International Journal of
Information
and
Management
Sciences

International Journal of Information and Management Sciences
30 (2019), 185-202. DOI:10.6186/IJIMS.201906_30(2).0006

The Impact of Environmental Factors on Housing Prices: A Case Study of Taipei Housing Transactions

Pei-De Wang and Mingchin Chen

Fu Jen Catholic University

Abstract

Most research on housing price modeling only utilizes a single environmental factor. The goals of this paper are to select the appropriate factors and to identify the influencing patterns for 3 major types of real estates through model building that includes 49 housing factors. The datasets were composed by 33,027 transactions in Taipei City from July 2013 to the end of 2016. The models utilized were Decision Tree (DT), Artificial Neural Networks (ANN), Random Forest (RF), Model Tree (MT), and Multiple Regression (MR). The importance of each factor derived from the above 5 models is thus analyzed and ranked for the 3 housing types. Also, this paper adopts Generalized Additive Models (GAM) to derive the patterns of important factors influencing housing prices that includes increasing, decreasing, and non-linear relationships.

Keywords: Data mining, housing prices, influencing patterns, housing type.

1. Introduction

Similar types of homes command varying prices in different regions and neighborhoods (see Visser et al. [23]). Hanink et al. [14] found significant spatial variations in the impacts of structural amenities and locational context on housing prices. Zoppi et al. [34] analyzed the relationship between such housing values and a set of determinants that were related to both the urban environment and the housing market's structural factors. To build upon these analyses, this paper adopts both structural and environmental factors to analyze Taipei's housing prices. These environmental factors are points of interest (POIs), which on a map have economic, social, and cultural values (see Bency et al. [3]). Such POIs also have a non-linear relationship with respect to housing prices (see Kim et al. [16]).

In this paper and to describe environmental factors, applied are proximity (see Sah et al. [21]) and a number of specific factors (see Wang et al. [25]). These two measures play different roles in terms of their respective influences on housing prices. Many studies utilize the proximity concept about different POIs that people are usually concerned about when purchasing a home. Proximity affects the accessibilities of POIs, as well as travel times. Thus, proximity represents the degree of convenience to reach a POI.

Diao [9] defined a subway station's impact zone as the half-mile (800m) circle as the distance loosely corresponds to a comfortable walking distance. Wen and Tao [26] adopted the 1,000m circle to measure a living facility. In this paper, the 1,200m circle is selected due to proximity and a number of specific factors are included and investigated.

The number of specific factors represents the degree of density, which reveals the aggregate numbers of factors within a buffer or catchment area of an observed property location (see Wang et al. [25], Ferrari et al. [12]). Thus, the number of specific factors shows the maturity of factors in the vicinity.

Metropolitan Rapid Transit (MRT) systems are a critical factor of housing prices in metropolitan areas. Namely as such systems have drawn the attentions of real estate investors, academics, and policymakers (see Kim and Lahr [17]). A comprehensive public transport system enables the general populace to both easily access and travel to more places. Shyr et al. [22] demonstrated that the price premium for housing near transit stations is relatively positive in Taipei City. With a log-linear function, adopted was the distance to the nearest MRT station as an accessibility factor.

Schools have well-known impacts on housing valuation. Wen et al. [27] focused on distances to primary and junior high schools. Found was that the distance to junior high schools is significant by utilizing a log-linear function. Wang [24] concluded that every additional kindergarten, primary school, and secondary school built within 500m of communities improves housing prices. Owusu-Edusi et al. [19] found a net positive impact for school proximity by applying a hedonic pricing equation.

Emrath [11] proved that having satisfactory shopping centers within a mile increased housing prices substantially via the usage of the American Housing Survey's data. Pope and Pope [20] focused on the effectiveness of Walmart stores opening nearby. Concluded were that Walmart stores increased housing prices - which vary depending on distance (proximity) - and outweigh the costs imposed by any negative externalities.

Public parks are both a vital and representative type of green space with ecological, entertainment, recreational, social, and cultural functions (see Wu et al. [32]). Hammer et al. [13] are pioneers in the influence of parks on real estate. Much of current research has focused on the appraisal of various green spaces. Wu et al. [32] stated that the distance from a house to the nearest city park is positively related to housing prices.

The availability of public transport infrastructure (e.g. bus stops) can also significantly raise and promote the values of nearby property prices. Wang et al. [25] found strong evidence that the number of bus stops within walking distance (300-1500m) to a property is positively associated with the property's observed sale price. Both proximity and the number of factors were applied in that research.

The lifestyles of people in areas surrounding convenience stores have changed as such stores have replaced the roles provided by larger stores (e.g., supercenters, supermarkets). Thus, convenience stores are used as a key determinant of property prices in housing studies about both proximity and number (if a house has 2 or more convenience stores; see Chiang et al. [6]).

The aforesaid studies focus on specific environmental characteristics. Most of such studies utilize linear regression models. This paper encompasses more factors to determine the configuration of the housing prices of 3 major housing types in Taipei. Also, this paper analyzes the links between POIs and housing prices that include increasing, decreasing, and non-linear ones.

2. Data

This study's basic data was downloaded from the Taiwan Actual Price Registration (APR). The APR's factors are regarded as structural (shown in Appendix factor 1∼18). Environmental aspects in the vicinity of homes were retrieved from a designated distance circle (1,200m) that applied the house as the center. As shown in Table 1 below, this paper includes 14 environmental factors:

Table 1: Environmental categories.

Categories	factors
transport-related	MRTs, train stations, bus stops
academic-related	university, senior high school, high school, elementary, library
shopping-related	department stores, supermarkets, convenience stores
living-related	hospitals, gas stations, parks

If there were no specific POI in the circle, this paper thus employed 1,500m as the shortest distance for that POI. All the POIs and transacted houses have both latitudes and longitudes. Thus, the Haversian distances between POIs and a house (i.e., the great-circle distances) were determined. The POI elements (which capture environmental information) are able to improve predictions (see Bency et al. [3]).

