The current issue and full text archive of this journal is available on Emerald Insight at:
https://www.emerald.com/insight/0144-3585.htm

Enhancing property value prediction: a regression model with fine-grid geographic segmentation

Journal of Economic Studies

Hongjai Rhee
School of Business, Ajou University, Suwon, South Korea

Received 22 December 2024
Revised 15 January 2025
Accepted 1 February 2025

Abstract

Purpose – We propose a novel scalable regression model that incorporates two-dimensional location dummies based on longitudinal and latitudinal coordinates. This approach easily achieves comparable levels of explanatory power to ML models while avoiding the interpretability challenges inherent in non-parametric ML techniques.

Design/methodology/approach – Regression models have long been a fundamental and effective toolkit in economic analysis, facilitating the validation of theoretical predictions and the examination of policy effects. However, their explanatory power is often constrained by linearity constraints or limited information, which can, in turn, distort the estimated marginal effects of explanatory variables. While machine learning (ML) models can serve as an alternative, their improved fit often comes at the cost of increased complexity in interpretation and inference. This paper proposes that incorporating geocoded information into the regression models for property values can significantly enhance explanatory power to a level comparable to ML models while preserving the simplicity of regression approaches. As a case study to validate the proposed method, we examine the impact of a recent deregulation policy in Korea on the market prices of the corresponding properties.

Findings – By accounting for location-specific heterogeneity in its full magnitude, our model provides more accurate and reliable estimates for policy evaluation. As a case study, this paper examines the effect of deregulation on apartment redevelopment in Korea, demonstrating how the proposed methodology can be applied to real-world policy analysis.

Originality/value – We propose a novel regression model that incorporates grid dummies for the longitudinal and latitudinal coordinates of the properties. This approach easily achieves comparable levels of explanatory power to ML models while avoiding the interpretability challenges inherent in non-parametric ML techniques.

Keywords Regression model, Policy effects, Deregulation, Housing prices, Geocoding

Paper type Research paper

1. Introduction

“Location, Location, Location” is a popular British reality property programme, named after the cliché that underscores geographical location as the most critical factor influencing property prices. This paper emphasizes the importance of incorporating fine-grained geographic segmentation into regression models for analyzing policy effects on housing prices.

As a policy example, the Korean government announced significant deregulation of apartment reconstruction in January 2024 to address skyrocketing housing prices by increasing the housing supply. Previously, apartment reconstruction was largely restricted to buildings over 20 years old, subject to stringent safety inspections, and often required an extended period to complete. Consequently, the number of apartments over 50 years old has grown significantly in Seoul, creating substantial inconveniences for residents. At the same time, older apartments that successfully underwent reconstruction have commanded substantial premiums in the market. Under the deregulation plan, safety inspection requirements will be significantly relaxed, and administrative procedures streamlined. This is expected to increase the number of apartments eligible for reconstruction and significantly reduce the time required for the process.

Journal of Economic Studies
Vol. 53 No. 2, 2026
pp. 273-282

e-ISSN: 1758-7387
p-ISSN: 0144-3585
DOI 10.1108/JES-12-2024-0856

JEL Classification — C5, R3

The effect of apartment age on asset value reflects two opposing forces: depreciation due to aging and appreciation driven by reconstruction expectations. The market evaluates property values by balancing these two factors. Deregulation is expected to positively influence the asset prices of older apartments by increasing the likelihood of reconstruction and reducing associated temporal costs.

To analyze these policy effects, two types of models can be applied: the traditional hedonic regression model (Chau and Chin, 2003) and the more recently popular machine learning (ML) models. The primary objective of this study is to evaluate the impact of a recent deregulation policy targeting aging apartment complexes in Korea using a novel geocoded regression framework. Additionally, this study aims to demonstrate that the explanatory power of the proposed regression models is quite comparable to that of machine learning models while retaining simplicity and interpretability. Unlike conventional regression approaches, the proposed model leverages fine-grid geographic segmentation to capture localized spatial heterogeneity, offering new insights into property value dynamics.

