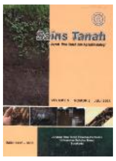
SAINS TANAH – Journal of Soil Science and Agroclimatology

Journal homepage: http://jurnal.uns.ac.id/tanah



Comparative performance of multiple linear regression and artificial neural network models in estimating solute-transport parameters

Mohammad Abdul Mojid[1], **A.B.M. Zahid Hossain** [2]

[1]Bangladesh Agricultural University, Mymensingh 2202, Bangladesh
[2]Bangladesh Rice Research Institute, Gazipur, Bangladesh

ARTICLE INFO

Keywords:
ANN
MLR
Prediction
Soil
Solute

Article history
Submitted: 2021-03-10
Accepted: 2021-05-08
Available online: 2021-06-30
Published regularly: June 2021

***** *Corresponding Author*
Email address:
ma_mojid@bau.edu.bd

ABSTRACT

Indirect estimate of solute-transport parameters through pedo-transfer functions (PTFs) is becoming important due to expensive and time-consuming direct measurement of the parameters for a large number of soils and solutes. This study evaluated the relative performance of PTFs of multiple linear regression (MLR) and Artificial Neural Network (ANN) models in predicting velocity (V), dispersion coefficient (D) and retardation factor (R) of $CaCl_2$, $NaAsO_2$, $Cd(NO_3)_2$, $Pb(NO_3)_2$ and $C_9H_9N_3O_2$ (carbendazim) in five agricultural soils. V, D and R of the solutes were determined in repacked soil columns under steady-state unsaturated water flow conditions. Textural class, particle size distribution, bulk density, organic carbon, relative pH, clay%, grain size, and uniformity coefficient of the soils were determined. MLR and ANN models were calibrated with the measured data of four soils and verified for another soil. Root-Mean Square Error (RMSE) is significantly smaller (0.015) and modelling efficiency (EF) is significantly larger (0.999) for ANN model than those (0.096 and 0.954, respectively) for MLR model. Negative Mean Absolute Error (MAE) (–0.0002) of MLR model indicates overestimation, while positive MAE (0.00003) of ANN model indicates minimal underestimation. The ANN model is less biased than the MLR model during prediction. Thus, the ANN model can significantly enhance pollution transport prediction through soils with good accuracy.

How to Cite: Mojid, M. A., Hossain, A. B. M. Z. (2021). Comparative performance of multiple linear regression and artificial neural network models in estimating solute-transport parameters [Research]. Sains Tanah Journal of Soil Science and Agroclimatology, 18(1): 27-35. https://dx.doi.org/10.20961/stjssa.v18i1.49207

1. Introduction

Continuous application of various agrochemicals and addition of industrial wastes pollute soils with heavy metals and pesticide residues. The polluted soils afterwards contribute to polluting groundwater through leaching. Therefore, characterizing the transport of soluble chemicals through soils is an essential part to assess the pollution of soil and groundwater resources (Amin Al Manmi et al., 2019; Chegenizadeh et al., 2014). Water flow and solute transport through the subsurface are normally simulated by mathematical models (Zhang et al., 2012), which require soil hydraulic parameters and solute-transport parameters as the major input data. The direct method of determining the solute-transport parameters includes measurement of solute breakthrough curves (BTCs) and fitting them to analytical solutions of the classical convection-dispersion equation. This method is time-consuming, laborious, expensive and practically impossible for a wide range of soil types and solutes to sample the temporal and spatial variations. Consequently, indirect approaches like pedo-transfer functions (PTFs) are getting importance as alternative techniques. The PTFs utilize easily measurable basic soil properties in predicting solute-transport parameters and also other soil hydraulic properties (Achat et al., 2016; Van Looy et al., 2017; Xu et al., 2021). Considering multifaceted limitations of direct measurements, fairly correct estimates of the solute-transport parameters can serve well for many practical applications.

The good potential has been reported in predicting flow velocity (V), dispersion coefficient (D) and dispersivity (λ) of solutes using PTFs from multiple linear regressions (MLR) (Alibuyog, 2007; Mojid, Hossain, Wyseure et al., 2019). In predicting V, D and λ using step-wise multiple regressions for a large number of soil textures, Perfect et al. (2002) explained more than 50% of the total variation in dispersivity in terms of parameters of the water retention curve. Artificial neural networks (ANNs), another class of PTFs, are now well-known techniques in many disciplines, such as engineering, medicine, biology, physics, etc. In contrast to MLR, ANNs are non-linear regression techniques and have the ability of mapping between input and output patterns. They have been applied for predicting soil hydraulic properties in many studies (e.g., Minasny et al., 2004; Schaap et al., 1998; Sihag, 2018; Sihag et al., 2019; Williams & Ojuri, 2021). In few occasions, ANN was also applied to predict the transport and distribution of solutes in groundwater (Almasri & Kaluarachchi, 2005; Morshed & Kaluarachchi, 1998). Recently, MLR and ANN models have been applied to predict transport parameters of heavy metal compounds and pesticides in agricultural soils (Mojid, Hossain, & Ashraf, 2019; Mojid, Hossain, Wyseure et al., 2019).

Due to entirely different working principles of MLR and ANN models, the PTFs based on these models also provide different predictions. The accuracy and reliability of the two sets of PTFs have been compared for various predicting purposes. For example, higher accuracy of the ANN model was reported in predicting field capacity and permanent wilting point of soils in terms of coefficient of determination (r^2), Mean Absolute Error (MAE), and Root-Mean Square Error (RMSE) (Taşan & Demir, 2020); monthly maximum rainfall in terms of r^2 (Ilaboya, 2019); water quality parameters in terms of r^2 and RMSE (Zare Abyaneh, 2014); changes in overall quality of cheese during storage in terms of r^2 and RMSE (Stangierski et al., 2019), and bonding strength of wood in terms of r^2 and MAE (Bardak et al., 2016). However, the literature search reveals no comparison of the MLR and ANN models for predicting solute-transport parameters, specifically reactive solutes, through agricultural soils. It is therefore important to identify the relative performance of the two sets of PTFs to choose the better one for practical applications. The objective of this study was therefore to evaluate and compare performances of the MLR and ANN models in predicting solute-transport parameters, specifically for reactive solutes.