Based on the works of Chen and Wang [5], the major housing types addressed in this paper included apartments that are condominiums of five stories or lesser in height and without an elevator (APT); buildings that are condominiums with elevators (BLD); and suites (SUT). In total, this paper applies 49 factors that are listed in the Appendix. The overall data includes 33,027 observations from July 2013 to end of 2016. About each housing type's amount, there were 8,891 of APT, 19,066 of BLD, and 5,070 of SUT.

Housing price was applied as the dependent variable. Other housing factors were applied as the independent variables. In this paper, conducted was a 5-fold cross validation.

3. Methodology

This paper employed 5 data mining techniques plus a GAM model. First, DT algorithm works by splitting a dataset to build a model that successfully classifies each record in terms of a target field or variable (see Woods and Kyral [31]). The splitting criteria for regression tree is $SS_T - (SS_L + SS_R)$, where $SS_T = \sum(y_i - \bar{y})^2$ is the sum of squares for the node, and SS_L, SS_R are the sums of squares for the right and left son respectively. The complexity parameter (cp) used for pruning the tress is 0.01. The minimum number of observations that must exist in a node is 20. The minimum number of observations in a terminal node is 7.

Second, MT is based on a divide-and-conquer approach in which it is possible to learn from a set of instances (see Witten et al. [28]). The output of a MT is represented by a tree-like structure in which it is possible to distinguish a root node, parent and child nodes, arches (or branches) and leaves (see Acciani et al. [1]). The greatest difference when compared with a decision tree is the leaf node's contents. In the model tree, each terminal node contains a linear regression model. Thus, it might provide a more precise estimation. The max number of rules is set to 100. The extrapolation parameter (being 1%) controls the extent to which predictions can fall outside the range of values observed in the training data.

Third, RF is an example of ensemble methods that combines a series of k base models (or trees) to co-create an improved composite model. Each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest (see Breiman [4]). The number of trees to grow is 500. To grow a single tree, selected are 6 random factors.

Fouth, ANN is an artificial intelligence model originally designed to replicate the human nervous system (see Bahia [2]). ANN consists of three main layers: input data (stimulation), the hidden, and the output. This paper uses a feed-forward neural networks with a single hidden layer. All the housing factors were taken as input. The housing price is output. The number of units in the hidden layer is 24 and the transfer function is sigmoid. The decay is 0. The final housing price prediction is detailed by the average of the output from an ensemble of 3 networks.

Fifth, the hedonic-based regression approach belongs to MR. There are many independent variables and one dependent variable in MR. Fixed independent variables derive the conditional expectations of the dependent variable that is an averaged value. Thus, MR is widely applied for prediction.

Lastly, Trevor Hastie and Robert Tibshirani developed GAM in 1986 (see Hastie [15]). GAM is a generalized linear model with a linear predictor involving a sum of the smooth function of covariates (see Wood [29]). GAM is an additive modeling technique in which the predictors' effectiveness is derived from smooth functions. In this research, the adopted smooth function are the thin plate regression splines. The argument of family in GAM is 'Gaussian' as the housing price will be on a normal distribution. The link function is set to 'identity' that is 'not using' a link function.

The research flow is shown in Figure 1 below. This paper utilizes the R 3.3.3 language. The vitalities of such factors (based on these data mining techniques except GAM) details the critical factors. Such techniques could be applied in the rminer 1.4.2 package (see Cortez [7]). Finally, GAM can be applied in the mgcv 1.8.17 package (see Wood [30]) to figure out non-linear relationships.

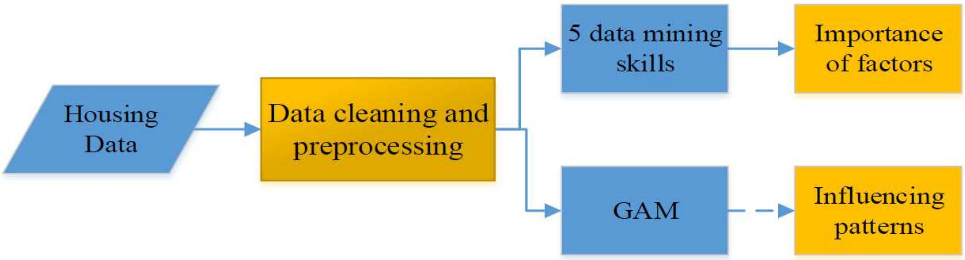


Figure 1: Research flow.

4. Results of Model Building

4.1. Data cleaning and preprocessing

Upon cleaning any potential inconsistencies, this paper recognizes more applicable factors by applying two measures on factors selection in this session. First, variance inflation factors (VIF) helps to investigate the collinearity. Second, the Akaike's information criterion (AIC) is adopted in forward, backward, and stepwise selection procedures to determine the modified model with the most suitable factors. The modified model is a model having the lowest adj_{R^2} with minimum number of factors. The VIF and AIC selection results are shown in Table 2 below. The number in parentheses after adj_{R^2} is the number of factors adopted.

Table 2: VIF and AIC selection Results.

Regression Types	APT	BLD	SUT
complete model adj_{R^2} (#)	0.5752(48)	0.7343(48)	0.8246(48)
VIF value	4	5	5
VIF model	0.5748(44)	0.7339(44)	0.8223(42)
Forward	0.5749(32)	0.734 (34)	0.8225(33)
AIC adj_{R^2} (#) backward	0.5749(33)	0.734 (33)	
stepwise	0.5749(31)		

The complete model encompasses all factors. The VIF model removes factors based on a stepwise selection on factors by using VIF value. This selection is suggested by Zainodin et al. [33] and uses VIF < 5 or even lower criteria as an exclusion rule that thus shows there is no serious collinearity problem. In this paper, the applied stop criterion used is to judge the adj_{R2} value that is obtained from regression in addition to the VIF value. If the adj_{R2} value of the new set of factors decreases much more than the previous set's, this paper considers the previous set of factors as a more suitable set of having no serious collinearity. For instance, the VIF threshold of buildings is 5 as shown in Table 3.

