MOHAMMAD REZA. NEYSHABURI*, HOSSEIN BAYAT**,
MOSTAFA RASTGOU**, KOUROSH MOHAMMADI***,
ANDREW S. GREGORY****, AND NADER NARIMAN-ZADEH*****

PARAMETRIC ESTIMATION OF WATER RETENTION USING MGMDH METHOD AND PRINCIPAL COMPONENT ANALYSIS

Abstract: Performing a primary analysis, such as principal component analysis (PCA) may increase accuracy and reliability of developed pedotransfer functions (PTFs). This study focuses on the usefulness of the soil penetration resistance (PR) and principal components (PCs) as new inputs along with the others to develop the PTFs for estimating the soil water retention curve (SWRC) using a multi-objective group method of data handling (mGMDH). The Brooks and Corey (1964) SWRC model was used to give a description of the water retention curves and its parameters were determined from experimental SWRC data. To select eight PCs, PCA was applied to all measured or calculated variables. Penetration resistance, organic matter (OM), aggregates mean weight diameter (MWD), saturated hydraulic conductivity (K_s), macro porosity (Mp), micro porosity (Mip) and eight selected PCs were used as predictors to estimate the Brooks and Corey model parameters by mGM-DH. Using PR or OM, K_s and MWD, improved the estimation of SWRC in some cases. Using the predicted PR can be useful in the estimation of SWRC. Using either the MP and Mip or the eight

* Prof. M. R. Neyshaburi, DSc.; Department of Soil Science, Faculty of Agriculture, Tabriz University, Tabriz, Iran.

** Assoc. Prof. H. Bayat, DSc., M. Rastgou, Ph. D. Student DSc.; Department of Soil Science, Faculty of Agriculture, Bu Ali Sina University, Hamedan, Iran. E-mail: h.bayat@basu.ac.ir, E-mail: mostafa.rastgo@gmail.com

*** Prof. K. Mohammadi, DSc.; Department of Irrigation and Drainage Engineering, Tarbiat Modares University, Tehran, Iran. E-mail: kourosh.mohammadi@gmail.com

**** Prof. A. S. Gregory, DSc.; Department of Sustainable Soils and Grassland Systems, Rothamsted Research, Harpenden, Hertfordshire AL5 2JQ, UK. E-mail: andy.gregory@rothamsted.ac.uk

***** Prof. N. Nariman-zadeh DSc.; Department of Mechanical Engineering, Engineering Faculty, Guilan University, P.O. Box 3756, Rasht, Iran. E-mail: nnzadeh@yahoo.com

PCs significantly improved the PTFs accuracy and reliability. It would be very useful to apply PCA on the original variables as a primary analysis to develop parametric PTFs.

INTRODUCTION

Pedotransfer functions (PTFs) are functional relationships that transfer easily, routinely, or cheaply measured properties (e.g. soil texture, structure and organic matter content) into missing soil properties (e.g. soil water retention curve,

Notation			
BD	bulk density	MWD	aggregates mean weight diameter
C	clay	OM	organic matter
PR	Penetration resistance	PC1-PC20	principal component 1 - principal component 20
c_1M	intercept of first domain of Millan's model	PTF1 – PTF4	pedotransfer function 1 - pedotransfer function 4
c_2M	intercept of second domain of Millan's model	PTF1 – PTF4	pedotransfer function 1 - pedotransfer function 4
CI	Cone index	RMSE	root mean squared error
D Bi	mass fractal dimension of Bird's model	R^2	coefficient of determination
d_cM	cutoff of the whole domain in Millan's model	S	Sand
d_g	geometric mean of particle diameter	Se^2B	variance of mass of aggregates for Bartoli's model
D_1M	fractal dimension of first domain in Millan's model	Se^2M	variance of cumulative mass of particles for Millan's model
D_2M	fractal dimension of second domain in Millan's model	s_g	geometric standard deviation of particle diameter
IME	integrated mean error	Si	silt
IRMSE	integrated root mean squared error	TP	total porosity
K_s	saturated hydraulic conductivity	θ_s	saturated water content
M_d	median diameter	θ_r	residual water content
MGN	Morgan-Granger-Newbold	α	reverse of air entry value
Mip	micro porosity	λ	pore size distribution index
Mp	macro porosity		

saturated hydraulic conductivity, etc.) [6]. Many attempts have been made to correlate the parameters of the soil hydraulic models of Brooks and Corey [8] [9, 10] and van Genuchten [51] [55] to basic soil properties. The studies include grouping soil into classes [30], developing alternative methods to derive or fit the PTFs [22, 40], finding the most influential soil attributes as predictors

[19, 44, 47] and finally using state-of-the-art techniques to develop PTFs [29]. About using state-of-the-art techniques regression trees technique has been used to estimate water retention by Pachepsky, *et al.* [31]. Nemes, *et al.* [25] and Nemes, *et al.* [27] used k-Nearest Neighbor technique to estimate water retention. Twarakavi, *et al.* [48] used support vector machines to derive a new set of PTFs and found that all the support vector machines-based PTFs performed better than the Rosetta PTF program. Jana and Mohanty [17] used Bayesian neural networks to predict the soil water retention.

The soil hydraulic properties are necessary in many fields such as hydrology and agronomy [46]. Also, water flow and solute transport models need hydraulic properties for accurate simulations [32]. In spite of many attempts that have been made to develop parametric PTFs [9, 10, 36, 38, 55], accurate prediction of SWRC remains a challenging object in soil physics.

Predicting some of the soil physical or chemical properties and using them as inputs for the parametric PTFs, along with performing some primary analysis could be helpful in raising the accuracy and reliability of the developed PTFs. In the light of the high spatial variability of penetration resistance (PR), it is difficult to collect data measured with an adequate accuracy [53]. Moreover, the soil databases are frequently devoid of PR values, which may be useful to predict SWRC [29]. On the other hand, some researchers such as To and Kay [45] and Bayat, *et al.* [3] used regression and artificial neural networks, respectively to estimate PR from the readily available soil data. The question of how reliable it is to apply the published PTFs to predict PR and then, using it as an input for the parametric PTFs, remains unanswered.

Many researchers have encountered the problem of multi-collinearity by using parametric PTFs [40, 43]. Van den Berg, *et al.* [50] suggested that it could be caused by interdependency among the input variables. Principal component analysis (PCA) can solve the problem of multi-collinearity [14] among the predictor variables. It is expected that using principal components (PCs) as predictors can improve the accuracy and reliability of estimates.