2. Material and Methods

This study utilized a part of comprehensive data sets measured by the authors to achieve the goal. A detailed description of the experiments and measurement of data sets are reported in Mojid et al. (2016). They are shortly described in the following sections. However, the readers seeking more details are referred to the cited source.

Table 1. Textural class, bulk density (γ, g cm^{-3}), organic carbon (OC%), relative pH (pH′), clay (fraction), median grain diameter (D_{50}, mm) and uniformity coefficient (C_u) of six soils used in calibrating and verifying MLR and ANN models

Sl. No.	Soil texture	γ (g cm^{-3})	OC %	pH′	Clay (fraction)	D_{50} (mm)	C_u
1	Silty clay loam	1.41	0.554	1.09	0.272	0.037	2.43
2	Silt loam	1.26	0.987	0.96	0.232	0.045	2.61
3	Sandy loam	1.54	0.288	1.14	0.156	0.073	2.99
4	Sandy loam	1.61	0.245	1.16	0.095	0.134	3.35
5	Loamy sand	1.63	0.134	1.20	0.049	0.299	3.64
6	Silt loam	1.33	0.760	0.97	0.170	0.070	2.85

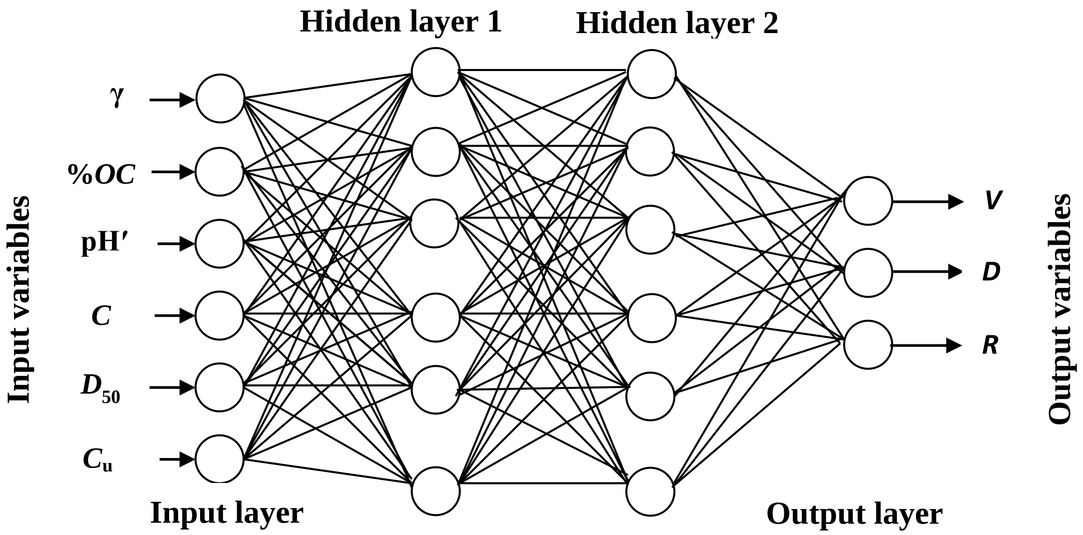


Figure 1. Structure of Artificial Neural Network (ANN) model for predicting solute-transport parameters from soil properties (Source: Mojid, Hossain & Ashraf (2019))

2.1. Measurement of solute transport

Six agricultural soils were sampled from the upper 0–15 cm soil layers from geographically distributed locations of Bangladesh. Particle fractions, grain size distribution, pH and organic carbon (OC) were determined following standard methods from sub-samples of the air-dried and sieved samples. Four soil columns were prepared for solute-transport experiment in 34-cm PVC columns that were sited separately on 1.2-m high soil columns. Two TDR sensors: one at 8 cm and the other at 28 cm below soil surface were inserted horizontally in each upper soil column (experimental soil column). This set-up was conditioned by leaching tap water following wetting-drying cycles during several months. After conditioning, steady-state water flow (0.32 ± 0.02 cm h^{-1}) condition through the soil columns was achieved with a cartridge pump and a 5-ml solution of $CaCl_2$, $NaAsO_2$, $Pb(NO_3)_2$, $Cd(NO_3)_2$ and $C_9H_9N_3O_2$ (carbendazim) was spread separately on each upper soil column evenly. Soil-water content and bulk electrical conductivity (EC) of the soils were recorded with a TDR100 and CR10X datalogger at suitable interval for each soil. Data recording continued until the applied $CaCl_2$ washed out of the upper soil columns completely with leaching water flux. Similar measurements were done consecutively for the six soils. After measurements of solute breakthrough data, soil samples were collected from each upper column and their physical and hydraulic properties were determined following standard methods. Soil pH, EC and OC were determined from another sample collected from each upper soil column.

Breakthrough curves, BTCs (normalized solute concentration versus solute breakthrough time), of the solutes were calculated from the measured time series of TDR-measured EC. The mean travel time (τ), mass-dispersion number (*N′* = *D*/*ZV*) and retardation factor, *R*, of the solutes were determined by analyzing the BTCs by a transfer-function method (Mojid et al., 2004). By using τ, *N′* and *Z* (distance between the input and response BTCs) the transport velocity, *V* (= *Z*/τ), and dispersion coefficient, *D* (= $VZN' = Z^2N/\tau$), of the solutes were calculated. *V* and *R* for reactive solutes being concentration-dependent are time-dependent. Note that $CaCl_2$ is a non-reactive solute and hence its retardation factor is typically assumed as unity.

