A Residential Real-Estate Valuation Model with Reduced Attributes

Bahar BULUT, Novruz ALLAHVERDİ, Humar KAHRAMANLI and Şükran YALPIR

Abstract: Prediction of real-estate prices by using its properties together is called as real-estate valuation. However, to predict approximate rate will be very complex problem. Since, every real-estate has many different features due to the location and the usage of it. Though, it is possible to find estimated values instead of exact values. In this paper, effects of attribute reducing on real-estate valuation were investigated. Real-estate data which were collected from Konya region were used to test the success of method. Attribute reducing has been applied to data. For this purpose correlation was used. Two data sets were created: First data set was created with all attributes and second data set was created with reduced attributes. 10 fold cross validation has been used for evaluation**.** Multiple regression analysis (MRA) was used for modeling. Statistical calculations showed that the prediction success of method which was created with reduced attributes is higher than the prediction success of method which was created with all attributes.

Keywords: Attribute reducing, correlation, cross-validation, multiple regression analysis, and real-estate valuation.

I. Introduction

REAL-ESTATE valuation is the task of appraising the prospective price of a site or building in the case of a sale. Such appraisals are important for investment decisions, for real-estate funds and project developments [1].

In general, real-estate valuation is independent, neutral and objective estimation of a real-estate, a real-estate project or rights and benefits of a real-estate, on the day of the valuation [2].

The value of many goods and services are generally determined with a single value due to ignore of the competitive conditions. However, there is no single value like that for real-estates that considered as an economic investment vehicle. And also the prediction of the rate of changes in values is very complex problem. Because of this, the different values come up with the same real-estates in our country. Whereas, a certain period of time, real-estate must have a single value. Even if the value of real-estate determined with different approaches, these values are expected to be close to each other in an acceptable range [3].

In fact, it is not possible to determine precisely the value of a real-estate. Because, every real-estate have different features due to the location and the usage of it and these features may vary from person to person in terms of quality and quantity and obtaining precise value becomes a complex problem. However, it is possible to find estimated values instead of exact values. Hence, some elected objective and subjective criteria, are evaluated separately. These valuations, done per unit area or volume are then reflected in the entire real-estate. For this reason primarily, real estate's value criteria should be classified [4]. Some of the criteria taken into account in the evaluation of any real-estate are given below:

1) Intended use of real-estate
2) Environmental Specifications
3) Location of real-estate
   - Transportation and proximity of the areas that given public service
   - The distance to shopping malls
   - The distance to the areas of education and worship
   - The distance to the unhealthy and harmful areas
   - Noise
   - The distance to historical sites
4) Locational Properties
   - Topographical structure
   - Soil structure
   - Shape and size of the parcel
   - Facade use
5) Scenery

Real-estate valuation is carried out by independent and professional valuation experts at developed western countries. The value which is determined by the experts may be such as the normal commerce value, the marketing price, the sale price, demand value etc. and this value is determined by using a variety of methods.

Residential real-estate valuation methods are classified into three as: traditional valuation methods, statistically valuation methods and advanced valuation methods [5].

1) Traditional Valuation Methods
   - Comparison Method
   - Income Method
   - Cost Method
   - Benefit Method
2) Statistically Valuation Method
   - Multiple Regression Method
   - Hedonic Pricing Method
   - Spatial Analysis Method
3) Advanced Valuation Methods
   - Artificial Neural Networks
   - Fuzzy Logic
   - Autoregressive Integrated Moving Average Method (ARIMA)

The traditional methods are usually based on the comparison principle and not appropriate for automatic valuation as they use a very limited data in similar groups of real-estates. Due to factors like the difficulty in finding residential real-estate similar to the considered real-estate, having many factors affecting the price, variability due to local regions and preferences, and the difficulty in defining the situation with common mathematical equations which are encountered in value estimation with traditional valuation methods, the statistical and the advanced valuation methods have been developed.

Artificial intelligence methods and application areas are divided into many different subjects. Of the advanced valuation methods using artificial intelligence methodologies (such as ANN, FIS, ANFIS etc.), ANN has been reported as an appropriate approach in determining the values of real-estate [6-13]. Regression analysis and correlation are applied in many fields such as Domestic Product, loan payment and Measuring Service Quality [14-17]. Regression, especially multiple regression analysis (MRA) is one of the well known statistical methods in which many criteria affecting the valuation of real-estates are involved, each of which having a different effect on the value [18-20]. MRA is a technique that allows additional factors to enter the analysis separately so that the effect of each can be estimated. It is valuable for quantifying the impact of various simultaneous influences upon a single dependent variable [21]. Linear MRA method has been generally used for this purpose in the literature.

