Geographically weighted machine learning for modeling spatial heterogeneity in off-campus student housing rents

SHIH-HUNG YANG[1*], HAN-LIANG LIN[1],YU-TUNG LIN[1], YA-ZHEN CAI[1], MEI-KUAN LI[1], HSIAO-TZU HSIUNG[1], WEN-YING CHEN[1], YEN-LIN CHEN[1], YU-HAN NIEN[2]
[1]Department of Urban Planning, National Cheng Kung University, 1, University Rd., East Dist., Tainan City 701401, Taiwan
[2]Department of Architecture, National Cheng Kung University, 1, University Rd., East Dist., Tainan City 701401, Taiwan
* Corresponding author: p26134016@gs.ncku.edu.tw

Abstract
This study employs Geographically Weighted Machine Learning (GW-ML) to analyze spatial patterns and influencing factors of off-campus rental prices near colleges in Taiwan. Using Moran's I, LISA, and K-means clustering, the research identifies rent clusters and segments the market. OLS, GWR, and GW-ML models were compared in terms of their performance, with Geographically Weighted LightGBM showing the best predictive accuracy. Key factors affecting rents include proximity to colleges, housing age, and access to public transportation. The GW-ML framework captures spatial heterogeneity and nonlinear relationships more effectively than traditional models. The results reveal that rental properties in city centres place more importance on rental area, while those in suburban areas focus more on building age for improving housing affordability and guiding policy in college neighborhoods.

Keywords: geographically weighted machine learning; machine learning; housing rents

1. introduction

In recent years, rising housing costs in Taiwan have highlighted challenges in property transactions, rental housing, and housing justice. For students living away from home, rising rents create financial burdens, compounded by information asymmetry and lack of market transparency, making the rental process stressful and complex. School districts, characterized by high rental demand, experience price fluctuations influenced by economic, spatial, and resource factors. Despite their importance, limited research addresses the spatial characteristics and factors influencing rental prices. This study investigates rental price spatial correlations and influencing factors near universities, thereby offering insights into rental market management.

2. Research Methodology and data sources

2.1. Study Area

This study investigates the spatial distribution characteristics of rental prices in areas surrounding school districts. To comprehensively analyse the factors influencing the rental market, the study encompasses eight administrative districts in Tainan City: West Central District, North District, East District, South District, Anping District, Annan District, Yongkang District, and Rende District. They are shown in Fig. 1.

Fig. 1. Location of study area.

2.2. Methods of Analysis

2.2.1. Global Spatial Autocorrelation

Moran's index method was adopted to evaluate the overall level of spatial autocorrelation of a specific attribute across a geographical area. It describes the spatial distribution characteristics of an observed variable, indicating whether it exhibits clustering, dispersion, or randomness within the study area. The calculation formula is as follows:

$$I = \frac{n}{\sum_i \sum_j W_{ij}} \frac{\sum_i \sum_j W_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2} \quad (1)$$

Where *n* represents the number of samples; x_i and x_j denote the attribute values at locations *i* and *j*; $\bar{x}$ represents the mean attribute value across all samples; W_{ij} is the spatial weight matrix between locations *i* and *j*.

The value of Moran's I ranges from -1 to 1. Moran 's I > 0 indicates a positive spatial autocorrelation, meaning the attribute demonstrates clustering. Conversely, Moran 's I < 0 suggests negative spatial autocorrelation, which is indicative of a dispersed pattern. When Moran's I = 0, it implies no spatial autocorrelation, signifying that the spatial distribution of the attribute is random.

2.2.2. Local Spatial Autocorrelation

Local Indicators of Spatial Association (LISA) is a measure of local spatial autocorrelation used to infer the extent of clustering within specific areas. It serves as a set of local indicators to identify the boundaries of clustered regions. The significance test primarily evaluates whether the clustering of spatial elements within a given study area is statistically significant. Higher levels of significance indicate more pronounced spatial clustering. Compared to the global indicator Moran's I, LISA focuses on detecting the variation of a specific area or unit and its association with neighbouring units. Proposed by Anselin (1995), the LISA indicator extends the concept of Global Moran's I by incorporating spatial weight considerations and calculating the Moran index for each spatial unit. The formula for calculating LISA can be expressed as follows:

$$I_i = Z_i^T \sum_j W_{ij} Z_j^T \quad (2)$$

Where Z_i^T and Z_j^T represent the standardized observed values (mean-centered) in vector form. W_{ij} is the spatial weight matrix indicating the adjacency relationship between spatial units *i* and *j*, W_{ij}=1: *i* and *j* are adjacent; =0: and are not adjacent.

Anselin (1995) categorized LISA into four quadrants based on the degree of spatial clustering, as shown in Fig. 2.: high-high (HH), low-high (LH), low-low (LL), and high-low (HL). This classification facilitates the analysis of spatial clustering intensity. HH and LL are also referred to as hot spots and cold spots, respectively. The quadrant diagram aids in analyzing the relationship between high or low observed values in a specific area and those in surrounding areas. It identifies spatial association patterns across the study area, highlighting local spatial anomalies and instabilities. Typically, the LH and HL categories reflect anomalies that are inconsistent or contradictory to their surrounding areas (Gray, 2011). These values indicate significant deviations and provide insight into local spatial heterogeneity.

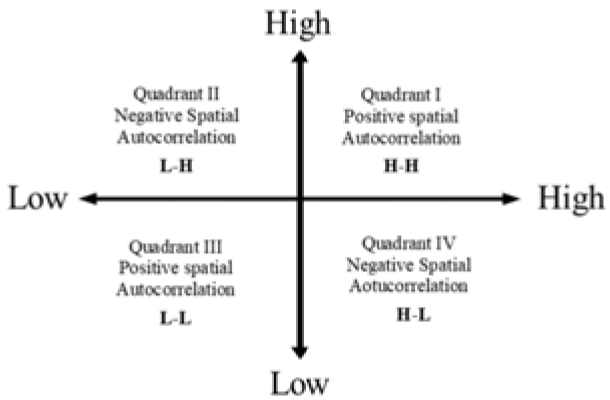


Fig. 2. LISA Local Spatial Autocorrelation Quadrant Map.
Source: Wang et al., 2019

2.2.3. OLS and HPM

The housing characteristic price model (HPM) often uses OLS for estimation, using OLS as the basic estimation technique to study how different characteristics of housing affect market prices. Its basic assumption is that housing prices can be predicted through the attributes of the unit's structure and location as independent variables, and it discusses housing prices as a heterogeneous product from a macro perspective (Ucal & Kaplan, 2020)using a hedonic price model with a large dataset and a single variable for locational attributes. Research Design & Methods: The analysis of consequent housing prices in İstanbul's counties with hedonic price modelling and the extrapolation of results by comparing the prices to the human development level of counties. We use multiple regression and Ordinary Least Squares (OLS. Many studies have proposed that various characteristics do indeed affect rent, and a regression model framework has been established to create a system for housing prices, further analyzing the effective analytical structure and characteristics of regional real estate prices. According to the research by Sirmans et al. (1989) , the impact of various local amenities and building characteristics on housing prices was studied, Davis et al. (2023) assessed the impact of green infrastructure on housing prices through a characteristic price model, including analysis of artificial wastewater wetlands, crystal gardens, ponds, etc., to further understand the property value of the ecosystem services they provide. The OLS calculation formula is as follow:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon \quad (3)$$

The HPM calculation formula is as:

$$lnP = \alpha_0 + \sum_{i=1}^{n} \alpha_i z_i + \epsilon \quad (4)$$

Where *P* represent the asset price, *lnP* is the natural logarithm of *P*, Z_i is the price characteristic, α_0 is the intercept, α_i denotes the coefficient of the price characteristic, and ϵ is the error term.

