Gazi University Journal of Science
GU J Sci
29(3):627-633 (2016)

The Modelling of Residential Sales Prices with Kriging Using Different Distance Metrics in Different Correlation Functions

Semra Erpolat Taşabat[1,♠], Olgun Aydın[2]

[1,2] *Mimar Sinan University, Department of Statistics, Istanbul, Turkey,34380*

Received: 10/01/2016 Accepted: 28/07/2016

ABSTRACT

The modelling and estimation of sales prices based on economical conditions are important for housing sector especially in developing countries. Analysts are focused on the subject to analyze price movements and estimate the future trend of the sales prices for housing sector. In this study, we tried to generate a robust and efficient model related to the subject. Firstly, we investigated economic variables affecting housing sales prices and then created a kriging model for housing sales prices in Dubai. To determine a better correlation function structure for creating a powerful kriging model, we used Euclidean and Canberra distances for both Exponential and Gaussian correlation functions. Simulation studies were applied to obtain optimum correlation functions. To detect the normality of response values, Focused Information Criteria(FIC) was used. By using cross validation criteria, we selected the best performed correlation function with the best performed distance metric for the kriging model.

Keywords:*Kriging, Euclidean Distance, Canberra Distance, Maximum Likelihood Estimation, Housing Price, Economical Condition, Dubai Housing Market.*

1.INTRODUCTION

Over a period of half a century the city state of Dubai has progressed from pre-industrial to industrial to post-industrial status. Change is evident in the economic, social and cultural characteristics of the city and, most visibly, in the scale, pace and nature of real estate development[1]. Dubai has developed into a city of regional importance, with a planned objective of becoming a city of significance within the global urban economic system. Development of the city has been energized by a synergistic relationship between global and local forces embedded within a particular historical and geographical contex[2].

Because of that, analyzing, making prediction for housing sector is important especially for investors. Hepsen et al(2011) has analyzed housing sales prices in Dubai and created prediction model about future trends of the prices[3]. Since real estate is one of the key sectors especially in emerging countries, so many different statistical modelling techniques has been applied to predict real estate price trends. There are also some published papers about application of kriging for real estate market in literature. In these studies generally, kriging was used as spatial estimators by using variograms for predicting housing prices. For example, Pace et al(1998) made overview about spatial data in real estate[4].

[♠] Corresponding author, e-mail: semra.erpolat@msgsu.edu.tr

Dubin et al(1999) described that how spatial statistical techniques can be used to improve the accuracy of market value estimations [5].

Zang et.al (2009) produced spatial decision support system for affordable housing. They used kriging for griding sample dots [6].

Olmu (2007) estimated housing location price by using kriging methods, isotopic data cokriging, and heterotopic data cokriging methods. This study also prediction was made by using location based data [7].

On the other hand, Cruerio(2007) worked on variograms and covariance functions with non-euclidean distances and made comparision between kriging predictions [8]. There is no any previous study analyzing residential price movements and macroeconomic indicators by using different correlation functions with different distance metrics for real estate data. At this point, this paper is the first attempt that compares different correlation functions with different distance metrics to model relationship between residential sales prices and macroeconomical indicators.

2. KRIGING METAMODELLING

Kriging, originally comes from geostatistics. It was found by Krige who is mine engineer. But in recent years so many different forms of Kriging metamodelling has been proposed. With the important progress in computer experiments, kriging has modified for using in computer experiments.

Applying linear regression to a computer code has some major practical and conceptual problems. The practical problem is lack of information about what functional form to specify for the regression terms. With Kriging without information about the functional form to specify for the regression terms, researchers could create robust model [9].

Kriging metamodeling technique that can be mathematically expressed as below;

$$\hat{y} = \hat{\mu} + \mathbf{r}'\mathbf{R}^{-1}(\mathbf{Y} - \mathbf{1}\hat{\mu}) \qquad (1)$$

In this equation, $\hat{y}$ is the predicted response value of **x** which is not observed before and $\hat{\mu}$ is explained as mean of the stochastic process. **Y** is the response value. **R** is corellation matrix gives the correlations between design points (Eq.2). **r** gives the correlations between **x**, which is not observed before, and design points (Eq.3). **1** is the unit matrix [10].

