A NEW SUBMARKET APPROACH USING DISTANCES TO TRANSIT LINES FOR THE PREDICTION OF REAL ESTATE PRICES

by

Muhittin Tan

M.S., Civil Engineering, Boğaziçi University, 2016

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Master of Science

Graduate Program in Civil Engineering

Boğaziçi University

2019

ACKNOWLEDGEMENTS

I would like to express my deep and sincere gratitude to my thesis supervisor Assoc. Prof. Ilgın Gökaşar for her guidance, support and encouragement throughout the preparation of this thesis.

I would like to thank the members of my thesis committee, Assistant Professor Mehtap Işık and Assistant Professor Gürkan Günay.

I would like to thank also Dr. Onur Şahin and Alperen Timuroğulları for their precious contribution to my thesis.

Finally, I would like to thank to my lovely girlfriend Esra Yalçınsoy, my cool brother Muhammed Kadir Tan, my colleagues and my supervisor Fugen Demirtaş in Boğaziçi University Computer Center Web Unit, my mother, and my father who believe in me all the time, for their sincere and continuous support.

ABSTRACT

A NEW SUBMARKET APPROACH USING DISTANCES TO TRANSIT LINES FOR THE PREDICTION OF REAL ESTATE PRICES

Prediction of a price of a real estate has been one of the trend topics in recent years. There are many studies conducted on both prediction of a value of a real estate or the affecting parameters. In this study, the prediction of a real estate price using submarket near transit lines is studied on two neighbor counties in Istanbul, Beylikduzu and Esenyurt. So, the data should be analyzed into parts to investigate altered dynamics of different districts, which is also called submartket analysis. After the whole data of real estates in these counties are collected and analyzed, the data are divided into three parts (Esenyurt, Beylikduzu and transition zone) and analyzed in order to investigate altered dynamics of different districts. A total of 3487 real estate data with one dependent variable and 13 independent variables collected from aforementioned districts are analyzed with two machine learning (Multiple Linear Regression (MLR) and Spatial Auto Regression (SAR) and one deep learning tool (MultiLayer Perceptron (MLP)). According to the results of the both whole data and submarket analysis, Spatial Auto Regressive model is superior to the others in a metric of R-squared. Moreover, with the submarket analysis, prediction power of all the algorithms (MLR, SAR, and MLP) are significantly increased. Significant independent variables of each model differ from each other so that it can be concluded that submarket analysis in real estate prediction is improving the prediction models and showing different dynamics of each specific district of a county.

ÖZET

GAYRİMENKUL FİYATLARININ TAHMİNİNDE KULLANILMAK İÇİN ULAŞIM HATLARINI KULLANARAK YENİ BİR ALT MARKET OLUŞTURMA YAKLAŞIMI

Gayrimenkul fiyatlarını tahmin etmek, son yıllarda trend konularından biri olmuştur. Hem bir taşınmazın değerini öngören hem de bu fiyatlara etki eden parametreler üzerine birçok çalışma yapılmıştır. Bu çalışmada, İstanbul, Beylikdüzü ve Esenyurt ilçelerinde bulunan iki komşu ilçede, ulaşım hatlarına yakın alt pazar kullanarak gayrimenkul fiyatlarının tahmini analiz edilmiştir. Bu yaklaşıma göre farklı bölgelerin iç dinamiklerini anlamak için veri bölümlere ayrılarak incelenmelidir ki bu yaklaşım alt market olarak adlandırılır. Bu ilçelerdeki gayrimenkullerin tüm verileri toplandıktan ve analiz edildikten sonra, veriler üç bölgeye (Esenyurt, Beylikdüzü ve geçiş bölgesi) ayrılarak farklı bölgelerin değişen dinamiklerini incelemek için analiz edilir. Yukarıda belirtilen bölgelerden bir bağımlı değişken ve 13 bağımsız değişkenden oluşan toplam 3487 gayrimenkul verisi iki makine öğrenmesi (Çoklu Doğrusal Regresyon (MLR) ve Mekansal Otomatik Regresyon (SAR) ve bir derin öğrenme aracı (MLP) ile analiz edilmiştir. Hem tüm verilerin hem de alt pazar analizinin sonuçlarına göre, Spatial Auto Regressive model, R-kare ölçüsündeki bir metrikte diğerlerine göre daha üstün performans göstermiştir. Ayrıca, alt pazar analizi ile, tüm algoritmaların öngörü gücü (MLR, SAR, ve MLP) önemli ölçüde artmıştır. Her modelin önemli bağımsız değişkenleri birbirinden farklıdır, bundan yola çıkarak gayrimenkul fiyat tahmininde alt marketler kullanmanın tahmin modellerini iyileştirdiği ve bu alt marketlere ait farklı dinamikleri gösterdiği sonucuna varılabilir.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS iii
ABSTRACT iv
ÖZET v
LIST OF FIGURES vii
LIST OF TABLES ix
LIST OF SYMBOLS x
LIST OF ACRONYMS/ABBREVIATIONS xi
1. INTRODUCTION 1
1.1. Motivation and Problem Description 1
1.2. Objectives of the Study 2
1.3. Outline of the Thesis 2
2. LITERATURE REVIEW 3
2.1. Regression Methods for Real Estate Price Prediction 3
2.2. Effect of Submarkets to Regression Models 4
2.3. Real Estate Price Prediction Factor Analysis 5
3. THEORY 7
3.1. Linear Regression 7
3.2. Coefficient of Determination 7
3.3. Ordinary Least Squares 8
3.4. Multiple Regression 8
3.5. Spatial Autoregression 10
3.6. Artificial Neural Network 10
4. METHODOLOGY 13
4.1. Data 13
4.2. Methodology 16
4.3. Results 17
5. CONCLUSION 25
REFERENCES 27

LIST OF FIGURES

Figure 3.1. Example Neural Network Setup 11

Figure 4.1. Esenyurt and Beylikduzu Metro and Metrobus Stations (Sahin [1]) 13

Figure 4.2. Study Area in Geographic Information System Software (Sahin [1]) 14

Figure 4.3. MRA Prediction Results of All Data 19

Figure 4.4. MRA Prediction Results of Esenyurt 19

Figure 4.5. MRA Prediction Results of All Beylikduzu 20

Figure 4.6. MRA Prediction Results of Transition Zone 20

Figure 4.7. MLP Prediction Results of All Data 21

Figure 4.8. MLP Prediction Results of Esenyurt 21

Figure 4.9. MLP Prediction Results of Beylikduzu 21

Figure 4.10. MLP Prediction Results of Transition Zone 22

Figure 4.11. SAR Analysis Residuals for All Data 22

Figure 4.12. SAR Analysis Residuals for Esenyurt 23

Figure 4.13. SAR Analysis Residuals for Beylikduzu 23

Figure 4.14. SAR Analysis Residuals for Transition Zone 24

LIST OF TABLES

Table 4.1. Descriptive Statistics of All Data 15

Table 4.2. R squared scores for different submarkets and different regression models 17

Table 4.3. MLP and SAR analysis results for all data 18

LIST OF SYMBOLS

R^2	Coefficient of determination
X	Explanatory variable
VIF	Variance inflation factor
Y_i	Dependent variable
α	Intercept of the model
β	Coefficient of an explanatory variable
ε	Error term
ω	Weight matrix

LIST OF ACRONYMS/ABBREVIATIONS

ANN	Artificial neural network
BP	Back propagation
MAPE	Mean absolute percentage error
MLP	Multilayer Perceptron
MRA	Multiple Regression Analysis
OLS	Ordinary Least Squares
SAR	Spatial Autoregression

1. INTRODUCTION

1.1. Motivation and Problem Description

Real estate market has lots of restrictions and specifications. Land is limited, properties of the land and surrounding lands are very diverse and structure features also differs a lot. So, putting a price label on a real estate is a hard job to accomplish and there is an expertise field for this. Those experts are called appraisers. According to 2019 1st Quarter Analysis of Turkish Appraisers Association, there are 7,980 real estate appraisers who prepared 206,697 appraisal reports. 60% of those appraisal reports are for residential units. Considering minimum price of a report as 100 dollars (in third quarter of 2019) [2], it makes an economy of more than 12 million dollars for one quarter.