OLS serves as a method for estimating and calculating global regression coefficients, testing whether there is a relationship between variables and the dependent variable, and checking the independence of residuals. It has been widely used in real estate price research, but it cannot explain the interactions of spatial data with heterogeneous characteristics. Therefore, most studies use GWR to compensate for the shortcomings of OLS (Koohpayma & Argany, 2021; Liu & Strobl, 2023).

2.2.4. Geographically Weighted Regression

Geographically Weighted Regression (GWR) is a local analysis technique for spatial heterogeneity that is non-stationary and conforms to the first law of geography, primarily used to understand the relationships of spatial changes, calculating each local regression through data estimated by distance decay, using the function of spatial location as coefficients in the GWR model, and allowing them to vary spatially within the study area (Fotheringham et al., 2002), thereby identifying differences in the impact and intensity of features across the overall space. For example, Gutiérrez-Posada et al. (2017) use GWR to understand whether the determining factors of local population growth in Spain have uniform or non-uniform effects over time and space, analysing the local effects of spatial differences to plan appropriate local policies.

GWR has the core concept that regression coefficients are not fixed constants but vary according to different geographical spaces. Thus, it can capture the relationship between dependent and independent variables in space and the non-stationarity of housing prices, achieving higher accuracy and fitting effects than classical regression models(Cellmer et al., 2020; Iliopoulou & Feloni, 2022)as well as of real estate prices and values. The main aim of the study was to analyse an effect of these socio-demographic and environmental factors on average housing property prices and on the number of transactions in a spatial approach. In previous research conducted on a national scale, usually all variables were treated in a similar way, i.e., as global or local variables. During the research, an attempt was also made to answer the question of which of the variables adopted for analysis have a local impact on prices and market activity, and which are global. The study was conducted in Poland and used data from the year 2018 on 380 counties (Local Administrative Units. The impact range of spatial heterogeneity and observation points has now been widely applied in various fields, such as economics, environmental science, public health, and the housing prices. The key point to note is the need to choose appropriate distance weighting functions and bandwidth (Brunsdon et al., 1996), and it can only describe continuous heterogeneity, still requiring other methods to assist in describing complete spatial heterogeneity, such as spatial autocorrelation analysis. The GWR calculation formula is as follows:

$$y_i = B_0(u_i, v_i) + \sum_{k=1}^{p} B_k(u_i, v_i) x_{ik} + \varepsilon_i \quad (5)$$

Where y_i is the rental price at the *i*-th observation point; B_0 (u_i , v_i)is the intercept term at the *i*-th observation point, varying with spatial location (u_i , v_i); B_k (u_i , v_i) represents the regression coefficient of the *k*-th independent variable at the *i*-th observation point; x_{ik} is the value of the -th independent variable at the *k*-th observation point; ε_i is the error term at the *i*-th observation point.

To estimate the location-specific coefficients in equation (5), GWR constructs a diagonal weight matrix W_i centred on each observation *i*. The most widely adopted choice is the Gaussian kernel, which assigns weights according to

$$W_{ij} = \exp\left(-\frac{d_{ij}^2}{2h^2}\right) \tag{6}$$

where d_{ij} denotes the Euclidean distance between observations *i* and *j*, and *h* is the bandwidth that produces smoother spatial surfaces, whereas a smaller value allows stronger local variation. The optimal bandwidth is typically selected by leave-one-out cross-validation or the corrected Akaike information criterion (AICc) (Brunsdon et al., 1996). The Gaussian kernel provides continuous, strictly positive weights, which stabilise local estimations and avoid singular design matrices (Fotheringham et al., 2002).

2.2.5. Applications of Machine Learning in Rental Market Analysis

The geographical distribution and influencing factors of rental market prices have been focal research areas. Traditional regression models, such as linear regression and geographically weighted regression (GWR), face limitations in handling high-dimensional features and non-linear relationships. Machine learning models, including Random Forest, Support Vector Regression (SVR), KNN, and gradient boosting frameworks (XGBoost, LightGBM), are increasingly replacing traditional methods, offering improved prediction accuracy and efficiency.

XGBoost and LightGBM stand out in particular. Research indicates that XGBoost effectively avoids overfitting and enhances computational efficiency through regularization and parallel computation, while LightGBM further accelerates large-scale data processing with improved accuracy (Chen & Guestrin, 2016; Ke et al., 2017). For instance, Takahiro Yoshida et al. (2021) demonstrated that XGBoost outperforms traditional regression models in accuracy and efficiency on large datasets. Similarly, Yue Ming (2020) verified XGBoost's precise prediction of Chengdu's housing prices, with a mean squared error (MSE) of just 0.04. These studies highlight the significant advantages of XGBoost and LightGBM in handling non-linear and spatially heterogeneous rental market data.

Taking advantage of its ability to evaluate spatial impacts and analyze nonlinear relationships, geographically weighted machine learning methods have been applied in research recently. Wang et al. (2024) used geographically weighted random forest (GW-RF) model to explore the local associations between crash frequency and various influencing factors in the US. Yang et al. (2023) introduced geographically weighted support vector regression (GWSVR) for analysing soil metal chromium (Cr) and PM2.5 concentrations. Assigning sample weights according to Gaussian kernel values—paralleling Geographically Weighted Regression—enhances model performance. Hence, the present study employs a geographically weighted LightGBM to describe spatial heterogeneity in the rental housing market. The geographically weighted approach is expected to outperform a LightGBM model that does not incorporate geographic weighting.

2.3. The factor selection and data sources

This study aims to explore the spatial distribution patterns of rental prices within school districts and examine the impact of various influencing factors on these rental prices, as shown in Table 1. Taking rental price as the research subject, the study adopts price per ping as the measurement standard for rental prices. The data are sourced from the Real Estate Transaction Price Inquiry System of the Ministry of the Interior, encompassing rental housing transaction records within the study area. The research period spans from January to September 2024, covering a complete three-season interval. The dependent variable for the GWR analysis is defined as rental price (NTD/ping).

In selecting potential influencing factors, this study considers the rental demands of student populations, which are typically constrained by commuting distance to schools and daily convenience. Starting from this perspective, the research identifies and analyzes nine potential factors that may affect rental prices, based on the residential convenience and accessibility needs of students, as well as the functional and structural characteristics of rental housing. These factors include: (1) distance to the nearest university, (2) distance to nearest convenience store, (3) distance to nearest supermarket, (4) distance to the nearest bus stop, (5) distance to the nearest bike stop, (6) house age, (7) rental area, (8) elevators, and (9) furniture. Table 1 shows the details of each factor. These factors serve as independent variables in the GWR analysis, which is employed to calculate the relative impact and spatial heterogeneity of these factors on rental prices within school districts.