The structure of the paper is as follows: Chapter 2 reviews recent literature on housing prices and compares the strengths and weaknesses of ML and regression approaches. Chapter 3 introduces the data used in the analysis. Chapter 4 summarizes the estimation results from various regression models with different levels of geographic segmentation, particularly introducing a novel model that uses geolocation coordinates in a grid format. Chapter 5 applies the grid model to analyze the effects of deregulation, and Chapter 6 concludes the paper.

2. Literature

Economic studies have heavily relied on regression-based models to decompose variables of interest into potential sources of variation. However, due to limitations such as insufficient information or linear constraints, these models often suffer from a lack of explanatory power. Consequently, there has been a growing trend in recent years toward the adoption of machine learning (ML) models (Athey and Imbens, 2019), which exhibit superior explanatory capabilities.

A common theme in the ML literature on property values is the demonstrated superiority of ML models over traditional regression methods. For example, Zaki *et al*. (2022) illustrated how the "Extreme Gradient Boosting" model achieved a remarkable R-squared of 0.84 for Boston's housing prices, nearly doubling the performance of the conventional hedonic regression model. Similarly, Kim and Kim (2023) applied ML models to develop a transaction-based apartment price index in Seoul, Korea, demonstrating that the random forest (RF) model achieved an impressive R-squared of 0.997, compared to 0.846 from the regression model. Rico-Juan and de La Paz (2021) analyzed housing market asking prices in Alicante, Spain, and found that most ML models outperformed regression models, with the RF model achieving the highest predictive accuracy, boasting an R-squared value of 0.97 compared to 0.80 for regression models. In another study, Soltani *et al*. (2022) analyzed 428,000 transaction records in Australia, revealing that an ensemble tree model provided the best fit for large datasets. Tchuente and Nyawa (2022) conducted a comparative analysis of various ML models for housing price data across nine French cities, showing that while ML models performed greatly, their effectiveness varied depending on location and the choice of covariates, including geocoding coordinates of the houses. Similarly, Ho *et al*. (2021) compared the performance of various ML models in predicting housing prices in Hong Kong. Zhang *et al*. (2023) also employed ML algorithms combined with temporal regression to offer a comprehensive understanding of housing price trends.

The general distinctions between ML and regression models are well-documented (Breiman, 2001). In summary, ML models are non-parametric and data-driven, excelling with large datasets and offering enhanced explanatory fit. In contrast, regression models are parametric and theory-driven, offering simpler interpretation and inference. However, ML models pose challenges and the risk of overfitting (Lundberg, 2017). While ML models excel in predictive performance, their non-parametric nature often obscures the intuitive understanding of how individual features influence the target variable. In contrast, regression models provide clear and interpretable insights into the marginal effects of policy variables by directly evaluating the magnitude and significance of estimated parameters.

In the context of housing policy evaluation, where understanding the specific impact and statistical significance of policy interventions is crucial, this interpretability becomes a decisive advantage. Consequently, this paper emphasizes regression models that achieve explanatory power comparable to ML models while maintaining simplicity and interpretability. Adhering to the principle of Occam's razor, model simplicity is particularly valuable in policy-driven analyses, where actionable and transparent insights are essential for informed decision-making.

Specifically, we propose a scalable regression model that incorporates two-dimensional location dummies based on longitudinal and latitudinal coordinates. This approach easily achieves comparable levels of explanatory power to ML models while avoiding the interpretability challenges inherent in non-parametric ML techniques. Moreover, while traditional regression models often suffer from limited explanatory power when applied to spatially heterogeneous data, and machine learning models, though more powerful in terms of fit, face challenges related to interpretability, our model combines the strengths of both approaches. It incorporates fine-grid geographic segmentation to capture spatial heterogeneity, while maintaining the interpretability of regression methods. In sum, our approach aims to achieve both robust explanatory power and actionable interpretability, addressing limitations not fully addressed in existing literature.

