Comp. Meth. Civil Eng. 2010, Vol. 1 No. 1 pp. 85-94

CMCE
Computational Methods in Civil Engineering

Rough sets theory in site selection decision making for water reservoirs

M.A. Lashteh Neshaei[a,*], M. Pirouz[b]

[a] *Department of Civil Engineering, University of Guilan, P.O. Box 3756, Rasht, Iran*
[b] *Pardis International Unit, University of Guilan, Rasht, Iran.*

Received 7 October 2010, accepted in revised form 22 December 2010

Abstract

Rough Sets theory is a mathematical approach for analysis of a vague description of objects presented by a well-known mathematician, Pawlak (1982, 1991). This paper explores the use of Rough Sets theory in site location investigation of buried concrete water reservoirs. Making an appropriate decision in site location can always avoid unnecessary expensive costs which is very important in construction projects such as water reservoirs. The proposed site location investigation approach is illustrated using a case study data of a semi-buried concrete reservoir with the capacity of 15000m^3 which is under construction in the North of Iran (Guilan Province, Maklavan). In this approach, the decision rules are derived from conditional attributes in Rough Sets analysis, in accounting for data vagueness and uncertainty in potentially reducing data collection. The results of study indicate that using this method can reduce unnecessary costs in water reservoirs construction.

Keywords: Site location; Data classification; Rough Sets; Uncertainties; Decision making; Water reservoirs.

1. Introduction

Water reservoirs are commonly used in nearly all water supply systems. Water storage is provided to ensure the reliability of supply, maintain pressure, balance pumping and treatment rates, reduce the size of transmission mains, and improve operational flexibility and efficiency. Site location investigations should be performed to select the appropriate site for reservoirs. In this paper, the Rough Sets method which is introduced by Pawlak (1982) [1] is applied as a machine learning method to select the most appropriate site for constructing a reservoir. The process is reviewed in details and the results are compared with those of a case study of a 15000m^3 semi-buried concrete water reservoir under construction in the North of Iran (Guilan province, Maklavan).

[*]Corresponding author.
E-mail addresses: maln@guilan.ac.ir (M.A. Lashteh Neshaei)

2. Rough Sets theory

The Rough Sets theory [2] has often proved to be an excellent mathematical tool for the analysis of a vague description of objects .The vague, referring to the quality of information, means inconsistency or ambiguity which follows from information granulation. The Rough Sets philosophy is based on the assumption that with every object of the universe there is associated a certain amount of information (data, knowledge), expressed by means of some attributes used for object description. Objects having the same description are indiscernible (similar) with respect to the available information. The indiscernible relation thus generated constitutes a mathematical basis of the Rough Sets theory; it induces a partition of the universe into blocks of indiscernible objects, called elementary set, which can be used to build knowledge about a real or abstract world. The use of the indiscernible relation results in information granulation. Some important characteristics of the Rough Sets approach make this a particularly interesting tool in a number of problems. With respect to the input information, it is possible to deal with both quantitative and qualitative data, and inconsistencies need not to be removed prior to the analysis. With reference to the output information, it is possible to acquire a posteriori information regarding the relevance of particular attributes and their subsets to the quality of approximation considered in the problem at hand, without any additional inter-attribute preference information. Moreover, the final result in the form of ``if..., then..." decision rules, using the most relevant attribute, are easy to interpret. The original concept of approximation space in Rough Sets can be described as follows. Given an approximation space is: *apr = (U, A),* where *U* is the universe which is a finite and non-empty set, and *A* is the set of attributes. Then based on the approximation space, we can define the lower and upper approximations of a set. Let *X* be a subset of *U* and the lower approximation of *X* in *A* is

$$\overline{apr}(A) = \{x \mid x \in U, U / ind(A) \subset X\} \tag{1}$$

The upper approximation of X in A is:

$$\underline{apr}(A) = \{x \mid x \in U, U / ind(A) \cap X \neq \varphi\} \tag{2}$$

Where:

$$U / ind(a) \{(x_i, x_j) \in U.\ U, f(x_{i,a}) = f(x_i, a) = \forall a \in A\} \tag{3}$$

