2019 5th International Conference on Web Research (ICWR)

Using Web Mining in the Analysis of Housing Prices: A Case study of Tehran

Rahimberdi Annamoradnejad
Associate Professor, Geography and Urban Planning
University of Mazandaran
Babolsar, Iran
Corresponding author, r.moradnejad@umz.ac.ir

Taher Safarrad
Assistant Professor, Geography and Urban Planning
University of Mazandaran
Babolsar, Iran
t.safarrad@umz.ac.ir

Issa Annamoradnejad
Department of Computer Engineering
Sharif University of Technology
Tehran, Iran
i.moradnejad@gmail.com

Jafar Habibi
Associate Professor, Computer Engineering
Sharif University of Technology
Tehran, Iran
jhabibi@sharif.edu

***Abstract*— There have been many previous works to determine the determinants of housing prices. All of these works relied on a relatively small set of data, mostly collected with the help of real estate agencies. In this work, we used web mining methods to generate a big, organized dataset from a popular national brokerage website. The dataset contains structural characteristics of more than 139,000 apartments, alongside their location and price. We provided our full dataset for the article, so that other researchers can reproduce our results or conduct further analyses. Using this dataset, we analyzed housing prices of Tehran in order to identify its major determinants. To this aim, we examine the dynamics of housing prices at the district levels of Tehran using Hedonic Price model. Our results highlight a number of points, including: Base area of an apartment is positively correlated with price per square meter (r=0.89), showing a two-folded impact on the overall price. Air quality is in a positive, and floor level is in negative correlation with housing prices.**

Keywords*— Hedonic Price model; web mining; web crawler; housing prices; real estate market; Tehran

I. INTRODUCTION

There have been many previous works to determine the determinants of housing prices. All of these works relied on a relatively small set of data, mostly collected with the help of real estate agencies. In this work, we used web mining methods to generate a big, organized dataset from the contents of a national brokerage website. This dataset contains multiple structural characteristics of more than 139,000 apartments, alongside their location and price. We provided our full dataset for the article [1], so that other researchers can reproduce our results or conduct further analyses.

Based on our aggregated dataset, this paper aims to examine dynamics of housing market at the district levels of Tehran, as a unique multidimensional city. More specifically, we created a web crawling program to parse the web pages of the largest national real estate brokerage website[1] in order to retrieve the overall price and attributes of dwellings. Furthermore, using the Hedonic Price model, we quantitatively and comprehensively examine the structural, locational and environmental factors of housing prices.

The structure of this paper is laid out as follows: Section 1 is the introduction. Section 2 presents a brief review of relevant literature on the determinants of housing price using the Hedonic Price model, a brief description of web crawling, and the study area. Section 3 will elaborate on the data and the methodology, and Section 4 explores our findings on the housing prices and their variables to give insights into the external and internal factors of housing market. The last section is conclusion for the study.

II. BACKGROUND

A. Determinants of housing prices

The Hedonic model is widely used to analyze the price of a good, by assuming that the final price is a function of the good's characteristics [2]. This model is based on Lancaster's theory of consumer demand [3]. Rosen [4] was the first to apply Hedonic utility on pricing a good, in which the value of an item is judged by its characteristics, and the total price of the item is estimated by the sum of prices for all homogeneous attributes. In this model, each attribute has an implicit value in an equilibrium market, and therefore, by using regression analysis on the value of each attribute, the model can determine the way in which each attribute contributes to the overall price of the item [5].

Many studies in the field of housing prices used the Hedonic model to determine the contribution of various variables on the final prices. Reference [6] investigated the determinants of housing prices in Istanbul. Their data was collected with help from real estate agents during the summer of 1997, which includes 1468 observations from 26 sub-districts. They showed that at the metropolitan level, the most influential factors of housing prices are sub-market, floor area, and sea view. They also showed that at the district level, housing prices vary from district to district according to various locational, socio-economic and property characteristics. In addition, they were able to show a positive correlation between planning for districts and housing prices of the area. Finally, they concluded that because of higher tax revenues, restructuring squatter areas and revitalizing inner city areas benefit the city as well as the individuals [6].

