Journal of Data Science,17(1). P. 131 - 144,2019
DOI:10.6339/JDS.201901_17(1).0006

# THE NON-LINEAR RELATIONSHIPS OF NUMERIC FACTORS ON HOUSING PRICES BY USING GAM

Pei-De Wang[*1], Mingchin Chen[2]

[1,2]*Graduate Institute of Business Administration, Fu Jen Catholic University*

## Abstract

Most research on housing price modeling utilize linear regression models. These research mostly describe the actual contribution of factors in a linear way on magnitude, including positive or negative. The goal of this paper is to identify the non-linear 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 non-linear patterns present in the combination manner of a sequence of uptrends and downtrends that are derived from Generalized Additive Models (GAM).

**Keywords**: Non-linear pattern, GAM, housing factor

[*] Corresponding author: Pei-De Wang
email: kmpeterwang@gmail.com

# 1 Introduction and motivation

Similar types of homes command varying prices in different regions and neighborhoods (see Visser et al. [1]). Therefore, it is necessary to encompass the environmental factors for investigating housing factors. Hanink et al. [2] and Zoppi et al. [3] 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 also adopts both structural and environmental factors to analyze housing prices in Taipei city.

For describing environmental factors, based on the works of Chen and Wang [4], this paper also applied are proximity (see Sah et al. [5]) and a number of specific factors (see Wang et al. [6]). Proximity represents the degree of convenience to reach a POI and the number of specific factors shows the maturity of factors in the vicinity.

There are many studies working on environmental factors. Kim and Lahr [7] and Shyr et al. [8] worked on Metropolitan Rapid Transit (MRT) systems. Wen et al. [9] , Wang [10] and Owusu-Edusi et al. [11] found the distance to different types of schools are significant. Emrath [12] ,and Pope and Pope [13] focused on shopping centers. Public parks were held the attention by Wu et al. [14] and Hammer et al. [15]. Chiang et al. [16] were interested in convenience stores.

The aforesaid studies discuss on specific environmental characteristics. Most of such studies utilize linear regression models. This paper encompasses more factors in order to figure out the interesting patterns of factors of 3 major housing types in Taipei.

# 2 Data

This study's fundamental data was downloaded from the Taiwan Actual Price Registration (APR). The APR's factors are treated 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.

Following Chen and Wang [4], the major housing types addressed in this paper also included apartments (APT); buildings (BLD); and suites (SUT). In total, this paper applies 49 factors that are listed in the Appendix 1. 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

Trevor Hastie and Robert Tibshirani developed GAM in 1986 (see Hastie [17]). GAM is a generalized linear model with a linear predictor involving a sum of the smooth function of covariates (see Wood [18]). GAM is an additive modeling technique in which the predictors' effectiveness is derived from smooth functions. In general the model has a structure like equation:

$$g\big(E(Y_i)\big) = \alpha + X_i^*\theta + \sum_{i=1}^{i=p} \beta_i s_i(x_i) + \varepsilon_i \qquad (1)$$

where $Y_i$~ some exponential family distributions (Gaussian, Gamma, Poisson and Binomial) and is a the dependent variable, $E(Y_i)$ represents the expected value, g() denotes the link function that links $E(Y_i)$ to the predictor variables $x_i$, α denotes the intercept or mean, $X_i^*$ is a row of the model matrix for parametric model components (category), θ is the corresponding parameter vector, $s_i(x_i)$is the smooth, nonparametric function, $\beta_i$ are coefficients of the smooth, p is the number of factors (numeric), and $\varepsilon_i$ is the residuals.

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. GAM can be applied in the mgcv 1.8.17 package (see Wood [19]) to figure out non-linear relationships.