3. Data

The dataset used in this study comprises nearly all apartment sales records in Seoul, Korea, for 2023 and 2024, obtained from publicly available records on the Seoul Metropolitan Government's website [1]. In Korea, real estate transactions must legally report the purchase price and physical details within a specified period, with incomplete submissions requiring resubmission. Consequently, the dataset includes comprehensive information such as the actual transaction price (in ten thousand KRW), transaction month, property location (districts "gu" and "dong"), apartment size (in square meters), floor level, year of construction, apartment name, and street address.

After excluding approximately 1% of records with missing values in some variables used for analysis, a total of 87,130 valid records were utilized. A summary of the dataset is provided in Table 1.

Table 1. Descriptive statistics

Variable	Mean and standard deviation (in parenthesis) Year 2023	Year 2024
No. of observations	35,020	52,110
Price (in 10 million KRW)	106.24 (77.5)	118.88 (87.3)
Size (in square meter)	74.32 (30.2)	76.30 (27.4)
Floor level	9.94 (6.57)	9.83 (6.48)
Age (in years)	19.37 (11.49)	20.74 (11.25)
Brand (D)	0.24 (0.42)	0.25 (0.43)

Source(s): Author's own creation

The key variable “Age” represents the number of years that have passed since the first construction year of each apartment. Additionally, the “Brand = 1” is a dummy variable, taking a value of 1 if the apartment belongs to one of the top 8 popular construction companies.

Spatial location is likely to be one of the most significant sources of price levels and price appreciations (Anselin and Lozano-Gracia, 2009; Copiello, 2020). The administrative districts of Seoul are divided into “gu” and “dong.” In total, there are 25 gu’s (as shown in Figure 1), further subdivided into 426 dong’s, which we may utilize as location variables.

4. Analysis

To fully account for potential differences in housing price determinants between 2023 and 2024, we applied separate regression models for each year rather than pooling the data. To enhance model flexibility, we incorporated nonlinear effects for all numerical variables (Size, Floor, and Age) by including their squared and cubed terms as additional predictors. Furthermore, 11 monthly dummy variables were introduced to capture within-year trends and variations in apartment prices. Across all models, the dependent variable was specified as the log of sales price (in 10 million KRW).

To explore the optimal model specification, we employed a sequential locational segmentation strategy. The simplest model [1] includes only the apartment-specific characteristics without any locational information. Model [2] incorporates dummy variables for 24 administrative districts (“gu”), while model [3] uses more granular subdistricts (“dong”) as dummy variables.

We then developed model [4], the most refined regression model in the analysis, leveraging precise locational information derived from the latitude and longitude coordinates of apartments. These coordinates were obtained using Google’s geocoding API applied to the address data in the original dataset. To utilize this information parametrically, we devised a novel algorithm to divide the latitude and longitude coordinates into K = 100 intervals each, generating grid dummy variables. Although this process theoretically results in 10,000

Source(s): Author’s own creation

Figure 1. The geographic map of Seoul

geographic grid dummies, the lack of observations in certain grids led to the creation of 2,482 geographic grid dummy variables for 2024 and 2,414 grid dummy variables for 2023. Unlike subdistrict dummies, the grid-based approach is not constrained by administrative boundaries, enabling a more accurate capture of spatial variation in housing prices.

Tables 2 and 3 summarize the regression models for 2023 and 2024, respectively [2].

Parameter estimates were observed to vary across regression models, with one of the most significant changes related to the parameters associated with age. Notably, in model [4], the effect of age exhibited substantial differences compared to previous models. This discrepancy is likely due to the inclusion of detailed location information in model [4], which may correlate with the age variable. It is evident that estimates from a model with richer information and higher explanatory power are more reliable.