There have been a few studies discussing the determinants of housing prices for the cities of Iran. Reference [7] applied the Hedonic price approach in residential properties of Mashhad, the second largest city of the country. Specifically, they investigated various variables for 775 houses and showed a strong positive correlation between the incomes of residents and their property values. They analyzed spatial autocorrelation in residuals of the

[1]Can be accessed from http://iranfile.ir.

978-1-7281-1431-6/19/$31.00 ©2019 IEEE

city, in which they were unable to find any homogeneous distribution [7]. A recent study [8] examined the impact of a nearby lagoon, as an environmental factor, on housing prices of Rasht. They used Hedonic pricing method for this northern city and showed that Eynak lagoon has a negative influence on the housing prices of the nearby area. Another recent study [9] used Hedonic Price model and an Artificial Neural Network model in analyzing the determinants of housing prices to predict accurate prices for the city of Ahvaz, a city in western part of Iran. They analyzed 27 variables on a relatively small sample of 286 cases and concluded that house price in Ahvaz is mainly influenced by structural variables such as land area and age of building.

In previous studies, locational determinants usually involved CBD, where it was associated with mixed results (positive and negative correlation). A few studies discussed the effect of air quality on housing prices over multiple cities. Reference [10], conducted a study in 85 cities in China, finding that for every 10% decrease in the inflow of air pollutants around a city, housing prices increase by 1.8%. Another study [5] examined air pollution effects on real estate values of Beijing, in which they demonstrated a negative correlation between housing prices and Air Quality Index (AQI).

B. Web Crawler

In order to collect the data necessary for this study, we created a web crawling program which parses the web pages of a national real estate brokerage website and retrieves full information about the structural attributes of the submitted apartments. In this section, we briefly explain a web crawler and the extent that it can be used for scientific researches.

A web crawler, robot or spider is a program or a collection of programs that will iteratively and automatically download web pages, extract URLs from their HTML and fetch the contents of all URLs [11]. A sophisticated web crawler may also perform additional calculations during the crawl in order to identify irrelevant pages to the crawl purpose, reject pages as duplicates of the ones previously visited or to perform data mining tasks to extract further information from the web page [12]. Information scientists and others who wish to extract information from large numbers of web pages require the services of a web crawler or web-crawler-based tool [13]. The potential power of web mining is illustrated by the Google search engine, which uses mathematical calculations, such as PageRank, on a huge matrix of data in order to extract meaningful relationships from the link structure of the web [14]–[16].

C. Study Area

Tehran is the political and economic center of Iran, and the largest and most populous city in Western Asia with more than 8.8 million residents in the city and 15 million in the larger metropolitan area [17]. It has one of the highest betweenness and closeness centrality among the cities of Iran, regarding national road and air routes [18]. As of 2018, the city of Tehran is divided into 22 municipal districts and 354 neighborhoods (Fig. 1) with an area of 574 km^2 [19]. In general, the northern districts comprise the majority of the commercial centers; the crowded city center contains bazaars, government ministries and headquarters, and the southern parts are mostly residential neighborhoods for city's newcomers. It is widely known in Iran that the wealthiest people and companies of Iran reside in the Northern districts of Tehran [20] which can be the result of its elevated height, less air pollution, natural destinations, ski resorts, and historical palaces.

Tehran is located in the northern part of Iran, towering Alborz Mountains to its north and the country's central desert to its south. Because of the vast area of the city, there are significant differences in elevation among various districts, ranging from 900 m above sea level in the south to over 1830 m in the northern districts, which can even rise up to 2,000 m at the end of northern regions. While elevation cannot be considered as a considerable determinant of housing prices for a small city or for a single district, it can be a factor in determining housing prices of multiple districts, especially in a city like Tehran with its wide range of elevation values.

The city faces the recurring issue of air pollution for many decades and there have been multiple official plans to relocate the capital of Iran to another place. While these plans are due to a few major environmental and economic issues (such as fault lines and traffic), the main reason has been the high values of air pollution [21], [22]. A few reasons have been attributed to this low quality of air, such as the activities of more than 2 million active vehicles, 5 thousand industrial units and 70 percent of the country's services [23]. Furthermore, the majority of industrial units are located to the west of the city which is in the path of in-flow winds.

