EasyChair Preprint

№ 7774

Fuzzy Clustering to Study Various Regression Models with Fractional Brownian Motion Errors

Mohammad Reza Mahmoudi, Mohammad Hossein Heydari and Kim-Hung Pho

April 12, 2022

Fuzzy Clustering to Study Various Regression Models with Fractional Brownian Motion Errors

Mohammad Reza Mahmoudi [1, 2], Mohammad Hossein Heydari [3], and Kim-Hung Pho[4]

[1] Institute of Research and Development, Duy Tan University, Da Nang 550000, Vietnam

[2] Department of Statistics, Faculty of Science, Fasa University, Fasa, Fars, Iran

[3] Department of Mathematics, Faculty of Science, Shiraz University of Technology

[4] Fractional Calculus, Optimization and Algebra Research Group, Faculty of Mathematics and Statistics, Ton Duc Thang University, Ho Chi Minh City, Vietnam

Abstract.

Background: Clustering numerous regression models fitted on the dataset is one of the most ubiquitous issues in different fields of sciences. This research aims to compare and to classify different regression models with fractional Brownian motion errors which can be used for a dataset. Our primary objective is to cluster these models based on fuzzy clustering and then to detect a subset of inexpensive predictors to predict a response variable reasonably well.

Results: The results indicate that the power of our proposed approach is very close to the one, specially when the sample size increases. In other hands, the power studies show the excellent effectiveness of our proposed approach.

Conclusion: In this research, Fuzzy clustering method is used to cluster regression models with fractional Brownian motion errors that can be fitted on a dataset. Thereafter the performance of proposed approach is studied in simulated and real situations. The results verify that the introduced technique had excellent power to cluster the models. It indicates that our proposed method obtain many advantages. The performance of proposed technique is allowable. In addition, the algorithm is not so complicated. Furthermore, this method can be employed to compare many models (both linear models and nonlinear models).

Keywords: Fuzzy Clustering; Fractional Brownian Motion, Data Modeling; Regression.

1. Background

Regression models are often applied to explore the relationships between a quantitative dependent (response) variable and one or more independent (or explanatory or predictor) variables. It will be known that one can apply different regression models on a dataset. In numerous disciplines, e.g., biology, climatology, economic, electronic, finance, hydrology, and management studies, we can fit the models that were performed in another datasets. For instance, they adjusted the previous models for their datasets and would like to compare these different models. The authors performed the values of R-square (R^2) and root mean square error (RMSE) to compare and rank models and choose the model with the largest R^2 or the model with the smallest RMSE as the best. For instance, in biological and environmental research, Bahrami et al. [1] executed different kinetic and isotherm models to adapt the experimental data of caffeine removal using multi-wall Carbon nanotubes. Also, in agricultural and hydrological studies, Zarei and Mahmoudi [2] studied the changes in RDIst index affected by different potential evapotranspiration (PET) calculation methods. In addition, Zarei et al. [3-4] investigated the changes in spatial sample and trend of drought using different autoregressive models. Although the ranking is true, it is not the only thing of interest. We are often interested in the model with few parameters and the smallest costs. In other words, the researchers are looking for the best choice. Hence if the models are statistically equal, then one can adapt the best model with few parameters and the smallest cost. For instance, investigate the following situation.

Stachys pilifera is an endemic plant in Iran. It is a Perennial plant that belongs to the Lamiaceae. Its distribution is related to several weather factors such as Rain, Temperature and Evaporation and several soil properties such as EC, OC, Silt, N, Fe, Zn, and Cu. The dataset contained the values of the distribution of Stachys pilifera, the weather factors and the soil properties for 25 samples (fields). We can model and predict the distribution of Stachys pilifera based on each of these 10 variables using a set of linear regression models. The estimated parameters of the linear regression models are presented in Table 1. These factors can determine from 16.2 to 66.22 percent of variation in distribution of Stachys pilifera. From Table 1, it can be observed that the factors Rain, Temperature, EC, Cu and Evaporation have positive effects on distribution of Stachys pilifera (coefficients are positive and p-values are less than 0.05). Also, other factors have negative effects on distribution of Stachys pilifera (coefficients are negative and p-values are less than 0.05). It can be seen that the model based on Fe has the best fitness to the distribution of Stachys pilifera, based on higher R^2 and less RMSE.

Table 1. Estimated parameters of the linear regression models

Factor	Coefficient	Standard Coefficient	P-Value (p)	R^2	*RMSE*
Rain	0.060	0.620	0.001	0.385	5.385
Temperature	0.829	0.507	0.010	0.257	5.915
EC	5.410	0.402	0.046	0.162	6.285
OC	-5.689	-0.513	0.009	0.264	5.891
Silt	-0.963	-0.530	0.006	0.281	5.822
N	-20.496	-0.781	<0.001	0.610	4.284
Fe	-1.354	-0.813	<0.001	0.662	3.993
Zn	-10.436	-0.740	<0.001	0.548	4.615
Cu	12.434	0.448	0.025	0.201	6.136
Evaporation	0.027	0.596	0.002	0.355	5.313

Therefore if we classify these 10 models, then the models in each cluster are statistically equal, and one can execute the model with the smallest cost to presage and simulate the distribution of Stachys pilifera.

All of previous works are about comparing two or several regression models. The references [5-9] developed procedures to compare the correlation of X and Y in the two populations. The references [6, 10-14] presented some approaches to compare the relationship of X and Y with the relationship of X and W. The references [12, 15-16] discussed techniques to compare the correlation of X and Y with the correlation of W and Z.

This article aims to cluster numerous regression models with fractional Brownian motion errors which may be adapted on a dataset. Our primary objective is to cluster these models and then to detect a subset of inexpensive predictors to predict a response variable reasonably well. In addition, the fuzzy clustering method will be applied to cluster the considered models.

