Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Spatial mechanisms of tree-based models in housing price mass appraisal

Jengei Hong
Handong Global University
Doojin Ryu
sharpjin@skku.edu
Sungkyunkwan University

Research Article

Keywords: Automatic valuation model, Housing prices, Mass appraisal, Machine learning, Random forest, XGBoost

Posted Date: May 28th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8871515/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Spatial mechanisms of tree-based models in housing price mass appraisal

Jengei Hong[1], Doojin Ryu[2,*]
[1] School of Management and Economics, Handong Global University, Pohang, Republic of Korea
[2] Department of Economics, Sungkyunkwan University, Seoul, Republic of Korea
*Corresponding Author. Email: sharpjin@skku.edu
https://sites.google.com/view/ryudoojin/home/research-international

ORCIDs
Jengei Hong (hwgh024@handong.edu): https://orcid.org/0000-0002-0591-3285
Doojin Ryu (doojin.ryu@gmail.com): https://orcid.org/0000-0002-0059-4887

Highlights

·Tree-based models capture housing price patterns by partitioning geographic space without assuming functional relationships.
·Geographic coordinates provide most of the information needed for accurate mass appraisal.
·The same spatial mechanism also introduces risks in extrapolation and interpretation.

Abstract

Tree-based ensemble models are increasingly used in housing price mass appraisal, yet their mechanisms for capturing locational complexity remain insufficiently understood. We examine how key challenges in the analysis of urban housing prices, particularly those arising from spatial heterogeneity, are addressed by these models. We conduct a twofold quantitative analysis. First, simulation experiments isolate locational effects from confounding real-world factors and demonstrate that these models capture nonlinear price surfaces, omitted locational factors, and latent housing submarkets through coordinate-based spatial partitioning. Second, analyses of a comprehensive dataset covering all officially recorded multi-family housing transactions in Seoul evaluate model performance in a realistic mass appraisal setting. Models incorporating geographic coordinates achieve consistently high empirical accuracy, while adding conventional locational variables yields little incremental benefit. Our findings suggest that the effectiveness of tree-based models in urban housing price analysis stems primarily from their ability to internalize spatial heterogeneity via partitioning, while also highlighting important limitations related to extrapolation and structural interpretation.

Keywords: Automatic valuation model; Housing prices; Mass appraisal; Machine learning; Random forest; XGBoost
JEL Classification: C53, C63, R30, R32

1. Introduction

Large-scale housing price assessment has become a central task in urban governance. In this context, ensemble decision tree models—most notably Random Forest and XGBoost—are increasingly adopted in academic research, taxation, development planning, and housing

market management (Pita, de Carvalho, Barbosa, Cury & Mendes, 2026). Although their superior predictive performance relative to traditional linear regression–based models is now well documented (Antipov & Pokryshevskaya, 2012; Pace & Hayunga, 2020; Clapp & Lindenthal, 2022), their growing use in practice has outpaced our understanding of the mechanisms through which the effectiveness of tree-based models arises and the conditions under which it can be reliably sustained (Shaamala, Yigitcanlar, Nili, & Nyandega, 2025). As these models begin to inform decisions with direct policy and financial outcomes, understanding not only whether they perform well, but why they do so and where their limitations lie, becomes essential. Without a clear understanding of the mechanisms underlying their predictive success, the growing reliance on tree-based models risks overextension, distorted interpretation, and unintended policy bias—particularly in contexts where fairness, transparency, and stability are central concerns.

Against this background, this study investigates how tree-based models address key challenges in housing price estimation. In particular, we show that the effectiveness of tree-based models stems from their ability to internalize locational complexity through recursive spatial partitioning. A fundamental difficulty in cross-sectional housing price analysis lies in the complexity of locational effects. This complexity arises from various factors such as abrupt spatial discontinuities, locally varying nonlinear relationships, unobserved neighborhood attributes, and latent housing submarkets that are difficult to represent using smooth, globally specified functional forms. Traditional models struggle in such environments because they rely on explicit assumptions about functional relationships and predefined locational variables. Instead, tree-based models rely on coordinate-based spatial partitioning, which can flexibly absorb locally embedded price patterns without imposing strong structural assumptions.

Our contribution is to jointly characterize the source and the boundaries of the effectiveness of tree-based models in housing price estimation. Moving beyond performance comparisons, we provide an analytical perspective on how these models capture locational complexity, when this mechanism is effective, and when it becomes fragile or misleading. To this end, we adopt a twofold empirical strategy. First, we conduct controlled simulation experiments designed to isolate key dimensions of locational complexity, including unpredictable nonlinear price surfaces, omitted spatial variables, and latent housing submarkets. These experiments demonstrate that tree-based models can effectively capture spatial price variation through coordinate-based spatial partitioning. At the same time, the simulations reveal conditions under which this mechanism becomes unstable or fails to recover meaningful spatial relationships, particularly when predictions rely on extrapolation beyond the support of the training data. Second, we examine whether the mechanisms identified in the simulation experiments are also present in real-world housing markets. Using a comprehensive dataset covering all officially recorded multi-family housing transactions in Seoul from January 2011 to January 2024, comprising over 1.2 million observations, we evaluate the performance of tree-based models in a realistic mass appraisal setting. Specifically, we compare models that share the same set of structural housing characteristics but differ in how locational information is incorporated: (i) models that include only geographic coordinates, (ii) models that include both coordinates and a rich set of traditional locational variables, and (iii) models that include traditional locational variables while excluding geographic coordinates.

Our empirical results yield two main findings. First, tree-based models that incorporate geographic coordinates achieve consistently high predictive accuracy over time, with Mean Absolute Percentage Error below 10% in most cases. Given the inherent transaction-level noise in housing prices, this level of accuracy suggests that tree-based models are able to capture most of the relevant complexity arising from locational effects in real-world housing markets. Second, augmenting coordinate-based models with additional locational variables provides little incremental predictive benefit, while excluding geographic coordinates—despite the inclusion of rich locational variables—leads to a noticeable deterioration in performance. In addition, drawing on the real-world analysis, we document a distinct pattern that aligns with the simulation results. As near-term forecasting necessarily involves extrapolation beyond historically observed price patterns, predictive bias can arise when forecasts extend beyond the support of the training data, underscoring inherent limitations of tree-based models.

Our findings have several implications for the application and interpretation of tree-based models in housing price analysis. Spatial partitioning based on geographic coordinates allows tree-based models to internalize much of the spatial complexity inherent in housing markets, rendering many traditional locational variables informationally redundant.[1] From a practical perspective, these findings suggest that tree-based models are particularly well suited for applications in which predictive accuracy is the primary objective, rather than structural interpretation. The key advantage is that these models provide a parsimonious yet highly effective approach to housing price prediction, without requiring the construction and maintenance of extensive sets of locational variables.

This study also calls for caution in the growing application of simulation-based XAI methods to tree-based models in housing price analysis. In housing markets, where prices are strongly governed by locational heterogeneity, simulation-based predictions used in XAI analyses are prone to extrapolation beyond the support of the training data. As a result, interpretative conclusions drawn from such XAI analyses may be unstable or misleading, even when the underlying model achieves high predictive accuracy. Together, these results emphasize that while tree-based models are powerful predictive tools, their outputs should not be interpreted as structural representations of housing markets.

The remainder is organized as follows. Section 2 establishes a conceptual contrast between traditional analytical models and tree-based models, clarifying why locational complexity poses a fundamental challenge for the former but can be absorbed by the latter through spatial partitioning. Section 3 uses controlled simulation experiments to isolate key sources of locational complexity and to identify the conditions under which spatial partitioning effectively captures these features—or breaks down. Section 4 then demonstrates that both the effectiveness and the limitations are not artifacts of simulation by validating them in a real-world mass appraisal setting using large-scale transaction data. Section 5 concludes by summarizing the main findings and discussing their methodological implications for prediction, interpretation, and structural inference in housing price analysis.

---

[1] The inclusion of such variables can distort interpretative analyses, such as feature importance assessments, by arbitrarily redistributing explanatory weight across correlated locational inputs.

2. Key Feature of Tree-Based Models

In the analysis of housing prices, traditional empirical models have been designed to explicitly investigate the endogenous relationships among variables. The most recognized model, the hedonic pricing model, decomposes a house's value into the values of its individual characteristics. According to the consumer theory of Lancaster (1966), a product is viewed as a collection of distinct attributes. Rosen (1974) introduces the hedonic pricing model based on the assumption that the market value of a good reflects the value of all its utility-bearing attributes. Under this model, a house's value is seen as a composite of utilities from various intrinsic features, such as location, building structure, and environmental qualities. If the values of houses and their relevant features are accurately represented by transaction prices and observable variables, the hedonic pricing theory can provide an equation that describes the relationship between housing prices and these variables, making it suitable for regression analysis (Taylor, 2008).

Models developed within this tradition—whose primary objective is to specify the endogenous structure between a target variable and its explanatory variables—can be characterized as analytical models. These models simplify reality, facilitating the interpretation of complex and non-intuitive data through specified functional relationships. An essential aspect of the hedonic model is that it only considers variables that align with its functional form (Sheppard, 1999). For example, although the location of a house significantly affects its value, the model cannot directly use geographical coordinates (latitude and longitude) because their functional relationship cannot be expressed in a regression-based model. The hedonic theory deduces the relationship between housing prices and desirable location attributes, such as reduced commuting costs or proximity to urban amenities. Researchers measure these attributes through observable variables, such as the distance to the central business district (CBD) or nearby parks.

The hedonic model is widely used not only in academic research but also in practical contexts such as housing valuation (Malpezzi, 2003). For purposes such as property taxation, loan-to-value (LTV) assessment (Berry & Bednarz, 1975; Pace & Gilley, 1990; Kang, Zhang, Peng, Gao, Rao, Duarte, & Ratti, 2021; Kim, Cho, & Ryu, 2018), construction project (Ugur, Kanit, Erdal, Namli, Erdal, Baykan & Erdal, M, 2019), and the calculation of housing price indices (Hill, 2013), accurate estimation of housing values is crucial because individual market values are not directly observable. Unlike financial assets or consumer goods, housing is highly heterogeneous, as each property occupies a unique location, making simple price comparison difficult. Thus, mass appraisal—evaluating the value of houses with unobservable prices using available data—holds significant academic and practical importance.

However, the hedonic model often lacks high predictive accuracy for individual property prices as it is basically a simplification that relies on a set of arbitrary assumptions to approximate complex market realities (Ekeland, Heckman, & Nesheim, 2004). Especially, the challenges in capturing the complexities from locational effects are crucial. There are several reasons. First, there may be discrepancies between the actual changes in factors that affect prices and the observable changes in real-world data. For example, according to the model of Alonso (1964), house values are inversely proportional to commuting time, indicating a

potential negative relationship between house prices and distance from the CBD. However, this distance is often used as a simplified proxy for commuting time and may not fully capture the complexity of actual variations in commuting costs, which can be influenced by the specifics of commuters' travel routes or the transportation system. Additionally, actual prices may reflect varying willingness to pay among diverse market participants, making it difficult to capture the full complexity of locational effects using a theoretical pattern alone.

Second, several factors that can affect housing prices are either challenging to measure accurately in the mass appraisal or are not yet theoretically identified. For example, choices related to neighborhoods (Quigley, 1985; Durlauf, 2004; Hardman & Ioannides, 2004; Tita, Petras, & Greenbaum, 2006; Ioannides & Zabel, 2008), government policy (Baum-Snow & Marion, 2009; Chen, Glaeser, & Wessel, 2023; Ding & Itoh, 2023), educational factors (Black & Machin, 2011; Feng & Lu, 2013; Avery & Pathak, 2021; Wang & Li, 2022), pollution (Bajari, Fruehwirth, Kim, & Timmins, 2012; Zhang, Chen, & Wang, 2021; Ou, Zheng, & Nam, 2022; Hansson, 2024), and racial or cultural differences (Kain & Quigley, 1972; Follain & Malpezzi, 1981; Anas, 1990; Sheppard, 2010) can theoretically or intuitively influence house prices. However, identifying appropriate variables to quantitatively measure these factors often proves difficult in the practice of real-world mass appraisal, leading to their exclusion from the model.

Third, the existence of housing submarkets. Submarkets are groups of houses that are substitutable within the same group but relatively less so with houses in other submarkets. This implies that the endogenous relationships between housing price determinants and prices can vary across submarkets. While housing submarkets are generally known to be closely associated with location (Adair, Berry, & McGreal, 1996; Bourassa, Hamelink, Hoesli, & MacGregor, 1999; Bourassa, Hoesli, & Peng, 2003), in the real world, submarkets driven by locational factors and those driven by housing characteristics are often intricately intertwined (Goodman & Thibodeau, 2007). Hence, it is almost impossible to fully identify the structure of submarkets when defining a traditional model.

The growing interest in applying machine learning models to housing price analysis stems from the need for more accurate and nuanced models that can better capture the complexity of the real estate market. With significant advancements in computing power, a variety of machine-learning models have been rapidly adopted for housing price assessments. Among these, decision tree ensembles (tree-based models) have become one of the most frequently used approaches (Mullainathan & Spiess, 2017; Yilmazer & Kocaman, 2020; Rico-Juan & de La Paz, 2021). Their increasing popularity is primarily because of their high predictive accuracy. Recent studies have emphasized the superior predictive performance of tree-based models compared with that of traditional analytical models, thereby expanding the discussion on their potential applications. For example, Mullainathan and Spiess (2017) compare the predictive performance of Ordinary Least Squares (OLS) and RF using a dataset of house prices from the 2011 American Housing Survey. Their analysis shows that OLS, which only accounts for main effects without interactions, achieves an R-squared of 41.7% on the hold-out sample. In contrast, the RF model, which automatically captures complex interactions and non-linearities, attains a higher R-squared of 45.5%, outperforming OLS. The study further highlights that while OLS provides a simpler and more interpretable model, it fails to capture the intricate patterns present in the data, leading to inferior predictions compared to RF.

Yilmazer and Kocaman (2020) provide additional evidence by comparing traditional multiple regression analysis (MRA) with RF for mass appraisal in a complex urban area in Ankara, Turkey. Their study reveals that the RF model surpassed the MRA in terms of predictive accuracy, with an adjusted R-squared value of 0.734 compared with 0.696 for MRA. This emphasizes the benefits of machine learning approaches in managing nonlinear interactions and automatically selecting relevant variables for property valuation. Pace and Hayunga (2020) support these findings by comparing traditional hedonic models with tree-based methods, such as Classification And Regression Tree (CART), boosting, and bagging. Their research shows that bagging, in particular, significantly reduces prediction errors and outperforms both global and local spatiotemporal models, underlining the potential of machine learning algorithms to enhance real estate price assessments.

The primary distinction between these tree-based models and traditional models is that they are not intended for analytical purposes. They are constructed to perform specific functional tasks, enabling us to categorize them conceptually as functional models. To understand the difference between analytical and functional models, consider a simple example involving data on house prices and their characteristics. If our goal is to identify the best linear combination that explains the relationship between the price and these characteristics (i.e., using linear regression to measure the relationships between variables), we will use an analytical model. If this model is accurate, it allows us to form an intuitive understanding of the relationships between variables. Furthermore, based on this understanding (through the measured linear function), we can perform the functional task of predicting the expected prices of houses for which the market prices are not directly observable. Alternatively, to predict the prices of these houses without this analytical process, we might assign the average price of the three closest houses with known market prices. This approach serves as an example of a functional model. Although it does not provide an intuitive understanding of the relationships between variables, it still fulfills the purpose of predicting the expected prices for houses with unobserved market prices.

Tree-based models fundamentally align with the functional model approach, as they emphasize task performance, such as prediction, rather than focusing on explaining the relationships between variables. To understand tree-based models, it is essential to first comprehend the nature of a decision tree. A decision tree is an algorithm that establishes conditions to separate data into groups that are as homogeneous as possible, based on specified criteria. This algorithm operates similarly to a branching tree, continuously creating conditions that further divide the data as examples in Figure 1. The number of hierarchical conditions generated in the tree is referred to as its depth, with each condition represented as a node. Ultimately, the samples classified according to the hierarchical condition structure and used for prediction are referred to as leaves. The level of detail and the number of conditions generated are determined by the researcher through the specification of hyperparameters.

[Figure 1 here]

A decision tree generates nodes until the classified sample groups can no longer be further divided under the given conditions. This means that the conditions are designed to either

minimize the variance within the classified data or maximize the purity of the information, ensuring that each resulting group is as homogeneous as possible. A decision tree is like narrowing down answers in a game of 20 questions. In this game, the objective is to reach the correct answer with as few questions as possible by identifying a group that cannot be subdivided further. Similarly, a decision tree algorithm seeks to identify the conditions that best explain the data. For instance, in a 20-question game aimed at guessing an animal, the questions might begin broadly by asking if the animal is a mammal and then progress to more specific inquiries, such as whether it is a lion.