III. Data and Method

A. Data

We created an automated web crawler to mine the web pages of a national real estate brokerage website. The tool iterated over the web pages dedicated to apartment exchange in the city of Tehran. It parsed the HTML code of each web page, extracted the relevant information about each entry and inserted a single row for each entry in the dataset. Using this tool, we generated a dataset composed of 139,751 entries that accounts for the past four years of website's activity. Each entry consists of entry Id, submission date, exact address, neighborhood name, base area, floor level, age of building, price per square meter and total price. We provided this dataset as supplementary data for other researchers to reproduce the results of this study or to conduct further studies.

For the purposes of this study, we separated the submitted entries of 2017 (39,257 entries) from the rest of the dataset. In addition, by using a dictionary function, the physical address of each entry was mapped to the district that the house resides in. As shown in Fig. 2, the selected entries are scattered unevenly among the 22 districts. District 13 with more than 14500 cases and district 19 with only 46 cases have the highest and lowest number of samples. Furthermore, since all of the prices were in Iran's national currency, Iranian Rial, we converted these values to US Dollar for ease of understanding by foreign readers. Average prices in districts are shown in Fig. 2.

To determine average land price of each neighborhood, we collected 685 entries from the same website which accounts for all of the land items in the past four years. The difference between the amounts of the two types clearly shows that apartment trades are more frequent than land trades. For the purposes of the next sections, we separated land entries of each neighborhood and calculated their average price per square meter.

