

## RESEARCH ON FACTORS AFFECTING REAL ESTATE VALUES BY DATA MINING

Filiz ERSOZ<sup>1</sup>, Taner ERSOZ<sup>2</sup>, Muhammet SOYDAN<sup>3</sup>

<sup>1, 2</sup> Karabük University, Karabük, Turkey

<sup>3</sup> Sakarya University, Sakarya, Turkey

Corresponding author's e-mail: [fersoz@karabuk.edu.tr](mailto:fersoz@karabuk.edu.tr)

**Abstract.** Construction sector has an important place in Turkey's economy. Real estate sales for the sector are increasing in parallel. However, the purchase cost is also important for those who are willing to buy a real estate. In the acquisition of real estate, factors such as size, location and age of the house are taken into consideration. The aim of the article is to conduct research on factors affecting real estate values by data mining. In this study, the most important variables that determine the value of the real estate have been investigated by data mining methods. The research has been carried out in Karabük and the variables determined according to the opinions of real estate experts. As classification methods, CHAID and C&RT algorithms have been used. It has been evaluated that both algorithm estimation results can be used. Within the framework of the study, the variables that have the most impact on the unit price have been determined, such as the size of the real estate, the distance to the city centre, the popularity, and the age of the building. The use of advanced technologies, such as statistical modelling and machine learning in real estate valuation and automatic value estimation, is of importance in determining the real value of the real estate.

**Keywords:** CHAID, classification, construction sector, C&RT, data mining, decision tree, real estate valuation.

## INTRODUCTION

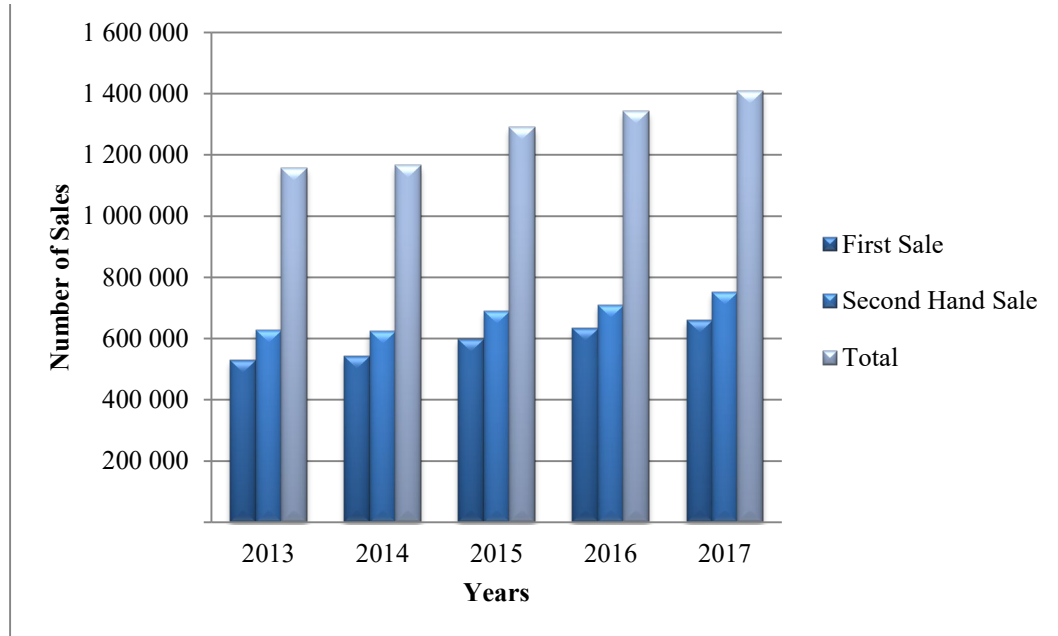
Housing and construction sector in Turkey show a structure parallel to economic development. This sector is also important in terms of employment creation. With the rapid population growth in Turkey, the need is arising for housing. For this reason, it is necessary to apply various housing finance models.

In Turkey, the government usually loads the cost of removing the need for housing and plays an active role in giving direction in the sector (Özkurt, 2007).

The valuation of real estate is the process of having a monetary counterpart and valuation depending on such factors as the structural state of the real estate, its characteristics and ease of use. This process is based on the principle of objectivity and impartiality.

Today, real estate appraisal procedures are carried out by experts who have relevant documents. At this point, apart from the experts' experience, an objective approach and future predictive skills are also important. Having enough information about the real estate appraisal area is also a very important issue.

First and second-hand housing sales have started in Turkey in recent years (see Fig. 1) (General Directorate of Land Registry and Cadastre).



**Fig. 1.** House sales in detail of first sale and second-hand sale by years, in Turkey, 2013–2017 (developed by the authors).

Different methods are monitored to determine the residential sales in Turkey. The three most commonly used methods in valuing real estate are (Açlar & Çağdaş, 2002) as follows.

- Comparison method according to a similar real estate object. It is the most used method. It is fast and easy to understand. The real estate value existing in the market is taken to determine the price of the real estate.
- Income reduction method. It is a frequently used method that takes place in the acquisition of real estate, which may be an economic gain in years. In this method, it is estimated how much the real estate can gain value in the coming years. This method, which is mostly used in the valuation of the parcels, is frequently applied in the rented business places and houses.
- Cost approach method. The method of price analysis and average cost is shaped by the price method. In the price analysis method, all properties of the real estate are valued individually and priced. According to the average cost price method, the building quality is essential. Unit prices per square metre determined according to this quality are used in the average cost price method. In houses, a cost approach method is preferred by official institutions and establishments as well as industrial structures.

## 1. FACTORS AFFECTING REAL ESTATE VALUE

Each building has its own unique identity. The better this identification is, the greater the price estimate for the property. Each building is absolutely separated from each other in terms of the quality of the material used and the quality of the workmanship. It is unlikely that the prices of the two different houses with the same characteristics, or even the two different houses in different regions, will always be the same.

However, in the case of an appraisal, certain criteria must be taken into consideration. The main factors affecting the value of the real estate are the age of the building, the distance to the centre of the city, the income situation and the availability of the area in which the real estate is located, the availability of parking lots and gardens, the location of the property in question, the elevator, the details addressed, the number of façades, the landscape, the type of heating and the area of the house (Özkan et al., 2007). These criteria are as follows.