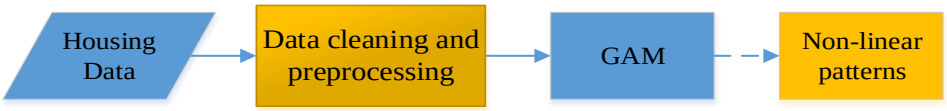



Figure 1: Research flow

## 4 Results of model building

### 4.1 Factors selection procedures

For digging out more applicable factors, this paper utilizes two factors selection procedures. First, variance inflation factors (VIF) helps to discover the factors having higher collinearity and then to remove them. 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 defined in this paper is a model having the lowest adj_R2 with fewest number of factors. The results of factors selection are depicted in Table 1 below. The number in parentheses after adj_R2 is the number of factors chose.

Table 1: 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) |
| AIC adj_$R^2$ (#) | Forward | 0.5749(32) | 0.734 (34) | 0.8225(33) |
| | backward | 0.5749(33) | 0.734 (33) | |
| | stepwise | 0.5749(31) | | |

The complete model includes 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. [20] and uses VIF < 5 or even lower criteria as an exclusion rule that thus shows there is no serious collinearity problem.

### 4.2 Descriptive Statistics

Table 2 show 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 2: 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 3 demonstrates 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.

Table 3: All adj-$R^2$

| APT | BLD | SUT |
|---|---|---|
| 0.62 | 0.80 | 0.87 |

## 5 The non-linear patterns

Chen and Wang [4] have shown some patterns, such as floor area, housing age, etc. This paper presents other interesting patterns, especially those are non-linear ones having the shapes of V, Λ, L, etc.

### 5.1 Apartments

The pattern of floor numbering for apartments represents the 'L' shape as shown in Figure 2. First floor is the most valuable and second floor is the cheapest. Other floors are about the same. The positive impact on housing price is first floor, others are all negative.

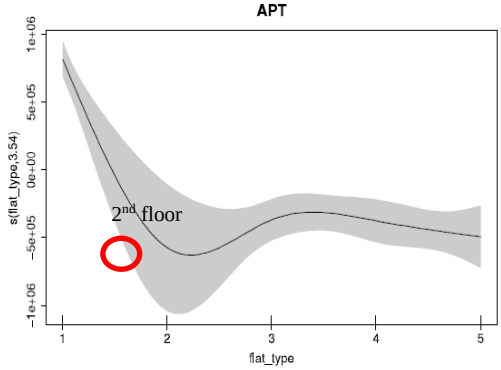
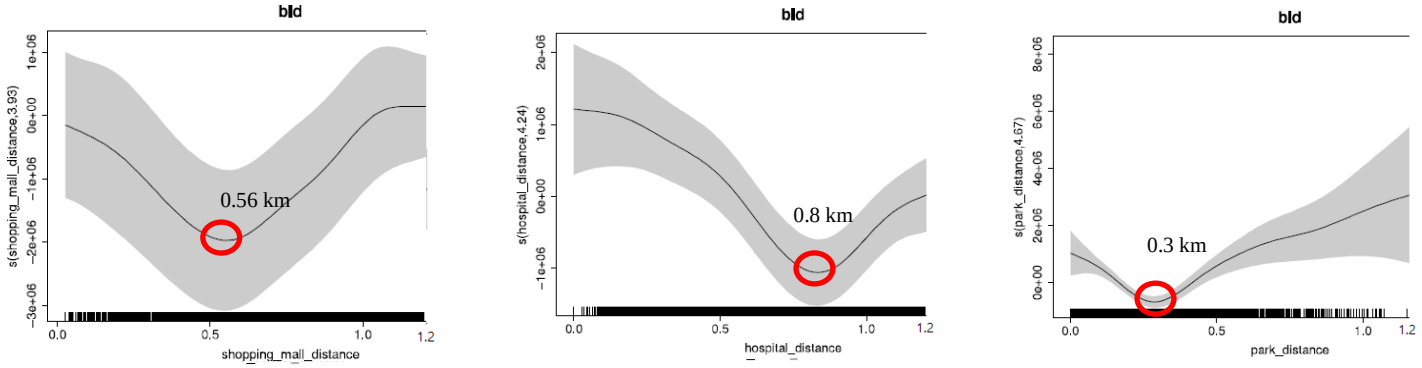



Figure 2: The non-linear patterns of APT

### 5.2 Buildings

The patterns of distances to shopping malls, hospital and parks all have 'V' shapes for buildings as shown in Figure 3 (a), (b) and (c) respectively. These factors' effectiveness in proximity for buildings are that housing prices are higher at places either near these factors or farther from them. The lowest housing prices appear at the distance of 0.56 km, 0.8 km, and 0.3 km far from shopping malls, hospital and parks respectively.




