PEDOMODELS FITTING WITH FUZZY LEAST SQUARES REGRESSION

J. MOHAMMADI AND S. M. TAHERI

ABSTRACT. Pedomodels have become a popular topic in soil science and environmental research. They are predictive functions of certain soil properties based on other easily or cheaply measured properties. The common method for fitting pedomodels is to use classical regression analysis, based on the assumptions of data crispness and deterministic relations among variables. In modeling natural systems such as soil system, in which the above assumptions are not held true, prediction is influential and we must therefore attempt to analyze the behavior and structure of such systems more realistically. In this paper we consider fuzzy least squares regression as a means of fitting pedomodels. The theoretical and practical considerations are illustrated by developing some examples of real pedomodels.

1. Introduction

Design and choice of appropriate soil management systems to bring about desired on soil quality depend on knowledge of interrelationships among soil properties. Mostly, classical regression models are used to relate properties of importance to production and resource management to basic, easily or cheaply measured soil properties. In soil science, these models are called pedotransfer functions, coined by Bouma [1], as translating data we have into what we need.

Different types of functions, mainly based on conventional regression, have been developed to predict physical and chemical properties of soil [7]. One of the main research area is developing alternative methods to fit pedotransfer functions. For example, one of the recent approaches for fitting such functions is to use artificial neural networks [4]. Although, some researchers found that artificial neural networks perform better than classical regression methods, however, both procedures are based on assumption of data crispness and deterministic relations among variables.

In modeling natural systems like soil system, in which the above assumptions are not held and, therefore, estimation is influential, we must attempt to analyze behavior and structure of such systems in a more realistic manner. In these situations, we may deal with a fuzzy structure, which can be represented as a function whose parameters are given by fuzzy numbers, and/or with non-precise data [3,8].

The fuzzy parameters of the obtained models mean a possibility distribution, which corresponds to the vagueness of the phenomenon. These models are typically called fuzzy regression models [8].

Received: February 2004; Accepted: August 2004
Keywords and phrases : Pedomodels, Pedotransfer Functions, Fuzzy Least Squares, Fuzzy Regression.

To avoid confusion with conventional regression functions, we introduced the term *pedomodels* (derived from Greek root of *pedo* as soil) implying modeling soil characteristics as a fuzzy natural system.

Fuzzy regression aims to provide regression models in fuzzy environments, i.e., when the relationships among variables are non-precise and/or when the observations are non-exact. From the conceptual and methodological point of view, several approaches exist to solve model-fitting problems under fuzzy circumstances [3,10,11].

In this paper, a least squares approach was used to develop fuzzy pedomodels. In this approach, observations of the dependent (response) variables are considered as non-precise data, and the mathematical pedomodel is considered as a fuzzy model. This approach is based on the notion of distance between the predicted fuzzy outputs and the observed fuzzy inputs. We used a goodness of fit index to reliability analysis in obtained fuzzy models. Sensitivity analysis, based on the amount of vagueness in data, is studied. In addition, to investigate outliers in obtained fuzzy pedomodels, a method is introduced and used in numerical examples.

2. Theory

We review some basic concepts in fuzzy arithmetic, as well as, a particular distance between fuzzy numbers, based on the integral of distance of every level set. A method of fuzzy least squares regression, due to Xu and Li [11] is described. It turns out that, their approach has some drawbacks in practice. To deal with such problems, a complement procedure is proposed. This approach is illustrated through a few case studies followed by soil characteristics, in Section 3.

2.1 Preliminaries

Definition 2.1. A fuzzy number $\widetilde{A}$ is said to be a triangular fuzzy number if its membership function can be expressed as:

$$\widetilde{A}(x)=\begin{cases}\dfrac{x-\left(a-s_a^L\right)}{s_a^L} & a-s_a^L\le x\le a\\ \dfrac{\left(a+s_a^R\right)-x}{s_a^R} & a\le x\le a+s_a^R\\ 0 & \textit{otherwise}\end{cases}$$

We write $\widetilde{A}=\left(a,s_a^L,s_a^R\right)$, where a is the mean value of $\widetilde{A}$, and s_a^L and s_a^R are called left and right spreads, respectively. In special case, if $s_a^L=s_a^R=s_a$, then $\widetilde{A}$ is called symmetric triangular fuzzy number and we write $\widetilde{A}=\left(a,s_a\right)$. Denote by $\Im_s$ the set of all symmetric triangular fuzzy numbers.

For the linear operations on triangular fuzzy numbers, we have the following proposition [12]:

Proposition 2.2. *Let* $\widetilde{A}=\left(a,s_a^L,s_a^R\right)$ *and* $\widetilde{B}=\left(b,s_b^L,s_b^R\right)$ *be two triangular fuzzy numbers. Then*

$$1a)\ \lambda\otimes\widetilde{A}=\left(\lambda a,\lambda s_a^L,\lambda s_a^R\right),\lambda>0$$

$$1b)\ \lambda\otimes\widetilde{A}=\left(\lambda a,-\lambda s_a^R,-\lambda s_a^L\right),\lambda<0$$

$$2)\ \widetilde{A}\oplus\widetilde{B}=\left(a+b,s_a^L+s_b^L,s_a^R+s_b^R\right)$$

In order to find an optimal fuzzy regression model, defining a distance between two fuzzy numbers is necessary. We use the following distance, which is introduced by Xu [10], to establish the multivariate least squares fitting fuzzy model.

Definition 2.3. Suppose that $\widetilde{A}$ and $\widetilde{B}$ are two fuzzy numbers. The distance between $\widetilde{A}$ and $\widetilde{B}$ according to function f(α) is defined as:

$$d\left(\widetilde{A},\widetilde{B}\right)=\left[\int_0^1 f(\alpha)d^2\left(\widetilde{A}_\alpha,\widetilde{B}_\alpha\right)d\alpha\right]^{\frac{1}{2}} \tag{2}$$

in which

$$d^2\left(\widetilde{A}_\alpha,\widetilde{B}_\alpha\right)=\left[a_1(\alpha)-b_1(\alpha)\right]^2+\left[a_2(\alpha)-b_2(\alpha)\right]^2, \tag{3}$$

and $\widetilde{A}_\alpha=\left[a_1(\alpha),a_2(\alpha)\right]$, $\widetilde{B}_\alpha=\left[b_1(\alpha),b_2(\alpha)\right]$ are α-cuts of $\widetilde{A}$ and $\widetilde{B}$, respectively; and f(α) is an increasing function on [0,1] for which f(0)=0 and $\int_0^1 f(\alpha)d\alpha=0.5$.