Table 1. Potential impact factors on rental prices and valuation methods

Categories	Variables	Measurements	Data sources	Reference
Educational characteristics	Distance to nearest university	Straight line distance (meters)	Ministry of Education	Wen et al.,2015 Wen et al.,2018
Location characteristics	Distance to nearest convenience store	Straight line distance (meters)	Administration of Commerce	Ti-Ching Peng, 2021
	Distance to nearest supermarket	Straight line distance (meters)	Administration of Commerce	Wang et al., 2022
	Distance to nearest bus stop	Straight line distance (meters)	Gist Transport data	Villada-Medina, 2022
	Distance to nearest bike stop	Straight line distance (meters)	Gist Transport data	Jiang et al., 2017
Buildings characteristics	House age	year	Dept of Land Administration	Cao et al., 2017 Wen et al.,2015
	Rental area	Number of square meters	Dept of Land Administration	Cao et al., 2017 Wen et al.,2015
	Elevator	yes=1 no=0	Dept of Land Administration	Ti-Ching Peng, 2021
	Furniture	yes=1 no=0	Dept of Land Administration	Zhao et al., 2024

3. Result

3.1. Spatial Autocorrelation analysis

3.1.1. Global Moran's I

Moran's I is used to determine whether the distribution characteristics of rent unit price in the overall space are agglomeration, dispersion, or random distribution. The value is between -1 and 1. A positive value indicates that the rental unit price is clustered in space, and near points have similar prices; a negative value indicates that the unit price of adjacent points has a large difference, which indicates a scattered data distribution; approaching 0 means that the space is randomly distributed.

The results of global Moran's I are presented in Fig. 3. It shows that the Moran's I value of the rental unit price in the school district is 0.3402, the Z value is 45.9976, and the P value is 0.000, passing the significance test. It shows that there is indeed spatial autocorrelation of the rental unit price. To have a more straightforward picture presentation for the following LISA analysis, we conduct Moran's I at the neighborhood scale to confirm whether the phenomenon of spatial autocorrelation still exists if the rental unit price is presented using polygon data. Calculate the average unit rental price in the neighborhood using the representative data here. The results are shown in Fig. 4. Moran's I value is 0.101, Z value is 3.684, and P value is 0.000, passing the significance test, the space shows a positive correlation. Although the value is closer to a random distribution, the significance test still shows a non-random spatial distribution of rental unit prices, and the polygon data can help present a clearer picture.

Both point data and polygon data illustrate that rents are spatially autocorrelated. The factors affecting unit rents may vary significantly across different regions. When modeling unit rents, the bandwidth and intensity of the impact of each variable on the local scale should be considered.

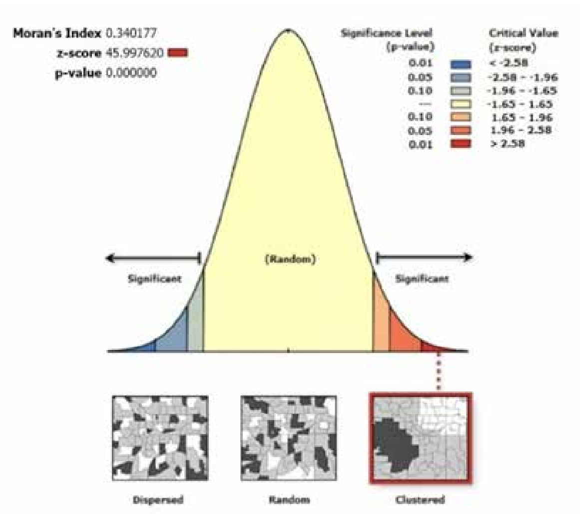


Fig. 3. Rental price data in point format; Results of spatial autocorrelation analysis by ARCGIS software.

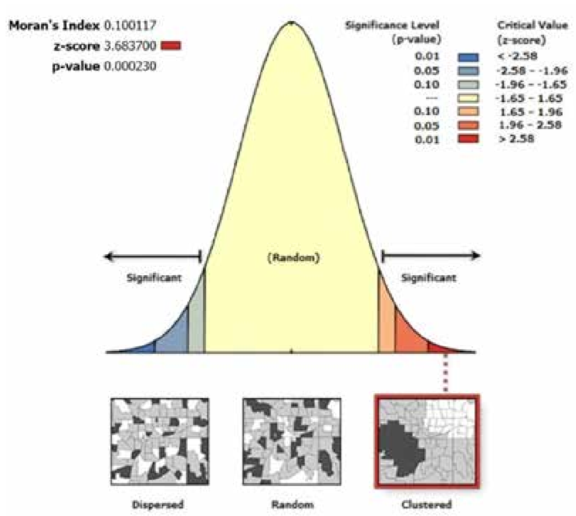


Fig. 4. Rental price data in polygon format; Results of spatial autocorrelation analysis by ARCGIS software.

3.1.2. Local autocorrelation analysis (LISA)
After confirming the spatial autocorrelation of rental unit prices, the neighbourhood scale was used to calculate the facet data of average rental unit prices. The LISA values were calculated to observe the clustering phenomenon of high and low values in each area, and the results are shown in Fig. 5.

High rent unit price clusters (H-H) are in the East, West Central, and North Districts, which are the city centre of Tainan. With complete living functions, and several colleges are clustered here, including National Cheng Kung University, Tainan Nursing College, and the University of Tainan, etc., leading to high rental demand, the rental price thereby increase; The low-rent unit price agglomeration area (L-L) is located near the sea in South District and Yongkang District adjacent to Xinhua District. It is a suburban area far away from the core development areas and schools of Tainan City. The convenience of life is low, resulting in the phenomenon of low-rent unit price clusters. In addition, the areas with a mix of high and low rent prices, H-L points, are located at the junction of the Yongkang District. It is speculated that the proximity to the Tainan Industrial Park and the Yongkang Interchange of National Highway No. 1 has resulted in higher rents in this area than in other areas. However, there is less rental data in this area, so bias due to insufficient sample size cannot be ruled out; L-H points are located in the West Central District and Annan District, which is the transition zone between the city centre and the suburbs. It results in a spatial distribution phenomenon where high-rent and low-rent properties coexist.

Then, we analysed the spatial cluster phenomenon on the main variables that affect housing prices. There are two variables in total, namely house age and floor area. The results are shown in Fig. 6. and Fig. 7. Aged housing areas are in the city centre and overlap with high-rent areas. Although the results of subsequent model construction show that the overall impact of house age on unit rent is negative, the unit rental price in areas with older houses is still higher in the city centre due to the location value and the impact of the specific rental market structure. Smaller floor area clusters also overlap with high-rent areas. Land resources in the city centre are scarce, and the unit area design tends to be smaller to improve space utilization efficiency. Furthermore, in response to the special target groups around the city centre, such as students, young tenants, etc., the rental price of the units is relatively high.

