ISSN: 1648-715X / eISSN: 1648-9179

VILNIUS TECH
Vilnius Gediminas Technical University

INTERNATIONAL JOURNAL OF STRATEGIC PROPERTY MANAGEMENT

2025
Volume 29
Issue 4
Pages 232–244
https://doi.org/10.3846/ijspm.2025.23638

A STUDY ON HOUSE PRICE INDEX PERFORMANCE: MIX ADJUSTMENT AND HIERARCHICAL LINEAR GROWTH REPEAT-SALES MODELS

Chun-Chang LEE[1*], Cheng-Yen CHUANG[1], Wen-Chih YEH[2], Pei-Syuan LIN[3]

[1] Department of Real Estate Management, National Pingtung University, No. 51, Mingsheng East Road, Pingtung, Taiwan
[2] Department of Real Estate Management, HungKuo Delin University of Technology, No. 1, Lane 380, Qingyun Road, New Taipei, Taiwan
[3] Department of Land Resources, Chinese Culture University, No. 55, Hwakang Road, Taipei, Taiwan

Article History:
- received 10 September 2024
- accepted 24 February 2025

Abstract. In this study, we examined the differences between three house price indexes constructed using hedonic price, mix adjustment, and hierarchical linear growth repeat-sales modeling. The data consisted of housing sales across 13 administrative districts in Kaohsiung City from the third quarter of 2013 to 2022. The predictions were compared using the mean standard error, mean absolute percentage error, mean absolute error, and root-mean-square error. The results revealed that the hedonic price index performed the best; its prediction scores, as reflected by the four aforementioned metrics were 0.072, 1.176, 0.181, and 0.181, respectively. The index with the second best performance was the mix adjustment model, with scores of 0.154, 1.905, 0.293, and 0.293. The worst-performing index was the repeat-sales model, with scores of 0.309, 2.804, 0.439, and 0.439. After comparing the annual prediction errors of the three models, it became apparent that the hedonic price index had the best performance, followed by the mix adjustment index, and then the hierarchical linear growth repeat-sales index.

Keywords: house price index, hedonic price model, mix adjustment model, hierarchical linear growth repeat-sales model, *k*-means.

[*] Corresponding author. E-mail: *lcc@mail.nptu.edu.tw*

1. Introduction

A house price index is a basic economic indicator for investors, lending institutions, policymakers, and economic analysts. The main purpose of this indicator is to track house price variation trends, show the directions of market fluctuations, and aid market actors in determining whether a market is flourishing or declining. It provides price information and helps these market actors understand market trends and engage in long-term planning. Tracking a house price index over a long period of time assists financial institutions in managing credit risk, enables policymakers to grasp market demands, and allows for the monitoring of economic well-being and the stability of the housing market. An accurate house price index is an important risk and benefit assessment tool for investors.

An effective house price index should be representative of the market, taking into account the heterogeneity of different house types, locations, and attributes. It should also be stable and consistent over various time periods to avoid disruptions arising from abnormal data. Furthermore, the index should be transparent so that market actors can accurately interpret and apply its data. The hedonic price model (HPM) and repeat-sales model (RSM) are two common approaches in house price index development. The former is suitable for stable markets while the latter is suitable for rising markets with active sales activities. However, the applicability and accuracy of these models are challenged by highly heterogeneous and rapidly changing markets. A suitable house price index is particularly important in cities undergoing economic transformation.

In response to the heterogeneity and data structure complexity of the real estate market, many house price prediction techniques such as hierarchical clustering, *k*-means clustering, and hierarchical linear growth modeling have been proposed and extensively utilized. These techniques effectively capture the characteristics of different types of houses as well as the price variation trends at the market level. For example, in hierarchical clustering, houses are categorized according to their geographic location, area, number of floors, and facilities. This effectively segments the market into highly homogenous clusters, thereby reducing the impact of market heterogeneity on the prediction results, as well as accurately predicting the price variation in each cluster. Hierarchical linear growth modeling analyzes the long-term impacts of regional and housing attributes on price through hierarchical structure analysis. By accounting for multilevel factors such as time and location, it effectively reflects the dynamic changes in the market. Although these techniques have been applied in house price prediction, they remain underutilized in house price index development.

This study aimed to apply house price prediction techniques and develop a house price index that compensates for the shortcomings of the HPM and RSM, and better represents market dynamics. The HPM is particularly useful in markets with a stable sales volume, as it is able to handle diversified housing characteristics in the market. However, it is susceptible to multicollinearity and heteroscedasticity when the market characteristics are strongly associated. To resolve this issue, we proposed a clustering median adjustment model (CMAM) that categorizes the housing market through stratification and clustering so that houses with highly similar attributes are placed into the same cluster. This simplifies the complexity of the market attributes, reduces the impacts of multicollinearity and heteroscedasticity on the estimation consistency and prediction accuracy, and further optimizes the process of constructing a house price index.

The CMAM provides buyers with an index that more clearly reflects the relative value of different house types, allowing buyers to better understand the market dynamics and make more informed house-buying decisions. The model explicitly reveals the relative values and price trends by adjusting the price median and ratio after clustering. Unlike the HPM, which may conceal the distinctive values of some house types, the clustering technique of the CMAM incorporates market heterogeneity so that buyers can gain a more precise understanding of the market trends of specific house types (e.g., small apartments or high-end villas) and assess whether their house-buying decisions align with the existing market values. Moreover, the CMAM helps buyers identify house types with the fastest sales price growth and highest stability, thus making it a strong reference tool when searching for houses with high investment or purchase potential.

The RSM is mainly used to analyze the price variation of the same property at different time periods based on the assumption that its quality remains unchanged across the sales times. However, in reality, house prices often vary due to depreciation and land value appreciation. These factors are not fully considered in the traditional RSM, thereby limiting its ability to reflect the actual market dynamics. To overcome this problem, this study introduced a hierarchical linear growth model (HLGM) and combined it with the RSM to form a hierarchical linear growth repeat-sales model (HLGRM). With its hierarchical structure, the HLGRM effectively analyzes market heterogeneity and combines spatial and temporal factors to provide a more accurate interpretation of the market. The first level included the variables of sales time, house age, and the square of house age to reflect the house price growth rate across different sales times, as well as to distinctly quantify the sales price trend over time. The sales time variable also overcomes the RSM's inability to treat time series variations, while the house age and square of house age variables illustrate the decline in house price over time and address the RSM's failure to account for the variation in house quality. The second level included location-based random effects to analyze the effects of land value appreciation in different locations on house price. For example, city centers may have a higher land value appreciation rate because they experience rapid developments, while rural areas have a lower land value appreciation rate.

The aforementioned characteristics of the HLGRM make it an innovative tool that combines dynamic market properties with heterogeneity analysis. It is practical for homeowners, property developers, and investors. The hierarchical structure of the HLGRM concurrently takes into account the variations in market heterogeneity and house quality and offers a more accurate analysis of house price variations over time. It helps market actors quantify the trends in house price variations over time and gain a better understanding of the long-term market trends. The HLGRM also showcases the composite effects of the depreciation and land value appreciation of older urban areas and houses, providing market actors with a baseline index that aligns with real-world conditions and facilitates reliable decision-making. To summarize, the HLGRM consolidates the strengths of market heterogeneity analysis and dynamic modeling, thus enhancing the interpretability and prediction accuracy of the house price index and serving as an important reference for investment and development in dynamic markets.