2. Methods

Let $X = (X_1, \dots, X_k)$ and Y be the k-dimensional contingent predictors and response variable, respectively. Also suppose $B_H(t)$ is a fractional Brownian motion with Hurst index $H \in (0,1)$, defined by

$$B_H(t) = \frac{1}{\Gamma\left(H + \frac{1}{2}\right)} \int_0^t (t-s)^{H-\frac{1}{2}} dB(s),$$

such that B and Γ are respectively Brownian motion process and gamma function, and integration is with respect to the white noise measure $dB(s)$. It should be noted that $B_H(t)$ and $B_H(s)$ are zero-mean processes with auto-covariance function

$$\gamma(s,t) \coloneqq Cov\big(B_H(s), B_H(t)\big) = \frac{1}{2}\left(|t|^{2H} + |s|^{2H} - |t-s|^{2H}\right).$$

The case $H = \frac{1}{2}$, is Wiener process or standard Brownian motion. The cases $H > \frac{1}{2}$ and $H < \frac{1}{2}$ indicate the positive and negative autocorrelation between increments, respectively.

In this research, the structures of *m* linear or nonlinear regression models are expressed as follows.

$$Y = f_i(X) + B_{H_i}, \quad i = 1, \dots, m, \tag{1}$$

such that $B_{H_i}, i = 1, \dots, m,$ are independent fractional Brownian motion errors, and $f_i, i = 1, \dots, m,$ are functions with unknown parameters. In case of *m = 2,* one may have the two regression models

$$Y = \beta_0 + \beta_1 X + B_{H_1},$$

and

$$Y = \beta_0 + \beta_1 e^{\beta_2 X} + B_{H_2}.$$

Assume that there are *n* observations from (X, Y). For these observations, the equations of the regression models can be represented by

$$\boldsymbol{Y} = \boldsymbol{f}_i(\boldsymbol{X}) + \boldsymbol{B}_{H_i}, \quad i = 1, \dots, m, \tag{2}$$

Let $\boldsymbol{Y} = (y_1, \dots, y_n)^T$ as the values of response variable Y, $\boldsymbol{x} = [\boldsymbol{x}_1, \dots, \boldsymbol{x}_n], \boldsymbol{x}_j = (x_{j1}, \dots, x_{jn})^T$, as to the values of the predictors X, and $\boldsymbol{B}_{H_i} = (B_{H_{i1}}, \dots, B_{H_{i}n})^T, i = 1, \dots, m,$ as m independent fractional Brownian motion processes.

Now, for each regression model, the corresponding equation can be estimated by

$$\widehat{\boldsymbol{Y}}_i = \widehat{\boldsymbol{f}}_i(\boldsymbol{x}), \quad i = 1, \dots, m,$$

where $\widehat{\boldsymbol{Y}}_i = (\hat{y}_{i1}, \dots, \hat{y}_{in})^T, i = 1, \dots, m,$ are estimators of Y**.** Because $B_{H_i}, i = 1, \dots, m,$ are zero-mean processes, thus $\hat{y}_{i1}, \dots, \hat{y}_{in}, i = 1, \dots, m,$ are unbiased estimators of $f_i(x), i = 1, \dots, m,$ respectively.

Remark 1: In estimation procedure, for linear and nonlinear regression models, the least squares approach and the Levenberg-Marquardt algorithm were respectively employed to compute $\widehat{\boldsymbol{Y}}_i$.

The procedure for fuzzy clustering of $f_1, \dots, f_m$ can be described as follows.

Step (1): The *m* regression models are estimated by

$$\hat{f}_i(\boldsymbol{x}), \quad i = 1, \dots, m.$$

Step (2): The predicted values of Y is simulated based on all *m* models by

$$\widehat{\boldsymbol{Y}}_i = (\hat{y}_{i1}, \dots, \hat{y}_{in})^T, i = 1, \dots, m,$$

Thereafter, we get *m* predicted values datasets, for dataset $y_1, \dots, y_n$.

Step (3): The fuzzy clustering method [17-18] is applied by using the couples $(\hat{y}_{1,1}, \dots, \hat{y}_{1,n}), \dots, (\hat{y}_{m,1}, \dots, \hat{y}_{m,n})$.

3. Results

This section is divided into two parts. In the first subsection, the ability of the proposed method is studied based on different simulated datasets. A case study is also given in the second subsection.

3.1. Numerical Results

In this subsection, several data sets were drawn to study the performance of our proposed approach. The simulations are performed by using the *R 3.6.1* software. The number of repetitions in this simalation is 1000.

The simulation procedure can be expressed as follows.

Step 1: For each parameter setting (model), we separately generated a sample of size n. In other words, for each model, a sample of size n was provided.

Step 2: For each sample, the corresponding model was fitted on generated dataset.

Step 3: The values of $\widehat{\boldsymbol{Y}}_i = (\hat{y}_{i1}, \dots, \hat{y}_{in})^T, i = 1,2,3,$ were computed for each model.

Step 4: The fuzzy clustering method was applied to cluster the values of $\widehat{\boldsymbol{Y}}_i, i = 1,2,3$**.**

Step 5: Previous steps were repeated 1000 times.

Step 6: The estimated power of the method ($\hat{\pi}$) was computed by

$$\hat{\pi} = \frac{T}{1000},$$

where T is the number of the runs for which the proposed method can correctly cluster the models.

Remark 2: When the parameter settings of three models are similar, the number of clusters is equal to 1 (all models are in 1 cluster). As the parameter settings of two models are analogous and another model has different parameter setting, the number of clusters is equal to 2 (two models are in cluster 1 and other model in cluster 2). When the parameter settings of three models are various, the number of clusters is equal to 3 (each model has distinct cluster).

Example 1: Consider the homoscedastic error model given by:

$$Y = \beta X + B_H.$$

For the random variable X and the Hurst parameter H, we consider the cases $X \sim Normal\ (0,0.25)$, $X \sim Exponential\ (5)$, and $H \in \{0.25, 0.75\}$. In this example, we use $\beta = 1, \beta \in \{1, 2\}$, and $\beta \in \{1, 2, 3\}$, for the three corresponding models.