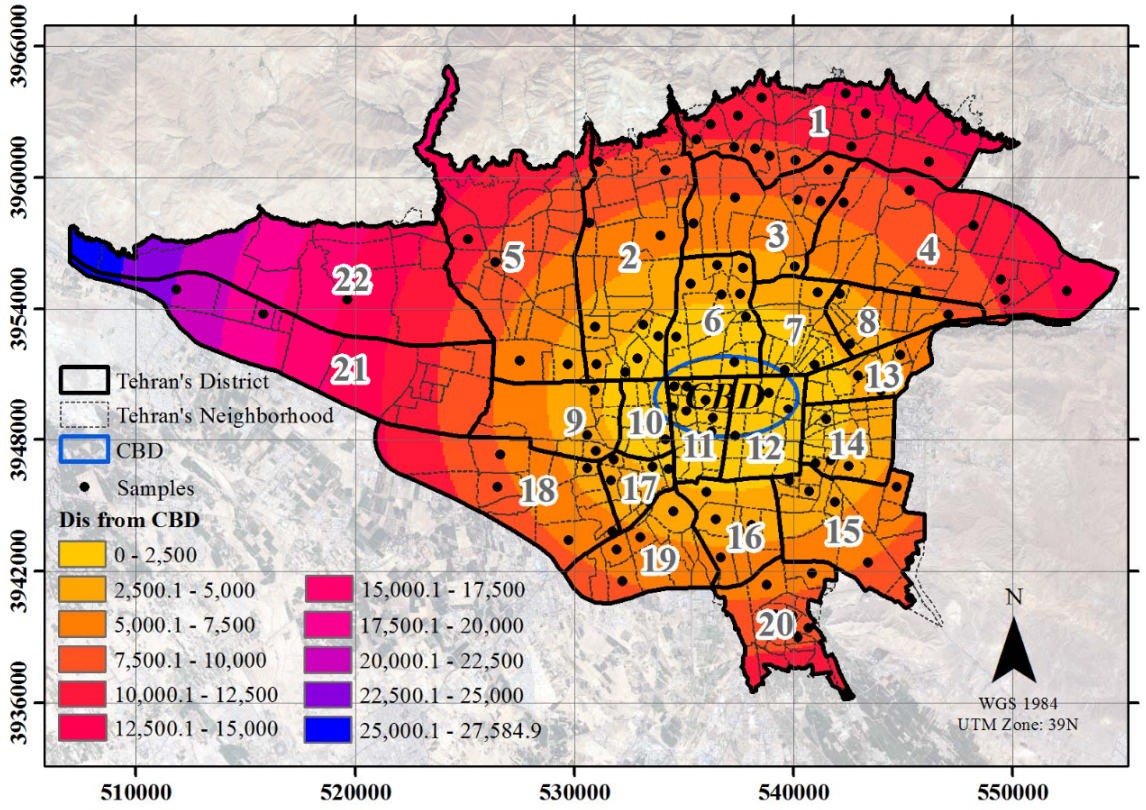

Fig. 1. Districts and neighborhoods of Tehran and their distance to CBD (in meters)

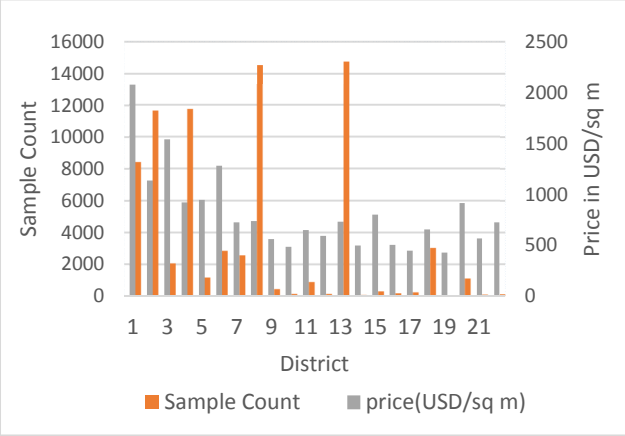

Fig. 2. Average prices and apartment sample count in 22 districts

B. Method

We picked 8 structural, locational and environmental variables to be assessed in this study. In order to determine the impact of each variable on the overall price of a house, we use the Hedonic Price model alongside Stepwise multiple regression as our core analysis methods. For better understanding of the variables and our results, we classified the selected variables into three main groups, which are:

1) *Structural attributes* (S) which include age of the apartment, floor level, and base area;
2) *Locational attributes* (L) which include distance to CBD, district population density and average land price;
3) *Environmental attributes* (E) that include air quality and green space per capita.

Table I displays details and some general statistics for the aforementioned attributes.

Hedonic Price model is our core analysis tool. The Hedonic Method assumes that a commodity is characterized by the set of all its characteristics, supposedly denoted by a vector $x = (x_1, x_2, x_3, \dots, x_n)$, which can be used to determine the preferences of the economic actors with respect to commodities. Therefore, the task is to find a functional relationship between the price and characteristics vector to create a unique utility value for any given vector. By considering transaction price as the utility value for an apartment and the values for structural, locational and environmental attributes as the apartment's vector, a functional relationship can be developed.

The Hedonic method takes place in two stages. In the first stage, we differentiate the price function with respect to any of the aforementioned characteristics to find the implicit utility function for that particular characteristic. It is called implicit, since the utility function is indirectly revealed to us using the extent actors are willing to go in order to obtain a better quality of that attribute. In the second stage, we apply regression models to these implicit prices against the actual quantities/qualities, so that we attain the marginal willingness of actors to pay for the amenity. The results of this stage will explicitly determine changes in the overall property price for a unit change for a characteristic, given that all other characteristics remain constant.

TABLE I. SELECTED PROPERTY VARIABLES AND THEIR ASSOCIATED CATEGORIES

Category	Name	Definition	Min	Max
Structural	Floor level	Floor level of the house above ground level	-1	43
	Age	Number of years after the construction date to the transaction date	0	30
	Base area	Base area of the apartment in square meters (m^2)	14	1800
Locational	Distance to CBD	Distance of the shortest path from district center to CBD (Fig. 1) in meters	447	16254
	District population density	Population density of the district where the apartment resides in (people/hectare)	29	399
	Land price	Average price of land (USD per square meter) for the neighborhood that the apartment is located	19	19047
Environmental	Air quality	Air Pollution Index in the district where the apartment is located (particles smaller than $PM10^3$)	53	129
	Green space per capita	Green space area per capita for the district (in square meters per person)	2.8	66