This paper is a part of Ph. D. thesis presented by Bayat, *et al.* [3]. Several parts of the research have been published previously. The indicated paper and other reported parts of the research are complementary of one another. Fractal parameters were used to estimate water retention through the Brooks and Corey [7] model by Bayat, *et al.* [2] (parametric method) using multi-objective group method of data handling and artificial neural networks (ANN). Also, fractal parameters have been used as predictors to develop point PTFs using ANNs with adding penetration resistance and principal components to the input parameters [4, 5]. More recently Neyshaburi, *et al.* [28] developed PTFs (point method) to estimate the soil water retention using fractal parameters and multi-objective group method of data handling. The new feature of this article was using multi-objective group method of data handling for the point estimation of water

content and comparing its results with those of the PTFs developed by ANNs. But, the effect of the principal components and the penetration resistance in the estimation of the parameters of the Brooks and Corey [7] model has not been reported so far. Therefore, this research is designated to complete the previously reported researches by using penetration resistance and principal components as inputs to estimate the parameters of the Brooks and Corey [7] model. Then, in order to be able to compare the results, the same data set was used in this study.

The objectives of the present work aimed at investigating: (i) the utility of applying PCA on original variables and using PCs as predictors to estimate the Brooks and Corey water retention parameters, and (ii) the efficiency of the predicted PR to improve the parametric PTFs.

MATERIALS AND METHODS

Sampling and Measurements

In this study a data set of 148 samples were used to develop the PTFs in order to estimate the Brooks and Corey model parameters. The study areas representing a wide range of the soil properties were two provinces of Iran: Hamedan, 124 data (east of Iran) and Guilan, 24 data (north of Iran). This study is a resumption of the research in which PTFs were developed by ANNs and a multi-objective group method of data handling (mGMDH) to predict the Brooks and Corey water retention parameters [2] and all of the data were taken from Bayat, *et al.* [2], Bayat, *et al.* [4], Bayat, *et al.* [5]. Therefore, the details of all measurements for the soil properties are given in the cited papers and supplementary materials, and we do not repeat them here. The mineralogical composition of the clay were greatly (2:1 clay types) vermiculites and illite for Guilan [11] and Hamedan [16] provinces, respectively.

Penetration resistance (PR) was measured in -100 kPa matric potential, only on 24 core samples (5.5 cm in diameter and a height of 4.0 cm) collected from Guilan province. For the other 124 samples collected from Hamedan province, PR was estimated using the PTF that had been developed by Bayat, *et al.* [3] using BD, TP, θ and relative saturation, that are usually available in many databases. Bayat, *et al.* [3] developed several PTFs to estimate PR. We describe the PTFs used in this study. A data set of 293 data from Hamedan province, Iran, was divided into training and validation groups. The first set of relative saturation (θ_v TP^{-1}), BD and θ_m served as inputs to ANN$_S$ in order to predict PR, and the second group used for it's validation. The RMSE and coefficient of correlation of PR prediction were 202.2 kPa and 0.80, respectively.

The Brooks and Corey [8] model was used to give a description of the water retention curves:

$$(1) \quad S = \frac{\theta - \theta_r}{\theta_s - \theta_r} = [\alpha h]^{-\lambda} \qquad h > \frac{1}{\alpha}$$

$$(2) \quad S = 1 \qquad\qquad h \leq \frac{1}{\alpha}$$

where   is the volumetric water content at the matric suction h and S is the degree of saturation. The parameters to be determined are: the residual (θ_r), saturated (θ_s) water contents (m³ m⁻³), the inverse of the air entry pressure α (1 cm⁻¹), and the pore size distribution index λ (-). The model parameters were estimated with the RETC program [52] using experimental soil water retention data.

Because some of the parameters exhibited non-normal distributions, the following transformations were made: ln (α), ln (λ), log (K_s) and log (PR).

Principal Component Analysis

Principal component analysis was applied to all measured and calculated variables and 8 PCs were selected to be used as inputs in the PTFs. Detailed information about the application of principal component analysis and selecting 8 PCs was given in Bayat, *et al.* [4] and supplementary materials.

Developing PTFs

We followed the procedure by Minasny and McBratney [22], who developed "neuropath" software to create PTFs, in developing mGMDH models. 100 data were selected randomly for development and remaining 48 data were used for validation. The PTFs were developed using a multi-objective group method of data handling (mGMDH) technique [24] to estimate θ_r, θ_s, α and λ in the Eq. 1.

The parameters were predicted by the combination of mGMDH and bootstrap method [13] using 50 bootstrap data sets. Details are given in Bayat, *et al.* [4] and supplementary materials.

The mGMDH algorithm uses multi-objective optimization. The mGMDH algorithms are given in detail by Atashkari, *et al.* [1] and Nariman-Zadeh, *et al.* [24]. Genetic algorithms have been used in a multi-objective GMDH neural network for each neuron searching its optimal set of connections with the preceding layer [1]. For the multi-objective optimization evolutionary algorithms were applied. The training and testing mean square errors were two objective functions which determined to be minimized in order to find the most accurate and reliable PTFs. In order to perform the mGMDH algorithm, a program was written in MATLAB (The MathWorks, Natick, MA).*

———————
* The models would be available upon requesting from corresponding author.

Four primary PTFs (from Step 1 to Step 4) were developed before creating the PTF1. Detailed information about the aforementioned four steps is given in the earlier study [2] and supplementary materials. In this paper we focus on the effect of PR, MWD, OM, K_s, Mp, Mip, and PCs in the estimation of SWRC.

For each parameter in Eq. 1 four PTFs were developed using various input variables including PCs. To our knowledge no one has used the estimated PR to predict SWRC parameters. Then, PR was used as a new predictor to develop PTF1, along with the inputs in the previous steps (from Step 1 to Step 4). Therefore, to make the inputs clear for each PTF, they were shown in the related tables. PTF2 was built by using OM, MWD, and K_s as new inputs. If the addition of an input significantly improved the reliability of the PTF, it was used as an input in all of the following steps; if not, it was not used as a predictor in the following steps (it was also applied to develop PTF1-PTF4). To develop PTF3, Mp and Mip were included in the list of inputs. To develop PTF4, eight selected PCs were used as new inputs.