Note that $d(\widetilde{A}_\alpha,\widetilde{B}_\alpha)$ defines a distance between the α level sets of fuzzy numbers $\widetilde{A}$ and $\widetilde{B}$. Function f(α) can be regarded as a weighting factor for $d^2(\widetilde{A}_\alpha,\widetilde{B}_\alpha)$. Monotonic increasing behavior of f(α) leads to placing more importance on higher membership degrees to determine distance between two fuzzy numbers. The conditions f(0)=0 and $\int_0^1 f(\alpha)d\alpha=0.5$ ensure that the above mentioned distance will be only a generalization of a conventional distance in R. In the following, f(α)=α was used as the weighting function. In this paper, we use the above distance to establish a method of multiple least squares model fitting.

Using distance d, we can draw the following result about the distance of symmetric triangular fuzzy numbers.

Theorem 2.4. *If* $\widetilde{A}=(a,s_a)$ *and* $\widetilde{B}=(b,s_b)$ *are two symmetric triangular fuzzy numbers, then:*

$$d^2(\widetilde{A},\widetilde{B})=(a-b)^2+\frac{1}{6}(s_a-s_b)^2$$

Proof. The membership function of symmetric triangular fuzzy number $\widetilde{A}$ is:

$$\widetilde{A}(x)=\begin{cases}\dfrac{x-(a-s_a)}{s_a} & a-s_a\le x\le a\\ \dfrac{(a+s_a)-x}{s_a} & a\le x\le a+s_a\\ 0 & \textit{otherwise}\end{cases}$$

So the level sets of $\widetilde{A}$ is:

$$\widetilde{A}_\alpha=[(\alpha-1)s_a+a,(1-\alpha)s_a+a].$$

Similarly, we have

$$\widetilde{B}_\alpha=[(\alpha-1)s_b+b,(1-\alpha)s_b+b].$$

Now, based on Equation (3), we have

$$\begin{aligned}d^2(\widetilde{A}_\alpha,\widetilde{B}_\alpha)&=[(\alpha-1)s_a+a-(\alpha-1)s_b-b]^2+[(1-\alpha)s_a+a-(1-\alpha)s_b-b]^2\\&=[\alpha(s_a-s_b)+(a-b)-(s_a-s_b)]^2+[\alpha(s_b-s_a)+(a-b)+(s_a-s_b)]^2.\end{aligned}$$

Therefore, using $f(\alpha)=\alpha$,

$$\begin{aligned}d(\widetilde{A},\widetilde{B})&=\int_0^1\alpha[\alpha(s_a-s_b)+(a-b)-(s_a-s_b)]^2\,d\alpha+\int_0^1\alpha[\alpha(s_b-s_a)+(a-b)+(s_a-s_b)]^2\,d\alpha\\&=\tfrac{1}{4}(s_a-s_b)^2+\tfrac{1}{2}[(a-b)-(s_a-s_b)]^2+\tfrac{2}{3}(s_a-s_b)[(a-b)-(s_a-s_b)]\\&\quad+\tfrac{1}{4}(s_b-s_a)^2+\tfrac{1}{2}[(a-b)-(s_a-s_b)]^2+\tfrac{2}{3}(s_b-s_a)[(a-b)-(s_a-s_b)]\\&=(a-b)^2+\tfrac{1}{6}(s_a-s_b)^2.\end{aligned}$$

2.2 Formulation of fuzzy least squares regression

The general fuzzy regression analysis, studied and applied in this paper, can be stated as follows. Given the set of observed data $(\tilde{y}_i,x_{i1},\ldots,x_{in})$, i=1,…,m in which $\tilde{y}_i\in\Im_s$,

i=1,…,m, and $x_{ij} \in R$, j=1,…,n, i=1,…,m; find an optimal model with fuzzy parameters such as:

$$\tilde{Y} = \tilde{A}_0 + \tilde{A}_1 x_1 + ... + \tilde{A}_n x_n \ , \qquad \tilde{A}_j \in \Im_s , j = 1,..., n \tag{4}$$

Note that, here it is assumed that data on dependent variable, i.e. $\tilde{y}_i$, i=1,…,m, and the coefficients $\tilde{A}_j$, j=1,…,n, are symmetric triangular fuzzy numbers, and independent variables are assumed to be crisp.

2.3 Estimation of model parameters

In either cases of fuzzy regression, a criterion needs to be selected so as to obtain an optimal model of the form of (4). We do this by a least squared error type method, using distance *d* in (2), as a measure of discrepancy between $\tilde{Y}_i$ and $\tilde{y}_i$. This can be achieved by minimizing the sum of squared errors between $\tilde{Y}_i$ and $\tilde{y}_i$, i=1,…,m, i.e.

$$SSE\left(\tilde{A}_0, \tilde{A}_1, ..., \tilde{A}_n\right) == \sum_{i=1}^{m} d^2\left(\tilde{Y}_i, \tilde{y}_i\right) = \sum_{i=1}^{m} d^2\left(\tilde{A}_0 + \tilde{A}_1 x_{i1} + ... + \tilde{A}_n x_{in}, \tilde{y}_i\right).$$

Based on Proposition 1,

$$\tilde{A}_0 + \tilde{A}_1 x_{i1} + ... + \tilde{A}_n x_{in} = (a_0 + a_i x_{i1} + ... + a_n x_{in} \ , \ \sigma_0 + \sigma_1 x_{i1} + ... + \sigma_n x_{in} \ ),$$

therefore, the minimization problem can be rewritten as:

$$\begin{aligned} \textit{Minimize} \quad SSE\left(\tilde{A}_0, \tilde{A}_1, ..., \tilde{A}_n\right) = & \sum_{i=1}^{m} \left(a_0 + a_i x_{i1} + ... + a_n x_{in} - y_i\right)^2 + \\ & \frac{1}{6} \sum_{i=1}^{m} \left(\sigma_0 + \sigma_1 x_{i1} + ... + \sigma_n x_{in} - s_i\right)^2 . \end{aligned} \tag{5}$$