The innovative contribution of this study is the introduction of two house price prediction techniques (the CMAM and HLGRM) that integrate the HPM and RSM. Both the CMAM and HLGRM aim to resolve the limitations of existing house price index calculation approaches with respect to sample selection bias, non-uniform sales time distribution, and market heterogeneity. Using real estate sales data in Kaohsiung City, this study evaluated the accuracies of the HPM, CMAM, and HLGRM. The CMAM accounts for market heterogeneity through clustering and effectively reflects the price trends of different types of houses, thus providing buyers and developers with a more practical house price baseline. The HLGRM accurately captures long-term market dynamics by accounting for house depreciation and land value appreciation and is particularly suitable for rapidly changing market environments. These results provide a reliable house price prediction baseline for policymakers and investors in Kaohsiung City while also serving as a practical reference for developing a house price index for cities with similar market conditions.

2. Literature review

2.1. Hedonic price model (HPM) and clustering median adjustment model (CMAM)

The HPM is a common approach in constructing a house price index. It estimates the house price by analyzing the attributes of the house (such as geographic location, area, floor number, and construction material). However, the

HPM faces several statistical challenges in practice, mainly multicollinearity, heteroscedasticity (Kennedy, 2008; Rahman et al., 2019), and sample selection bias. Multicollinearity occurs when the attribute variables are excessively correlated, thus affecting the stability and accuracy of the regression model. Heteroscedasticity can cause the error term to fluctuate according to the variation of the attributes, thus reducing the stability of the model. This problem is particularly prominent when there is significant market heterogeneity (Studenmund, 2014). Heterogeneity may impact the simple means of constructing a house price index according to the median and may generate bias when the market components vary. Moreover, sample selection bias occurs when the model is unable to effectively include each house type in the market. Consequently, certain housing attributes are overestimated or underestimated, thus diminishing the representativeness and accuracy of the house price index.

To resolve these problems, clustering techniques are increasingly utilized in house price index development. The core objective of clustering is to segment the market into subsets that share similar characteristics, thereby enhancing the stability and accuracy of the model. Clustering reduces the impacts of multicollinearity and heteroscedasticity by segmenting the housing market into groups that are less heterogeneous. This increases the model's explanatory power and prediction accuracy. Clustering also effectively decreases sample selection bias and prevents house price variations arising from changes in the sample compositions. Therefore, clustering reduces the correlation between the attribute variables, increases the representativeness of the house price index, and reflects the actual variations in each housing attribute on the market.

The mix adjustment model is the most common median-based house price index. Prasad and Richards (2008) proposed an adjustment approach that stabilizes the house price index by adjusting the ratio of houses with different price levels. They found that the median house price may deviate from actual market trends due to the variation in the sales ratio of high-priced and low-priced locations. Therefore, the approach must be adjusted by incorporating market compositions. Subsequent studies such as Miller and Maguire (2020) included housing attribute levels (e.g., the number of rooms) to further improve the mix adjustment model, thereby demonstrating its superior stability and representativeness compared to the traditional HPM.

Clustering or stratification techniques have been successfully applied in HPMs for predicting house prices, thus enhancing their ability to handle market heterogeneity. For example, Kim and Irakoze (2022) applied *k*-means clustering and the partitioning around medoids (PAM) algorithm on housing sales data in Seoul in 2018 to analyze the green premium of Green Standard Energy and Environmental Design-certified apartments. The authors revealed that the two clustering methods yielded different green premium estimations: 12.2% through *k*-means clustering and 17.8% through PAM. This shows that the choice of clustering method significantly affects the market segmentation results and prediction accuracy. Similarly, Kwon et al. (2017) used *k*-means clustering to resolve house price heterogeneity in Seoul. The authors pointed out that traditional administrative district segmentation is not suitable for property market analysis because it does not effectively account for house price heterogeneity. By segmenting Seoul into 16 clusters and consolidating geographic attributes and sales data, the authors significantly improved the accuracy of house price prediction.

To address the shortcomings of the HPM, this study proposed a clustering algorithm to construct an HPM-based house price index called the CMAM. Although clustering algorithms are widely used to construct median-based indexes and HPM-based house price prediction models, their inclusion in HPM-based house price indexes is rarely mentioned in the existing literature. The objective of the CMAM is to reduce the impacts of heterogeneity and outliers on the model and increase the stability and accuracy of the index by incorporating market segmentation and median adjustment techniques. Through agglomerative and *k*-means clustering, the model first segments the sample into several subsets according to the housing attributes and market heterogeneity. Next, the median price in each cluster is calculated, and the median price combinations are used to develop the final price index. This process thoroughly captures the price variations across different levels in the market and reduces the potential bias present in traditional weighted models, thus providing a more stable and representative approach to constructing a house price index.

2.2. Traditional repeat-sales model (RSM) and hierarchical linear growth repeat-sales model (HLGRM)

The RSM is a classic and widely applied approach in house price index development. Its core concept is to calculate the price variations of the same house across different sales times, and thus reduce the impacts of housing heterogeneity on index estimation. Because it simply relies on repeat-sales samples, the RSM is able to reflect the variations in house price to a certain extent. There are several other prominent weaknesses in this method as well. First, because it relies on repeat-sales samples, the sample size is limited, reducing the representativeness of the index. Second, sample selection bias may be present in the repeat-sales sample. Houses with a higher sales frequency may have different attributes, which causes the index to be oversensitive to changes in the market for these types of houses and thereby neglects the price trends of other houses. Moreover, the uneven intervals between the sales times may result in biased house price variation estimations.

Researchers have proposed various solutions to resolve the aforementioned problems. For example, the Case–Shiller index is a classic type of RSM. It simplifies house price variations into a function of price variation within sales times. However, in the traditional RSM framework, time effects and house depreciation (age) effects are often entangled and difficult to separate. Consequently, the house price variation estimation may be biased. This model assumes that the quality of the house remains unchanged, but in reality, the impacts of house depreciation may be mixed into the time effects. To resolve this problem, Cannaday et al. (2005) proposed the multivariate repeat-sales model (MRSM), which separately controls the time effects and house depreciation effects by including the variable of house age. This creates a pure time price index (time-constant, age-varying) and a depreciation price index (age-constant, time-varying). Although the MRSM model has its strengths in controlling depreciation, it is still limited in handling the heterogeneity of house price growth.

To overcome the shortcomings of the traditional RSM in markets with high heterogeneity and data sparsity, Francke and Van de Minne (2017) introduced the hierarchical repeat-sales model (HRSM), which is based on hierarchical structures. Price trends are divided into common trends and cluster-specific trends. This approach can handle diversified attributes on the market and provides stable index estimations in the presence of data sparsity. The t-distribution in the HRSM resolves the impacts of outliers and further enhances the stability and accuracy of the index.

Inspired by the aforementioned studies, we propose an HLGRM that improves the RSM and addresses the shortcomings of house price index development. Although there is evidence supporting the strengths of HLGMs in house price prediction (Lee et al., 2013, 2023; Tan et al., 2019), it has yet to be applied in house price index development. In this study, we constructed a hierarchical model that generates a house price index by stratifying housing sales data according to geographic location and sales time and estimating the growth trajectories of house prices. The HLGRM remedies the RSM's lack of sample representativeness, non-uniform time intervals, and inconsistent quality, thus enhancing the stability and accuracy of the index. This approach not only generates a more accurate house price index but also provides a new theoretical and practical reference for property market analysis. The HLGRM differs from Cannaday et al.'s (2005) method of separating time and depreciation effects, as it is able to capture the heterogeneity of house price growth and handle multilevel variance in the data. Compared to Francke and Van de Minne's (2017) HRSM, our model focuses on controlling the impacts of depreciation on the index by constructing multilevel price trend structures. In addition to controlling the impact of depreciation, the HLGRM also takes into account the impacts of housing attributes such as the square of house age on price growth, thus enhancing its flexibility.

