most important variable while *C* ranked fourth in controlling *V* for $CaCl_2$.

Clay ranked as the most prominent input variable in controlling dispersion coefficient, *D*, of $NaAsO_2$ and carbendazim. But, it ranked as the second, third and fourth most influential variable in controlling *D* for $Cd(NO_3)_2$, $Pb(NO_3)_2$ and $CaCl_2$, respectively. Bulk density ranked as the most influencial variable in case of $Pb(NO_3)_2$ and $Cd(NO_3)_2$. It however ranked as the second, third and fourth most influential variable in case of carbendazim, $CaCl_2$ and $NaAsO_2$, respectively. These results were in partial agreement with the findings of Bromly et al. (2007), who, by using step-wise multiple regressions, predicted *D* with pore-water velocity, clay content, silt content and bulk density with an adjusted coefficient of determination of 0.735. Since velocity of the solutes was related to *D*, γ and *C* also influenced *D*. For $CaCl_2$, D_{50} controlled the dispersion coefficient as the most prominent input variable; it ranked third in case of $Cd(NO_3)_2$ and carbendazim, and fourth in case of $Pb(NO_3)_2$. Relative soil pH, exerted second largest impact in predicting *D* for $CaCl_2$ and $Pb(NO_3)_2$, and third largest impact in case of $NaAsO_2$. Uniformity coefficient ranked as the second most dominant variable in predicting *D* for $NaAsO_2$. Organic carbon was less influential since it ranked fourth and fifth in case of carbendazim and $CaCl_2$, respectively.

Bulk density was the most impact-generating variable and exerted positive contribution in predicting retardation factor, *R*, of the reactive solutes under investigation. Clay content was the second most leading variable in controlling *R* of $NaAsO_2$ and $Pb(NO_3)$, and third most important variable in case of $Cd(NO_3)_2$ and carbendazim. For $Cd(NO_3)_2$ and carbendazim, D_{50} put the second largest impact in modeling *R*; it however ranked third and fourth in case of $NaAsO_2$ and $Pb(NO_3)_2$, respectively. Relative pH of the soils ranked third in case of $Pb(NO_3)_2$, and fourth in case of $NaAsO_2$ and carbendazim; it however exerted only minor influence in modeling *R*. Organic carbon, *OC*, ranked as fifth main variable, also exerted less impact on the retardation factor of $NaAsO_2$ and carbendazim. The low rank of *OC* seems unexpected since it was believed to be the second most dominant factor for sorption after soil pH, and most solutes exhibit high affinities for soil organic matter (Springob and Böttcher, 1998). The low ranking of *OC* could be since most of the 13 soils contained relatively low organic carbon (0.134–0.987%, Table 1), and bulk density and clay content of the soils might dominantly controlled sorption of the solutes. Based on the ratio of *RMSE*s, the orders of sensitivity of the MLR model outputs (*V*, *D* and *R*) to the input variables were – (i) $CaCl_2$: $D>V$, (ii) $NaAsO_2$: $R>V>D$, (iii) Pb $(NO_3)_2$: $R>V>D$, (iv) Cd $(NO_3)_2$: $V>R>D$ and (v) carbendazim: $R>V>D$.

Conclusion

Pedo-transfer functions, PTFs, in the form of multiple linear regression, MLR, models were developed for estimating transport velocity, *V*, dispersion coefficient, *D*, and retardation factor, *R*, of $NaAsO_2$, $Pb(NO_3)_2$, $Cd(NO_3)_2$, carbendazim and $CaCl_2$ in 14 Bangladeshi soils. Bulk density and clay content of the soils were the most sensitive/impact-generating input parameters to the MLR models. Based on root-mean-square error, *RMSE*, in estimating the transport parameters, the orders of sensitivity of the model outputs to the input variables for the solutes were – $CaCl_2$: $D>V$, $NaAsO_2$: $R>V>D$, $Pb(NO_3)_2$: $R>V>D$, $Cd(NO_3)_2$: $V>R>D$ and carbendazim: $R>V>D$. The *RMSE*, mean error, *ME*, and bias components of overall error, *BOE*, were appreciably small except for the retardation factor, for which *BOE* was considerably large (3.23–74.41%,) that indicated necessity of further improvement of the model. The model efficiencies were noticeably large (0.93–1.00) for the reactive solutes. Thus, the developed MLR models could fairly predict transport velocity, dispersion coefficient and retardation factor of the reactive solutes under investigation, and hence they can be utilized for practical applications at local scales. The MLR models, however, need to be improved for predicting the retardation factor, possibly, by including additional input variable(s). Also, data of only 14 soils were used in this study and the developed MRL models were verified with the data of only one soil. This is a drawback of our study that needs to be addressed in future studies.