Setting the partial derivatives of the SSE with respect to a_j and σ_j to 0, leads to the following equations:

$$\begin{aligned} & a_0 \sum_{i=1}^{m} x_{i0} x_{ij} + a_1 \sum_{i=1}^{m} x_{i1} x_{ij} + ... + a_n \sum_{i=1}^{m} x_{in} x_{ij} = \sum y_i x_{ij}, \quad j = 0,1,...,n \\ & \sigma_0 \sum_{i=1}^{m} x_{i0} x_{ij} + \sigma_1 \sum_{i=1}^{m} x_{i1} x_{ij} + ... + \sigma_n \sum_{i=1}^{m} x_{in} x_{ij} = \sum s_i x_{ij}, \quad j = 0,1,...,n \end{aligned} \tag{6}$$

where x_{i0}=1 , i=1,…,m.

Equations (6) can be represented in matrix form:

$$Aa = y \ , \quad A\sigma = s \ , \tag{7}$$

where

$$A = X'X \quad , \quad X = \begin{pmatrix} 1 & x_{11} & \dots & x_{1n} \\ 1 & x_{21} & \dots & x_{2n} \\ \dots & \dots & \dots & \dots \\ 1 & x_{m1} & \dots & x_{mn} \end{pmatrix}_{m\times(n+1)}$$

$$a = (a_0, a_1, ..., a_n)^T \quad , \quad y = \left(\sum_{i=1}^{m} y_i x_{i0}, \sum y_i x_{i1}, ..., \sum y_i x_{in} \right)^T ,$$

$$\sigma = (\sigma_0, \sigma_1, ..., \sigma_n)^T \quad , \quad s = \left(\sum_{i=1}^{m} s_i x_{i0}, \sum s_i x_{i1}, ..., \sum s_i x_{in} \right)^T .$$

According to [11], the following results about the existence and uniqueness of solution for system (7), hold:

Proposition 2.5. *If Rank(X) = n+1, then matrix A is positive definite.*

Proposition 2.6. *If Rank(X) = n+1, then the system of equations has unique solution as:*

$$a = A^{-1}y \ , \ \sigma = A^{-1}s \tag{8}$$

Proposition 2.7. *If Rank(X) = n+1 and $A^{-1}s \geq 0$, then the minimization problem has a unique solution which can be achieved by (8).*

2.4 The case in which $A^{-1}s<0$

The condition $A^{-1}s\geq 0$, in Proposition 4, guarantees that spread of fuzzy parameters, i.e., σ_i, i=1,…,n, will be non-negative. However, it is possible to encounter conditions that $A^{-1}s<0$. Under this circumstance, the Proposition 4 cannot be used. Xu and Li [11] considered only the situation in which $A^{-1}s\geq 0$. Thus, it seems to be necessary to consider the case in which $A^{-1}s<0$ when it occurs.

We suggest the following procedure to remove this difficulty. If we encounter negative values for one or more σ_i, spread of parameters, we may run the model again but with considering such parameters as crisp, i.e., the spread of fuzzy coefficient is set to zero. Thus, the center values, a_i, are estimated as before, but the spreads should be calculated by $\sigma^*=A^{*-1}s^*$, where s^* is a vector like s in which elements corresponding to crisp coefficients are omitted, and A^* is a matrix like A in which its rows corresponding to crisp coefficient are removed. We use and explain this procedure by way of some numerical examples, in Section 3.

2.5 Reliability analysis

In order to evaluate the goodness of fit of the model, we use the following goodness of fit index introduced in [9].

Definition2.8. Let $\widetilde{A}$ and $\widetilde{B}$ be two fuzzy numbers, and

$$\widetilde{A} \circ \widetilde{B} = \sup_{x \in R} \{\min[\widetilde{A}(x), \widetilde{B}(x)]\},$$
$$\widetilde{A} \bullet \widetilde{B} = \inf_{x \in R} \{\max[\widetilde{A}(x), \widetilde{B}(x)]\}.$$

Then

$$I(\widetilde{A}, \widetilde{B}) = \min(\widetilde{A} \circ \widetilde{B}, 1 - (\widetilde{A} \bullet \widetilde{B})), \tag{9}$$

is called the goodness of fit of $\widetilde{A}$ and $\widetilde{B}$.

This index has some properties, which are summarized in the following proposition.

Proposition 2.9.

1) $0 \le I(\widetilde{A}, \widetilde{B}) \le 1,$
2) $\widetilde{A} = \widetilde{B} \Rightarrow I(\widetilde{A}, \widetilde{B}) = 1,$
3) $I(\widetilde{A}, \widetilde{B}) = I(\widetilde{B}, \widetilde{A}),$
4) $\widetilde{A} \subseteq \widetilde{B} \subseteq \widetilde{C} \Rightarrow I(\widetilde{A}, \widetilde{C}) \le \min(I(\widetilde{A}, \widetilde{B}), I(\widetilde{B}, \widetilde{C}))$.

Proposition 2.9 indicates that $I(\widetilde{A}, \widetilde{B})$ is measure of similarity of $\widetilde{A}$ and $\widetilde{B}$, and $I(\widetilde{A}, \widetilde{B}) = 1$ when $\widetilde{A} = \widetilde{B}$. Therefore, we use this index as the measure of goodness of fit between the observed value $\tilde{y}_i$ and the predicted value $\tilde{Y}_i$, in a fuzzy regression model.

Remark 2.10. Sadeghpour and Gien presented and employed a different goodness of fit index to evaluate a fuzzy model with fuzzy observations on dependent variable. For a comparison between their index and index d, see [6].