In real estate valuation, many criteria are taken in the account. This process also leads to complexity. Therefore, attribute reducing needed. In this study, for attribute reducing correlation was used, the success of the method was tested by using Multiple Regression Analysis (MRA).

II. Materials and Methods

Mathematical modeling was performed in the study. MRA was the mathematical tool for the modeling approaches.

A. Materials

In this study the workspace is Konya city which has the biggest area in Turkey. In Fig.-1 Konya has been shown as green on the Turkey map. In Fig.-2 the area has been shown where the data have been collected from. This area is the largest district of Konya and the name of this district is Selçuklu.

The data belonging to 190 flats have been collected from Konya (Turkey) District, have been used for modeling. The data have been collected from land agencies.

The original data have 14 attributes (inputs) and 1 output namely number of rooms, the size of the house, floor information, number of floors, facade, parking status, the age of the building, warm-up status, road conditions, the distance to public transport, the distance to education sites, the distance to health centers, the distance to police-stations, the distance to parks and market value. The attribute reducing has been applied to data and it has concluded that, 5 attributes namely number of rooms, number of floors, parking status, warm-up status, the distance to police-stations, have been removed from the database. After this process, remaining 9 attributes have been used as input variables and the market value as output variable. Maximum, average, minimum and standard deviation values of the input and output data used in the models were presented in Table 1. Attributes that have been shown as bold are remaining attributes for second data set.

TABLE I

MAXIMUM, AVERAGE, MINIMUM AND STANDARD DEVIATION VALUES OF REAL-ESTATE ATTRIBUTES

	Max	**Average**	**Min**	**Standard Deviation**
Number of rooms	5	4	3	1
Number of floors	14.0	6.7	3.0	2.7
Warm-up status	2.0	1.4	0.0	0.9
Parking status	1.0	0.9	0.0	0.3
Distance to police-stations(m)	1115.0	600.9	64.0	224.5
Size of the house(m^2)	**220.0**	**136.9**	**90.0**	**24.8**
Floor information	**11.0**	**3.6**	**0.0**	**2.1**
Facade	**3.0**	**1.3**	**0.0**	**1.1**
Age of the building	**40.0**	**25.4**	**5.0**	**10.1**
Road conditions	**1.0**	**0.5**	**0.0**	**0.5**
Distance to public transport(m)	**265.0**	**86.9**	**10.0**	**70.2**
Distance to education sites(m)	**1203.0**	**462.7**	**26.0**	**276.3**
Distance to health centers(m)	**1562.0**	**739.5**	**51.0**	**361.6**
Distance to parks(m)	**1543.0**	**530.7**	**10.0**	**371.6**
Market value(TL)	**320000.0**	**110039.5**	**55000.0**	**39418.7**

Fig.-1: Map of Turkey

Fig.-2: Map of Selçuklu District in Konya

A. Methods

The correlation is one of the most common and most useful statistics. A correlation is a single number that describes the degree of relationship between two variables.

The main result of a correlation is called the correlation coefficient (or "r"). It ranges from -1.0 to +1.0. The closer r is to +1, the more closely the two variables are related.

If r is close to 0, it means there is no relationship between the variables. If r is positive, it means that as one variable gets larger the other gets larger. If r is negative it means that as one gets larger, the other gets smaller (often called an "inverse" correlation).

Correlation coefficient will be calculated by using different formulas. One of them is given below:

$$\rho = \frac{n(\sum(x_i y_i) - (\sum x_i)(\sum y_i)}{\sqrt{n(\sum x_i^2) - (\sum x_i)^2}\sqrt{n(\sum y_i^2) - (\sum y_i)^2}} \quad (1)$$

Estimating the accuracy of a classifier induced by supervised learning algorithms is important not only to protect its future prediction accuracy, but also for choosing a classifier from a given set (model selection), or combining classifiers [22].

The cross-validation estimate is a random number that depends on the division into folds [23].

Cross-validation is a method for model selection according to the predictive ability of the models. Suppose that n data points are available for selecting a model from a class of models. The data set is split into two parts. The first part contains n_c data points used for fitting a model (model construction), whereas the second part contains $n_v = n - n_c$ data points reserved for assessing the predictive ability of the model (model validation). Strictly speaking, model validation is carried out using not just n_v, but all the $n = n_v + n_c$ data. There are $\binom{n}{n_v}$ different ways to split the data set. Cross-validation, as its name indicates, selects the model with the best average predictive ability calculated based on all (or some) different ways of data splitting [24].