3. Methods

In this study, the house price index was developed using a HPM, repeat-sales model, and mix adjustment. The metrics used to form comparisons included the MSE, RMSE, MAE, and MAPE. The coefficient of dispersion (COD) is a measurement standard that indicates the dispersion of the probability distribution. It quantifies whether a cluster of observed events is stretched or squeezed and reflects the dispersion of the predicted values around the median. It can be described as the mean percentage deviation of the quotient of the actual and predicted values relative to the median ratio. Mix adjustment is seldom used in academic research because of its high data demand. However, a wide range of sales data has been made available since the launch of the Ministry of the Interior's actual price registration system for real estate properties on August 1, 2012. Furthermore, from July 1, 2021 onwards, the system requires the disclosure of house numbers and other information, thus improving the accuracy of the house price index. Additionally, data homogeneity can be increased by rigorously stratifying housing attributes such as house area, house type, and house age.

3.1. Index settings and methodology for the hedonic price model

A clear advantage of hedonic price modeling is its ability to effectively account for the impact of asset heterogeneity by considering asset attributes. This results in a derived index that can reliably track and monitor price variations over time (Owusu-Ansah, 2018). In this study, we developed the house price index using a log-linear HPM. The annual regression coefficients were multiplied by the hedonic mean of the baseline year, ensuring that the prices of all years were standardized to the same benchmark. The model is presented in Equation (1):

$$\begin{aligned}\ln P &= \beta_0 + \beta_1 Age + \beta_2 Age^2 + \beta_3 Floor_1 + \\ &\beta_4 Floor_4 + \beta_5 BArea + \beta_5 LArea + \beta_7 Parking + \\ &\beta_8 Room + \beta_9 LivRoom + \beta_{10} BathRoom + \\ &\beta_{11} Type1 + \beta_{12} Type2 + \beta_{13} ESchool + \beta_{14} JSchool + \\ &\beta_{15} HSchool + \beta_{16} Metro + \beta_{17} Tra + \sum_{i=1}^{12} \theta_i Location_i + \varepsilon,\end{aligned} \tag{1}$$

where: ln*P* is the logarithm of the grand housing sales price; *Age* is the house age; *Age*^{2} is the square of house age; *Floor*$_1$ and *Floor*$_4$ are the transferred floor numbers of the house; *BArea* is the total transferred house area; *LArea* is the total transferred land area; *Parking* is the availability of parking lots; *Room* is the number of rooms; *LivRoom* is the number of living rooms; *BathRoom* is the number of bathrooms; *Type*1 and *Type*2 are the house types; *ESchool* is the distance of the house to the nearest elementary school; *JSchool* is the distance of the house to the nearest junior high school; *HSchool* is the distance of the house to the nearest senior high school; *Metro* is the distance of the house to the nearest MRT station; *Tra* is the distance of the house to the nearest railway station; *Location* is the township or city in which the house is located; β is the coefficient of an independent variable; θ is the coefficient of the dummy variable of *Location*; ε is the error term. The variables are detailed in Table 1.

Table 1. Definitions of the hedonic price variables

Variable	Definition
Logarithm of grand sales price	The logarithm of the grand sales price of a housing sale, measured in units of NT$10,000
House age	The house age measured in units of years
Square of house age	The square of the house age
Transferred floor number Floor1, 4	The floor number of the house. We used three types of floor numbers–first floor, fourth floor, and floors other than the first or fourth–with floors other than the first or fourth serving as the baseline. Two dummy variables were defined. For Floor1, houses on the first floor were assigned a value of 1, while others were assigned a value of 0. For Floor4, houses on the fourth floor were assigned a value of 1, while others were assigned a value of 0
Transferred house area	The total transferred house area measured in units of ping
Transferred land area	The total transferred land area measured in units of ping
Number of rooms	The number of rooms in a house
Number of living rooms	The number of living rooms in a house
Number of bathrooms	The number of bathrooms in a house
House type (Type)	House types include apartment buildings, luxury condos, and condominiums, with condominiums serving as the baseline. Two dummy variables were defined. For Type1, houses in apartment buildings were assigned a value of 1, while others were assigned a value of 0. For Type2, houses in luxury condos were assigned a value of 1, while others were assigned a value of 0
Parking lots	A dummy variable. Houses with parking lots were assigned a value of 1, while those without were assigned a value of 0
Distance to the nearest elementary school	The distance to the nearest elementary school measured in units of meters
Distance to the nearest junior high school	The distance to the nearest junior high school measured in units of meters
Distance to the nearest senior high school	The distance to the nearest senior high school measured in units of meters
Distance to the nearest MRT station	The distance to the nearest MRT station measured in units of meters
Distance to the nearest railway station	The distance to the nearest railway station measured in units of meters
House location Location	Location of the house in the 13 administrative districts of Kaohsiung City. There are 12 dummy variables (Location1~Location12), with Yancheng District serving as the baseline

Afterwards, we performed regression to estimate the annual regression coefficients, multiplied them by the mean of each variable, and then substituted them into Equation (1) to derive the estimated annual price. The estimated prices were then substituted into Equation (2) to derive the house price index for a particular year.

$$PI_t = \frac{P_t}{P_0} \times 100, \tag{2}$$

where: PI_t is the house price index for a particular year; P_t is the estimated price for that year; P_0 is the estimated price for the baseline year.

3.2. Index settings and methodology for the mix adjustment model

In the mix adjustment model, the housing sales data were stratified and then the median of each level was used to calculate the house price index. In contrast to manual stratification methods, which have been used in previous studies (Prasad & Richards, 2008), we opted for a more efficient approach. First, we performed agglomerative hierarchical clustering with Ward's linkage, with intervals measured using the squared Euclidean distance. The plotted dendrogram was then used to determine the required number of *k*-means clusters, so as to expedite the stratification process and minimize the data variance in the same level. We used *k*-means clustering due to its unsupervised learning nature, computational efficiency, and ease of interpretation. Moreover, *k*-means clustering exhibits its strengths more prominently as the dimensionality of the clusters increases, rendering it suitable for a wide array of applications. In comparison, manual clustering tends to be more susceptible to human error and can involve an extremely taxing and lengthy process when the data size is large.

In the mix adjustment model, the ratio of each house type in the house price levels was adjusted in order to reflect the relative importance of different house types in the market. The model is expressed in Equation (3) as follows:

$$I = P_1Q_1 + P_2Q_2 + P_3Q_3 + P_4Q_4 + P_5Q_5 + P_6Q_6, \tag{3}$$

where: I is the estimated house price; P_1 is the median house price of the first group; Q_1 is the current ratio of the house type in the first group, and so forth. The estimated annual price was then substituted into Equation (2) to deduce the annual mix adjustment price index.

3.3. Index settings and methodology for the repeat-sales model

3.3.1. Single-level repeat-sales model

There are two basic types of repeat-sales models: the original repeat-sales (ORS) model proposed by Bailey et al. (1963) and the weighted repeat-sales (WRS) model proposed by Case and Shiller (1987, 1989). In the ORS, on the basis of a set of time dummy variables, the price index is obtained by using the ratio of the second sales price to the first sales price obtained through ordinary least squares regression. The model is expressed in Equation (4) as follows:

$$\ln\left(\frac{P_{nt}}{P_{n\tau}}\right) = \sum c_t D_{nt} + e_{n\tau t}, \qquad (4)$$

where: $\frac{P_{nt}}{P_{n\tau}}$ is the sales price of a property n at period t and at period τ, which precedes t; D_{nt} is a dummy variable that is equal to –1 at period τ (the time of the initial (first) sale in period), 1 at period t, or 0 otherwise; c_t the logarithm of the cumulative price index in period t; $e_{n\tau t} = e_{nt} - e_{n\tau}$ is an error term.