In housing price assessments, a decision tree constructs a framework of conditions that classify groups with similar prices. This decision tree examines data containing both housing prices and their attributes, creating a hierarchy of conditions that separate the data into subsets where price differences are no longer explained by attribute differences. When predicting a house's price, the decision tree uses these established conditions to identify a group of samples in the training data that most closely match the specified attributes. The predicted price is then determined by calculating either the mean or median of that group.

As explained, tree-based models do not assume a predefined functional relationship between prices and variables. Instead of analyzing or simplifying the endogenous structure of locational attributes, these models use information as inputs to identify price differences; they do not measure the separable effect of each variable on price. Consequently, the analytical use of decision trees is inevitably subject to several limitations. This is not because the structure of a tree model is more complex than our intuition, but rather because they are not inherently designed to perform such analytical tasks.

Nevertheless, the flexibility of tree-based models makes them particularly effective at capturing the complexities arising from locational effects. Unlike traditional models, tree-based models are not restricted to using only variables explicitly known to affect housing prices directly. Since tree-based models adapt flexibly to the data's structure without any prior assumptions about the nature of these relationships, they can use any information that captures price differences resulting from various attributes, even if those attributes are not directly observable. This means that tree-based models directly use spatial coordinates (latitude and longitude) as variables to capture variations in housing prices.

In tree algorithms, locational information serves as a partitioning criterion to divide the overall sample population. The specific points where splits occur, and the number of splits, are determined during the optimization process, tailored to the training data. Figure 2 illustrates examples of how a single decision makes predictions using only locational information in the simplest cases. The blue solid line represents an arbitrary one-dimensional function defined between 0 and 1. The red cross-dotted line (left), the magenta round-dotted line (middle), and the orange star-dotted line (right) depict the predictions made by the decision trees at depths of 1, 2, and 3, respectively. The decision tree uses only the domain x as the variable. As shown in the figure, when the condition is created, split points are generated to divide the domain, and predictions are made based on these splits. As the depth of the tree increases, additional and finer split points are created, leading to more detailed predictions.

[Figure 2 here]

Location is not merely one of many attributes but the fundamental source of all locational characteristics. Essentially, all locational attributes are functions of the location itself. Therefore, capturing price differences based on location inherently accounts for the effects of all locational attributes (including unobserved attributes) simultaneously. Furthermore, when analyzing the price determination structure of housing markets, various other variables, in addition to locational variables, are typically considered. As a result, much more intricate hierarchical structures of conditions are generated in a tree. This highlights that tree-based algorithms can effectively capture the complex and often unpredictable structure of housing submarkets. The reason housing submarkets undermine the explanatory power of traditional models lies in the fact that the price determination structure in one submarket may differ significantly from that in another. When submarkets exist, failing to identify the criteria that separate them in advance can reduce the accuracy of estimated coefficients. In contrast, tree models automatically identify and generate conditions to segment submarkets during the process of dividing samples. In doing so, they can generate different conditions within different subsets, allowing them to flexibly capture the unique price determination structures across various locations.

Another practical advantage of tree-based models is that they require less information while still performing effectively. As previously mentioned, defining locational attributes is complex and subjective, and constructing locational variables in analytical models usually requires more than just the property coordinates. For instance, to measure distance variables, an analyst needs to know the coordinates of each property and the locations of various facilities. However, obtaining sufficient locational data for these various facilities is often challenging. Therefore, when the goal is to make accurate assessments, the tree-based approach is not only effective but also more convenient.

In most cases of housing price analysis based on the decision tree algorithm, the approach involves generating multiple decision trees and aggregating their numerical outputs, rather than relying on a single decision tree. By doing so, the discreteness inherent in the predictions of individual decision trees is mitigated and smoothed, effectively addressing potential overspecification issues and enabling much more accurate and stable predictions. These approaches are commonly known as ensemble techniques, which are broadly categorized into two types: bagging (bootstrap aggregating) and boosting. Bagging trains multiple models of the same type on different subsets of the training data, generated through random sampling with replacement, as shown in Figure 3. The final prediction is usually an average of the predictions made by the multiple trees. A well-known model employing this technique is the RF. This method allows each decision tree to learn different patterns in the data, thus reducing the risk of any single tree overfitting to the noise in a specific sample. While a single decision tree may be overly sensitive to minor noise in the data, combining the outputs of multiple independent trees through averaging generally produces more reliable and stable predictions because the ensemble is more resilient to the effects of such noise.

[Figure 3 here]

Numerous studies have shown that RF performs well in predicting housing. For instance, Antipov and Pokryshevskaya (2012) find that RF outperforms other models in predicting real estate prices in Saint Petersburg, Russia, using a dataset of 2,848 apartment records. Similarly, Adetunji, Akande, Ajala, Oyewo, Akande, and Oluwadara (2022) confirm the strong predictive power of RF in their analysis of 1,978 housing prices in Boston, emphasizing the critical role of independent variable preprocessing in enhancing model performance. Ho, Tang, and Wong (2021) demonstrate that both RF and gradient-boosting machine models provided the most accurate predictions in their study of 40,000 housing transactions in Hong Kong over 18 years, surpassing support vector machines in terms of both accuracy and error rates.

Boosting sequentially trains new decision trees to correct the errors made by previously trained trees, unlike bagging, in which trees are created independently through random sampling. In this process, each tree addresses the errors of its predecessors, gradually improving predictive accuracy by reducing individual tree sensitivities. Prominent examples of boosting models are XGBoost, LightGBM, and CatBoost. Several studies have used boosting techniques for housing price analysis, including the works of Ming, Zhang, Qi, Liao, Wang, and Zhang (2020), Hong and Kim (2022), and Jafary, Shojaei, Rajabifard, and Ngo (2024). These studies highlight the predictive power of tree-based models in predicting real estate prices.

Although most studies employing tree-based models have focused primarily on their predictive accuracy for individual housing prices, more recently, some studies have shifted their attention to providing an analytic understanding of market structures (Dou, Gu, & Fan, 2023; Krämer, Nagl, Stang, & Schäfers, 2023; Mamre & Sommervoll, 2024; Zeng & Rao, 2025). With advancements in techniques for conducting analytical tasks using trained machine learning models, such as Explainable AI, an increasing number of studies are expected to focus on leveraging the interpretability of tree-based models alongside their high predictive performance. Explainable AI refers to the methods that interpret and explain the predictions of machine learning models. These techniques simulate model behavior to assess feature importance and relationships between variables. For example, the method of SHapley Additive exPlanations (SHAP) simulates all possible feature combinations to determine their contribution to the model's predictions, and the method of Local Interpretable Model-agnostic Explanations (LIME) generates small input variations to fit simpler, interpretable models that approximate complex ones. Similarly, partial dependence plots (PDPs) and individual conditional expectation (ICE) plots simulate changes in one or more features to assess their effect on predictions.

Nevertheless, considerable caution is required when results obtained from simulations of tree-based models are extended to infer endogenous relationships among attributes. This caution stems from the inherent limitations of tree-based models in generating predictions for observations that fall outside the support of the training data. Unlike traditional econometric models, which extrapolate relationships among attributes through explicitly specified parametric structures, tree-based models generate predictions by aggregating patterns observed within the training sample.

As discussed in greater detail in subsequent sections, when applied to observations beyond the empirical range of the data, tree-based models tend to return predictions based on the

nearest available data points, rather than systematically adjusting for differences in attribute values. This characteristic becomes particularly relevant in cross-sectional housing price analyses, where observations are dispersed across a wide and heterogeneous set of locations. In such contexts, such predictions often involve a form of local extrapolation driven by limited nearby observations. When simulation-based interpretability methods are used to assess the effects of individual variables, these limitations become especially consequential. Many XAI techniques rely heavily on simulations of the trained model, and thus any instability or bias embedded in the model's predictions could be directly transmitted to the interpretation results, increasing the risk of misleading inferences.

These characteristics highlight the strengths and limitations of tree-based models relative to traditional methods. A significant advantage of tree-based models is their ability to capture the complexities from locational effects such as unpredictable non-linearity, omitted variables, and submarkets, without the need for explicit modeling. However, there are several direct and indirect limitations to using tree-based models for analytical purposes. Understanding these characteristics is essential for constructing a reliable model for housing price analysis. The following section offers a quantitative analysis of the capacity of tree-based models, using simulation approaches.

3. Simulation Analyses

The previous section discusses the features of tree-based models, focusing on their ability to capture the complexities of locational effects. This section examines the effectiveness of tree-based models in addressing such complexities through simulation-based analysis. The motivation for employing this simulation-based approach lies in its ability to clearly define the analytical focus. Real-world housing price data often include a wide range of attributes, potential observation errors, and collection biases, which make it challenging to isolate and analyze specific factors of interest. However, this simulation approach enables a direct comparison between predicted and true functions, providing a clearer evaluation of how effectively tree-based models handle the complexities considered in each scenario.

To begin, we examine the simplest cases in a one-dimensional space, making the prediction results visually easy to understand. In the baseline model, housing prices are assumed to be determined solely by locational effects. As discussed in the previous section, these locational effects are assumed to consist of components that can be expressed through specific functional forms and observable variables, as well as components that cannot be explicitly expressed in such a way. We use arbitrary polynomial functions as examples of nonlinear and continuous target functions. These functions are defined over the closed interval (0,1], representing the housing locations. The functions are expressed as follows:

$$p_i = \alpha_{1,s}|0.5 - x_i| + \alpha_{2,s}{x_i}^3 + \alpha_{3,s}{x_i}^2 + \alpha_{4,s}x_i + \alpha_{5,s}, \quad (1)$$

where p_i and x_i represent the housing price and location of property i, respectively. The first term on the right-hand side of the equation represents the theoretically known and measurable effects of locational attributes. We assume the midpoint of the domain as the location of urban amenities or the CBD, and the distance from this midpoint is defined as a measurable locational

attribute. In reality, the relationship between these locational attributes and housing prices is often nonlinear. However, in the simulation, we consider the simplest case to ensure that the regression-based analytical model captures these effects. The remaining terms represent complexities of the real world that cannot be observed directly. This represents the various factors discussed in the previous section—such as neighborhood effects, educational environments, or cultural amenities—that are difficult to quantify, not directly observable, or not yet fully understood, and thus capture the full range of locational complexities. These complexities are modeled as an arbitrary cubic polynomial, capable of taking various nonlinear forms, depending on its coefficients. This equation serves as a simplified representation of situations that analysts might encounter in practice.

We illustrate the predictive behavior of linear regression models and a tree-based model in estimating the housing value p_i over a grid of the domain, using samples drawn from Equation (1). For the linear models, two specifications are considered. The first model includes only the observable locational effect (i.e., the distance from the midpoint, $|0.5 - x_i|$); the second model also incorporates locational dummy variables[2], $I_j(x_i)$. The formulations of these two models are presented below:

$$p_i = \beta_0 + \beta_1|0.5 - x_i| + \varepsilon_i, \text{ where } \varepsilon_i \sim N(\mu, \sigma^2), i.i.d., \tag{2}$$

$$p_i = \beta_0 + \beta_1|0.5 - x_i| + \sum \beta_j I_j(x_i) + \varepsilon_i, \text{ where } \varepsilon_i \sim N(\mu, \sigma^2), i.i.d. \tag{3}$$

For the tree-based model, we employ the RF algorithm, which is one of the most fundamental and widely used ensemble tree methods.[3] In the RF model, only the location x_i is used as an input feature. In numerical prediction, the RF constructs multiple decision trees via bootstrapping and generates predictions by averaging the outcomes from each tree. The model requires various hyperparameter settings, such as the size of subsamples, the number of decision trees, and the criteria for splitting the nodes in each tree. The settings used in this study are provided in Table 1.

[Table 1 here]

Because the specific form of the function $f(x)$ varies depending on the coefficient values $\alpha_{1,s}$, $\alpha_{2,s}$, $\alpha_{3,s}$, $\alpha_{4,s}$, and $\alpha_{5,s}$, we can evaluate the model's predictive power across various scenarios by randomly selecting coefficient sets. We assume that the coefficients $\alpha_{1,s}$, $\alpha_{2,s}$, $\alpha_{3,s}$,

2 For real-world data, zip code or administrative district information is usually used. Recently, Ryu, Hong, and Jo (2024) propose a model that directly generates locational dummy variables using latitude and longitude instead of administrative information.

3 The simulation results remain qualitatively unchanged when alternative tree-based models, such as XGBoost, are employed.

$\alpha_{4,s}$, and $\alpha_{5,s}$ follow the uniform distribution between –10 and 10.[4] The simulation process is as follows. First, we draw a set of coefficients to define an arbitrary nonlinear function. Second, we randomly sample n points from this function, which we designate as the training sample. Third, we train each model using this training sample. Fourth, we predict the function values over a grid defined on the function's domain and compare these predictions to the actual function values. We repeat this simulation N times.

Figure 4 shows examples from four different simulations. The qualitative characteristics observed in these results are consistent across all simulations. The figures illustrate the characteristics of linear regression models and tree-based models side by side. Linear regression models rely on variables whose relationships with the dependent variable are explicitly specified by the analyst; consequently, any complexity that falls outside these predefined structures appears as prediction error. The left-most figures display predictions from the distance-only linear model from equation (2), illustrating how nonlinear price patterns can be oversimplified under a linear interpretation. The middle figures display simulation results from a linear model that incorporates locational dummy variables. While this specification improves predictive power relative to the first model, it still fails to fully capture the nonlinearity of the true function. Moreover, in this setting, the distance variable may inadvertently capture nonlinear price effects driven by unobserved factors, reflecting a standard omitted-variable concern in regression analysis. By contrast, tree-based models do not impose a predefined functional relationship between prices and covariates. Instead, they exploit locational coordinates to distinguish price variation across space. Because all locational complexities—both observable and unobservable—ultimately operate through location itself, tree-based models can recover the nonlinear price shape that emerges from differences across locations driven by the aggregate of these effects. As shown in the figures on the right, the tree-based model provides a close approximation to the true function.

[Figure 4 here]

To quantitatively compare the predictive power of each model, we need to establish a quantitative metric. In this study, we use R-squared, a commonly used metric for assessing the similarity between predictions and actual outcomes.[5] *R*-squared measures how well the predicted values of a model explain the actual variability, with values ranging from 0 to 1. A value closer to 1 indicates that the model accounts for a larger portion of the actual variability. The definition of R-squared is as follows:

---

4 These parameters are selected for illustrative purposes in the quantitative simulations, and the implications of quantitative results remain consistent under alternative parameter choices.
5 Other metrics, such as the mean squared error (MSE) and MAPE, are also commonly considered. However, in the repeated simulations involving functions with arbitrary ranges, MSE and MAPE are not appropriate because they are sensitive to the absolute magnitude of function values and unit variations.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(p_i - \hat{p}_i)^2}{\sum_{i=1}^{n}(p_i - \bar{p}_i)^2}, \quad (4)$$

where $\hat{p}_i$ and $\bar{p}_i$ are the predicted value and the sample mean, respectively, of p_i.

Figure 5 shows the distribution of R-squared values obtained from 100 repeated simulations. In general, a model that achieves higher average accuracy across repeated function approximations and exhibits lower variance in results is considered to perform better. The figure reveals that the median R-squared values increase in the following order: the linear model with the locational factor, the linear model with the locational factor and locational dummy variables, and the RF model. Simultaneously, the variance decreases in the same sequence. This pattern indicates how effectively each model captures the nonlinearity of locational effects, aligning with the findings presented in Figure 4. Notably, the median R-squared value for the tree-based model is close to 1, suggesting that, with sufficient sample size, this model can closely approximate any arbitrary nonlinear continuous function defined on the given coordinates.

[Figure 5 here]

Does the addition of the distance variable $|0.5 - x_i|$ to the tree-based model, alongside the location coordinates x_i, significantly affect the prediction of the model? Figures 6 and 7 compare the results from simulations that use both the location coordinates and the distance variable with those that use only the location coordinates. In both scenarios, the R-squared values are close to 1, indicating that the two approximations are virtually identical. This implies that the location coordinates alone are sufficient for capturing the relevant nonlinear structure, and that adding the distance variable does not improve the model's predictive power. Therefore, incorporating specific locational attributes, such as distance measurements from particular locations, along with coordinates in tree-based models, is not recommended. When location coordinates are already included, adding variables derived from these coordinates introduces redundant information. This redundancy does not enhance predictive power but complicates the interpretation of the machine learning models in subsequent steps, such as deriving variable importance[6] or conducting simulations on synthetic or hypothetical cases.