Acknowledgement

This study was done under the 'VLIR–Own Initiatives' program with the Bangladesh Agricultural University at Mymensingh. The authors gratefully acknowledge the funding support of the Belgian Directorate General for Development Cooperation (DGDC) through the 'Vlaamse Interuniversitaire Raad' (VLIR; Flemish Interuniversity Council).

References

Achat, D.L., Pousse, N., Nicolas, M., Brédoire, F., Augusto, L., 2016. Soil properties controlling inorganic phosphorus availability: general results from a national forest network and a global compilation of the literature. *Biogeochemistry* 127(2–3): 255–272.

Alibuyog, N.R., 2007. Development of pedotransfer functions for predicting soil hydraulic properties and solute-transport parameters using artificial neural network analysis. Ph.D. Thesis in Agricultural Engineering, University of the Philippines Los Baños, Philippines.

Black, C.A., 1965. Method of soil analysis. Part-I and II. Agronomy No. 9. American Society of Agronomy, Madison, Wisconsin, USA.

Bouma, J., 1989. Using soil survey data for quantitative land evaluation. In: Advances in Soil Science. Springer, New York, NY. pp. 177–213.

Bromly, M., Hinz, C., Aylmore, L.A.G., 2007. Relation of dispersivity to properties of homogeneous saturated repacked soil columns. *European Journal of Soil Science* 58(1): 293–301.

BS 1377, 1990. Methods of Test for Soils for Civil Engineering Purposes. British Standards Institution, London. 2004.

Dian-qing, L.V., Wang, H., Pan, Y., Wang, L., 2010. Effect of bulk density changes on soil solute transport characteristics. *Journal of Natural Science of Hunan Normal University* 33(1): 75–79.

Draper, N.R., Smith, H., 1981. Applied Regression Analysis. 2nd edn. John Wiley and Sons. New York, USA.

Geman, S., Bienenstock, E., Doursat, R., 1992. Neural networks and the bias/variance dilemma. *Neural Computation*, 4(1): 1–58.

Gonçalves, M.C., Leij, F.J., Schaap, M.G., 2001. Pedotransfer functions for solute transport parameters of Portuguese soils. *European Journal of Soil Science* 52(4): 563–574.

Gonçalves, M.C., Pereira, L.S., Leij, F.J., 1997. Pedo-transfer functions for estimating unsaturated hydraulic properties of Portuguese soils. *European Journal of Soil Science* 48(3): 387-400.

Horn, A.L., Reiher, W., Düring, R.A., Gäth, S., 2006. Efficiency of pedotransfer functions describing cadmium sorption in soils. *Water, Air and Soil Pollution* 170(1–4): 229–247.

Jackson, M.L., 1962. Soil Chemical Analysis. Prentice Hall, Inc. Englewood Chiffs, Ny, USA.

Kodešová, R., Kočárek, M., Kodeš, V., Drábek, O., Kozák, J., Hejtmánková, K., 2011. Pesticide adsorption in relation to soil properties and soil type distribution in regional scale. *Journal of Hazardous Materials* 186(1): 540–550.

Moeys, J., Bergheaud, V., Coquet, Y., 2011. Pedotransfer functions for isoproturon sorption on soils and vadose zone materials. *Pest Management Science* 67(10): 1309–1319.

Mojid, M.A., Hossain, A.B.M.Z., Cappuyns, V., Wyseure, G.C.L., 2016. Transport characteristics of heavy metals, metalloids and pesticides through major agricultural soils of Bangladesh as determined by TDR. *Soil Research* 54(8): 970-984.

Mojid, M.A., Hossain, A.B.M.Z., Wyseure, G.C.L., 2018. Relation of reactive solute-transport parameters to basic soil properties. *Eurasian Journal of Soil Science* 7(4): 326–336.