2.2. Measurement of soil properties

Soil textural classes, pH and OC were determined by Hydrometer method (Black, 1965), glass electrode pH meter (Jackson, 1962) and Walkley-Black method (Jackson, 1962), respectively following standard protocols. Grain size distribution was determined with sieve analysis following British Standards, BS 1377 (1990), and median grain diameter (D_{50}) and uniformity coefficient (C_u) of the soils were calculated there from. The bulk density of the soils was measured by drying soil samples in core samplers in an oven at 105°C for 24 h. The textural class, bulk density, organic carbon, 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 listed in Table 1.

2.3. MLR and ANN models

The general form of multiple linear regression, MLR, models for predicting solute-transport parameters can be expressed by Equation 1.

$$\overline{Y} = b_0\overline{X_0} + b_1\overline{X_1} + b_2\overline{X_2} + \cdots + b_k\overline{X_k} \quad \text{.................[1]}$$

with $\overline{Y}$ being mean of the response/dependent variables (Y), $\overline{X}$ mean of predictors (independent variables) (X) and b_s the regression coefficients.

A four-layer ANN model for the solutes under investigation (Figure 1) was developed by Mojid, Hossain & Ashraf (2019). With input variables I_1, …, I_n (i = 1, …, n) and hidden units H_j, …, H_k (j = 1,…, k), the total weighted input (X_j) is expressed by the sum of products of inputs and weights (w_{ij}) of the connections between the input and hidden units as Equation 2.

$$X_j = \sum_{i=1}^{n} I_i w_{ij} \quad \text{.....[2]}$$

By using X_j a transfer function gives activity level of the hidden unit that is given by Equation 3.

$$H_j = \frac{1}{1+e^{-X_j}} \quad \text{.....[3]}$$

The hidden layer's activity and weight are multiplied together to generate output (prediction) of the neural network, P_v, as Equation 4.

$$P_v = \sum_{j=1}^{k} H_j w_{jv} \quad \text{.....[4]}$$

The observed output (O_v) and predicted output, P_v, are compared using root-mean-square error (RMSE) as Equation 5.

$$RMSE = \left[\frac{1}{n}\sum (O_v - P_v)\right]^{1/2} \quad \text{.....[5]}$$

To minimize deviation between the observed and predicted outputs, the weights, w_{ij}, are adjusted in the calculation system by using a learning function given by Equation 6.

$$\Delta w_{ij}(t+1) = \varepsilon\delta_j o_i + \alpha\Delta w_{ij}(t) \quad \text{.....[6]}$$

with Δw_{ij} (*t+1*) being magnitude of weight-change, ε learning rate, δ_j a local error gradient, α a momentum coefficient, o_i output of the previous unit, and $w_{ij}(t)$ current weight. Following Haykin (1994), Mojid, Hossain & Ashraf (2019) 0.1 was adopted for learning rate and 0.3 for momentum coefficient. The error gradient for output units is expressed by Equation 7.

$$\delta_j = \sigma^1(H_j)(O_v - P_v) \quad \text{.....[7]}$$

and for hidden units, it is expressed by Equation 8.

$$\delta_j = \sigma^1(H_j)\sum(O_v - P_v)w_{jv} \quad \text{.....[8]}$$

with $\sigma^1(H_j)$ being the derivative of the network's hidden layer activity. ANN model becomes appropriate when RMSE reaches its least value that is obtained by continuously re-calculating RMSE after each adjustment of the weight.

2.4. Model calibration and verification

The MLR models (Eq.1) were calibrated by fitting them to the measured solute-transport parameters of five soils (#1 to 5, Table 1). Attainment of least errors between the measured and estimated solute-transport parameters provided desired values of the regression coefficients, b_s (Eq.1). Similarly, the ANN model was calibrated with the same set of observed data of the five soils and the weights (w_{ij}, Eq.2) were determined. Both models were verified by using the data set of soil #6 (Table 1) to evaluate the accuracy levels of their predictive capability.

2.5. Model performance indices

Any model encounters a number of errors while making predictions, the most important of which are expressed in terms of RMSE, modelling efficiency (EF), Mean Absolute Error (MAE), Bias Error (BE) and Mean Square Error (MSE). Consequently, the performances of our MLR and ANN models were assessed in terms of these performance indices following Piegorsch & Bailer (2005), Sarmah et al. (2005), and Phillips (2006). RMSE is expressed by Equation 9.

$$RMSE = \left[\sum_{v=1}^{n} (P_v - O_v)^2 / n \right]^{1/2} \quad \text{.................................} [9]$$

where P_v is predicted and O_v is measured/observed solute-transport parameters, and n is observation number. The lesser an RMSE the superior is the performance of the model; for perfect matching of the measured and predicted values, RMSE becomes 0 (zero). EF indicates overall agreement between the observed and predicted outputs and calculated by Equation 10.

$$EF = \frac{\sum_{v=1}^{n} (O_v - O_m)^2 - \sum_{i=1}^{n} (P_v - O_v)^2}{\sum_{v=1}^{n} (O_v - O_m)^2} \quad \text{................} [10]$$

where O_m is the average of observed outputs, and O_v and P_v are observed output and predicted output, respectively (Eq.5). EF must be positive for a good model; for identical values of the observed and predicted outputs, EF becomes unity. Negative EF infers a prediction level that is worse than simply adopting the observed mean as the best estimate of output. MAE measures the size and identifies sign of bias error in prediction, thus quantifying the magnitude of over-or under-estimation of measurements. MAE is expressed by Equation 11.

$$MAE = \sum_{v=1}^{n} (O_v - P_v) / n \quad \text{...} [11]$$

Bias is a persistent positive or negative deviation of predicted value from actual value. BE is usually calculated as a percentage of overall error and expressed by Equation 12 (Geman et al., 1992).

$$BE = \frac{ME^2}{MSE} \times 100 \quad \text{..} [12]$$

with MSE being the average squared difference between the predicted and actual outputs and calculated by Equation 13.

$$MSE = (RMSE)^2 \quad \text{...} [13]$$

3. Results

3.1. Model performance in predicting solute-transport velocity

Predicted velocity, *V*, of the five solutes under investigation ($CaCl_2$, $NaAsO_2$, $Cd(NO_3)_2$, $Pb(NO_3)_2$ and $C_9H_9N_3O_2$) agrees fairly well with the measured velocity both for MLR and ANN models (Figure 2).