Equation (1) represents the least composed set in A containing X, called the best upper approximation of X in A , and equation (2) represents the greatest composed set in A contained in X , called the best lower approximation. After constructing upper and lower approximations, the boundary can be represented as:

$$BN(A) = \overline{apr}(A) - \underline{apr}(A) \tag{4}$$

According to the approximation space, we can calculate reducts and decision rules. Given an information system *I = (U, A)* then the reduct, *RED (B),* is a minimal set of attributes $B \subseteq A$ such that $r_B(U) = rA(U)$ where:

$$r_B(U) = \frac{\sum card\left(\underline{B}(X_i)\right)}{card(U)} \tag{5}$$

Denotes the quality of approximation of U by B. Once the redacts have been derived, overlaying the redacts on the information system can induce the decision rules. A decision rule can be expressed as $\varphi \Rightarrow \theta$, where φ denotes the conjunction of elementary conditions, $\Rightarrow$ denotes ‘indicates', and θ denotes the disjunction of elementary decisions. The advantage of the induction based approaches is that it can provide the intelligible rules for decision-makers (DMs). These intelligible rules can help DMs to realize the contents of data sets.

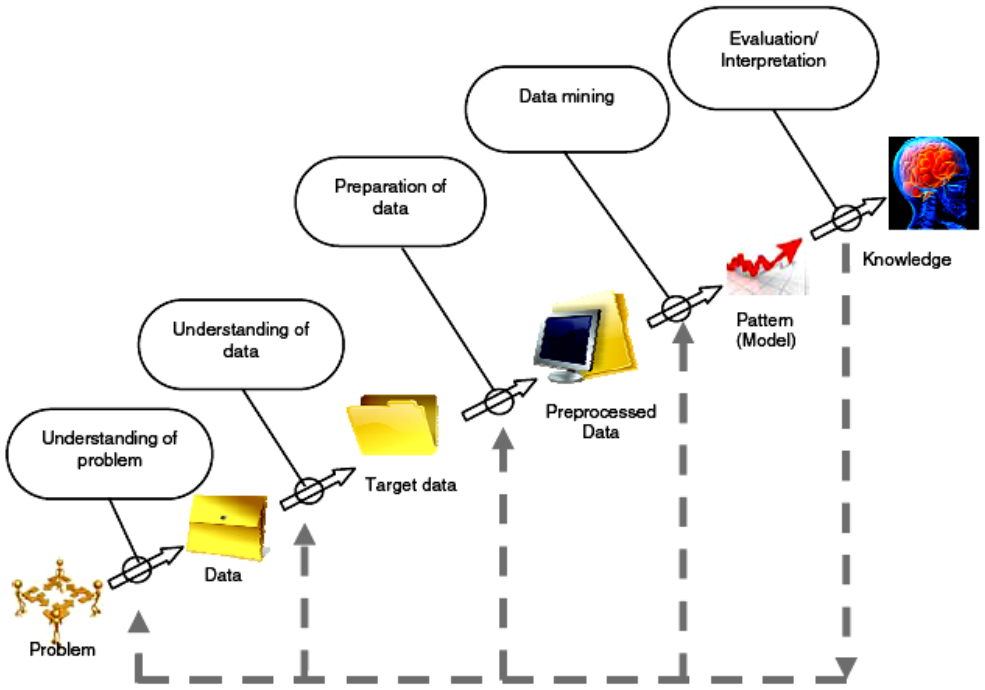

Figure 1. The steps of knowledge discovery in Rough Sets theory [12].

3. Application of Rough Sets theory in buried water reservoir site selection

In Rough Sets data modeling all the information should be categorized in a table, in which each column represents a characteristic or a property and the last column indicates the decision level. All the knowledge available about the site is the corresponding row in the table. Data tables are usually difficult to evaluate. They may store an enormous quantity of data, which is hard to manage for decision making. Due to geotechnical and civil uncertainties, some data may not be reliable for the project [3]. One of the main objectives of Rough Sets data analysis is to reduce data size. In order to show the applicability of the Rough Sets theory, the information table was constructed, according to standards that are prepared by experts in water industry [4]. Thus, table 1 can be obtained with 13 conditional attributes, such as (a) Topography; (b) Geology and tectonics, and (c) Accessibility of electric power. The collected data is categorized as a table in which each row indicates the specifications of a particular site and each column of the table indicates one of the characteristics considered for the location and the last column shows the suitability of the site for the project. For application of Rough Sets theory to analyze the information, the data should be classified. Consequently, each conditional attribute is provided with 4 classes, which show high, medium, low and no severity and the decision parameter (attribute) is classified by three levels, which describes high, medium and low suitability conditions: H, M and L, and also N which indicates none suitable condition. The classification of all attributes has been carried out by defining the specific levels and assigning a code to each specified attribute in the rows of the table. Table 2 shows the class numbers of conditional attributes and severity levels for 22 selected sites for a water reservoir in Maklavan (Guilan Province, North of Iran). For instance, site S_1 is classified into class number 2 of conditional attribute (a), class number 2 of conditional attribute (b) and class number 1 of conditional attribute (c)