Regression analysis is a statistical tool for the investigation of relationships between variables. Multiple linear regression is one of the main tools of statistical modeling widely used for estimation of a dependent variable by its predictors. Regressions are very effective for prediction, but are not always useful for the analysis and interpretation of the individual predictors input due to multicollinearity effects [25].

Regression analysis is one of the basic tools of scientific investigation enabling identification of functional relationship between independent and dependent variables. In the classical regression analysis both the independent and dependent variables are given as real numbers [26].

In its simplest form regression analysis involves finding the best straight line relationship to explain how the variation in an outcome (or dependent) variable, Y, depends on the variation in a predictor (or independent or explanatory) variable, X. Once the relationship has been estimated we will be able to use the equation:

$$Y = b_0 + b_1 X \quad (2)$$

And in multiple regression we will be able to use the equation below:

$$y_i = \beta_0 + \beta_{1x_{i1}} + \cdots + \beta_{kx_{ik}} \quad (3)$$

III. Statistical Analysis Of The Models Performances

In order to determine the models performances comparatively, the average approximation (AA) ratio (equation (4)), standard deviation (SD in equation 5), standard deviation percentage (equation 6) and the coefficient of determination (R^2 in equation 7) were calculated.

$$AA\% = \frac{\sum_{i=1}^{n}(1 - \frac{|x_p - x_i|}{x_p})}{n} * 100 \quad (4)$$

$$SD = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x_i})^2}{(n-1)}} \quad (5)$$

$$SD\% = 100 * \frac{SD_{model}}{SD_{output}} \quad (6)$$

$$R^2 = 1 - \frac{\sum(x_p - x_i)^2}{\sum_i^n (x_i - \bar{x})^2} \quad (7)$$

Where, x_p is Market value, x_i is Value of the model, i: $\{1,2,3\ldots.n\}$$n$ is Total number of the residential real-estates in the data set, SD is Standard deviation of the approximations in the data set, AA_i is Approximation value.

IV. Evaluation

In the first stage of this study, attribute reducing has been applied to data. For this purpose, correlation was used. The correlations between flat attributes have been calculated. The high correlation between two attributes means that when one of the attribute is grows the other grows too. Because of this characteristic of correlation it was considered in this study that using two attributes in a model together which are high correlated, will reduce the success of the model. For example; a 150m^2 house consists of 4 rooms and the price of this house is approximately 70.000TL. In the same region, another house which is 200m^2, will consists of 5 rooms most likely. For this reason, creating a model according to each attributes will increase the price of the second house unnecessarily. As shown in this example, removing one of these two related features, will give better results. Here, the most important problem is to decide which attribute will remove. There are price differences between a 170m^2 house and a 150m^2 house which has the same characteristics. Hence, estimating according to the size of the house, gives more successful results to estimating according to the information of the number of rooms. Thus, the size of the house value was preferred between these two attributes. As a result of the calculated correlations between the other attributes, the number of rooms, the distance to police stations, warm-up status, number of floors and parking status attributes were removed from the database. Eventually, remaining 9 attributes were used as input values and market value was used as output value.

In the second stage, the data have been normalized to [0, 1]. After then, two data sets were created. First data set was created with all attributes and second data set was created with reduced attributes.

In the third stage, the data in each data set have been separated into 10 folds to use in modeling. Thus, every fold had 19 data. 10-fold cross validation has been used for evaluation. 10 models have been created with 171 training data and tested with 19 testing data. The MRA has been used for modeling.

According to the results of statistical evaluations, it has seen that the model's success that created with reduced attributes was higher than the model which is obtained from the first form of data without reducing the attributes.

As known, R^2 value of a successful model must be close to 1. As seen from Fig.-3 and Fig.-4, R^2 value of the model which was created with reduced attributes is closer to 1 than R^2 value of the model which was created with all attributes.

In Table-2, it can be seen that, in most of the folds, AA (%) values of second data set is higher than first data set.