Table 2. Summary of regression results for 2023

	Model [1]	Model [2]	Model [3]	Model [4]
Dependent variable	Log price	Log price	Log price	Log price
Intercept	2.188	1.867	1.132	2.073
Size	0.054	0.053	0.052	0.043
Size^2	−0.0003	−0.0004	−0.0004	−0.0003
Size^3	7.9e−7	8.2e−7	8.1e−7	6.5e−7
Floor	0.0085	0.011	0.012	0.011
Floor^2	0.0002	−6.5e−5	−0.0003	−0.0004
Floor^3	−3.0e−6	1.1e−6	4.4e−6	6.4e−6
Age	−0.031	−0.043	−0.044	−0.039
Age^2	0.0007	0.0016	0.0017	0.0014
Age^3	1.7e−6	−1.4e−5	−1.7e−5	1.6e−5
Brand (D)	0.163	0.187	0.196	0.151
Location (D)	None	24 “gu” dummies	309 “dong” dummies	2,414 spatial grid dummies
R-squared for train data (Adj. *R*-squared)	0.631 (0.631)	0.841 (0.841)	0.901 (0.900)	0.968 (0.966)

Source(s): Author’s own creations

Table 3. Summary of regression results for 2024

	Model [1]	Model [2]	Model [3]	Model [4]
Dependent variable	Log price	Log price	Log price	Log price
Intercept	2.279	1.864	1.325	2.083
Size	0.057	0.057	0.055	0.044
Size^2	−0.0004	−0.0004	−0.0004	−0.0003
Size^3	9.3e−7	9.8e−7	9.4e−7	7.5e−7
Floor	0.0072	0.0081	0.0095	0.0104
Floor^2	0.0002	4.4e−5	−0.0002	−0.0004
Floor^3	−1.5e−6	−4.4e−8	3.3e−6	5.5e−6
Age	−0.034	−0.041	−0.046	−0.023
Age^2	0.0004	0.0011	0.0014	0.0006
Age^3	9.2e−6	−5.3e−6	−1.2e−5	−2.9e−6
Brand (D)	0.158	0.158	0.165	0.166
Location (D)	None	24 “gu” dummies	307 “dong” dummies	2,482 spatial grid dummies
R-squared for train data (Adj. *R*-squared)	0.582 (0.581)	0.826 (0.826)	0.899 (0.899)	0.970 (0.968)