Correct appraisal has a significant importance in mortgage credits to prevent banks from lending more. But, are the appraisals accurate? 25 years of sales data from 1984 to 2010 is analyzed by Cannon and Cole and they found that appraisal prices are 12% up or down by the market prices. They also found that appraisals lag the accurate prices. In the hot market prices are below and in the cold markets above the real market prices [3]. Also, according to Private Real Estate Report by MSCI [4] weighted average absolute difference of the reports and the market prices happened to be 9% in 2018. MSCI publishes the same report annually for more than a decade.

Turkish Banking Regulation and Supervision Agency published a decree for protecting banks from inaccurate appraisals or maybe short-term price fluctuations and prevented banks from lending more than 75% value of the residential units [5]. This restriction also protects appraisers because it provides 25% tolerance. Appraisal also protects borrower from paying more than the property value because they cannot borrow if the estimated price is lower than the demanded credit.

1.2. Objectives of the Study

The main objective of this study is to find a proper technique for creating submarkets for increasing real estate price prediction model performances by finding answers to following questions:

- Which regression method among multiple regression analysis (MRA), Spatial Autoregression (SAR) and Artificial Neural Network (ANN) perform better for real estate price predictions using Beylikduzu and Esenyurt residential unit data?
- Does introduction of submarkets affect the performance of regression models?
- Is the submarket effect the same for regression methods MRA, SAR and ANN?
- Which submarket creation method performs better, using governmental borders or distance to transportation lines?

1.3. Outline of the Thesis

- Chapter 2 provides information about previous studies where regression methods used for real estate price predictions. Those methods include MRA, ANN and SAR.
- In Chapter 3, theory of hedonic methods including MRA and SAR and theory of ANN is provided.
- Chapter 4 analyzes the available data and the applied methodology is provided. Then the results are given and discussed.
- Chapter 5 discusses the conclusions of the study and transferability of the methods to other studies.

2. LITERATURE REVIEW

Hedonic method is quite popular for real estate price predictions [6–10]. Haas [11] was the first to apply the hedonic method [12]. Hedonic models generally use MRA or SAR. Although popularity of hedonic pricing methods in real estate sector, according to Do and Grudnitski [13], Tay and Ho [14], Lai Pi-ying [15], Nguyen and Cripps [16] and Selim [17] ANN performs better than multiple regression analysis for appraisals. Nguyen and Cripps [16] also compared MRA and ANN with different data sizes and found that ANN performs better for moderate and large data sizes. Success of ANN is considered to be related to its performance with existence of outliers and non-linearities in the factors. On the other hand, they might hinder the performance of hedonic price models [18, 19]. According to Steven and Albert [20] hedonic models will always be exposed to sampling error simply because they predict means from a big group of samples.

2.1. Regression Methods for Real Estate Price Prediction

ANN requires expertise, because there are no rules for calculating the optimum network structure, optimum data size and the best factors to use [18, 21]. So, when the literature is analyzed, researchers devised their own unique networks based on lack of data or some priori research results. Do [13] used 163 single-family houses sell data in southwestern part of San Diego to apply ANN. His input variables are all internal. There is no external variable like location or closeness to transportation. Do explains this situation as the residential unit subject to data are all somewhat uniform with respect to locational properties like population density, ease of access to transportation and major public facilities. He used 58 data for training and remaining 105 residential unit data for testing. His application of mean absolute percentage error (MAPE) found to be 6.9. He also applied MRA with the same data and it resulted with MAPE of 11.25. These are very good results considering prices are changing from 105.000 dollars to 288.000 dollars. This might be due to very low locational differences as Do suggests.

Lai Pi-ying [15] applied ANN to Kaohsiung city using a dataset of 2471 residential units. She used 70 percent of this data as training dataset and remaining as test dataset. In this dataset location is evaluated in a special way. Since she knows the area well, she divided the city into two groups of districts. Also, she again divided the relative position of the land into two. One group for good positions namely; close to or in the street, in corner and another group for bad positions namely; others. Another interesting variable she used is road width. She thinks that it is related to the residential unit prices. Her ANN application MAPE found to be 19.02.

Nguyen [16] used four factors as: square footage, number of bedrooms, number of bathrooms and building age. Although other factors exist, only these factors are used due to the lack of data. The importance of the work does not come from the accuracy of the created model, it comes from a deep comparative analysis of ANN and MRA models for different number of datasets ranging from 306 to 3706. It is found that although the sample sizes differ a lot, MRA's performance is almost constant. On the other hand, performance of the ANN model gets better as the sample size increases.

Dubin [22] compared MRA with spatial regression models; namely SAR, best linear unbiased prediction, conditional autoregressive model. Spatial models found to be performed better by providing the minimum mean squared error.

2.2. Effect of Submarkets to Regression Models

Typically, a submarket is defined as a set of dwellings that are reasonably close substitutes for one another, but relatively poor substitutes for dwellings in other submarkets [23]. Dividing residential units into submarkets is proposed by many researchers; [24–27]

Bourassa [25] compared different methods for defining submarkets. These methods include the classifications by real estate agents and analytical method. For the analytical method principal component analysis used to identify the relevant factors and then K means clustering technique is used to create submarkets. According to this

research creating submarkets improved the accuracy of predictions than the overall market equation. Also, K means clustering technique weighted mean square error is found to be significantly lower than others.

Bourassa [24] compared eight different techniques for submarket detection. Among those; four of them are geostatistical models, two of them are ordinary least squares (OLS) methods and last two are lattice methods. Results of this research suggests that the geostatistical methods perform better than the simple OLS, but OLS model with submarkets is almost as good as geostatistical methods. Adding submarkets dummy variables to simple OLS model increases the ex-sample predictions percentage within 10% of the real sale price from 39.8

Basu [28] used 5320 transactions between 1991 and 1993 in Dallas, Texas. 8 submarkets created and the spatial dependence of factors is analyzed. Substantial spatial autocorrelation is found. But, controlling the age and size is eliminated spatial autocorrelation in some markets and decreased significantly in others.

Can [26] used data for the Columbus, Ohio MSA and defined neighborhood quality index using variables that are considered to be related to neighborhood effects within the study area. It is found that the type of exterior, the lot size of the house, the availability of a 2-car garage and the presence of a basement are positively correlated with neighborhood quality index. This could be due to the locational governmental regulations which enforce similar structures in the same neighborhood.