Table 2. Comparison of root–mean square error (RMSE), modelling efficiency (EF), mean absolute error (MAE) and bias components of error (BE) of the MLR and ANN models

Solutes	Variables	MLR Model				ANN Model			
		RMSE	EF	MAE	BOE (%)	RMSE	EF	MAE	BE (%)
$CaCl_2$	*V*	0.110	0.990	-0.0060	0.300	0.013	0.999	-0.00002	0.0001
	D	0.046	0.999	0.0000	0.000	0.058	0.999	0.00010	0.0003
$NaAsO_2$	*V*	0.091	0.993	0.0027	0.087	0.018	0.999	-0.00003	0.0003
	D	0.373	0.927	-0.0041	0.012	0.030	0.999	0.00050	0.0281
	R	0.004	0.992	0.0035	74.409	0.000	0.999	0.00000	0.0001
$Pb(NO_3)_2$	*V*	0.126	0.987	0.0029	0.051	0.015	0.999	0.00001	0.0000
	D	0.213	0.963	0.0124	0.338	0.030	0.999	-0.00002	0.0000
	R	0.007	0.965	-0.0051	54.816	0.003	0.995	0.00000	0.0000
$Cd(NO_3)_2$	*V*	0.114	0.990	-0.0065	0.326	0.007	0.999	-0.00004	0.0028
	D	0.169	0.939	0.0090	0.283	0.011	0.999	0.00001	0.0000
	R	0.006	0.987	0.0042	51.402	0.003	0.998	-0.00001	0.0030
Carben- dazim	*V*	0.084	0.994	-0.0084	0.997	0.010	0.999	-0.00001	0.0001
	D	0.074	0.983	-0.0021	0.082	0.021	0.999	0.00000	0.0000
	R	0.003	0.997	-0.0023	59.597	0.002	0.998	0.00000	0.0000
Average		0.096	0.954	-0.0002	16.395	0.015	0.999	0.00003	0.0024

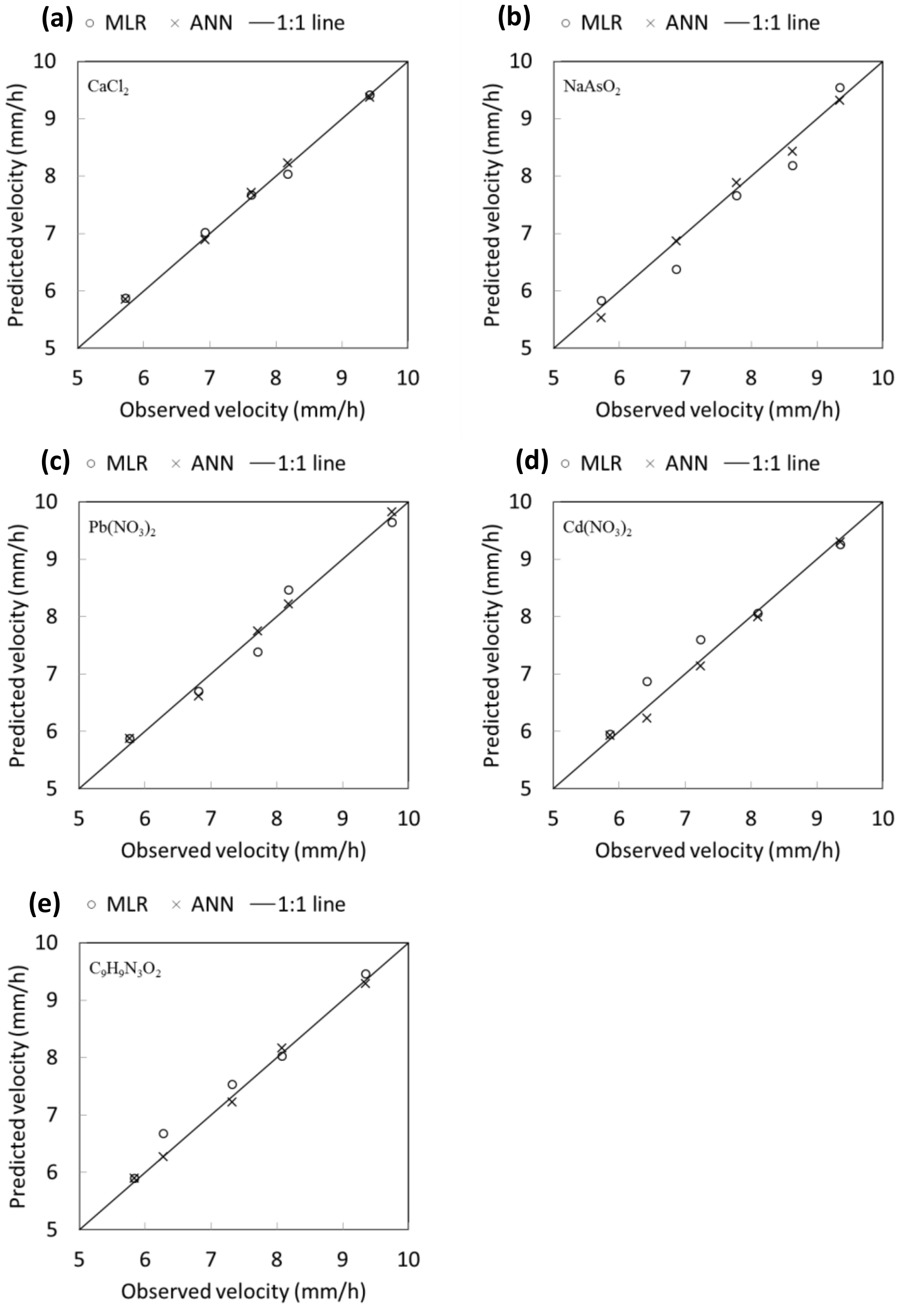


Figure 2. Predicted velocities of $CaCl_2$ **(a)**, $NaAsO_2$ **(b)**, $Pb(NO_3)_2$ **(c)**, $Cd(NO_3)_2$ **(d)** and $C_9H_9N_3O_2$ (carbendazim) **(e)** by MLR and ANN models versus their measured velocities