3. Case studies

One of the classical problems in soil science is the measurement of some physical, chemical and/or biological soil properties. The problem results from the difficulty, time and cost of direct measurements. In this article, three pedomodels including one, two, and three independent variables were studied to develop relationships between different chemical and physical soil properties by means of fuzzy least squares regression technique. The study area is a part of Silakhor plain situated in the province of Lorestan (west of Iran), between the cities of Broujerd and Durood. The 100

hectares field, located in the middle of the plain, were sampled on a grid with intersections at 200 m interval. A total of 25 core samples were obtained from 0.0 to 25-cm depth. Different soil physical and chemical properties were measured using standard procedures [5]. The data set was given in Table 1.

3.1 Pedomodel of ESP-SAR

The first model provides relationship between exchangeable sodium percentage (ESP), as a dependent variable, and sodium absorption ratio (SAR), as an independent variable. The exchange sodium percentage, ESP, governs the source/sink phenomenon for ionic constituents, i.e., sodium, as a contaminant in sodic soils, is calculated from ratio of exchangeable sodium, Na_x, to cation exchangeable capacity, CEC. All these soil parameters measured on soil colloidal surface, in turn are time consuming and costly. Due to close relationship between distribution of cations in exchange and solution phases, it is preferred to estimate ESP from sodium adsorption ratio, SAR, i.e., $Na/\left(Ca+Mg/2\right)^{0.5}$, in soil solution [2,5].

TABLE 1. Measured soil properties at sampling locations

No. of Obs.	CEC (Cmol(+)/kg)	ESP (%)	SAR	OM (%)	SAND (%)	SILT (%)	SP (%)
1	16.5	3.08	0.78	0.88	35	45	38.0
2	18.6	2.86	0.64	1.13	37	42	41.0
3	19.3	6.25	0.62	1.31	27	43	47.5
4	20.3	4.11	0.49	1.98	29	41	51.0
5	17.3	1.04	1.10	1.02	38	39	35.0
6	20.4	2.71	0.61	1.29	32	39	43.0
7	19.3	4.45	0.74	1.52	29	37	54.0
8	21.9	6.92	1.15	1.33	18	45	52.0
9	15.9	7.41	1.08	1.71	40	38	45.0
10	18.3	9.08	0.38	2.00	28	46	50.0
11	22.6	6.56	0.61	1.68	13	40	58.6
12	23.7	5.05	0.98	2.15	19	41	62.0
13	24.4	5.23	0.71	3.52	31	41	60.0
14	21.8	5.16	0.51	2.33	31	42	52.0
15	23.8	11.10	0.77	1.71	17	50	52.0
16	20.8	4.74	0.99	1.14	14	53	49.0
17	17.5	28.84	3.56	0.99	19	44	49.0
18	17.8	9.43	0.86	1.14	28	43	44.0
19	20.2	4.50	0.61	1.46	26	44	49.0
20	20.0	9.30	0.64	1.81	32	42	50.3
21	22.8	9.48	0.71	1.38	10	49	52.0
22	19.1	3.65	0.61	0.84	38	43	42.0
23	12.1	10.14	0.63	1.48	49	35	40.0
24	12.8	3.00	1.13	1.08	42	44	37.0
25	5.3	2.00	Missing	0.36	79	14	21.2

In this case, ESP is considered as cost and time variable, therefore the need for less expensive indirect measurement is emphasized. Measurements of SAR have been related to ESP due to low cost, simplicity, and the possibility of relating measurements to the quantity and quality parameters.

3.1.1 Estimation of model parameters

To develop a relationship between ESP (as a dependent variable) and SAR (as an independent variable) based on 24 observations, using fuzzy least squares regression, first, values of dependent variable, y_i, were fuzzified using a symmetric triangular fuzzy number which it's right and left spreads proportional to y_i. For all case studies, fuzzification was performed using $0.10y_i$. This starting value for the amount of vagueness in observations can be based upon expert opinion and might be considered as the acceptable level of uncertainty. Since this value is selected subjectively, we considered different amount of fuzzification during sensitivity analysis. This allowed us to evaluate the effects of different fuzzification scenarios on selected pedomodels.

Considering $s_i = 0.10y_i$, the system of equations becomes:

$$X = \begin{pmatrix} 1 & x_{11} \\ 1 & x_{21} \\ \dots & \dots \\ 1 & x_{m1} \end{pmatrix} = \begin{pmatrix} 1 & 0.78 \\ 1 & 0.64 \\ \dots & \dots \\ 1 & 1.13 \end{pmatrix}_{24\times 2}$$

$$A = X'X = \begin{pmatrix} 24 & 20.9 \\ 20.9 & 26.81 \end{pmatrix}$$

$$a = (a_0, a_1)^T \ , y = \left(\sum_{i=1}^{24} y_i x_{i0}, \sum_{i=1}^{24} y_i x_{i1} \right)^T = (164.09, 202.32)^T,$$

$$\sigma = (\sigma_0, \sigma_1)^T \ , s = \left(\sum_{i=1}^{24} s_i x_{i0}, \sum_{i=1}^{24} s_i x_{i1} \right)^T = (16.41, 20.23)^T.$$

Since Rank(X)=2, the Proposition 2.5 is fulfilled, i.e., the system of equations $Aa=y$, $A\sigma=s$ have unique solution as:

$$a=A^{-1}y = (0.8265, 6.5021)^T,$$
$$\sigma=A^{-1}s= (0.0827, 0.6902)^T.$$

Since $A^{-1}s \geq 0$, according to Proposition 2.7, the minimization problem of the error sum of squares results in unique solution as above. Consequently, the optimal model is:

$$\tilde{y} = \tilde{A}_0 + \tilde{A}_1 x_1 = (0.8265, 0.0827) + (6.5021, 0.6902) x_1 \ .$$

For example, if x= 0.75, then the predicted value of ESP will be:

$$\tilde{y} = (0.8265, 0.0827) + (6.5021, 0.6902) = (5.70, 0.60) \ .$$

It means that the predicted value of ESP is about 5.70 with possible minimum and maximum values 5.10 and 6.30, respectively.