2.3. Real Estate Price Prediction Factor Analysis

For Turkish real estate market Ozsoy and Sahin [29] and Selim [6] studied to find the significant factors that have an effect on the prices of residential unit prices.

Selim [6] made a comprehensive analysis for Turkey housing market to find the determinant factors for residential unit rent prices. Hedonic model applied using 2004 Household Budget Survey Data for the analysis and it is found that water system,

pool, type of house, number of rooms, house size, type of the building and locational characteristic are the most significant parameters.

Ozsoy and Sahin [29] used residential unit data located in Istanbul which is collected from different internet pages of real estate agencies in 2007 Classification and regression tree technique is applied to determine the relevant factors to the residential unit prices. It is found that size, elevator existence, security availability, central heating unit type and view are found to be the most significant factors for housing market in Istanbul. Selim [6] made a comprehensive analysis for Turkey housing market to find the determinant factors for residential unit rent prices. Hedonic model applied using 2004 Household Budget Survey Data for the analysis and it is found that that water system, pool, type of house, number of rooms, house size, type of the building and locational characteristic are the most significant parameters.

This paper contributes to the literature by:

- Analyzing submarket effects using distances to transit lines to predict real estate prices in a very diverse urban area Esenyurt and Beylikduzu districts of Istanbul, Turkey.
- Comparing the prediction resuls of two machine learning (Multiple Linear Regression (MLR) and Spatial Auto Regression (SAR) and one deep learning tool (MultiLayer Perceptron (MLP)).

3. THEORY

In this study various analysis techniques, namely OLS, MRA, SAR and ANN are used. In this section these techniques are briefly described.

3.1. Linear Regression

Revealed preference method is widely used in economics to determine consumer preferences by observing their purchasing behavior. Hedonic regression is a revealed preference method used to discover real estate prices and parameters' relevance to the prices [1,30–33]. It simply evaluates the price of a commodity as weighted combination of commodity related parameters. OLS, MRA and SAR are hedonic regression methods and they estimates the coefficients of the parameters of a linear function by minimizing the sum of the squares of the error term. From a geometric perspective, they produce a straight line through a dataset which has the minimum total squared distances to each data point.

3.2. Coefficient of Determination

R^2 value, also known as coefficient of determination, measures the closeness of the data to the fitted regression line. In other terms, R^2 means explained percent of the given data by the model. So, R^2 value could be between 0 and 1. If R^2 is 0, this means that the regression cannot explain any variance in the data. On the other hand, if R^2 is 1, this means that the regression model explains all the variance in the data, in other words all the data points are located on the regression line. R^2 can be formulated as:

$$R^2 = 1 - \frac{SS_{regression}}{SS_{total}} \tag{3.1}$$

Having a high R^2 does not always mean a good fit. For a good fit, residual plot could be used. The values on the residual plot should be randomly scattered. For example, if residuals are always positive for some data interval, then the model does not have a good fit. This could happen if the model does not have an important independent variable.

3.3. Ordinary Least Squares

The basic analysis used for investigation of a relationship between the dependent variable and one independent variable is called OLS. If there are two or more independent variables for the investigation of the relationship with the dependent variable, it is called MRA. A general form of an OLS model is presented as follows:

$$Y_i = \alpha + \beta X + \varepsilon_i \tag{3.2}$$

where Y_i is the dependent variable, α is the intercept of the model, X is the explanatory variable, β is the coefficient of an explanatory variable and ε_i is the error term. The parameters of the model, α indicates the value where the regression line crosses the y-axis and β indicates the slope of the regression line, respectively.

3.4. Multiple Regression

A general form of MRA model is presented as follows:

$$Y_i = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \ldots + \beta_k X_k + \varepsilon_i \tag{3.3}$$

Where we have k independent variables explaining the variance of Y_i with β_k and X_k values. X_k is the dependent variable with a slope of β_k. MRA adjusts the β_k values by minimizing ε_i.

For MRA to have a good fit, our independent variables X_k must be able to explain the dependent variable Y_i with a linear relationship. Existence of high intercorrelations among independent variables called multicollinearity. Multicollinearity must not exist among independent variables for a good fit. Because when multicollinearity exists among 2 variables, one of them tends to explain a big portion of variance in the dependent variable. So, the other variable ends up with a little variance to explain. In order to discover multicollinearity among variables variance inflation factor (VIF) is used. It measures the impact of collinearity among the variables. VIF for each factor can be defined as:

$$VIF_i = \frac{1}{1 - R_i^2} \tag{3.4}$$

There is not a predefined VIF value for discovery of multicollinearity. VIF values greater than 10 are often considered as existence of multicollinearity. If R^2 and VIF values are high for any of the variables in the model, multicollinearity probably exists. Another way of determining multicollinearity may be removing a factor from the model. If this results in sign change for any of the slope factors β_k, again, multicollinearity may exist.

For linear models, significance of each factor can be determined analyzing p values. p value is the result of the null hypothesis that the coefficient is equal to zero. So, having a small p value means that the factor has significant effect on the dependent variable.

3.5. Spatial Autoregression

The investigation of the location based datapoints is usually more successful when a spatial analysis technique is applied [1, 31, 33]. One of these spatial analysis techniques is called SAR, which utilizes the nearby data points in order to adjust the prediction of the estimates in the model. SAR is also a hedonic regression method and it is a common analysis tool for spatial economics and spatial statistics. It is a special variable correlation analysis method. Getis [34] defined the difference of SAR from other correlation analysis methods as "Whereas correlation statistics were designed to show relationships between or among variables, spatial autocorrelation shows the correlation within variables across georeferenced space." SAR model evaluates output as dependent on the spatially close data. Using SAR models selecting correct tolerance value is important for getting more accurate results.

3.6. Artificial Neural Network

ANN is a computer mock of biological neural networks. It consists connected neurons. There are nine type of ANN setups. These are feedforward, regulatory feedback, radial basis function, recurrent neural network, modular, physical, other types, dynamic and memory networks. Feedforward neural network is historically the first type used. In this type of ANN, the information flows from the input layer to the output layer through any number of hidden layers if exist. If there is no hidden layer in between input and output layers, this network is called single layer perceptron.

If there are at least one hidden layer of neurons, this network is called MLP. In MLP, there must be at least three layers of neurons. MLP is a very popular ANN structure which falls under feedforward ANN. The first layer is the input layer, the last layer is the output layer and in between are hidden layers. Every neuron is connected with every neuron in the adjacent layers with a weight. MLP networks can be constructed with different types of neurons. Neurons receives one or more inputs and calculates weighted sum of inputs to produce output based on activation function which is also called transfer function. This neuron action can be formulated as:

$$z = \sum_{i=1}^{m} \omega_i x_i + bias \tag{3.5}$$

where z is sum of the dot product of input data x_i and weight ω_i which is summed with bias. x_i represents the input data and ω_i is the weight for that input. Then, z is put into activation function $f()$ and we get the output of the neuron.

$$h_{jk} = f\left(z\right) \tag{3.6}$$

where h_{jk} is kth neuron of the jth layer and output of the neuron. Activation functions generally have sigmoid shape.