The coefficient of determination (r^2) between the measured and predicted velocities of 0.990 to 0.998 for MLR model and ≥0.99 for ANN model for the solutes demonstrates almost 1:1 relation of the solute velocities for both models. Root-mean-square error, RMSE (Eq.9), in the prediction ranges between 0.084 and 0.126 for MLR model and between 0.007 and 0.018 for ANN model (Table 2). Although RMSEs of both models are small in terms of accuracy of prediction, ANN model provides significantly smaller RMSE than MLR model. Both models predict velocity of the solutes with high efficiency, EF, ranging from 98.7% to 99.4% for MLR model and 99.9% for ANN model. So, overall agreement between the measured and predicted velocities is almost perfect, with only minimal deviation for MLR model. The mean absolute error, MAE, of estimate by MLR model varies from –0.0084 to 0.0029, with negative error for $CaCl_2$, $Cd(NO_3)_2$ and $C_9H_9N_3O_2$ and positive error for $NaAsO_2$ and $Pb(NO_3)_2$. Negative MAE indicates that the model overestimates solute-transport velocity during verification, while positive MAE reveals underestimation of velocity as visualized in Figure 2. The observed errors in both directions (over-or under-estimations) are however very small to affect prediction by the models. The bias component of overall error, BE, is also small, which ranges between 0.051% and 0.997% for MLR model. The mean absolute error of ANN model is negative (–0.00004 to –0.00001) for $CaCl_2$, $NaAsO_2$, $Cd(NO_3)_2$ and $C_9H_9N_3O_2$ (Table 2), implying that the model slightly over-predicted transport velocity of these solutes. The positive MAE (0.00001) for $Pb(NO_3)_2$ indicates a tendency for marginal underestimation. The bias error of prediction of solute velocity by ANN model ranges between 0 and 0.0028% for the five solutes.

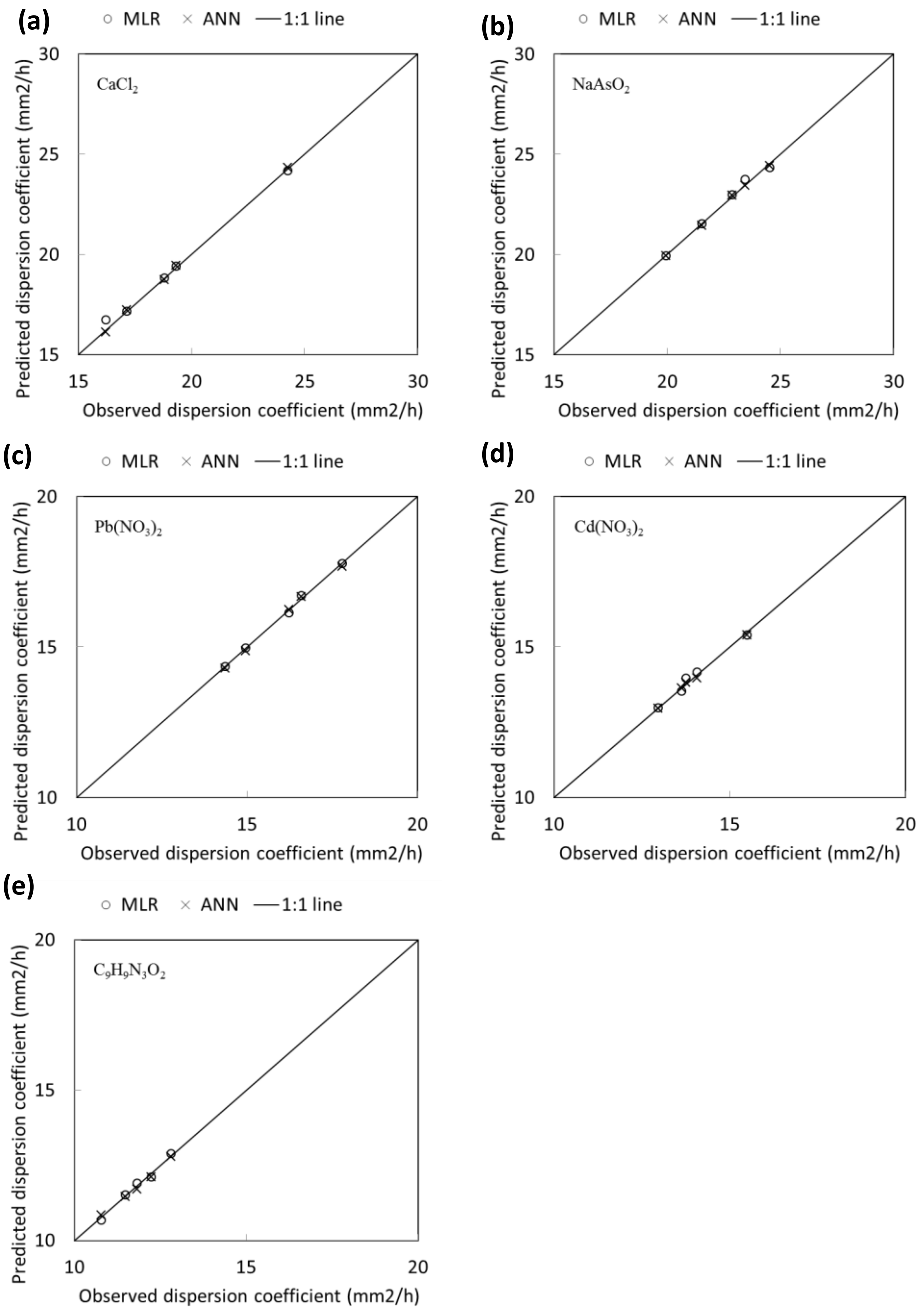


Figure 3. Predicted dispersion coefficients of $CaCl_2$ **(a)**, $NaAsO_2$ **(b)**, $Pb(NO_3)_2$ **(c)**, $Cd(NO_3)_2$ **(d)** and $C_9H_9N_3O_2$ **(**carbendazim), and **(e)** by MLR and ANN models versus their observed dispersion coefficients