3.1.2 Reliability and sensitivity analysis

The performance of the pedomodel predicting the ESP was assessed using $I(\tilde{y}_i, \tilde{Y}_i)$, $i=1,\ldots,m$. The results are given in Table 2. As it is shown, the I index is very small for almost all-individual data points implying that the considered model, using $s_i=0.10y_i$, is not suitable for fitting data. This table also showed the same performance criterion using different amount of vagueness in y_i. The corresponding optimal models are given in Table 3. The results show that the relative improvement is achieved by increasing the width of y_i. However, the better fitting is achieved at the expense of increasing vagueness in predicted y_i. Although the value of I index can be increased by increasing the width of fuzzified y_i, but it is not sensible to fuzzify more than $0.25y_i$, since, in measuring ESP, vagueness more than such amount is rarely happened.

Selecting the model with $s_i=0.25y_i$, results in I values which are more than 0.80 for almost all individual data points.

TABLE 2. Goodness of fit index, I, for different amount of fuzzifications in y_i's.

No. of obs.	$s_i=0.05y_i$	$s_i=0.10y_i$	$s_i=0.15y_i$	$s_i=0.20y_i$	$s_i=0.25y_i$	$s_i=0.25y_i$ (After removing outliers)
1	0.0000	0.0000	0.0064	0.0585	0.1626	0.1305
2	0.0000	0.0002	0.0214	0.1149	0.2504	0.2150
3	0.0172	0.3623	0.6369	0.7758	0.8501	0.8923
4	0.9454	0.9861	0.9938	0.9965	0.9978	0.9929
5	0.0000	0.0000	0.0000	0.0000	0.0001	--------
6	0.0000	0.0001	0.0181	0.1048	0.2361	0.2047
7	0.0003	0.1297	0.4034	0.6001	0.7212	0.6515
8	0.0040	0.2510	0.5410	0.7078	0.8016	0.7060
9	0.2917	0.7349	0.8721	0.9259	0.9519	0.8915
10	0.0000	0.0000	0.0001	0.0064	0.0395	--------
11	0.0010	0.1781	0.4645	0.6497	0.7588	0.8076
12	0.0000	0.0176	0.1661	0.3643	0.5240	0.4377
13	0.4391	0.8140	0.9126	0.9499	0.9676	0.9337
14	0.0303	0.4172	0.6780	0.8037	0.8695	0.8939
15	0.0000	0.0003	0.0253	0.1264	0.2662	0.3193
16	0.0000	0.0039	0.0851	0.2501	0.4119	0.3350
17	0.1997	0.6685	0.8361	0.9042	0.9376	0.9932
18	0.0000	0.0663	0.2993	0.5073	0.6477	0.7289
19	0.2816	0.7285	0.8687	0.9239	0.9506	0.9199
20	0.0000	0.0004	0.0315	0.1431	0.2881	0.3313
21	0.0000	0.0023	0.0667	0.2181	0.3774	0.4361
22	0.0000	0.0782	0.3222	0.5288	0.6651	0.6119
23	0.0000	0.0000	0.0094	0.0722	0.1860	0.2171
24	0.0000	0.0000	0.0000	0.0029	0.0236	--------

TABLE 3. Optimal fuzzy least squared models, in terms of the amount of vagueness in y_i

Amount of vagueness in y_i	Models
$s_i = 0.05y_i$	$\tilde{y}$ = (0.8265,0.0413) + (6.5021,0.3451)x
$s_i = 0.10y_i$	$\tilde{y}$ = (0.8265,0.0827) + (6.5021,0.6902)x
$s_i = 0.15y_i$	$\tilde{y}$ = (0.8265,0.1239) + (6.5021,1.0353)x
$s_i = 0.20y_i$	$\tilde{y}$ = (0.8265,0.1654) + (6.5021,1.3804)x
$s_i = 0.25y_i$	$\tilde{y}$ = (0.8265,0.2068) + (6.5021,1.7255)x

3.1.3 Investigating outliers

Among 24 data points, there are still few data points with small I index for optimal selected model (column 6 in Table 2). They can be regarded as possible outliers. The values of index d for data points with numbers 5, 10, and 24 are very small and close to zero. To investigate the effects of possible outliers on model performance, the three data points were removed and a new model, similar to that presented in Subsection 3.1.1, based on s_i=0.10y_i was fitted as:

$$\tilde{y} = \tilde{A}_0 + \tilde{A}_1 x = (0.5535,0.0553) + (7.6188,0.7919)x \ .$$

In addition, for s_i=0.25y_i, the new model is

$$\tilde{y} = (0.5535,0.1384) + (7.6188,1.9047)x \ .$$

The results of reliability analysis for s_i=0.25y_i are given in column 7 of Table 2. As seen in this table, removing outliers results in improving the performance of the model. Particularly, the average value for I index increased from 0.54 of the original model to 0.60 after removing outliers.

3.2 Pedomodel of CEC-OM-SAND

The second model provides relationship between cation exchange capacity (CEC), as a function of two soil variables including percentage of sand content (SAND), and organic matter content (OM). In the soil, organic matter can enhance the CEC, while the sand content has negative relation with cation exchange capacity [2,5].

3.2.1 Model parameter estimation

After fuzzification of observations, with s_i =0.10y_i, the system of equations in matrix form was constructed as follow:

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ \vdots & \vdots & \vdots \\ 1 & x_{m1} & x_{m2} \end{pmatrix} = \begin{pmatrix} 1 & 35 & 0.88 \\ 1 & 37 & 1.13 \\ \vdots & \vdots & \vdots \\ 1 & 79 & 0.36 \end{pmatrix}_{25\times 3}$$

$$A = X'X = \begin{pmatrix} 25 & 761 & 37.24 \\ 761 & 27913 & 1065.58 \\ 37.24 & 1065.58 & 64.67 \end{pmatrix}$$

$$a=(a_0,a_1,a_2)^T\ ,y=\left(\sum_{i=1}^{25}y_ix_{i0},\sum y_ix_{i1},\sum y_ix_{i2}\right)^T=(472.5,13159.2,741.70)^T,$$

$$\sigma=(\sigma_0,\sigma_1,\sigma_2)^T\ ,s=\left(\sum_{i=1}^{25}s_ix_{i0},\sum s_ix_{i1},\sum s_ix_{i2}\right)^T=(47.25,1315.92,74.17)^T.$$

In the current case study, Rank(X)=3, thus according to Proposition 2.5, the unique solution was found for system of Aa=y, Aσ=s as follows:

$$a=A^{-1}y=(22.9810,-0.2223,2.4741)^T,$$
$$\sigma=A^{-1}s=(2.1982,-0.0222,0.2474)^T.$$