Table 3: Factors Selection Results by VIF for Type_BLD.

Model	Removed factors	VIF value	adj_{R^2}
Complete model	-	-	0.7343
with VIF < 10	target_tp num_of_convenience_store	15.42321 10.41656	0.7342
with VIF < 5	train_distance prk_sold	8.871036 5.137639	0.7339
with VIF < 4	num_of_shopping_mall flr_area num_of_H_school	4.316431 4.294665 4.093504	0.559

4.2. Descriptive statistics

Table 4 displays the means and standard deviations of the 3 housing types' respective housing prices. On average, buildings are the most expensive housing type in Taipei (worth over NTD 27 million), followed by apartments (about NTD 14 million), and suites (about NTD 9 million). The housing prices of apartments hold the largest range.

Table 4: Descriptive statistics of 3 housing types.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	SD
APT	10,000	9,450,000	13,000,000	13,910,000	16,800,000	99,000,000	6,890,961
BLD	396,400	14,900,000	22,750,000	27,660,000	34,100,000	536,000,000	21,082,126
SUT	546,000	6,125,000	8,500,000	9,205,000	11,500,000	40,000,000	4,328,882

4.3 Model performance

Table 5 shows the MAPE (Mean Absolute Percentage Error) and adj_{R2} values of the 3 housing types. In this paper, the factors for 3 housing types utilized are from the modified models. These modified models have different factors that are shown in Appendix without - label. RF performed best among apartments and buildings, and ANN performed best at suites. GAM explored more relationships and was not limited to linear analysis. The last model was DT. Even though adj_{R2} held no greater improvements, MAPE got better upon incorporating environmental factors.

Table 5: All adj_{R2}.

Factors	Measurement	Model	APT	BLD	SUT
Structural	MAPE	MR	72.15	38.02	18.42
		DT	68.89	38.58	23.50
		RF	67.01	25.57	15.46
		MT	69.81	26.92	16.14
		ANN	69.52	29.44	15.64
	adj_{R2}	MR	0.52	0.71	0.78
		DT	0.43	0.69	0.67
		RF	0.56	0.82	0.83
		MT	0.51	0.81	0.81
		ANN	0.54	0.81	0.84
Structural and environmental	MAPE	MR	61.19	36.63	16.94
		DT	68.72	38.59	23.54
		RF	58.30	23.02	13.12
		MT	58.93	24.84	14.05
		ANN	59.99	26.22	13.02
	adj_{R2}	MR	0.57	0.73	0.82
		DT	0.44	0.69	0.67
		RF	0.62	0.85	0.87
		MT	0.57	0.82	0.83
		ANN	0.58	0.84	0.88
		GAM	0.62	0.80	0.87

Apartments had the worst prediction of the 5 data mining models. This could be due to its aforesaid largest range. Buildings had a better prediction as urban growth has risen in time. Developers seek to provide residents with more benefits by decreasing transportation costs in a city like Taipei that includes housing, work, shopping, and other amenities altogether. These buildings may have been built due to urban renewal in areas that used to contain older apartments.

5 Discussion on Important Factors

5.1. The importance of housing factors

Each housing type has its own ranking or focused factors. Insights might be gained by applying the important values of each factor in a model that can be derived from rminer. All factors in a model constitute 100% importance. Thus, it is an easy way to distinguish the important factors by their importance or even rank them. This importance analysis extracts human understandable knowledge from supervised learning black box models such as ANN, SVM and RF (see Cortez and Embrechts [8]).

By applying the top 10 averages of the factors' importance, it became possible to detail the 3 housing types' respective housing prices as shown in Figure 2. The floor area is the most significant factor; the importance percentages for apartments, buildings, and suites are 46%, 63% and 43% respectively. None of the remaining factors was greater than 10%. Unequivocally, floor area was a dominant factor in terms of housing prices.

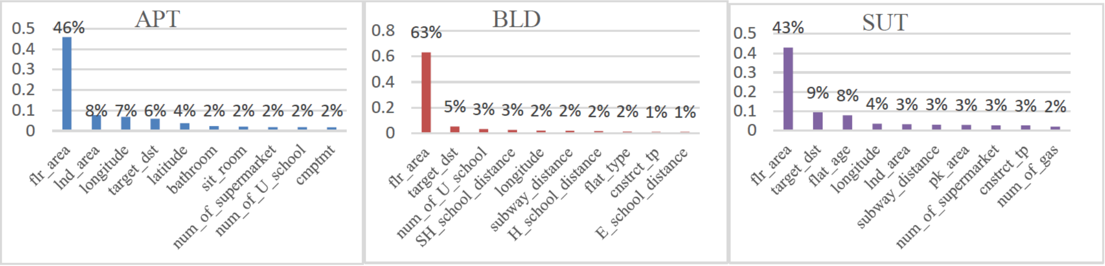
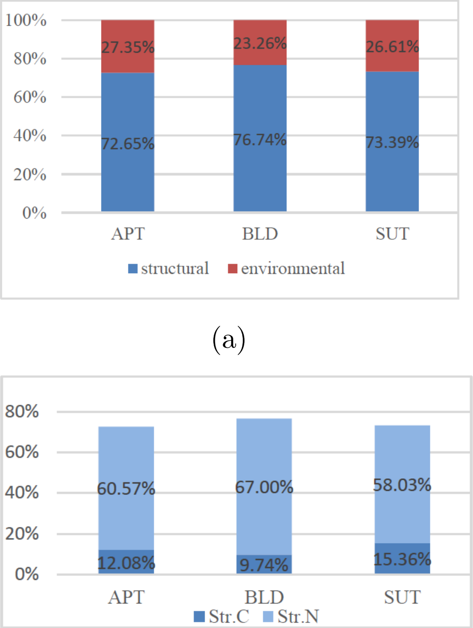


Figure 2: The factors importance of 3 housing types.