and its severity level is diagnosed as "L". This table shows the relationship between the classes numbers of the conditional attributes of each site and its decision attribute. This relationship is named "decision rules" and such a table is called "a decision table".

4. Minimal decision algorithm

Principally, in this stage, the compatibility of the decision levels with 13 conditional attributes in Table 2, should be checked. The decision rules of all sites should be checked to find non-deterministic rules or sites which were classified into one and the same class under every conditional attribute but were assigned different decision levels. Non-deterministic rules were not found in Table 2, thus, the decision level confirmed subordinate to conditional attributes. If non-deterministic rules were found in such a decision table, it means that the number of conditional attributes in the decision table is not sufficient and new conditional attributes have to be added to existing ones. In the process of extracting a minimal decision algorithm, it is required, to use trial and error to determine non-deterministic rules, and make a decision table free from contradictions. In order to find the insignificant conditional attributes in the diagnoses, a number of conditional attributes should be removed each time and the decision table should be examined to make sure no contradiction has occurred. For illustration, if one removes the conditional attributes (a), (c), (e) and (m), the decision rules of sites S14 and S15 will be contradictory to each other, which denote that the decision level of locations S14 and S15 is subordinate to one of the conditional attributes (a), (c), (e), and (m). Consequently, these conditional attributes cannot be eliminated simultaneously. Each combination was eliminated from Table 2, and then, it was checked to see if any contradiction occurred within the decision rules. Seven combinations of conditional attributes, cases 1 to 7, have been shown in Table 3. All of the combinations consist of the minimum number of conditional attributes, but still are able to diagnose the problem. For illustrating the procedure for extracting the minimal decision algorithm, the method used for case 3 is presented. In case 3 of Table 3, the conditional attributes, other than those of (a), (c), and (e) of case 1, were removed from Table 2 and consequently Table 4 was acquired. Finally, the classes of the conditional attributes in Table 4 should be checked. All of the class numbers of each conditional attribute were eliminated one by one and checked for any contradiction. If any contradiction happened, it means that the removed class number is significant in the suitable site selection. Otherwise, it is not significant in the analysis. Table 5 is the "minimal decision algorithm", in which there is no conditional attribute or class removable without causing contradiction.

Table 1. Conditional attributes for decision ranking of selected sites.

Conditional Attributes	Classification of Individual Situations	Decision Levels
(a) Topography	1-flat area 2- low hills 3-fairly high hills 4- mountains or relatively high hills	**H** **M** **L** **N**
(b) Geology and tectonics	1-very fine compacted soil 2- sedimentary layers 3-possiblity of layers movement or landslide 4-low quality soils with harmful minerals ,fault zone area	**H** **M** **L** **N**
(c) Accessibility of electric power	1- availability near the site 2- relatively near the site 3- far from the site	**H** **M** **L**
(d) Distance from water resource	1-water resource near to the site 2- relatively near to the site 3- relatively long distance 4-very far from the site	**H** **M** **L** **N**
(e) Ground flooding	1-without any structure 2- existence of roads 3-existance of farmland houses 4-existanse of farms, houses and other facilities	**H** **M** **L** **N**
(f) Flood risk	1-no risk 2- low risk 3- high risk	**H** **M** **N**
(g) Seismicity background	1- low severity 2- medium severity 3- high severity 4-very high severity	**H** **M** **L** **N**
(h) Bearing capacity of soil	1- high bearing capacity 2- quite good bearing capacity 3- relatively low bearing capacity 4-very low bearing capacity	**H** **M** **L** **N**
(i) water distribution system	1- gravity 2- gravity and pumping 3- pumping 4-very difficult to transit	**H** **M** **L** **N**
(j) Future developments considerations	1- can be developed with no problem 2- need to destroy some roads 3- need to destroy farmland houses 4- need to destroy buildings or other facilities	**H** **M** **L** **N**
(k) Access to project	1-Existance of access road 2-Quite easy access to site by minor roads 3-Need to construct access road to site	**H** **M** **N**
(l) Environmental impacts	1-no environmental impacts 2-creating some solvable environmental problems 3-sever environmental impacts	**H** **M** **N**
(m) Economical considerations	1-Very Economic 2- Partly Economic 3-None Economical	**H** **L** **N**