[Figure 6 here]

[Figure 7 here]

For instance, Figure 8 compares the variable importance derived from models that use only location coordinates with those that include both the coordinates and the distance variable. The predictive performance of the two models is essentially identical, as demonstrated above,

6 Importance can be defined by calculating the proportion of total information contributed by each variable during the sample partitioning process and ranking them accordingly.

indicating that the inclusion of the distance variable does not enhance predictive power. However, we observe that the information attributed to the distance variable overlaps with that already captured by the location coordinates. This redundancy occurs because, in certain cases, the two variables convey nearly identical information. This finding suggests that the inclusion of multiple variables providing overlapping information complicates the interpretation of results. It provides guidance to researchers on which variables should or should not be included when designing a tree-based model—if location coordinates are included, other locational attributes may be unnecessary.

[Figure 8 here]

Subsequently, we examine a broader range of scenarios to verify the consistency of the results. Consider the case where noise is added to the target function. The actual function for housing values is expressed as follows:

$$p_i = \alpha_{1,s}|0.5 - x_i| + \alpha_{2,s}{x_i}^3 + \alpha_{3,s}{x_i}^2 + \alpha_{4,s}x_i + \alpha_{5,s} + \varepsilon_i, \text{ where } \varepsilon_i \sim N(\mu, \sigma^2), i.i.d. \quad (5)$$

Equation (5) introduces the noise ε_i, representing observational error, to the original function shown in Equation (1). This noise is assumed to be unbiased (with a mean of $\mu = 0$) and follows a normal distribution with a standard deviation σ.

Figure 9 illustrates examples comparing the approximations from the linear regression models in Equations (2) and (3) and the RF to the true function using simulations. To clearly demonstrate the models' ability, the prediction targets are set to the expected values of the function, rather than to the function values with random errors. The figures show that the tree-based model can still almost perfectly approximate the true function, even in the presence of noise. This result is consistently reflected in the distribution of *R*-squared values across 100 simulations, as shown in Figure 10.

[Figure 9 here]

[Figure 10 here]

We consider a scenario in which both locational and structural influences simultaneously affect housing prices. In general, housing prices are influenced by two main types of attributes: locational and structural. Locational attributes refer to the access benefits associated with a specific location, whereas structural attributes represent the value derived from the building itself, such as the number of rooms, quality of landscaping, and availability of parking. We express these structural attributes in the target model as follows:

$$p_i = \alpha_{1,s}|0.5 - x_i| + \alpha_{2,s}{x_i}^3 + \alpha_{3,s}{x_i}^2 + \alpha_{4,s}x_i + \alpha_{5,s}w_i + \alpha_{6,s}, \quad (6)$$

where w_i represents the structural factors that influence price. We assume that the structural factors are independent of the locational effect and are observable.

Figures 11 and 12 show the results of simulations that incorporate both locational and structural effects. The actual function values scatter around the trajectory dictated by the locational effect because the structural factors differ for each evaluation point. The simulation results demonstrate that the tree-based model accurately approximates the actual values, even in the presence of structural factors. Figure 12 shows the distribution of R-squared values over 100 simulations, revealing that the model consistently maintains a high accuracy level across different scenarios.

[Figure 11 here]

[Figure 12 here]

Next, we consider the case where the relationship between structural factors and housing prices varies by location. This issue is closely related to housing submarkets. Housing submarkets can be defined as clusters of houses that share similar price-determining factors. The relationship between housing prices and their attributes may differ across submarkets, particularly because locational differences influence price determinants. To illustrate this, we modify Equation (6), expressing that the influences of structural attributes w_i on housing prices vary depending on location.

$$p_i = \alpha_{1,s}(0.5 - x_i) + \alpha_{2,s}{x_i}^3 + \alpha_{3,s}{x_i}^2 + \alpha_{4,s}x_i + \alpha_{5,s}x_i w_i + \alpha_{6,s}. \quad (7)$$

Figures 13 and 14 compare the prediction outcomes for cases where the effects of structural attributes vary depending on locational factors. First, it can be observed that the predictive accuracy of the hedonic model, particularly the one including only the distance variable and the structural attribute, deteriorates significantly in this scenario. As is well known, this issue arises from the endogeneity between explanatory variables. The varying effect of structural attributes is not properly captured by their coefficients but is instead mistakenly attributed to the effect of the distance variable. In contrast, the tree-based model demonstrates a consistent and remarkable ability to capture the complex effects resulting from the endogeneity between structural attributes and locational factors, even without prior knowledge of this endogeneity. As shown in Figure 14, the tree-based model consistently achieves an R-squared value exceeding 0.99 across repeated experiments.

[Figure 13 here]

[Figure 14 here]

We address a limitation to consider when applying and interpreting the predictions of tree-based models. Unlike analytical models, tree-based models make predictions by matching similar cases, rather than explicitly modeling the relationships. When new data fall outside the range of features observed in training, tree-based models may find the closest match without

properly adjusting for differences in variable values, leading to biased predictions, especially in cases of extrapolation.

Figures 15 and 16 demonstrate how a tree-based model's predictions may become biased when extrapolating. Models (2) and (3), and the RF are trained on the target function from Equation (1), but the training data are now confined to the interval from 0 to 0.8, rather than covering the full range of the domain. Consequently, within the interval (0, 0.8], the models' predictions resemble those in Figure 4. However, for the (0.8, 1] interval, the models must predict outcomes in an unobserved region. While the analytical model adjusts its predictions by accounting for differences in variable values, the tree-based model simply extrapolates from the closest observed cases, leading to potential bias.

[Figure 15 here]

[Figure 16 here]

These limitations require careful consideration when analyzing housing prices because location and transaction timing substantially affect property values. The unique nature of each location means that housing datasets for specific areas are often sparse over short periods, leading to an increased risk of biased simulations with tree-based models. This is particularly concerning because analytical approaches, such as XAI models, rely on the accuracy and robustness of these simulations.

The potential bias in predictions highlights the need for caution when designing automated valuation models, especially because the timing of predictions usually does not match the timing of the training data. In general, predictions target future events; however, tree-based models have limitations in accounting for changes that occur during the evaluation period, which increases the risk of biased forecasts.

4. Real-World Data Analyses

The previous section shows through controlled simulation experiments that tree-based models are capable of effectively internalizing locational complexity using spatial partitioning based solely on geographic coordinates. We now examine whether the key features identified in the simulation experiments are also observed in real-world housing markets, based on real-world data.

Our analysis rests on two main expectations. First, tree-based models that incorporate geographic coordinates are expected to exhibit not only superior predictive performance relative to traditional analytical models, but also a high level of absolute predictive accuracy, provided that the sample size and spatial coverage of the data are sufficiently rich. Second, as locational complexity is effectively captured through coordinate-based spatial partitioning, the inclusion of conventional locational variables—such as distances to amenities or zoning classifications—should provide little additional predictive value.

Evaluating predictive performance for this analysis requires a dataset that is sufficiently large and comprehensive. This is because housing prices are shaped by both spatial segmentation and temporal dynamics. From a spatial perspective, housing markets are widely understood to

consist of multiple submarkets. When a cross-sectional analysis relies on geographically narrow samples, strong predictive performance may simply reflect homogeneity within a specific submarket rather than the model's ability to capture broader locational complexity. From a temporal perspective, housing prices are subject to market cycles. Analyses conducted at a single point in time make it difficult to distinguish structural predictive patterns from outcomes driven by a particular phase of the cycle. In short, a credible empirical analysis requires not only a large number of observations but also data with sufficiently broad regional and temporal coverage to assess the consistency of the results.

Accordingly, we use an extensive dataset covering all officially reported multi-family housing transactions across Seoul, the capital and largest city of South Korea, from January 1, 2011, to January 31, 2024. Obtained from the Korea Real Estate Board,[7] this dataset consists of 1,292,025 transaction records. Spanning more than a decade, the dataset encompasses multiple phases of the housing cycle, allowing for an evaluation of the consistency of the results across different market conditions.

Multi-family housing in South Korea includes row houses, multiplex housing, and apartments, which together constitute the dominant form of residential housing in the country. This housing structure reflects South Korea's rapid industrialization and urbanization from the 1970s through the 2000s, during which apartments in particular played a central role in accommodating large-scale urban housing demand (Ambrose & Kim 2003). According to the 2024 Population and Housing Census, of Seoul's total 3.17 million housing units, multi-family housing accounted for 2.85 million, representing 90.1% of the city's total housing stock.

Unlike detached houses, which pose analytical challenges due to substantial variation in interior and exterior characteristics that are difficult to quantify consistently, multi-family housing is particularly well-suited for quantitative analysis. Key attributes—such as the average number of parking spaces or the floor area ratio—tend to be standardized within and across similar housing complexes, enabling consistent measurement and meaningful comparison. Moreover, the dataset encompasses a wide range of multi-family housing types, from affordable public housing to high-end luxury complexes. These properties exhibit substantial variation in scale, from small low-rise buildings to large-scale residential complexes.

Table 2 displays the variables used in this study, which include both the structural and locational characteristics of the properties. The structural attributes address factors such as property size (m^2), property age (years), property type, building materials, the number of households in the complex, the average number of parking spaces, building coverage ratio, floor area ratio, property orientation, floor level, exclusive floor area, the presence of an elevator, the share of land ownership, loft structure status, the presence of a private garden, and the residential brand of the construction firm. The locational attributes cover latitude; longitude;

7 The Korea Real Estate Board is a government agency responsible for collecting and managing real estate statistics in Korea. The data were provided to the researcher in the course of a 2024 project aimed at designing a machine learning–based Automatic Valuation Model. However, due to data-sharing restrictions imposed by the data provider, the dataset is not publicly available.

administrative district; zoning district; assigned elementary, middle, and high schools; and accessibility metrics such as the distance to the nearest subway station, schools, and university. The transaction date is also included with the structural and locational attributes.

[Table 2 here]

Table 3 presents descriptive statistics for the numerical variables included in the dataset, summarizing key structural and locational characteristics of multi-family housing. The average number of households per complex is approximately 619, with a large standard deviation. Property size averages 81.2 square meters, while the wide range of observed values reflects the diversity of housing units included in the sample. The average property age is 15.5 years, and parking availability averages 0.8 spaces per unit, suggesting heterogeneity in building age and infrastructure provision. Building coverage ratios average 37.6%, while floor-to-area ratios average 207.7%. Taken together, these statistics suggest that the dataset captures a broad range of housing characteristics relevant for quantitative analysis.

[Table 3 here]

Using the data, we primarily examine the performance of two representative ensemble tree-based models: Random Forest (RF), representing a bagging-based approach, and XGBoost, representing a boosting-based approach. These models are implemented to evaluate the role of geographic coordinates and conventional locational variables, following the simulation framework discussed earlier.

Specifically, for both RF and XGBoost, we compare predictive performance across three model specifications: (i) models that include only latitude and longitude as locational inputs, (ii) models that include both latitude and longitude together with a rich set of conventional locational variables, and (iii) models that exclude geographic coordinates and rely solely on conventional locational variables. All specifications use the same set of structural housing attributes. For reference, we additionally estimate a linear regression model as a benchmark. The linear model incorporates all structural attributes and a wide range of conventional locational variables, but excludes latitude and longitude. In the tree-based models, categorical variables are encoded as integers, whereas in the linear regression model, categorical variables are transformed into dummy variables.

We aim to test the models under conditions that closely replicate real-world mass appraisal scenarios. To achieve this, it is essential to separate training and test samples based on time. Although it is common to randomly divide data into training and test sets, this approach is not appropriate for predicting housing prices because each transaction is associated with a specific date. If the training set includes observations from within or beyond the test period, the measured performance may reflect reliance on information that would not be available in a real-world setting, leading to an overestimation of predictive accuracy. Accordingly, a model must be trained on past data and evaluated on future data to ensure a credible assessment of predictive performance.

Based on this consideration, we segment the dataset to evaluate predictive performance over an extended period. Specifically, January of each year (from January 1 to January 31) is designated as the test set, while the training set consists of transaction data from the three years preceding January 1 of the corresponding year. This rolling evaluation scheme generates multiple overlapping time points, where each test set is subsequently incorporated into the training sample, allowing the results to be assessed repeatedly across different periods. As a result, the model is evaluated repeatedly across different temporal settings, allowing this procedure to function not only as a sequence of out-of-sample tests but also as a form of cross-validation for assessing the robustness of the results over time.

This design reflects the fact that cross-sectional models for individual housing price prediction rely on the assumption that market characteristics observed in the training period remain relevant during the prediction period. As the length of the test period increases, the likelihood of substantial changes in housing market conditions also rises, potentially distorting the evaluation of a model's explanatory power. For similar reasons, excessively outdated observations are excluded from the training sample. Under this setting, the training set spans three years of transaction data, while the test set covers one month at a time.

To assess the effectiveness of models, we use four metrics: R-squared, MAPE, mean percentage error (MPE), and the hitting rates (the proportion of predictions within the ranges of 10%, 20%, 30%, 40%, and 50%). A widely recognized and intuitive metric, MAPE measures the average deviation between predicted and actual transaction prices as a percentage. It calculates the absolute value of each error to ensure that positive and negative deviations do not negate each other, subsequently averaging these absolute values. However, MPE averages the percentage errors directly, without converting them to absolute values, which helps in identifying any consistent bias in the predictions, be it overestimation or underestimation. The formulae for MAPE and MPE are given as follows:

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{p_i-\hat{p}_i}{p_i}\right|, \tag{8}$$

$$MPE = \frac{100}{n}\sum_{i=1}^{n}\frac{p_i-\hat{p}_i}{p_i}, \tag{9}$$

where p_i refers to the actual price of the i-th property, $\hat{p}_i$ represents the corresponding predicted price, and n is the total number of samples.

Table 4 reports the predictive performance of Random Forest (RF) and XGBoost under alternative specifications of locational information, along with a linear regression benchmark, evaluated using a rolling, time-based testing framework. The results are highly consistent across test periods and provide several noteworthy findings. First, the predictive performance of tree-based models is not only superior relative to the linear regression model, but also high in an absolute sense. Across most test periods, both RF and XGBoost achieve low MAPE values, typically within the range of approximately 5 to 10. In many periods, R^2 values for the tree-based models exceed 0.90 and frequently approach or surpass 0.95, indicating a high degree of explanatory power. Consistent with these results, hitting rates are also consistently high: the proportion of predictions falling within ±20% of actual transaction prices commonly exceeds 90%, and rates within ±30% often exceed 95%. Taken together, these metrics indicate

that average prediction errors are limited to single-digit percentages while a large majority of predictions remain close to observed prices. Given the inherent noise in housing transaction prices and the well-known limitations of linear models in capturing complex locational effects, this level of accuracy suggests that tree-based models are highly effective in approximating the underlying price determination process. In particular, the substantial performance gap relative to the linear model—whose R^2 values are markedly lower and whose MAPE values frequently rise into double-digit ranges—supports the interpretation that the strength of tree-based models stems from their ability to capture locational complexity that is difficult to represent within linear specifications.

[Table 4 here]

Second, a striking result is that predictive performance remains largely unchanged when locational inputs are limited to geographic coordinates—latitude and longitude. For both RF and XGBoost, the predictive performance of the "coordinates-only" specification is nearly identical to that of models that additionally incorporate a rich set of conventional locational variables. Across most test periods, differences in MAPE, R^2, and hitting rates between these two specifications are minimal and often negligible in practical terms. In contrast, when geographic coordinates are excluded, and only conventional locational variables are used, predictive performance deteriorates substantially and consistently. Under these specifications, MAPE values commonly increase into the double-digit range, R^2 declines noticeably, and hitting rates at tighter error thresholds fall sharply for both RF and XGBoost across nearly all test periods. Taken together, these results strongly suggest that coordinate-based spatial partitioning effectively substitutes for a wide range of traditional locational variables.

While the preceding discussion emphasizes the strong average predictive accuracy of tree-based models, an examination of the mean percentage error (MPE) reveals an important additional dimension of predictive performance. In particular, the behavior of MPE indicates that high predictive accuracy in terms of error magnitude does not rule out the emergence of systematic directional bias in certain periods. Unlike MAPE, which measures the size of prediction errors regardless of direction, MPE captures the bias of predictions. If predictions are unbiased, MPE should be close to zero. However, the results show that MPE varies considerably over time. In some periods, MPE remains close to zero, while in others it exceeds 10% in absolute value, indicating the presence of non-negligible prediction bias during specific phases of the market.

This temporal variation in predictive bias can be traced to a structural limitation of tree-based models discussed earlier. Because these models do not explicitly estimate a functional relationship between housing prices and their underlying attributes, they rely on local information from the training sample when extrapolation is required. In real-world mass appraisal settings, predictions are often generated for properties whose transaction dates fall outside the temporal range of the training data. To reflect this practical constraint, our empirical design enforces a strict temporal separation between training and test samples. Under such conditions, tree-based models are unable to infer market changes that occur between the final observation in the training sample and the timing of the prediction targets.