3.2. Model performance in predicting solute dispersion coefficient

Figure 3 demonstrates the level of agreement between the measured and predicted dispersion coefficients, *D*, for the five solutes. The coefficient of determination is between 0.93 and 0.98 for $NaAsO_2$, $Cd(NO_3)_2$, $Pb(NO_3)_2$ and $C_9H_9N_3O_2$, and 0.99 for $CaCl_2$ for MLR model and ≥0.99 for ANN model. These results imply that although MLR model can capture the variation in dispersion coefficient of non-reactive solute, it cannot adequately capture the variation for reactive solutes. ANN model perfectly captures the variation. RMSE varies between 0.213 and 0.373 for MLR model and between 0.011 and 0.058 for ANN model, with significantly larger error for MLR model. The efficiency of the modelling dispersion coefficient is 0.927 to 0.983 for the reactive solutes and 0.999 for non-reactive $CaCl_2$ with MLR model. EF of ANN model is 0.99. MLR model results in positive mean absolute error, MAE (0.000 to 0.0124), for $CaCl_2$, $Pb(NO_3)_2$, $Cd(NO_3)_2$ and $Cd(NO_3)_2$ but negative MAE (–0.0041 to –0.0021) for the other solutes. These MAEs reveal that MLR models underestimate dispersion coefficient for $CaCl_2$, $Pb(NO_3)_2$ and $Cd(NO_3)_2$ but overestimate it for the other solutes. ANN model results in small (0 to 0.0005) negative mean absolute errors except for $Pb(NO_3)_2$, thus revealing an affinity of the model to underestimate dispersion coefficient for $Pb(NO_3)_2$ and overestimate it for the other solutes. The bias error, BE, varies from 0 to 0.338% for MLR model and 0 to 0.0281 for ANN model among the five solutes. Small bias errors imply that both models are free from bias or only minimally biased in predicting solute-dispersion coefficient, with ANN model being almost no bias in the prediction.

3.3. Model performance in predicting solute retardation factor

Coefficient of determination between the measured and predicted retardation factors, *R*, of the reactive solutes varies from 0.984 to 0.999 for MLR model and 0.975 to 0.987 for ANN model. The level of agreement between retardation factors predicted by the two models is illustrated in Figure 4. The input parameters exert a less consistent impact on ANN model in determining the retardation factor compared to the other solute-transport parameters. RMSEs of 0.003 to 0.019 for MLR model and 0.0004 to 0.0029 for ANN model, both ranges being small, reveal good matching between the measured and predicted retardation factors of the solutes. Modeling efficiency, EF, of MLR model in predicting retardation factor of $NaAsO_2$, $Pb(NO_3)_2$, $Cd(NO_3)_2$ and $C_9H_9N_3O_2$ is 0.992, 0.965, 0.987 and 0.997, respectively. The modelling efficiency of ANN model in predicting retardation factor of these solutes is ≥0.98. In predicting the retardation factor with MLR model, the mean absolute error, MAE, of 0.0035 for $NaAsO_2$ and 0.0042 for $Cd(NO_3)_2$ imply a small degree of underestimation. For the other solutes, the mean absolute errors varying from –0.0023 to –0.0051 indicate a small degree of overestimation in the prediction. For ANN model, the mean absolute error of the solutes is 0 (zero) except for $Cd(NO_3)_2$ for which the mean absolute error is –0.00001. Bias component of error, BE, is large, ranging from 3.23% to 74.41%, for MLR. BE with ANN model being small, 0 to 0.003%, reveals no or negligible bias in the prediction of retardation factor by this model.

4. Discussion

During building of the MLR model for different solutes ($CaCl_2$, $NaAsO_2$, $Cd(NO_3)_2$, $Pb(NO_3)_2$ and $C_9H_9N_3O_2$) the input variables that exerted an insignificant impact on outputs were discarded to improve the efficiency of the model. This selection criteria of input variables avoided major over-parameterization of the model during calibration. In contrast, ANN model utilized all input variables for its construction and consequently permitted a certain degree of flexibility in terms of input variables. For constructing this model, functional relationships between input and output need to be known a priori from enough training examples.

In vast majority of cases, ANN model provides larger values of coefficient of determination compared to MLR model in the prediction of solute-transport parameters. These results are similar to that of Bardak et al. (2016), Stangierski et al. (2019), and Taşan & Demir (2020) who also obtained a larger coefficient of determination with ANN model than with MLR model. So, ANN model is more accurate in capturing variation in solute-transport parameters than MLR model. Various performance indices of the MLR and ANN models in predicting velocity, dispersion coefficient and retardation factor of $CaCl_2$, $NaAsO_2$, $Cd(NO_3)_2$, $Pb(NO_3)_2$ and $C_9H_9N_3O_2$ are compared in Table 2. The average RMSE in predicting velocity, dispersion coefficient and retardation factor of the solutes is 0.105, 0.175 and 0.008, respectively, with MLR model and 0.012, 0.030 and 0.002, respectively with ANN model.