- Building Age: The age of the building is inversely proportional to the value of housing. Earthquake regulations in construction have been considered since 1999 in Turkey, where the year 1999 should be taken as a critical date. 2007 is also critical, as the “2007 Earthquake Regulation” (the last earthquake regulation) entered into force.
- Distance: The distance to the city centre is inversely proportional to the housing value. The main thing to note here is that the busiest area of the city should not be taken as the centre. On the contrary, the number of centres taken as a centre is increased according to the structure of the city.
- Popularity: It is the level of popularity of the environment in which the residence is expected to affect housing value the most. This factor also reflects the cultural structures, economic and social levels of the people in that region. Regardless of all other variables, it is a variable that can be directly influenced by the result.
- Parking and Garden: Properties such as park and garden are important for buyers. The dwellings in the city centres are especially lacking the garden due to the adjacent construction.
- Location: Another important aspect of the apartment is the location of the apartment in which it is located. The prices of the apartments on the bottom, middle and top floors in the same building vary.
- Elevator: Elevator is compulsory in new buildings in Turkey. It is an important factor affecting a housing value. It is the first time that elderly and child families especially pay attention to housing.
- Building Heat Protection: At the first stage, mantle is not one of the first elements that comes to mind. However, sound and especially heat insulation is important in the mantle. Old buildings can be brought together with this method at the same time.
- Interior Features: Interior features are one of the first variables to multiply the space in a house. Changes can be made to the interior with subsequent restorations.

- Aspects of Building: While the number of directions of buildings in adjoining houses can be reduced to 1, the number of directions can be up to 4 in general, like a villa. In this study, there is no villa type construction in order to increase the accuracy of the results. The direction of the building varies from single-faced, double-faced and three-faced.
- Landscape: For some buyers the landscape can be the most important factor. Therefore, landscape is very influential in the value of the place.
- Heating: Although almost all of the buildings constructed today use natural gas as heating type, the use of stoves in the old buildings continues. As this study considers old buildings, the type of heating has been identified as one of the twelve titles that affect the value of the property. All the data consist of a set of samples, either natural gas or stoves.
- Area: The first factor that comes to mind in determining the value of an estate is the area in which it is owned.

## 2. LITERATURE OVERVIEW

Selim (2008), the factors that determine the price of housing in Turkey has been analysed using the 2004 Household Budget Survey. The main factors affecting the housing price were the type of house, type of building, number of rooms, size of the house and other structural variables. The results obtained with hedonic model and artificial neural networks were compared. It was found that the properties determined from these results did not show any multiple linearity between the explanatory variables.

Topçu and Kubat (2009) investigated the urban and spatial variables affecting the land value in residential areas in Istanbul. As a result of the study, it was determined that the distance to the sea, the distance to the central business area, the distance to the university, the distance to the health facility were the factors affecting the land change.

Çetintahra et al. (2011) tried to measure the effect of environmental aesthetics on house prices by a hedonic price method. Information was obtained from real estate brokers and surveys from a region selected from İzmir/Karşıyaka for many leased and purchased housing units. The results were analysed by a hedonic pricing method; used for renting and for sale housing separately. The study explored the economic values of urban design discipline, economic science, real estate and property sectors and the economic values of physical environment factors. As a result of the research, the physical criteria of the residence, “the desire to have housing”, influenced the purchase price and the “environmental” factors of the housing that has “complex” and “well-maintained” status affected the rental value.

Toktaş and Erdoğan (2012) produced the immovable value map of the city centre of Afyonkarahisar using the nominal valuation method. It is aimed at increasing the accuracy by performing a clustering analysis in the parcels taken before. As a result of the research, in the treasury sales, the values calculated from the nominal model and the values sold in the procurement procedure were compared, and in this study, the values calculated from the nominal model of treasury sales and the values of the procurement procedure were compared. In some

sales, it is seen that the closeness values have decreased to 20% and in some sales up to 90%.

Çakır and Sesli (2013) carried out a questionnaire study with the competent persons in the field of real estate appraisal and evaluated the results of the survey statistically. As a result of the analysis, the most important factors affecting the value of the land qualified properties were determined and a standard for these factors was proposed.

Yayar and Gül (2014) determined the factors affecting the value of the apartments in the city centre of Mersin using a hedonic estimation method. The usage area of the house, kitchen size, distance to the market, number of bathrooms, garage, central satellite system, private security and elevator variables increase the price of housing, the fact that the house has a garden, is within the site, away from public transportation vehicles and the former has reduced the price of housing.

Bulut et al. (2015) examined the factors affecting the housing prices of Samsun using a hedonic price model. Within this scope, 3 + 1 residences, which were ideal family size, were dealt with. E-commerce data were used. The authors predicted the rate at which the price changed according to the characteristics of the house using the computer program.

Kayral (2017), factors influencing changes in housing prices of the three most populous city in Turkey were examined. It was found that housing prices were affected by big cities and other cities in the region using logarithmic models. At the same time, it was determined that the change in housing prices in Istanbul was influenced by the changes in stock returns and the consumer price index in Istanbul.

In the study by Afşar et al. (2017), it was aimed at determining the effect levels of these factors within the framework of hedonic price model of the factors affecting the house sale prices in Odunpazarı and Tepebaşı districts located in Eskişehir. Semi-logarithmic form was used as functional form. As a result of the research, housing, the number of rooms, the number of bathrooms, the use of central heating, the presence of the car park, the presence of the built-in kitchen, the presence of the parental bathroom, the presence of the first floor of the residence were determined as the factors determining the sales prices in the dwellings.

Yilmazel and his team (2018), In the artificial neural network models, 19 different models were obtained by differentiating the secret layer neuron numbers and comparison of the performances of these models was made and the most suitable secret layer neuron number was determined. The size of the house, the number of rooms, whether it is located on the first floor or not, the total number of floors, central heating, the number of bathrooms, the elevators, the car park, the kitchen, the fibre internet, and the distance to the tramway were taken into consideration.

### **3. DATA MINING**

In this study, the factors that determine the value of real estate have been investigated by data mining methods. The main advantage of data mining methods in real estate valuation is the ability of the method to solve mixed problems. Data mining, as a semi-automatic knowledge discovery approach, includes identifying

the relationships among data collections via data patterns, managing changes and irregularities and determining the rules and structures of data sources (Larose, 2005).

At this stage of the whole process, an appropriate data mining algorithm is decided to be recruited to achieve the overall aim and objectives of the project. The aims may include data classification, clustering and analysis towards knowledge discovery needed for decision makers (Campos et al., 2015).

Decision trees are popular functionalities and tools for data classification and forecasting in which a dependent variable (the decision variable here) is related to independent variables via the branches of the tree (Samar Ali et al., 2018), where the branches make up the complete tree representing all relationships. The trees are constructed with a top-down approach, where the most important independent variables may split further sub-branches. A tree can grow very hugely depending on the complexity of the relationships. The decision tree algorithm is an approach, which stops the enlargement of the tree with irrelevant data and relationships (Silahtaroglu, 2013).