Case and Shiller (1987) posited that heteroscedasticity occurs when the sales times are different. This is because the holding period is often distributed non-uniformly in the repeat-sales data sample. Building on this, Costello and Watkins (2002) revealed that the significance of the repeat-sales index is low at the beginning and end of short holding periods. This results in heteroscedasticity that arises from regression disturbances. In light of such heteroscedasticity, Case and Shiller (1989) proposed using the WRS model to treat the heteroscedasticity associated with the ORS model. The WRS model takes into account the fact that house prices often increase over time. After examining Case and Shiller's study, Owusu-Ansah (2018) suggested that the log price ($\ln P_{nt}$) of house n at period t can be expressed in Equation (5) as follows:

$$\ln P_{nt} = I_t + H_{nt} + U_{nt}, \qquad (5)$$

where: I_t is the logarithm of the price level at period t; H_{nt} is the Gaussian random walk in which $E\left(H_{nt} - H_{n\tau}\right) = 0$, $E\left(H_{nt} - H_{n\tau}\right)^2 = \left(t - \tau\right)\sigma^2 H$; U_{nt} is the white noise in which $E\left(U_{nt}\right) = 0,\ E\left(U_{nt}\right)^2 = \sigma_U^2$.

Calculating the house price index in this model comprises three steps. First, the estimated residual ($\hat{e}_{nt} - \hat{e}_{n\tau}$) is derived through Equation (4) using the ORS procedure. Second, the squared residual $(\hat{e}_{nt} - \hat{e}_{n\tau})^2$ undergoes regression with respect to the time interval between the sales or the holding periods (t–t_{-1}):

$$(\hat{e}_{nt} - \hat{e}_{n\tau})^2 = \alpha_O + c_t\left(t - \tau\right), \qquad (6)$$

which yields the estimated variance ($\hat{\sigma}_H^2, \hat{\sigma}_U^2$). Lastly, the weighted least squares approach is used to perform re-estimations through Equation (4), with the diagonal element being $\sqrt{\hat{\sigma}_U^2 + \left(t - \tau\right)\hat{\sigma}_H^2}$.

When employing weighted regression, it is important to acknowledge that certain attribute variables in the repeat-sales model may be omitted. This results in omitted variable bias, which is associated with hedonic regression because these variables are not explicitly required in the repeat-sales estimation process. However, given that implicit prices change with time, using the ORS can produce biased results, as it lacks the capacity to control for these implicit prices (Owusu-Ansah, 2018). In contrast to previous studies that used single-level repeat-sales modeling to estimate house prices (Xu et al., 2018; Hill & Trojanek, 2022), we adopted the HLGM approach, in which the index was developed and the differences in repeat-sales house prices were analyzed using a two-level model.

3.3.2. Hierarchical linear growth model settings

Before performing HLGM, it is essential to assess its suitability through the use of a null model. The null model, in this context, takes the form of a one-way ANOVA with random effects. This model does not incorporate any independent variables into either of its two levels. Instead, it serves as an initial model to determine whether HLGM or traditional regression should be applied for the subsequent analysis. The primary objective of the null method is to detect for the presence of significant differences in the repeat-sales price variation of a single house across multiple house types. The null model is described by Equations (7) and (8):

$$\ln Price = \beta_0 + \varepsilon, \qquad (7)$$

where: $\ln Price$ is the logarithm of the sales price; β_0 is an intercept that signifies the first sales price of the house; ε is the error term.

$$\text{Level-2 Model: } \beta_0 = \gamma_{00} + \mu_0, \qquad (8)$$

where: β_0 is the grand mean of the first sales price of all housing sales; μ_0 is the error term.

Next, the HLG model settings were configured. In this context, the Level-1 variables were sales time (*Time*), house age (*Age*), and square of house age (*Age*^{2}), each of which indicate the extent of house depreciation. These Level-1 variables were chosen to capture the influence of the growth rate of housing sales prices and the age of the house on housing sales prices. It is important to note that the repeat-sales method does not treat house age at different sales times as a fixed quality. Therefore, the repeat-sales house price index can be considered to be nearly of fixed quality (Leishman & Watkins, 2002). The model settings are shown in Equations (9) to (13):

Level-1 Model:

$$\ln Price = \beta_0 + \beta_1 Time + \beta_2 Age + \beta_3 Age^2 + \varepsilon, \qquad (9)$$

where: *Time* is the sales year minus the baseline year (2013 in this study). After subtraction, the repeat-sales time variable consists of 10 time points (0 to 9). The initial time point is designated as 0 and serves as the reference point, denoted as the initial state. The variable *Age* corresponds to the house's age at the time of sale, while *Age*^{2} signifies the square of the house's age at the time of sale. Lastly, ε is the error term.

$$\text{Level-2 Model: } \beta_0 = \gamma_{00} + \mu_0; \tag{10}$$

$$\beta_1 = \gamma_{10}; \tag{11}$$

$$\beta_2 = \gamma_{20}; \tag{12}$$

$$\beta_3 = \gamma_{30}, \tag{13}$$

where: γ_{00} is the growth rate of the mean housing sales price from 2013 to 2022; γ_{10} is the variation in the different housing sales prices over time and represents the growth rate of the housing sales price; γ_{20} represents the impacts of house age on housing sales price; and γ_{30} is the coefficient of the square of house age; μ_0 is the error term. Substituting Equations (10) to (13) into Equation (9) yields the mix adjustment model, expressed in Equation (14) as follows:

$$\ln Price = \gamma_{00} + \gamma_{10} \times Time + \gamma_{20} \times Age + \gamma_{30} \times Age^2 + \varepsilon + \mu_0 . \tag{14}$$

The coefficient can be estimated through Equation (14). After calculating the annual estimated price, the annual house price index can be obtained by substituting the estimated price into Equation (2).

3.4. Comparison metrics

A prediction index serves as an important basis for measuring house price indexes. Widely used metrics for evaluating indexes include the MSE, MAPE, MAE, and RMSE. It is essential to note that lower values for these metrics indicate a higher degree of accuracy and reliability in the model's predictions (Ho et al., 2021; Nazemi & Rafiean, 2022).

3.4.1. Mean standard error

The MSE measures the square of the difference between the predicted and actual values. It only considers the mean size of the error. A smaller MSE indicates that the model has a better accuracy. The MSE is represented in Equation (15) as follows:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\left(f_i - y_i\right)^2, \tag{15}$$

where: f_i is the predicted value; y_i is the actual value; N is the sample size (all the sales data in each model).

3.4.2. Root-mean-square error

The RMSE measures the square root of the difference between the predicted and actual values. A larger RMSE indicates that the model has a poorer accuracy. The *RMSE* is represented in Equation (16) as follows:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(f_i - y_i\right)^2} . \tag{16}$$

An RMSE that is equal to 0 indicates that the model is a perfect prediction model.