The connection between prediction bias and extrapolation becomes clearer when MPE is examined alongside changes in housing prices over time. When housing prices rise during the evaluation period, predictions based on past data tend to underestimate realized prices; conversely, when prices decline, predictions are more likely to overestimate them. This relationship is reflected in a negative correlation between MPE and house price growth, which becomes more pronounced during periods of heightened price volatility. Figure 17 illustrates this pattern by comparing the MPE of both tree-based models with changes in the House Price Index (HPI), obtained from the Korea Real Estate Board, over time. The estimated correlation coefficients are –0.4 for RF and –0.55 for XGBoost, underscoring the importance of accounting for extrapolation-induced bias when interpreting or applying tree-based models in practice.

[Figure 17 here]

5. Concluding Remarks

This study examines how tree-based ensemble models capture the complex locational structures underlying cross-sectional housing prices. Rather than focusing solely on predictive performance, the paper aims to identify both the sources and the boundaries of spatial predictive power in these models. Conceptually, we describe how tree-based models internalize locational complexity not by explicitly modeling locational effects, but by recursively partitioning geographic space using coordinates. This mechanism allows the models to absorb nonlinearities, omitted locational factors, and latent housing submarkets without imposing functional assumptions. Through controlled simulation experiments, we demonstrate that this spatial partitioning mechanism is highly effective in capturing locational complexity, while also revealing the conditions under which it becomes fragile when predictions rely on extrapolation beyond the support of the training data. These insights are then validated using large-scale transaction data from Seoul in a realistic mass appraisal setting. The empirical results show that tree-based models incorporating geographic coordinates achieve consistently high levels of absolute predictive accuracy over time, while the inclusion of additional conventional locational variables yields little incremental benefit. At the same time, excluding geographic coordinates leads to a marked deterioration in performance, underscoring the central role of coordinate-based spatial partitioning in capturing locational complexity in real-world housing markets.

Nevertheless, our findings also highlight limitations. Because tree-based models do not estimate structural relationships between prices and their determinants, their predictions can exhibit systematic bias when applied to extrapolation tasks, particularly in periods of rapid market change. Moreover, the reliance on spatial partitioning complicates economic interpretation and poses challenges for simulation-based explanatory analyses. These limitations suggest that strong predictive performance should not be conflated with structural validity.

Taken together, the results position tree-based models as powerful predictive tools that complement, rather than replace, traditional analytical approaches. Their ability to internalize complex locational structures makes them particularly well-suited for applications such as large-scale urban valuation systems. However, their outputs should be interpreted with caution

when used for structural inference or policy analysis. By clarifying both the strengths and the boundaries of spatial predictive power in tree-based models, this study contributes to a more disciplined and informed use of machine learning methods in urban housing price analysis.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Conflict of interest

There is no conflict of interest in the study.

Funding

There is no external funding.

References

Adair, A.S., Berry, J.N., & McGreal, W.S. (1996). Hedonic modelling, housing submarkets and residential valuation. *Journal of Property Research*, 13(1), 67-83. https://doi.org/10.1080/095999196368899

Adetunji, A.B., Akande, O.N., Ajala, F.A., Oyewo, O., Akande, Y.F., & Oluwadara, G. (2022). House price prediction using random forest machine learning technique. *Procedia Computer Science*, 199, 806-813. https://doi.org/10.1016/j.procs.2022.01.100

Alonso, W. (1964). Location and land use: Toward a general theory of land rent. *Harvard University Press*. https://doi.org/10.4159/harvard.9780674730854

Ambrose, B.W., & Kim, S. (2003). Modeling the Korean Chonsei lease contract. *Real Estate Economics*, 31(1), 53-74. https://doi.org/10.1111/j.1080-8620.2003.00057.x

Anas, A. (1990). Taste heterogeneity and urban spatial structure: The logit model and monocentric theory reconciled. *Journal of Urban Economics*, 28(3), 318-335. https://doi.org/10.1016/0094-1190(90)90031-h

Antipov, E.A., & Pokryshevskaya, E.B. (2012). Mass appraisal of residential apartments: An application of Random forest for valuation and a CART-based approach for model diagnostics. *Expert Systems with Applications*, 39(2), 1772-1778. https://doi.org/10.2139/ssrn.1729653

Avery, C., & Pathak, P.A. (2021). The distributional consequences of public school choice. *American Economic Review*, 111(1), 129-152. https://doi.org/10.3386/w21525

Bajari, P., Fruehwirth, J.C., Kim, K.I., & Timmins, C. (2012). A rational expectations approach to hedonic price regressions with time-varying unobserved product attributes: The price of pollution. *American Economic Review*, 102(5), 1898-1926. https://doi.org/10.1257/aer.102.5.1898

Baum-Snow, N., & Marion, J. (2009). The effects of low income housing tax credit developments on neighborhoods. *Journal of Public Economics*, 93(5-6), 654-666. https://doi.org/10.1016/j.jpubeco.2009.01.001

Berry, B.J.L., & Bednarz, R.S. (1975). A hedonic model of prices and assessments for single-family homes: Does the assessor follow the market or the market follow the assessor? *Land Economics*, 51(1), 21-40. https://doi.org/10.2307/3145138

Black, S.E., & Machin, S. (2011). Housing valuations of school performance. *In Handbook of the Economics of Education* (Vol. 3, pp. 485-519). Elsevier. https://doi.org/10.1016/b978-0-444-53429-3.00010-7

Bourassa, S.C., Hamelink, F., Hoesli, M., & MacGregor, B.D. (1999). Defining housing submarkets. *Journal of Housing Economics*, 8(2), 160-183. https://doi.org/10.1006/jhec.1999.0246

Bourassa, S.C., Hoesli, M., & Peng, V.S. (2003). Do housing submarkets really matter?. *Journal of Housing Economics*, 12(1), 12-28. https://doi.org/10.1016/S1051-1377(03)00003-2

Chen, J., Glaeser, E., & Wessel, D. (2023). JUE Insight: The (non-) effect of opportunity zones on housing prices. *Journal of Urban Economics*, 133, 103451. https://doi.org/10.1016/j.jue.2022.103451

Clapp, J.M., & Lindenthal, T. (2022). Urban land valuation with bundled good and land residual assumptions. *Journal of Housing Economics*, 58, 101872. https://doi.org/10.1016/j.jhe.2022.101872

Ding, K., & Itoh, R. (2023). JUE Insight: The impact of the school admission restriction policy on the housing market in Shanghai. *Journal of Urban Economics*, 136, 103568. https://doi.org/10.1016/j.jue.2023.103568

Dou, M., Gu, Y., & Fan, H. (2023). Incorporating neighborhoods with explainable artificial intelligence for modeling fine-scale housing prices. *Applied Geography*, 158, 103032. https://doi.org/10.1016/j.apgeog.2023.103032

Durlauf, S.N. (2004). Neighborhood effects. *Handbook of Regional and Urban Economics*, 4, 2173-2242. https://doi.org/10.1016/s1574-0080(04)80007-5

Ekeland, I., Heckman, J.J., & Nesheim, L. (2004). Identification and estimation of hedonic models. *Journal of Political Economy*, 112(S1), S60-S109. https://doi.org/10.3386/w9910

Feng, H., & Lu, M. (2013). School quality and housing prices: Empirical evidence from a natural experiment in Shanghai, China. *Journal of Housing Economics*, 22(4), 291-307. https://doi.org/10.1016/j.jhe.2013.10.003

Follain, J.R., & Malpezzi, S. (1981). Another look at racial differences in housing prices. *Urban Studies*, 18(2), 195-203. https://doi.org/10.1080/00420988120080351

Goodman, A.C., & Thibodeau, T.G. (2007). The spatial proximity of metropolitan area housing submarkets. Real Estate Economics, 35(2), 209-232. https://doi.org/10.1111/j.1540-6229.2007.00188.x

Hansson, F. (2024). Pricing radioactive pollutants: Evidence from the Swedish housing market. *Journal of Real Estate Finance and Economics*, 69(4), 585-620. https://doi.org/10.1007/s11146-021-09885-7

Hardman, A., & Ioannides, Y.M. (2004). Neighbors' income distribution: Economic segregation and mixing in US urban neighborhoods. *Journal of Housing Economics*, 13(4), 368-382. https://doi.org/10.1016/j.jhe.2004.09.003

Hill, R.J. (2013). Hedonic price indexes for residential housing: A survey, evaluation and taxonomy. *Journal of Economic Surveys*, *27*(5), 879-914. https://doi.org/10.1111/j.1467-6419.2012.00731.x

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

Hong, J., & Kim, W. S. (2022). Combination of machine learning-based automatic valuation models for residential properties in South Korea. *International Journal of Strategic Property Management*, *26*(5), 362-384. https://doi.org/10.3846/ijspm.2022.17909

Ioannides, Y.M., & Zabel, J.E. (2008). Interactions, neighborhood selection and housing demand. *Journal of Urban Economics*, 63(1), 229-252. https://doi.org/10.1016/j.jue.2007.01.010

Jafary, P., Shojaei, D., Rajabifard, A., & Ngo, T. (2024). Automated land valuation models: A comparative study of four machine learning and deep learning methods based on a comprehensive range of influential factors. *Cities*, 151, 105115. https://doi.org/10.1016/j.cities.2024.105115

Kang, Y., Zhang, F., Peng, W., Gao, S., Rao, J., Duarte, F., & Ratti, C. (2021). Understanding house price appreciation using multi-source big geo-data and machine learning. *Land Use Policy*, 111, 104919. https://doi.org/10.1016/j.landusepol.2020.104919

Kain, J.F., & Quigley, J.M. (1972). Housing market discrimination, home-ownership, and savings behavior. *American Economic Review*, 62(3), 263-277. http://www.jstor.org/stable/1803375

Kim, H., Cho, H., & Ryu, D. (2018). Characteristics of mortgage terminations: An analysis of a loan-level dataset. *Journal of Real Estate Finance and Economics*, 57(4), 647–676. https://doi.org/10.1007/s11146-017-9620-5

Krämer, B., Nagl, C., Stang, M., & Schäfers, W. (2023). Explainable AI in a real estate context–exploring the determinants of residential real estate values. *Journal of Housing Research*, 32(2), 204-245. https://doi.org/10.1080/10527001.2023.2170769

Lancaster, K.J. (1966). A new approach to consumer theory. *Journal of Political Economy*, 74(2), 132-157. https://doi.org/10.1086/259131

Malpezzi, S. (2003). Hedonic pricing models: A selective and applied review. *Housing Economics and Public Policy/Blackwell Publishing*.

https://doi.org/10.1002/9780470690680.ch5

Mamre, M. O., & Sommervoll, D. E. (2024). Coming of age: renovation premiums in housing markets. *Journal of Real Estate Finance and Economics*, 69(2), 307-342. https://doi.org/10.1007/s11146-022-09917-w

Ming, Y., Zhang, J., Qi, J., Liao, T., Wang, M., & Zhang, L. (2020). Prediction and analysis of Chengdu housing rent based on XGBoost algorithm. *In Proceedings of the 3rd International Conference on Big Data Technologies* (pp. 1-5). https://doi.org/10.1145/3422713.3422720

Mullainathan, S., & Spiess, J. (2017). Machine learning: an applied econometric approach. *Journal of Economic Perspectives*, 31(2), 87-106. https://doi.org/10.1257/jep.31.2.87

Ou, Y., Zheng, S., & Nam, K.M. (2022). Impacts of air pollution on urban housing prices in China. *Journal of Housing and the Built Environment*, 37(1), 423-441. https://doi.org/10.1007/s10901-021-09845-w

Pace, R.K., & Gilley, O.W. (1990). Estimation employing a priori information within mass appraisal and hedonic pricing models. *Journal of Real Estate Finance and Economics*, 3, 55-72. https://doi.org/10.1007/bf00153706

Pace, R.K., & Hayunga, D. (2020). Examining the information content of residuals from hedonic and spatial models using trees and forests. *Journal of Real Estate Finance and Economics*, 60(1), 170-180. https://doi.org/10.1007/s11146-019-09724-w

Pita, R.P., de Carvalho, A.R., Barbosa, R.M., Cury, A.A., & Mendes, J.C. (2026). A Systematic review of the use of machine learning in the prediction of house pricing. *Computational Economics*, in press. https://doi.org/10.1007/s10614-025-10983-4

Quigley, J.M. (1985). Consumer choice of dwelling, neighborhood and public services. *Regional Science and Urban Economics*, 15(1), 41-63. https://doi.org/10.1016/0166-0462(85)90031-6

Rico-Juan, J.R., & 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*, 171, 114590. https://doi.org/10.1016/j.eswa.2021.114590

Rosen, S. (1974). Hedonic prices and implicit markets: product differentiation in pure competition. *Journal of Political Economy*, 82(1), 34-55. https://doi.org/10.2307/j.ctv1kz4h99.5

Ryu, D., Hong, J., & Jo, H. (2024). Capturing locational effects: Application of the K-means clustering algorithm. *Annals of Regional Science*, 73, 265-289. https://doi.org/10.1007/s00168-024-01263-4

Shaamala, A., Yigitcanlar, T., Nili, A., & Nyandega, D. (2025). Machine learning applications for urban geospatial analysis: A review of urban and environmental studies. *Cities*, 165, 106139. https://doi.org/10.1016/j.cities.2025.106139

Sheppard, S. (1999). Hedonic analysis of housing markets. *Handbook of Regional and Urban Economics*, 3, 1595-1635. https://doi.org/10.1016/S1574-0080(99)80010-8

Sheppard, S. (2010). Measuring the impact of culture using hedonic analysis. *Center for Creative Community Development*, 28. https://web.williams.edu/Economics/ArtsEcon/library/pdfs/UsingHedonicAnalysis2010.pdf

Taylor, L.O. (2008). Theoretical foundations and empirical developments in hedonic modeling. In *Hedonic methods in housing markets: Pricing environmental amenities and segregation* (pp. 15-37). New York, NY: Springer New York. https://doi.org/10.1007/978-0-387-76815-1_2

Tita, G.E., Petras, T.L., & Greenbaum, R.T. (2006). Crime and residential choice: A neighborhood level analysis of the impact of crime on housing prices. *Journal of Quantitative Criminology*, 22, 299-317. https://doi.org/10.1007/s10940-006-9013-z

Ugur, L. O., Kanit, R., Erdal, H., Namli, E., Erdal, H. I., Baykan, U. N., & Erdal, M. (2019). Enhanced predictive models for construction costs: A case study of Turkish mass housing sector. *Computational Economics*, 53, 1403-1419. https://doi.org/10.1007/s10614-018-9814-9

Wang, J., & Li, G. (2022). Pursuing educational equality and divergence in the housing market: How do educational equality policies affect housing prices in Shanghai?. *Cities*, 131, 104001. https://doi.org/10.1016/j.cities.2022.104001

Yilmazer, S., & Kocaman, S. (2020). A mass appraisal assessment study using machine learning based on multiple regression and random forest. *Land Use Policy*, 99, 104889. https://doi.org/10.1016/j.landusepol.2020.104889

Zeng, S., & Rao, D. (2025). Random forests with economic roots: Explaining machine learning in hedonic imputation. *Computational Economics*, 66, 2457-2481. https://doi.org/10.1007/s10614-024-10798-9

Zhang, H., Chen, J., & Wang, Z. (2021). Spatial heterogeneity in spillover effect of air pollution on housing prices: Evidence from China. *Cities*, 113, 103145. https://doi.org/10.1016/j.cities.2021.103145