Example 2: Consider the heteroscedastic error model given by:

$$Y = \beta_0 + \beta_1 X + \beta_2 X B_H.$$

For the random variable X and the Hurst parameter H, we consider the cases $X \sim Normal\ (0,1)$, $X \sim Exponential\ (1)$, and $H \in \{0.25, 0.75\}$. For these models, we utilize $(\beta_0, \beta_1, \beta_2) = (2,1,2)$, $(\beta_0, \beta_1, \beta_2) \in \{(2,1,2),\ (0,2,1)\}$, and $(\beta_0, \beta_1, \beta_2) \in \{(2,1,2),$ $(0,2,1), (3,2,1)\}$, for the three corresponding models.

Example 3: Consider the model with discrete covariate given by:

$$Y = 1 + \beta X + B_H.$$

For the random variable X and the Hurst parameter H, we consider the cases $X \sim \text{Geometric}\ (0.4)$, $X \sim Binomial\ (2, 0.7)$ and $H \in \{0.25, 0.75\}$. In this example, we assume that $\beta = 1$, and $\beta \in \{1, 2\}$, and $\beta \in \{1, 2{,}5\}$, for the three corresponding models.

Example 4: Consider the multiple linear regression model given by:

$$Y = \beta_0 + \beta_1 X_1 + 2\beta_2 X_2 + B_H.$$

For the random variables X_1 and X_2 and the Hurst parameter H, we consider the cases $X_1 \sim Uniform(0{,}2)$, $X_2 \sim Exponential(5)$, $X_2 \sim \text{Geometric}\ (0.3)$ and $H \in \{0.25, 0.75\}$. In this example, we assume that $(\beta_0, \beta_1, \beta_2) = (2{,}1{,}2)$, $(\beta_0, \beta_1, \beta_2) \in \{(2{,}1{,}2),\ (0{,}2{,}1)\}$, and $(\beta_0, \beta_1, \beta_2) \in \{(2{,}1{,}2), (0{,}2{,}1), (3{,}2{,}1)\}$, for the three corresponding models.

Example 5: Consider the simple nonlinear regression model given by:

$$Y = e^X + B_H.$$

For the random variable X and the Hurst parameter H, we consider the cases $X \sim Normal(0{,}0.5)$, $X \sim Poisson(5)$ and $H \in \{0.25, 0.75\}$. In this example, we use $Y = e^X + B_H$, $Y = \{e^X + B_H, 1 + \beta X + B_H\}$, and $Y \in \{e^X + B_H, 1 + \beta X + B_H, 2X + B_H\}$, for the three corresponding models.

The values of $\hat{\pi}$ for Examples 1-5, can be respectively observed in Tables 2-6. It can be concluded that the power of our proposed approach is very close to the one, specially when the value of n increases. In other words, the results show the excellent effectiveness of our proposed approach. It can be seen that our proposed method obtain many advantages. The performance of proposed technique is allowable. In addition, the algorithm is not so complicated. Furthermore, this approach can be employed to compare many models (both linear models and nonlinear models).

Table 2. Estimated power of method for Example 1

H	X	β			Number of Clusters	n			
		First model	Second model	Third model		20	50	75	100
0.25	$Normal$ (0,0.25)	1	1	1	1	0.952	0.954	0.963	0.992
0.25	$Exponential$ (5)	1	1	1	1	0.941	0.967	0.982	0.991
0.75	$Normal$ (0,0.25)	1	1	1	1	0.954	0.969	0.974	0.986
0.75	$Exponential$ (5)	1	1	1	1	0.946	0.966	0.971	1.000
0.25	$Normal$ (0,0.25)	1	1	2	2	0.952	0.954	0.969	0.992
0.25	$Exponential$ (5)	1	1	2	2	0.947	0.963	0.969	0.976
0.75	$Normal$ (0,0.25)	1	1	2	2	0.969	0.965	0.967	0.979
0.75	$Exponential$ (5)	1	1	2	2	0.946	0.960	0.966	0.975
0.25	$Normal$ (0,0.25)	1	1	3	2	0.952	0.963	0.979	0.983
0.25	$Exponential$ (5)	1	1	3	2	0.948	0.956	0.975	0.991
0.75	$Normal$ (0,0.25)	1	1	3	2	0.950	0.954	0.971	0.991
0.75	$Exponential$ (5)	1	1	3	2	0.953	0.970	0.979	0.994
0.25	$Normal$ (0,0.25)	1	2	1	2	0.953	0.970	0.978	1.000
0.25	$Exponential$ (5)	1	2	1	2	0.958	0.959	0.973	0.984
0.75	$Normal$ (0,0.25)	1	2	1	2	0.951	0.968	0.983	0.983
0.75	$Exponential$ (5)	1	2	1	2	0.956	0.954	0.972	1.000
0.25	$Normal$ (0,0.25)	1	2	2	3	0.949	0.959	0.982	0.986
0.25	$Exponential$ (5)	1	2	2	3	0.937	0.962	0.985	0.974
0.75	$Normal$ (0,0.25)	1	2	2	3	0.941	0.957	0.968	0.987
0.75	$Exponential$ (5)	1	2	2	3	0.938	0.959	0.962	0.984
0.25	$Normal$ (0,0.25)	1	2	3	3	0.947	0.953	0.975	0.985
0.25	$Exponential$ (5)	1	2	3	3	0.957	0.967	0.972	1.000
0.75	$Normal$ (0,0.25)	1	2	3	3	0.954	0.963	0.968	0.986
0.75	$Exponential$ (5)	1	2	3	3	0.961	0.975	0.986	0.977