Since Mp and Mip were calculated from the 4 kPa water retention, a question would rise about using it towards fitting a SWRC model or calculating pore size distribution index λ, rather than estimating them. The answer is that the SWRC models and/or λ cannot be fitted or calculated by using only one point of the water retention data. Instead, it's using could improve the SWRC estimation substantially.

Evaluation Criteria

In this study the accuracy and reliability of each PTF refer to the precision of the estimations for the development and validation data sets, respectively. The improvement of each PTF performance in comparison with the previous one has been evaluated using the evaluation criteria. The coefficient of determination (R^2), the root mean square error (RMSE) [57], and the Morgan-Granger-Newbold (MGN) test [12] have been used to evaluate the accuracy and reliability of PTFs. Detailed information about the MGN test is given in the Bayat, *et al.* [4] and supplementary materials. The *MGN* values for significant differences for accuracy and reliability are 1.99 and 2.01, respectively. It means that if the calculated *MGN* values are greater than the aforesaid values, the difference between two PTFs would be significant.

To evaluate the overall performance of the developed PTFs, the estimated parameters of Eq. 1 for each step were used to compute the volumetric water contents corresponding to the experimental values of matric suctions. These water contents were compared with the experimental ones on an individual basis (i.e. curve by curve) by computing the integral mean error, IME (m^3 m^{-3}) and the integral root mean square error, IRMSE (m^3 m^{-3}). The differences were calculated with the numerical quadrature of the following integrals [43]:

$$(3) \quad IME = \frac{1}{b-a} \int_a^b \left(\theta_p - \theta_m \right) d \log|h|$$

$$(4) \quad IRMSE = \left[\frac{1}{b-a} \int_a^b \left(\theta_p - \theta_m \right)^2 d \log|h| \right]^{\frac{1}{2}}$$

where h is the matric suction (kPa) and θ_p and θ_m are the PTF predicted and the measured water contents at that potential. Limits a and b define the pF range over which the experimental curve is measured.

RESULTS AND DISCUSSION

The statistics of the soil properties for the development and validation of data sets are shown in Table 1. Since both data sets were selected randomly, their parameter ranges were the same and the t-test demonstrated their similarity.

TABLE 1. STATISTICS OF THE SOIL PHYSICAL PROPERTIES FOR THE DEVELOPMENT AND VALIDATION OF DATA SETS.

Variable		Development data set [n=100]				Validation data set [n=48]			
		Mean	Standard Deviation	Minimum	Maximum	Mean	Standard Deviation	Minimum	Maximum
OM[*)]	%	1.26	0.98	0.07	6.02	1.16	1.02	0.23	6.81
K_s	cm hour^{-1}	7.85	13.13	0.02	98.85	6.89	10.55	0.01	61.27
PR	kPa	549.9	585.4	46.3	3793.3	501.1	379.7	60.2	1932.2
BD	Mg m^{-3}	1.42	0.16	0.90	1.75	1.45	0.15	1.05	1.77
Mp	m^3 m^{-3}	0.11	0.08	0.00	0.24	0.09	0.07	0.00	0.25
Mip	m^3 m^{-3}	0.40	0.12	0.14	0.67	0.40	0.12	0.14	0.66
C	%	23.72	9.69	5.24	55.37	25.77	10.26	9.58	53.64
Si	%	45.67	12.33	12.95	70.96	44.47	13.58	16.08	69.36
S	%	30.61	17.64	2.15	81.63	29.76	19.33	1.23	73.37
MWD	mm	1.68	1.61	0.15	5.82	1.28	1.22	0.22	5.48

[*)] Refer to the notation list for the variables description.

The results of PCA are shown in Table 2 and the discussion concerning them is given in the Bayat, *et al.* [4] and supplementary materials. The coefficients of the parameters show their effect on each PC (Table 3).

TABLE 2. THE FIRST EIGHT PRINCIPAL COMPONENTS OF ALL ORIGINAL VARIABLES.

Principal component	Eigenvalues	proportion	Cumulative proportion
		%	%
PC1 [*)]	13.07	39.62	39.62
PC2	5.32	16.13	55.75
PC3	3.37	10.21	65.96
PC4	2.10	6.37	72.33
PC5	1.84	5.59	77.92
PC6	1.33	4.02	81.94
PC7	1.27	3.83	85.77
PC8	0.94	2.85	88.62

[*)] Refer to the notation list for the variables description.

Saturated Water Content (θ_S)

Table 4 depicts the four developed PTFs to predict θ_s using the input variables including PR and PCs.

PTF1. By going over the step 1-4 of the previous study [2] only addition of detailed information of PSD (s_g and M_d) to the basic soil properties could improve the estimation of θ_s. Therefore, attributed to PTF1 was compared with the PTF with input variables of C, TP, BD, s_g and M_d. In line with the MGN value no significant improvement occurred when the PR was included in the list of inputs (Table 4). It can be estimated that PR did not improve the θ_s because high water PR content was relatively insensitive to soil structural parameters [23], and therefore PR could not differentiate the structural effects on θ_s. The indicated finding is similar to the one reported by Pachepsky, *et al.* [29], who found that PR did not lead to a significant improvement in θ_s estimates.

TABLE 3. COEFFICIENT OF EACH STANDARDIZED VARIABLE IN THE PCS.

Soil properties		PCs							
		1	2	3	4	5	6	7	8
OM[*)]	[%]	0.018	-0.139	0.062	0.173	0.023	0.094	0.120	0.032
log K_s	[log. cm h[-1]]	-0.038	-0.066	0.040	-0.138	0.178	0.267	0.075	-0.162
TP	[m³ m⁻³]	0.044	-0.110	0.027	0.048	0.016	0.165	0.258	0.250
log CI	[log. kPa]	0.024	0.056	0.036	0.295	-0.104	-0.267	-0.081	-0.152
BD	[Mg m⁻³]	-0.031	0.142	-0.012	0.032	-0.008	-0.207	-0.181	-0.134
Mp	[m³ m⁻³]	-0.041	-0.009	0.020	-0.297	0.123	0.136	0.070	-0.199
Mip	[m³ m⁻³]	0.054	-0.062	0.006	0.210	-0.069	0.008	0.118	0.276
$D_m Y$	-	0.074	0.013	-0.001	-0.072	0.012	0.055	-0.007	-0.085
$Se^2 Y$	-	-0.049	-0.074	-0.170	-0.060	-0.053	-0.085	-0.064	-0.007