Districts 10, 11, 6, 12 are considered as the central districts of the town, mostly for their governmental buildings and well-known markets such as Grand Bazaar. We considered parts of the above-mentioned districts as CBD (shown in Fig. 1), which is located close to the center of the geographical map. While the central districts contain more than 50% of the governmental buildings and 40% of industrial units of the city, a few areas outside these regions such as Tajrish in the north and Rey in the south, are turning into trade centers for the residents of the neighboring districts. While the standard urban economic model ([24], [25]) assumes a monocentric city that the majority of employment is located in CBD, the creation of multiple trading centers in Tehran is rejecting this assumption by transforming Tehran from a monocentric city into a multicentric city.

IV. FINDINGS

We represent our result in two sections. In the first section, we will discuss some general statistical findings, such as the distribution, average, min and max values for the characteristics at macro levels, such as the district level. In the second section, we will represent our results to find the impacts of the selected variables on housing prices in Tehran and will further examine the city's housing market.

A. Analysis of attributes at macro levels

In order to give a general view on housing prices and housing attributes, we analyzed our dataset at macro levels (district and neighborhood level) using aggregated observations. At the district level, the average price for an apartment is 983 USD/m^2, ranging from 430 USD/m^2 in district 19 to 2070 USD/m^2 in district 1, as it is displayed in Fig. 2. In addition, Zaferanieh with 3159 USD/m^2 and Afsarieh with 343 USD/m^2 have the highest and lowest average of prices among the neighborhoods of Tehran.

The average base area for the entries of the whole city is 92.38m^2, ranging from 68m^2 in district 14 to 209m^2 in district 1. At the neighborhood level, Isfahanak and Fereshteh, have the lowest and highest average base area with 60m^2 and 214m^2, respectively. The average age of buildings for the entries of our dataset is 7.5 years, which ranges from 2.2 to 26 years at the district level. The average floor level for the entire city is 2.7, ranging from 2.4 in district 7 to 6.6 in district 22. It is interesting to note that district 22 has the highest number of floors on average while being the youngest district of Tehran.

Analysis on the population density of the districts shows 141 people/hectare for the whole city, where district 10 at the city center with 399 people/hectare and district 22 with 29 people/hectare has the highest and lowest density among the 22 districts of Tehran. The average green space per capita for the whole city is 17.4m^2, which ranges from 2.8m^2 in district 10 to 66m^2 in district 22 (Tehran Parks & Green Space Organization 2018). It is interesting to note that district 10 has the highest population density and lowest green space per capita, while district 22 is in exact opposite situation. It can be inferred from the results that the significance of district 10 and its closeness to CBD had led the city planners to reduce open spaces to provide more space for residential apartments.

B. Determinants of housing prices

Base area of the apartment is the most influential factor among the selected variables, which has a strong positive correlation with the price per square meter (correlation coefficient of 89%). This means that base area effects the overall price of a house in two separate ways; first by its area size and second by its correlation with unit price. This finding can be explained by the fact that while the wealthy districts of 1 and 3 have the highest base area average among the 22 districts of Tehran, more affordable districts have lower base areas in average.

Floor level is another determinant of housing prices in Tehran. It is in a negative correlation with the total price of an apartment (correlation coefficient of -0.68), which relates to the findings of a previous study about Istanbul [6]. One cause of this relationship can be attributed to the special case of district 22. As we briefly discussed this district in the previous sections, it is the youngest district of Tehran and has been formed in a response to the increasing need of the city's newcomers for residence. The district has the highest floor level average among the districts, which because of its low qualities of urban amenities and welfare services has become one of the cheapest and affordable districts of Tehran (Fig. 2). As we will discuss this shortly later, district 22 has the highest distance from CBD, which is another sign of lower quality of welfare services for the residents of the district.

We were unable to show a robust connection between the age of a building and its total price, as the correlation analysis for this variable showed a weak positive correlation with a coefficient of 30%. In contrary, the land price was positively correlated with housing prices with a correlation coefficient of 50.5%. Based on our observations, the impact of land price increases with the age of a building, meaning that the older an apartment is the higher the effect of land price becomes.