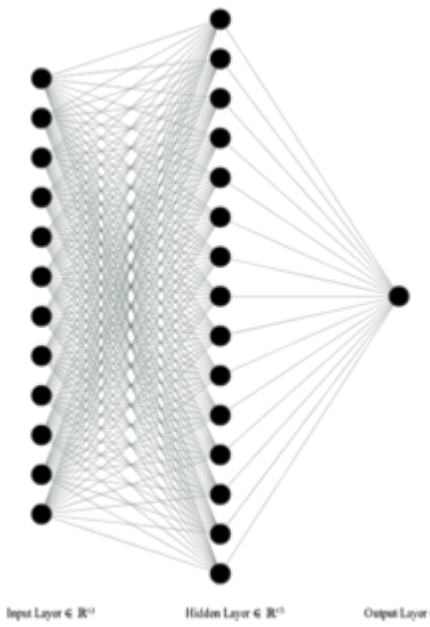

Figure 3.1. Example Neural Network Setup

For a start point neuron weights are assigned randomly. Then the network is fed from the input layer, the output layer will generate a prediction. Then, the correct output of the prediction is given to the network. This process is called back propagation (BP). Based on the correct output, the network adjusts the weights of connections between neurons according to activation functions to reduce the error. This process is called training. After training process, the model can predict given scenarios. The performance of the ANN is dependent upon the network setup and neuron activation functions. Network setup is ambitious. Number of hidden layers and number of neurons in each layer should be selected correctly. There is no rule for selecting the correct number of layers and neurons. This is accomplished by trial and error. An example network setup can be seen in Figure 3.1.

4. METHODOLOGY

4.1. Data

A total of 3,478 residential unit sale price data including the asking prices in the last quarter of 2018 and 2019 are collected from the study area. Residential units are in Esenyurt and Beylikduzu, Istanbul, Turkey. Esenyurt is located in the west side of Istanbul and 30 kilometers away from the city center (See Figure 4.1). Beylikduzu is located south of Esenyurt (Figure 4.1)

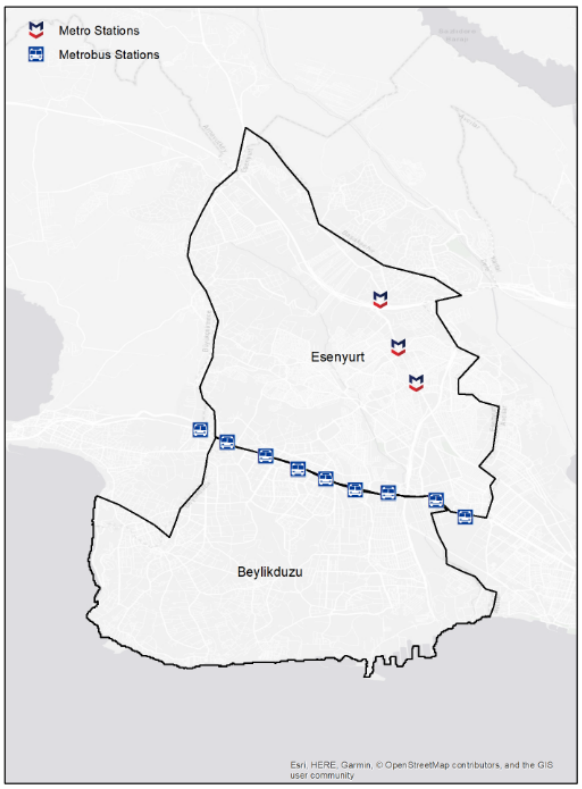

Figure 4.1. Esenyurt and Beylikduzu Metro and Metrobus Stations (Sahin [1])

Istanbul is in the top 20 most crowded cities in the world with a population more than 15 million. According to Turkish Statistical Institute 2018 Report for Address Based Population Registration System Results, Esenyurt is the most crowded district in Istanbul with a population of 891.120 and Beylikduzu is the 22th most crowded

district with a population of 331.525 among 39 districts. Both districts are located in the west side of Istanbul. The Metrobus line goes through E5 highway which is the border between Esenyurt and Beylikduzu. According to Istanbul Municipality 2016 Transportation Report Istanbul Metrobus line have more than 1 million trips per day. Also, there is an ongoing metro construction in Esenyurt which is planned to be finished in 2020 August. There is a subway project for Beylikduzu too, but it is not decided yet. It is observed that new transportation lines could affect residential unit prices negatively by Bohman and Nilsson [35] and Beimer and Maenning [32]. On the other hand, Mulley [31], Paul and Cohen [36] and Dziauddin [33] found that new transportation lines effect the residential unit prices positively. Therefore, the effect of an upcoming project is subjected to change over the space. There are more than 10 shopping malls in Esenyurt. There are less than 10 shopping malls in Beylikduzu and they are close to Esenyurt border. So, there are lots of different powerful factors which may affect the residential unit sale prices in the districts (Figure 4.2).

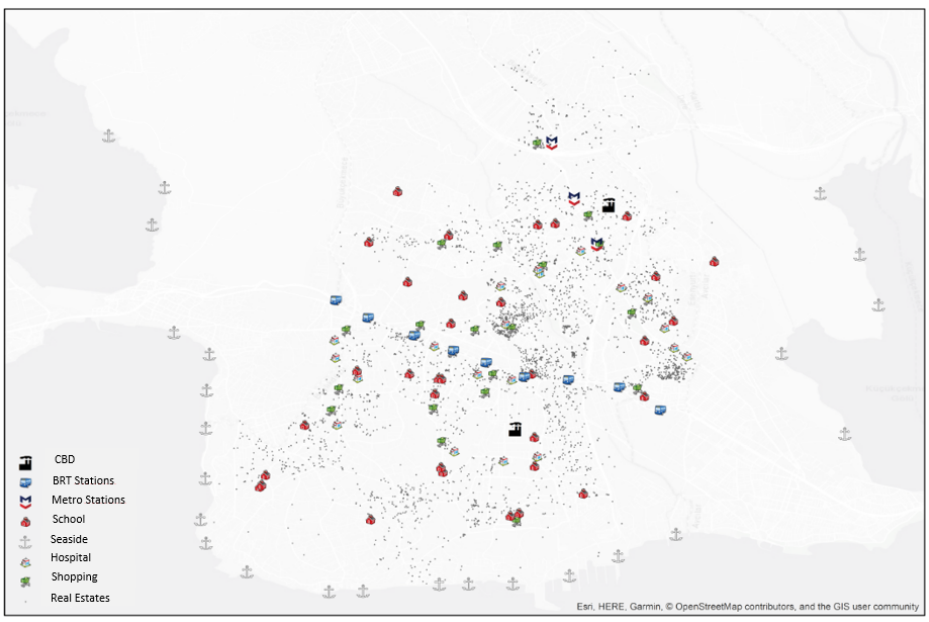

Figure 4.2. Study Area in Geographic Information System Software (Sahin [1])

A transition zone between Esenyurt and Beylikduzu is constructed. This zone includes neighborhoods of both Esenyurt and Beylikduzu districts which are adjacent to the Metrobus. Those neighbourhoods are Adnan Kahveci, Barış, Beylikdüzü,

Büyükşehir, Cumhuriyet, Yakuplu, Zafer, Namık Kemal, Mehterçeşme, Mehmet Akif Ersoy, Güzelyurt and Alevler.