Soil properties		PCs							
		1	2	3	4	5	6	7	8
c Bi	[g μm^{D-3}]	0.071	0.038	0.062	-0.059	-0.002	0.032	-0.098	0.112
D Bi	-	0.070	0.044	0.065	-0.034	0.034	0.084	0.022	-0.057
Se^2Bi	-	-0.067	0.029	0.051	0.034	0.016	0.043	0.008	0.123
c_1M	[g μm^{D1-3}]	0.064	0.066	0.087	-0.091	0.043	0.051	-0.111	0.077
c_2M	[μm^{D2-3}]	0.073	-0.013	-0.011	-0.046	-0.063	-0.027	-0.099	0.119
d_cM	[μm]	0.074	0.013	0.021	-0.067	-0.033	0.030	-0.049	0.051
D_1M	-	0.009	0.121	0.157	-0.011	0.094	0.084	-0.018	0.010
D_2M	-	0.073	-0.026	-0.048	-0.020	-0.030	-0.011	0.016	-0.128
Se^2M	-	-0.053	-0.080	-0.149	-0.023	-0.068	-0.072	-0.072	-0.022
d_g	[mm]	-0.063	0.028	0.108	0.045	-0.030	-0.026	-0.047	0.320
s_g	-	-0.003	0.046	0.114	0.250	0.100	0.078	0.404	-0.388
Clay	[%]	0.067	0.052	0.063	-0.096	0.029	0.072	-0.125	0.070
Silt	[%]	0.048	-0.087	-0.140	-0.015	-0.107	-0.100	0.004	-0.014
Sand	[%]	-0.070	0.033	0.064	0.063	0.059	0.030	0.065	-0.028
M_d	[μm]	-0.060	0.036	0.109	0.020	-0.009	-0.007	-0.076	0.375
MWD	[mm]	0.001	-0.077	0.042	0.211	0.090	0.145	-0.440	-0.209
D_mT	-	-0.002	0.146	-0.139	-0.017	-0.109	0.106	0.152	0.101
Se^2T	-	0.003	-0.129	0.146	-0.034	0.122	-0.194	-0.192	-0.171
AR	[cm^{Dn}]	-0.021	-0.068	0.161	-0.093	-0.156	-0.150	0.153	0.317
D_nR	-	0.020	0.129	-0.130	0.031	0.007	-0.066	0.218	-0.055
Se^2R	-	-0.016	-0.007	-0.124	0.120	-0.050	0.479	-0.078	0.082
k_mB	[g cm^{-Dm}]	0.000	0.062	-0.108	0.063	0.400	-0.052	-0.087	0.323
D_mB	-	0.027	-0.021	-0.124	0.109	0.393	-0.059	-0.024	0.202
Se^2B	-	0.017	-0.056	-0.008	-0.097	0.224	-0.350	0.266	-0.014

[*] Refer to the notation list for the variables description.

TABLE 4. DEVELOPMENT AND VALIDATION RESULTS AND ERROR STATISTICS OF Θ_S ESTIMATES MADE WITH DISTINCT INPUTS BY MGMDH PTF2.

Variables selected by mGMDH	MGN[b]	RMSE	R^2	θ_s [a]
		$m^3\ m^{-3}$		
		Development		
C, TP, BD, s_g, M_d, logPR	0.47	0.026	0.881	PTF1
S, Si, C, TP, BD, d_g, s_g	0.94	0.028	0.870	PTF2
Si, C, TP, BD, s_g, Mp, Mip	2.38	0.023	0.908	PTF3
S, C, TP, s_g, PC1, PC3, PC4, PC5, PC7, PC8	0.32	0.026	0.883	PTF4
		Validation		
	0.70	0.033	0.842	PTF1
	1.54	0.035	0.823	PTF2
	1.31	0.029	0.880	PTF3
	2.33	0.029	0.887	PTF4

[a] Refer to the notation list for the variables description.
[b] MGN values for significant differences ($P < 0.05$) for accuracy and reliability are 1.99 and 2.01 respectively.

PTF2. Using OM, K_s and MWD did not lead to a significant improvement either (Table 4). It could result from the inter dependencies between these parameters and the other input variables (e.g. correlation between BD and OM) [35] or the pronounced effect of OM, K_s and MWD on water retention at matric suctions other than 0 kPa (e.g. θ_{FC} or θ_r) [18]. Mayr and Jarvis [21], also found that OM was excluded as a predictor for θ_s by regression technique. Starks, *et al.* [42] also reported that using K_s improved the water retention estimates at some sites, but deteriorated the estimates at some other sites.

PTF3. Using the Mp and Mip caused an improvement, even though it was not significant in the case of PTF reliability (Table 4). This is in agreement with many authors who successfully used water contents at one or two matric suctions to predict SWRC [34, 39, 56].

PTF4. A significant improvement of the PTF reliability occurred when 8 selected PCs were included in the list of inputs (Table 4). Using original variables themselves did not lead to a significant improvement of PTF reliability, whereas using PCs values did. This may explain the complexity of the soil system that is difficult to capture even by the mGMDH technique. It could be the reason why Schaap and Bouten [37] could not estimate θ_s from basic soil properties by ANNs. However, Ungaro, *et al.* [49] found the simplest structure of GMDH network for θ_s.

Residual Water Content (θ_r)

Table 5 depicts the five developed PTFs to predict θ_r using input variables including PR and PCs.

TABLE 5. DEVELOPMENT AND VALIDATION RESULTS AND ERROR STATISTICS OF Θ_R ESTIMATES MADE WITH DISTINCT INPUTS BY MGMDH PTF1 AND PTF2.

Variables selected by mGMDH	MGN [b]	RMSE	R^2	θ_r [a]
		$m^3\ m^{-3}$		
		Development		
S, Si, C, TP, C_2M, d_cM, D_1M, D_2M, Se^2M, logPR	0.34	0.017	0.562	PTF1
Si, C, TP, BD, C_1M, C_2M, d_cM, D_1M, Se^2M, OM	0.34	0.017	0.563	PTF2
S, Si, C, TP, C_1M, C_2M, d_cM, D_1M, Se^2M, Mp, Mip	0.99	0.016	0.603	PTF3
Si, C, TP, C_1M, d_cM, D_1M, Se^2M, PC1, PC2, PC3	0.32	0.017	0.560	PTF4
PC1, PC2, PC3, PC4, PC6, PC9, PC11, PC12, PC13, PC14, PC17, PC20	7.14	0.011	0.820	PTF5
		Validation		
	0.95	0.018	0.481	PTF1
	0.10	0.016	0.597	PTF2

Variables selected by mGMDH	MGN [b]	RMSE	R^2	θ_r [a]
	0.09	0.016	0.584	PTF3
	0.22	0.017	0.549	PTF4
	7.98	0.008	0.912	PTF5

[a] Refer to the notation list for the variables description.