However, in this case, there is no justification for Proposition 2.7, since $A^{-1}s\geq 0$. In fact, the negative spread of σ_1=-0.0222 for fuzzy parameter of x_1 cannot be accepted. To present a solution to this problem, according to procedure proposed in Section 2.4, a new model was fitted to data considering σ_1=0. To determine values of σ_0 and σ_2, the relation of $\sigma^*=A^{*-1}s^*$ used with:

$$s^*=\left(\sum_{i=1}^{25}s_ix_{i0},\ \sum_{i=1}^{25}s_ix_{i2}\right)^T=(47.25,1315.92)^T\ ,$$
$$\sigma^*=(\sigma_0,\sigma_2)^T\ ,$$
$$A^*=\begin{pmatrix}\sum x_{i0}x_{i0} & \sum x_{i2}x_{i0}\\ \sum x_{i0}x_{i2} & \sum x_{i2}x_{i2}\end{pmatrix}=\begin{pmatrix}25 & 37.24\\ 37.24 & 64.67\end{pmatrix}.$$

In such circumstances, the condition $\sigma^*=A^{*-1}s^*\geq 0$ is fulfilled, thus Proposition 2.7 can be applied. Based on $s_i=0.10y_i$, the following results obtained:

$$a=A^{-1}y=(22.9810,-0.2223,2.4741)^T,$$
$$\sigma^*=A^{*-1}s^*=(1.2766,0.4118)^T.$$

Consequently, the optimal model was obtained:

$$\tilde{y}=\tilde{A}_0^{\,*}+\tilde{A}_1^{\,*}x_1+\tilde{A}_2^{\,*}x_2=(22.981,1.2766)-0.2223x_1+(2.4741,0.4118)x_2.$$

3.2.2 Reliability and sensitivity analysis

According to the different amount of vagueness, optimal models were obtained. The results are given in Table 4. Furthermore, the goodness of fit index, $I(\tilde{y}_i,\tilde{Y}_i)$, was calculated for different models. It is shown in Table 5.

TABLE 4. Optimal fuzzy least squared models, in terms of the amount of vagueness in y_i

Amount of vagueness in y_i	Models
$s_i = 0.05y_i$	$\tilde{y} = (22.9810,0.6383) - 0.2223x_1 + (2.4741,0.2059)x_2$
$s_i = 0.10y_i$	$\tilde{y} = (22.9810,1.2766) - 0.2223x_1 + (2.4741,0.4118)x_2$
$s_i = 0.15y_i$	$\tilde{y} = (22.9810,1.9148) - 0.2223x_1 + (2.4741,0.6177)x_2$
$s_i = 0.20y_i$	$\tilde{y} = (22.9810,2.5561) - 0.2223x_1 + (2.4741,0.8236)x_2$
$s_i = 0.25y_i$	$\tilde{y} = (22.9810,3.1914) - 0.2223x_1 + (2.4741,1.0295)x_2$

3.2.3 Investigating outliers

Among 25 data points, there are still few data with small I index for optimally selected model, column 3 in Table 5. They can be regarded as possible outliers. The values of index I for data with numbers 17, 22, 23, and 24 are small. To investigate the effects of possible outliers on model performance, those four data points were removed and new model, with $s_i = 0.10y_i$, was fitted as:

$$\tilde{y} = \tilde{A}_0 + \tilde{A}_1 x = (22.0176,1.3176) - 0.2159x_1 + (2.4650,0.4088)x_2 \ .$$

The results of reliability analysis are given in column 7 of Table 5. As seen in this table, removing outliers results in the improvement of the performance of the model. For half of the cases, the value of I index increases, while the increase is not substantial for other cases. On average, the I index increased from 0.85, for the original model, to 0.93 after outlier rejection.

TABLE 5. Goodness of fit index, I, for different amount of fuzzifications in y_i's.

No. of obs.	$s_i=0.05y_i$	$s_i=0.10y_i$	$s_i=0.15y_i$	$s_i=0.20y_i$	$s_i=0.25y_i$	$s_i=0.10y_i$ (After removing outliers)
1	0.9945	0.9986	0.9994	0.9997	0.9998	0.9985
2	0.2744	0.7238	0.8662	0.9224	0.9496	0.7862
3	0.9982	0.9995	0.9998	0.9999	0.9999	0.9990
4	0.9958	0.9990	0.9995	0.9997	0.9998	0.9934
5	0.5911	0.8768	0.9433	0.9677	0.9792	0.9243
6	0.2276	0.6907	0.8484	0.9116	0.9425	0.7443
7	1.0000	1.0000	1.0000	1.0000	1.0000	0.9972
8	0.9063	0.9757	0.9891	0.9939	0.9961	0.9855
9	0.5311	0.8537	0.9321	0.9612	0.9750	0.8014
10	0.2236	0.6876	0.8467	0.9106	0.9418	0.6500
11	0.9104	0.9768	0.9896	0.9942	0.9963	0.9694
12	0.9272	0.9813	0.9916	0.9953	0.9970	0.9888
13	0.9473	0.9866	0.9940	0.9966	0.9978	0.9941
14	0.8325	0.9552	0.9798	0.9886	0.9927	0.9733
15	0.6749	0.9064	0.9573	0.9757	0.9844	0.9238
16	0.8056	0.9474	0.9763	0.9866	0.9914	0.9346
17	0.0833	0.5373	0.7587	0.8561	0.9054	--------
18	0.8233	0.9525	0.9786	0.9879	0.9923	0.9268
19	0.9615	0.9902	0.9956	0.9975	0.9984	0.9975
20	0.8996	0.9739	0.9883	0.9934	0.9958	0.9889
21	0.9679	0.9919	0.9964	0.9980	0.9987	0.9878

22	0.0203	0.3773	0.6484	0.7837	0.8556	--------
23	0.0533	0.4805	0.7220	0.8326	0.8893	--------
24	0.0600	0.4949	0.7316	0.8388	0.8936	--------
25	0.9999	1.0000	1.0000	1.0000	1.0000	0.9270

3.3 Pedomodel of SP-SILT-SAND-OM

The amount and status of water in soil is described by different constants like SP, which shows soils saturated by water. Although, measuring SP is not expensive but it is time consuming. Both mineral and organic colloids; i.e., percentage of sand and soil organic matter, can increase water capacity of soils [2].