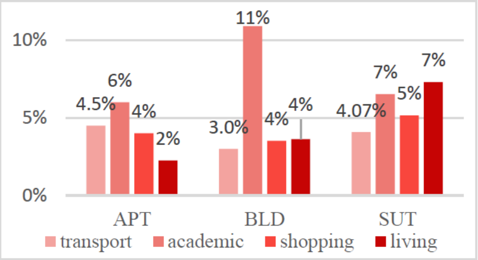
Buildings gave the prominence to academic-related factors (shown in Table 1), such as the number of universities, distances to senior high school, high school, and elementary ones. More important for suites were the relative conveneices of the subway, supermarkets, and gas stations.

It is vital to aggregate the factors to discover housing patterns when dealing with this paper's many factors. Figure 3(a) shows the combined importance of structural factors is over 70% in all the 3 housing types. Buildings consider structural factors most that comprise 76%. This indicates that the buildings' buyers prefer a house's more comfortable structure and buildings always have public utilities at mean time, normally 30%∼40% of floor area.



(a)

(b)



(c)

Figure 3: The aggregated importance of 3 housing types.

Differing structural factors types have differing importance. Factors include category and numeric ones. For instance, floor area, land area, the number of rooms, sitting room, bathroom, etc. are numeric types (Str.N). This numeric type's importance occupies over 58% as shown in Figure 3(b). The category type's importance (Str.C)—such as administrative district, type of land usage, etc.—occupies under 16%. The numeric type holds the buyer's attention much more.

Different homes types have different emphases on environmental factors. As shown in Figure 3(c), the environmental factors continue to divide. The results about both apartments and buildings agree that academic-related factors are more important than living, shopping, and transport-related ones. These are coupled with the strong need for a family. Namely as academic-related factors occupy over 10% for buildings. About environmental factors, the suites' living-related factors are considered as the first priority. Obviously, the needs of hospitals, public parks, and gas stations are more important for individuals with a single marital status. About transport, apartments raise more concerns than both buildings and suites. Such concerns are due to apartments occupying better locations than those two types of housing types in advance.

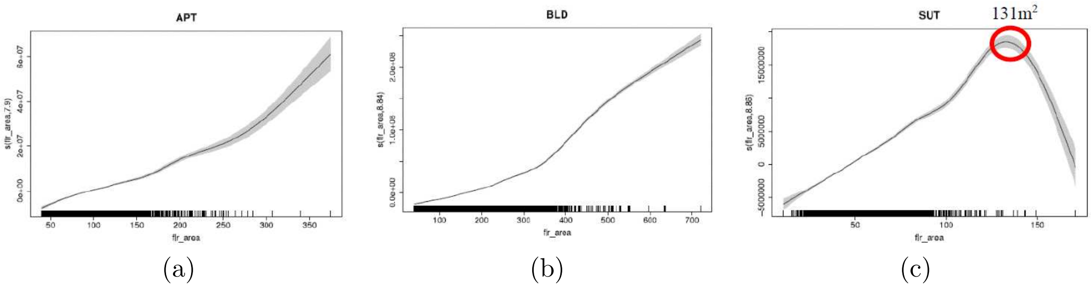
5.2. The relationship between housing prices and important factors

The following factors describing patterns includes the floor area, house's age, administrative districts, along with the proximities of the subway and post-secondary institutions. These interesting patterns hold very high attention in this paper.

5.2.1. Structural factors

As shown in Figure 4, all 3 housing types have similar relationships between housing price and floor area. Found was that the larger floor area correlated to a higher housing price. The empirical results also support these expectations, as proven by Lee et al. [18], who demonstrated that a larger living area had more households having a preference to purchase such a type.

However, there is a turning point at floor area 131m^2 in suites identified by a red circle as shown in Figure 4(c). After that point, the prices of the suites decline. Floor

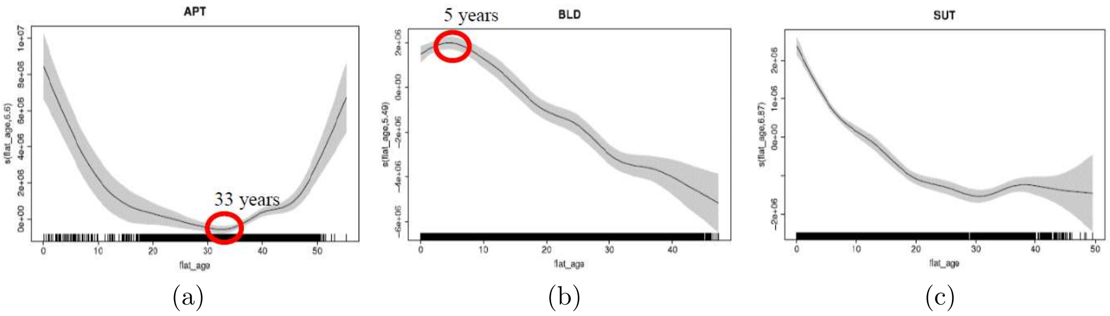


(a) (b) (c)

Figure 4: The relationship of floor area with housing prices.

areas that are too large do not increase the suites' prices, but instead decrease them. This phenomenon is not visible in multiple regression.