Mojid, M.A., Rose, D.A., Wyseure, G.C.L., 2004. A transfer-function method for analysing breakthrough data in the time domain of the transport process. *European Journal of Soil Science* 55(4): 699–711.

Mojid, M.A., Vereecken, H., 2005. On the physical meaning of retardation factor and velocity of a nonlinearly sorbing solute. *Journal of Hydrology* 302(1-4): 127–136.

Nelson, D.W., Sommers, L.E., 1982. Total carbon, organic carbon, and organic matter. In: Methods of Soil Analysis, Part 2. Chemical and Microbiological Properties. Page, A.L, Miller, R.H., Keeney, D.R. (Eds.). 2nd Edition. Agronomy Monograph, vol. 9. ASA and SSSA, Madison, WI, USA. pp. 539-579.

Perfect, E., Sukop, M.C., Haszler, G.R., 2002. Prediction of dispersivity for undisturbed soil columns from water retention parameters. *Soil Science Society of America Journal* 66(3): 696–701.

Phillips, I.R., 2006. Modelling water and chemical transport in large undisturbed soil cores using HYDRUS-2D. *Soil Research* 44(1): 27–34.

Piegorsch, W.W., Bailer, A.J., 2005. Quantitative risk assessment with stimulus-response data. In: Analyzing Environmental Data, Piegorsch, W.W., Bailer, A.J. (Eds.). Chichester, West Sussex, UK. pp. 171–214.

Porro, I., Wierenga, P.J., Hills, R.G., 1993. Solute transport through large uniform and layered soil columns. *Water Resources Research* 29(4): 1321–1330.

Rashid, M.A., 1999. On the linearity of multiple regression model. *Bangladesh Journal of Agricultural Engineering* 10 (1–2): 67–76.

Rose, D.A., Abbas, F., Adey, M.A., 2006. Limitations in the use of electrical conductivity to monitor the behaviour of soil solution. *Soil Research* 44(7): 695–700.

Sarmah, A.K., Close, M.E., Pang, L., Lee, R., Green, S.R., 2005. Field study of pesticide leaching in a Himatangi sand (Manawatu) and a Kiripaka bouldery clay loam (Northland). 2. Simulation using LEACHM, HYDRUS-1D, GLEAMS, and SPASMO models. *Australian Journal of Soil Research* 43(4): 471–489.

Schaap, M.G., Leij, F.J., 1998. Database-related accuracy and uncertainty of pedotransfer functions. *Soil Science* 163(10): 765–779.

Soil Survey Staff, 1975. Soil taxonomy. USDA Agriculture Handbook No. 436. Washington, D.C., U.S. Government Printing Office. p. 754.

Springob, G., Böttcher, J., 1998. Parameterization and regionalization of Cd sorption characteristics of sandy soils. I. Freundlich type parameters. *Journal of Plant Nutrition and Soil Science* 161(6): 689–696.

Touil, S., Degre, A., Chabaca, M.N., 2016. Sensitivity analysis of point and parametric pedotransfer functions for estimating water retention of soils in Algeria. *Soil* 2(4): 647–657.

Van Looy, K., Bouma, J., Herbst, M., Koestel, J., Minasny, B., Mishra, U., Montzka, C., Nemes, A., Pachepsky, Y., Padarian, J., Schaap, M.G., Tóth, B., Verhoef, A., Vanderborght, J., van der Ploeg, M.J., Weihermüller, L., Zacharias, S., Zhang, Y., Vereecken, H., 2017. Pedotransfer functions in Earth system science: challenges and perspectives. *Reviews of Geophysics* 55(4): 1199–1256.

Vereecken, H., 1992. Derivation and validation of pedotransfer functions for soil hydraulic properties. In: Indirect methods for estimating the hydraulic properties of unsaturated soils. van Genuchten, M.T., Leij, F.J., Lund, L.J. (Eds.). University of California, Riverside, CA. pp. 473–488.

Ward, A.L., Elrick, D.E., Kachanoski, R.G., 1994. laboratory measurements of solute transport using time-domain reflectometry. *Soil Science Society of America Journal* 58(4): 1031–1039.