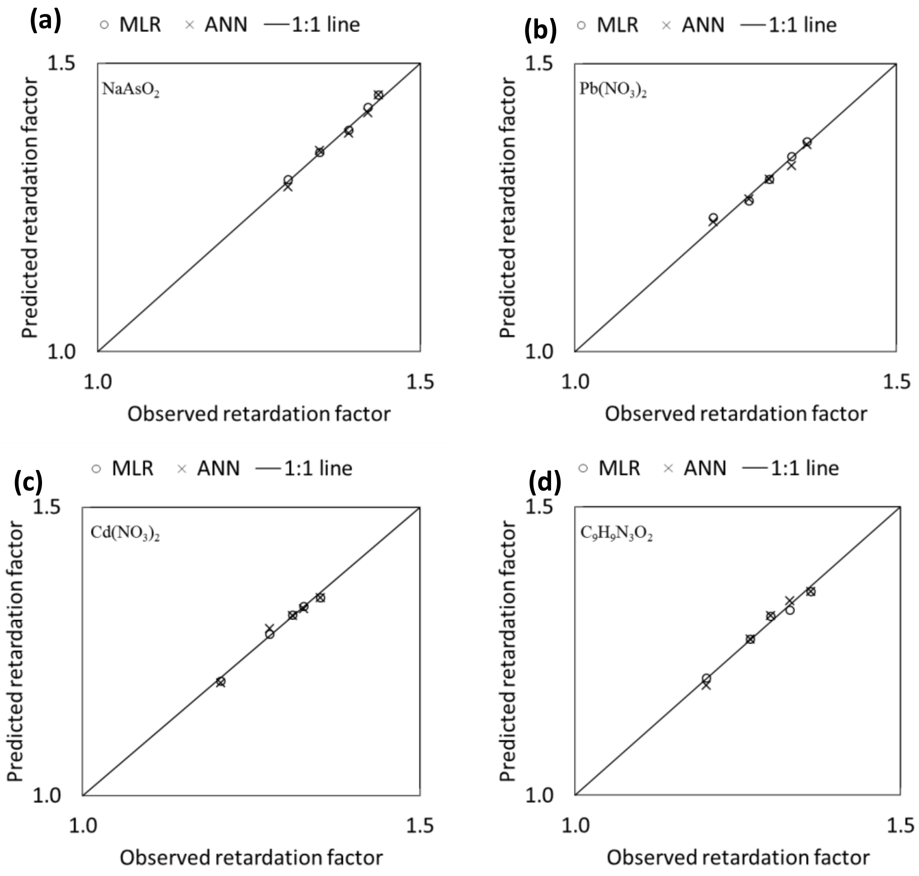


Figure 4. Predicted retardation factors of $NaAsO_2$ **(a)**, $Pb(NO_3)_2$ **(b)**, $Cd(NO_3)_2$ **(c)** and $C_9H_9N_3O_2$ **(**carbendazim, **d)** by MLR and ANN models versus their retardation factors.

When compared, the overall average RMSE is much larger (0.096) for MLR model compared to ANN model (0.015). Higher prediction accuracy of ANN model than MLR model in terms of RMSE has been also reported by Stangierski et al. (2019) and Taşan & Demir (2020) for application of these models in product (cheese) quality and soil hydraulic property prediction, respectively. Thus, our results reveal that ANN technique produces more accurate model than MLR method for predicting solute-transport parameters. Overall efficiency of modelling for MLR model (0.954) is much lower than that for ANN model (0.999). So, in terms of modelling efficiency, ANN technique also provides more accurate models than MLR method. Based on the mean absolute error, MAE, MLR model with average MAE of –0.0002 overestimates solute-transport parameters, while ANN model with an average MAE of 0.00003 underestimates them slightly. When MAE expressed as a percentage of corresponding RMSE, the average bias component of error accounts for 16.40% of overall error in case of MLR model, whilst this error accounts for 0.0024% only in case of ANN model. Smaller MAE by ANN model than MLR model was also obtained by Bardak et al. (2016) and Taşan & Demir (2020). So, ANN model is not bias in predicting solute-transport parameters but MLR model is significantly bias for this prediction. MLR method however retains the advantage of the physical interpretation of the solute-transport parameters, while ANN technique does not retain this property. ANN model is capable of capturing non-linearity in data. Hence, improved prediction of solute-transport parameters by ANN model is more likely compared to MLR model. This inference is fully in agreement with Zare Abyaneh (2014), who obtained better performance of ANN model compared to MLR model while predicting water quality parameters by both models.

5. Conclusion

Multiple linear regression, MLR, models, utilizing selective inputs, although avoid over-parameterization they considerably overestimates solute-transport parameters. ANN model permits flexibility in selecting input variables but it slightly underestimates the parameters. ANN technique provides more accurate models with improved root-mean square error, RMSE, and modelling efficiency, EF, compared to MLR models. ANN model is less biased than MLR model in predicting solute-transport parameters. Thus, ANN model can significantly enhance the prediction of pollution transport through soils by providing essential input parameters.

Declaration of Competing Interest

The authors declare no competing financial or personal interests that may appear and influence the work reported in this paper.

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), 255-272. https://doi.org/10.1007/s10533-015-0178-0

Alibuyog, N. R. (2007). *Development of pedo-transfer functions for predicting soil hydraulic properties and solute-transport parameters using artificial neural network analysis* [PhD Thesis, Agricultural Engineering, University of the Philippines Los Baños].

Almasri, M. N., & Kaluarachchi, J. J. (2005). Modular neural networks to predict the nitrate distribution in ground water using the on-ground nitrogen loading and recharge data. *Environmental Modelling & Software*, *20*(7), 851-871. https://doi.org/10.1016/j.envsoft.2004.05.001

Amin Al Manmi, D. A. M., Abdullah, T. O., Al-Jaf, P. M., & Al-Ansari, N. (2019). Soil and groundwater pollution assessment and delineation of intensity risk map in Sulaymaniyah City, NE of Iraq. *Water*, *11*(10), 2158. https://www.mdpi.com/2073-4441/11/10/2158

Bardak, S., Tiryaki, S., Bardak, T., & Aydin, A. (2016). Predictive Performance of Artificial neural network and multiple linear regression models in predicting adhesive bonding strength of wood. *Strength of Materials*, *48*(6), 811-824. https://doi.org/10.1007/s11223-017-9828-x

Black, C. A. (1965). *Methods of Soil Analysis. Part-I and II.* . American Society of Agronomy, Inc, Publisher, Madison, Wisconsin USA.