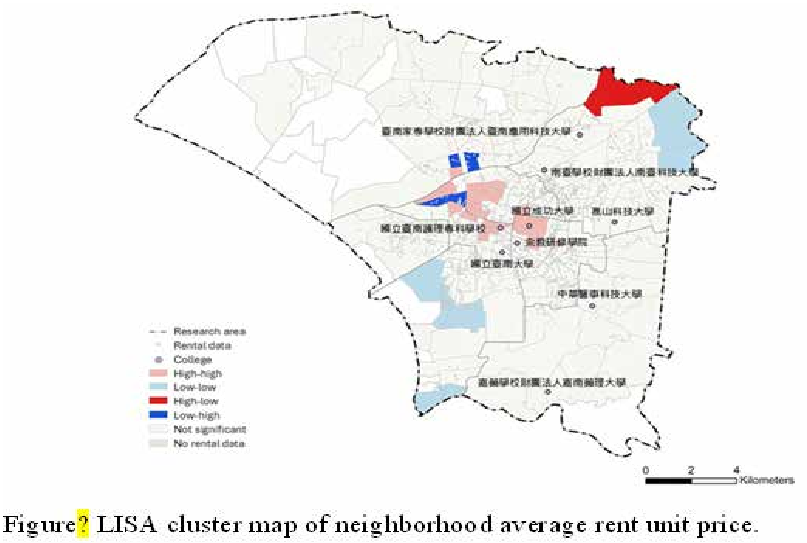
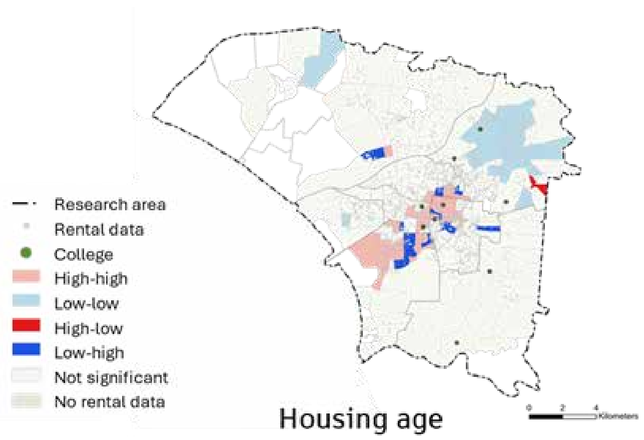


Fig. 5. LISA cluster map of neighbourhood average rent unit price.



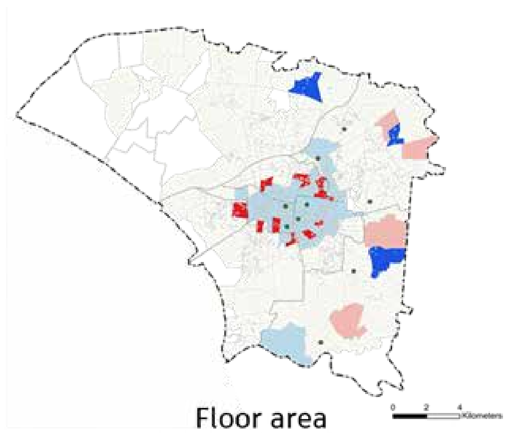


Fig. 6. LISA cluster map of average housing age.

Fig. 7. LISA cluster map of average floor area.

3.2. Ordinary Least Squares (OLS)

Ordinary Least Squares (OLS) is a statistical method used to determine the relationship between two or more variables by minimizing the sum of the squared differences between predicted values and actual observed values. This method identifies the best linear relationship. The OLS estimation results are shown in Tables 2 and 3.

From the regression statistics, the model is based on an analysis of 4,942 samples, which is a sufficient sample size. The multiple R value is 0.289, indicating the correlation between variables ranging from 0 to 1. This suggests a relatively low correlation, meaning the model has limited explanatory power for the dependent variable. The R-squared value explains the proportion of variance in the dependent variable accounted for by the model, which is only 8.35%. This indicates that the explanatory variables in the model have limited ability to explain housing prices, and other important variables may not have been included. The adjusted R-squared is slightly lower than the original R^2 because it accounts for the number of variables, providing a more conservative measure of the model's explanatory power. The standard error, which represents the range of prediction errors, is 104.209. A smaller value is preferable, but the relatively high standard error indicates that the model's accuracy is suboptimal.

In the regression coefficient analysis, the coefficients show the direction of the impact of each variable on the dependent variable. When all independent variable values are zero, the predicted value of the dependent variable is 298.431. Among the independent variables analyzed, most were statistically significant, except for the distance to elevators and supermarkets, which were not significant.

Overall, the findings indicate that as housing age, distance to schools, convenience stores, bus stops, and bicycle lanes increase, housing prices decrease. Although the model is statistically significant, the R^2 value is only 0.084. It suggests that the model may be missing important variables, such as location, school district quality, and transportation. This causes limitations for the research.

Table 2. Result of OLS regression statistics

regression statistics	result
Multiples of R	0.289
R-squared	0.084
Adjustable R Square	0.082
Standard Errors	104.209
Number of observed values	4942

Table 3. Result of OLS regression model

	coefficient	Standard Errors	t Stats	P-value	Lower Limit 95%	Up to 95%
intercept	298.431	5.942	50.227	0.000	286.782	310.079
Age	-0.346	0.040	-8.588	0.000	-0.426	-0.267
Area	-0.067	0.006	-10.719	0.000	-0.079	-0.055
Furniture	-10.589	3.621	-2.925	0.003	-17.687	-3.491
Manage	-46.400	5.299	-8.756	0.000	-56.789	-36.011
Elevator	6.696	5.761	1.162	0.245	-4.598	17.990
Coll_Dist	-0.009	0.001	-6.920	0.000	-0.012	-0.006
PX_Dist	0.002	0.004	0.428	0.668	-0.006	0.010
Conv_Dist	-0.052	0.016	-3.323	0.001	-0.082	-0.021
Bus_Dist	-0.036	0.015	-2.358	0.018	-0.065	-0.006
Bike_Dist	-0.034	0.008	-4.171	0.000	-0.050	-0.018

3.3. Geographically Weighted Regression (GWR)

As shown in Table 4. The overall model's R^2 reaches 0.22, significantly improving the performance compared to the linear model. Based on the t-values, The primary influencing variables are Age, Area, Coll_Dist, Manage, Conv_Dist, and Bike_Dist. The secondary influencing variables are Furniture and Bus_Dist. Elevator and PX_Dist are identified as non-significant variables.

Table 4. Calculation results of the GWR model

Variable	Standard Deviation	Mean (Coefficient)	Maximum	Minimum	t value
Intercept	5.9416	298.4305	377.5660	265.3773	2.00E-16
Age	0.0403	-0.3465	-0.1252	-0.6986	2.00E-16
Area	0.0062	-0.0668	-0.0220	-0.8823	2.00E-16
Coll_Dist	0.0013	-0.0090	0.0010	-0.0278	5.08E-12
Furniture	3.6207	-10.5890	5.5771	-44.9303	0.003465
Manage	5.2994	-46.4001	1.2113	-75.2109	2.00E-16
Elevator	5.7609	6.6956	33.3501	-48.9803	0.245192
PX_Dist	0.0041	0.0018	0.0735	-0.0323	0.668474
Conv_Dist	0.0155	-0.0516	0.0601	-0.2283	0.000897
Bus_Dist	0.0151	-0.0356	0.0730	-0.1372	0.018406
Bike_Dist	0.0082	-0.0342	0.0228	-0.1068	3.08E-05

1.1.1. The impact of educational characteristics on rental housing prices

By comparing the distance map of each point to the nearest university (Fig. 8) with the results of the Geographically Weighted Regression (GWR) analysis (Fig. 9), it is evident that the closer the distance to a university, the higher the housing prices.