Table 3. Estimated power of method for Example 2

H	X	$(\beta_0, \beta_1, \beta_2)$			Number of Clusters	n			
		First model	Second model	Third model		20	50	75	100
0.25	$Normal$ (0,1)	(2,1,2)	(2,1,2)	(2,1,2)	1	0.946	0.964	0.981	0.985
0.25	$Exponential$ (1)	(2,1,2)	(2,1,2)	(2,1,2)	1	0.949	0.963	0.965	0.974
0.75	$Normal$ (0,1)	(2,1,2)	(2,1,2)	(2,1,2)	1	0.953	0.960	0.979	0.996
0.75	$Exponential$ (1)	(2,1,2)	(2,1,2)	(2,1,2)	1	0.954	0.965	0.976	0.987
0.25	$Normal$ (0,1)	(2,1,2)	(2,1,2)	(0,2,1)	2	0.965	0.971	0.964	0.979
0.25	$Exponential$ (1)	(2,1,2)	(2,1,2)	(0,2,1)	2	0.944	0.967	0.961	0.993
0.75	$Normal$ (0,1)	(2,1,2)	(2,1,2)	(0,2,1)	2	0.955	0.964	0.968	0.986

0.75	*Exponential* (1)	(2,1,2)	(2,1,2)	(0,2,1)	2	0.953	0.955	0.974	0.989
0.25	*Normal* (0,1)	(2,1,2)	(2,1,2)	(3,2,1)	2	0.933	0.972	0.977	0.988
0.25	*Exponential* (1)	(2,1,2)	(2,1,2)	(3,2,1)	2	0.941	0.973	0.979	0.976
0.75	*Normal* (0,1)	(2,1,2)	(2,1,2)	(3,2,1)	2	0.936	0.970	0.984	1.000
0.75	*Exponential* (1)	(2,1,2)	(2,1,2)	(3,2,1)	2	0.961	0.963	0.979	1.000
0.25	*Normal* (0,1)	(2,1,2)	(0,2,1)	(2,1,2)	2	0.948	0.966	0.982	0.980
0.25	*Exponential* (1)	(2,1,2)	(0,2,1)	(2,1,2)	2	0.933	0.961	0.976	0.997
0.75	*Normal* (0,1)	(2,1,2)	(0,2,1)	(2,1,2)	2	0.951	0.967	0.978	0.976
0.75	*Exponential* (1)	(2,1,2)	(0,2,1)	(2,1,2)	2	0.954	0.954	0.973	0.985
0.25	*Normal* (0,1)	(2,1,2)	(0,2,1)	(0,2,1)	3	0.935	0.968	0.980	0.980
0.25	*Exponential* (1)	(2,1,2)	(0,2,1)	(0,2,1)	3	0.958	0.978	0.973	0.996
0.75	*Normal* (0,1)	(2,1,2)	(0,2,1)	(0,2,1)	3	0.938	0.964	0.973	0.980
0.75	*Exponential* (1)	(2,1,2)	(0,2,1)	(0,2,1)	3	0.954	0.971	0.968	0.993
0.25	*Normal* (0,1)	(2,1,2)	(0,2,1)	(3,2,1)	3	0.936	0.969	0.967	0.986
0.25	*Exponential* (1)	(2,1,2)	(0,2,1)	(3,2,1)	3	0.940	0.967	0.980	0.993
0.75	*Normal* (0,1)	(2,1,2)	(0,2,1)	(3,2,1)	3	0.968	0.963	0.975	0.989
0.75	*Exponential* (1)	(2,1,2)	(0,2,1)	(3,2,1)	3	0.954	0.968	0.972	0.981

Table 4. Estimated power of method for Example 3

H	X	β			Number of Clusters	n			
		First model	Second model	Third model		20	50	75	100
0.25	Geometric (0.4)	1	1	1	1	0.951	0.970	0.986	0.986
0.25	*Binomial* (2, 0.7)	1	1	1	1	0.941	0.954	0.964	0.990
0.75	Geometric (0.4)	1	1	1	1	0.938	0.972	0.975	0.983
0.75	*Binomial* (2, 0.7)	1	1	1	1	0.938	0.953	0.980	0.987
0.25	Geometric (0.4)	1	1	2	2	0.939	0.964	0.983	1.000
0.25	*Binomial* (2, 0.7)	1	1	2	2	0.939	0.965	0.983	0.996
0.75	Geometric (0.4)	1	1	2	2	0.935	0.969	0.967	0.993
0.75	*Binomial* (2, 0.7)	1	1	2	2	0.945	0.956	0.981	0.981
0.25	Geometric (0.4)	1	1	5	2	0.948	0.958	0.976	0.974
0.25	*Binomial* (2, 0.7)	1	1	5	2	0.944	0.958	0.974	0.997
0.75	Geometric (0.4)	1	1	5	2	0.963	0.972	0.968	0.988
0.75	*Binomial* (2, 0.7)	1	1	5	2	0.953	0.973	0.970	0.978
0.25	Geometric (0.4)	1	2	1	2	0.939	0.963	0.983	0.998
0.25	*Binomial* (2, 0.7)	1	2	1	2	0.964	0.972	0.978	1.000
0.75	Geometric (0.4)	1	2	1	2	0.957	0.958	0.977	0.992
0.75	*Binomial* (2, 0.7)	1	2	1	2	0.951	0.972	0.974	0.986
0.25	Geometric (0.4)	1	2	2	3	0.966	0.957	0.983	0.988
0.25	*Binomial* (2, 0.7)	1	2	2	3	0.948	0.957	0.969	0.992
0.75	Geometric (0.4)	1	2	2	3	0.954	0.965	0.985	0.983

0.75	*Binomial* (2, 0.7)	1	2	2	3	0.947	0.968	0.982	0.995
0.25	Geometric (0.4)	1	2	5	3	0.945	0.972	0.978	0.976
0.25	*Binomial* (2, 0.7)	1	2	5	3	0.953	0.972	0.978	0.990
0.75	Geometric (0.4)	1	2	5	3	0.955	0.956	0.979	0.986
0.75	*Binomial* (2, 0.7)	1	2	5	3	0.938	0.959	0.978	0.976