$$\mathbf{R} = exp\left[-\sum_{h=1}^{k}\theta_h\left|x_h^{(i)} - x_h^{(j)}\right|^{p_h}\right],\ \ \theta_h \geq 0, p_h \in [1,2] \qquad (2)$$

$$\mathbf{r} = exp\left[-\sum_{h=1}^{k}\theta_h\left|x_h\ - x_h^{(i)}\right|^{p_h}\right],\ \theta_h \geq 0, p_h \in [1,2] \qquad (3)$$

Mostly used correlation functions are exponential and gaussian. Exponential correlation functions could be written as equation 4 and 5.

$$\mathbf{R} = exp\left[-\sum_{h=1}^{k}\theta_h\left|x_h^{(i)} - x_h^{(j)}\right|\ \ \right],\ \theta_h \geq 0 \qquad (4)$$

$$\mathbf{r} = exp\left[-\sum_{h=1}^{k}\theta_h\left|x_h\ - x_h^{(i)}\right|\ \ \right],\ \theta_h \geq 0 \qquad (5)$$

Gausssian correlation functions could be written as equation 6 and 7. Also ${x_h}^{(i)} - x^{(j)}$ and $x_h - {x_h}^{(i)}$ parts of correlation functions gives simple distances.

$$\mathbf{R} = exp\left[-\sum_{h=1}^{k}\theta_h\left|x_h^{(i)} - x_h^{(j)}\right|^2\right],\ \theta_h \geq 0 \qquad (6)$$

$$\mathbf{r} = exp\left[-\sum_{h=1}^{k}\theta_h\left|x_h\ - x_h^{(i)}\right|^2\right],\ \theta_h \geq 0 \qquad (7)$$

Maximum Likelihood Estimation is used for predicting θ parameters. Assuming **Y** has a normal distribution, likelihood function (L) can be written as follows

$$L = \frac{1}{(2\pi)^{\frac{n}{2}}(\sigma^2)^{\frac{n}{2}}|R|^{\frac{1}{2}}} exp\left[\frac{(y-1\beta)\prime R^{-1}(y-1\beta)}{2\sigma^2}\right] \qquad (8)$$

After obtaining θ parameters which maximize likelihood function, Kriging model must be validated. For this purpose well known Cross-validation method is used for validation. In this method, a prediction is generated with one data point excluded from the data set. Then check whether that data point falls within a certain confidence interval for the prediction. If the test fails, appropriate transformations such as log or inverse may be applied to the response values [11].

3. MAXIMUM LIKELIHOOD ESTIMATION (MLE)

In statistics, maximum-likelihood estimation (MLE) is a popular and well-known method of estimating the parameters of some statistical model. Somtimes the MLE could be used for estimation parameters of some statistical distributions.

To use this method of maximum likelihood, the joint density function for all observations must be specified. For an independent and identically distributed sample, the joint density function is written as following equation.

$$f(x_1, x_2, \ldots, x_n|\theta) = f(x_1|\theta) \times \ldots \times f(x_n|\theta) \qquad (9)$$

$\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_\mathrm{n}$ in equation 9 are observed values and θ is parameter to be estimated and likelihood function for this situation could be written as follow [12].

$$L(\theta; x_1, x_2, \ldots, x_n) = f(x_1, x_2, \ldots, x_n|\theta)\ \prod_{i=1}^{n} f(x_i|\theta) \qquad (10)$$

4. DISTANCE METRICS

A metric or distance function d(x,y) defines the distance between elements of a set of non-negative real numbers. When distance is zero between two elements, it means that the elements are equal. Distance functions is also used to measure how close two elements are. In mathematics, the Euclidean distance is ordinary distance between two points in Euclidean space. With this distance, Euclidean space becomes a metric space. The associated norm with this distance is called the

Euclidean norm. Euclidean distance is calculated by following equation [13].

$$\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (11)$$

The Canberra distance was introduced by Lance and Williams as a software metric. The metric is a weighted format of the classic L_1 [14].

$$\sum_{i=1}^{n}\frac{|x_i - y_i|}{|x_i|+|y_i|} \quad (12)$$

5. APPLICATION

To model relationship between macroeconomical indicators and housing prices ind Dubai, Oil prices, Gold prices and Dubai Financial Market General Index (FMGI) were selected as macroeconomical indicators and calculated pearson's correlation coefficient for the variables. Macroeconomical indicators were obtained from Dubai Statistical Centre, Dubai housing sales prices were obtained from REIDIN Real Estate Information Services for the time period between January 2009 and March 2015 period [15].