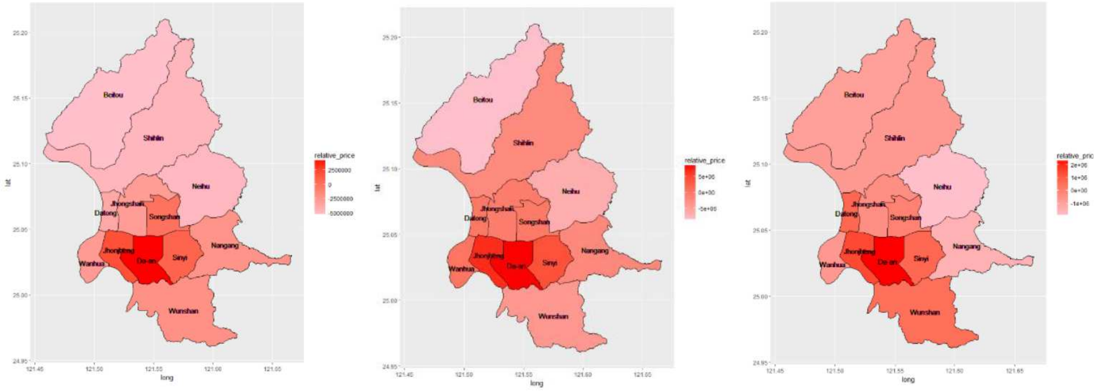
In terms of a house's age, both the buildings and suites' housing prices present an inverse ration, as shown in Figures 5(b) and (c). Yet, the apartments' ages present a U-shaped curve as shown in Figure 5(a). About aging speed, housing prices' decrease links to the older a home is (this is truer for apartments than for both buildings and suites), namely for apartments younger than 10 years old. However, there is a turning point at 33 years. After that point, the prices of apartments rebounded that could serve as one reason for urban renewal. There is a small "age penalty" for buildings younger than 5 years. Namely as building prices abnormally rise during this period. This phenomenon indicates that the best policy for selling buildings is at 5 years, but not newer.



(a) (b) (c)

Figure 5: The relationships f housing age with housing prices.

About administrative districts' housing prices, differing housing types have differing expanding configurations. The three most equally expensive administrative districts are the Daan, Jhonjheng, and Xinyi ones across all 3 housing types, as shown in Figure 6. For apartments, the three cheapest districts are Shihlin, Neihu, and Beitou, as shown in Figure 6(a). The housing prices of apartments in the southern part of Taipei are more expensive than in the north. The higher prices of apartments highly centralize around the Daan district. Compared to apartments, the higher housing prices of buildings radiate out of Daan district as shown in Figure 6(b). The higher prices of suites extend most, except for both the Nangang and Neihu districts, as shown in Figure 6(c). This configuration provides a clear concept and urban development trend.

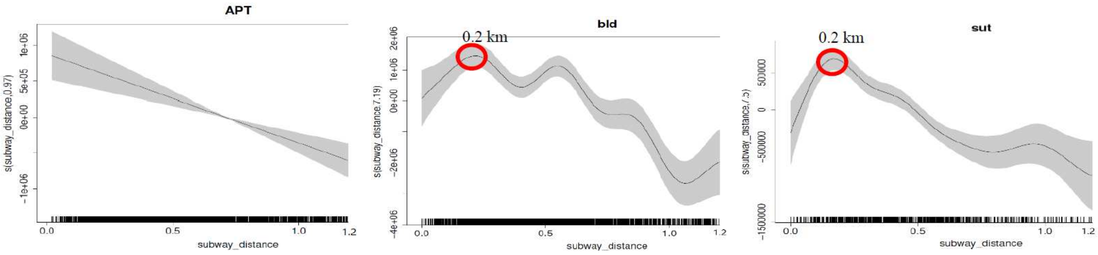


(a) (b) (c)

Figure 6: The relationship of administrative districts with housing prices.

5.2.2. Environmental factors

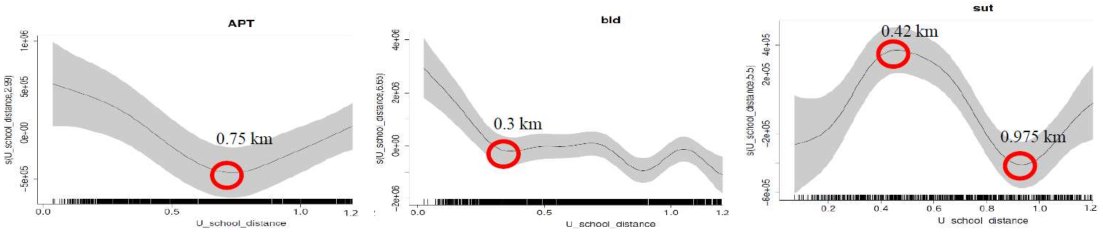
Diao et al. [10] recognized the significant capitalization of the closest MRT station into housing prices in Singapore, namely for households living within a 400-meter radius. In Taipei City, distances to the nearest MRT all comprise a significant and roughly inverse ratio with housing prices for all 3 housing types, as shown in Figure 7. And distance to the apartments' nearest MRT reveals a linear-like relationship as shown in Figure 7(a). There is a proximity penalty within about 0.2 km for buildings as shown in Figure 7(b). And the suites' prices increases with the distance to the nearest MRT until about 0.2 km, after which the suites' prices begin to fall as shown in Figure 7(c).



(a) (b) (c)

Figure 7: The relationships of subway proximity with housing prices.