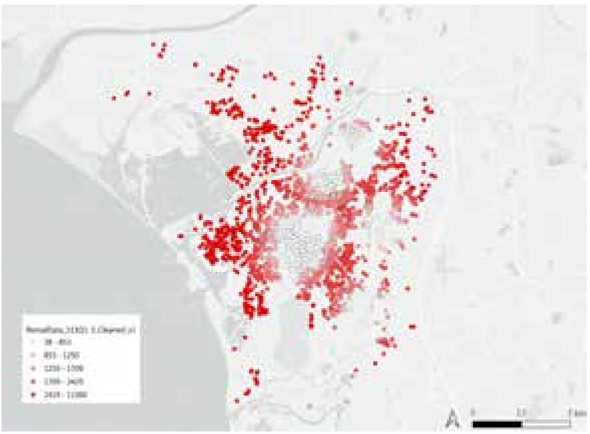


Fig. 8. The distance map of each point to the nearest university.

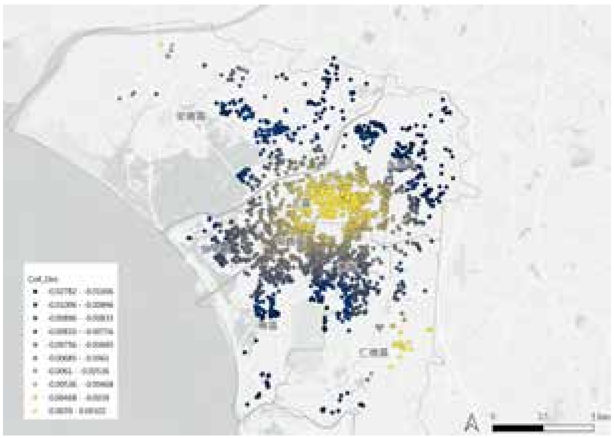
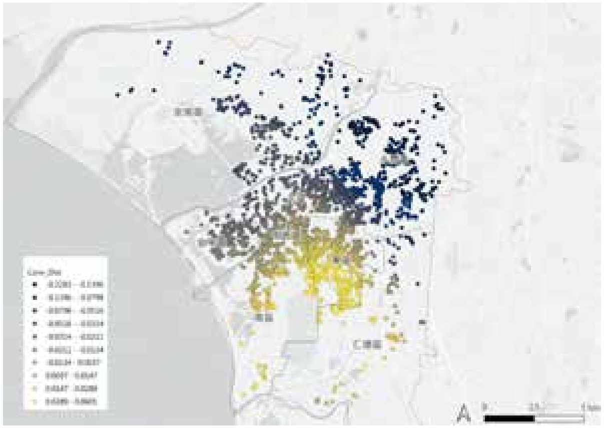


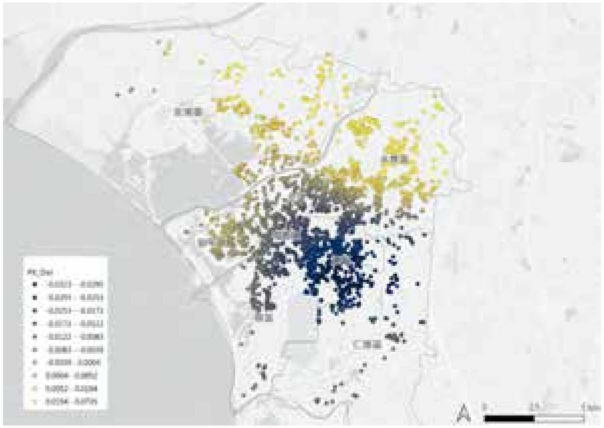
Fig. 9. The impact of educational characteristics on rental housing prices.

1.1.2. The impact of location characteristics on rental housing prices

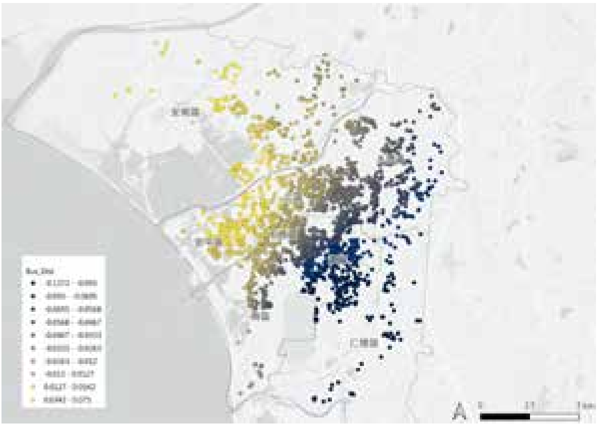
As shown in Fig. 10(a-d), the distances to the five major convenience stores and shared bicycle stations are the primary variables affecting housing prices, while the distance to the nearest bus stop is a secondary variable. The negative impact of the distance to convenience stores on housing prices is more pronounced in Yongkang District and Annan District, which can be attributed to the density of convenience stores in these areas. The negative impact of the distance to shared bicycle stations on housing prices is more significant in the East District, likely due to the demand for bicycles and the number of station installations in this area.



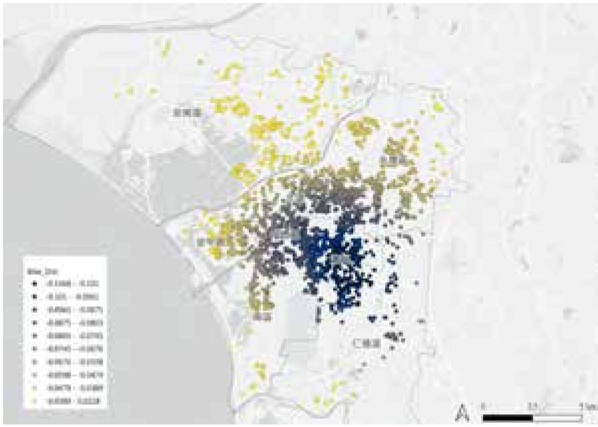
a. Conv_Dist



b. PX_Dist



c. Bus_Dist

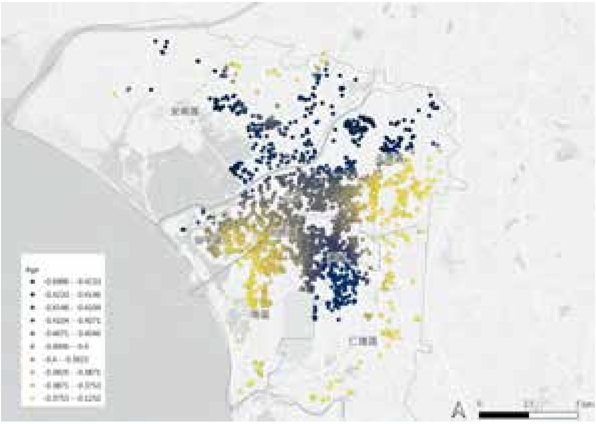


d. Bike_Dist

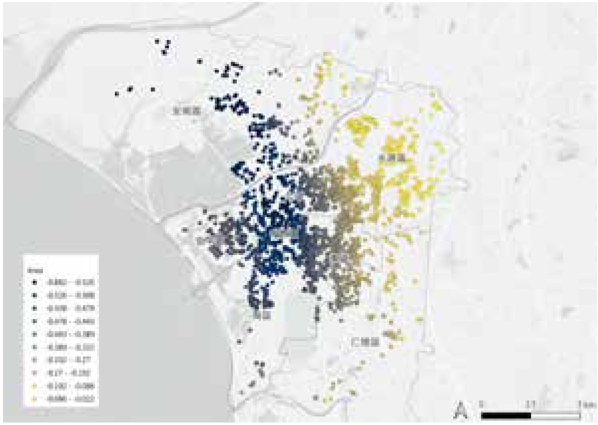
Fig. 10. The impact of location characteristics on rental housing prices.