Table 4.1. Descriptive Statistics of All Data

	Minimum	**Maximum**	**Mean**	**Standard Deviation**
Area	30	450	120	41
Value	50000	4000000	322733	193548
Rooms	1	10	3.37	1.01
Age	1	29	5.12	5.30
Floor	0	35	4.07	4.45
Parking	0	1	0.34	0.47
DistShopping	8	5343	1077	669
DistHospital	1	5752	1217	951
DistBRT	319	9445	3277	1654
DistSchool	1	5172	1001	708
DistCBD	33	7441	2950	1154
DistSea	53	7684	4526	1152
DistMetro	9	11328	4050	2547

According to the descriptive analysis of the data (Table 2), mean area of residential units is 120m2 differs from small studio houses (30m2) to big residential units (450 m2). Pricing of houses also has big difference of scale; minimum value is 1/80 of the maximum value having 322.733 averages. Rooms are converging to standard of 2+1 and 3+1. The average residential unit age is 5.12 years which means these districts have relatively younger residential units. 35% of the residential units have parking lot. When the distances are compared, residential units are not so far from hospital, shopping centers, school. (Means ¡ 1500) Most of the data is collected far from metro and sea which are north and south limits of data set, relatively which means the data points are located in-between of these boundaries.

4.2. Methodology

A general form of the used MRA model for our data could be defined as:

$$P = f(Area, Rooms, Age, Floor, Parking, Credibility, DistShopping, \\ DistHospital, DistBRT, DistSchool, DistCBD, DistSea, DistMetro) + \varepsilon \tag{4.1}$$

where P indicates the price of the real estate property, "Area" indicates for area of the residential unit in terms of square meters, "Rooms" indicates the number of rooms in the residential unit, "Age" indicates the number of years since the residential unit is constructed, "Floor" indicates for the floor level of the residential unit, "Parking" indicates the existence of a parking lot available for the real estate property, "Credit Viability" indicates that the condition of a real estate property's credit viability for transaction process, "DistShopping" indicates the distance to the closest shopping center in terms of meters, "DistHospital" indicates the distance to the closest hospital in terms of meters, "DistBRT" indicates the distance to the closest bus rapid transport station in terms of meters, "DistSchool" indicates the distance to the closest school in terms of meters, "DistCBD" indicates the distance to the closest business center in terms of meters, "DistSea" indicates the distance to seaside in terms of meters, "DistMetro" indicates the distance to the closest metro station in terms of meters and ε indicates the error term.

All the data and submarkets are analyzed using ANN, SAR and MRA. Three different submarkets are defined. Beylikduzu and Esenyurt data considered to be two submarkets and then a transition zone is defined and analyzed separately. Python 3 coding language is used for ANN analysis. To create an ANN scikit-learn library MLPRegressor function is used [37]. Using MLPRegressor function all the network settings can be set. Relu activation function and Adam solver is used. For all of the methods, training dataset and testing dataset randomly created. 75% of the data is used to train the network and 25% of the data is used as testing. Only one hidden layer is used but different number of neurons for this hidden layer is tested. Model is

tested with 32, 64, 128 and 256 neurons in the hidden layer and 256 neurons found to perform better using all data so we stick with 256 neurons for submarket analyses.

Model performances are measured by comparing R^2 values. Models with higher coefficient of determination, in other words, larger R^2 values are considered to be superior. Submarkets are created based on districts and distances of data points to the BRT line. Those submarkets are considered to increase models' performances.

For linear regression models, factors with p values smaller than 0.05 considered to be not significant. Significant factors of different submarkets and different linear regression models, namely SAR and MRA, are compared.

R coding language is used for both SAR model. For SAR model, lagsarlm function is used with a very low tolerance for detecting linear dependencies value (1e-17). For MRA analysis ols function of python 3 software language is used.

4.3. Results

R^2 scores of ANN, SAR and MRA for three submarkets and all data are compared (Table 4.2).

Table 4.2. R squared scores for different submarkets and different regression models

	ANN	**SAR**	**MRA**
All Data	0.40	0.47	0.44
Beylikduzu	0.40	0.46	0.44
Esenyurt	0.48	0.53	0.49
Transition Zone	0.50	0.57	0.53

ANN performed worst for all datasets. SAR on the other hand performed the best for all cases. According to MRA analysis Area, Floor, Parking, Credibility, Dist-Shopping, DistBRT and DistCBD found to be the most significant while Rooms found

to be the least significant parameter for all datasets. According to SAR results Area, Parking, Credibility and DistHospital found to be the most significant factors. Area, Parking and Credibility are the common significant factors for SAR and MRA.

Table 4.3. MLP and SAR analysis results for all data

	MLP		**SAR**	
Parameter	**Coefficient**	**P**	**Coefficient**	**P**
(Intercept)	-95100.536	0	-129820	2.2E-16
Area	2547.075	0	2380.4	2.2E-16
Rooms	-10410.606	0.016	-6291.1	0.1062328
Age	1003.406	0.079	683.31	0.2181965
Floor	1489.841	0.015	914.14	0.1171554
Parking1	95487.033	0	80660	2.2E-16
Credibility1	36958.618	0	36966	9.432E-08
DistShopping	-37.313	0	-18.796	0.0002827
DistHospital	8.525	0.101	4.8128	0.1952788
DistBRT	7.831	0.003	2.3127	0.1966065
DistSchool	42.604	0	27.098	1.914E-08
DistCBD	-6.86	0.01	-6.1461	0.014221
DistSea	1.68	0.51	0.62113	0.5478568
DistMetro	16.097	0	6.8713	0.0002349

When new markets introduced ANN, SAR and MRA models' relative performances did not changed. So, when establishing submarkets, only one of the methods could be used for assessing established submarkets' relative performances.

All analysis methods had shown the best performance using the data collected from the transition zone. Also, except for Beylikduzu submarket, DistBRT found to be a significant factor for MRA. This could imply that popular transportation lines establish submarkets around.

In prediction / real value illustrations, the more data gathered around x=y line, the more quality of prediction model. Because of the fact that there are too many data points, it is hard to observe the quality of the model. In first look, all the data points at dense areas in plot are around x=y line. To determine the quality of a model thickness of areas that data points are densely gathered will be used because more thickness means that there are data points drifting apart from the line. Thus, according to the full data plot (Figure 4.3), the model does not allow outlier values as much as the other models do but it fails to collect the whole predictions around the x=y line as much as the other models achieve.