Data mining facilitates to obtain meaningful and sound knowledge from huge piles/collections of data. The knowledge acquired as a result of long-lasting and difficult processes can be obtained using data mining technologies in shorter periods of time. The acquired knowledge can be used for strategic decision making and/or objective evaluation, as it helps analyse the enterprise data resources and forecast the outcome of various business approaches (Ersöz, 2017).

### **3.1. Data Mining Steps**

Data mining begins with the retrieval of data from databases. This phase, called the collection of the data, follows the steps of clearing and integrating the data, transforming and reducing the data, respectively. If the data from different data sources are required, they are converted, or incorrect data are extracted. After data cleansing, the research model is established. The model is tested with the most appropriate algorithms in the direction of need, and the interpretation of the results of the decision maker becomes a meaningful result.

The steps of data mining are briefly described below (Ersöz, 2017).

#### **1. Identification of the Problem**

Expertise is important to identify the research problem. The aim of the project, which is decided to be solved by data mining, describes the situation at the starting point and the plans based on solving the problem. At the same time, the successful completion of the process determines what problems can be eliminated.

#### **2. Data Collection**

It is the stage of data extraction and preparation of the necessary information from reliable and appropriate sources. The larger the data to be used in the study, the greater the accuracy of the model result. Data mining is associated with large data.

### 3. Data Preparation

At this stage, the items are cleaned and, if necessary, combined to make the model ready. Incorrect or missing data may be found in the collected data set, which may corrupt the system. In addition, incorrectly entered data, called noisy data, can be included in the data set of extremely large or extremely small extremes. When the specified incorrect data are cleared, the values of the missing data can be filled in. The missing value is an accurate step in the name of trying to keep the dataset wide to give the average value of the variable.

Extraction of extreme values from the system is particularly important for a successful result. The ability to predict the system is improved by overloading the extreme values. In integration of the data, the data from different sources are combined in a single environment. In this study, repeated data are deleted and cleaned to ensure integrity.

### 4. Data Converting and Data Reduction

It is the stage where the final shape is given before the model is set up. In order to be able to use data mining at maximum capacity, the data are improved by following the corrective, merging, generalising and normalising steps while retaining the relevant properties.

The characteristics of interest in the study are defined as variables. As the model is stronger, the size can be reduced by various methods and related variables are modelled.

### 5. Creation of Model

Within the data mining methods, it is the stage of choosing the most appropriate model for research. For the reliability of the research results, appropriate models should be used. In Table 1, data mining methods and methods included in these methods are described using the defined algorithms (Ersöz, 2017).

**Table 1.** Data Mining Techniques

Techniques	Methods/Algorithms
Predictive methods	Neural Net Decision Tree      CHAID, C&RT, C5.0, etc. Linear & Logistic Regression Bayes Classifiers
Clustering/segmentation	Kohonen Network K-Means Clustering Two-Step Clustering
Association rules	Apriori GRI Carma Sequential Association

In the present study, CHAID and C&RT algorithms have been used in the analysis of data. It is also well known that there are many different mathematical and statistical approaches applicable for deciding tasks of analysis and decision making.

**CHAID Algorithm:** This method was developed by Kaas in 1980. Estimation variables are based on the best class calculation technique by combining a statistically significant difference in a pair matching the target variable. The Chi-square test is used when the best class is selected. CHAID classifies the groups that differ according to their level of relationship by chi-square analysis, and the leaves of the tree branch by the number of different structures on the verse, not the binary.

The CHAID algorithm is widely used because it can work with all types of variables, both continuous and categorical. In this method, continuous predictor variables are automatically categorised by this application as appropriate for analysis purposes.

**C&RT Algorithm:** CRT is widely used in decision tree algorithms and this algorithm was proposed by Breiman et al. in 1984. The algorithm is used for classification and regression analysis. The CRT is an algorithm that sets the classification model, while the dependent variable is categorical, and the estimation model is continuous (Larose, 2005). The CRT uses the entropy technique to select the best decimal separation criteria as in the ID3 algorithm (Dunham, 2003).

#### **4. MATERIAL AND METHODS**

In the study, the effects of the data mining techniques on the factors affecting the value of the real estate are examined.

The extent to which the unit price (TRY/m<sup>2</sup>) shaped by these factors affects the specified components has been investigated by classification and regression trees (C&RT) and CHAID algorithms, which are classifier models. At the same time, the success of the C&RT and CHAID algorithms in this study has been compared among themselves and the best predictive model determined.

The present study aims at producing solutions by using the data mining approach to the problems that the real estate appraisal experts have in determining the real estate value.

Determining from the files of historical reports in the process of determining the value of real estate is time consuming and ineffective. Instead, it is possible to store real estate information in the computer environment as data and to produce more effective and practical solutions using the digital data.

Real estate value is based on house prices. The data in the study have been taken from real estate agents in Karabük. The reports of the experts are used for determining the value of real estate operating in this province.

In the study, 128 real estate data were used. The variables related to various structural and environmental factors are taken as a basis. Attribute information:














- property island / parcel;
- real estate (land, residential, shop);
- building age;
- distance to city centre;

- popularity-demand;
- parking and garden;
- number of floors;
- real estate on the floor;
- lift;
- building heat protection;
- the number of aspects of the building;
- interior features;
- landscape;
- heating type;
- building group (3A/3B etc.);
- area.

Variables that have no effect on real estate value are not analysed. These variables, which are not analysed: real estate island / parcel and building group variables. The effect of the number of floors on the value of the property is very small, among other factors. Therefore, the number of floors has been removed from the model. In the model, only the places where the households reside have been taken as data and analysed.

## 5. RESEARCH FINDINGS

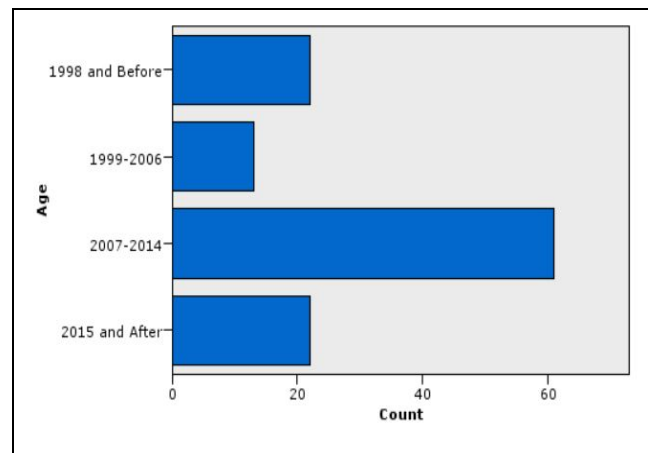
In the present study, the effects of the data mining techniques on the factors affecting the distribution of the descriptive statistics are given in Fig. 1 below.