Table 5. Estimated power of method for Example 4

H	X_2	$(\beta_0, \beta_1, \beta_2)$			Number of Clusters	n			
		First model	Second model	Third model		20	50	75	100
0.25	*Exponential* (5)	(2,1,2)	(2,1,2)	(2,1,2)	1	0.944	0.968	0.966	0.992
0.25	Geometric (0.3)	(2,1,2)	(2,1,2)	(2,1,2)	1	0.961	0.956	0.966	0.999
0.75	*Exponential* (5)	(2,1,2)	(2,1,2)	(2,1,2)	1	0.952	0.961	0.972	0.992
0.75	Geometric (0.3)	(2,1,2)	(2,1,2)	(2,1,2)	1	0.940	0.961	0.972	0.985
0.25	*Exponential* (5)	(2,1,2)	(2,1,2)	(0,2,1)	2	0.935	0.960	0.982	0.981
0.25	Geometric (0.3)	(2,1,2)	(2,1,2)	(0,2,1)	2	0.963	0.963	0.986	0.994
0.75	*Exponential* (5)	(2,1,2)	(2,1,2)	(0,2,1)	2	0.934	0.974	0.967	0.987
0.75	Geometric (0.3)	(2,1,2)	(2,1,2)	(0,2,1)	2	0.959	0.973	0.977	0.983
0.25	*Exponential* (5)	(2,1,2)	(2,1,2)	(3,2,1)	2	0.962	0.964	0.969	1.000
0.25	Geometric (0.3)	(2,1,2)	(2,1,2)	(3,2,1)	2	0.940	0.969	0.970	0.987
0.75	*Exponential* (5)	(2,1,2)	(2,1,2)	(3,2,1)	2	0.956	0.961	0.976	0.994
0.75	Geometric (0.3)	(2,1,2)	(2,1,2)	(3,2,1)	2	0.953	0.968	0.970	0.997
0.25	*Exponential* (5)	(2,1,2)	(0,2,1)	(2,1,2)	2	0.959	0.961	0.977	0.986
0.25	Geometric (0.3)	(2,1,2)	(0,2,1)	(2,1,2)	2	0.962	0.974	0.987	0.984
0.75	*Exponential* (5)	(2,1,2)	(0,2,1)	(2,1,2)	2	0.931	0.953	0.981	0.989
0.75	Geometric (0.3)	(2,1,2)	(0,2,1)	(2,1,2)	2	0.954	0.958	0.968	0.996
0.25	*Exponential* (5)	(2,1,2)	(0,2,1)	(0,2,1)	3	0.952	0.965	0.981	0.988
0.25	Geometric (0.3)	(2,1,2)	(0,2,1)	(0,2,1)	3	0.945	0.979	0.980	0.986
0.75	*Exponential* (5)	(2,1,2)	(0,2,1)	(0,2,1)	3	0.967	0.953	0.988	0.979
0.75	Geometric (0.3)	(2,1,2)	(0,2,1)	(0,2,1)	3	0.951	0.958	0.980	1.000
0.25	*Exponential* (5)	(2,1,2)	(0,2,1)	(3,2,1)	3	0.938	0.952	0.984	0.986
0.25	Geometric (0.3)	(2,1,2)	(0,2,1)	(3,2,1)	3	0.957	0.961	0.965	0.986
0.75	*Exponential* (5)	(2,1,2)	(0,2,1)	(3,2,1)	3	0.960	0.971	0.972	0.985
0.75	Geometric (0.3)	(2,1,2)	(0,2,1)	(3,2,1)	3	0.948	0.964	0.971	0.992

Table 6. Estimated power of method for Example 5

H	X	Y			Number of Clusters	n			
		First model	Second model	Third model		20	50	75	100

0.25	$Normal\ (0,0.5)$	e^X	e^X	e^X	1	0.934	0.976	0.972	0.983
0.25	$Poisson\ (5)$	e^X	e^X	e^X	1	0.938	0.963	0.970	0.998
0.75	$Normal\ (0,0.5)$	e^X	e^X	e^X	1	0.959	0.962	0.977	0.986
0.75	$Poisson\ (5)$	e^X	e^X	e^X	1	0.948	0.954	0.963	0.998
0.25	$Normal\ (0,0.5)$	e^X	e^X	$1+\beta X$	2	0.947	0.960	0.980	0.979
0.25	$Poisson\ (5)$	e^X	e^X	$1+\beta X$	2	0.950	0.956	0.970	0.995
0.75	$Normal\ (0,0.5)$	e^X	e^X	$1+\beta X$	2	0.937	0.957	0.964	0.990
0.75	$Poisson\ (5)$	e^X	e^X	$1+\beta X$	2	0.958	0.960	0.963	0.991
0.25	$Normal\ (0,0.5)$	e^X	e^X	$2X$	2	0.947	0.963	0.987	1.000
0.25	$Poisson\ (5)$	e^X	e^X	$2X$	2	0.962	0.966	0.967	0.996
0.75	$Normal\ (0,0.5)$	e^X	e^X	$2X$	2	0.946	0.977	0.974	0.989
0.75	$Poisson\ (5)$	e^X	e^X	$2X$	2	0.965	0.971	0.977	0.990
0.25	$Normal\ (0,0.5)$	e^X	e^X	e^X	2	0.949	0.958	0.987	0.996
0.25	$Poisson\ (5)$	e^X	e^X	e^X	2	0.956	0.969	0.980	0.983
0.75	$Normal\ (0,0.5)$	e^X	e^X	e^X	2	0.947	0.962	0.979	0.995
0.75	$Poisson\ (5)$	e^X	e^X	e^X	2	0.953	0.976	0.981	0.986
0.25	$Normal\ (0,0.5)$	e^X	$1+\beta X$	$1+\beta X$	3	0.935	0.969	0.983	0.993
0.25	$Poisson\ (5)$	e^X	$1+\beta X$	$1+\beta X$	3	0.944	0.970	0.982	0.993
0.75	$Normal\ (0,0.5)$	e^X	$1+\beta X$	$1+\beta X$	3	0.963	0.967	0.967	0.991
0.75	$Poisson\ (5)$	e^X	$1+\beta X$	$1+\beta X$	3	0.935	0.962	0.972	0.992
0.25	$Normal\ (0,0.5)$	e^X	$1+\beta X$	$2X$	3	0.948	0.972	0.970	0.990
0.25	$Poisson\ (5)$	e^X	$1+\beta X$	$2X$	3	0.945	0.968	0.975	0.982
0.75	$Normal\ (0,0.5)$	e^X	$1+\beta X$	$2X$	3	0.952	0.961	0.965	0.986
0.75	$Poisson\ (5)$	e^X	$1+\beta X$	$2X$	3	0.964	0.970	0.971	0.998