We were unable to find any strong correlation between distance to CBD and housing prices. This result can be explained by the fact that the majority of dwellings in these central districts are located in crowded business neighborhoods with many timeworn structures. This reason alongside with the higher quality of air in some districts and ease of access to CBD through public and private transportation have led many people to relocate to northern or other suburban districts. In contrast to northern districts, district 22 has the highest distance from CBD with an average of 16.2 km, while having low housing prices in average.

Based on correlation analysis[2], air pollution has a negative impact on housing prices with a correlation coefficient of -0.48. Therefore, central and southern districts that have higher air pollution levels are lower in price. Based on an article from 2017 [26], districts 18, 9 and 21 had the highest levels of air pollution with more than 100 $PM10^3$. These districts are located in the south-east of Tehran where the western wind flows into the city. Even though it has been suggested that the cause of air pollution in these areas are external to the activities of the city [27], blockage of air flow in the northern and north-eastern part of the city by mountains range can be considered as a potential reason for the congestion of air pollutants in the south-western areas.

We were unable to find any relationship between the two remaining factors -green space per capita and population density- with housing prices.

V. Conclusion

In this study, we examined dynamics of housing market in the 22 districts of Tehran. We selected Tehran for its multiple interesting characteristics, including its political and economic significance, population density, ever-increasing air quality issues and the wide range of elevation across its districts. We used Hedonic Price model in order to assess the structural, locational and environmental determinants of housing prices in the city of Tehran. Furthermore, we provided some general information about housing prices and the selected variables at the district level.

We developed a web crawling application to parse the web pages of the largest national real estate brokerage website to retrieve the overall price and structural attributes of all submitted apartments in Tehran. Using this tool, we collected a dataset that spans for the information of the past four years (close to 140,000 entries). Each entry consists of entry Id, submission date, exact address, neighborhood name, house age, base area, price per square meter and total price. By providing this dataset as supplementary data, other researchers are able to reproduce the results of this study or conduct further analyses.

To examine the determinants of housing price using the Hedonic model, we selected 8 factors, composed of 3 structural, 3 locational and 2 environmental characteristics. Base area was the most influential factor, showing a strong positive correlation with price per square meter (r = 89%), which can be associated with the fact that the wealthy districts of 1 and 3 have larger base area than crowded and cheaper districts. In addition, floor level was negatively correlated with the total price with a correlation coefficient of -0.68 and age of building was in a weak positive correlation with the total price.

Although the impacts of structural determinants are more concrete than those of external variables, results confirm that housing price is not only affected by the physical characteristics of a building, but also by its external factors, such as air pollution and average land price of the neighborhood. In this regard, land price was positively correlated with the total price of apartment with a correlation coefficient of 50.5% and air pollution level was negatively correlated with the same statistic with coefficient of -48%.

Public and private city planners can use the results of this study to further develop the districts of the city, possibly by redistributing urban amenities and welfare services. Homeowners, buyers, and investors can also use our results for a higher quality and more productive participation in the real estate market, and future researchers can use the methodology, the dataset, or the final results of this article to examine more variables and models on the housing market of Tehran or other cities.

[2]For the purposes of our analysis, we considered particles smaller than $PM10^3$.

References

[1] R. Annamoradnejad, T. Safarrad, I. Annamoradnejad, and J. Habibi, "Structural attributes, locational information and prices of more than 139 thousand apartment real estate listings in the city of Tehran," *Mandeley Data*, 2018. .

[2] A. Beamonte, P. Gargallo, and M. Salvador, "Analysis of housing price by means of STAR models with neighbourhood effects: A Bayesian approach," *J. Geogr. Syst.*, 2010.

[3] K. J. Lancaster, "A New Approach to Consumer Theory," *J. Polit. Econ.*, 1966.

[4] S. Rosen, "Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition," *J. Polit. Econ.*, 1974.

[5] Y. Xiao, S. Orford, and C. J. Webster, "Urban configuration, accessibility, and property prices: a case study of Cardiff, Wales," *Environ. Plan. B Plan. Des.*, 2016.