1.1.3. The impact of building characteristics on rental housing prices.

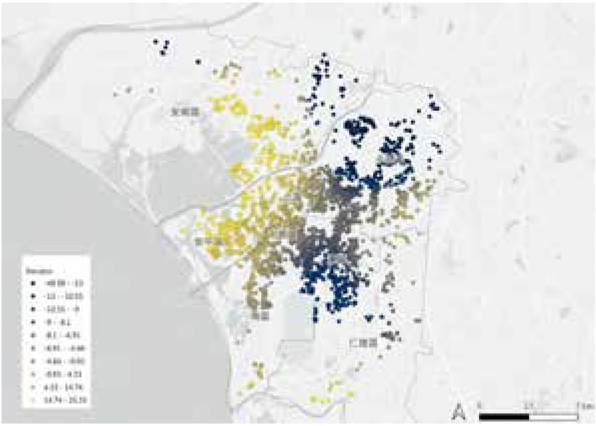
As shown in Fig. 11 (a-d), building age and floor area are the primary variables influencing housing prices, with the presence of furniture being a secondary factor. The negative impact of building age on housing prices is more pronounced in the East District and Yongkang District. Additionally, there is a significant north-south disparity in Yongkang District, with Provincial Highway 20 serving as a dividing line; this is likely because the southern area consists of older neighbourhoods, while the northern area features more newly constructed houses, resulting in a greater influence. The impact of floor area on the negative effects of building age is more significant in Annan District and the West Central District. Cross-referencing the analysis of building age, it can be inferred that Annan District contains newly constructed houses, leading to noticeable differences in floor area.



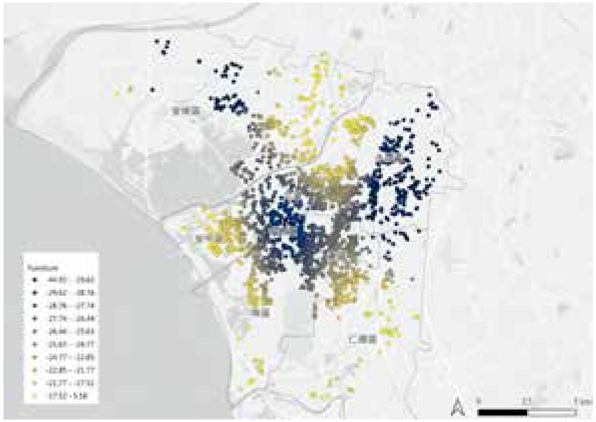
a. Age



b. Area



c. Elevator



d. Furniture

Fig. 11. The impact of building characteristics on rental housing prices.

3.4. Geographically Weighted Machine Learning

The study employed geographically weighted LightGBM (GW-LGBM) to capture spatial heterogeneity and nonlinear relationships more effectively than traditional models. The bandwidth selection and Gaussian kernel weighting scheme were used in the same way as geographically weighted regression (bandwidth = 0.023), so that each observation receives a distance-decay weight relative to the prediction location—the farther an observation lies from the test point, the smaller its weight becomes, as illustrated in Fig. 12.

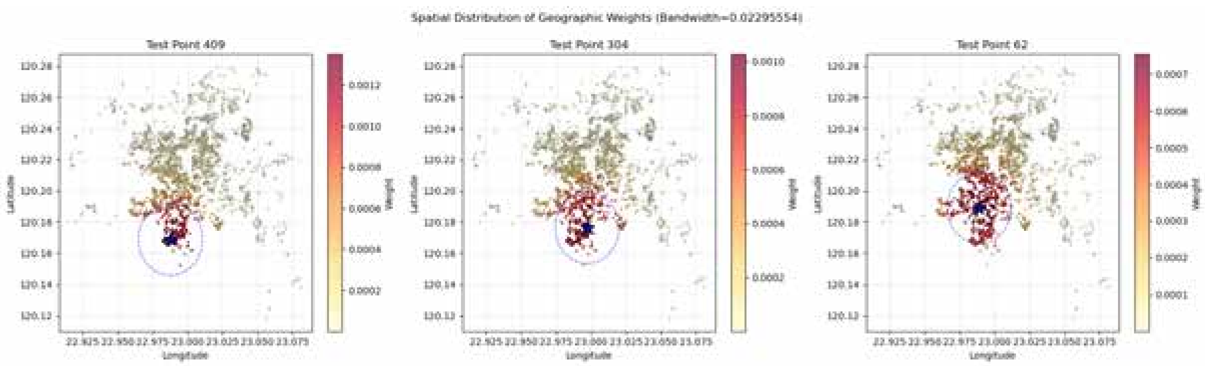


Fig. 12. Spatial Distribution of Geographic Weights.

The geographically weighted LightGBM (GW-LGBM) model attains a root-mean-square error (RMSE) 68.53 in predicting unit rental prices. Feature-importance analysis identifies floor area, building age, and furnishing status as the three most influential predictors, with floor area exerting a markedly greater effect than the others. SHAP analysis shows that its contribution to explaining unit rent diminishes as floor area increases. This attenuation is likely due to a neighbourhood-level rent ceiling. Once that threshold is approached, additional square metres generate progressively smaller rent increases, making unit-rent estimation more challenging for the model. Consequently, lower building age shows a positive association. Newer properties improve the model performance for the unit rental price. Furnishing status displays a highly polarised pattern: because most Taiwanese rental units are furnished, an unfurnished listing becomes a distinctive indicator that enhances predictive accuracy.

Spatial mapping of SHAP values demonstrates that predictor importance varies across the research area. In the city centre, the floor area contributes most strongly to the model's discrimination of unit rents, whereas in suburban areas, building age becomes the dominant explanatory factor. Furnishing status shows far less spatial variation. These results indicate that the GW-LGBM framework yields superior predictive performance and captures the rental market's inherent spatial heterogeneity when combined with SHAP diagnostics.