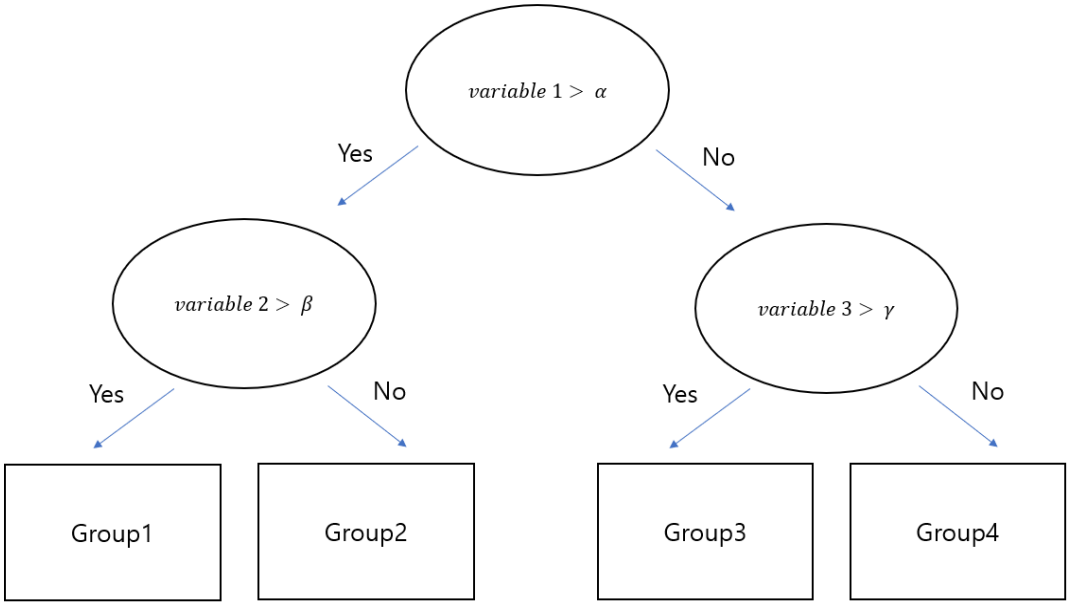
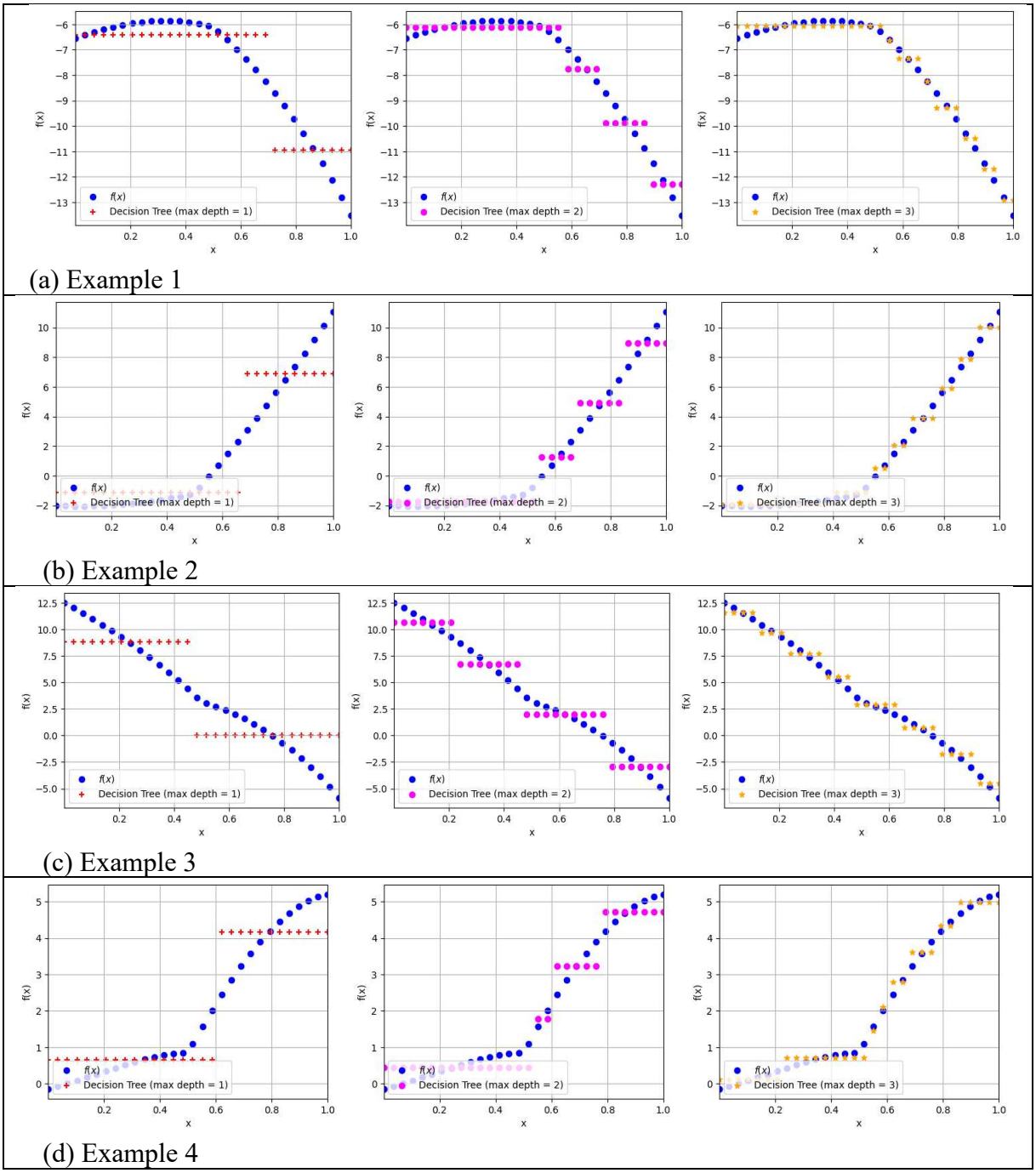


Fig. 1. Structure of a decision tree
Note: This figure shows an example of how a decision tree segments samples.



(a) Example 1

(b) Example 2

(c) Example 3

(d) Example 4

Fig. 2. Examples of predictions generated by a decision tree at varying depths

Note: This figure illustrates examples of predictions from a decision tree. The horizontal axis represents the location x. The vertical axis represents either the arbitrary non-linear function value ($f(x)$) or the predicted value. The blue dotted line shows the actual function values. The red cross-dotted line (left), the magenta round-dotted line (middle), and the orange star-dotted line (right) indicate the predictions from the decision tree at depths of 1, 2, and 3, respectively.

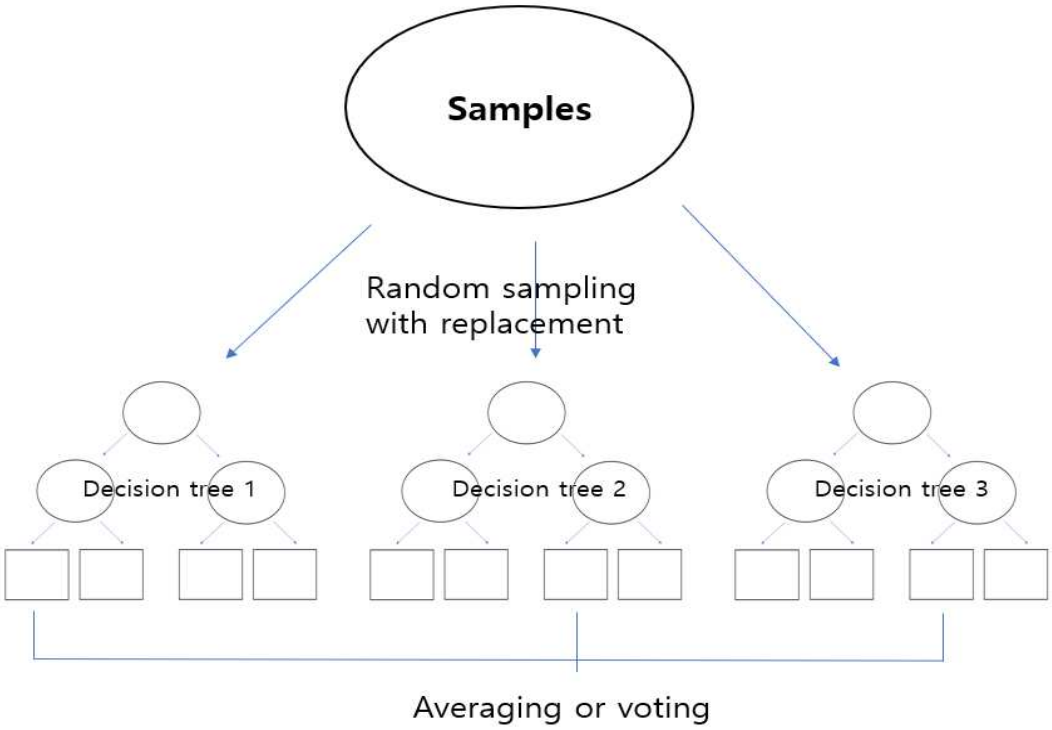
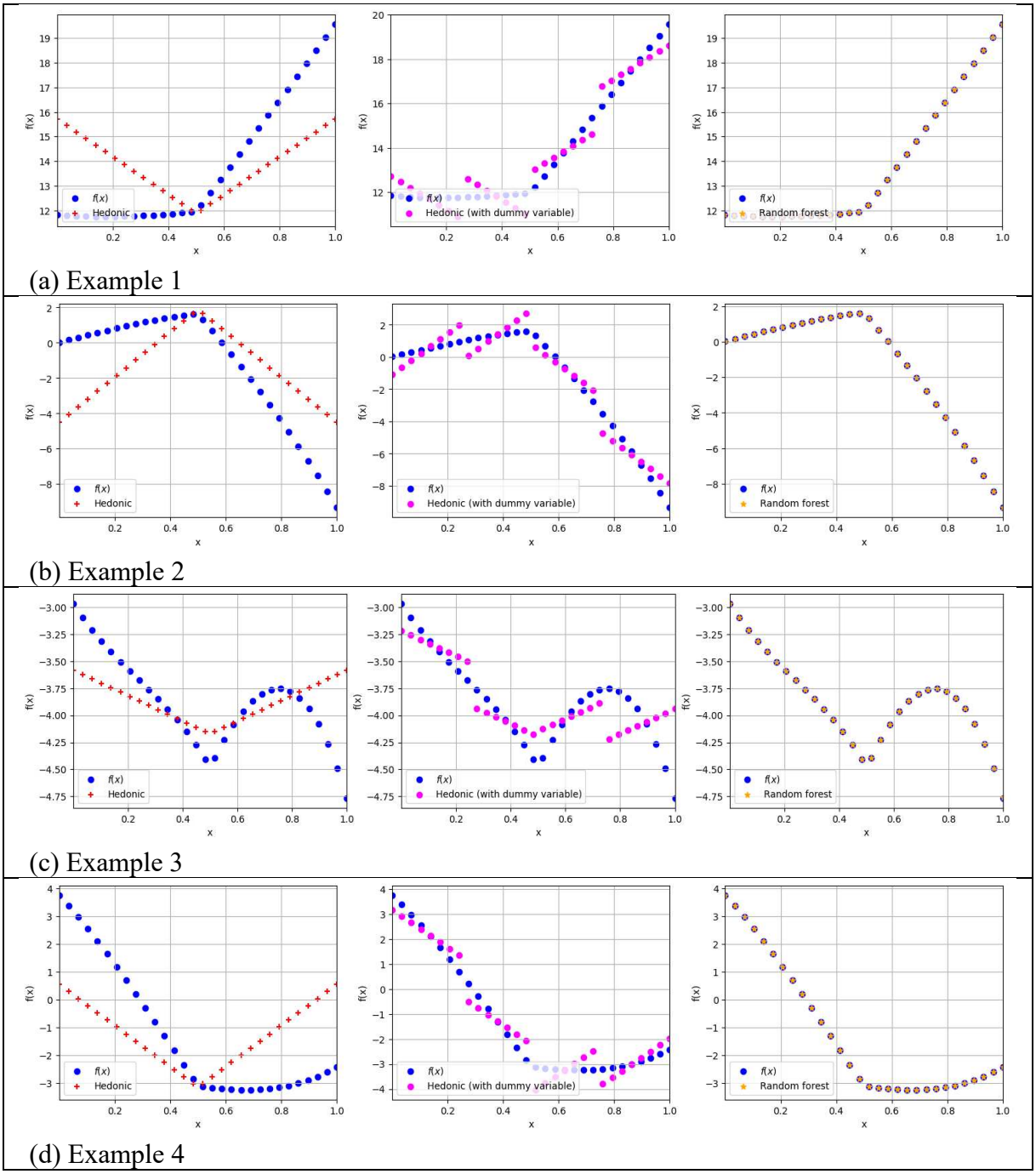


Fig. 3. Structure of a random forest model
Note: This figure illustrates an example of a bagging model.



(a) Example 1

(b) Example 2

(c) Example 3

(d) Example 4

Fig. 4. Comparison between linear models and a tree-based model: Baseline model

Note: This figure presents examples of baseline simulations. The horizontal axis indicates the location x. The vertical axis indicates the function value ($f(x)$) or the predicted value. The blue dotted line shows the actual function values. The red cross dotted line (left) represents the hedonic model's predictions using only the distance. The magenta round dotted line (middle) shows the hedonic model's predictions using both distance and dummy variables. The orange star dotted line (right) indicates the random forest's predictions.

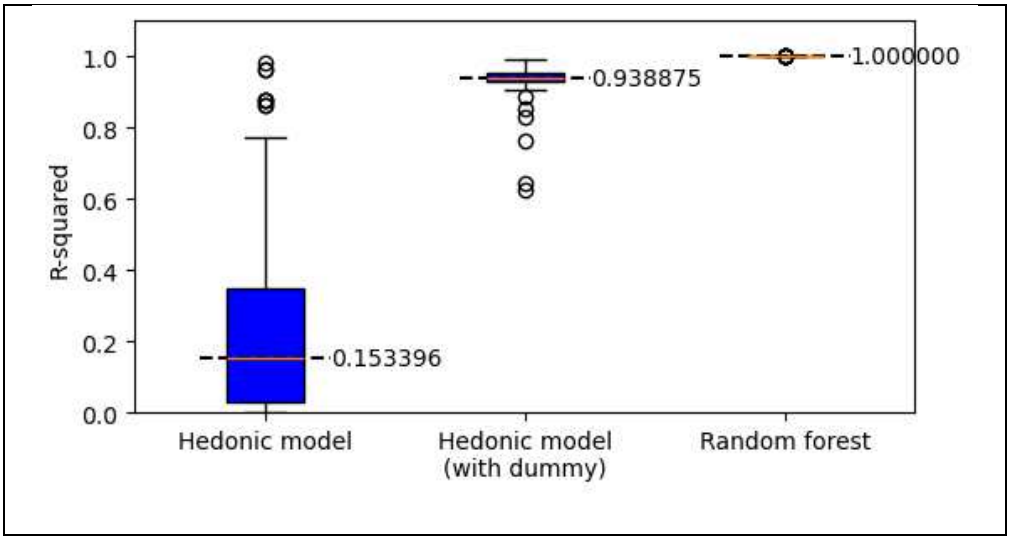


Fig. 5. Distribution of *R*-squared values for conventional models and a tree-based model: Baseline model

Note: The above distributions are derived from 100 repeated experiments. The lines positioned on each bar represent the median of each distribution.

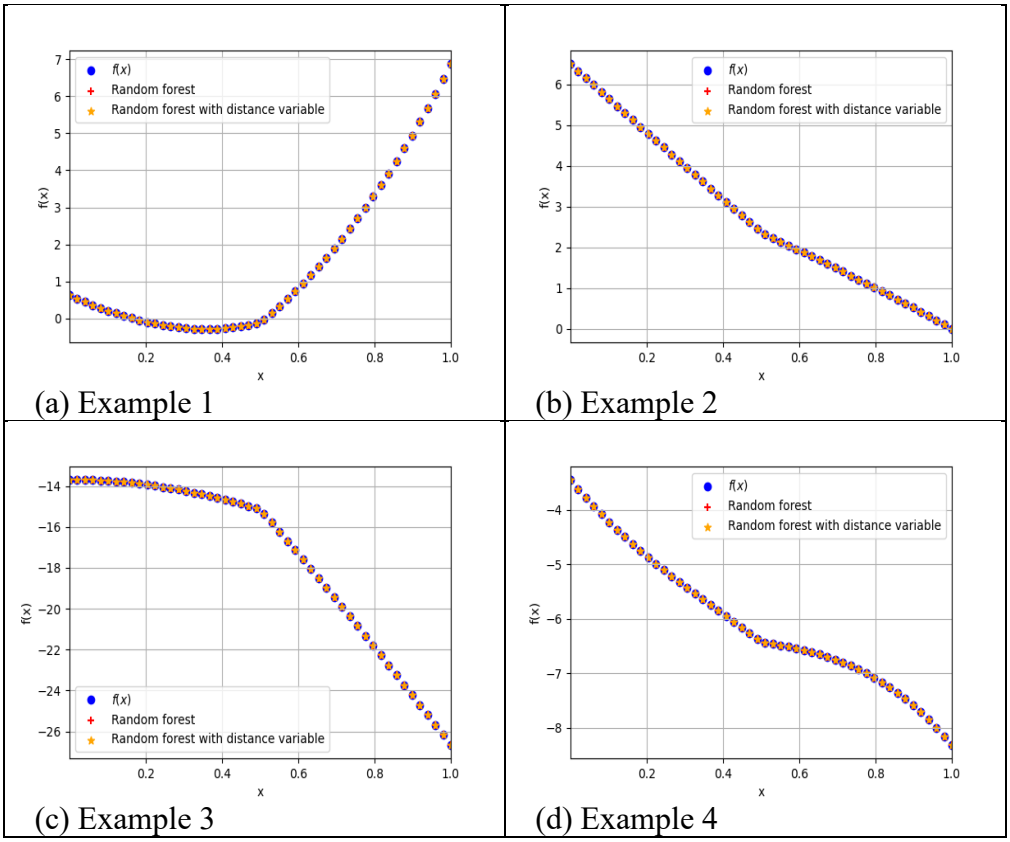


Fig. 6. Comparison of tree-based models: Coordinates only versus coordinates with distance
Note: The horizontal axis indicates the location x. The vertical axis indicates the function value ($f(x)$) or the predicted value. The blue dotted line shows the actual function values. The red cross dotted line is the prediction of the random forest model, including only the location x. The orange star dotted line indicates the predictions of the random forest model, including both the location x and the distance from the center point.