TABLE 2

R2, AA (%), SD (%) VALUES OF THE MODEL WHICH WAS CREATED WITH ALL ATTRIBUTES AND THE MODEL WHICH WAS CREATED WITH REDUCED ATTRIBUTES

	Results with All Attributes			***Results with Reduced Attributes***		
	R^2	***AA (%)***	***SD%***	R^2	***AA (%)***	***SD%***
1st Fold	0,9241	99,473	97,8296	0,9346	98,3001	97,5622
2nd Fold	0,8024	90,1485	19,5563	0,8437	91,8998	38,7324
3rd Fold	0,8646	89,576	76,5074	0,8867	90,2216	74,2765
4th Fold	0,3019	90,3919	78,6923	0,3547	92,8566	79,2678
5th Fold	0,8605	86,8715	45,0482	0,8727	85,8492	46,7925
6th Fold	0,2902	94,044	50,0233	0,4882	99,4603	66,7844
7th Fold	0,9834	91,2521	97,2807	0,9845	89,75304	95,3849
8th Fold	0,9318	88,5339	70,3985	0,9459	91,6803	81,254
9th Fold	0,7905	85,5378	64,8246	0,7838	84,7188	62,6968
10th Fold	0,2103	82,2104	81,9431	0,247	88,5542	77,0353

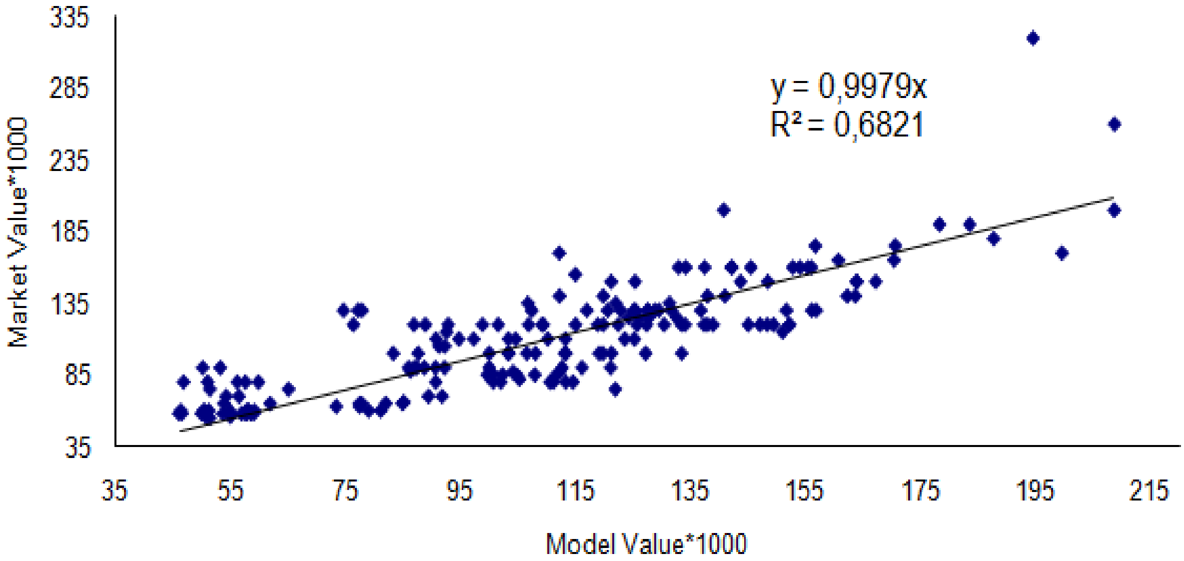

Fig.-3.The regression line between the market values and model values with all attributes