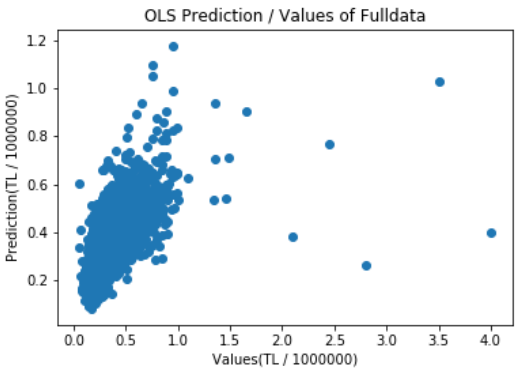

Figure 4.3. MRA Prediction Results of All Data

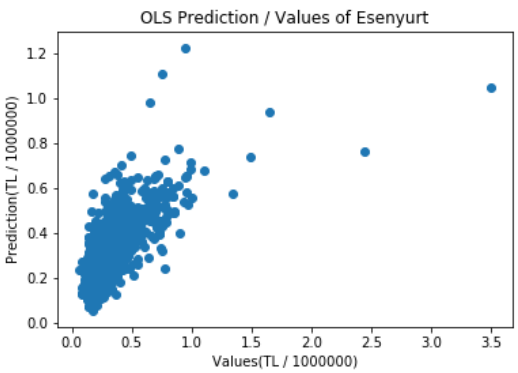

Figure 4.4. MRA Prediction Results of Esenyurt

So, market segmentation of the data is improved models for Esenyurt (Figure 4.4). Because of the fact that the max value of transition zone is 200.000, the thickness

of it looks large while it is almost half of the full data model. So, it can be said that segmentation is improved model in also Transition Zone (Figure 4.6).

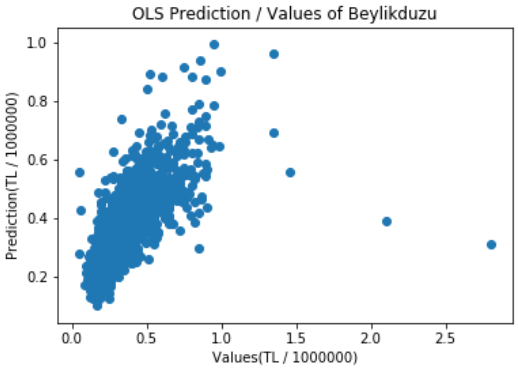

Figure 4.5. MRA Prediction Results of All Beylikduzu

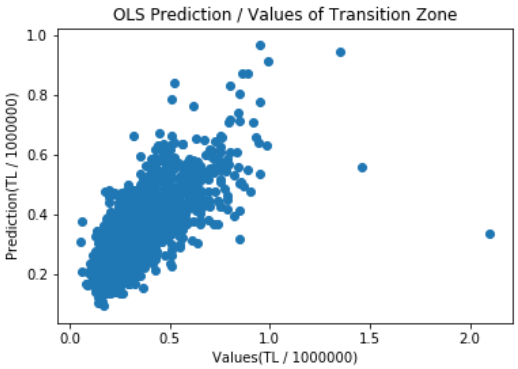

Figure 4.6. MRA Prediction Results of Transition Zone

The explanations of the MLP prediction/ real value plots are as same as the explanation of OLS results. It can be seen that full data analysis does not allow outlier except one case where the others have some outliers. (Figure 4.7 4.8 4.9 4.10) However, according to the R-square values of the models, the impact of outliers is compensated by the values collected around x=y line. Thus, it can be concluded that the segmentation of market explains altered pricing dynamics in different districts better than using the whole dataset in Esenyurt and Transition zone. (Table :4.2)

In order to validate SAR model has a good fit, residual figures are analyzed. Because SAR performed better for all markets, residuals of SAR analysis price predictions

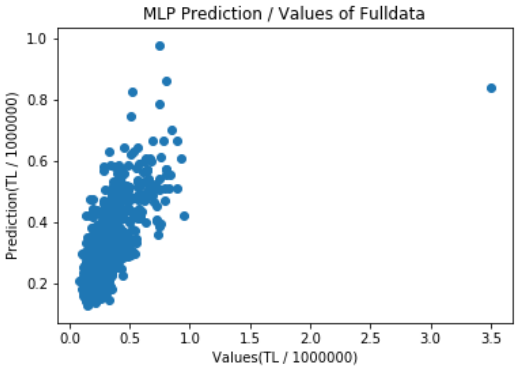

Figure 4.7. MLP Prediction Results of All Data

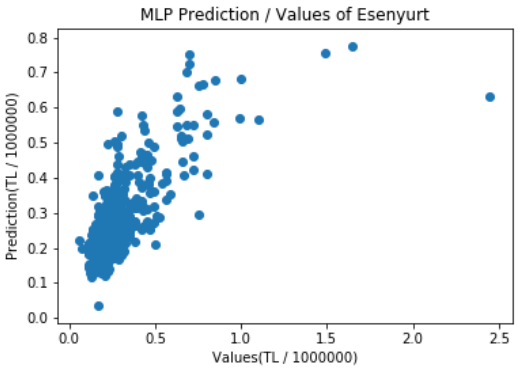

Figure 4.8. MLP Prediction Results of Esenyurt

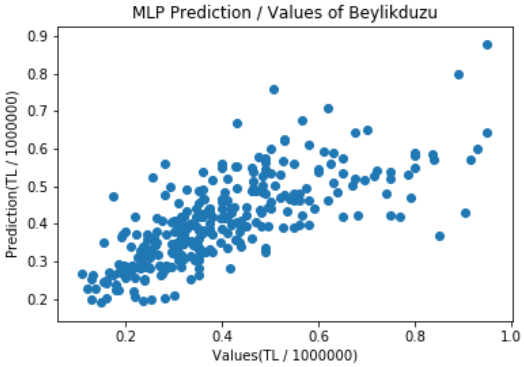

Figure 4.9. MLP Prediction Results of Beylikduzu

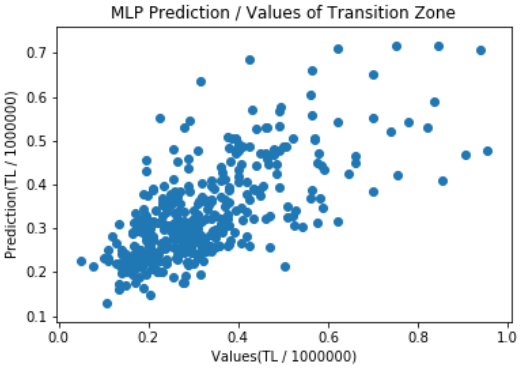

Figure 4.10. MLP Prediction Results of Transition Zone

are given in Figures 4.11 4.13 4.12 4.14.

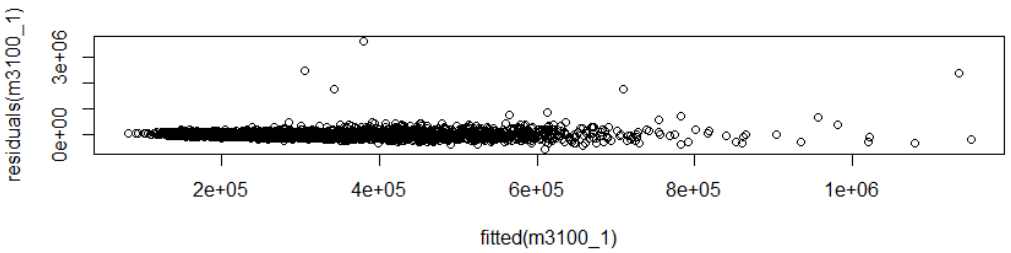

Figure 4.11. SAR Analysis Residuals for All Data

On the plots above (Figure 4.11 4.13 4.12 4.14), x axis stands for real prices of data and y axis is the error. According to the results of SAR analysis, when market is segmented into districts it can be noticed the fact that the residual errors are gathered around 0 in all the price levels except Beylikduzu. In Beylikduzu district, after 114.000$ (in third quarter of 2019) price level, the error become larger. However, in the SAR residual plot of full data, predictor model is having errors in both positive and negative direction. Even if it does not look like the data gathered around of y=0, the max error limit of full data is 1.5 times of the others. Thus, the result that the market segmentation using distances provide more stable predictions except outlier cases could be reached.