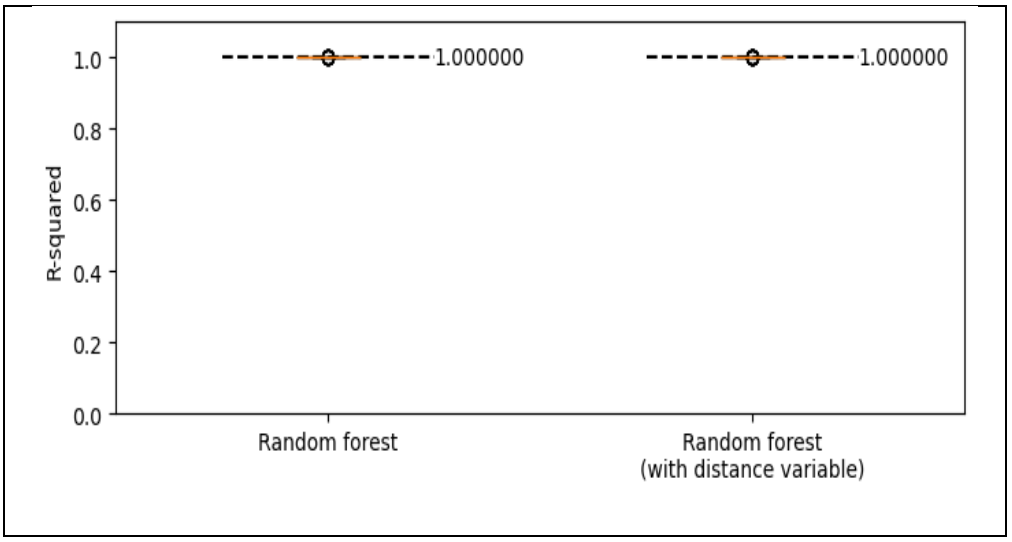


Fig. 7. Distribution of *R*-squared values: Coordinates only versus coordinates with distance
Note: The left plot reports the R-squared obtained when only the location coordinates are included, while the right plot reports the R-squared obtained when both the location coordinates and the distance variable are included. Since the R-squared is already close to 1 when only the coordinates are used, adding the distance variable does not lead to any further improvement in explanatory power. The above distributions are derived from 100 repeated experiments. The lines positioned on each bar represent the median of each distribution.

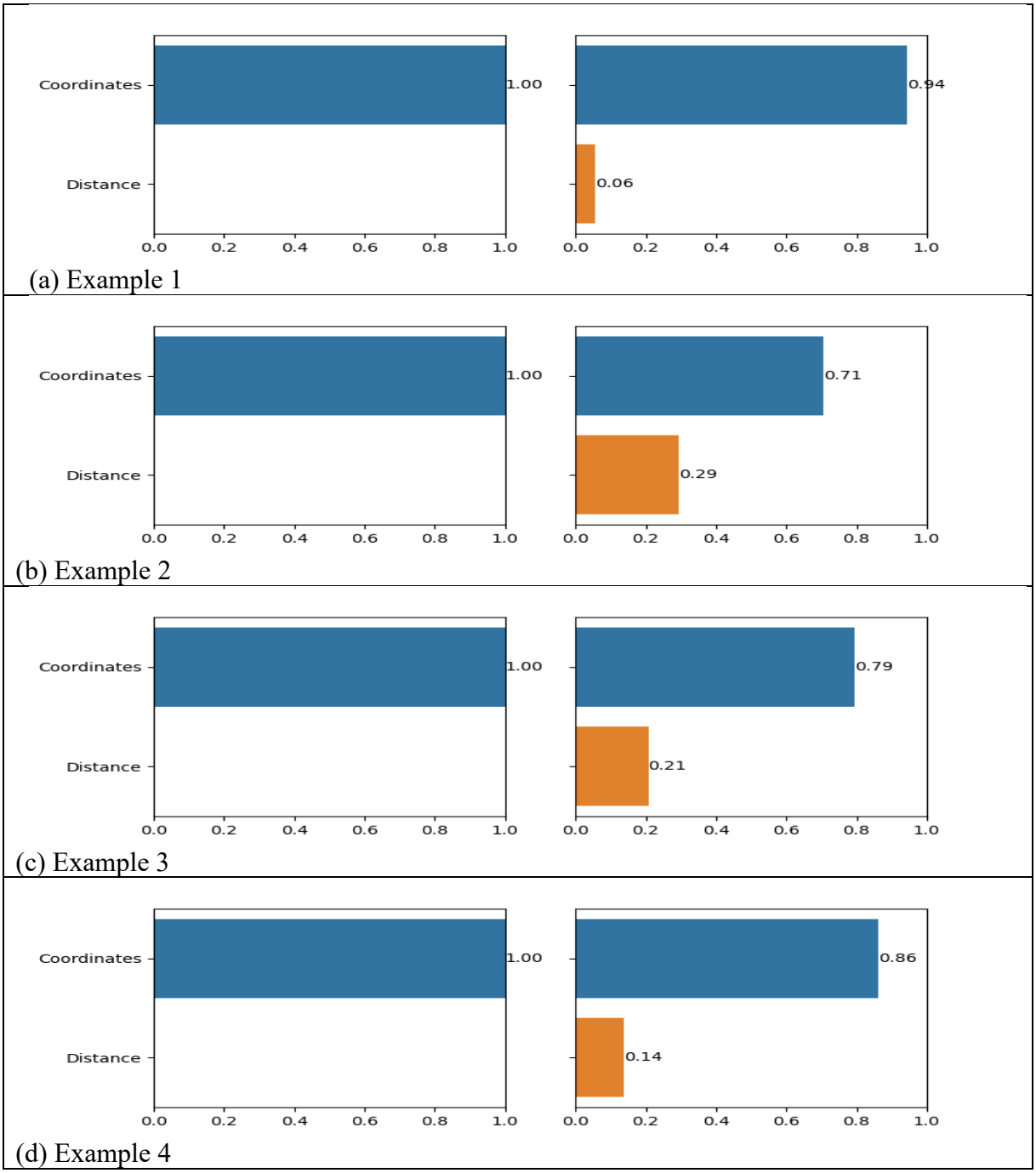
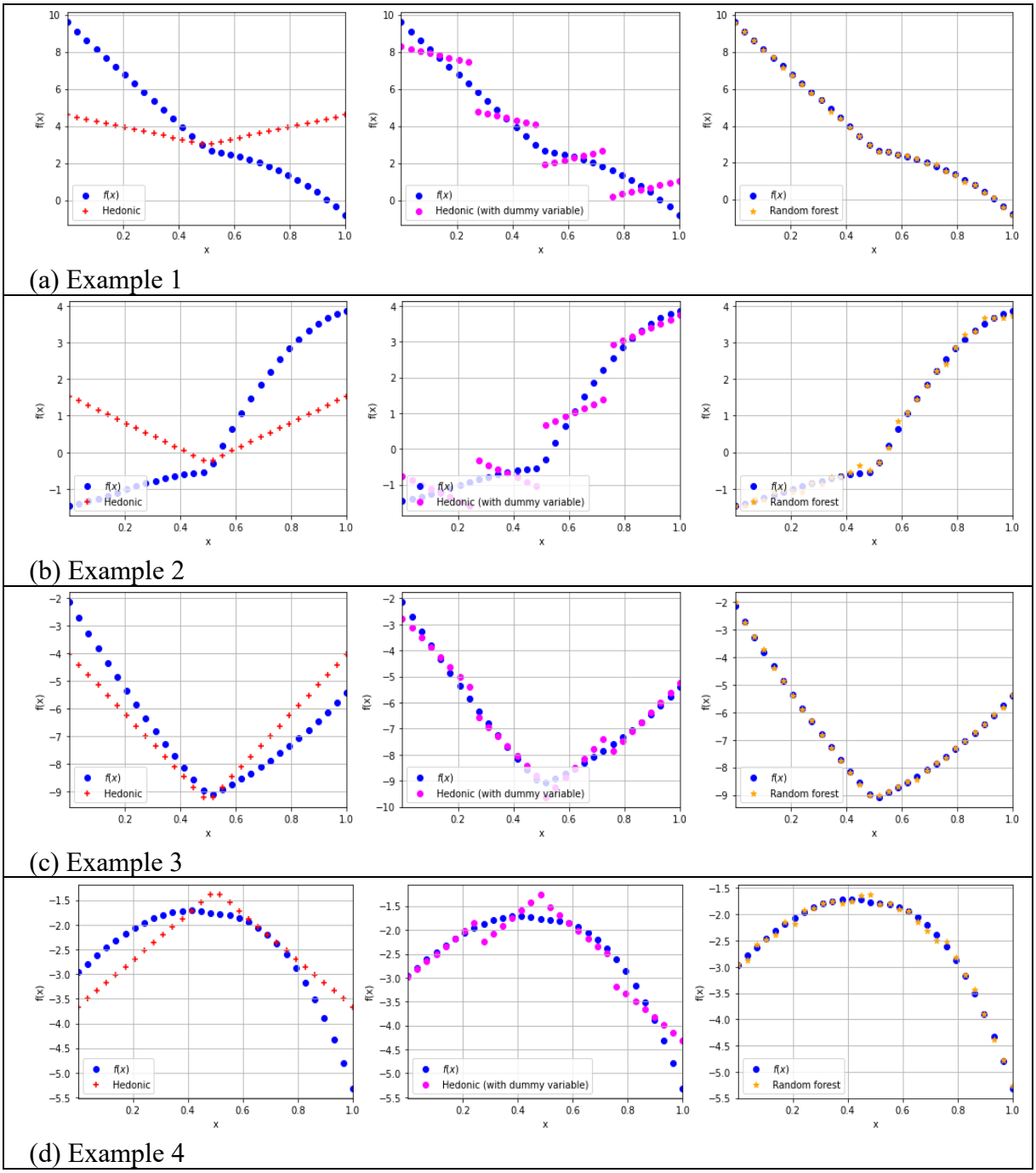


Fig. 8. Comparison of importance measurements: Coordinates only versus coordinates with distance

Note: This figure presents examples of feature importance tests. The results on the left are derived from a model that includes only location coordinates. Those on the right are from a model that incorporates both location coordinates and the distance from the midpoint. The blue bars represent the importance metric of the location coordinates. The orange bars represent the importance metric of the distance from the midpoint.



(a) Example 1

(b) Example 2

(c) Example 3

(d) Example 4

Fig. 9. Comparison between linear models and a tree-based model, with a noise factor

Note: This figure presents examples of the simulations with a noise factor. The horizontal axis indicates the location x. The vertical axis indicates the function value ($f(x)$) or the predicted value. The blue dotted line shows the actual function values. The red cross dotted line represents the hedonic model's predictions using only the distance. The magenta round dotted line shows the hedonic model's predictions using both distance and dummy variables. The orange star-dotted line indicates the random forest's predictions.

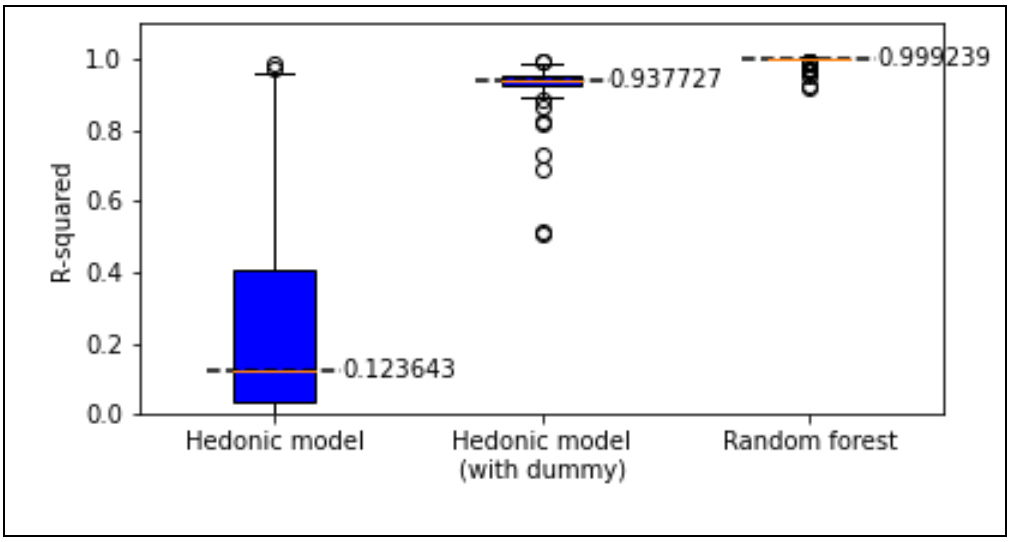
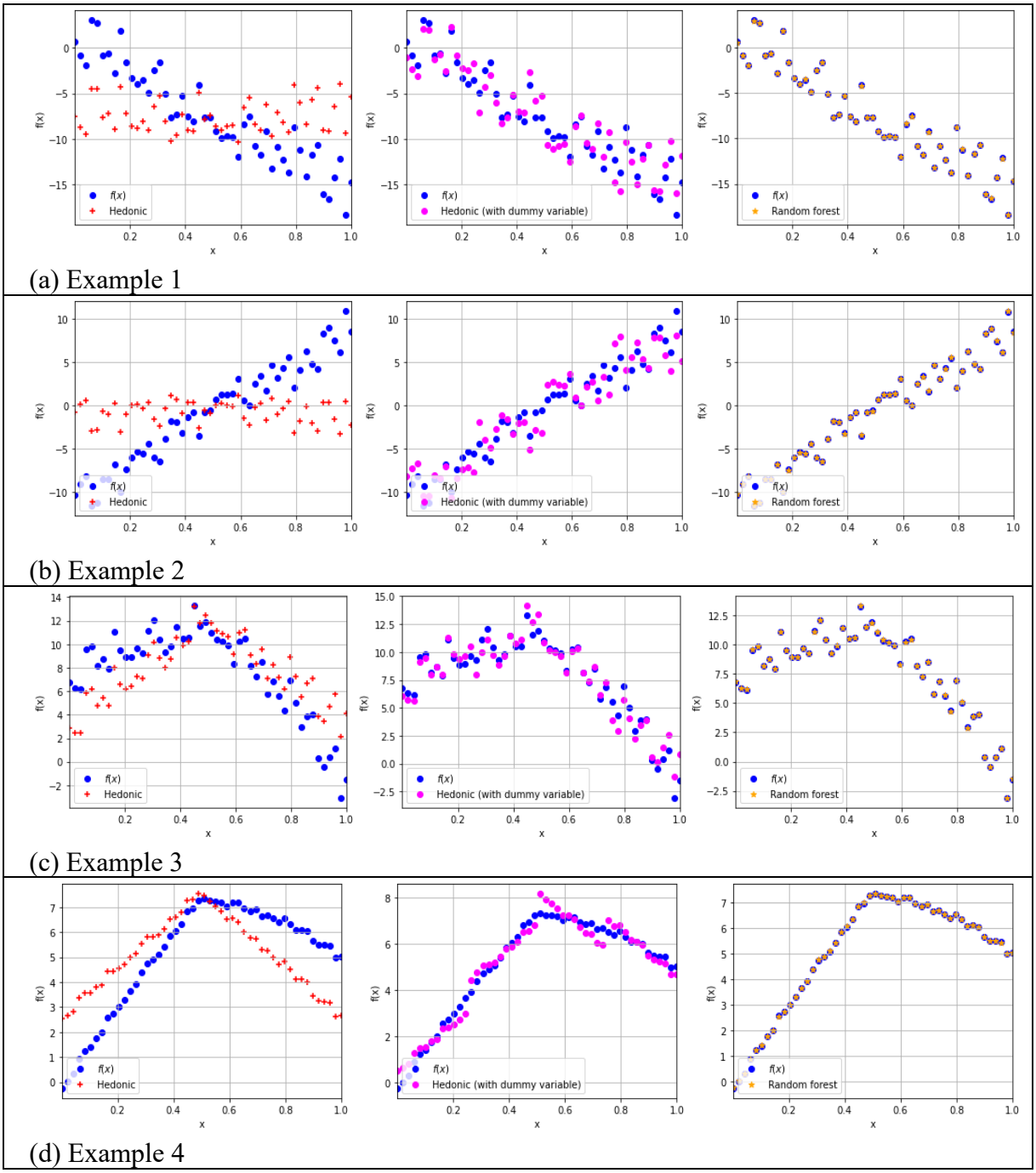


Fig. 10. Distribution of *R*-squared values for conventional models and a tree-based model, with a noise factor

Note: The above distributions are derived from 100 repeated experiments. The lines positioned on each bar represent the median of each distribution.