The POIs' proximities can have both positive and negative effects on housing prices (see Kim et al. [16]). These phenomena also appear in university apartments and suites in Taipei, as shown in Figure 8. The universities' effects on nearby apartments are presented as a proximity premium. The closer an apartment is to a university, the higher its price until the distance exceeds 0.75 km, a turning point after which an apartment's price rises once again. About buildings, a proximity premium occurs within 0.3 km and prices go down sharply, after which prices go down softly. Generally, distances to the nearest university for buildings are roughly inverse ratio. About suites, a proximity penalty occurs within 0.42 km, after which prices go down until the distance exceeds 0.975 km. Sah et al. [21] also measured school proximity effects on nearby residential housing and thus recognized a "school proximity penalty."



(a) (b) (c)

Figure 8: The relationships of subway proximity with housing prices.

5.2.3. Overview of the trends of important housing factors

According to the factors' patterns, this paper partitioned the factors into 3 categories: increasing (I), decreasing (D), and non-linear (N) as listed in Table 6. Increasing factors are ones that have the direct ratios' main patterns. Conversely, decreasing factors own inverse ratios. And non-linear factors have non-linear relationships. Detailed non-linear patterns are shown in the Appendix.

It is easier to judge the patterns' effectiveness by either proximity or number. With the public park factor as an example, proximity is vital for both apartments and buildings, but the number of parks is not vital. The prices of apartments are increasing with the distance to parks. The parks' effectiveness in proximity for buildings are that housing prices are higher at places either near public parks or farther from public parks due to the 'V' main pattern. Whereas, the parks' proximity are not significant for suites, but the number of parks is increasing. Overall, the number of public parks increase the prices of suites.

Perhaps a more interesting phenomenon is the effectiveness of having both proximity and number of factors. Take buildings as an example, the relationship of distance to gas station is increasing and the number of gas station is increasing too. This implies that the better choice is to have the proximity of gas station in a far-enough distance and have as many gas stations as possible. The relationship of distance to senior high school is increasing and the number of senior high school is yet decreasing. This implies that the better choice is to have the proximity of senior high school in a far-enough distance and have as less senior high schools as possible. The relationship of distance to supermarket is decreasing and the number of supermarket is yet increasing. This implies that the better choice is to have a closer distance from supermarkets to the house and have more supermarkets in the vicinity of the house. Finally, the relationship of distance to elementary school is decreasing and the number of elementary school is decreasing too. This implies that the better choice is to have a closer distance from the elementary

Table 6: The patterns of housing factors.

Type	Category	Factors	APT	BLD	SUT
Structural	N	floor_area	I	I	N
		lnd_area	I	I	I
		flat_age	N	D	D
		flat_type	N	I	I
		sit_room	I	I	
		bathroom	I	I	
		room	D	I	I
		pk_area		D	D
		total_flat		I	N
Environmental	transport	subway_distance	D	D	D
		num_of_subway	I		
		train_distance	I		
		bus_distance			D
	academic	U_school_distance	N	D	D
		num_of_U_school	I		
		SH_school_distance		I	N
		num_of_SH_school	D	D	N
		H_school_distance			I
		num_of_H_school	I		I
		E_school_distance		D	
		num_of_E_school		D	D
		library_distance		D	I
		num_of_library	I	I	I
	shopping	supermarket_distance		D	D
		num_of_supermarket	I	I	N
		shopping_mall_distance		N	N
		num_ of_ shopping_mall		I	
		convenience_store_distance	I	D	
	living	gas_distance		I	D
		num_of_gas	I	I	
		park_distance	I	N	
		num_of_park			I
		hospital_distance		N	N
		num_of_hospital			N

*Empty: not significant factors

school to the house and have lesser elementary schools in vicinity of the house. Having an elementary school in the house's vicinity—not the number of—is more important.

6. Conclusion

Overall, this paper leverages different data mining methods and applies their advantages to detail the major relationships and factors involving apartments, buildings, and suites. As follows, this paper's main contributions are to:

(1) Incorporate 14 environmental factors that include proximity and number ones.

(2) Raise the vitalities of housing factors such as structure versus environment, along with distance versus number. Such factors allow the categorizations of the differing critical factors for all the 3 housing types.

(3) Adopt GAM to detail the relationships between crucial and interesting factors such as floor area, administrative districts, housing age, etc.

Appendix

1. If the p-value of one term is smaller than 0.001, this paper significantly believes that that term is "not zero," rejects H0, and marks ***. The criterion of such a small p-value helps determine the truly significant factors. The significant codes, including each notation and corresponding p-value, are denoted as the followings: *** 0.001, ** 0.01, * 0.05, . 0.1, + 1.
2. The notation of - refers to the removed factors from the complete models.
3. Type includes category (C) and numeric (N).
4. I stands for increasing; D for decreasing. Non-linear (N) is represented by shapes, such as $\vee$, $\wedge$, $\cup$, $\cap$, etc.

	Factors	Type	Description	APT	BLD	SUT
1	target_dst	C	Administrative districts: Songshan (1), Sinyi (2), Da-an (3), Jhongshan (4), Jhonjheng (5), Datong (6), Wanhua (7), Wunshan (8), Nangang (9), Neihu (10), Shihlin (11) and Beitou (12).	***	***	***
2	target_tp	C	With (1) or without (2) parking place	-	-	-
3	lnd_area	N	Occupied land area of the house(M^2)	*** I	*** I	*** I
4	lndusg_tp	C	Type of land usage: Residential (1), Commercial (2), Industrial (3), Others (4)	*	***	*
5	ym_sold	C	Year and month when the house has been sold	***	***	*
6	prk_sold	N	Number of parking places sold	-	-	-