Field	Graph	Type	Mn	Max	Mean	Std. Dev	Skewness	Median	Mode	Unique	Valid
Area		Range	35	214	107.941	37.717	0.427	105	90	--	118
Aspects_Building		Set	1	3	--	--	--	--	2	3	118
Building_Age		Set	--	--	--	--	--	--	2007-2014	4	118
Building_heat_protection		Set	--	--	--	--	--	--	yes	2	118
Distance		Range	0.200	8.200	2.940	2.279	0.639	2.100	0.800 <sup>1</sup>	--	118
Elevator		Set	--	--	--	--	--	--	yes	2	118
Heating		Set	--	--	--	--	--	--	natur.gas	2	118
Interior_Features		Set	--	--	--	--	--	--	good	4	118
Landscape		Set	--	--	--	--	--	--	medium	3	118
Location		Set	--	--	--	--	--	--	middle	3	118
Parking_Garden		Set	--	--	--	--	--	--	no	2	118
Popularity		Set	--	--	--	--	--	--	medium	4	118
Unit_Value		Range	700	2000	1351.881	291.310	0.758	1300	1300	--	118

**Fig. 2.** Descriptive statistics and distribution of data using the data mining package program (developed by the authors).

As shown in Fig. 2, the average distance from the centre of the real estate to the city centre is  $2.94 \text{ m} \pm 2.27 \text{ m}$ . It is seen that the popularity of real estate is at the middle level; there is no garden and park, natural gas is preferred for heating, the average size of the real estate is  $107.94 \text{ TRY} \pm 37.71 \text{ TRY}$ . Elevators and building protection are preferred. The unit price of the real estate is  $1351.88 \text{ TRY} \pm 291.31 \text{ TRY}$ .

It is observed that the period used in the study is between 1983 and 2017, and the age distribution of buildings is shown in Fig. 3.



**Fig. 3.** Distribution of building age (developed by the authors).

As it can be seen in Fig. 3, the buildings between 2007 and 2014 are most preferred.

IBM SPSS Modeler Package Program has been used to prepare data for the model and to work with high quality data. To build a model with data mining, the data have been cleared from extreme and outlier values, and a total of 118 data models have been established. At this stage, three datasets have been determined as an area size and a unit price of the building, and the results are given in Fig. 4.

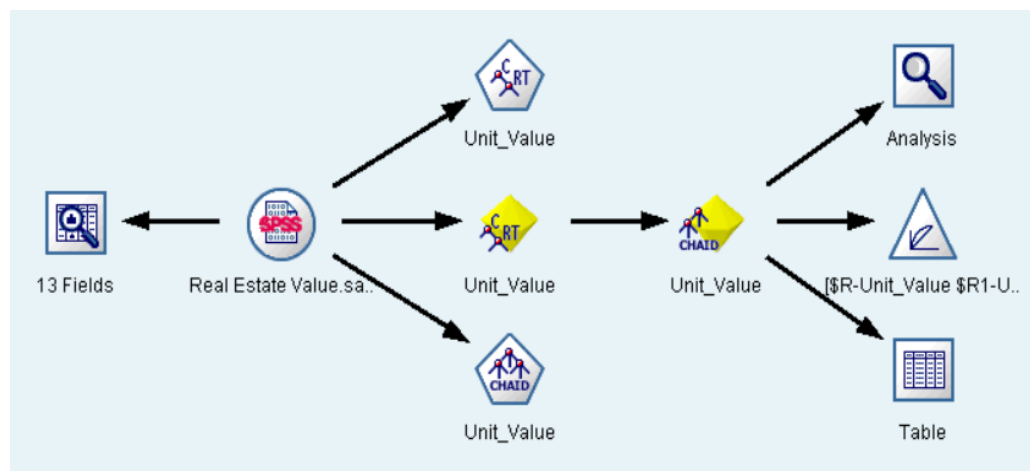
Field	Type	Outliers	% Complete	Valid Records
Age	Set	--	100	121
Distance	Range	0	100	121
Popularity	Set	--	100	121
ParkingGard...	Set	--	100	121
Location	Set	--	100	121
Elevator	Set	--	100	121
heat_protection	Set	--	100	121
Interior_Features	Set	--	100	121
Aspects_Building	Set	--	100	121
Landscape	Set	--	100	121
Heating	Set	--	100	121
Area	Range	1	100	121
UnitValue	Range	2	100	121

**Fig. 4.** Results of data quality (developed by the authors).

According to Fig. 4, three datasets have been determined to clean. The data have been cleared from extreme and outlier values. The model has been analysed with a total of 118 data units.

Data were analysed by CHAID and C & RT algorithms in IBM SPSS Modeler Decision Tree algorithm.

IBM SPSS Decision Trees enable the exploration of results and the visual determination of how the model flows. This helps find specific subgroups and relationships that are impossible to obtain with more traditional statistics” (Samar Ali et al., 2018). The results of IBM SPSS Modeler Data Mining are shown in Fig. 5.



**Fig. 5.** IBM SPSS Modeler data mining program view (developed by the authors).

### 5.1. C&RT Analysis Results of Data Mining

Classification and regression trees (C&RT) analysis within classification methods has been performed to investigate the factors affecting real estate values by means of data mining techniques. The results of C&RT decision tree analysis are given in Fig. 6 as a general perspective.

The decision tree consists of 29 nodes. All the nodes of C&RT decision tree are given in Fig. 7.



Fig. 6. C&RT model results for a real estate value (developed by the authors).