(a) Example 1

(b) Example 2

(c) Example 3

(d) Example 4

Fig. 11. Comparison between linear models and a tree-based model, with a structural factor
Note: This figure presents examples of the simulations with a structural attribute. The horizontal axis indicates the location x. The vertical axis indicates the function value ($f(x)$) or the predicted value. The blue dotted line shows the actual function values. The red cross dotted line represents the hedonic model's predictions using only the distance. The magenta round dotted line shows the hedonic model's predictions using both distance and dummy variables. The orange star-dotted line indicates the random forest's predictions.

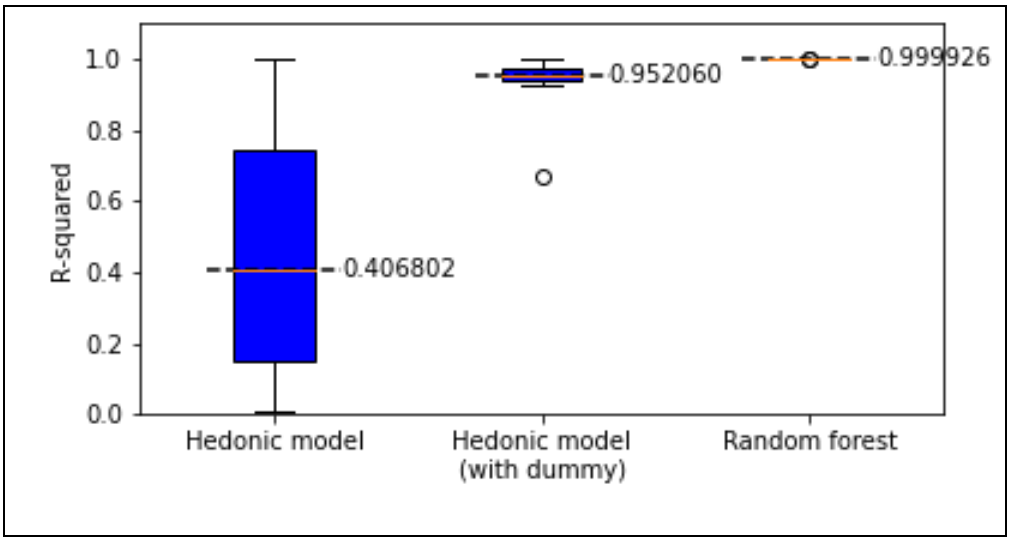
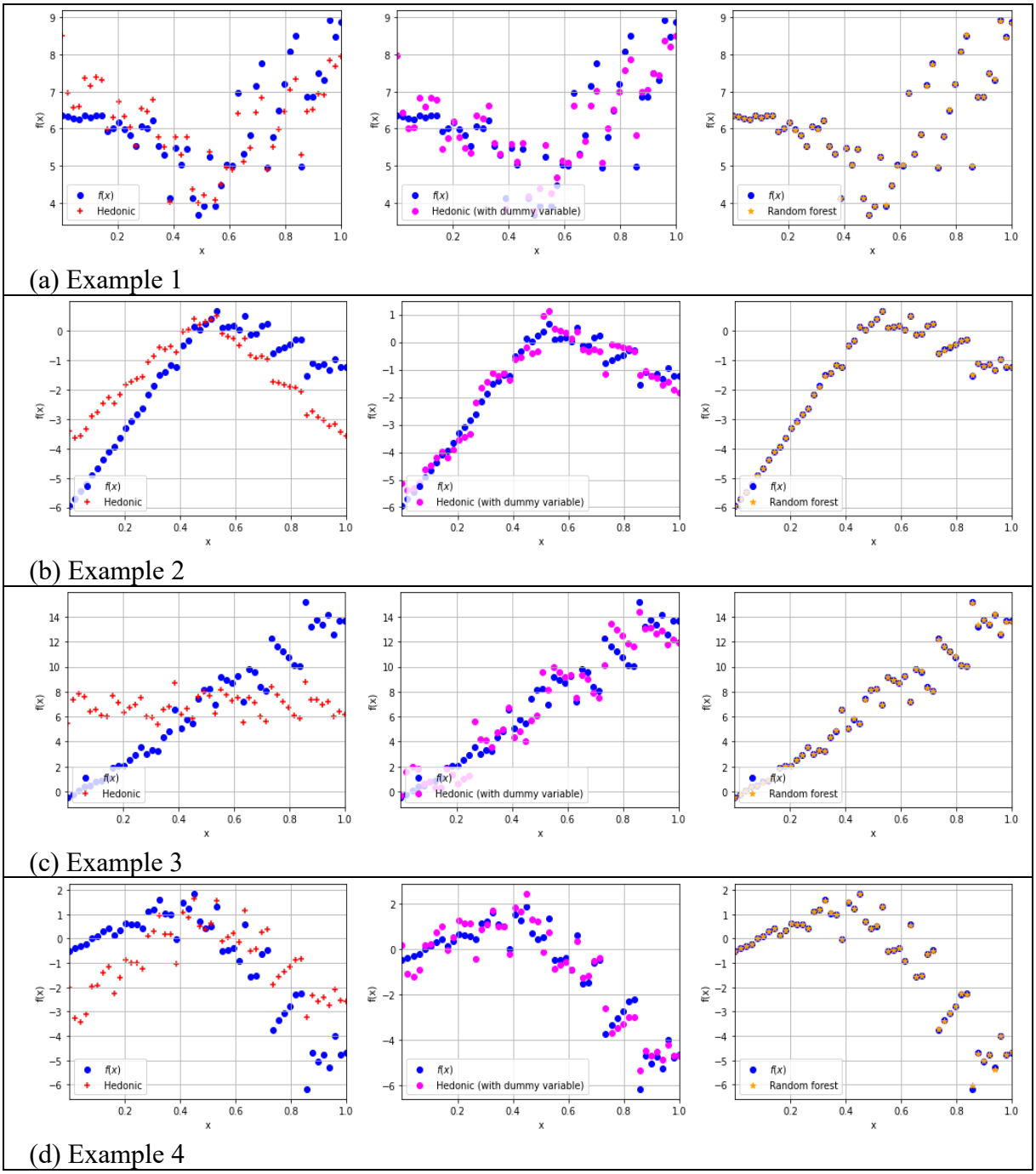


Fig. 12. Distribution of *R*-squared values for conventional models and a tree-based model, with a structural factor

Note: The above distributions are derived from 100 repeated experiments. The lines positioned on each bar represent the median of each distribution.



(a) Example 1

(b) Example 2

(c) Example 3

(d) Example 4

Fig. 13. Comparison between linear models and a tree-based model, with a varying structural factor

Note: This figure presents examples of the simulations with a structural attribute. The horizontal axis indicates the location x. The vertical axis indicates the function value ($f(x)$) or the predicted value. The blue dotted line shows the actual function values. The red cross dotted line represents the hedonic model's predictions using only the distance. The magenta round dotted line shows the hedonic model's predictions using both distance and dummy variables. The orange star-dotted line indicates the random forest's predictions.

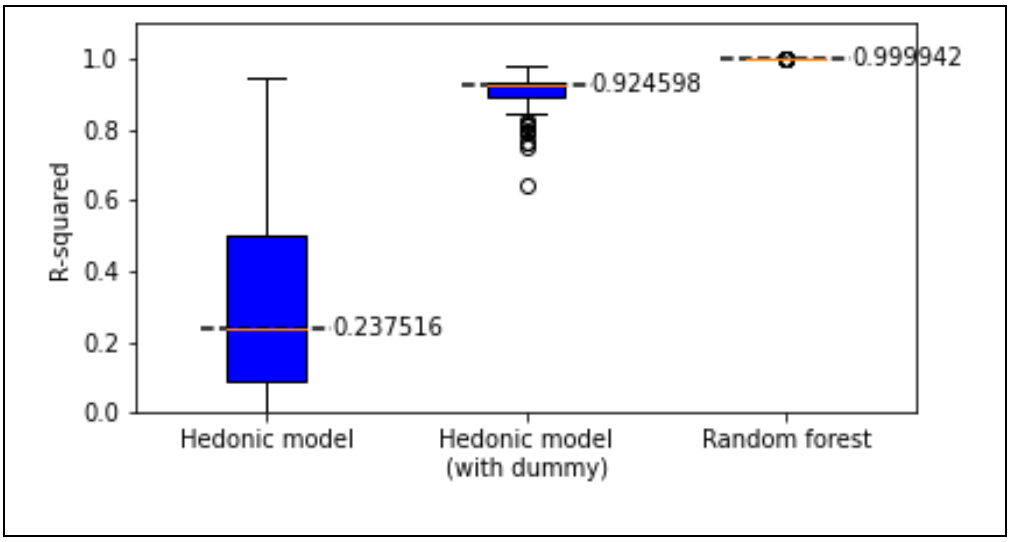
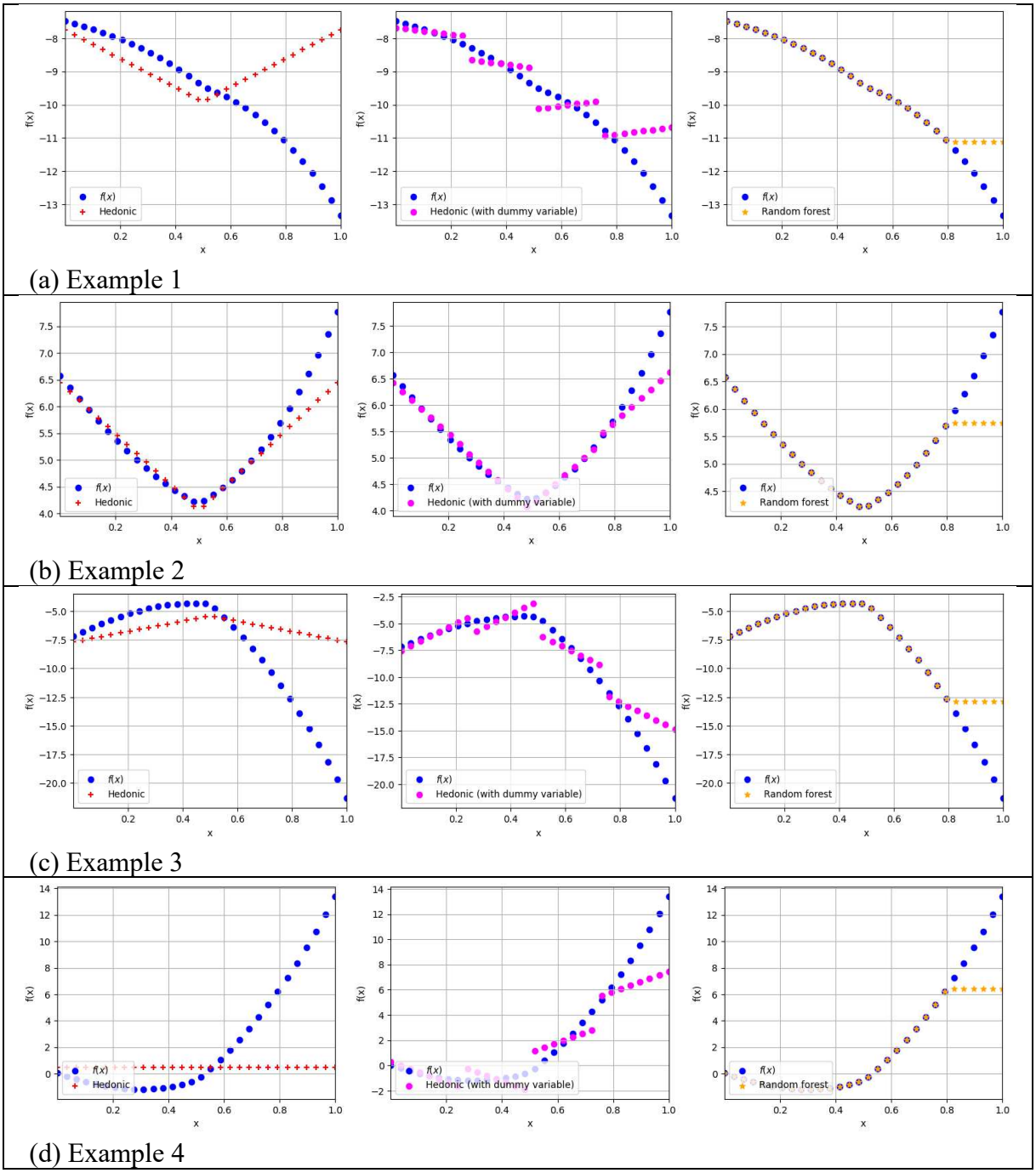


Fig. 14. Distribution of *R*-squared values for conventional models and a tree-based model, with a varying structural factor

Note: The above distributions are derived from 100 repeated experiments. The lines positioned on each bar represent the median of each distribution.



(a) Example 1

(b) Example 2

(c) Example 3

(d) Example 4

Fig. 15. Comparison between conventional models and a tree-based model: Extrapolation case
Note: This figure presents examples of the extrapolative simulations. The horizontal axis indicates the location x. The vertical axis indicates the function value ($f(x)$) or the predicted value. The blue dotted line shows the actual function values. The red cross dotted line represents the hedonic model's predictions using only the distance. The magenta round dotted line shows the hedonic model's predictions using both distance and dummy variables. The orange star-dotted line indicates the random forest's predictions.

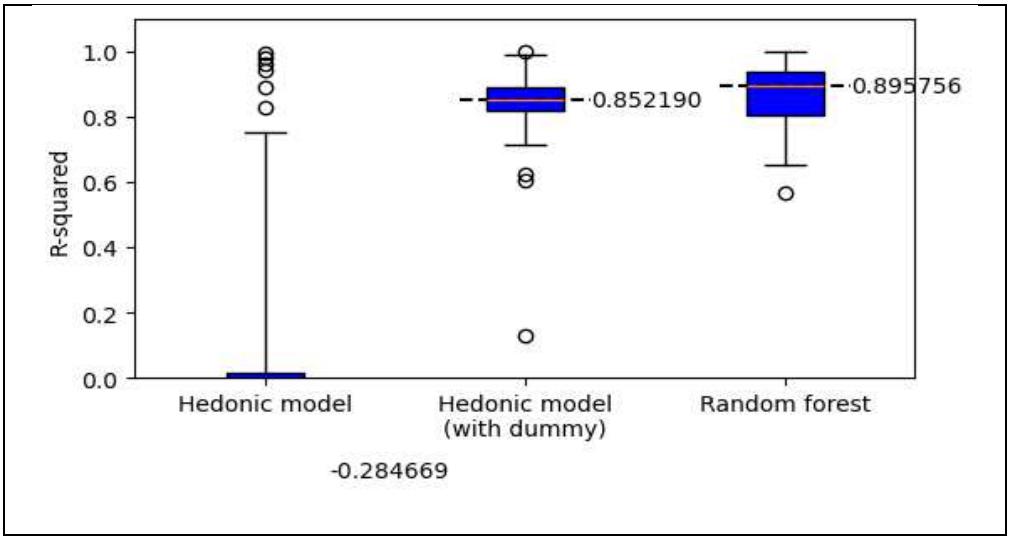
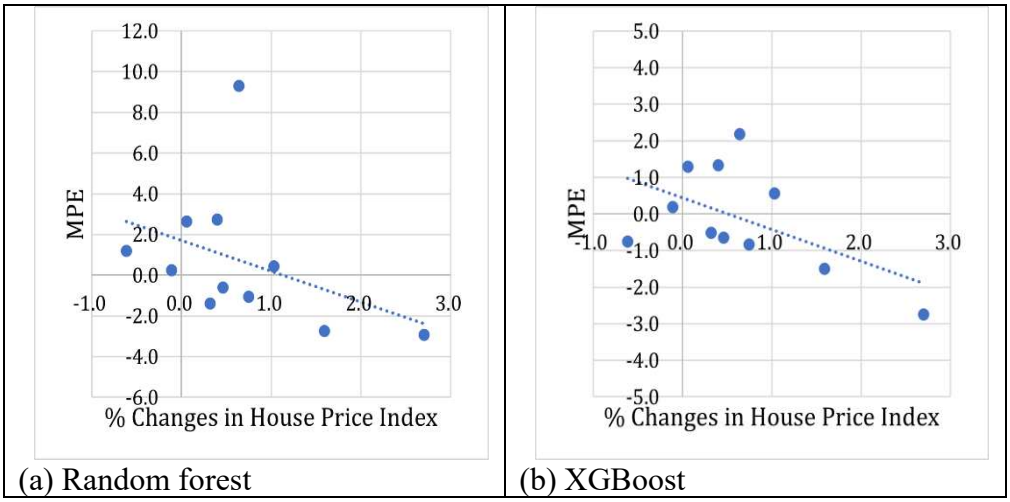


Fig. 16. Distribution of *R*-squared values for conventional models and a tree-based model: Extrapolation case

Note: The above distributions are derived from 100 repeated experiments. The lines positioned on each bar represent the median of each distribution.