	Factors	Type	Description	APT	BLD	SUT
7	flat_type	N	Floor numbering	*** L	*** I	*** I
8	total_flat	N	Total floor level of a building		*** I	*** ∼
9	cnstrct_tp	C	Types of construction methods: Reinforced concrete (1), Reinforced brick structure (2), Referring to building occupation permit (3) Steel reinforced concrete (5), Referring to other registrations (6).		***	***
10	flr_area	N	Area of the house (M^2)	*** I	*** I	*** ∧
11	room	N	Number of rooms	*** D	*** I	*** I
12	sit_room	N	Number of living and/or dining rooms	*** I	*** I	+
13	bathroom	N	Number of bathrooms	*** I	*** I	**
14	cmptmt	N	Compartment (1) or not (2)	*	+	-
15	mgt_cmt	C	Having (1) or not having (2) a management committee	+	***	*
16	pk_type	C	Parking type: On the ground floor (1), Lifting plane (2), Lifting machinery (3), Ramp (4), Ramp machinery (5), Tower (6), Others (7), No parking space (None)	**	***	*
17	pk_area	N	Parking area (M^2)	+	*** D	*** D
18	flat_age	N	Housing age (year)	*** ∪	*** D	*** D
19	latitude	N	latitude of the house	*** I	*** I	*** I
20	longitude	N	longitude of the house	*** I	*** ∩	*** I
21	subway_distance	N	Distance to the nearest MRT(km) within 1.2km	*** D	*** D	*** D
22	num_of_subway	N	Number of MRT within 1.2km	*** I	+	+
23	U_school_distance	N	Distance to the nearest university MRT(km) within 1.2km	*** ∨	*** D	*** D

	Factors	Type	Description	APT	BLD	SUT
24	num_of_U_school	N	Number of universities within 1.2km	*** I	**	**
25	SH_school_distance	N	Distance to the nearest senior high school(km) within 1.2km	*	*** I	*** ∨
26	num_of_SH_school	N	Number of senior high schools within 1.2km	*** D	*** D	*** ∩
27	H_school_distance	N	Distance to the nearest high school (km) within 1.2km	+	**	*** I
28	num_of_H_school	N	Number of high schools within 1.2km	*** I	+	*** I
29	E_school_distance	N	Distance to the nearest elementary(km) within 1.2km	*	*** D	+
30	num_of_E_school	N	Number of elementary within 1.2km	*	*** D	*** D
31	hospital_distance	N	Distance to the nearest hospital within 1.2km	*	*** ∨	*** ∨
32	num_of_hospital	N	Number of hospitals within 1.2km	+	*	*** ∧
33	supermarket_distance	N	Distance to the nearest supermarket(km) within 1.2km	**	*** D	*** D
34	num_of_supermarket	N	Number of supermarkets within 1.2km	*** I	*** I	*** ∩
35	train_distance	N	Distance to the nearest train station(km) within 1.2km	*** I	-	-
36	num_of_train	N	Number of train stations within 1.2km	-	**	.
37	shopping_mall_distance	C	Distance to the nearest shopping mall (km) within 1.2km	.	*** ∨	*** ∨
38	num_of_shopping_mall	N	Number of shopping malls within 1.2km	**	*** I	-
39	library_distance	N	Distance to the nearest library(km) within 1.2km	.	*** D	*** I
40	num_of_library	N	Number of libraries within 1.2km	*** I	*** I	*** I
41	gas_distance	N	Distance to the nearest gas station(km) within 1.2km	+	*** I	*** D
42	num_of_gas	N	Number of gas stations within 1.2km	*** I	*** I	*

	Factors	Type	Description	APT	BLD	SUT
43	park_distance	N	Distance to the nearest park(km) within 1.2km	*** I	*** ∨	+
44	num_of_park	N	Number of parks within 1.2km	+	+	*** I
45	bus_distance	N	Distance to the nearest bus station(km) within 1.2km	+	**	*** D
46	num_of_bus	N	Number of bus stations within 1.2km	+	+	*
47	convenience_store_distance	N	Distance to the nearest convenience store (km) within 1.2km	*** I	*** D	**
48	num_of_convenience_store	N	Number of convenience stores within 1.2km	-	-	-
49	price	N	Total price (NTD)			

References

[1] Acciani, C., Fucilli, V. and Sardaro, R. (2011). *Data Mining in Real Estate Appraisal: a Model Tree and Multivariate Adaptive Regression Spline Approach1*, Aestimum, 27-45.

[2] Bahia, I. S. H. (2013). *A Data Mining Model by Using ANN for Predicting Real Estate Market: Comparative Study*, International Journal of Intelligence Science, Vol.3, No.4, 162-169.

[3] Bency, A. J., Rallapalli, S., Ganti, R. K., Srivatsa, M. and Manjunath, B. S. (2017). *Beyond Spatial Auto-Regressive Models: Predicting Housing Prices with Satellite Imagery*, In Applications of Computer Vision (WACV), 2017 IEEE Winter Conference, 320-329.

[4] Breiman, L. (2001). *Random Forests*, Machine Learning, Vol.45, No.1, 5-32.

[5] Chen, M. and Wang, P. (2017). *A Roadmap to Determine the Important Factors of the House Value: a Case Study by Using Actual Price Registration Data of Taipei Housing Transactions*, Independent Journal of Management and Production, Vol.9, No.1, 245-261.

[6] Chiang, Y. H., Peng, T. C. and Chang, C. O. (2015). *The Nonlinear Effect of Convenience Stores on Residential Property prices: A Case Study of Taipei, Taiwan*, Habitat International, Vol.46, 82-90.

[7] Cortez, P. (2016). Package 'rminer', Available at https://cran.r-project.org/web/packages/rminer/rminer.pdf

[8] Cortez, P., and Embrechts, M. J. (2013). *Using Sensitivity Analysis and Visualization Techniques to Open Black Box Data Mining Models*, Information Sciences, Vol.225, 1-17.

[9] Diao, M. (2015). Selectivity, *Spatial Autocorrelation and the Valuation of Transit Accessibility*, Urban Studies, Vol.52, No.1, 159-177.