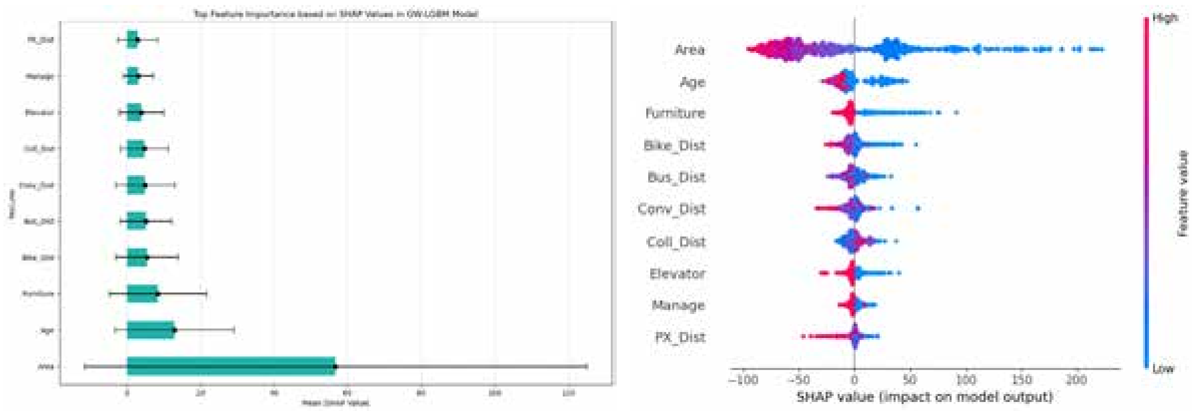


Fig. 13. Feature Importance on SHAP Value

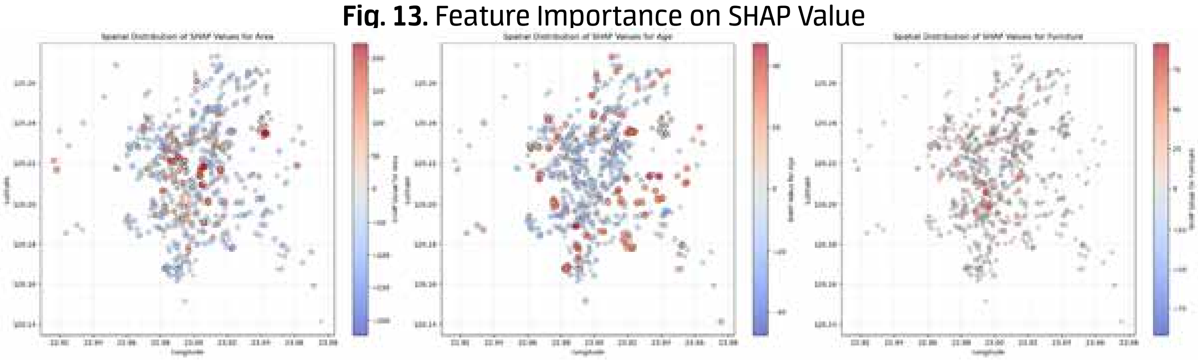


Fig. 14. Spatial distribution of SHAP Value for Area, Age, and Furniture

3.5. Model Comparison

In this study, geographically weighted models show better performance than original models. This study employed OLS, GWR, and GW-ML models to compare their performance. In order to evaluate the effectiveness of these models, a comparison was also made with general machine learning methods that do not incorporate geographical weighting. The comparison results were shown in Fig. 15. Previous studies have primarily compared linear regression models with GWR, demonstrating that GWR is a superior analytical method. This phenomenon is also confirmed in the present experiment, where the RMSE of the linear regression model is 112.36, while that of GWR is 96.23, indicating a significant improvement in model performance. However, upon including machine learning models in the comparison, it was observed that the RMSE values of all machine learning models outperformed those of GWR. This suggests that machine learning provides a more effective approach to analysing unit rental prices.

The superior performance of machine learning models may be attributed to the inherent characteristics of GWR, which is a spatial statistical method specifically designed to capture spatial heterogeneity. Factors that influence unit rental prices exhibit limited spatial characteristics, rendering GWR effective in enhancing linear regression but insufficient for addressing the details of rental data. Consequently, machine learning models demonstrate better performance than GWR in this context.

The results showed that the geographically weighted machine learning model (GW-LGBM) achieved the lowest RMSE of 68.53, outperforming all other models. It significantly surpassed both the linear regression model and state-of-the-art machine learning models. This indicates that when data exhibits spatial heterogeneity, applying geographical weighting to the model can notably improve its performance.

Table 5. RMSE for all models

Model Type	Model Name	RMSE
OLS	Linear Regression	112.36
	Bayesian Ridge	112.38
	Lasso	112.36
Machine Learning	Support Vector Regression (SVR)	108.84
	Random Forest (RF)	75.64
	Gradient Boosted Decision Tree (GBDT)	76.88
	eXtreme Gradient Boosting (XG-Boost)	78.70
	Light Gradient Boosting Machine (LightGBM)	75.08
	K-Nearest Neighbor (KNN)	94.35
GWR	Geographically Weighted Regression	96.23
GW-ML	**Geographically Weighted LightGBM (GW-LGBM)**	**68.53**
	Geographically Weighted Random Forest (GW-RF)	70.81

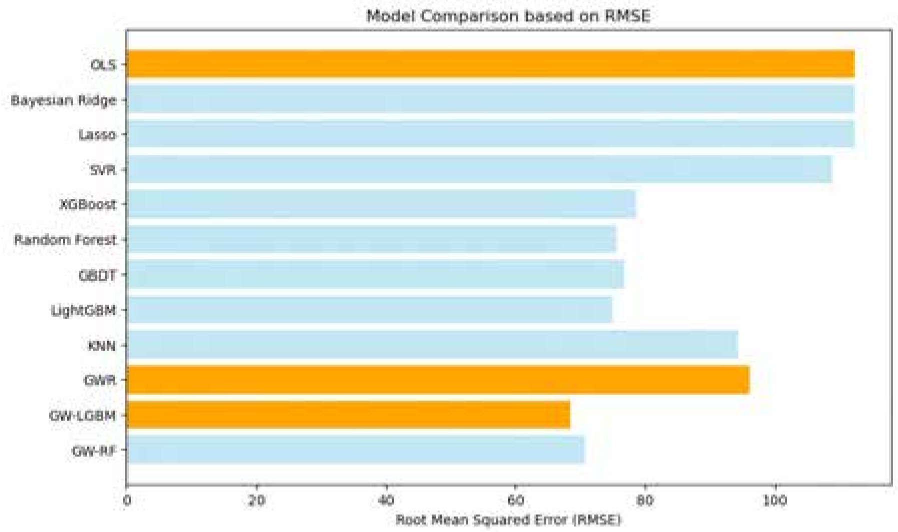


Fig. 15. Model Comparison

The aim of this study is to analyse the spatial characteristics of rental housing using various analytical methods, including Geographically Weighted Regression (GWR) and Geographically Weighted Machine Learning (GW-ML). The GW-ML model outperforms linear regression, machine learning, and GWR models due to its ability to simultaneously accommodate spatial sample weight adjustments and non-linear relationships that describe interactions with rental prices. The results reveal that, within the GWR model, factors such as rental area, building age, distance to universities, and the availability of management staff are the most significant determinants of rental prices. Meanwhile, in the GW-LGBM model, rental area, building age, and furniture availability are the most important factors influencing model performance. Both methods highlight rental area and building age as key factors impacting rental prices. From the GW-LGBM model, it was found that rental properties in city centres place more importance on rental area, while those in suburban areas focus more on building age. These findings can assist urban planners in understanding the preferences and spatial heterogeneity of rental populations, enabling the development of policies that provide appropriate subsidies or other measures to enhance overall social welfare.