(a) Random forest (b) XGBoost

Fig. 17. Correlation between the MPE and changes in the House Price Index

Note: The horizontal axis indicates the MPE estimated for each test set, and the vertical axis indicates the change rate in the actual House Price Index. The House Price Index is the repeat sales price index, published by the Korea Real Estate Board.

Table 1. Hyperparameters for the random forest simulation

Hyperparameter	Value	Description
$n_estimators$	200	The number of trees in the random forest. A higher number of trees generally improves accuracy but increases computation time.
$min_samples_leaf$	1	The minimum number of data points required in a group for it to be considered final (no further splitting). A value of 1 allows splitting until each group has only one data point.
$min_samples_split$	2	The minimum number of data points required to split a group into smaller ones. A split occurs only if a group has at least two points.
$max_features$	1	The number of variables considered at each split. A value of 1 means the model considers only one variable at a time when deciding how to split the data.
max_depth	None	The maximum number of times the data can be split. If set to "None," the model keeps splitting until it cannot improve further.
max_leaf_nodes	None	The maximum number of final groups (endpoints) allowed. "None" means there is no limit on how many final groups the model can create.

Note: This table presents the hyperparameters used for the random forest in the simulations. These hyperparameters are tuned to optimize predictive power.

Table 2. Variables used in real-world data analysis

Variable	Attribute	Description	Unit	Type
Transaction price	Dependent variable	Transaction price per square meter	Local currency (KRW)	Numerical
Property size	Structural attribute	Total floor area, including common-use space	m^2	Numerical
Property age		Year of transaction minus the year of construction	Year	Numerical
Property type		Three types of property: row house, multiplex house, apartment	Type	Categorical
Building materials		Materials used in the building's structure	Type of materials	Categorical
Number of households		Number of households in the complex	Number	Numerical
Parking space		Total parking spaces per household	Number	Numerical
Building coverage ratio		Percentage of the lot area covered by buildings	%	Numerical
Floor-to-area ratio		Ratio of total floor area to lot area	%	Numerical
Floor level		Floor level of the property	Number	Numerical
Property orientation		Orientation of the property	Cardinal direction	Categorical
Exclusive floor area		Area of exclusive-use space	m^2	Numerical
Elevator		Presence of an elevator	1 or 0	Indicator
Land ownership		Share of land ownership in the complex	m^2	Numerical
Loft structure		Presence of a loft	1 or 0	Indicator
Private garden		Presence of a private garden	1 or 0	Indicator
Residential brand		Brand of the constructor or developer for the apartment complex	Brand name	Categorical
Entrance elevation		Presence of a slope at the entrance	1 or 0	Indicator
Transaction date	Temporal attribute	Transaction date as an integer relative to the start date of the initial training sample (0 value)	Number	Numerical
Latitude	Locational attribute	Latitude in coordinates	Number	Numerical
Longitude		Longitude in coordinates	Number	Numerical
Administrative district		The smallest administrative unit in Seoul (252 in total), Dong	Dong	Categorical
Zoning district		Type of zoning district	Type	Categorical
Assigned elementary, middle, and high schools		The school assigned based on the property's address	School name	Categorical
Accessibility to subway station, elementary school, middle school, high school, and university		Euclidean distance to the nearest subway station, elementary school, middle school, high school, or university	m	Numerical

Note: This table presents the variables used in the real-world analysis. The dataset is obtained from the Korea Real Estate Board, the government institution responsible for assessing the property tax base in Korea.

Table 3. Descriptive Statistics for numerical variables

	Average	Standard deviation	Minimum	25th percentile	50th percentile	75th percentile	Maximum
Number of households	619.5	948.8	2	14	217	850	9,510
Property size	81.2	36.7	4.3	54.9	77.2	102.1	688.0
Property age	15.5	10.1	0	7	15	23	65
Parking space	0.8	1.0	0.0	0.2	0.8	1.2	92.4
Building coverage ratio	37.6	17.7	0.9	21.1	30.1	57.8	98.0
Floor-to-area ratio	207.7	154.5	0.0	144.8	212.1	272.4	1472.0
Exclusive floor area	67.3	30.5	4.3	46.8	60.0	84.9	431.6
Land ownership	34.8	19.2	0.0	23.9	31.8	41.9	4880.3
Floor level	6.9	5.9	-4	3	5	10	69
Access to elementary schools	378.2	201.7	18	224	344	494	1,974
Access to middle schools	508.5	290.1	20	290	455	664	2,000
Access to high schools	734.3	411.2	35	406	666	986	2,000
Access to universities	1119.6	487.0	26	725	1,112	1,523	2,000
Access to subway stations	485.5	213.8	19	324	459	636	1,000

Note: This table presents the statistics of the numerical variables. Accessibility measurements are limited to a maximum distance of 2000 meters (2 km). Negative values for floor level indicate underground floors. All values are rounded to the third decimal place.

Table 4. Estimated predictive performance

Panel A - Training sample: 2011–2013; Test sample: January 2014

Model		R^2	MAPE	COD	MPE	Hitting rates				
Algorithm	Type					10%	20%	30%	40%	50%
Random forest	Coordinates	0.953	6.768	6.694	0.444	84.88	95.14	97.48	98.33	98.64
	Coordinates + Locational variables	0.952	6.856	6.789	0.584	84.67	94.97	97.34	98.24	98.66
	Locational Variables	0.899	10.773	10.812	2.721	65.39	86.96	94.13	96.74	97.91
XGBoost	Coordinates	0.948	7.010	7.001	0.559	82.36	94.44	97.33	98.15	98.96
	Coordinate + Locational variables	0.949	7.009	7.000	0.794	82.77	94.45	97.16	98.22	98.93
	Locational Variables	0.940	7.853	7.873	2.121	80.27	92.18	96.04	97.59	98.56
Linear	Locational Variables	0.874	11.100	11.071	-1.047	61.18	87.36	95.6	97.69	98.69

Panel B - Training sample: 2012–2014; Test sample: January 2015

Model		R^2	MAPE	COD	MPE	Hitting rates				
Algorithm	Type					10%	20%	30%	40%	50%
Random forest	Coordinates	0.951	5.933	5.714	-1.064	85.4	96.16	98.38	99.12	99.52
	Coordinates + Locational variables	0.950	5.970	5.734	-1.095	85.6	96.03	98.46	99.22	99.56
	Locational Variables	0.888	9.462	9.497	0.348	67.52	89.29	95.7	98.02	98.93
XGBoost	Coordinates	0.943	6.081	5.977	-0.831	84.19	95.3	98.12	99.14	99.53
	Coordinate + Locational variables	0.945	5.913	5.801	-0.895	85.24	95.58	98.11	99.05	99.58
	Locational Variables	0.935	6.630	6.641	0.883	82.55	93.92	97.34	98.71	99.23
Linear	Locational Variables	0.840	10.612	10.677	-0.508	62.04	87.2	95.18	97.54	98.57

Panel C - Training sample: 2013–2015; Test sample: January 2016

Model		R^2	MAPE	COD	MPE	Hitting rates				
Algorithm	Type					10%	20%	30%	40%	50%
Random forest	Coordinates	0.943	6.721	6.660	-0.617	80.71	94.66	97.93	99.09	99.43
	Coordinates + Locational variables	0.944	6.751	6.629	-0.966	80.72	94.85	97.95	99.06	99.52
	Locational Variables	0.875	10.345	10.391	0.142	63.99	87.35	95.05	97.46	98.55
XGBoost	Coordinates	0.937	6.893	6.868	-0.645	79.29	93.79	97.84	98.95	99.51
	Coordinate +	0.939	6.675	6.646	-0.645	80.96	94.43	97.84	99.00	99.39

	Locational variables									
	Locational Variables	0.928	7.382	7.389	1.344	78.20	92.47	96.66	98.36	99.04
Linear	Locational Variables	0.837	11.933	11.939	2.109	57.97	84.48	93.18	96.47	98.03

Panel D - Training sample: 2014–2016; Test sample: January 2017

Model		R^2	MAPE	COD	MPE	Hitting rates				
Algorithm	Type					10%	20%	30%	40%	50%
Random forest	Coordinates	0.939	7.458	7.458	0.250	78.30	92.95	96.96	98.42	99.12
	Coordinates + Locational variables	0.938	7.549	7.514	-0.105	77.73	93.00	96.98	98.40	99.09
	Locational Variables	0.882	11.209	11.268	1.772	60.76	85.63	93.47	96.44	97.84
XGBoost	Coordinates	0.936	7.469	7.470	0.191	77.81	92.67	96.79	98.34	99.06
	Coordinate + Locational variables	0.936	7.421	7.420	0.156	78.02	92.51	96.80	98.37	99.06
	Locational Variables	0.925	8.156	8.159	1.871	76.32	91.12	95.72	97.35	98.32
Linear	Locational Variables	0.832	13.918	13.471	5.302	52.98	79.55	89.88	94.45	96.41

Panel E - Training sample: 2015–2017; Test sample: January 2018

Model		R^2	MAPE	COD	MPE	Hitting rates				
Algorithm	Type					10%	20%	30%	40%	50%
Random forest	Coordinates	0.949	7.109	6.491	-2.747	77.57	95.40	98.36	99.33	99.66
	Coordinates + Locational variables	0.941	7.476	6.727	-3.082	75.89	94.87	98.22	99.26	99.64
	Locational Variables	0.887	11.711	11.679	-0.750	57.20	84.58	93.78	96.96	98.28
XGBoost	Coordinates	0.948	6.959	6.812	-1.493	77.92	95.05	98.32	99.33	99.68
	Coordinate + Locational variables	0.948	6.983	6.849	-1.227	78.76	94.61	97.80	99.13	99.56
	Locational Variables	0.933	8.070	8.064	-0.067	74.09	92.67	96.90	98.39	99.11
Linear	Locational Variables	0.777	14.446	14.550	1.021	46.38	76.34	90.75	95.75	97.51

Panel F - Training sample: 2016–2018; Test sample: January 2019

Model		R^2	MAPE	COD	MPE	Hitting rates				
Algorithm	Type					10%	20%	30%	40%	50%
Random forest	Coordinates	0.927	10.035	10.088	1.197	66.59	87.86	95.38	97.63	98.55
	Coordinates + Locational variables	0.924	10.355	10.409	1.321	66.42	87.86	94.51	97.15	98.09

	Locational Variables	0.862	15.324	15.133	5.304	50.02	77.18	88.20	93.40	95.53
XGBoost	Coordinates	0.922	10.020	9.993	-0.748	64.75	88.27	95.02	97.78	98.84
	Coordinate + Locational variables	0.922	10.213	10.234	0.163	66.73	87.98	94.71	97.10	98.28
	Locational Variables	0.914	11.669	11.642	3.687	63.42	84.99	92.24	96.13	97.34
Linear	Locational Variables	0.772	23.016	19.339	15.429	34.50	60.15	75.56	84.28	89.94

Panel G - Training sample: 2017–2019; Test sample: January 2020

Model		R^2	MAPE	COD	MPE	Hitting rates				
Algorithm	Type					10%	20%	30%	40%	50%
Random forest	Coordinates	0.945	7.877	7.628	-1.392	76.12	93.55	97.53	98.79	99.28
	Coordinates + Locational variables	0.935	8.246	7.989	-1.261	75.83	92.53	96.81	98.48	98.95
	Locational Variables	0.854	13.835	13.909	3.970	56.29	80.43	89.50	93.85	96.28
XGBoost	Coordinates	0.939	7.934	7.879	-0.516	75.41	92.53	97.11	98.57	99.11
	Coordinate + Locational variables	0.939	7.975	7.926	-0.137	76.08	92.13	96.78	98.36	98.95
	Locational Variables	0.917	9.762	9.801	2.421	71.31	88.41	94.17	96.57	98.02
Linear	Locational Variables	0.751	20.418	18.475	11.416	38.07	65.39	80.15	87.59	91.88

Panel H - Training sample: 2018–2020; Test sample: January 2021

Model		R^2	MAPE	COD	MPE	Hitting rates				
Algorithm	Type					10%	20%	30%	40%	50%
Random forest	Coordinates	0.945	10.031	9.631	-2.928	64.29	88.70	95.71	98.17	99.06
	Coordinates + Locational variables	0.937	10.224	9.699	-3.266	63.53	88.42	95.43	98.14	98.94
	Locational Variables	0.882	14.210	14.369	0.227	50.50	78.37	89.86	94.83	97.11
XGBoost	Coordinates	0.939	9.868	9.581	-2.743	65.77	88.39	95.58	98.02	98.97
	Coordinate + Locational variables	0.937	9.882	9.669	-2.173	66.29	88.25	95.25	97.86	98.92
	Locational Variables	0.924	11.140	11.193	0.339	63.01	85.72	93.28	96.66	98.10
Linear	Locational Variables	0.791	22.424	23.032	4.735	28.46	59.03	77.94	86.62	91.44

Panel I - Training sample: 2019–2021; Test sample: January 2022

Model		R^2	MAPE	COD	MPE	Hitting rates				
Algorithm	Type					10%	20%	30%	40%	50%

Random forest	Coordinates	0.926	12.429	12.304	2.646	55.09	81.96	92.46	96.24	97.88
	Coordinates + Locational variables	0.912	12.204	12.203	1.554	54.47	83.30	92.97	96.64	97.91
	Locational Variables	0.861	15.524	15.213	4.521	47.92	73.74	86.60	93.80	96.62
XGBoost	Coordinates	0.926	11.658	11.694	1.292	57.85	84.46	93.42	96.67	98.36
	Coordinate + Locational variables	0.920	11.532	11.573	1.274	58.74	83.60	93.69	97.10	98.36
	Locational Variables	0.899	12.611	12.530	2.624	56.13	81.56	92.24	96.05	97.80
Linear	Locational Variables	0.737	28.207	22.554	19.758	28.00	51.38	68.35	77.58	84.35

Panel J - Training sample: 2020–2022; Test sample: January 2023

Model		R^2	MAPE	COD	MPE	Hitting rates				
Algorithm	Type					10%	20%	30%	40%	50%
Random forest	Coordinates	0.900	13.653	10.662	9.297	45.38	76.27	91.65	96.25	98.47
	Coordinates + Locational variables	0.893	14.297	10.839	10.168	43.59	74.22	91.10	96.29	98.30
	Locational Variables	0.805	21.248	15.430	16.108	31.57	60.25	79.16	89.18	92.76
XGBoost	Coordinates	0.946	10.099	9.863	2.183	62.89	88.37	94.84	98.04	98.98
	Coordinate + Locational variables	0.949	9.448	9.333	2.348	67.19	88.54	95.48	97.87	98.89
	Locational Variables	0.927	11.696	10.761	5.727	59.48	83.72	92.76	95.87	97.53
Linear	Locational Variables	0.587	35.172	16.831	32.582	15.17	30.72	50.62	67.02	79.25

Panel K - Training sample: 2021–2022; Test sample: January 2024

Model		R^2	MAPE	COD	MPE	Hitting rates				
Algorithm	Type					10%	20%	30%	40%	50%
Random forest	Coordinates	0.958	8.869	8.717	2.724	71.91	90.02	95.29	97.59	98.56
	Coordinates + Locational variables	0.957	9.266	9.083	3.309	70.70	89.27	94.73	97.31	98.44
	Locational Variables	0.917	14.069	13.414	7.237	58.14	79.78	88.48	92.99	95.50
XGBoost	Coordinates	0.957	8.673	8.660	1.330	71.93	90.13	95.84	98.07	99.07
	Coordinate + Locational variables	0.957	8.742	8.756	1.242	71.74	89.48	95.70	98.35	99.05
	Locational Variables	0.945	10.625	10.421	4.419	68.40	86.42	92.55	95.64	97.42
Linear	Locational Variables	0.869	16.541	16.266	3.122	43.79	71.19	85.07	92.45	95.98

Note: This table presents the predictive performance for each test set. All values are rounded to the third decimal place.