[10] Diao, M., Fan, Y. and Sing, T. F. (2017). *A New Mass Rapid Transit (MRT) Line Construction and Housing Wealth: Evidence from the Circle Line*, Journal of Infrastructure, Policy and Development, Vol.1, No.1, 64-89.

[11] Emrath, P. (2002). *Explaining House Prices*, Housing Economics, Vol.50, No.1, 9-13.

[12] Ferrari, L., Berlingerio, M., Calabrese, F. and Curtis-Davidson, B. (2012). *Measuring the Accessibility of Public Transport Using Pervasive Mobility Data*, IEEE Pervasive Computing, Vol 12, No.1, 26-33.

[13] Hammer, T. R., Coughlin, R. E. and Horn IV, E. T. (1974). *The Effect of a Large Urban Park on Real Estate Value*, Journal of the American Institute of Planners, Vol.40, No.4, 274-277.

[14] Hanink, D. M., Cromley, R. G. and Ebenstein, A. Y. (2012). *Spatial Variation in the Determinants of House Prices and Apartment Rents in China*, The Journal of Real Estate Finance and Economics, Vol.45, No.2, 347-363.

[15] Hastie, T. and Tibshirani, R. (1986). *Generalized Additive Models*, Statistical Science, Vol.1, No.3, 297-318.

[16] Kim, H. G., Hung, K. C. and Park, S. Y. (2015). *Determinants of Housing Prices in Hong Kong: a Box-Cox Quantile Regression Approach*, The Journal of Real Estate Finance and Economics, Vol.50, No.2, 270-287.

[17] Kim, K. and Lahr, M. L. (2014). *The impact of HudsonBergen Light Rail on Residential Property Appreciation*, Papers in Regional Science, Vol.93 (Suppl.), S79-S97.

[18] Lee, C. C., Ho, Y. M. and Chiu, H. Y. (2016). *Role of Personal Conditions, Housing Properties, Private Loans, and Housing Tenure Choice*, Habitat International, Vol.53, 301-311.

[19] Owusu-Edusei, K., Espey, M. and Lin, H. (2007). *Does Close Count? School Proximity, School Quality, and Residential Property Values*, Journal of Agricultural and Applied Economics, Vol.39, No.1, 211-221.

[20] Pope, D. G. and Pope, J. C. (2015). *When Walmart Comes to Town: Always Low Housing Prices? Always?*, Journal of Urban Economics, Vol.87, 1-13.

[21] Sah, V., Conroy, S. J. and Narwold, A. (2016). *Estimating School Proximity Effects on Housing Prices: the Importance of Robust Spatial Controls in Hedonic Estimations*, The Journal of Real Estate Finance and Economics, Vol.53, No.1, 50-76.

[22] Shyr, O., Andersson, D. E., Wang, J., Huang, T. and Liu, O. (2013). *Where do Home Buyers Pay Most for Relative Transit Accessibility? Hong Kong, Taipei and Kaohsiung Compared*, Urban Studies, Vol.50, No.12, 2553-2568.

[23] Visser, P., Van Dam, F. and Hooimeijer, P. (2008). *Residential Environment and Spatial Variation in House Prices in the Netherlands*, Tijdschrift voor economische en sociale geografie, Vol.99, No.3, 348-360.

[24] Wang, X. Y. (2006). *A Study on the Housing Hedonic Price of Shanghai Based on the Hedonic Model*, Shanghai: Tongji University.

[25] Wang, Y., Potoglou, D., Orford, S. and Gong, Y. (2015). *Bus Stop, Property Price and Land Value Tax: A Multilevel Hedonic Analysis with Quantile Calibration*, Land Use Policy, Vol.42, 381-391.

[26] Wen, H. and Tao, Y. (2015). *Polycentric Urban Structure and Housing Price in the Transitional China: Evidence from Hangzhou*, Habitat International, Vol.46, 138-146.

[27] Wen, H., Zhang, Y. and Zhang, L. (2014). *Do Educational Facilities Affect Housing Price? An Empirical Study in Hangzhou, China*, Habitat International, Vol.42, 155-163.

[28] Witten, I. H., Frank, E., Hall, M. A. and Pal, C. J. (2016). Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann.

[29] Wood, S. N. (2006). Generalized Additive Models: an Introduction with R, Boca Raton: Chapman & Hall/CRC, 119-138.

[30] Wood, S. N. (2018). Package 'mgcv',
Available at https://cran.r-project.org/web/packages/mgcv/mgcv.pdf

[31] Woods, E. and Kyral, E. (1997). Ovum Evaluates: Data Mining, London:Ovum.

[32] Wu, C., Ye, X., Du, Q. and Luo, P. (2017). *Spatial Effects of Accessibility to Parks on Housing Prices in Shenzhen, China*, Habitat International, Vol.63, 45-54.

[33] Zainodin, H. J., Khuneswari, G., Noraini, A. and Haider, F. A. A. (2015). *Selected Model Systematic Sequence via Variance Inflationary Factor*, International Journal of Applied Physics and Mathematics, Vol.5, No.2, 105-114.

[34] Zoppi, C., Argiolas, M. and Lai, S. (2015). *Factors Influencing the Value of Houses: Estimates for the City of Cagliari, Italy*, Land Use Policy, Vol.42, 367-380.

Graduate Institute of Business Administration, Fu Jen Catholic University, Taiwan.

E-mail: kmpeterwang@gmail.com

Major area(s): Data mining, taxation.

Graduate Institute of Business Administration, Fu Jen Catholic University, Taiwan.

E-mail: 081438@mail.fju.edu.tw

Major area(s): Production and operations management, reliability and maintainability, data mining.

(Received February 2018; accepted December 2018)