3.4.3. Mean absolute error

The MAE first measures the absolute of the difference between the predicted and actual values, deduces the mean, and then expresses the value as a percentage. The larger the value, the higher the dispersion, and the poorer the predictive power of the model. The *MAE* is represented in Equation (17) as follows:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|f_i - y_i\right| \times 100\%. \tag{17}$$

3.4.4. Mean absolute percentage error

The MAPE first measures each individual difference between the predicted and actual values divided by the actual value, sums up the absolute value of this ratio, and then divides the sum by the sample size. The value is expressed as a percentage. The larger the value, the higher the dispersion, and the poorer the predictive power of the model. The *MAPE* is represented in Equation (18) as follows:

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{f_i - y_i}{f_i}\right| \times 100\% . \tag{18}$$

4. Data sources and processing

4.1. Data sources

The data used in this study were acquired from the Ministry of the Interior's actual price registration system for real estate properties. The 390,324 pieces of data consisted of housing sales data in Kaohsiung City for the third quarter in the years 2013 to 2022. This time frame was selected because the Taiwan government mandated the disclosure of real estate sales prices and housing attributes starting in 2012. Data treatment is a cumbersome process because the data obtained from the actual price registration system for real estate properties must be collated, organized, reviewed, and screened. Moreover, the metrological estimations and model development must be performed on three models, which increases the complexity of the data treatment process. We focused on 13 urban administrative districts in Kaohsiung due to their significance in industrial transformation and urban development. As the largest city in southern Taiwan, the city has been founded for more than a century and plays a key role in the manufacturing, export, petrochemical, and shipbuilding industries. In recent years, the city has undergone industrial transformation through the development of industrial parks and the introduction of high-tech industries, tourism, and filmmaking, fostering a more livable environment. These efforts, along with government planning and private investment, have contributed to rising housing prices. According to the Ministry of the Interior, Kaohsiung's Q3 2024 house price index was **149.49** (baseline year = 2016). To refine the dataset for analysis, we initially excluded data related to rural administrative districts. This left us with the data from 13 districts (Fongshan, Renwu, Nanzi, Zuoying, Sanmin, Niaosong, Qianzhen, Xinxing, Lingya, Yancheng, Gushan, Qiaotou, and Qianjin), comprising a total of 286,582 data pieces. We also excluded data that may have interfered with the authenticity of the

Table 2. Descriptive statistics of the sample

Total sample size (*N* = 137,585)	Mean	Standard deviation (SD)	Minimum	Maximum
Price (in NT$10,000)	795.725	686.976	10	38,500
House age	13.90	12.990	0	63
Square of house age	362.00	481.424	0	3906
House area	42.659	23.654	0.082	2121.617
Land area	5.736	3.997	0.000	565.100
Number of rooms	2.82	1.057	1	5
Number of living rooms	1.77	0.507	1	5
Number of bathrooms	1.78	0.899	1	5
Distance to the nearest MRT station	1229.414	840.577	8.668	5087.302
Distance to the nearest railway station	1846.055	1260.789	6.754	7689.875
Distance to the nearest elementary school	479.862	252.206	2.499	1987.084
Distance to the nearest junior high school	671.775	375.601	23.587	2475.596
Distance to the nearest senior high school	908.613	540.999	10.327	3674.075
	Number of data pieces	Percentage	Cumulative percentage	
House type				
Apartment building	108,829	79.1%	79.1%	
Luxury condo	11,281	8.2%	87.3%	
Condominium	17,475	12.7%	100%	
Floor level				
First floor	11,832	8.6%	8.6%	
Fourth floor	10,869	7.9%	16.5%	
Other floors	114,884	83.5%	100%	

findings, such as non-housing sales data (e.g., pure land lot or parking lot sales); house types other than apartment buildings, luxury condos, and condominiums; non-arm's-length transactions (those involving housing sales between relatives of the first and second degrees of kinship, employees, or other special relations); and houses with additional structures, unregistered buildings, and extended balconies. The descriptive statistics of the remaining 137,585 pieces of data are presented in Table 2.

To obtain the repeat-sales data, we specifically targeted houses that had been sold two or more times while maintaining consistent attributes (such as house area, number of rooms, living rooms, and bathrooms) from the 137,585 pieces of data. After removing data that exhibited variations in housing attributes, we were left with 23,992 pieces of sales data remaining, nested across 11,491 houses sold.

4.2. Statistics of the repeat-sales sample

The Level 1 repeat-sales sample in this study contained 23,992 repeat-sales records across 10 time points. Regarding the dependent variables, the 23,992 repeat-sales records were nested across 11,491 houses sold. Regarding the sales price, the logarithmic mean of the 23,992 repeat-sales records was 15.519 and the standard deviation (SD) was 0.592. Regarding the variables, the mean house age at the time of sale was 16.165 years and the SD was 12.270 years (Table 3).

Table 3. Descriptive statistics of the Level 1 sample (*n* = 23,992)

Variable	Mean	SD	Minimum	Maximum
(ln*Price*) Sales price	15.519	0.592	12.612	18.373
(*Age*) House age	16.165	12.270	0	54

5. Empirical results

5.1. Hedonic price index

When developing the hedonic price index, we separately regressed the data for each year. To ensure that the comparison baseline was the same in every year, we used the mean of the variable in the baseline year as the baseline and multiplied the value by the regression coefficients of the variables in each year. The estimated annual price was then divided by the estimated price of the baseline year and multiplied by 100, thus yielding the annual house price index. We used the variance inflation factor (VIF) to check for the presence of extreme multicollinearity between

Table 4. The hedonic price index of house prices in Kaohsiung City

Year	Price index	Growth rate	Estimated price
2013	100	0	15.387
2014	101.007	1.007%	15.542
2015	101.501	0.494%	15.618
2016	101.482	−0.019%	15.615
2017	102.333	0.851%	15.746
2018	101.820	−0.513%	15.667
2019	101.885	0.065%	15.678
2020	102.366	0.481%	15.751
2021	102.847	0.481%	15.825
2022	104.237	1.390%	16.039

the explanatory variables. In general, a VIF smaller than 5 indicates the absence of extreme multicollinearity between the explanatory variables. All the VIFs were smaller than 5 except for house age, the square of house age, and several dummy variables, indicating the absence of extreme multicollinearity. The hedonic price index is shown in Table 4.

According to Table 4, the hedonic house price index for house prices in Kaohsiung City rose 1.007% from 2013 to 2014, fell 0.019% from 2015 to 2016, rose significantly by 0.851% from 2016 to 2017, fell 0.513% from 2017 to 2018, and then rose continuously from 2018 onwards, with a cumulative growth rate of 2.352% from 2019 to 2022. Despite several fluctuations during this period, generally speaking, it is apparent that the hedonic price index of house prices in Kaohsiung City has grown steadily over the years.

5.2. Clustering median adjustment index

First, we performed hierarchical cluster analysis by setting the six variables of the logarithm of the sales price (ln*Price*), house age (*Age*), house area (*Area*), number of rooms (*Room*), number of living rooms (*LivRoom*), and number of bathrooms (*BathRoom*) as cluster variables. To avoid excessive complexity caused by a large number of housing attribute variables, we clustered the data according to typical and important attribute variables. We then proceeded with *k*-means analysis at a *k* value of 6, as there were six clusters. Afterwards, we performed variance analysis, in which a larger *F*-value indicates a larger between-group variance and a smaller within-group variance. The *F*-values in this study were all considerably significant, which attests to the rationality and suitability of dividing the sales data into six clusters.

Next, we developed the index by multiplying the median price in each cluster (level) by the cluster's percentage of representation in a particular year (reflecting the relative importance of the level), thus yielding the mix-adjusted price in each level. By summing up these mix-adjusted prices, we arrived at an estimated price for a particular year. This estimated price was then subsequently substituted into Equation (2) to compute the price index. To illustrate this process, let us consider the example of 2013 and 2014, as shown in Tables 5 and 6. From these tables, we can see the data of the baseline year (2013) and the mix-adjusted data of each level in 2014. The estimated price is calculated by multiplying the median house price in each level with the level's percentage of representation and then taking the sum of the house price in all levels to derive the estimated price of a particular year. Substituting this estimated price into Equation (2) yields the price index, as shown in Table 7.