Table 2. Data inspection for analysis of site selection decision ranking

Sites	Conditional Attributes													Decision Levels
	a	***b***	***c***	***d***	***e***	***f***	***g***	***h***	***i***	***j***	***k***	***l***	***m***	
S_1	2	3	1	3	3	1	2	1	1	2	1	1	2	L
S_2	2	2	1	1	2	2	2	2	3	2	1	2	1	M
S_3	1	2	1	3	3	1	2	2	2	1	1	1	3	L
S_4	1	3	2	2	1	1	2	1	1	1	1	1	1	H
S_5	1	3	1	3	3	2	2	2	1	2	1	2	3	L
S_6	2	2	2	2	2	1	2	1	1	1	2	1	2	M
S_7	1	2	2	2	2	1	2	1	1	1	2	1	2	M
S_8	1	2	1	1	2	1	2	1	1	1	2	1	2	H
S_9	2	2	1	3	3	1	2	2	1	2	3	1	3	L
S_{10}	1	2	1	1	2	1	2	1	1	1	1	1	1	H
S_{11}	2	3	2	3	3	1	2	2	2	3	3	1	3	L
S_{12}	1	2	1	1	2	1	2	1	1	2	1	1	1	H
S_{13}	2	2	1	2	3	3	2	3	2	3	2	2	3	L
S_{14}	1	2	1	1	1	1	2	1	2	1	1	1	1	H
S_{15}	2	2	1	2	3	1	2	1	2	1	1	1	2	L
S_{16}	1	3	2	1	2	1	2	2	3	3	2	1	2	M
S_{17}	1	1	2	2	2	2	2	1	2	1	1	2	1	M
S_{18}	1	3	1	3	3	2	2	2	1	2	3	2	2	L
S_{19}	1	1	2	1	1	1	2	1	1	1	2	1	2	H
S_{20}	1	1	2	2	1	1	2	1	1	1	1	1	1	H
S_{21}	1	2	1	2	1	1	2	1	1	1	2	1	2	H
S_{22}	3	2	2	2	1	2	2	3	2	2	2	2	1	L

Table 3. Arrangements of conditional attributes

Case Number	Conditional Attributes		
1	a	c	e
2	c	e	f
3	c	e	h
4	c	e	i
5	d	e	i
6	d	e	h

Table 4. Decision table for case 1.

Sites	Conditional Attributes			Decision Levels
	a	c	e	
S_1	2	1	3	L
S_2	2	1	2	M
S_3	1	1	3	L
S_4	1	2	1	H
S_5	1	1	3	L
S_6	2	2	2	M
S_7	1	2	2	M
S_8	1	1	2	H
S_9	2	1	3	L
S_{10}	1	1	2	H
S_{11}	2	2	3	L
S_{12}	1	1	2	H
S_{13}	2	1	1	L
S_{14}	1	1	1	H
S_{15}	2	1	3	L
S_{16}	1	2	2	M
S_{17}	1	2	2	M
S_{18}	1	1	3	L
S_{19}	1	2	1	H
S_{20}	1	2	1	H
S_{21}	1	1	1	H
S_{22}	3	2	1	L

Table 5. Rules generated by Rough Sets analysis.