[b] MGN values for significant differences ($P < 0.05$) for accuracy and reliability are 1.99 and 2.01 respectively.

PTF1 and PTF2. Since in the step 1-4 of the previous study (Bayat et al., 2011) only addition of fractal parameters of Millan's model (C_2M, d_cM, D_1M, D_2M and Se^2M) to the basic soil properties could improve the estimation of θ_r therefore, attributed to PTF1 was compared with the PTF with input variables of S, Si, C, TP, C_2M, d_cM, D_1M, D_2M and Se^2M. According to the MGN value no significant improvement occurred when the PR or OM, K_s and MWD were included in the list of inputs (Table 5).

PTF3. No significant improvement occurred when Mp and Mip were included in the list of inputs (Table 5). In physical terms this may suggest that the θ_r dominated by hygroscopic water, depends only on the micropores size distribution, whereas the threshold suction between Mp and Mip is 4 kPa, that could improve the estimation of θ_s and water content at lower suctions, but which did not improve θ_r estimates.

PTF4. No significant improvement occurred when 8PCs were included in the list of inputs (Table 5). It could result from the fact that no significant correlation was found between θ_r and 8PCs (data is not shown). Also, poor correlation of the θ_r with other soil physical properties was reported by Luckner, *et al.* [20] and Tietje and Tapkenhinrichs [43]. We thought it might be in the light of the complexity of the relationship between θ_r and soil physical properties or PCs, thus using more related variables such as clay mineralogy may help to solve the problem. Since, the estimation of θ_r was not improved, using different input variables in various PTFs then, we decided to estimate θ_r using 20 PCs, explaining 99.48% of the total variation of original variables.

PTF5. Only twenty PCs, were used as predictors (without any other original variable), and caused a highly significant improvement of PTF accuracy and reliability (Table 5). Particular features of the soil properties may affect θ_r in such a way that they could not be recognized by using original variables as inputs. Nevertheless, PCA could put these features in some PCs that explain a high percentage of the total variance, and using them as predictors increased development and validation R^2 from 0.543 and 0.608 to 0.82 and 0.912, respectively, which are quite high. Some of the values of R^2 for θ_r estimates found in the literature are 0.722 [55], 0.41 [38], 0.791 [46], 0.68 [49] and 0.599 [41].

Reverse of Air Entry Value (α)

Table 6 depicts the three developed PTFs to predict α using input variables including PR.

TABLE 6. DEVELOPMENT AND VALIDATION RESULTS AND ERROR STATISTICS OF LN A ESTIMATES MADE WITH DISTINCT INPUTS BY MGMDH PTF1.

Variables selected by mGMDH	MGN [b]	RMSE	R^2	ln α [a]
		ln. cm^{-1}		
		Development		
TP, BD, S, Si, C, Se²B, logPR	0.29	1.876	0.449	PTF1
TP, S, Si, Se²B, OM, logKs, MWD	0.94	1.768	0.511	PTF2
TP, BD, S, Se²B, Mp,Mip	9.83	0.988	0.847	PTF3
		Validation		
	0.05	1.868	0.490	PTF1
	0.54	1.780	0.512	PTF2
	8.24	0.962	0.866	PTF3

[a] Refer to the notation list for the variables description.
[b] MGN values for significant differences ($P < 0.05$) for accuracy and reliability are 1.99 and 2.01 respectively.

PTF1. In the step 1-4 of the previous study [2] only TP, BD, S, Si, C and Se²B were selected as input variables to estimate the α. Therefore, attributed to PTF1 was compared with the PTF with those input variables. According to the MGN value no significant improvement occurred when PR was included in the list of inputs (Table 6). A reason could be that PR was predicted from other soil physical properties (e.g. TP, BD) and they had been included in the list of inputs. It is conjectured that the mGMDH procedure, in effect, used TP and BD values to estimate PR, and these estimates had improved estimates of α, thus using PR as another input did not induce further improvement.

PTF2. No significant improvement of PTF accuracy and reliability occurred when OM, K_s and MWD were included in the list of inputs (Table 6).

Selecting OM, K_s and MWD in the equation to estimate α explains that there are correlations between α and the indicated parameters, the correlations, nevertheless, are not strong enough to induce a significant improvement.

PTF3. A highly significant improvement was achieved when Mp and Mip were included in the list of inputs (Table 6). Air entry values were in the range between 0.125 and 607 kPa for 70 % of samples and the threshold between Mp and Mip is 4 kPa. It is inferred from the literature [26, 34] that when water content at a matric suction is used to estimate water content at other matric suctions, the closer the two points are the more accurate the estimates. Therefore

the closeness of 4 kPa with air entry values could explain the highly significant effect of Mp and Mip on the α estimates. Using 8PCs did not induce a significant improvement. This is the reason why results have not been shown.

PORE SIZE DISTRIBUTION INDEX (Λ)

Table 7 depicts the four developed PTFs to predict λ using input variables including PR and PCs.

TABLE 7. DEVELOPMENT AND VALIDATION RESULTS AND ERROR STATISTICS OF LN Λ ESTIMATES MADE WITH DISTINCT INPUTS BY MGMDH PTF1.

Variables selected by mGMDH	MGN [b]	RMSE	R^2	ln λ [a]
		-		
		Development		
S, Si, C, TP, BD, logPR	0.73	0.378	0.395	PTF1
S, Si, C, TP, BD, logKs, MWD	0.28	0.386	0.371	PTF2
S, C, TP, BD, Mp, Mip	2.67	0.333	0.531	PTF3
S, Si, C, TP, Mp, Mip, PC1, PC2, PC3, PC4, PC6, PC7	2.12	0.296	0.629	PTF4
		Validation		
	0.83	0.375	0.360	PTF1
	0.42	0.385	0.330	PTF2
	2.05	0.334	0.493	PTF3
	1.47	0.298	0.606	PTF4

[a] Refer to the notation list for the variables description.
[b] MGN values for significant differences ($P < 0.05$) for accuracy and reliability are 1.99 and 2.01 respectively.