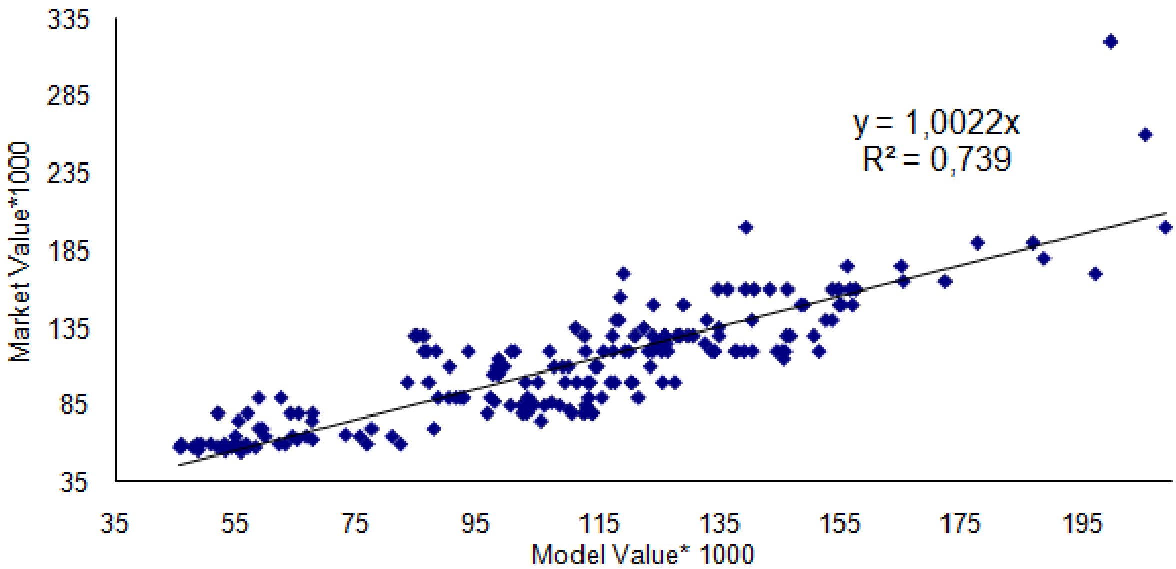

Fig.-4. The regression line between the market values and model values with reduced attributes

.

TABLE 3
THE COMPARISON OF REAL-ESTATE VALUATION

	Results with all attributes	**Results with reduced attributes**
AA (%)	84.59	86.28
SD	36.79	36.01
SD%	93.32	91.05

As seen from Table-3, *AA (%)* value of the model which was created with reduced attributes is more successful than the model which was created with all attributes. Standard deviation of the data set is 39.41866. In first model which was created with all attributes, *SD* value was 93.32%.

V. Conclusion

Making a reasonable estimation is a quite complex task on real-estate valuation which is very significant for the economies of countries. In real-estate valuation it is not possible to talk about exact methods and exact attributes. Therefore, to determine the number of most appropriate variables, affects the operating speed of the model and increases the model's performance. Thence, before deciding the real-estate valuation model, the attributes which affects valuation, must be detected.

In this study, the data belonging to 190 flats have been collected with 14 attributes. This database was used for first data set and in second data set attribute reducing was performed with the help of the correlation. These data sets were used in MRA modeling.

The results of this study can be summarized as follows:

- R^2 value of the model which is created with reduced attributes is closer to 1 than R^2 value of the model which is created with all attributes.
- *AA (%)* value of the model which is created with reduced attributes is more successful than the model which is created with all attributes.
- SD (%) value of the model which is created with all attributes is more successful than the model which is created with reduced attributes.

As a result of statistical calculations, the model produces values, closer to the real results after attribute reducing (*R^2, slope and, AA (%) values*). As seen that, *SD (%)* value of the first data set has caught the *SD (%)* value of market value. However, the difference between *SD (%)* values was very little. The purpose of this study was increasing the model's performance and operating speed and it was provided by using MRA with attribute reducing in real-estate valuation.

Acknowledgment

This study is supported by the Scientific Research Projects of Selcuk University.

References

[1] R. Schulz, "Real-Estate Valuation According to Standardized Methods: An Empirical Analysis," Center for Applied Statistics and Economics, 2002.

[2] V. Tecim, U. Cagatay, "An example study to produce real estate value maps through the geographical information system based real estate valuation studies," in 4th GIS Days in Türkiye, İstanbul, 2006.

[3] T. Yomralıoglu, R. Nisancı, "Türkiye'de harita mühendisliğinin Taşınmaz Değerlemesindeki Yeri ve Rolü (Unpublished work style)," unpublished.

[4] T. Yomralıoglu, "Taşınmazların Değerlendirilmesi ve Kat Mülkiyeti Mevzuatı," JEFOD-Kentsel Alan Düzenlemelerinde İmar Planı Uygulama Teknikleri, Trabzon, 1997, pp. 153–169.

[5] G. Özkan, S. Yalpir, "Taşınmaza Ekonomik Bakış ve Değerlendirmesi," TMMOB Harita ve Kadastro Mühendisleri Odası 10. Türkiye Harita Bilimsel ve Teknik Kurultayı, Ankara, 2005.

[6] E. Pagourtzi, V. Assimakopoulos, T. Hatzichristos and, N. French, "Practice Briefing Real-Estate Appraisal: A Review of Valuation Methods," Journal of Property Investment and Finance, vol. 21, no. 4, 2003.

[7] J. Gallego, M. Esperanza, "Artificial Intelligence Applied To Real Estate Valuation An Example For The Appraisal Of Madrid," Catastro, April, http://www. Eurocadastre. Org/Pdf/Jgallegoct50ing.Pdf, 2004.