	Gold	Oil	FMGI	HOUSING PRICES
Gold	1	0.828(**)	0.277(*)	0.421(**)
Oil	0.828(**)	1	0.382(**)	0.374(**)
FMGI	0.277(*)	0.382(**)	1	0.929(**)
HOUSING PRICES	0.421(**)	0.374(**)	0.929(**)	1

* *Correlation is significant at the 0.05 level (2-tailed).*

** *Correlation is significant at the 0.01 level (2-tailed).*

Table 1. Correlation Coefficients

Table 1 shows that house price rise and falls are mostly related with Dubai Financial Market General Index (FMGI), which is a stock exchange located in Dubai and a free-float market capitalization weighted price index comprising stocks of listed companies. Also, significant correlation is observed between housing prices and FMGI with a 0.929 coefficient. According to these results, decided to select FMGI as independent macroeconomic variable.

Before building kriging model FIC applied to response value for detecting that if the response values have gaussian distribution or non-parametric distributions.

	Mean	Sd	RMSE	Rank
Nonparametric	10663	495.0708	495.0708	1
Gaussian	11560.02	385.8852	934.1257	2

Table 2. Summary Statistics of FIC Test for Normality

According to Table 2, response value has nonparametric distribution. After applying inverse transformation to response value, the response values were converted to normal distribution. Results are shown in Table 3.

	Mean	Sd	RMSE	Rank
Nonparametric	0.0000938	0.00000468	0.0000047	2
Gaussian	0.0000908	0.00000271	0.0000030	1

Table 3. Summary Statistics of FIC Test for Normality

In this study MATLAB was used for all calculations. For calculations, created following MATLAB functions.

function [d_R R r F L_final ind y1hat_euclidean_gaussian s_euclidean_gaussian]=findtheta (x,y,d)

function [d_R R r F L_final ind y1hat_canberra_gaussian s_canberra_gaussian]=findtheta2 (x,y,d)

function [d_R R r F L_final ind y1hat_ euclidean _exponential s_ euclidean _exponential]=findtheta3 (x,y,d)

function [d_R R r F L_final ind y1hat_canberra_ exponential s_canberra_ exponential]=findtheta4 (x,y,d)

findtheta functions were designed for 1000 steps iteration with user defined increments value. Users could define initial value to theta according to structure of correlation functions and distance metrics. After that, the functions find theta which maximizes log likelihood function.

Log likelihood function, when using exponential correlation function with euclidean distance, reached maximum at 554 iteration, Log likelihood function, when using exponential correlation function with canberra distance, reached maximum at 104 iteration. After applying cross validation for all likelihood functions with euclidean and canberra distance, standardized error was obtained. According to graph shown belown, kriging model with canberra distance, obtained smaller standardized residual than euclidean distance.

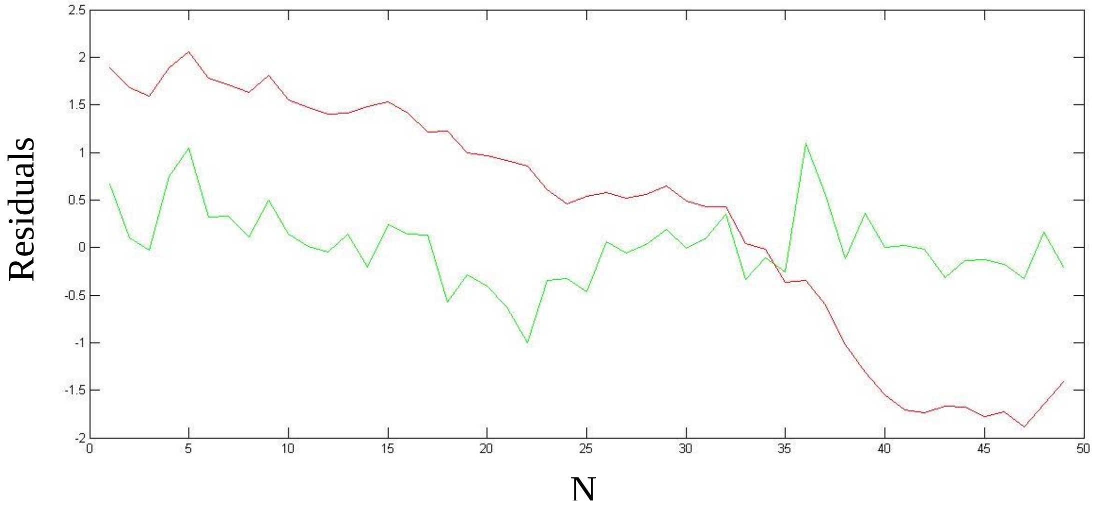
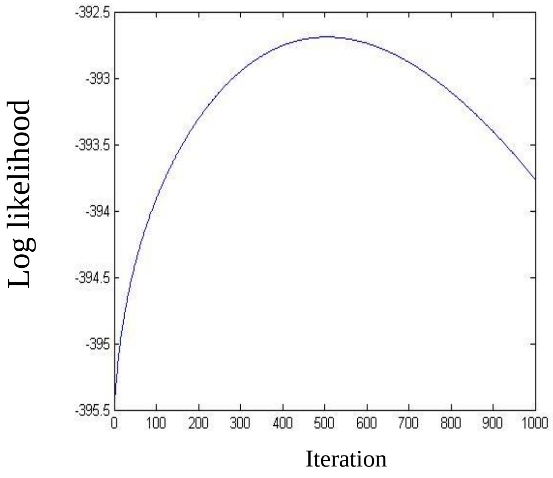