Deterministic Rules	
Rule 1	***(e = 1) => (Decision = H)***
Rule 2	***(d = 1) & (h = 1) => (Decision = H)***
Rule 3	***(a = 3) => (Decision = H)***
Rule 4	***(c = 2) & (e = 2) => (Decision = M)***
Rule 5	***(i = 3) => (Decision = M)***
Rule 6	***(e = 3) => (Decision = L)***

5. Comparing Rough Sets and linear regression method

Linear Regression analysis [12] is a theoretically simple method to explore relationships among variables which is used where there are independent variables, $X_1, X_2, \ldots, X_p$ presumed to measure a cause, one dependent variable, Y, presumed to measure an effect, and the relationship between the two is linear. The relationship between Y and $X_1, X_2, \ldots, X_p$, can be estimated by the regression equation:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 x_2 + \ldots + \beta_P X_P, + \varepsilon \quad (6)$$

where, $\beta_0, \beta_1, \ldots, \beta_P$ called the regression parameters or coefficients, are unknown constants to be estimated from the data and ε is assumed to be a random error representing the discrepancy in the approximation. In this paper, it has been tried to find the effective attributes in site selection for water reservoirs, so the dependent parameters should be the indicators of the decision levels and the independent parameters should represent the characteristics of the sites. However, there are still problems in the formation of regression equations which are used to select the best site, because using all the parameters and factors to form the regression equation is practically difficult or maybe impossible. Consequently, the stepwise method is used to determine the shortest and the most suitable combinations of attributes. In this method, different parameters are used to develop the best linear correlation with the highest value of R^2 by dependent parameters. In this procedure, first, the value of the correlation coefficient between each independent parameter and dependent variables is estimated. This is accomplished to determine which independent parameter can give the highest degree of correlation coefficient with the dependent parameter. The process is continued by adding each independent parameter to the primary one, in the form of a linear regression equation with two independent variables and at each step, the value of R^2 is evaluated. This trend continues until the best secondary parameter of the independent attributes is developed. This process continues until, with the addition of another independent parameter to the model, changes in R^2 are negligible. Accordingly, the presented parameters in the linear regression equation acquired in this procedure, are regarded as the most significant defined parameters that can be used for the site selection.

Table 6. Stepwise regression equations.

Step	Parameters	Equations	R^2
1	Ground flooding	Decision= 0.243 + 0.862 *e*	60.5%
2	Topography	Decision= -0.485 + 0.755 *e* + 0.675 *a*	79.3%
3	Distance from water resource	Decision= -0.714 + 0.525 *d* + 0.609 *a* + 0.412 *e*	87.3%
4	water distribution system	Decision= -1.14+0.434 *e* + 0.468 *a* + 0.562 *d* + 0.347 *i*	92.4%

The analysis of the information has been accomplished by stepwise method and the results indicate that the ground flooding (e) is the most significant attribute, providing the highest degree of correlation coefficient with a dependent parameter. By the same manner, the parameters regarding topography (a) and distance from water resource (d) are regarded as the second and third mostly important parameters. The above mentioned parameters form a four variable regression equation which provides the highest value of R^2 between the other four-variable equations. The results of linear regression equations are shown in Table 6. It can be seen that the water distribution system parameter (i), after the three above-mentioned parameters, is the most significant factor among the independent parameters. It is observed that by adding and replacing other additional parameters in the outcome structure of the

equation of these three independent parameters, slight changes happen in the rate of R^2 which can be neglected. This indicates that applying these independent parameters with high value of correlation can make an appropriate site selection for the project and the set of these four parameters resulted from the stepwise regression analysis is considered as the best algorithm. The results of the decision-making algorithms developed from the rough set analysis and the algorithm resulted from the stepwise linear regression analysis accomplished by statistical software through the regression method, are listed in Table 7.

Table 7. Coefficient of determination, quality and accuracy of approximation of different algorithms.

Algorithm	R^2	Accuracy of Approx.	Quality of Approx.
{a ,c ,e}*	81.6%	1	1
{c ,e ,f}*	72.4%	1	1
{c ,e ,h}*	81.4%	1	1
{d ,e ,h}*	86.9%	1	1
{d ,e ,i}*	84.8%	1	1
{c ,e ,i}*	68.8%	1	1
{a ,d ,e ,j}*	88.2%	1	1
{d ,e ,f ,j}*	81.1%	1	1
{d ,e ,h ,j}*	87.2%	1	1
{c ,e ,i, k}*	70.2%	1	1
{d ,e ,i ,k}*	85.7%	1	1
{a ,c ,e ,i, k}*	82.3%	1	1
{a ,d ,e ,i ,k}*	92.5%	1	1
{a ,d ,e , i}**	92.4%	0.66	0.8

** The shortest decision-making algorithms resulted from Rough Sets analysis*
*** The algorithm resulted from stepwise regression analysis*
a = Topography,
b = Geology and tectonics,
c = Accessibility of electric power,
d = Distance from water resource,
e = Ground flooding,
f = Flood risk,
g = Seismicity background,
h = Bearing capacity of soil,
i = water transition system,
j = Future developments considerations,
k = Access to project,
l = Environmental impacts,
m= Economical considerations
The R^2 value resulted from the linear regression of all parameters (all the 13 parameters) is 93.3%.