[8] E. Pagourtzi, K. Metaxiotis, K. Nikolopoulos, K. Giannelos and, V. Assimakopoulos, "Real Estate Valuation with Artificial Intelligence Approaches," International Journal of Intelligent Systems Technologies and Applications, vol. 2, pp. 50-57, 2007.

[9] J. Guan, J. Zurada, and A. S. Levitan, "An Adaptive Neuro-Fuzzy Inference System Based Approach to Real Estate Property Assessment," JRER, vol.30, no. 4, 2008.

[10] E. Worzala, M. Lenk and, A. Silva, "An Exploration of Neural Networks and Its Application to Real Estate Valuation," The Journal of Real Estate Research, vol. 10, no. 2, 1995.

[11] V. Limsombunchai, C. Gan, and M. Lee, "House Price Prediction: Hedonic Price Model vs. Artificial Neural Network," American Journal of Applied Sciences, vol. 1, pp. 193-201, 2004.

[12] R.A Borst, W. J. McCluskey, "The Role of Artificial Neural Networks in the Mass Appraisal of Real Estate," The Third European Real Estate Society Conference, Belfast, June 26-28, 1996.

[13] H. James, E. Lam, "The Reliability of Artificial Neural Networks for Property Data Analysis," The Third European Real Estate Society Conference, Belfast, June 26-28, 1996.

[14] A. Bărbulescu, "The Analysis of Correlation of Some Ions Concentration in Rainwater in an Urban Area," International Journal Of Mathematical Models And Methods In Applied Sciences, Issue 2, vol. 4, 2010.

[15] D. Aydın, "Comparison of Regression Models Based on Nonparametric Estimation Techniques: Prediction of GDP in Turkey," International Journal Of Mathematical Models And Methods In Applied Sciences, Issue 2, vol. 1, 2007.

[16] B. Ruzgar, N. S. Ruzgar, "Rough Sets and Logistic Regression Analysis for Loan Payment," International Journal Of Mathematical Models And Methods In Applied Sciences, Issue 1, vol. 2, 2008.

[17] A. Andronov, N. Kolmakova and, I. Yatskiv, "A Quasi Regression Model for Polytomous Data and Its Application for Measuring Service Quality," International Journal Of Mathematical Models And Methods In Applied Sciences, Issue 2, vol. 4, 2010.

[18] P. Rossini, " Improving the Results of Artificial Neural Network Models for Residential Valuation," Fourth Annual Pacific-Rim Real Estate Society Conference Perth, Western Australia, 19-21 January, 1998.

[19] S. Yalpir, G. Özkan and, G. Tezel, "The Investigation of Usability of Artificial Neural Networks in The Determination of Residential Real-Estate Prices," The Association for Modeling and Simulation in Enterprises (AMSE), Turkey, 2006.

[20] H. R. Isakson, "Using Multiple Regression Analysis in Real Estate Appraisal", Appraisal Journal, October 01, 2001.

[21] A. O. Sykes, "An Introduction to Regression Analysis, Chicago Working Paper in Law & Economics," Chicago Working Paper in Law & Economics, http://www.law. uchicago.edu /files/files/ 20.Sykes_.Regression.pdf (last access 28.07.2010).

[22] D. H. Wolpert, "Stacked Generalization," Neural Networks 5, 241-259, 1992.

[23] R. Kohavi, "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection," International Joint Conference on Artificial Intelligence, 1995.

[24] J. Shao, "Linear Model Selection by Cross-Validation," Journal of the American Statistical Association, vol. 88, pp. 486-494, 1993.

[25] S. Lipovetsky, "Linear Regression with Special Coefficient Features Attained via Parameterization in Exponential, Logistic and Multinomial-Logit Forms," Mathematical and Computer Modelling, vol. 49, pp. 1427–1435, 2009.
[26] A. Bargiela, W. Pedrycz, and T. Nakashima, "Multiple regression with fuzzy data," Fuzzy Sets and Systems, vol. 158, pp. 2169–2188, 2007.

Bahar BULUT (1986, Turkey): She is a research assistant in Selcuk University, Technical Education Faculty, Department of Electronic and Computer Systems Education.

Novruz ALLAHVERDİ (1948, Azerbaijan): He is a professor in Selcuk University, Technical Education Faculty, Department of Electronic and Computer Systems Education.

Humar KAHRAMANLI (1974, Azerbaijan): She is a lecturer in Selcuk University, Technical Education Faculty, Department of Electronic and Computer Systems Education.

Şükran YALPIR (1974, Turkey): She is a research assistant in Selcuk University, Faculty of Engineering and Architecture, Department of Geomatic Engineering.