Figure 1. Residuals for exponential correlation function with euclidean and canberra distances.

(Red line for euclidean, green line for canberra)

Iteratively changes of log likelihood functions for euclidean and canberra distances as following figures.



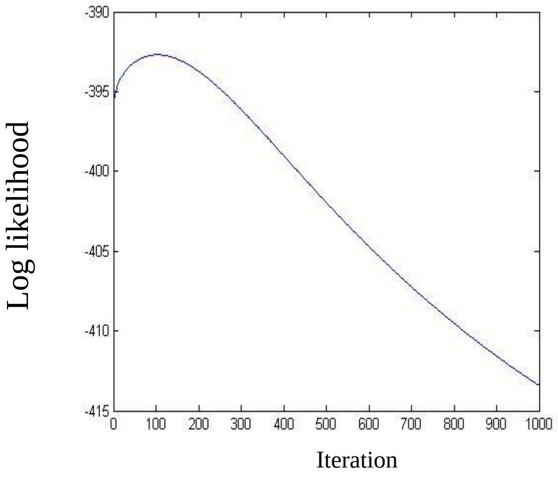


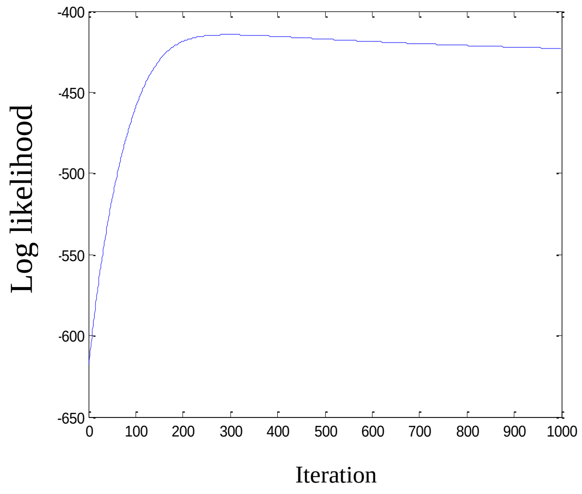
Figure 2.

Log likelihood function for exponential correlation function with euclidean distance

Figure 3.

Log likelihood function for exponential correlation function with canberra distance

Log likelihood function, when using gaussian correlation function with euclidean distance, reached maximum at 298 iteration, Log likelihod function, when using exponential correlation function with canberra distance, reached maximum at 279 iteration. Iteratively changes of log likelihood function for euclidean and canberra distances as following figures.



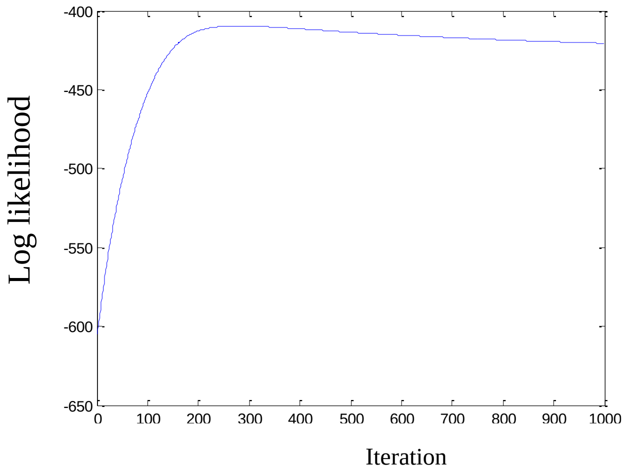


Figure 3.