According to Table 4, the mix adjustment house price index of house prices in Kaohsiung City rose steadily from 2013 to 2017 at a cumulative growth rate of 2.424%, fell 0.862% from 2017, with a cumulative growth rate of 2.424%. It fell by 0.862% from 2017 to 2018, and then rose steadily again from 2018 to 2021, with a cumulative growth rate of 1.361%. However, it dropped by 0.207% from 2021 to 2022.

Table 5. The clustering median adjustment data at each level (cluster) in the baseline year (2013)

Level	Price median	*N*	Percentage of representation	CMAM
Level 1	14.648	4,610	0.252	3.690
Level 2	15.407	8,215	0.449	6.916
Level 3	15.895	4,228	0.231	3.672
Level 4	16.732	1,003	0.055	0.917
Level 5	17.210	214	0.012	0.201
Level 6	17.630	31	0.002	0.030
Total		18,310	100%	15.426

Table 6. The clustering median adjustment data at each level (cluster) in 2014

Level	Price median	*N*	Percentage of representation	CMAM
Level 1	14.845	4,209	0.274	4.067
Level 2	15.511	6,472	0.421	6.533
Level 3	15.961	3,622	0.236	3.762
Level 4	16.792	818	0.053	0.894
Level 5	17.346	212	0.014	0.239
Level 6	17.398	32	0.002	0.036
Total		15365	100%	15.532

Table 7. Annual estimation results of the clustering median adjustment data in Kaohsiung City

Year	Price index	Growth rate	Estimated price
2013	100.000	0	15.426
2014	100.687	0.687%	15.532
2015	100.758	0.071%	15.543
2016	101.653	0.895%	15.681
2017	102.424	0.771%	15.800
2018	101.562	−0.862%	15.667
2019	101.757	0.194%	15.697
2020	102.023	0.266%	15.738
2021	102.924	0.901%	15.877
2022	102.716	−0.207%	15.845

5.3. Hierarchical linear growth repeat-sales index

According to the null model, the differences between the mean prices of all houses were significant, indicating that the HLG model was well suited for our analysis. The estimation results obtained from the HLG model diverged from those of the hedonic price and mix adjustment price models, which had computed sales data separately. To investigate the growth in sales prices and other relevant data for the same house across different time periods, we regressed all the repeat-sales data without annual regression estimates. The empirical results of the HLG model are presented in Table 8. The estimated coefficient of the sales price (*Price*) of each house was 15.779 and achieved a significance level of 1%. The estimated coefficient of the time of repeat-sales (*Time*) of each house was 0.062 and achieved a 1% level of significance. This shows that the sales price of the same house grew annually by 6.2% from 2013 onwards. Meanwhile, the estimated coefficient of house age (*Age*) was –0.038 and achieved a significance level of 1%. Moreover, the estimated coefficient of the square of house age was 0.0002 and achieved a 1% level of significance. These results suggested that the relationship between house price and house age was non-linear. The annual house price coefficients are presented in Table 9.

Table 8. Estimation results of the HLGRM

Variable	Coefficient	Standard error	T	*p*-value
ln*Price*	15.779	0.007	2471.699	0.001***
Time	0.062	0.0006	97.876	0.001***
Age	–0.038	0.0007	–84.762	0.001***
Age^2	0.0002	0.00002	12.932	0.001***

Note: *** denotes $p < 0.01$.

Table 9. The annual house price coefficients estimated by the HLGRM

Year	ln*Price*
2013	15.270
2014	15.330
2015	15.390
2016	15.450
2017	15.520
2018	15.570
2019	15.625
2020	15.690
2021	15.750
2022	15.810

To calculate the estimated annual house prices in Kaohsiung City, we substituted the data provided in Table 10 into Equation (14) (for example, P_2013 = 15.270 + 0.062(0) – 0.038(13.222) + 0.0002(288.025) = 14.825; P_2014 = 15.330 + 0.062(1) – 0.038(14.624) + 0.0002(347.708) = 14.906). The estimated annual prices was then substituted into Equation (2) to derive the annual house price index, as shown in Table 11.

Table 10. Descriptive statistics of the repeat-sales data in Kaohsiung City

Year	2013		2014	
Variable	Mean	SD	Mean	SD
ln*Price*	15.362	0.602	15.425	0.566
Age	13.222	10.641	14.624	11.571
Age^2	288.025	358.624	347.708	407.269
Year	2015		2016	
ln*Price*	15.447	0.560	15.531	0.574
Age	15.118	11.286	14.001	12.041
Age^2	355.870	404.160	341.075	428.925
Year	2017		2018	
Variable	Mean	SD	Mean	SD
ln*Price*	15.586	0.536	15.454	0.602
Age	13.015	10.397	17.904	12.642
Age^2	277.430	347.035	480.280	511.261
Year	2019		2020	
ln*Price*	15.509	0.582	15.543	0.568
Age	18.518	12.992	19.302	12.817
Age^2	511.664	548.418	536.771	560.133
Year	2021		2022	
ln*Price*	15.706	0.588	15.832	0.598
Age	18.258	13.005	19.184	13.228
Age^2	502.454	569.794	542.858	598.386

Note: Because the variable Time refers to the year minus the baseline year, its mean merely increases by 1 annually; therefore, the mean of Time was excluded from this table.

Table 11. Annual estimation results of the HLGRM data in Kaohsiung City

Year	Price index	Growth rate	Estimated price
2013	100	0	14.825
2014	100.546	0.546%	14.906
2015	101.255	0.708%	15.011
2016	102.341	1.086%	15.172
2017	103.400	1.059%	15.329
2018	103.177	−0.223%	15.296
2019	103.852	0.675%	15.396
2020	104.540	0.688%	15.498
2021	105.585	1.046%	15.653
2022	106.226	0.641%	15.748

According to Table 11, the repeat-sales house price index of house prices for Kaohsiung City rose steadily from 2013 to 2017 at a cumulative growth rate of 3.4%, fell by 0.223% from 2017 to 2018, and then rose steadily again from 2018 to 2022 at a cumulative growth rate of 3.049. Despite several fluctuations during this period, generally speaking, it is apparent that the repeat-sales house price index of house prices for Kaohsiung City has grown steadily over the years.

6. Discussion

Figure 1 shows the house price indexes in Kaohsiung City created using the three aforementioned methods. The trends of the three indexes were extremely similar, with a steady rise from 2013 to 2015 (hedonic price index = 1.501%, mix adjustment index = 0.758%, HLG repeat-sales index = 1.255%, respectively). The first trend difference occurred in the 2015–2016 period, during which the hedonic price index dropped by 0.019% while the mix adjustment index and the HLG repeat-sales index rose by 0.895% and 1.086%, respectively. Then, from 2016 to 2021, the trends of the three indexes became extremely similar once again, rising by 0.851%, 0.771%, and 1.059% from 2016 to 2017, respectively, then falling by 0.513%, 0.862%, and 0.223% from 2017 to 2018, and rising by 1.027%, 1.362%, and 2.408% from 2018 to 2021. The second trend difference occurred in the 2021–2022 period, in which the hedonic price index and the HLG repeat-sales index rose to record highs of 104.237 and 106.226, respectively, while the mix adjustment index fell to 102.695. Despite falling on several occasions, the three indexes exhibited a steady long-term growth from 2013 to 2022.

The Ministry of the Interior calculates its house price index quarterly using hedonic price modeling, taking 2016 as the baseline year. To provide a clearer representation of the trend, we generated an annual price index for the Ministry by averaging the indexes for the four quarters within each year, setting 2013 as the baseline year. The trend graph is presented in Figure 2. Our hedonic price index closely aligns with the Ministry's index trend, notably reflecting two distinct price drops: one occurring between 2015 and 2016, and another from 2017 to 2018.