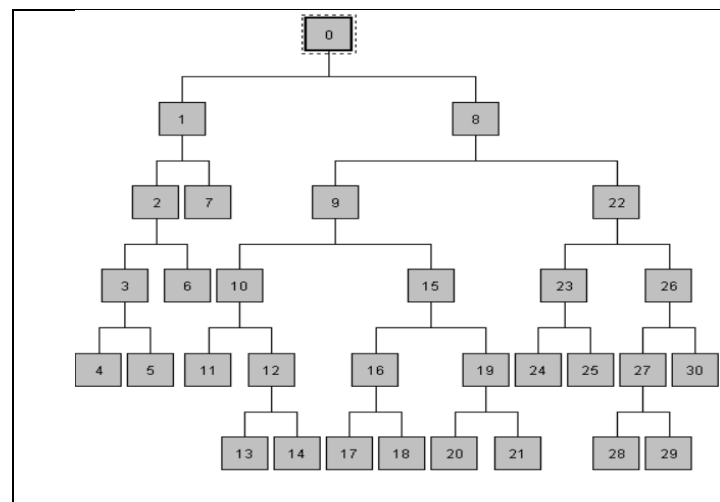
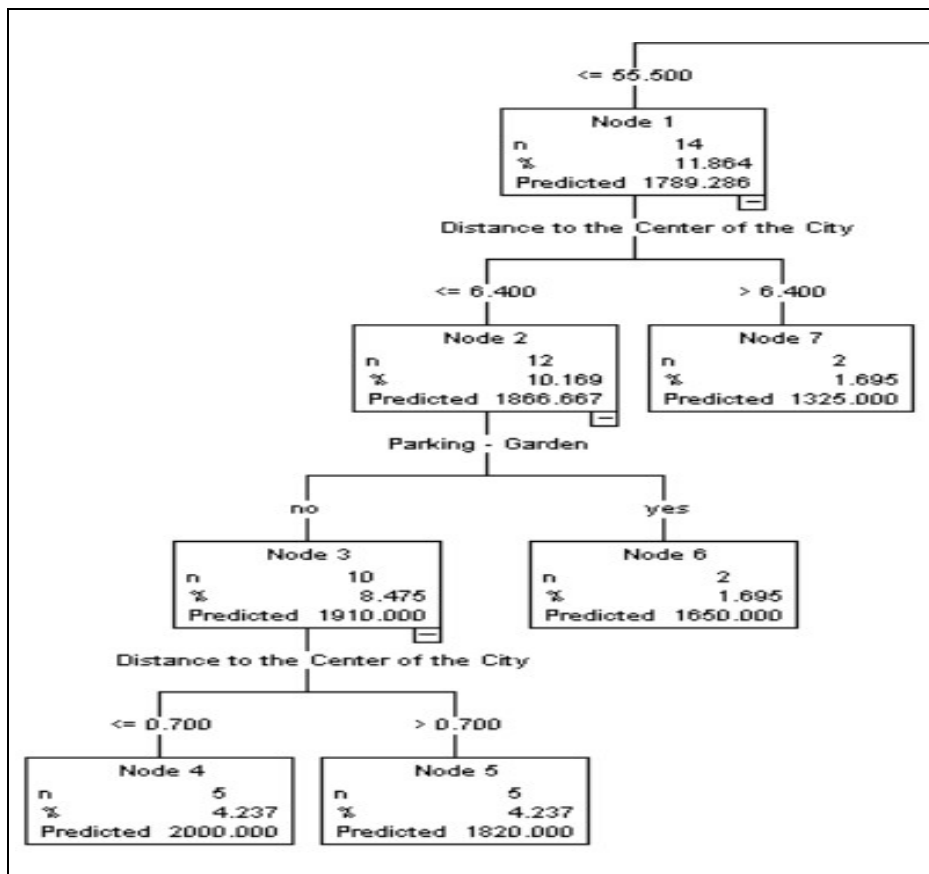


Fig. 7. All nodes of C&RT decision tree (developed by the authors).

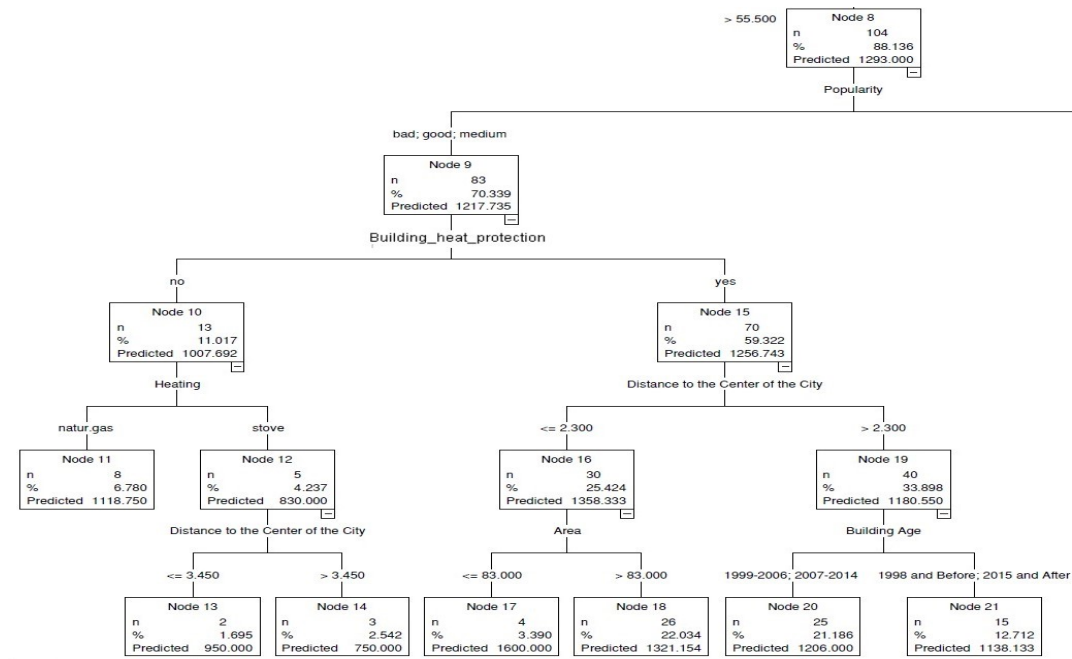
In Karabük province, the most important variable affecting the house prices per unit is the size of the house. The unit price of houses less than 55 m<sup>2</sup>, without garden and park, and less than 0.7 km to the city centre, is estimated to be 2000 TRY/m<sup>2</sup>. Karabük province is a student city. The proximity to the city centre and the area of real estate have emerged as an important element. House size is 55 m<sup>2</sup>, and the distance to the centre is less than 2.3 km, the unit price is estimated to be 1600 TRY/m<sup>2</sup>. The unit price is estimated to be 1118.75 TRY/m<sup>2</sup>, if the house does not have the protection system but there is a natural gas heating system. The price of the house is estimated to be 750 TRY/m<sup>2</sup>, if the house is over 55 m<sup>2</sup> in size, warmed by the stove and the distance to the centre is 3.48 km.

Node 4 is the node that maximises the real estate unit value. According to this node; It was determined that the buildings which are smaller than 55 m<sup>2</sup> and less than 0.7 km away from the city centre and haven't park - garden are the most valuable buildings. According to the results of the C&RT algorithm, the distance of 0.7 km to the city centre is most valuable in terms of the real estate unit value. The values of decision tree node 1 are given in Fig. 8.