[6] E. Ozus, "Determinants of office rents in the Istanbul metropolitan area," *Eur. Plan. Stud.*, 2009.

[7] M. Rahnama and R. Asadi, "Analysis of spatial distribution of Housing price in Mashhad," *Geogr. Res. Q.*, vol. 30, no. 1, pp. 37–52, 2015.

[8] H. Amirnejad, M. Nabizadeh Zolpirani, and R. Heydari Kamalabadi, "The impact of Rasht Eynak Lagoon on housing price of the region by using hedonic pricing method," *J. Urban Econ. Manag.*, vol. 4, no. 16, pp. 33–48, 2016.

[9] S. Ghorbani and S. M. Afgheh, "orecasting the House Price for Ahvaz City: the Comparison of the Hedonic and Artificial Neural Network Models," *J. Urban Econ. Manag.*, vol. 5, no. 19, pp. 29–44, 2017.

[10] S. Zheng and M. E. Kahn, "Understanding China's Urban Pollution Dynamics," *J. Econ. Lit.*, 2013.

[11] T. Y. Chun, "World Wide Web Robots: An Overview," *Online Inf. Rev.*, 1999.

[12] M. Thelwall, "A web crawler design for data mining," *J. Inf. Sci.*, 2001.

[13] J. Bar-Ilan, "Data collection methods on the web for informetric purposes - A review and analysis," *Scientometrics*, 2001.

[14] S. Brin and L. Page, "Reprint of: The anatomy of a large-scale hypertextual web search engine," *Comput. Networks*, 2012.

[15] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank Citation Ranking," *World Wide Web Internet Web Inf. Syst.*, 1998.

[16] D. Sulivan, "What Is Google PageRank? A Guide For Searchers & Webmasters - Search Engine Land," *Search Engine L.*, 2017.

[17] Statistical center of Iran, "Tehran 2018 report," 2018. [Online]. Available: http//amar.sci.org.ir. [Accessed: 15-Feb-2019].

[18] I. Annamoradnejad, R. Annamoradnejad, "An Analysis of Centrality Measures of Iran's Province Capitals based on Road and Air Connections (Using Gephi)," *J. Res. Urban Plan.*, vol. 8, no. 30, pp. 19–34, 2017.

[19] The Statistics of Tehran city in 2016, "Tehran Municipality's Information and Communication Technology Organization," 2016. [Online]. Available: http://tmicto.tehran.ir. [Accessed: 15-Feb-2019].

[20] S. H. I. Moeini, M. Arefian, B. Kashani, and G. Abbasi, "A Genealogy of Tehran's Art Galleries: A History of the (Home-) Studio," in *Urban Culture in Tehran*, Springer International Publishing, 2018, pp. 123–165.

[21] Tamaddon, "Iran Moots Shifting Capital from Tehran," 2010. .

[22] Aljazeera, "Iran mulls plan to move capital from Tehran," 2013. [Online].

Available: https://www.aljazeera.com/news/middleeast/2013/12/iran-mulls-plan-move-capital-from-tehran-201312241916725897.html. [Accessed: 15-Feb-2019].

[23] Iran's Department of Environment, "Tehran Air Pollution Report," 2018. [Online]. Available: www.tehran-doe.ir/fa/Pollution. [Accessed: 15-Feb-2019].

[24] E. S. Mills, *Studies in the Structure of the Urban Economy*. 1972.

[25] W. Alonso, *Location and land use. Toward a general theory of land rent.* 1964.

[26] A. Najafpoor, A. Jonidi, and S. Dousti, "Trend analysis of Air Quality Index criteria pollutants (CO, NO2, SO2, PM10 and O3) concentration changes in Tehran metropolis and its relationship with meteorological data, 2001-2009," *J. Heal. F.*, vol. 3, no. 2, pp. 15801–15810, 2017.

[27] Tehran Air Quality Control Company, "Tehran's air quality report," 2017. [Online]. Available: http://airnow.tehran.ir/home/AQIArchive.aspx. [Accessed: 15-Feb-2019].