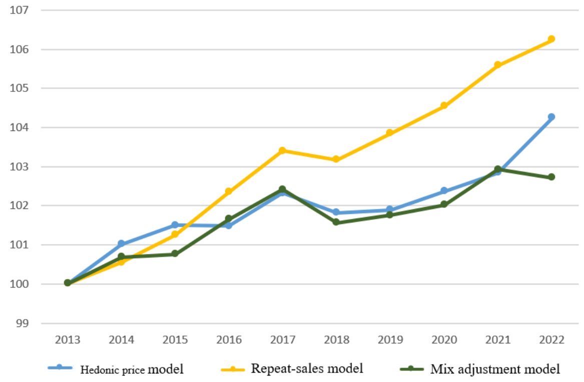

Figure 1. The three house price indexes in this study

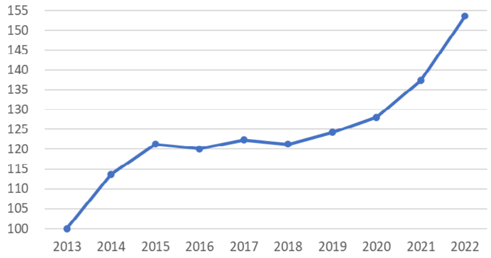

Figure 2. The house price index of the Ministry of the Interior

The prediction errors of the indexes in this study were calculated by subtracting the predicted value in the sample with the actual value. Our approach is distinguished from those adopted in previous studies, which have tended to use a single metric to compare indexes. For example, Owusu-Ansah (2018) only used the MSE to compare the performances of the HPM and the single-level regression repeat-sales model. Similarly, other studies like that of Prasad and Richards (2008) focused primarily on a single metric, RMSE, to compare price indexes developed using different stratification methods, like administrative district stratification and price stratification. In that context, the RMSE of the price-stratified index was 1.15%, which proved to be superior to the administrative district-stratified index (1.95%). However, our study adopted a more comprehensive approach by employing four metrics for index comparison: MSE, MAPE, MAE, and RMSE, which scored 0.072, 1.176, 0.181, and 0.181 on the hedonic price index, respectively. Meanwhile, the MSE, MAPE, MAE, and RMSE of the mix adjustment price index were 0.154, 1.905, 0.293, and 0.293, respectively. Finally, the MSE, MAPE, MAE, and RMSE of the HLG repeat-sales price index were 0.309, 2.804, 0.439, and 0.439, respectively.

As mentioned in the Introduction, even though the HPM suffers from multicollinearity, this issue can paradoxically enhance the model's predictive power (Mundfrom et al., 2018). In our study, the hedonic price index exhibited the lowest prediction error, underscoring its strong performance. Furthermore, although the performance of the mix adjustment index was lower than that of the hedonic price index, it was still better than the HLG repeat-sales model. This performance may be associated with the selected level (cluster) variables and the number of levels (clusters).

According to de Haan and Diewert (2011), mix adjustment indexes are extremely sensitive to the changes in the attribute combinations of sold houses. To a certain extent, this problem is similar to the attribute omission problem in hedonic price modeling. In reality, though, resolving this problem through rigorous stratification would only reduce the number of observed housing sales in each level (cluster). Consequently, the sample index has a higher standard deviation and lower accuracy. The two-level regression results of our HLG repeat-sales index were unsatisfactory, with it being the worst-performing index with respect to the single-metric or overall prediction errors. The main drawback of the repeat-sales method is sample selection bias. According to Clapp and Giaccotto (1992), if the data contains houses that are sold frequently, then these houses may be overrepresented in the sales data. This is caused by several factors, such as young house owners frequently upgrading their houses and relatively cheaper "starter" houses being sold more frequently. Furthermore, the sample used to construct a repeat-sales index often excludes "brand-new" houses, as they cannot be sold repeatedly unless they were already sold upon completion (Costello & Watkins, 2002). The ratio of houses that were sold repeatedly may be very low in the repeat-sales index sample, which is relatively smaller than that of the hedonic

Table 12. Prediction error of each index

Model	MSE	MAPE	MAE	RMSE
HPM	0.072	1.176	0.181	0.181
CMAM	0.154	1.905	0.293	0.293
HLGRM	0.309	2.804	0.439	0.439

price index sample. Moreover, the repeat-sales method does not specifically consider the effects of house depreciation on house prices. When adopting this approach, assuming that there are no physical alterations to the house as it is sold repeatedly, its age will still increase between each sale, meaning the repeat-sales index may underestimate the appreciation of house prices (Clapp & Giaccotto, 1992). In summary, the hedonic price index had the lowest overall prediction error and single-metric prediction error as well as the best performance among the three indexes, followed by the mix adjustment price index, and then the HLG repeat-sales index. More detailed data can be found in Table 12.

7. Conclusions and recommendations

7.1. Conclusions

Existing studies have demonstrated that the HPM outperforms the RSM in terms of prediction error rate (Anthony, 2018; Hill & Trojanek, 2022). Consistent with these studies, our HPM also had the best prediction accuracy despite having a similar market trend performance to the CMAM and HLGRM.

Although the CMAM and HLGRM are less accurate than the HPM, both models are indispensable for analyzing the diversity of and dynamic fluctuations in the market, as well as meeting the demands of different stakeholders. By analyzing market segmentation through agglomeration, the CMAM reflects the price trends of different house types and has a high utility when addressing the diverse demands of buyers. The CMAM and HLGRM can help investors identify market heterogeneity, analyze the price variation trends of different house types, support their decision-making, and lower risks. These methods accurately portray the market structures, guiding policymakers in developing targeted policies that align with actual needs. Despite the HPM's superior accuracy, the ability of the CMAM and HLGRM to analyze market segmentation and assess dynamic fluctuations provides an additional reference for understanding market changes and decision-making.

The CMAM and HLGRM are theoretically suitable for addressing market heterogeneity and dynamic fluctuations. However, our findings of HPM's better overall accuracy suggest that even though the CMAM and HLGRM are able to thoroughly analyze market structures, they are more suitable as assistive instruments in specific scenarios. In stable markets, the HPM remains the top choice due to its high prediction accuracy. However, the CMAM and HLGRM can yield more valuable insights in the presence of larger market fluctuations or heterogeneity. For example, the CMAM effectively reveals the price trends in a segmented market in clusters pertaining to special house types (townhouses or apartment buildings) or special house locations; the HLGRM accurately reflects the price variation trends in long-term price predictions and provides important information for policymakers and long-term investors.

To summarize, rather than replacing the HPM or RSM, the CMAM and HLGRM instead expand the former two models, particularly when there is a need to thoroughly analyze market fluctuations and heterogeneity. The CMAM and HLGRM assist analysts in identifying the key factors that affect house prices in ever-evolving markets and thus increase their prediction accuracy. They also support more effective policy formulation and market decision-making.

7.2. Limitations and recommendations for future research

In this study, using data sourced from the Ministry of the Interior's actual price registration system for real estate properties, we analyzed the performance of three different house prices indexes in an area encompassing the 13 administrative districts with the highest house prices in Kaohsiung City. A house price index is designed to track dynamic market variations influenced by market structure, economic conditions, and policy environments. Previous studies suggest that HPM performs well in stable markets, while the RSM is more suitable for new markets with active trading, though it has its own limitations. The CMAM and HLGRM models proposed in this study address weaknesses in existing approaches, such as sample selection bias, non-uniform sales distribution over time, and market heterogeneity, thereby improving house price index development. However, we cannot generalize whether these models will yield the same outcomes in all markets. It is crucial to acknowledge that housing markets can exhibit significant variations across different counties and cities, and that, therefore, the findings of this study may not be representative of housing sales on a national scale. Hence, future studies could expand the study area to cover all counties and cities in Taiwan in order to explore whether index performances differ regionally.