However, the R-squared value for the GWR model in this study is only 0.22, indicating that it cannot fully explain the variations in unit rental prices. Although the GW-LGBM model performs better, some errors remain. Future research could leverage Graph Neural Networks (GNNs) or other models to further investigate the underlying causes of rental behaviour and price fluctuations.

REFERENCES

Anselin, L. (1995). Local Indicators of Spatial Association—LISA. Geographical Analysis, 27(2), 93–115. https://doi.org/10.1111/j.1538-4632.1995.tb00338.x

Brunsdon, C., Fotheringham, A. S., & Charlton, M. E. (1996). Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity. Geographical Analysis, 28(4), 281–298. https://doi.org/10.1111/j.1538-4632.1996.tb00936.x

Can, A. (1992). Specification and estimation of hedonic housing price models. Regional Science and Urban Economics, 22(3), 453–474. https://doi.org/10.1016/0166-0462(92)90039-4

Cao, K., Diao, M., & Wu, B. (2019). A Big Data–Based Geographically Weighted Regression Model for Public Housing Prices: A Case Study in Singapore. Annals of the American Association of Geographers, 109(1), 173–186. https://doi.

org/10.1080/24694452.2018.1470925

Cellmer, R., Cichulska, A., & Bełej, M. (2020). Spatial Analysis of Housing Prices and Market Activity with the Geographically Weighted Regression. ISPRS International Journal of Geo-Information, 9(6), Article 6. https://doi.org/10.3390/ijgi9060380

Davis, J., Hagen, B., Mahid, Y., & Pijawka, D. (2023). Evaluating Property Value Impact From Water-Related 'Green Infrastructure': A Hedonic Modeling Approach. Journal of Green Building, 18(1), 3–16. https://doi.org/10.3992/jgb.18.1.3
Fotheringham, A., Brunsdon, C., & Charlton, M. (2002). Geographically Weighted Regression: The Analysis of Spatially Varying Relationships. John Wiley & Sons, 13.

Gray, D. (2012). District House Price Movements in England and Wales 1997–2007: An Exploratory Spatial Data Analysis Approach. Urban Studies, 49(7), 1411-1434. https://doi.org/10.1177/0042098011417020

Gutiérrez-Posada, D., Rubiera-Morollon, F., & Viñuela, A. (2017). Heterogeneity in the Determinants of Population Growth at the Local Level: Analysis of the Spanish Case with a GWR Approach. International Regional Science Review, 40(3), 211–240. https://doi.org/10.1177/0160017615589009

Iliopoulou, P., & Feloni, E. (2022). Spatial Modelling and Geovisualization of House Prices in the Greater Athens Region, Greece. Geographies, 2(1), Article 1. https://doi.org/10.3390/geographies2010008

Koohpayma, J., & Argany, M. (2021). Estimating the price of apartments in Tehran using extracted compound variables. INTERNATIONAL JOURNAL OF HOUSING MARKETS AND ANALYSIS, 14(3), 569–595. https://doi.org/10.1108/IJHMA-05-2020-0050

Li-Chueh Lin. (2018). A Spatial Autocorrelation Analysis of Housing Affordable Limit - A Case Study of Six Taiwan Cities. Master's thesis, National Cheng Kung University. https://hdl.handle.net/11296/n6h5yr

Liu, N., & Strobl, J. (2023). Impact of neighborhood features on housing resale prices in Zhuhai (China) based on an (M)GWR model. Big Earth Data, 7(1), 146–169. https://doi.org/10.1080/20964471.2022.2031543

Mocanu, T., & Tremacoldi-Rossi, P. (2023). The impact of international students on housing markets. Canadian Journal of Economics/Revue Canadienne d'économique, 56(2), 647–675. https://doi.org/10.1111/caje.12651

Sirmans, S., Sirmans, C., & Benjamin, J. (1989). Determining Apartment Rent: The Value of Amenities, Services and External Factors. Journal of Real Estate Research, 4(2), 33–43. https://doi.org/10.1080/10835547.1989.12090581

Ti-Ching Peng. (2021). The Effect of The Spatial Dispersion of Convenience Stores on House Prices: A Spatial Analysis in Taipei Metropolis. Taiwan Journal of Applied Economics, 110: 51-104. https://doi.org/10.3966/054696002021120110002

Tzu-Ting Kao. (2015). Research on High Housing Vacancy Rate in Taiwan by Spatial Analysis. Master's thesis, National Taiwan University. https://hdl.handle.net/11296/29kewa.

Ucal, M., & Kaplan, U. (2020). Housing Prices in a Market Under Years of Constant Transformation: A County-Based Analysis of Istanbul. Entrepreneurial Business and Economics Review, 8(2), 71–91. https://doi.org/10.15678/

EBER.2020.080204

Villada-Medina, H. D. (2022). The real estate market and public transportation systems in developing countries: The case of Medellín, Colombia. Urbe. Revista Brasileira de Gestão Urbana, 14, e20210447. https://doi.org/10.1590/2175-3369.014.e20210447

Wang, S., Gao, K., Zhang, L. et al (2024). Geographically weighted machine learning for modeling spatial heterogeneity in traffic crash frequency and determinants in US. Accident Analysis and Prevention, 199. http://dx.doi.org/10.1016/j.aap.2024.107528

Wang, L., Wang, G., Yu, H., & Wang, F. (2022). Prediction and analysis of residential house price using a flexible spatiotemporal model. Journal of Applied Economics, 25(1), 503–522. https://doi.org/10.1080/15140326.2022.2045466
Wang, W. -C., Chang, Y. -J., & Wang, H. -C. (2019). An Application of the Spatial Autocorrelation Method on the Change of Real Estate Prices in Taitung City. ISPRS International Journal of Geo-Information, 8(6), 249. https://doi.org/10.3390/ijgi8060249

Wen, H., Xiao, Y., & Zhang, L. (2017). School district, education quality, and housing price: Evidence from a natural experiment in Hangzhou, China. Cities, 66, 72–80. https://doi.org/10.1016/j.cities.2017.03.008

Wen, H., Zhang, Y., & Zhang, L. (2015). Assessing amenity effects of urban landscapes on housing price in Hangzhou, China. Urban Forestry & Urban Greening, 14(4), 1017–1026. https://doi.org/10.1016/j.ufug.2015.09.013

Yang, W., Deng, M., Tang, J. et al.(2023). Geographically weighted regression with the integration of machine learning for spatial prediction. J Geogr Syst 25, 213–236 . https://doi.org/10.1007/s10109-022-00387-5

Yingjie Zhang, Hongqiang Wang, Yu Xu(2020) A quantitative study on the influencers of rental prices in first-tier cities--taking downtown Shanghai as an example. Acta Geograph. Sin., 76 (8) (2020), pp. 72-75

Ying-Hui Chiang, Yu-Ting Chuang, & Chin-Oh Chang. (2017). THE IMPACT OF PUBLIC BIKE STATION ON RESIDENTIAL HOUSING PRICE IN TAIPEI CITY. Transportation Planning Journal, 46, 4, pp. 399-428.

Zhao, W., Zhong, J., & Lv, J. (2024). Influencing factors and spatial differentiation of rental housing in a smart city: A GWR model analysis. Measurement: Sensors, 33, 101126. https://doi.org/10.1016/j.measen.2024.101126