PTF1. In the step 1-4 of the previous study [2] only basic soil properties were selected as input variables to estimate the pore size distribution index. Therefore, attributed to PTF1 was compared with the PTF with input variables of S, Si, C, TP and BD. Using PR improved the PTF accuracy and reliability, however its effect was not significant in line with the MGN value (Table 7). It could result from the correlations between PR and the basic soil physical properties [15] or the error in the PR estimates. It is suggested that using the PR measured values can improve the λ estimates.

PTF2. No significant improvement was achieved by the inclusion of OM, K_s and MWD in the list of inputs (Table 7). The situation is similar to that pointed out by Vereecken, *et al.* [54] who found that OM and K_s were not selected in the regression to estimate λ.

PTF3. A significant improvement of the estimates occurred when Mp and Mip were included in the list of inputs (Table 7). This suggests that λ has been strongly influenced by the pore size distribution. Water content at one or two matric suctions have been successfully used to estimate the WRC [26, 29, 33, 38, 39].

PTF4. A significant improvement of PTF accuracy was achieved when 8PCs were included in the list of inputs; however the improvement of PTF reliability was not significant (Table 7). This suggests that applying PCA on the original variables and using PCs as predictors to estimate λ was very useful. Nevertheless using most of original variables did not induce a significant improvement, whereas using PCs did.

The effect of each input variable on the PTF estimated water content has been evaluated by IRMSE and IME (Table 8).

TABLE 8. THE COMPARISON OF THE PTF ESTIMATED AND MEASURED WATER CONTENTS FOR EACH STEP.

Validation			Development			
R^2	IRMSE	IME	R^2	IRMSE	IME [a]	
0.944	0.0618	-0.0193	0.949	0.0677	-0.0210	Step4 [b]
0.946	0.0581	-0.0155	0.947	0.0691	-0.0228	PTF1
0.948	0.0637	-0.0150	0.957	0.0782	-0.0115	PTF2
0.984	0.0472	-0.0062	0.985	0.0482	-0.0063	PTF3
0.986	0.0364	-0.0106	0.987	0.0355	-0.0045	PTF4

a) Refer to the notation list for the variables description.
b) Results of the step4 in the study of Bayat, *et al.* [2].

Using the predicted PR led to a small improvement of the reliability. It shows that using the predicted PR can be useful in the estimation of SWRC and further investigation is needed. Using OM, K_s and MWD did not improve the water content estimates, whereas Mp and Mip or 8PCs did. The most reliable and accurate water contents estimates were achieved when 8PCs were included in the list of inputs, and using 8 PCs considerably improved the PTFs.

The comparison of the accuracy and reliability of the PTFs which were developed in this study and by Bayat, *et al.* [2], Bayat, *et al.* [5], Bayat, *et al.* [4] and Neyshaburi, *et al.* [28] is shown in Fig. 1.
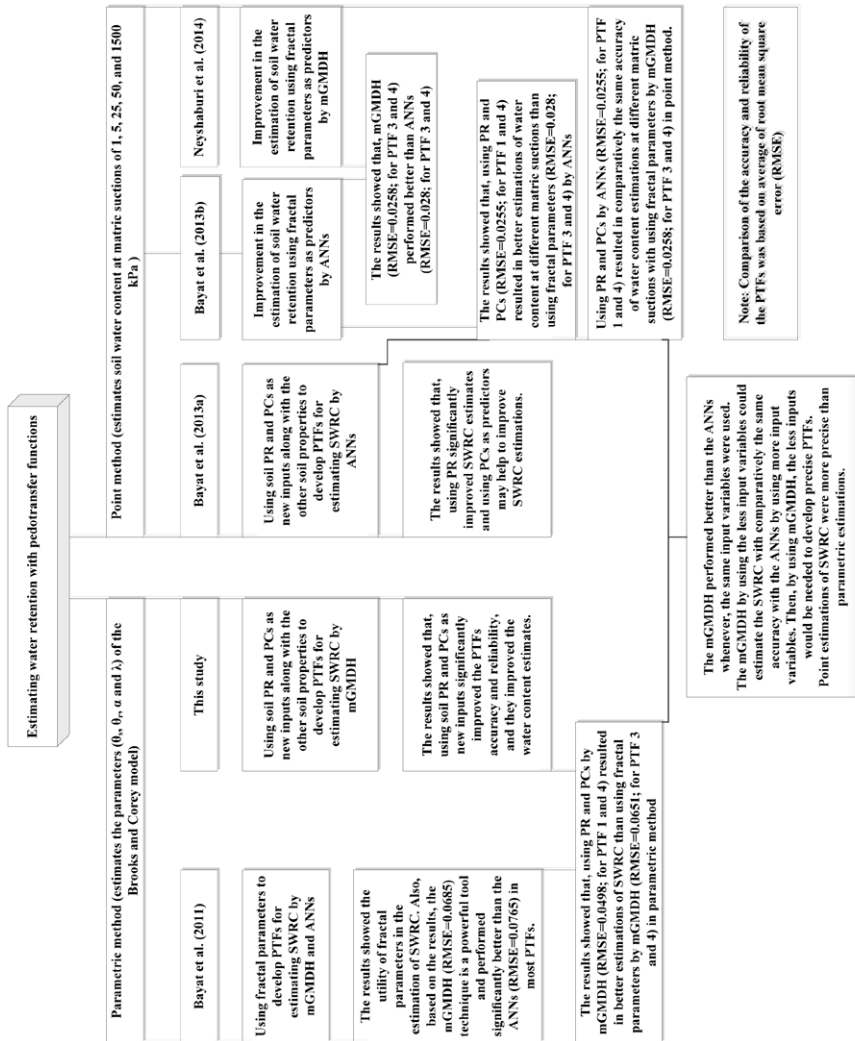
Fig. 1. The comparison of the accuracy and reliability of the PTFs which were developed in this study and by Bayat, *et al.* [2], Bayat, *et al.* [5], Bayat, *et al.* [4] and Neyshaburi, *et al.* [28].

CONCLUSION

1. Using the predicted PR or OM, K_s and MWD improved the estimation of the Brooks and Corey water retention parameters in some cases but they did not induce a significant improvement in the water content estimates.