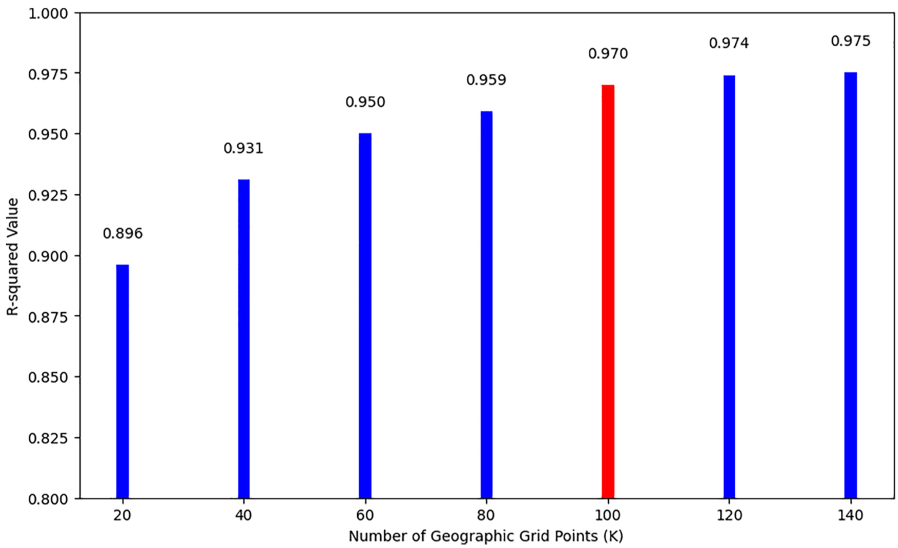
Source(s): Author’s own creation

For both years, it was observed that finer geographic segmentation by grid dummies leads to higher explanatory power in the models. The R-squared value of the district-level model was 0.826, while that of the subdistrict-level model was 0.899. These results align well with the explanatory power reported in existing literature for traditional regression models (Kim and Kim, 2023). The inclusion of geolocation coordinates in model [4] substantially enhanced the model's explanatory power, with the R-squared value increasing significantly to 0.968 for 2023 and 0.970 for 2024. Moreover, all parameter estimates in model [4] were found to be statistically significant.

To assess the sensitivity of model performance to the number of geographic grid points (K), we examined the changes in R-squared values as K varied. Figure 2 illustrates the relationship between explanatory power and the number of grid points in 2024. When K = 20, the explanatory power was comparable to that of the subdistrict-based model. As K increased, corresponding to finer spatial segmentation, the R-squared value steadily improved, reflecting enhanced model performance. However, beyond approximately K = 120, no significant improvement was observed. Consequently, we adopted K = 100 for the subsequent analysis.

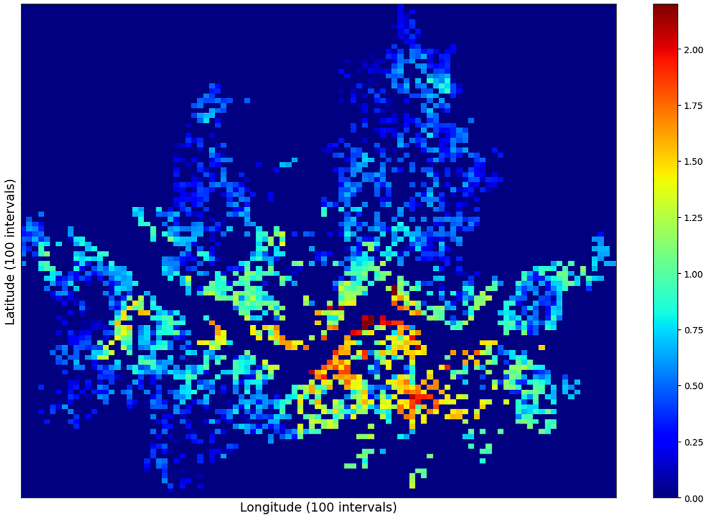
To assess how the high explanatory power of model [4] compares to that of ML models, we implemented an auxiliary random forest (RF) model (Rico-Juan and de La Paz, 2021) using exactly the same variables as the regression analysis, including geographic coordinates. For example, in the case of the 2024 data, the R-squared of the RF model was 0.987, only slightly surpassing that of model [4].

Figure 3 represents the coefficient estimates for geographic dummy variables in 2024 as a heatmap on a two-dimensional coordinate grid. The heatmap closely resembles the actual map of Seoul in Figure 1, with empty areas indicating mountain or riverside regions where residential areas are less common or where there were no apartment transactions in the year. The areas with a concentration of high-priced apartments, such as Gangnam district and regions near the Han River, exhibit significantly elevated coefficient values. This suggests that apartments in these regions enjoy extra locational advantages in addition to those attributable to their physical attributes.



Source(s): Author's own creation

Figure 2. Changes in *R*-squared values of model [4] by the number of grid points