BS 1377. (1990). *Methods of Test for Soils for Civil Engineering Purposes. Classification Tests. Parts 2 and 5*.

Chegenizadeh, A., Ghadimi, B., & Nikraz, H. (2014). The prediction of contaminant transport through soil: a novel two-dimensional model approach. *Journal of Civil & Environmental Engineering*, *4*, 1-6. https://doi.org/10.4172/2165-784X.1000138

Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural Networks and the Bias/Variance Dilemma. *Neural Computation*, *4*(1), 1-58. https://doi.org/10.1162/neco.1992.4.1.1

Haykin, S. (1994). *Neural Networks: A Comprehensive Foundation*. Macmillan, New York: Macmillan College Publishing.

Ilaboya, I. (2019). Performance of multiple linear regression (MLR) and artificial neural network (ANN) for the prediction of monthly maximum rainfall in Benin City, Nigeria. *International Journal of Engineering Science and Application*, *3*(1), 21-37.

Jackson, M. L. (1962). *Soil chemical analysis: Advanced course*. Inc. Englewood Chiffs, Ny, USA.

Minasny, B., Hopmans, J. W., Harter, T., Eching, S. O., Tuli, A., & Denton, M. A. (2004). Neural networks prediction of soil hydraulic functions for alluvial soils using multistep outflow data. *Soil Science Society of America Journal*, *68*(2), 417-429. https://doi.org/10.2136/sssaj2004.4170

Mojid, M. A., Hossain, A. B. M. Z., & Ashraf, M. A. (2019). Artificial neural network model to predict transport parameters of reactive solutes from basic soil properties. *Environmental Pollution*, *255*, 113355. https://doi.org/10.1016/j.envpol.2019.113355

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. https://doi.org/10.1071/SR15367

Mojid, M. A., Hossain, A. Z., Wyseure, G. C., & Ashraf, M. A. (2019). Pedo-transfer functions with multiple linear regressions to predict solute-transport parameters. *Eurasian Journal of Soil Science*, *8*(3), 196-207.

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. https://doi.org/10.1111/j.1365-2389.2004.00636.x

Morshed, J., & Kaluarachchi, J. J. (1998). Application of artificial neural network and genetic algorithm in flow and transport simulations. *Advances in Water Resources*, *22*(2), 145-158. https://doi.org/10.1016/S0309-1708(98)00002-5

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. https://doi.org/10.2136/sssaj2002.6960

Phillips, I. R. (2006). Modelling water and chemical transport in large undisturbed soil cores using HYDRUS-2D. *Soil Research*, *44*(1), 27-34. https://doi.org/10.1071/SR05109

Piegorsch, W. W., & Bailer, A. J. (2005). Quantitative risk assessment with stimulus-response data. In *Analyzing Environmental Data* (pp. 171-214). https://doi.org/10.1002/0470012234.ch4

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. *Soil Research*, *43*(4), 471-489. https://doi.org/10.1071/SR04040

Schaap, M. G., Leij, F. J., & van Genuchten, M. T. (1998). Neural network analysis for hierarchical prediction of soil hydraulic properties. *Soil Science Society of America Journal*, *62*(4), 847-855. https://doi.org/10.2136/sssaj1998.03615995006200040001x

Sihag, P. (2018). Prediction of unsaturated hydraulic conductivity using fuzzy logic and artificial neural network. *Modeling Earth Systems and Environment*, *4*(1), 189-198. https://doi.org/10.1007/s40808-018-0434-0

Sihag, P., Tiwari, N. K., & Ranjan, S. (2019). Prediction of unsaturated hydraulic conductivity using adaptive neuro- fuzzy inference system (ANFIS). *ISH Journal of Hydraulic Engineering*, *25*(2), 132-142. https://doi.org/10.1080/09715010.2017.1381861

Stangierski, J., Weiss, D., & Kaczmarek, A. (2019). Multiple regression models and Artificial Neural Network (ANN) as prediction tools of changes in overall quality during the storage of spreadable processed Gouda cheese. *European Food Research and Technology*, *245*(11), 2539-2547. https://doi.org/10.1007/s00217-019-03369-y

Taşan, S., & Demir, Y. (2020). Comparative analysis of MLR, ANN, and ANFIS models for prediction of field capacity and permanent wilting point for Bafra Plain Soils. *Communications in Soil Science and Plant Analysis*, *51*(5), 604-621. https://doi.org/10.1080/00103624.2020.1729374

Van Looy, K., Bouma, J., Herbst, M., Koestel, J., Minasny, B., Mishra, U., Montzka, C., Nemes, A., Pachepsky, Y. A., 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. https://doi.org/10.1002/2017RG000581

Williams, C. G., & Ojuri, O. O. (2021). Predictive modelling of soils' hydraulic conductivity using artificial neural network and multiple linear regression. *SN Applied Sciences*, *3*(2), 152. https://doi.org/10.1007/s42452-020-03974-7

Xu, X., Li, H., Sun, C., Ramos, T. B., Darouich, H., Xiong, Y., Qu, Z., & Huang, G. (2021). Pedotransfer functions for estimating soil water retention properties of northern China agricultural soils: Development and needs*. *Irrigation and Drainage*, *n/a*(n/a). https://doi.org/10.1002/ird.2584

Zare Abyaneh, H. (2014). Evaluation of multivariate linear regression and artificial neural networks in prediction of water quality parameters. *Journal of Environmental Health Science and Engineering*, *12*(1), 40. https://doi.org/10.1186/2052-336X-12-40

Zhang, R., Qian, X., Yuan, X., Ye, R., Xia, B., & Wang, Y. (2012). Simulation of water environmental capacity and pollution load reduction using QUAL2K for water environmental management. *International Journal of Environmental Research and Public Health*, *9*(12), 4504-4521. https://www.mdpi.com/1660-4601/9/12/4504