(a) (b) (c)

Figure 3: The non-linear patterns of BLD

### 5.3 Suites

The patterns of distances to senior high schools, hospital and shopping malls also have 'V' shapes for suites as shown in Figure 4 (a), (b) and (c) respectively. The lowest housing prices locate at 0.55 km, 0.8 km, and 0.6 km far from senior high schools, hospital and shopping malls respectively.

The patterns of the numbers of senior high schools, hospitals and supermarkets for suites have 'Ω' shapes as shown in Figure 4 (d), (e) and (f) respectively. These factors' effectiveness in the number for suites are that housing prices are higher with the proper number of such factors. In other words, too few or too many such factors both lead to lower housing prices. The proper numbers of senior high schools, hospital and shopping malls are 2 to 6, 3 and 15 separately.

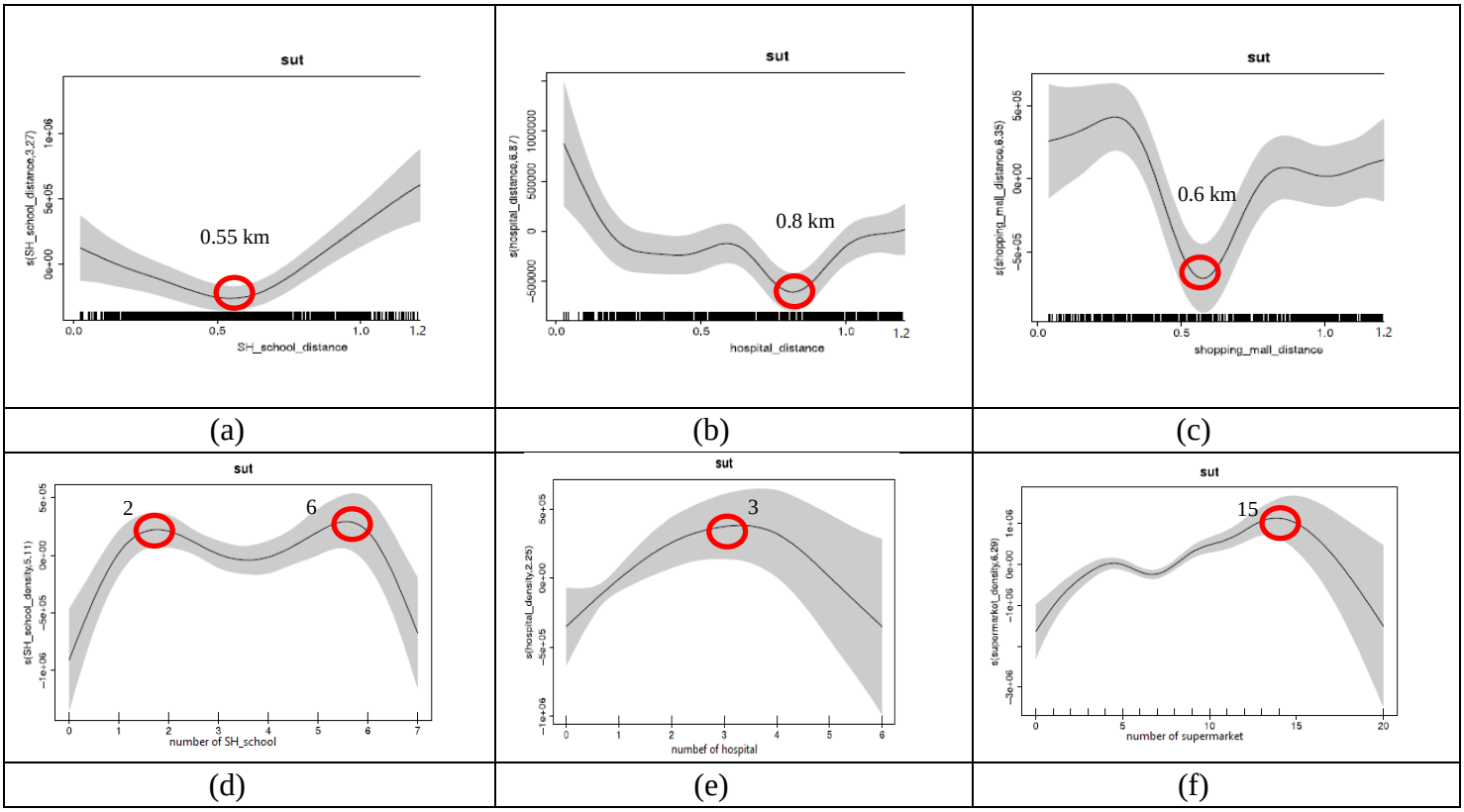



Figure 4: The non-linear patterns of SUT

## 6 Summary

Overall, this paper leverages GAM and applies its advantages to detail the major relationships and factors involving apartments, buildings, and suites. This paper's main contribution is to adopt GAM to detail the relationships between interesting factors. Those relationships are nonlinear.

## References

[1] 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* **99**, No. 3, 348-360.

[2] 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* **45**, No. 2, 347-363.

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

[4] Wang, P., & Chen, M. The impact of environmental factors on housing prices: A case study of Taipei housing transactions. *International Journal of Information and Management Sciences (IJIMS)*, **submitted**.

[5] 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* **53**, No. 1, 50-76.

[6] 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* **42**, 381-391.

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

[8] 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* **50**, No. 12, 2553-2568.

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

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

[11] 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* **39**, No. 1, 211-221.

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

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

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

[16] 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* **40**, No. 4, 274-277.

[17] 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* **46**, 82-90.