2. Using either the Mp and Mip or the 8 PCs significantly improved the PTF accuracy and reliability, and they improved the water content estimates. This suggest that using direct information on pore size distribution could be useful to develop parametric PTFs.

3. Applying PCA on the original variables could be used as a primary analysis to develop parametric PTFs.

4. Overall, using PCs, Mp and Mip (if available) as inputs along with the basic soil properties, provide better water retention estimates.

REFERENCES

[1] A t a s h k a r i K., N a r i m a n - Z a d e h N., P i l e c h i A., J a m a l i A., Y a o X., 2005. Thermodynamic Pareto optimization of turbojet engines using multi-objective genetic algorithms. International Journal of Thermal Sciences, 44, 1061–1071.

[2] B a y a t H., N e y s h a b o u r i M., M o h a m m a d i K., N a r i m a n - Z a d e h N., 2011. Estimating water retention with pedotransfer functions using multi-objective group method of data handling and ANNs. Pedosphere, 21, 107–114.

[3] B a y a t H., N e y s h a b u r i M.R., H a j a b b a s i M.A., M a h b o u b i A.A., M o s a d d e g h i M.R., 2008. Comparing neural networks, linear and nonlinear regression techniques to model penetration resistance. Turkish Journal of Agriculture and Forestry, 32, 42–433.

[4] B a y a t H., N e y s h a b u r i M.R., M o h a m m a d i K., N a r i m a n - Z a d e h N., I r a n - n e j a d M., 2013. Improving water content estimations using penetration resistance and principal component analysis. Soil and Tillage Research, 129, 83–92.

[5] B a y a t H., N e y s h a b u r i M.R., M o h a m m a d i K., N a r i m a n - Z a d e h N., I r a n - n e j a d M., G r e g o r y A.S., 2013. Combination of artificial neural networks and fractal theory to predict soil water retention curve. Computers and Electronics in Agriculture, 92, 92–103.

[6] B o u m a J., 1989. Using soil survey data for quantitative land evaluation. Journal, 177–213.

[7] B r o o k s R., C o r e y T., 1964. HYDRAU UC PROPERTIES OF POROUS MEDIA.

[8] B r o o k s R.H., C o r e y A.T., 1964. Hydraulic properties of porous media. Hydrology Papers, Colorado State University.

[9] C l a p p R.B., H o r n b e r g e r G.M., 1978. Empirical equations for some soil hydraulic properties. Water resources research, 14, 601–604.

[10] C o s b y B., H o r n b e r g e r G., C l a p p R., G i n n T., 1984. A statistical exploration of the relationships of soil moisture characteristics to the physical properties of soils. Water Resources Research, 20, 682–690.

[11] D a v a t g a r N., K a v o o s i M., A l i n i a M.H., P e y k a n M., 2006. Study of Potassiun Status and Effect of Physical and Chemical Properties of Soil on it in Paddy Soils of Guilan-Province. JWSS - Isfahan University of Technology, 9, 71-89.

[12] D i e b o l d F.X., M a r i a n o R.S., 2002. Comparing predictive accuracy. Journal of Business & economic statistics, 20.

[13] E f f r o n B., T i b s h i r a n i R.J., 1994. An Introduction to the Bootstrap (Monographs on Statistics and Applied Probability). Journal.

[14] F r e u n d R., L i t t e l l R., 1991. SAS system for regression, 1991. SAS Institute, Cary, NC.

[15] G r u n w a l d S., R o o n e y D., M c S w e e n e y K., L o w e r y B., 2001. Development of pedotransfer functions for a profile cone penetrometer. Geoderma, 100, 25–47.

[16] J a l a l i M., 2007. A study of the quantity/intensity relationships of potassium in some calcareous soils of Iran. Arid Land Research and Management, 21, 133–141.

[17] J a n a R.B., M o h a n t y B.P., 2011. Enhancing PTFs with remotely sensed data for multi-scale soil water retention estimation. Journal of hydrology, 399, 201–211.

[18] J a n a R.B., M o h a n t y B.P., S p r i n g e r E.P., 2007. Multiscale pedotransfer functions for soil water retention. Vadose Zone Journal, 6, 868–878.

[19] K h o d a v e r d i l o o H., H o m a e e M., v a n G e n u c h t e n M.T., D a s h t a k i S.G., 2011. Deriving and validating pedotransfer functions for some calcareous soils. Journal of hydrology, 399, 93–99.

[20] L u c k n e r L., V a n G e n u c h t e n M.T., N i e l s e n D., 1989. A consistent set of parametric models for the two–phase flow of immiscible fluids in the subsurface. Water Resources Research, 25, 2187–2193.

[21] M a y r T., J a r v i s N., 1999. Pedotransfer functions to estimate soil water retention parameters for a modified Brooks–Corey type model. Geoderma, 91, 1–9.

[22] M i n a s n y B., M c B r a t n e y A., 2002. The Method for Fitting Neural Network Parametric Pedotransfer Functions. Soil Science Society of America Journal, 66, 352–361.

[23] M u l q u e e n J., S t a f f o r d J., T a n n e r D., 1977. Evaluation of penetrometers for measuring soil strength. Journal of Terramechanics, 14, 137-151.

[24] N a r i m a n - Z a d e h N., A t a s h k a r i K., J a m a l i A., P i l e c h i A., Y a o X., 2005. Inverse modelling of multi-objective thermodynamically optimized turbojet engines using GMDH-type neural networks and evolutionary algorithms. Engineering Optimization, 37, 437–462.

[25] N e m e s A., R o b e r t s R., R a w l s W.J., P a c h e p s k y Y.A., V a n G e n u c h t e n M.T., 2008. Software to estimate− 33 and− 1500kPa soil water retention using the non-parametric< i> k</i>-Nearest Neighbor technique. Environmental Modelling & Software, 23, 254–255.

[26] N e m e s A., S c h a a p M., W ö s t e n J., 2003. Functional evaluation of pedotransfer functions derived from different scales of data collection. Soil Science Society of America Journal, 67, 1093–-1102.

[27] N e m e s A., T i m l i n D., P a c h e p s k y Y.A., R a w l s W., 2009. Evaluation of the Pedotransfer Functions for their Applicability at the US National Scale. Soil Science Society of America Journal, 73, 1638–1645.

[28] N e y s h a b u r i M.R., B a y a t H., M o h a m m a d i K., N a r i m a n - Z a d e h N., I r a n - n e j a d M., 2014. Improvement in estimation of soil water retention using fractal parameters and multiobjective group method of data handling. Archives of Agronomy and Soil Science, 61, 257–273.