3.2. Case Study

This subsection is devoted to real world problem (provided in the first section) to study the ability of the proposed approach in real situations. As mentioned above, the distribution of Stachys pilifera can be modeled and predicted based on weather factors such as Rain, Temperature and Evaporation and soil properties consist of EC, OC, Silt, N, Fe, Zn, and Cu. If we cluster these 10 models, then the models in each cluster are not significantly different. Therefore we need to use the model with the smallest cost to presage and simulate the distribution of Stachys pilifera.

We now execute our proposed method to cluster these models. The results for the fuzzy clustering method are provided in Table 7 and Figures 1 and 2. It can be observed that, there are significant differences between these models and they can be clustered in some clusters. Based on Kaiser Index (the number of eigen-values of correlation matrix that are more than 1), the number of clusters is determined to be 3 clusters. Based on Table 7 and Figures 1 and 2, these clusters are as follow:

First cluster: Temperature, OC, Silt, EC and Cu.

Second cluster: Rain and Evaporation.

Third cluster: N, Fe and Zn.

Therefore, in future researches, when an environmental scientist must predict the distribution of Stachys pilifera, based on only one variable of first cluster, he can select the variable that its measuring is simplest and has minimum costs (for example, Temperature). The candidate in second and third clusters can be Rain and N, respectively. It should be noted that in practical cases, first we should cluster the models and select the most powerful cluster based on R^2 and *RMSE* values (Third cluster in this real data example) and then select the inexpensive predictor in this cluster as the final choice (N in this real data example). For future studies investigation of further application and case studies, e.g., [19-41] are suggested to better validate the proposed method.

Table 7. The precents of membership in different clusters based on Fuzzy clustering method to classify the regression models fitted on the distribution of Stachys pilifera

Factor	Cluster 1	Cluster 2	Cluster 3
Rain	0.06446546	**0.8677285**	0.06780603
Temperature	**0.76507691**	0.1751447	0.05977839
EC	**0.74223111**	0.1835854	0.07418353
OC	**0.62210716**	0.2935968	0.08429606
Silt	**0.71223446**	0.2113327	0.07643281
N	0.05497585	0.1231782	**0.82184600**
Fe	0.06552710	0.1807743	**0.75369859**
Zn	0.07368989	0.1861433	**0.74016681**
Cu	**0.62671675**	0.2793840	0.09389929
Evaporation	0.15605870	**0.4833335**	0.36060782

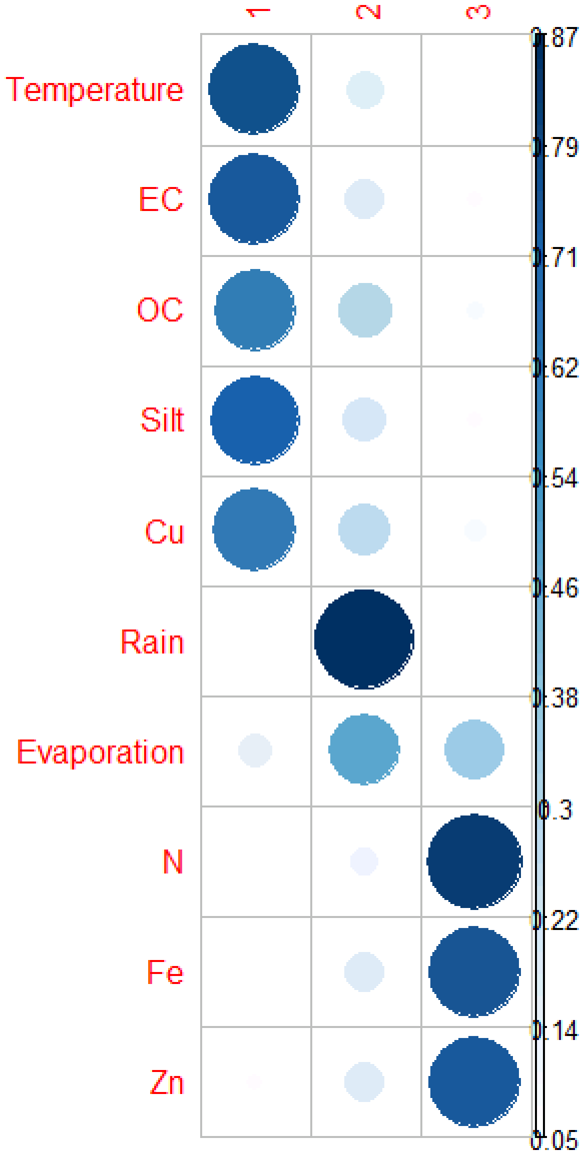

Figure 1. Fuzzy clustering method to classify the regression models fitted on the distribution of Stachys pilifera