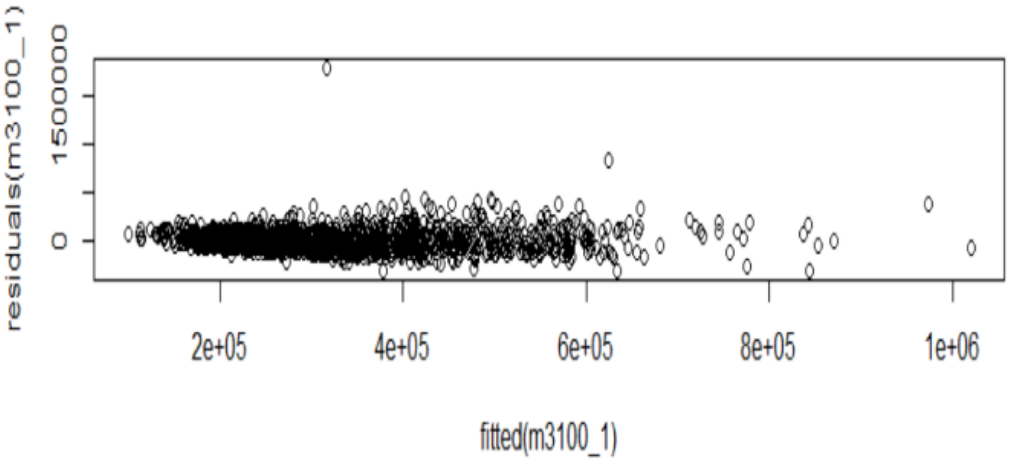

Figure 4.12. SAR Analysis Residuals for Esenyurt

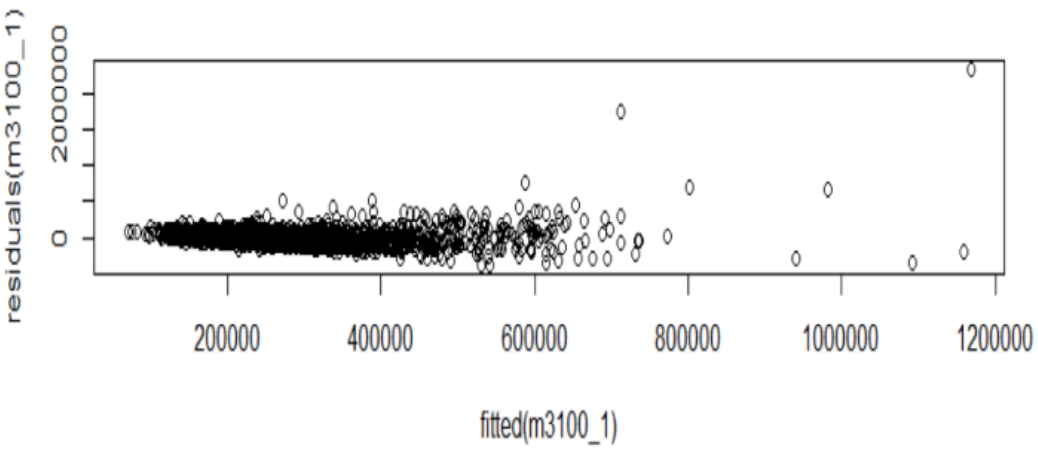

Figure 4.13. SAR Analysis Residuals for Beylikduzu

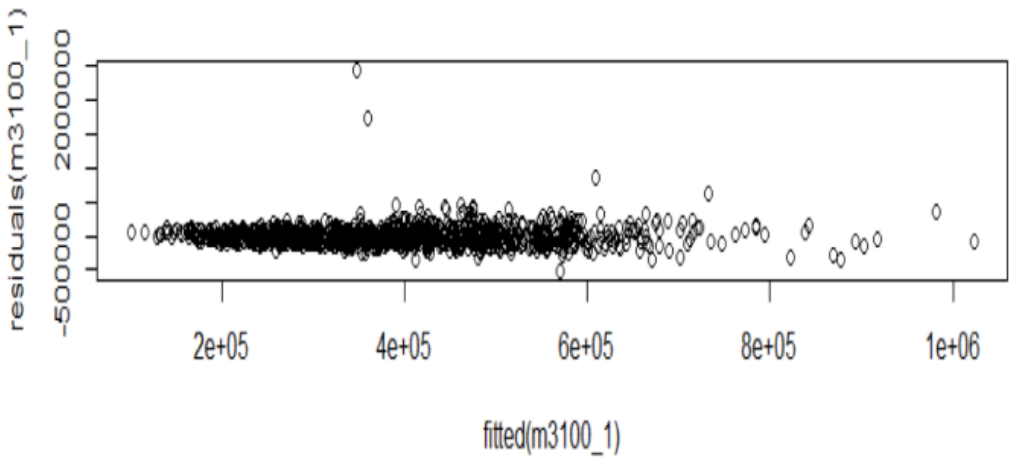

Figure 4.14. SAR Analysis Residuals for Transition Zone

5. CONCLUSION

This study is important for comparing prediction performances of ANN, SAR and MRA methods and increasing real estate price prediction model performances through a new sub market approach. Real estate price prediction models may be applied for validating real estate appraisal reports which are used for mortgage credits to prevent banks from lending more. Also, these models could be used for analyzing the real estate market.

Two neighbor counties of Istanbul, Esenyurt and Beylikduzu residential unit data is collected for analysis. A total of 3487 residential unit data with sale prices are collected from both regions. In addition to internal properties, distances to various key locations also added to the data factors. These key locations are shopping centers, hospitals, bus rapid transport (Metrobus), schools, business centers, sea and metro.

Machine learning tools are good predictors for estimating real estate prices. ANN, SAR and MRA regression tools are used for predicting residential unit prices. Then, performances of these models compared using R^2 scores. After this comparison, submarkets are introduced and their effect on regression R^2 scores are analyzed. Beylikduzu and Esenyurt submarkets are constructed using governmental borders and transition zone submarket constructed using distance to bus rapid transportation (Metrobus, DistBRT) factor. Introducing submarkets found to increase the R^2 scores of all regression models significantly. Furthermore, establishing a submarket using distance to major transportation line factor, namely DistBRT, performed better than submarkets established using governmental borders. This implies that DistBRT could be used to establish submarkets for better regression models. Also, it is found that SAR performed the best, while ANN performed the worst for predicting residential unit prices.

MRA analysis significant factors found to be Area, Floor, Parking, Credibility, DistShopping, DistBRT and DistCBD. SAR analysis significant factors found to be Area, Parking, Credibility and DistHospital. For both of the linear regression models

Area, Parking and Credibility found to be the common significant factors.

Transferability of the results to other residential unit price prediction tools can be through creation of submarkets considering distance to transportation lines for regression models. Such a tool would increase existing prediction model performances.