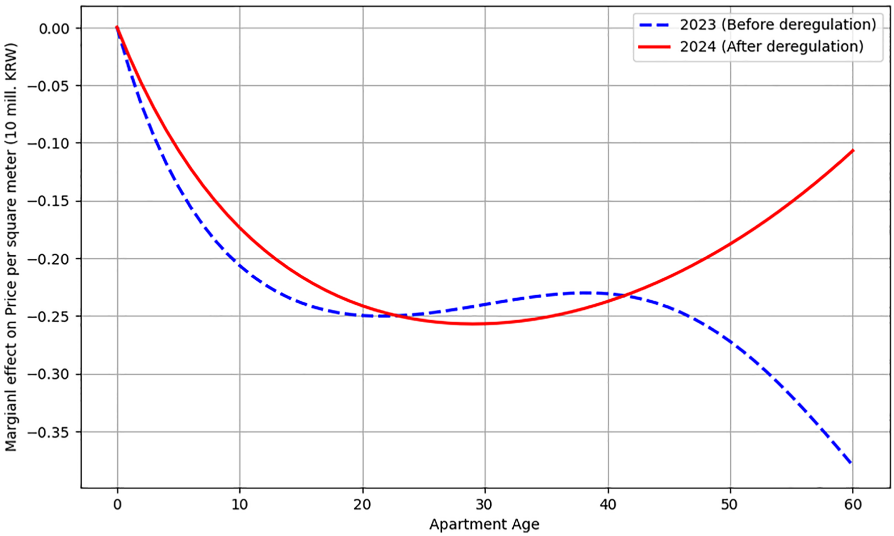


Source(s): Author's own creation

Figure 3. Heatmap of geographic grid dummy estimates for 2024

5. Policy effect

We now analyze the effects of the apartment redevelopment deregulation policy introduced earlier. The most notable difference between the estimation results for the two years, as shown in Tables 2 and 3, lies in the parameter associated with apartment age. To illustrate this difference, Figure 4 depicts how apartment age impacts prices in each year, based on polynomial calculations using the respective model parameters.



Source(s): Author's own creation

Figure 4. Marginal effect of apartment age by year ("grid" model)

In 2023, the effect exhibits a modest increase beyond the 30-year mark, followed by a sharp decline after 40 years. By contrast, the 2024 results show a rapid rebound after reaching a trough at approximately 30 years. This divergence underscores a substantial shift in market sentiment, indicating that expectations regarding redevelopment became significantly more optimistic following the announcement of the deregulation policy in January 2024 [3].

To examine the effects of deregulation from a different perspective, we designated a hypothetical "test" apartment and used regression models to predict its price under various aging scenarios. The predicted prices, initially in log scale, were converted back to the original scale (10 million KRW). The test apartment is assumed to be located in Grid #1004, with a size of 100 square meters, a floor level of 10, a brand dummy value of 1, and a transaction date of September 2024.

We constructed two predictions. For price prediction (1), we applied the parameters from the 2024 estimates of model [4] in Tables 3. For price prediction (2), the age parameters were replaced with the 2023 estimates in Table 2 to enable a counterfactual scenario with no deregulation in 2024.

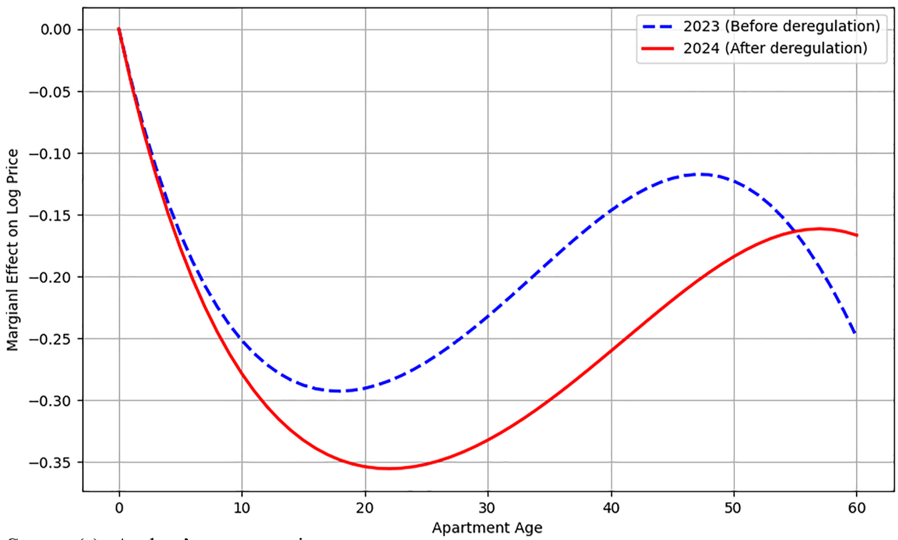
Table 4 presents the predicted prices and their differences for the test apartment. Obviously, the increase in housing values resulting from deregulation is predicted to be concentrated primarily in very old apartments, specifically those aged around 45 years or more. For instance, the price of a 60-year-old apartment is projected to rise substantially by as much as 1.74 billion KRW after the deregulation.