Log likelihood function for gaussian correlation function with canberra distance

Figure 4.

Log likelihood function for gaussian correlation function with euclidean distance

After applying cross validation for both likelihood function with euclidean and canberra distance, standardized error was obtained. According to Figure 5, kriging model with euclidean distance obtained smaller standardized residual than canberra distance.

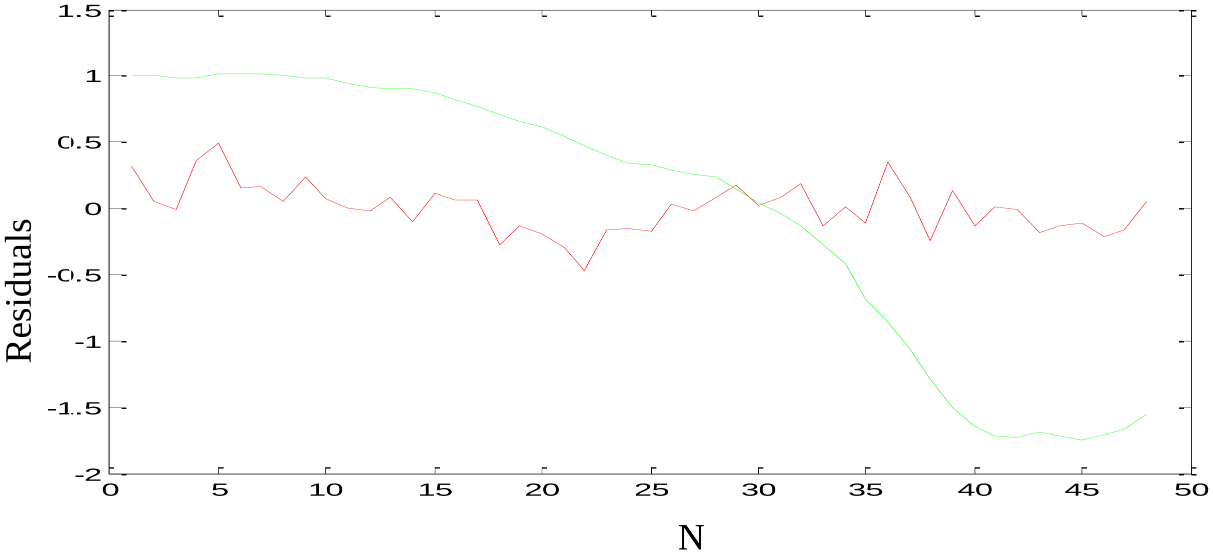


Figure 5. Residuals for gaussian correlation function with euclidean and canberra distances.

(Red line for euclidean, green line for canberra)

6. CONCLUSION

By this study, relationship between residential price movements and macroeconomic indicators were modeled by using different correlation functions with different distance metrics. Canberra and Euclidean distances were used in both gaussian and exponential correlation functions.

	Canberra_gaussian	Euclidean_gaussian	Canberra_exponential	Euclidean_exponential
Min	-1.7503	-0.4705	-0.7424	-2.9854
Mean	-0.0074	0.0002	0.0059	-0.9127
Max	1.0174	0.4963	0.9993	2.4478
Std	1.0311	0.1855	0.3321	2.0393

Table 4. Summary Statistics for Cross Validated Standardized Errors

According to Table 4 and Figure 6, exponential correlation function with canberra distance has minimum residuals rather than exponential correlation function with euclidean distance. Gaussian correlation function with euclidean distance has minimum residuals rather than gaussian correlation function with canberra distance.