Collected data cannot explain the residential units completely. Interior design, sold price, for sale duration, roof type, window type, view and availability of daylight are also very important factors affecting the residential unit prices. So, prediction accuracy is limited for the available data. If more detailed data could be collected, better price predictions could be accomplished.

REFERENCES

1. Sahin, O., *Investigation of the effects of transportation investments on real estate prices: case study Beylikduzu and Esenyurt*, PhD Thesis, 2019.

2. of Turkey, C. M. B., *Capital Market Board Decision on Appraisal Report Prices*, 2018, `http://www.resmigazete.gov.tr/eskiler/2018/12/20181231M4-29.pdf`, accessed in August 2019.

3. Cannon, S. and R. Cole, "How Accurate Are Commercial Real Estate Appraisals? Evidence from 25 Years of NCREIF Sales Data", *The Journal of Portfolio Management*, 2011.

4. MSCI, *Private Real Estate: Valuation and Sale Price Comparison*, 2019, `https://bit.ly/2ZcxByl`, accessed in August 2019.

5. Regulation, T. B. and S. A. Decree, *Banking Regulation and Supervision Agency of Turkey*, 2010, `http://www.resmigazete.gov.tr/eskiler/2010/12/20101218-15.htm`, accessed in August 2019.

6. Selim, S., "Determinants of House Prices in Turkey: A Hedonic Regression Model", *Doğuş Üniversitesi Dergisi*, Vol. 9, No. 1, 2011.

7. Kaya, A. and M. Atan, "Determination of the Factors That Affect House Prices in Turkey by Using Hedonic Pricing Model", *Journal of Business Economics and Finance*, 2014.

8. Baen, J. and R. Guttery, "The Coming Downsizing of Real Estate: Implications of Technology", *Journal of Real Estate Portfolio Management*, Vol. 3, No. 1, pp. 1–18, 1997, `https://aresjournals.org/doi/abs/10.5555/repm.3.1.a44378x504x2uj4l`.

9. Kershaw, P., R. Kooymans and P. Rossini, "Micro-Computer Based Real Estate Decision Making and Information Management-An Integrated Approach", *2nd Australasian Real Estate Educators Conference, Adelaide*, 1992.

10. H., D. J. and R. E. Radigan, *Computer-Assisted Real Estate Appraisal: A Tool for the Practicing Appraiser*, The Appraisal Journal, 1996.

11. Haas, G. C., "Sales Prices as a Basis for Farm Land Appraisal", *Technical Bulletin No*, Vol. 9, 1922.

12. Colwell, P. and G. Dilmore, "Who Was First? An Examination of an Early Hedonic Study", *Land Economics*, p. 2307/3147070, 1999.

13. Do, A. Q. and G. Grudnitski, "A neural network approach to residential property appraisal", *The Real Estate Appraiser*, Vol. 58, pp. 38–45, 1992.

14. Tay, D. P. and D. K. Ho, "Artificial intelligence and the mass appraisal of residential apartments", *Journal of Property Valuation and Investment*, Vol. 10, No. 2, pp. 525–540, 1991.

15. P., L., "Analysis of the Mass Appraisal Model by Using Artificial Neural Network in Kaohsiung City", *Journal of Modern Accounting and Auditing*, 2011.

16. Nguyen, N. and A. Cripps, "Predicting Housing Value: A Comparison of Multiple Regression Analysis and Artificial Neural Networks", *Journal of Real Estate Research*, 2001.

17. Selim, H., "Determinants of house prices in Turkey: Hedonic regression versus artificial neural network", *Expert Systems with Applications*, 2008.

18. M., L. M., E. M. Worzala and A. Silva, "High-tech Valuation: Should Artificial Neural Networks Bypass the Human Valuer?", *Journal of Property Valuation and Investment*, 1997.

19. C., O. and J. Howard, "Estimation Realisation Price (ERP) by Neural Networks: Forecasting Commercial Property Values", *Journal of Property Valuation & Investment*, 1998.

20. P., S. and B. Albert, "Neural Network Hedonic Pricing Models in Mass Real Estate Appraisal", *Journal of Real Estate Research, American Real Estate Society*, 2009.

21. Bishop, C. M., *Neural Networks for Pattern Recognition*, Oxford University Press, 1995.

22. Dubin, R., R. K. Pace and T. Thibodeau, "Spatial Autoregression Techniques for Real Estate Data", *Journal of Real Estate Literature*, 1999.

23. Grigsby, W., M. Baratz, G. Galster and D. Maclennan, "The Dynamics of Neighborhood Change and Decline", *Progress in Planning*, 1987.

24. Bourassa, S. C., E. Cantoni and M. J. Hoesli, "Spatial Dependence, Housing Submarkets, and House Price Prediction", *Real Estate Finan Econ*, Vol. 35, p. 143, 2007.

25. C., B. S., F. Hamelink, M. Hoesli and B. D. Macgregor, "Defining housing submarkets", *Journal of Housing Economics*, Vol. 8, 1999.

26. Can, "A", *The Measurement of Neighborhood Dynamics in Urban House Prices, Economic Geography*, 1992.

27. Can, A. and I. Megbolugbe, "Spatial Dependence and House Price Index Construction", *The Journal of Real Estate Finance and Economics*, Vol. 14, p. 203, 1997.

28. Basu, S. and T. G. Thibodeau, "Analysis of Spatial Autocorrelation in House Prices", *The Journal of Real Estate Finance and Economics*, Vol. 17, pp. 61–85, 1998.

29. Özsoy, O. and H. Şahin, "Housing price determinants in Istanbul, Turkey: An application of the classification and regression tree model", *International Journal of Housing Markets and Analysis*, Vol. 2, No. 10., p. 1108/17538270910963090., 2009.

30. Seo, K., A. Golub and M. Kuby, "Combined impacts of highways and light rail transit on residential property values: a spatial hedonic price model for Phoenix", *Arizona. Journal of Transport Geography, 2014*, Vol. 41, 2014.

31. Mulley, C., L. Ma, G. Clifton, B. Yen and M. Burke, *Residential property value impacts of proximity to transport infrastructure: An investigation of bus rapid transit and heavy rail networks in Brisbane*, Australia, 2016.

32. Maennig, W. and W. Beimer, *Noise effects and real estate prices: A simultaneous analysis of different noise sources*, Transportation Research Part D, 2017.

33. Dziauddin, M. F., *Estimating land value uplift around light rail transit stations in Greater Kuala Lumpur: An empirical study based on geographically weighted regression (GWR)*, Research in Transportation Economics, 2019.

34. Getis, A., "A history of the concept of spatial autocorrelation: A geographer's perspective", *Geographical Analysis*, Vol. 40, No. 3, pp. 297–309, 2008.

35. Bohman, H. and D. Nilsson, "The impact of regional commuter trains on property values: Price segments and income", *Journal of Transport Geography*, Vol. 56, 2016.

36. Cohen, J. P. and C. M. Paul, "The impacts of transportation infrastructure on property values: A higher-order spatial econometrics approach", *Journal of Regional Science*, 2007.

37. E., P., "Scikit-learn: Machine Learning in Python", *Journal of Machine Learning Research*, Vol. 12, pp. 2825–2830, 2011.