Had we used the subdistrict-level model [3] instead of the grid-level model [4], we would not have been able to observe the clear effects of deregulation. Figure 5 illustrates the predicted

Table 4. Price prediction of a test apartment in billion KRW

Age	Prediction: A	Prediction: B	A-B	Age	Prediction: A	Prediction: B	A-B
10	3.86	3.67	0.19	45	3.61	3.45	0.16
20	3.45	3.40	0.05	50	3.78	3.26	0.52
30	3.36	3.46	−0.10	55	3.99	2.94	1.05
40	3.48	3.52	−0.04	60	4.24	2.50	1.74

Source(s): Author's own creation



Source(s): Author's own creation

Figure 5. Marginal effect of apartment age by year ("subdistrict" model)

effects of age from the subdistrict-level model [3], but unlike Figure 4, it shows no discernible differences in patterns between 2023 and 2024, before and after deregulation. In addition, the timing of the housing price rebound driven by redevelopment expectations appears differently in both years compared to model [4]. This disparity can undoubtedly be attributed to the subdistrict-level model's limited explanatory power, leading to misestimation of the marginal effects of apartment age. These findings further underscore the value of the grid-level segmentation proposed in this study.

Beyond the price effects of the regulation, several socioeconomic issues arise. While the deregulation policy appears to have increased the asset value of older apartments by fostering redevelopment expectations, it also raises broader concerns about affordability and equity. Specifically, the rapid price appreciation of redeveloped apartments may make these properties less accessible to first-time homebuyers or middle-income families, potentially exacerbating affordability issues in urban housing markets. Furthermore, the uneven distribution of redevelopment activity across regions could raise equity concerns. For example, redevelopment is more likely to occur in affluent neighborhoods with well-maintained infrastructure, while economically disadvantaged regions may face slower progress. Such disparities could deepen the existing socioeconomic divide and undermine the broader goal of equitable housing access.

6. Conclusion

This study demonstrates the significant improvement in explanatory power achieved by incorporating geolocation information into traditional regression models for housing price analysis. By leveraging precise latitude and longitude coordinates of property location, we developed a grid-level segmentation approach that enhances model flexibility without compromising interpretability. Our results show that models incorporating detailed location data offer substantial improvements in explanatory power compared to more traditional approaches such as district-level models. Furthermore, the performance of the proposed model was found to be comparable to that of machine learning (ML) models. This indicates that, while ML models may provide slightly superior fit, the regression model with geocoded segmentation offers a more interpretable alternative with comparable performance at least for the study of property values.

A key finding of the study is the significant impact of location information on the estimation of the marginal effects of target variables. By accounting for location-specific heterogeneity in its full magnitude, our model provides more accurate and reliable estimates for policy evaluation. As a case study, this paper examines the effect of deregulation on apartment redevelopment in Korea, demonstrating how the proposed methodology can be applied to real-world policy analysis. The findings suggest that the deregulation policy has had a significant impact on housing prices, particularly in older apartments, aligning with intuitive predictions. Had we only used district-level information, the true impact would likely have gone undetected.