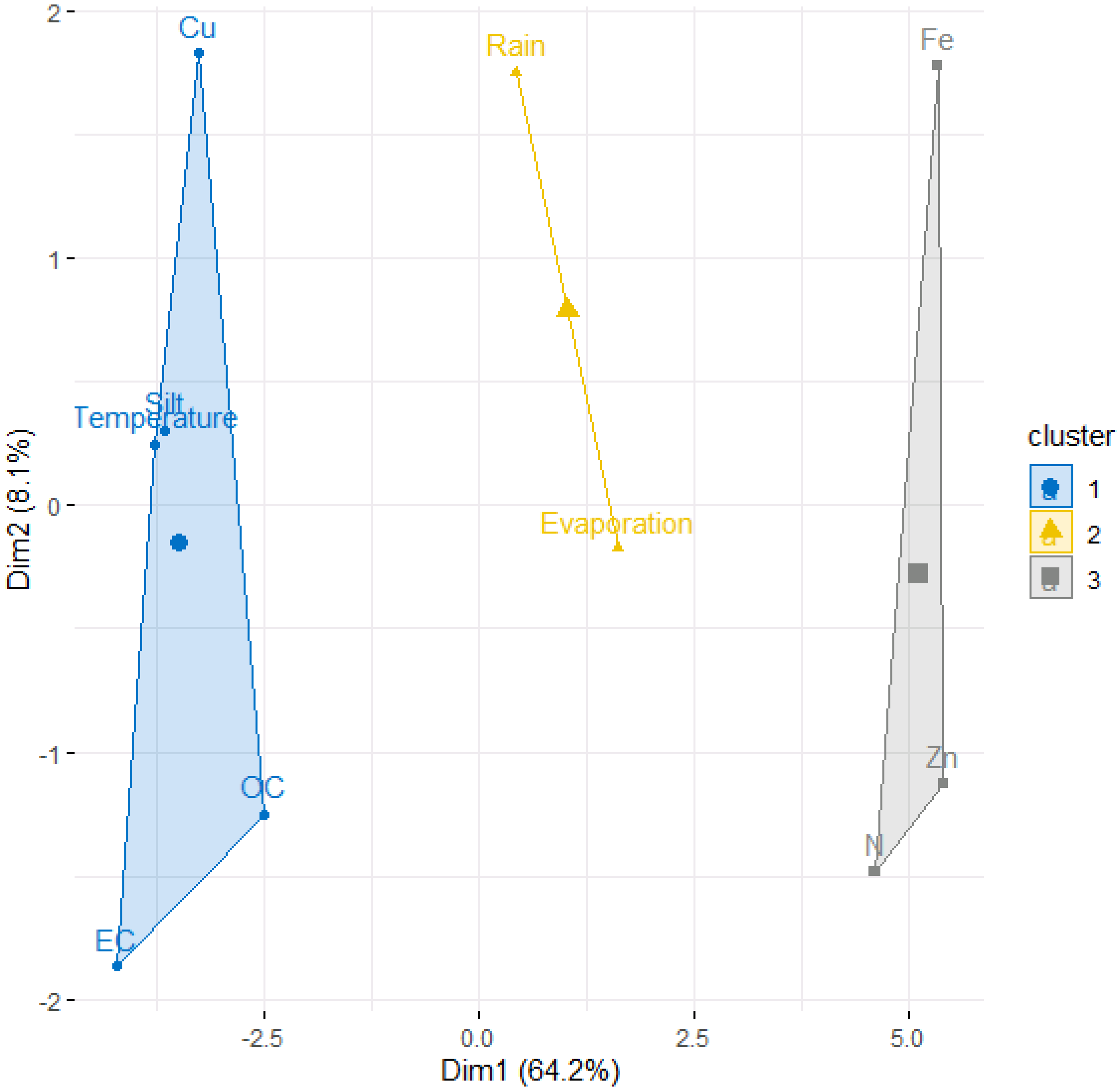

Figure 2. Fuzzy clustering plot to classify the regression models on the distribution of Stachys pilifera

4. Conclusion

Clustering numerous regression models adapted on the dataset is one of the most ubiquitous issues in data modeling and statistical inference. In this work, the fuzzy clustering of different regression curves with fractional Brownian motion errors, which can be used for a dataset, was considered. Our primary objective was to cluster these models and then to find a subset of inexpensive predictors to predict a response variable reasonably well. In this research, fuzzy clustering method was used to cluster regression models. Thereafter the performance of proposed method was studied in simulated and real situations. The results verified that the introduced technique had excellent power to cluster the models.

List of abbreviations

R^{2}**:** R-square

RMSE: root mean square error

PET: potential evapotranspiration

Availability of data and materials

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no competing interests.

Funding

This research received no external funding.

References

1. Bahrami, M.; Amiri, M. J.; Mahmoudi, M. R.; Koochaki, S. Modeling caffeine adsorption by multi-walled carbon nanotubes using multiple polynomial regression with interaction effects. *J Water Health* **2017**, 15(4):526-535.

2. Zarei, A. R.; Mahmoudi, M. R. Evaluation of changes in RDIst index effected by different Potential Evapotranspiration calculation methods. *Water Resour Manag **2017***, 31 (15), 4981–4999.

3. Zarei, A. R.; Moghimi, M. M.; Mahmoudi, M. RAnalysis of Changes in Spatial Pattern of Drought Using RDI Index in south of Iran. *Water Resour Manag **2016***, 30 (11), 3723- 3743.

4. Zarei, A. R.; Moghimi, M. M.; Mahmoudi, M. R. Parametric and Non-Parametric Trend of Drought in Arid and Semi-Arid Regions Using RDI Index. *Water Resour Manag **2016***, 30 (14), 5479-5500.

5. Fisher, R.A. On the Probable Error of a Coefficient of Correlation Deduced from a Small Sample. Metron **1921**, 1, 3-32.

6. Howell, D. C. *Statistical Methods for Psychology*. 6th edition, Thomson Wadsworth: Stamford, Connecticut, US, 2007.

7. Mahmoudi, M.R; Mahmoodi, M. Inferrence on the Ratio of Correlations of Two Independent Populations. *J Math Ext* **2014**, *7(4),* 71-82.

8. Mahmoudi, M.R.; Mahmoudi, M.; Nahavandi, E. Testing the Difference between Two Independent Regression Models. *Commun Stat Theory Methods* **2016**, *45(21),* 6284-6289.