After fuzzification of observations, $s_i=0.10y_i$, the following matrix system was developed:

$$X=\begin{pmatrix}1 & x_{11} & x_{12} & x_{13}\\ 1 & x_{21} & x_{22} & x_{23}\\ \vdots & \vdots & \vdots & \vdots\\ 1 & x_{m1} & x_{m2} & x_{m3}\end{pmatrix}=\begin{pmatrix}1 & 45 & 35 & 0.88\\ 1 & 42 & 37 & 1.13\\ \vdots & \vdots & \vdots & \vdots\\ 1 & 14 & 79 & 0.36\end{pmatrix}_{25\times 4}$$

$$A=X'X=\begin{pmatrix}25 & 1040 & 761 & 37.24\\ 1040 & 44442 & 29685 & 1572.12\\ 761 & 29685 & 27913 & 1065.58\\ 37.24 & 1572.12 & 1065.58 & 64.67\end{pmatrix}$$

$$a=(a_0,a_1,a_2,a_3)^T,\ y=\left(\sum_{i=1}^{25} y_i x_{i0},\sum y_i x_{i1},\sum y_i x_{i2},\sum y_i x_{i3}\right)^T=(1174.6,49677.9,33349.7,1846.31)^T,$$

$$\sigma=(\sigma_0,\sigma_1,\sigma_2,\sigma_3)^T,\ s=\left(\sum_{i=1}^{25} s_i x_{i0},\sum s_i x_{i1},\sum s_i x_{i2},\sum s_i x_{i3}\right)^T=(117.46,4967.79,3334.97,184.63)^T.$$

3.3.1 Estimation of model parameters

Since Rank(X)=4, according to Proposition 2.5, the unique solution was found for system of Aa=y, Aσ=s as follows:

$$a=A^{-1}y=(70.8366,\ -0.4115,\ -0.5751,\ 7.2488)^T,$$

$$\sigma=A^{-1}s=(7.0837,\ -0.0412,\ -0.0575,\ 0.7249)^T$$

Considering σ_2 and σ_3, the condition of $A^{-1}s\geq 0$ is violated. Thus, Proposition 2.7 could not used anymore. Using relation $\sigma^*=A^{*-1}s^*$ results in:

$$a=A^{-1}y=(70.8366,\ -0.4115,\ -0.5751,\ 7.2488)^T$$

$$\sigma^*=A^{*-1}s^*=\begin{pmatrix}25 & 37.42\\ 37.42 & 64.67\end{pmatrix}^{-1}\begin{pmatrix}117.46\\ 184.63\end{pmatrix}=(3.1403,1.0476)^T.$$

Consequently, the optimal model, based on $s_i=0.10y_i$, is:

$$\begin{aligned}\tilde{y} &= \tilde{A}_0^{\,*} + \tilde{A}_1^{\,*} x_1 + \tilde{A}_2^{\,*} x_2 + \tilde{A}_3^{\,*} x_3 \\ &= (70.8366,3.1403)_T - 0.4115x_1 - 0.5751x_2 + (7.2488,1.0476)_T\, x_3 \ .\end{aligned}$$

3.3.2 Reliability and sensitivity analysis

Optimal models were obtained in terms of the different amount of vagueness. The results are given in Table 6. Furthermore, the goodness of fit index, $I(\tilde{y}_i, \tilde{Y}_i)$, was calculated for different models, as shown in Table 7.

TABLE 6. Optimal fuzzy least squared models, in terms of the amount of vagueness in y_i

Amount of vagueness in y_i	Optimal fuzzy least quare models
$s_i = 0.05y_i$	$\tilde{y}$ = (70.8366,1.5701) - 0.4115x_1- 0.5751x_2 + (7.2488,0.5238)x_3
$s_i = 0.10y_i$	$\tilde{y}$ = (70.8366,3.1403) - 0.4115x_1- 0.5751x_2 + (7.2488,1.0476)x_3
$s_i = 0.15y_i$	$\tilde{y}$ = (70.8366,4.7104) - 0.4115x_1- 0.5751x_2 + (7.2488,1.5714)x_3
$s_i = 0.20y_i$	$\tilde{y}$ = (70.8366,6.2806) - 0.4115x_1- 0.5751x_2 + (7.2488,2.0952)x_3
$s_i = 0.25y_i$	$\tilde{y}$ = (70.8366,7.8507) - 0.4115x_1- 0.5751x_2 + (7.2488,2.6190)x_3

TABLE 7. Goodness of fit index, I, for different amount of fuzzifications in y_i's.

No. of Obs.	$s_i=0.05y_i$	$s_i=0.10y_i$	$s_i=0.15y_i$	$s_i=0.20y_i$	$s_i=0.25y_i$	$s_i=0.10y_i$ (After removing outliers)
1	0.9794	0.9948	0.9977	0.9987	0.9992	0.9988
2	0.9839	0.9960	0.9982	0.9990	0.9994	0.9933
3	0.9929	0.9982	0.9992	0.9996	0.9997	0.9983
4	0.9848	0.9962	0.9983	0.9990	0.9994	0.9948
5	0.1482	0.6204	0.8088	0.8875	0.9265	--------
6	0.6792	0.9078	0.9579	0.9761	0.9846	0.9012
7	0.5277	0.8523	0.9314	0.9608	0.9748	0.8720
8	0.9935	0.9984	0.9993	0.9996	0.9997	0.9990
9	0.9924	0.9981	0.9992	0.9995	0.9997	0.9983
10	0.9966	0.9991	0.9996	0.9998	0.9999	0.9997
11	0.9922	0.9980	0.9991	0.9995	0.9997	0.9923
12	0.7120	0.9186	0.9630	0.9790	0.9865	0.9328
13	0.9359	0.9836	0.9927	0.9959	0.9974	0.9805
14	0.9871	0.9968	0.9986	0.9992	0.9995	0.9965
15	0.9704	0.9925	0.9967	0.9981	0.9988	0.9947
16	0.9974	0.9993	0.9997	0.9998	0.9999	1.0000
17	0.9817	0.9954	0.9979	0.9988	0.9993	0.9965
18	0.9151	0.9781	0.9902	0.9945	0.9965	0.9791
19	0.9822	0.9955	0.9980	0.9989	0.9993	0.9953
20	0.8497	0.9601	0.9821	0.9899	0.9935	0.9595
21	0.6997	0.9146	0.9611	0.9779	0.9858	0.9129
22	0.2819	0.7287	0.8688	0.9239	0.9506	--------
23	0.9464	0.9863	0.9939	0.9966	0.9978	0.9860
24	0.9778	0.9944	0.9975	0.9986	0.9991	0.9855
25	0.8705	0.9659	0.9847	0.9914	0.9945	0.9356