We believe there is no universal approach to developing a house price index suitable for all regions, as the primary goal of such an index is to reflect market variations and to assist stakeholders in decision-making. Market environments, economic structures, and policy contexts differ across regions, influencing the applicability of house price index methodologies. Additionally, market participants (e.g., government agencies, developers, investors) have varying objectives and prioritize different market variables. Therefore, we emphasize that models should be tailored to specific market conditions rather than applied uniformly. The CMAM and HLGRM models demonstrate advantages in certain market settings by improving the accuracy and operability of house price indexes.

Although the 2013–2022 data period covers different market cycles, it remains relatively short and may influence the model's prediction error results. A longer data period would allow for a more comprehensive assessment of the long-term stability and predictive accuracy of house price indexes. In this study, we mainly compared the time period limitations on the prediction errors of different models using the same dataset, which may reflect only their short-term predictive performance rather than long-term trends. Due to data period constraints in this study, we recommend that future research extend the data period and compare model results using other cities or regions with similar characteristics to Kaohsiung. This would enhance our understanding of the CMAM and HLGRM models' applicability in rapidly developing historical cities and clarify how market structure, economic conditions, and policy environments affect model accuracy. Further optimization–such as addressing data limitations or integrating our models with other methodologies–could improve predictive performance and broaden applicability.

Next, because the variables in the mix adjustment model are clustered, the generated results may also be different. This means that clustering can be performed using different variables in order to delineate the differences between the performances of mix-adjusted price models. Additionally, incorporating the housing attributes into the second level of the HLG repeat-sales model could enhance its estimation performance. Finally, house price indexes can be developed using big data, machine learning, and artificial intelligence, and their performances can be compared with those of indexes developed through other methods.

References

Anthony, O. A. (2018). *Construction and application of property price indices*. Routledge.

Bailey, M. J., Muth, R. F., & Nourse, H. O. (1963). A regression method for real estate price index construction. *Journal of the American Statistical Association*, *58*(304), 933–942. https://doi.org/10.1080/01621459.1963.10480679

Cannaday, R. E., Munneke, H. J., & Yang, T. T. (2005). A multivariate repeat-sales model for estimating house price indices. *Journal of Urban Economics*, *57*(2), 320–342. https://doi.org/10.1016/j.jue.2004.12.001

Case, K. E., & Shiller, R. J. (1987). Prices of single-family homes since 1970: New indexes for four cities. *New England Economic Review*, *9*, 45–56. https://doi.org/10.3386/w2393

Case, K. E., & Shiller, R. J. (1989). The efficiency of the market for single-family homes. *American Economic Review*, *79*(1), 125–137.

Clapp, J. M., & Giaccotto, C. (1992). Estimating price indices for residential property: A comparison of repeat sales and assessed value methods. *Journal of the American Statistical Association*, *87*(418), 300–306. https://doi.org/10.1080/01621459.1992.10475209

Costello, G., & Watkins, C. (2002). Towards a system of local house price indices. *Housing Studies*, *17*(6), 857–873. https://doi.org/10.1080/02673030216001

de Haan, J., & Diewert, W. E. (Eds.) (2011). *Handbook on residential property price indexes*. Eurostat.

Francke, M. K., & Van de Minne, A. (2017). The hierarchical repeat sales model for granular commercial real estate and residential price indices. *The Journal of Real Estate Finance and Economics*, *55*, 511–532. https://doi.org/10.1007/s11146-017-9632-1

Hill, R. J., & Trojanek, R. (2022). An evaluation of competing methods for constructing property price indexes: The case of Warsaw. *Land Use Policy*, *120*, Article 106226. https://doi.org/10.1016/j.landusepol.2022.106226

Ho, W. K., Tang, B. S., & Wong, S. W. (2021). Predicting property prices with machine learning algorithms. *Journal of Property Research*, *38*(1), 48–70. https://doi.org/10.1080/09599916.2020.1832558

Kennedy, P. (2008). *A guide to econometrics*. John Wiley & Sons.

Kim, D. H., & Irakoze, A. (2022). Identifying market segment for the assessment of a price premium for green certified housing: A cluster analysis approach. *Sustainability*, *15*(1), Article 507. https://doi.org/10.3390/su15010507

Kwon, S., Kim, S., Tak, O., & Jeong, H. (2017). A study on the clustering method of row and multiplex housing in Seoul using K-means clustering algorithm and hedonic model. *Journal of Intelligence and Information Systems*, *23*(3), 95–118.

Lee, C. C., Huang, L. Y., & You, S. M. (2013). The changes and trends in urban land prices: An application of hierarchical growth modelling. *Asian Economic and Financial Review*, *3*(5), Article 579.

Lee, C. C., Wang, Y. C., Liang, C. M., & Yu, Z. (2023). Price changes of repeat-sales houses in Kaohsiung city: Analyses based on hierarchical linear growth models. *International Journal of Strategic Property Management*, *27*(5), 290–303. https://doi.org/10.3846/ijspm.2023.19935

Leishman, C., & Watkins, C. (2002). Estimating local repeat sales house price indices for British cities. *Journal of Property Investment and Finance*, *20*(1), 36–58. https://doi.org/10.1108/14635780210416255

Miller, R., & Maguire, P. (2020). A rapidly updating stratified mix-adjusted median property price index model. In *2020 IEEE Symposium Series on Computational Intelligence* (pp. 9–15), Canberra, ACT, Australia. IEEE. https://doi.org/10.1109/SSCI47803.2020.9308235

Mundfrom, D., Smith, M. D., & Kay, L. (2018). The effect of multicollinearity on prediction in regression models. *General Linear Model Journal*, *44*(1), 24–28. https://doi.org/10.31523/glmj.044001.003

Nazemi, B., & Rafiean, M. (2022). Modelling the affecting factors of housing price using GMDH-type artificial neural networks in Isfahan city of Iran. *International Journal of Housing Markets and Analysis*, *15*(1), 4–18. https://doi.org/10.1108/IJHMA-08-2020-0095

Prasad, N., & Richards, A. (2008). Improving median housing price indexes through stratification. *Journal of Real Estate Research*, *30*(1), 45–72. https://doi.org/10.1080/10835547.2008.12091213

Rahman, S. N. A., Maimun, N. H. A., Razali, M. N. M., & Ismail, S. (2019). The artificial neural network model (ANN) for Malaysian housing market analysis. *Planning Malaysia*, *17*(1), 1–9. https://doi.org/10.21837/pmjournal.v17.i9.581

Studenmund, A. H. (2014). *Using econometrics: A practical guide*. Pearson Education Limited.

Tan, R., He, Q., Zhou, K., Song, Y., & Xu, H. (2019). Administrative hierarchy, housing market inequality, and multilevel determinants: A cross-level analysis of housing prices in China. *Journal of Housing and the Built Environment*, *34*, 845–868. https://doi.org/10.1007/s10901-019-09690-y

Xu, Y., Zhang, Q., Zheng, S., & Zhu, G. (2018). House age, price and rent: Implications from land-structure decomposition. *The Journal of Real Estate Finance and Economics*, *56*, 303–324. https://doi.org/10.1007/s11146-016-9596-6

Zhang, X., Zheng, Y., Sun, L., & Dai, Q. (2019). Urban structure, subway system and housing price: Evidence from Beijing and Hangzhou, China. *Sustainability*, *11*(3), Article 669. https://doi.org/10.3390/su11030669