9. Mahmoudi, M. R.; Maleki, M.; Pak, A. Testing the Equality of Two Independent Regression Models, *Commun Stat Theory Methods* **2018**, 47 (12): 2919-2926.

10. Hotelling, H. The Selection of Variates for Use in Prediction with Some Comments on the General Problem of Nuisance Parameters. *Ann Math Statist* **1940**, 11 (4), 271-283.

11. Williams, E.G. The Comparison of Regression Variables. *J R Stat Soc Series B* **1959**, 21, 396-399.

12. Steiger, J.H. Tests for Comparing Elements of a Correlation Matrix. *Psychol Bull* **1980**, 87 (2), 245-251.

13. Meng, X.; Rosenthal, R.; Rubin, D. B. (1992). Comparing Correlated Correlation Coefficients. *Psychological Bulletin*, ***111***, 172-175.

14. Mahmoudi, M. R. On Comparing Two Dependent Linear and Nonlinear Regression Models. *J Test Eval* **2018**, In Press.

15. Peter, C.C.; Van Voorhis, W.R. *Statistical Procedures and Their Mathematical Bases.* McGraw-Hill : New York:, US, 1940.

16. Raghunathan, T. E.; Rosenthal, R.; Rubin, D.B. Comparing Correlated but Nonoverlapping Correlations. *Psychol Methods* **1996**, *1*, 178-183.

17. Soto, J., Flores-Sintas, A., Palarea-Albaladejo, J. Improving probabilities in a fuzzy clustering partition, *Fuzzy Sets and Systems* ***2006***, 159(4), 406–421.

18. Ferraro, M. B., Giordani, P. A toolbox for fuzzy clustering using the R programming language, *Fuzzy Sets and Systems* ***2017***, 279(4), 1–16.

19. Samadianfard, Saeed, et al. "Wind speed prediction using a hybrid model of the multi-layer perceptron and whale optimization algorithm." Energy Reports 6 (2020): 1147-1159.

20. Taherei Ghazvinei, Pezhman, et al. "Sugarcane growth prediction based on meteorological parameters using extreme learning machine and artificial neural network." Engineering Applications of Computational Fluid Mechanics 12.1 (2018): 738-749.

21. Qasem, Sultan Noman, et al. "Estimating daily dew point temperature using machine learning algorithms." Water 11.3 (2019): 582.

22. Mosavi, Amir, and Atieh Vaezipour. "Reactive search optimization; application to multiobjective optimization problems." Applied Mathematics 3.10A (2012): 1572-1582.

23. Shabani, Sevda, et al. "Modeling pan evaporation using Gaussian process regression K-nearest neighbors random forest and support vector machines; comparative analysis." Atmosphere 11.1 (2020): 66.

24. Ghalandari, Mohammad, et al. "Aeromechanical optimization of first row compressor test stand blades using a hybrid machine learning model of genetic algorithm, artificial neural networks and design of experiments." Engineering Applications of Computational Fluid Mechanics 13.1 (2019): 892-904.

25. Mosavi, Amir. "Multiple criteria decision-making preprocessing using data mining tools." arXiv preprint arXiv:1004.3258 (2010).

26. Karballaeezadeh, Nader, et al. "Prediction of remaining service life of pavement using an optimized support vector machine (case study of Semnan–Firuzkuh road)." Engineering Applications of Computational Fluid Mechanics 13.1 (2019): 188-198.

27. Asadi, Esmaeil, et al. "Groundwater quality assessment for sustainable drinking and irrigation." Sustainability 12.1 (2019): 177.

28. Mosavi, Amir, and Abdullah Bahmani. "Energy consumption prediction using machine learning; a review." (2019).

29. Dineva, Adrienn, et al. "Review of soft computing models in design and control of rotating electrical machines." Energies 12.6 (2019): 1049.

30. Mosavi, Amir, and Timon Rabczuk. "Learning and intelligent optimization for material design innovation." In International Conference on Learning and Intelligent Optimization, pp. 358-363. Springer, Cham, 2017.

31. Torabi, Mehrnoosh, et al. "A hybrid machine learning approach for daily prediction of solar radiation." International Conference on Global Research and Education. Springer, Cham, 2018.

32. Mosavi, Amirhosein, et al. "Comprehensive review of deep reinforcement learning methods and applications in economics." Mathematics 8.10 (2020): 1640.

33. Ahmadi, Mohammad Hossein, et al. "Evaluation of electrical efficiency of photovoltaic thermal solar collector." Engineering Applications of Computational Fluid Mechanics 14.1 (2020): 545-565.

34. Ghalandari, Mohammad, et al. "Flutter speed estimation using presented differential quadrature method formulation." Engineering Applications of Computational Fluid Mechanics 13.1 (2019): 804-810.

35. Ijadi Maghsoodi, Abteen, et al. "Renewable energy technology selection problem using integrated h-swara-multimoora approach." Sustainability 10.12 (2018): 4481.

36. Mohammadzadeh S, Danial, et al. "Prediction of compression index of fine-grained soils using a gene expression programming model." Infrastructures 4.2 (2019): 26.

37. Sadeghzadeh, Milad, et al. "Prediction of thermo-physical properties of TiO2-Al2O3/water nanoparticles by using artificial neural network." Nanomaterials 10.4 (2020): 697.

38. Choubin, Bahram, et al. "Earth fissure hazard prediction using machine learning models." Environmental research 179 (2019): 108770.

39. Emadi, Mostafa, et al. "Predicting and mapping of soil organic carbon using machine learning algorithms in Northern Iran." Remote Sensing 12.14 (2020): 2234.

40. Shamshirband, Shahaboddin, et al. "Developing an ANFIS-PSO model to predict mercury emissions in combustion flue gases." Mathematics 7.10 (2019): 965.

41. Salcedo-Sanz, Sancho, et al. "Machine learning information fusion in Earth observation: A comprehensive review of methods, applications and data sources." Information Fusion 63 (2020): 256-272.