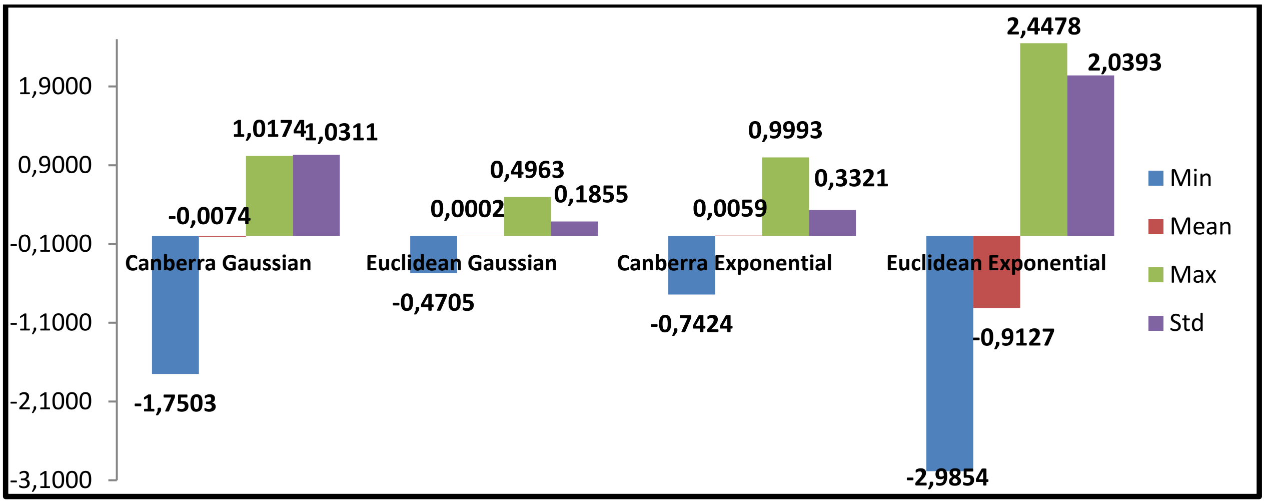


Figure 6. Summary Statistics for Cross Validated Standardized Errors

It could be said that gaussian correlation function with euclidean distance is better than gaussian correlation function with canberra distance and exponential correlation function with canberra distance is better than exponential correlation function with euclidean distance. For modelling housing prices with kriging, gaussian correlation function with euclidean distance should be used.

For future works, more distance metrics and correlation functions should be used to make high detailed comparision for real estate data. Also best kriging models should be investigated for different application areas for example biostatistics, finance, engineering etc.

CONFLICT OF INTEREST

No conflict of interest was declared by the authors.

REFERENCES

[1] Hepsen A., Vatansever M. (2012) Relationship between residential property price index and macroeconomic indicators in Dubai housing market, International Journal of Strategic Property Management, 16:1, pp. 71-84.

[2] Pacione, M. (2005) Dubai. Cities, Vol. 22, No.3, pp. 255-265.

[3] Hepsen, A., Vatansever, M. (2011) Forecasting future trends in Dubai housing market by using Box-Jenkins autoregressive integrated moving average, International Journal of Housing Markets and Analysis Vol. 4 No. 3, pp. 210-223.

[4] R. Kelley Pace, Ronald Barry, C. F. Sirmans (1998) Spatial Statistics and Real Estate, The Journal of Real Estate Finance and Economics, Volume 17, Issue 1, pp 5-13.

[5] Robin Dubin, Kelley Pace, And Thomas Thibodeau (1999) Spatial Autoregression Techniques for Real Estate Data, Journal of Real Estate Literature, Vol. 7, No. 1, pp. 79-95.

[6] Zuo Zhang, Jiangfeng Li (2009) Application of GIS and Spatial Decision Support System for Affordable Housing, Proceedings of 2009 4th International Conference on Computer Science & Education

[7] Jorge Chica-Olmo (2007) Prediction of Housing Location Price by a Multivariate Spatial Method: Cokriging. Journal of Real Estate Research, Vol. 29, No. 1, pp. 91-114.

[8] Frank C. Curriero (2006) On the Use of Non-Euclidean Distance Measures in Geostatistics, Mathematical Geology, November 2006, Volume 38, Issue 8, pp. 907-926.

[9] D. R. Jones, M. Schonlau, and W. J. Welch (1998) Efficient Global Optimization of Expensive Black-Box Functions, Journal of Global Optimization, 13, 455–492.

[10] Hepsen A., Aydin O., Vatandas O. (2015) Statistical Analysis For Impacts Of Economical Conditions on Housing Markets: An Example on Fragile 5 Countries, Journal of Business, Economics and Finance, 4,1, pp. 1-23.

[11] M. Schonlau(1997) Computer Experiments and Global Optimization,Phd. Thesis, University of Waterloo, pp. 8-11.

[12] Moti L. Tiku and Aysen D. Dikkaya (2004) Robust Estimation and Hypothesis Testing, New Age International Limited, Publishers, New Delhi, pp. 22-23.

[13] Math.NET Numerics, Accessed 19 July 2016.

[14] Schulz Jan., "Canberra distance". Code 10, Accessed 18 October 2015.

[15] REIDIN Real Estate Information Services.