3.3.3 Investigating outliers

Considering the model of $y_i=0.10y_i$, data points for 5 and 22 show smaller value for index I than other data points (column 3, Table 7). They can be regarded as possible outliers. To investigate the effects of these possible outliers on model performance, those data points were removed and a new model, with $s_i=0.10y_i$, was fitted as:

$$\begin{aligned}\tilde{y} &= \tilde{A}_0 + \tilde{A}_1 x + \tilde{A}_2 x + \tilde{A}_3 x \\ &= (73.6004,3.2071) - 0.4644x_1 - 0.5957x_2 + (7.3025,1.0191)x_3 \ .\end{aligned}$$

The results of reliability analysis after removing outliers, for $s_i=0.10y_i$, are given in column 7 of Table 7. As seen, removing outliers results in improving the performance of the model. Although the values of I index increased for about half of data, however it no obvious changes were observed for other half of data. This is mainly due to optimal model fitting performed on the original data. On average, the value of I index increased from 0.95 to 0.97 after removing two outliers.

4. Discussion and conclusion

Fuzzy least squares regression has been described in terms of practical soil science examples. We aimed to illustrate that fuzzy regression technique could be a useful and promising scientific tool for environmental scientists to describe relationships between variables when imprecise and vague observations and/or relations would preclude the use of classical regression analysis. We applied a least squares approach to derive fuzzy pedotransfer models, based on vague data. During pedomodels fitting, the problem of encountering negative width of fuzzy parameters may violate attempts for modeling. We proposed an approach for solving this problem.

The regression parameters could be calculated by minimizing the distance between two fuzzy numbers. In fact, a pedomodel can be fitted by minimizing the sum of square of the deviations between the observed fuzzy values (on dependent variable) and their predicted fuzzy values. The quality of the obtained model can be evaluated using goodness of fit criterion between the observed values and the estimated values.

In order to fuzzify the values of dependent variable, soil scientists experience should be used. It is shown that excessive vagueness in the regression model may indicate a poor choice of underlying relationship. Although, it may be a general tendency not to use fuzzy regression when the data quality is poor, but we showed that the problem of outliers can be handled through fuzzy regression. The index of goodness of fit can be used as a criterion to determine the possible outliers.

As a future research direction, it would be desirable to consider vagueness of not only observations on dependent variable but exploratory variables. Although the resulting pedomodels are best suited for incorporating in decision making process, however, the way of interpretation of results and combining with deterministic and stochastic pedomodels are still under question.

Acknowledgments. This research was sponsored by a grant from Isfahan University of Technology and Shahrekord University. The authors are grateful to M. Nasiri for her invaluable assistance with some computational calculations.

REFERENCES

[1] J. Bouma, Using soil survey data for qualitative land evaluation. In B.A. Stewart (Editor), *Advances in Soil Sciences*, Vol. 9. Springer-Verlag, New York, (1989) 177-213.
[2] R. L. Donahue, R. W. Miller, and J. C. Shickluna, *Soils, an introduction to soils and plant growth*. Prentice-Hall, (1983).
[3] K. J. Kim, H. Moskowitz, and M. Koksalan, *Fuzzy versus statistical linear regression*, Euro. J. Oper. Res., **92** (1996) 417-434.
[4] B. Minasny, and A. B. McBratney, *The neuro-m method for fitting neural network parametric pedotransfer functions,* Soil Sci. Soc. Am. J., **66** (2002) 352-361.
[5] A. L. Page, R. H. Miller, and D. R. Keeney, *Methods of soil analysis*, *Part 2*, Soil Science Society of Ameriac, Madison, Wisconsin, (1982).
[6] B. Sadeghpour, and D. Gien, *A goodness of fit index to reliability analysis in fuzzy model*. In A. Grmela (Editor), Advances in Intelligent Systems, Fuzzy Systems, Evolutionary Computation, WSEAS Press, Greece, (2002).
[7] E. Salchow, R. Lal, N. R. Fausey, and A. Ward, *Pedotransfer functions for variable alluvial soils in southern Ohio*, Geoderma, **73** (1996) 165-181.
[8] S. M. Taheri, *Trends in fuzzy statistics*, Austrian J. Stat., **32** (2003) 239-257.
[9] P. Wang, *Fuzzy sets and its applications*, Publishing House of Shanghai Science and Technology, Shanghai, (1983).
[10] R. Xu, *A linear regression model in fuzzy environment*, Adv. Modelling Simulation, **27** (1991) 31-40.
[11] R. Xu, and C. Li , *Multidimensional least-squares fitting with fuzzy model*, Fuzzy Sets and Systems, **119** (2001) 215-223.
[12] H. J. Zimmermann, *Fuzzy set theory and its applications*, Kluwer Academic, Boston, (1991).

JAHANGARD MOHAMMADI, SOIL SCIENCE DEPARTMENT, COLLEGE OF AGRICULTURE, SHAHREKORD UNIVERSITY, SHAHREKORD, IRAN.
E-mail address: **j_mohammadi@sku.ac.ir**

SYED MAHMOUD TAHERI *, SCHOOL OF MATHEMATICAL SCIENCES, ISFAHAN, UNIVERSITY OF TECHNOLOGY, ISFAHAN 84156, IRAN.
E-mail address: **Taheri@cc.iut.ac.ir**

* CORRESPONDING AUTHOR