**Fig. 8.** Decision tree node 1 and the estimated real estate unit value (developed by the authors).

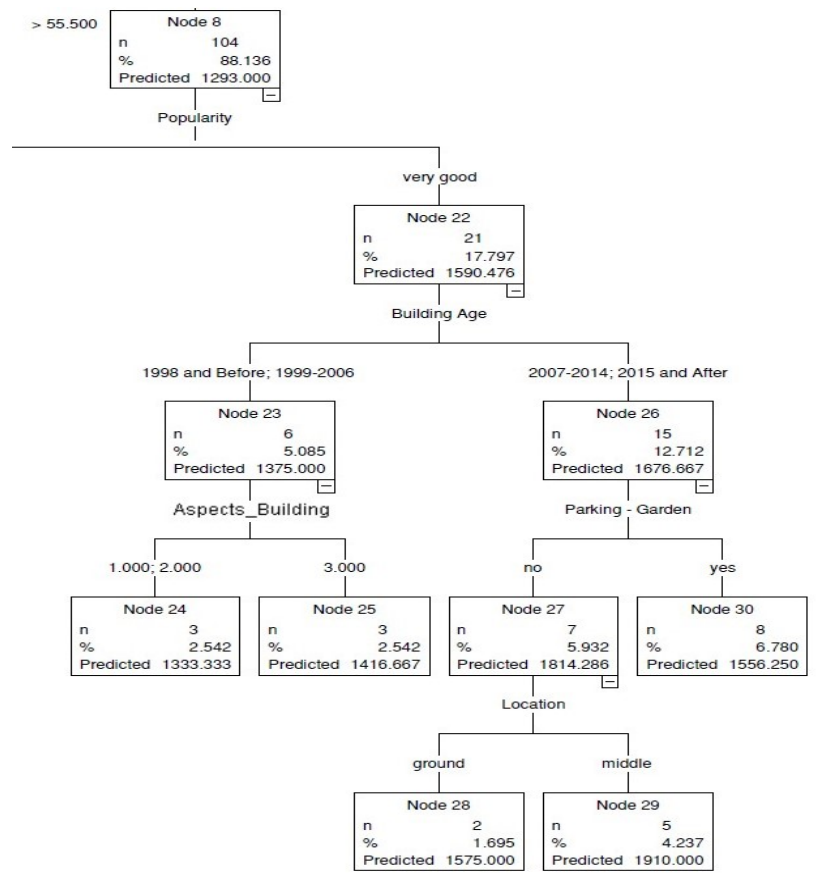
In case of houses larger than 55 m<sup>2</sup>, the unit price of houses with building protection system is 1256.74 TRY/m<sup>2</sup> and the unit price of houses without building protection system is 1007.69 TRY/m<sup>2</sup>. In addition, it is determined that the houses, which do not have building protection and natural gas heating and have more than 3450 km to the city centre, have an estimated unit price of 750 TRY/m<sup>2</sup>. The C&RT decision tree of an area that exceeds 55 m<sup>2</sup> and of areas the popularity of which is not very good is given in Fig. 9.



**Fig. 9.** Decision tree node 9 and the estimated real estate unit value (developed by the authors).

The houses with the most popular areas and over 55 m<sup>2</sup> area have a higher unit value. The houses built before and after 2007 make a difference in unit prices. The unit price of houses built after 2007 is estimated to be higher. In addition, if the house is open from three sides, the unit price is 1417 TRY/m<sup>2</sup>, and if the house has two sides, the unit price is 1333 TRY/m<sup>2</sup>.

Figure 10 demonstrates the decision tree results of the houses, which are larger than 55 m<sup>2</sup> and located in very popular regions.

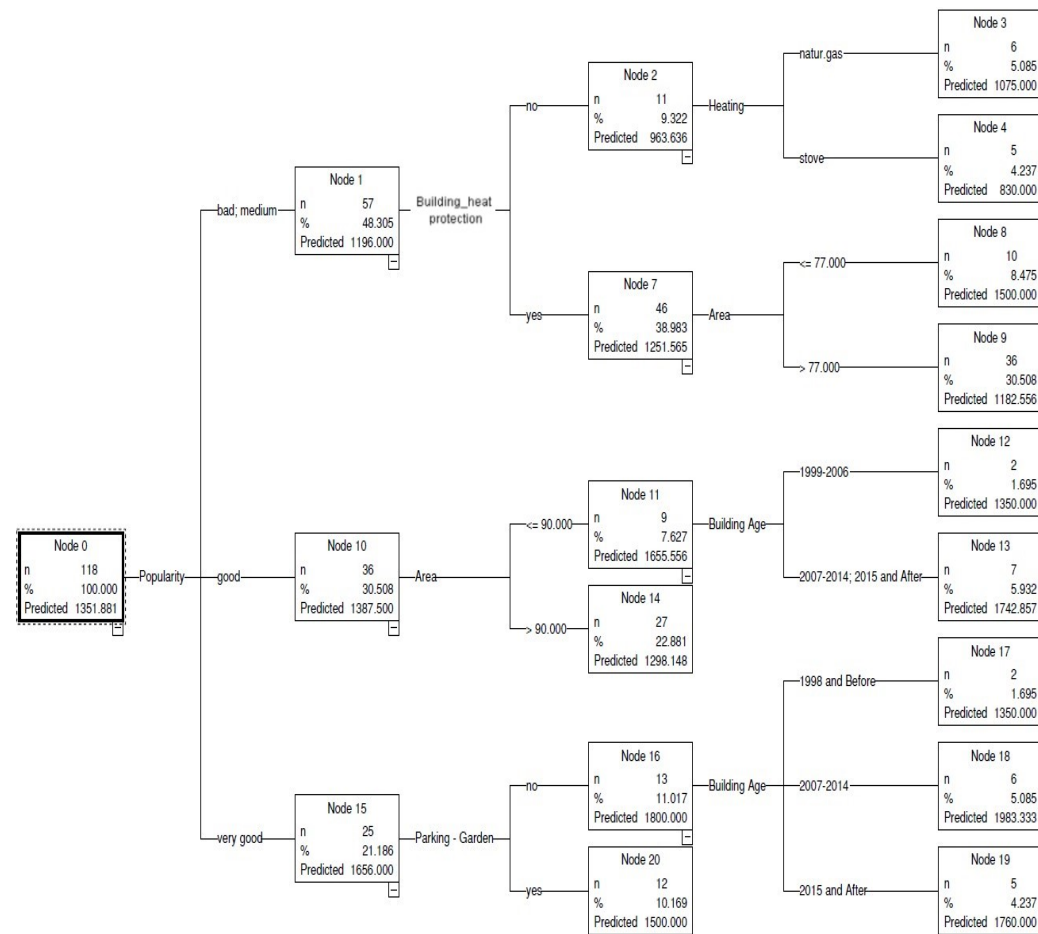


**Fig. 10.** Decision tree node 22 and the estimated real estate unit value (developed by the authors).

According to the results of C&RT general analysis, while the unit price of the house with the area of more than 55 m<sup>2</sup> is 1293 TRY/m<sup>2</sup>, the unit price of the house with an area that is less than 55 m<sup>2</sup> is 1789 TRY. The main indicators are the distance to the city centre and the popularity of the zone, in which the dwelling is located. If the location of the residence is close to the city centre (0.7 km), the price per unit is 2000 TRY/m<sup>2</sup>. The proximity to the centre causes an increase in price per unit.