In this table, the shortest and the most important algorithms obtained from the Rough Sets analysis are recorded. In addition, the rate of the correlation coefficient of some linear equations resulting from the two algorithms is evaluated. As shown in Table 7, the difference between the values of R^2 resulting from the decision-making algorithm of the Rough Sets and that of the stepwise linear regression algorithm can be neglected. Besides, the R^2 value obtained from the Rough Sets algorithm is slightly different from the value of R^2 observed for different evaluated algorithms of attributes. Therefore, it can be concluded that the algorithm resulting from Rough Sets analysis provides a suitable correlation coefficient comparing with the other evaluated algorithms. Therefore, the value of accuracy and the

quality of approximation were studied for different decision-making algorithms and consequently it was observed that other algorithms of attributes have less value of accuracy and quality of approximation in comparison to reductions resulting from Rough Sets analysis.

6. Conclusions

In this article, it was shown that Rough Sets theory could provide a useful approach for water reservoirs site selection investigations. A limited data sampling of offered sites and their characteristics for a concrete water reservoir in the North of Iran was used. The advantage of the Rough Sets modeling approach is that the decisions generated by the model are explicit and the modeling process is not limited to restrictive assumptions. Additional advantages of the Rough Sets approach include a method of reducing the cognitive complexity of the attribute space by finding residuals and the flexibility to decrease decision rules based on their strength, thus providing engineers with an additional valuable tool in finding essential attributes to select the suitable site for the project.

References

[1] Z. Pawlak, Rough sets: *Theoretical Aspects of Reasoning about Data*, Kluwer Academic Publishers, 1991.
[2] C.T. Hung, J.R. Chang, J. D. Lin and G. H. Tzeng, *Rough Set Theory in Pavement Maintenance Decision*, Springer-Verlag, 2009.
[3] M. Arabani, M.A. Lashteh Neshaei, Application of Rough Set theory as a new approach to simplify dams location, *Scientia Iranica*, Vol. 13, 2 (2006) 152-158.
[4] Iran bureau for water and wastewater engineering system and standards, code 436, *Guideline for hydraulics of water treatment plant*, 2008.
[5] P.V. Gorsevski, P. Jankowski, Discerning landslide susceptibility using rough sets, *Computers, Environment and Urban Systems*, Vol. 32 (2008)53-65.
[6] W. Fellin, H. Lessmann, M. Oberguggenberger, R. Vieider, *Analyzing Uncertainty in Civil Engineering*, Springer-Verlag, 2005.
[7] Y. Zhou, S. Fried, F. Stiemer, Engineering analysis with uncertainties and complexities, Using reasoning approaches, *Joint International Conference on Computing and Decision Making in Civil and Building Engineering*, Montreal, Canada, (June 14-16, 2006).
[8] J.R. Chang, C.T. Hung, G.H. Tzeng and W.H. Hsiao, Pavement maintenance and rehabilitation decisions derived by Rough Set theory, *Joint International Conference on Computing and Decision Making in Civil and Building Engineering*, Montreal, Canada, (June 14-16, 2006).
[9] Z. Kersner, D. Novak, J.L. Vítek, and B. Teply, Decision-making support based on virtual statistical modeling: Case study of large subway tunnel launching, *Joint International Conference on Computing and Decision Making in Civil and Building Engineering*, Montreal, Canada, (June 14-16, 2006).
[10] T.M.Walski, *Hydraulic Design of Water Distribution Storage Tanks*, Pennsylvania American Water Company, Wilkes-Barre, 2007.
[11] M. Miradi, A.A. Molenaar, and M.F.C. van de Ven, Knowledge discovery and data mining using artificial intelligence to unravel porous asphalt concrete in the Netherlands, *Intelligent and Soft Computing in Infrastructure Systems Engineering*, Vol. 259 (2009) 107-176. .
[12] S. Chattefuee, A.S. Hadi, *Regression Analysis by Example*, 4th edition, John Wiley & Sons, 2006.