There are several limitations to this study. While the model shows promise in the context of the Korean market, its applicability to other regions requires careful consideration. For example, the availability and quality of geolocation data may differ significantly across property markets, which could impact the model's accuracy and generalizability. Moreover, local market conditions, such as regulatory frameworks, cultural differences, and economic factors, may influence the relevance of the model's assumptions and results. Therefore, future research may focus on refining the model, particularly to account for interaction effects not addressed in this study, and exploring its applicability to other property markets where geolocation data is readily accessible.

Notes

1. The data is available at https://data.seoul.go.kr/dataList/OA-21275/S/1/datasetView.do. All the estimations in this paper were coded in Python. Links to view the codes can be provided upon request.
2. All estimates are statistically significant. The estimates for the monthly dummy variables are omitted from the tables for brevity.

3. There was a slight increase in the values of low-age apartments in 2024 compared to 2023. This is probably due to a decrease in the supply of such apartments in 2024, driven by a housing market downturn attributable to high interest rates.

References

Anselin, L. and Lozano-Gracia, N. (2009), "Spatial hedonic models", in *Palgrave Handbook of Econometrics: Volume 2: Applied Econometrics*, Palgrave Macmillan UK, London, pp. 1213-1250.

Athey, S. and Imbens, G.W. (2019), "Machine learning methods that economists should know about", *Annual Review of Economics*, Vol. 11 No. 1, pp. 685-725, doi: 10.1146/annurev-economics-080217-053433.

Breiman, L. (2001), "Statistical modeling: the two cultures (with comments and a rejoinder by the author)", *Statistical Science*, Vol. 16 No. 3, pp. 199-231, doi: 10.1214/ss/1009213726.

Chau, K.W. and Chin, T.L. (2003), "A critical review of literature on the hedonic price model", *International Journal for Housing Science and its Applications*, Vol. 27 No. 2, pp. 145-165.

Copiello, S. (2020), "Spatial dependence of housing values in Northeastern Italy", *Cities*, Vol. 96, 102444, doi: 10.1016/j.cities.2019.102444.

Ho, W.K., Tang, B.S. and Wong, S.W. (2021), "Predicting property prices with machine learning algorithms", *Journal of Property Research*, Vol. 38 No. 1, pp. 48-70, doi: 10.1080/09599916.2020.1832558.

Kim, J. and Kim, K.M. (2023), "A study on the machine learning-based apartment price index", *Journal of Korea Planning Association*, Vol. 58 No. 4, pp. 160-177, doi: 10.17208/jkpa.2023.08.58.4.160.

Lundberg, S. (2017), "A unified approach to interpreting model predictions", *arXiv preprint arXiv: 1705.07874*.

Rico-Juan, J.R. and de La Paz, P.T. (2021), "Machine learning with explainability or spatial hedonics tools? An analysis of the asking prices in the housing market in Alicante, Spain", *Expert Systems with Applications*, Vol. 171, p. 1145.

Soltani, A., Heydari, M., Aghaei, F. and Pettit, C.J. (2022), "Housing price prediction incorporating spatio-temporal dependency into machine learning algorithms", *Cities*, Vol. 131, 103941, doi: 10.1016/j.cities.2022.103941.

Tchuente, D. and Nyawa, S. (2022), "Real estate price estimation in French cities using geocoding and machine learning", *Annals of Operations Research*, Vol. 308 No. 1, pp. 571-608, doi: 10.1007/s10479-021-03932-5.

Zaki, J., Nayyar, A., Dalal, S. and Ali, Z.H. (2022), "House price prediction using hedonic pricing model and machine learning techniques", *Concurrency and Computation: Practice and Experience*, Vol. 34 No. 27, e7342, doi: 10.1002/cpe.7342.

Zhang, Y., Rahman, A. and Miller, E. (2023), "Longitudinal modelling of housing prices with machine learning and temporal regression", *International Journal of Housing Markets and Analysis*, Vol. 16 No. 4, pp. 693-715, doi: 10.1108/ijhma-02-2022-0033.

Corresponding author
Hongjai Rhee can be contacted at: hrhee@ajou.ac.kr