[29] Pachepsky Y., Rawls W., Gimenez D., Watt J., 1998. Use of soil penetration resistance and group method of data handling to improve soil water retention estimates. Soil and Tillage Research, 49, 117–126.

[30] Pachepsky Y.A., Rawls W., 1999. Accuracy and reliability of pedotransfer functions as affected by grouping soils. Soil Science Society of America Journal, 63, 1748–1757.

[31] Pachepsky Y.A., Rawls W., Lin H., 2006. Hydropedology and pedotransfer functions. Geoderma, 131, 308–316.

[32] P a n F., P a c h e p s k y Y., J a c q u e s D., G u b e r A., H i l l R.L., 2012. Data assimilation with soil water content sensors and pedotransfer functions in soil water flow modeling. Soil Science Society of America Journal, 76, 829–844.

[33] R a j k a i K., K a b o s S., V a n G e n u c h t e n M.T., 2004. Estimating the water retention curve from soil properties: comparison of linear, nonlinear and concomitant variable methods. Soil and Tillage Research, 79, 145–152.

[34] R a w l s W., B r a k e n s i e k D., S a x t o n K., 1982. Estimation of soil water properties. Trans. Asae, 25, 1316–1320.

[35] R a w l s W., P a c h e p s k y Y.A., R i t c h i e J., S o b e c k i T., B l o o d w o r t h H., 2003. Effect of soil organic carbon on soil water retention. Geoderma, 116**,** 61–76.

[36] S a x t o n K., R a w l s W.J., R o m b e r g e r J., P a p e n d i c k R., 1986. Estimating generalized soil-water characteristics from texture. Soil Science Society of America Journal, 50**,** 1031–1036.

[37] S c h a a p M.G., B o u t e n W., 1996. Modeling water retention curves of sandy soils using neural networks. Water Resources Research, 32**,** 3033-3040.

[38] S c h a a p M.G., L e i j F.J., v a n G e n u c h t e n M.T., 1998. Neural network analysis for hierarchical prediction of soil hydraulic properties. Soil Science Society of America Journal, 62**,** 847–855.

[39] S c h a a p M.G., L e i j F.J., v a n G e n u c h t e n M.T., 2001. Rosetta: A computer program for estimating soil hydraulic parameters with hierarchical pedotransfer functions. Journal of hydrology, 251**,** 163–176.

[40] S c h e i n o s t A., S i n o w s k i W., A u e r s w a l d K., 1997. Regionalization of soil water retention curves in a highly variable soilscape, I. Developing a new pedotransfer function. Geoderma, 78**,** 129–143.

[41] S h a r m a S.K., M o h a n t y B.P., Z h u J., 2006. Including topography and vegetation attributes for developing pedotransfer functions. Soil Science Society of America Journal, 70**,** 1430–1440.

[42] S t a r k s P.J., H e a t h m a n G.C., A h u j a L.R., Ma L., 2003. Use of limited soil property data and modeling to estimate root zone soil water content. Journal of Hydrology, 272**,** 131–147.

[43] T i e t j e O., T a p k e n h i n r i c h s M., 1993. Evaluation of pedo-transfer functions. Soil Science Society of America Journal, 57**,** 1088–1095.

[44] T i m l i n D., A h u j a L., P a c h e p s k y Y., W i l l i a m s R., G i m e n e z D., Rawls W., 1999. Use of Brooks-Corey parameters to improve estimates of saturated conductivity from effective porosity. Soil Science Society of America Journal, 63**,** 1086–1092.

[45] T o J., K a y B.D., 2005. Variation in penetrometer resistance with soil properties: The contribution of effective stress and implications for pedotransfer functions. Geoderma, 126**,** 261–276.

[46] T o m a s e l l a J., P a c h e p s k y Y., C r e s t a n a S., R a w l s W., 2003. Comparison of two techniques to develop pedotransfer functions for water retention. Soil Science Society of America Journal, 67, 1085–1092.

[47] T ó t h B., Makó A., G u a d a g n i n i A., T ó t h G., 2012. Water retention of salt-affected soils: quantitative estimation using soil survey information. Arid Land Research and Management, 26**,** 103–121.

[48] T w a r a k a v i N.K., Š i m ů n e k J., S c h a a p M., 2009. Development of pedotransfer functions for estimation of soil hydraulic parameters using support vector machines. Soil Science Society of America Journal, 73, 1443–1452.

[49] U n g a r o F., C a l z o l a r i C., B u s o n i E., 2005. Development of pedotransfer functions using a group method of data handling for the soil of the Pianura Padano–Veneta region of North Italy: water retention properties. Geoderma, 124, 293–317.

[50] V a n d e n B e r g M., K l a m t E., Van R e e u w i j k L., S o m b r o e k W., 1997. Pedotransfer functions for the estimation of moisture retention characteristics of Ferralsols and related soils. Geoderma, 78**,** 161–180.

[51] V a n G e n u c h t e n M.T., 1980. A closed-form equation for predicting the hydraulic conductivity of unsaturated soils. Soil Science Society of America Journal, 44**,** 892–898.

[52] V a n G e n u c h t e n M.T., Leij F., Yates S., 1991. The RETC code for quantifying the hydraulic functions of unsaturated soils.

[53] Vaz C.M., Bassoi L.H., Hopmans J.W., 2001. Contribution of water content and bulk density to field soil penetration resistance as measured by a combined cone penetrometer–TDR probe. Soil and Tillage Research, 60**,** 35–42.

[54] Vereecken H., Diels J., Van Orshoven J., Feyen J., Bouma J., 1992. Functional evaluation of pedotransfer functions for the estimation of soil hydraulic properties. Soil Science Society of America Journal, 56**,** 1371–1378.

[55] Vereecken H., Maes J., Feyen J., Darius P., 1989. Estimating the soil moisture retention characteristic from texture, bulk density, and carbon content. Soil science, 148**,** 389–403.

[56] Wagner B., Tarnawski V., Wessolek G., Plagge R., 1998. Suitability of models for the estimation of soil hydraulic parameters. Geoderma, 86**,** 229–239.

[57] Wösten J., Pachepsky Y.A., Rawls W., 2001. Pedotransfer functions: bridging the gap between available basic soil data and missing soil hydraulic characteristics. Journal of hydrology, 251**,** 123–150.