## 5.2. CHAID Analysis Results of Data Mining

CHAID analysis within classification methods has been performed to investigate the factors affecting real estate values by means of data mining techniques. The results of CHAID analysis are given in Fig. 11 as a general perspective.



**Fig. 11.** CHAID algorithm results (developed by the authors).

According to the CHAID algorithm results, the average unit price of real estate is 1351 TRY/m<sup>2</sup>. The most important variable in this model is the popularity of the place where the order is located. In places where the popularity is very good, and the building age is between 2007 and 2014, the price of the unit is 1983 TRY/m<sup>2</sup>.

The most important variable in the CHAID algorithm is whether the area where the houses are located is popular. Other important variables include building protection system, size and park & garden. Houses with the lowest unit price; it has been seen in the houses where the popularity is defined as medium to bad, there is no building protection system and the stove is used as heating system and the estimated unit price of these houses is 830 TRY/m<sup>2</sup>.

### 5.3. Model Comparison and Selection

The most important variables affecting the value of real estate in the study have been researched using different classifier models (CHAID and C&RT), and the results are provided below. Data mining has been studied to determine which of the decision tree models is the best model.

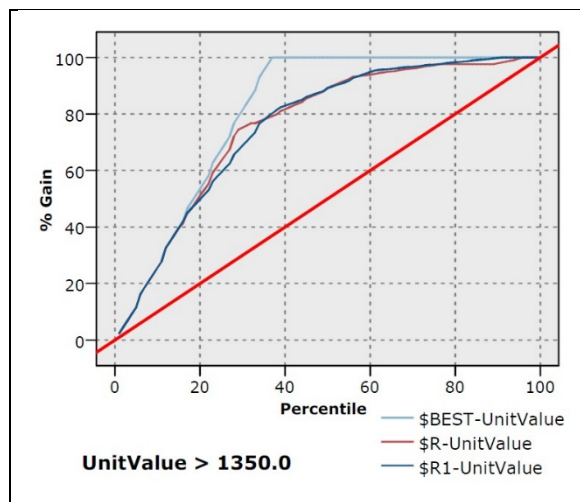
The results of performance analysis of C&RT and CHAID algorithms are given in Fig. 12.

Comparing \$R-UnitValue with UnitValue	
Minimum Error	-250
Maximum Error	444
Mean Error	0,051
Mean Absolute Error	72,22
Standard Deviation	101,951
Linear Correlation	0,937
Occurrences	118
Comparing \$R1-UnitValue with UnitValue	
Minimum Error	-400
Maximum Error	500
Mean Error	-0,093
Mean Absolute Error	97,619
Standard Deviation	136,22
Linear Correlation	0,884
Occurrences	118

**Fig. 12.** The performance analysis of C&RT and CHAID algorithms (developed by the authors).

According to the mean absolute error and standard deviation value in the CRT model, it is considered to give good estimation results when the linear correlation showing the prediction success of the developed model is considered. The analysis has shown that the success rate is higher in the resulting C&RT algorithm. In C&RT (*R*), the mean absolute error value is 72.22, while in CHAID (*R1*) this value is 97.62.

Graphical comparison of data mining modelling success is given in Fig. 13.



**Fig. 13.** C&RT (*R*) and CHAID (*R1*) comparison chart (developed by the authors).

As can be seen in Fig. 10, although the C&RT model shows the best unit value line, it can be seen that the model performance results are close to each other, and both decision trees give good estimation results.

## CONCLUSION

The aim of the study has been to examine the effects of the properties of the houses on the value of the real estate using data mining methods. The research has been carried out in Karabük and the variables determined according to the opinions of real estate experts. The data have been taken randomly from the real estate agents.

In the model, only flats have been used under the name of real estate and these apartments were completed between 1983 and 2017. In data mining, CHAID and C&RT algorithms have been used. When model performance has been investigated, maximum absolute error rates and graphical performance have been examined, and it has been evaluated that both algorithm estimation results can be used.

In the study, while the unit value of the real estate was the result variable, the age of the real estate, distance to the city centre, region's popularity, whether the park and garden, the type of heat protection, interior features, aspects building, location, landscape, heating, area were determined as independent variables. According to the results of the analysis, it has been determined that there is a statistically significant relation between these independent variables and the value of real estate unit.

According to the C&RT algorithm, the most important variables affecting the unit price of real estate are: the area of the housing, the distance to the centre of the city, and the popularity of the region it is located in. On the other hand, according to the CHAID algorithm, popularity, building protection system, the area of the housing has emerged as important variables.

The variables that have the most impact on the unit price have been determined. These are the size of the real estate, the distance to the city centre, the popularity, and the age of the building.

An important detail of the research is that the construction in the city centre does not have privileges such as parks and gardens due to the lack of site identity of buildings near the centre. It has been observed that the distance to the city centre is important, and despite the absence of parking lots and gardens in these buildings, they seem to have a higher unit price.

It has been determined that the houses with the highest unit prices are apartments located up to 700 metres from the city centre. It has also been determined that the houses with the lowest unit price are larger than 55 m<sup>2</sup>, they are not "very good" in popularity, they do not have a mantle and natural gas heating system, and they are more than 3.5 km away from the city centre.

The use of advanced technologies such as statistical modelling and machine learning in real estate valuation and automatic value estimation will be important in determining the real value of the real estate.

The construction sector is directly or indirectly related to approximately 200 sectors, which directly affects both industry and the economy. Today, Turkey is contributing to production in the economic structure and leadership makes investment in the construction sector, has the distinction of being the leading sector in Turkey.

It is needed to create efficient added value in other sectors that provide input to the construction sector and, if necessary, supported by the government.