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

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

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

[21] 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* **5**, No. 2, 105-114.

## 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 $H_0$, 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 V, Λ, U, Ω, 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 | *** | *** | * |

| | Factors | Type | Description | APT | BLD | SUT |
|---|---|---|---|---|---|---|
| 6 | prk_sold | N | Number of parking places sold | - | - | - |
| 7 | flat_type | N | Floor numbering | *** L | *** I | *** I |
| 8 | total_flat | N | Total floor level of a building | . | *** I | *** S |
| 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 | *** Λ |
| 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 | + | *** | * |

| | Factors | Type | Description | APT | BLD | SUT |
|---|---|---|---|---|---|---|
| 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 (M²) | + | *** D | *** D |
| 18 | flat_age | N | Housing age (year) | *** U | *** D | ** D |
| 19 | latitude | N | latitude of the house | *** I | *** I | *** I |
| 20 | longitude | N | longitude of the house | *** I | *** Ո | *** 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 | *** V | *** D | *** D |
| 24 | num_of_U_school | N | Number of universities within 1.2km | *** I | ** | ** |
| 25 | SH_school_distance | N | Distance to the nearest senior high | * | *** I | *** V |

| | Factors | Type | Description | APT | BLD | SUT |
|---|---|---|---|---|---|---|
| | | | school(km) within 1.2km | | | |
| 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 | * | *** V | *** V |
| 32 | num_of_hospital | N | Number of hospitals within 1.2km | + | * | *** Λ |
| 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 | *** I | - | - |

| | Factors | Type | Description | APT | BLD | SUT |
|---|---|---|---|---|---|---|
| | | | station(km) within 1.2km | | | |
| 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 | . | *** V | *** V |
| 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 | * |
| 43 | park_distance | N | Distance to the nearest park(km) within 1.2km | *** I | *** V | + |
| 44 | num_of_park | N | Number of parks within 1.2km | + | + | *** I |
| 45 | bus_distance | N | Distance to the nearest bus | + | ** | *** D |

| | Factors | Type | Description | APT | BLD | SUT |
|---|---|---|---|---|---|---|
| | | | station(km) within 1.2km | | | |
| 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) | | | |