## REFERENCES

- Açlar, A., & Çağdaş, V. (2002). *Taşınmaz (Gayrimenkul) Değerlemesi*. TMMOB Harita ve Kadastro Mühendisleri Odası, Ankara.
- Afşar, A., Yılmazel, Ö., & Yılmazel, S. (2018). Konut Fiyat Tahmininde Yapay Sinir Ağları Yönteminin Kullanılması. *Uluslararası İktisadi ve İdari İncelemeler Dergisi*, 4–10. <https://doi.org/10.18092/ulikidince.341584>
- Bulut, H., Öner, Y., & İslamoğlu, E. (2015). *Alphanumeric Journal, The Journal of Operations Research, Statistics, Econometrics and Management Information Systems*, 3(2).
- Çakır, P., & Sesli, F. A. (2013). Arsa Vasıflı Taşınmazların Değerine Etki Eden Faktörlerin ve Bu Faktörlerin Önem Sıralarının Belirlenmesi, Determination of Factors Affecting the Real Estate Value of Land Property and the Rank of Importance of these Factors, *2013*(13), 1–16.
- Campos, V., Bueno, C., Brancher, J., Matsunaga, F., and Negrao, R. (2015). Knowledge Discovery Using an Integration of Clustering and Classification to Support Decision-Making in E-commerce. *Advances in Economics and Business*, 3(8), 329–336. <https://doi.org/10.13189/aeb.2015.030805>
- Çetintahra, G. E., & Çubukçu, E. (2011). Çevre Estetiğinin Konut Fiyatlarına Etkisi. *ITU Journal Series A: Architecture, Planning, Design*, 10(1), 3–12.
- Dunham, M. H. (2003). *Data Mining Introductory and Advanced Topics*. Upper Saddle River, NJ: Pearson Education, Inc.
- Ersöz, F. (2017). Veri Madenciliği Teknikleri ve Uygulamaları, 72 Tasarım Ltd.Şti. Yayın Evi.
- Kayral, İ. E. (2017). İstanbul, Ankara ve İzmir Konut Fiyat Değişimlerini Etkileyen Faktörlerin Araştırılması. *Çukurova Üniversitesi İİBF Dergisi*, 21(1), 65–84.
- Larose, D. T. (2005). *Discovering Knowledge in Data*. New Jersey: Wiley Publication.
- Özkan, G., Yalpir, S., & Uygunol, O. (2007). An Investigation on the Price Estimation of Residable Real-Estates by Using ANN and Regression Methods. In: *The 12th International Conference on Applied Stochastic Models and Data Analysis (ASMDA 2007)*. Chania, Crete, Greece.
- Özkurt, H. (2007). Türkiye Ekonomisinde Konut Sektörü: Gelişimi ve Alternatif Finansman Modelleri. *İstanbul Üniversitesi Sosyal Bilimler Dergisi*, 2007(1), 159–173.
- Samar Ali, S., Kaur, R., Ersöz, F. et al. (2018). Journal of Remanufacturing. <https://doi.org/10.1007/s13243-018-0053-y>
- Selim, S. (2008). Determinant of House Price in Turkey: A Hedonic Regression Model, *Doğuş Üniversitesi Dergisi*, 9(1), 65–76.
- Silahtaroglu, G. (2013). *Veri Madenciliği: Kavram ve Algoritmaları*. İstanbul: Papatya Yayıncılık Eğitim.
- Toktaş, M., & Erdoğan, S. (2012). Generating GIS Aided Value Maps for Public Properties Real Estate: Afyankarahisar Example. *Electronic Journal of Map Technologies*, 4(3), 23–38.
- Topçu, M., & Kubat, A. S. (2009). Fen, İ.T.Ü., Enstitüsü, B., & Planlama, B. Konut bölgelerinde Mekansal Yapı- Arazi Değer İlişkisi, (332), 17–26.
- Yayar, R., & Gül, D. (2014). Mersin Kent Merkezinde Konut Piyasası Fiyatlarının Hedonik Tahmini Hedonic Estimation of Housing Market Prices in Mersin City Centre. *Anadolu Üniversitesi Sosyal Bilimler Dergisi*, 14(3), 87–100.
- Yılmazel, S., Afşar, A., & Yılmazel, Ö. (2017). Analysis of Apartments for Sale in Turkey Based on City and Region by Using Big Data Technologies. *The Sakarya Journal of Economics*, 6(3), 1–21.

## AUTHORS' SHORT BIOGRAPHIES

**Filiz Ersoz** is a Professor at Karabük University, Department of Industrial Engineering, Turkey. Dr. Filiz Ersoz holds BSc. degree (1989) in Statistics from Anadolı University, MSc. (1992) and Ph. D. (1998) degrees in Biostatistics from Ankara University. She also received the rank of Associate Professor in Quantitative Decision-Making Methods from the Turkish Inter-University Council (UAK) in October 2011 and the rank of Professor in Industrial Engineering from Karabük University (2017). Her current research deals with decision making, statistics, data mining, simulation and

modelling, statistical quality control, multi-criteria decision-making techniques. Dr. Ersoz has published widely in the field, with more than 90 publications and 5 books. She acted as a Project Manager and Advisor in more than 15 projects. Furthermore, she regularly works as reviewer for about 10 different scientific journals.

**Taner Ersoz** is a Lecturer at Karabük University, Department of Actuary and Risk Management, Turkey. He holds BSc. degree (1988) in Statistics and MSc. degree (1993) from Hacettepe University and Ph. D. degree (2000) in Biostatistics from Ankara University. He has published widely in the field, with more than 50 publications and 3 books on parametric and nonparametric statistics (I-II), statistics with IBM SPSS. He worked as a Project Manager and Advisor in more than 7 projects. He was granted the Certificate of Best Presentation Award from WASET-World Academy of Science, Engineering and Technology in Sweden, 2017.

E-mail: [tanerersoz@karabuk.edu.tr](mailto:tanerersoz@karabuk.edu.tr)

ORCID iD: 0000-0001-5523-5120

Researcher iD: P1020-2018

**Muhammet Soydan**, Civil Engineer, was born in 1994 in Kastamonu. He started his undergraduate studies at the Department of Civil Engineering at Sakarya University in 2012. In 2016, he completed his undergraduate education with modelling of a structure with different carrier systems-graduate degree project. In the same year, he started his postgraduate education at Sakarya University, Department of Civil Engineering – Department of Construction Management. With his master's degree, he worked as a Control Engineer at a private structure audit company in Sakarya. He left to do his military service for a year. During his military service, he continued his graduate studies. In this process, he focused his studies on data mining. He is currently working on real estate valuation and data mining and pursuing his master's degree at the Sakarya University.