Enhancing real estate decision-making: a similarity-based approach for property valuation and recommendation

CHANGRO LEE

Kangwon National University, Department of Real Estate, South Korea; e-mail: spatialstat@naver.com

ABSTRACT Although similarity metrics have proven valuable in information retrieval and recommendation systems, their application to real estate has been limited. This study examines 1,014 benchmark land lots in Gwangjin-gu, Seoul, and identifies similar lots using Euclidean distance and cosine similarity. Nine land attributes, including site area, road width, and geographical descriptions are used to compute similarity. Being textual data, geographical descriptions are converted into numerical representations using an embedding model. It was found that the land lots identified as similar by Euclidean distance and cosine similarity are nearly identical, suggesting the effective applicability of both metrics to the real estate industry. In addition, we used industry standards to rigorously evaluate the performance of a similarity-based approach and demonstrated that it outperforms the traditional regression model. Our findings indicate that these metrics can enhance property valuation and recommendation by reducing subjectivity and increasing the efficiency of selecting similar land lots.

KEY WORDS similarity metrics – Euclidean distance – cosine similarity – property valuation – property recommendation – Seoul

LEE, CH. (2026): Enhancing real estate decision-making: a similarity-based approach for property valuation and recommendation. Geografie, 131, 1, 27–45.
https://doi.org/10.37040/geografie.2026.001
Received June 2025, accepted January 2026.

1. Introduction

Similarity metrics such as Euclidean distance and cosine similarity have been extensively applied across various industries. They have been useful for information retrieval, particularly in improving the relevance of search results by facilitating document comparisons (Wang, Dong 2020). The most notable commercial success is in their application to recommendation systems (Patel, Patel 2020), wherein these metrics suggest items such as books, movies, or dating matches.

Despite their potential applicability, similarity metrics have been understudied in the real estate sector. For instance, in property valuation, the sales comparison approach is commonly used to estimate prices. This method begins by selecting recently sold properties that match the subject closely in terms of size, zoning, location, and other characteristics. The selected properties, known as sales comparables, are then used to determine the value of the subject property. The accuracy of this valuation depends entirely on the relevance of the selected properties. Currently, practitioners select the sales comparables manually, which can lead to bias and inconsistency because the selection criteria may vary between practitioners. This subjective aspect of property selection is also present on real estate website platforms when candidate houses are recommended to users. This study aims to enhance the use of similarity metrics and reduce subjectivity in the real estate sector.

Using the concepts of Euclidean distance and cosine similarity, we identify similar lots from a sample of 1,014 benchmark land lots in Gwangjin-gu, Seoul. Nine attributes, including site area, road width, and geographical descriptions like "Near the southern side of the National Hospital," are used to compute similarity. Our findings demonstrate the effectiveness of applying similarity metrics to the real estate industry.

This study makes two key contributions. First, it comprehensively examines similar and dissimilar sales comparables and provides a nuanced perspective on how different metrics shape the process of selecting sales comparables. Second, using industry standards, the study evaluates the performance of instance-based valuation and demonstrates their potential as a robust alternative to conventional model-based valuation. By addressing these understudied areas, this study contributes to the development of more refined valuation and recommendation systems.

The remainder of this paper is structured as follows. Section 2 reviews similarity metrics and conversion techniques for textual data. Section 3 describes the dataset and methodology used in this study. Section 4 presents the results of the similarity metrics and discusses their implications for the real estate sector. Finally, Section 5 concludes the paper and suggests directions for future research.

2. Literature review

2.1. Similarity metrics

A similarity metric is used to quantify the similarity between two objects or data points. It quantifies the distance between points in a multidimensional space. Numerous similarity metrics have been developed in the literature; however, there is no universal similarity metric that is relevant to all tasks (Faisal, Zamzami 2020; Legendre, Legendre 2012). Therefore, a relevant metric must be selected by considering the characteristics of a specific task. Table 1 lists some commonly-used similarity metrics.

We used two similarity metrics: Euclidean distance and cosine similarity. Both metrics involve relatively straightforward calculations and provide intuitive interpretations (Leskovec, Rajaraman, Ullman 2020). In addition, both metrics are widely used in a broad range of fields, including natural language processing and recommendation systems (Fauzan et al. 2022; Singh et al. 2020).

Table 1 – Commonly-used similarity metrics

Name	Description	Usage
Euclidean distance	Straight-line distance between points in an n-dimensional space	Used in financial decision-making, demonstrating how Euclidean distance can be integrated into investment selection and evaluation problems (Merigo, Casanovas 2011).
Manhattan distance	Sum of absolute differences between coordinates, such as city block paths	Applied to recommendation systems for customer behavior analysis and personalized hotel recommendations (Uyanık, Orman 2023).
Canberra distance	Weighted version of Manhattan distance	Permata et al. (2025) employed Canberra distance within k-means clustering to segment customers and compared the results with those obtained through alternative distance metrics.
Chord distance	Similarity measure after normalizing vector lengths to unit lengths	Used to quantify dissimilarity between companies' accounting information as part of clustering for stock picking and investment analysis (Thrun 2022).
Hamming distance	Number of positions with different symbols (used in binary data)	Kuswardana et al. (2025) used Hamming distance to segment customers based on transaction attributes (e.g., payment method and order type).
Cosine similarity	Cosine of angle between vectors, indicating directional similarity	Albone (2024) applied cosine similarity to identify groups of customers with similar preferences and behaviors.
Jaccard similarity	Similarity of sets by comparing the intersection to union	Kosub (2019) demonstrated that Jaccard similarity satisfies the triangle inequality and, therefore, constitutes a valid metric on finite sets.
Dot product	Sum of element-wide multiplications, indicating magnitude and direction	Used in collaborative filtering and matrix factorization models to predict user preferences (Gunathilaka et al. 2025).

Euclidean distance is the straight-line distance between two vectors in a multidimensional space. The equation for calculating the Euclidean distance between two vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ is as follows:

$$d(\boldsymbol{a},\boldsymbol{b}) = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + \cdots + \ (a_n - b_n)^2} \tag{1}$$

It is a straightforward similarity metric in that it literally reflects the distance between each of the values of the vectors being compared; if the Euclidean distance is small, then the values of each coordinate in the vectors are close. This does not hold for cosine similarity in general. The Euclidean distance works well when the magnitude of the vectors and spatial relationships are important (Singh, Singh 2021). However, this method is sensitive to outliers and scale (Huang et all. 2020).

Cosine similarity is a measure of the angle between two vectors (Lahitani, Permanasari, Setiawan 2016). It is computed by taking the dot product of the vectors and dividing it by the product of their magnitudes, as follows:

$$similarity(\boldsymbol{a},\boldsymbol{b}) = \cos\theta = \frac{\boldsymbol{a}\cdot\boldsymbol{b}}{\|\boldsymbol{a}\|\|\boldsymbol{b}\|} \tag{2}$$

where a and *b* are the vectors being compared, "•" stands for the dot product, and ||a|| and ||b|| stand for the vectors' lengths. This metric is useful when the direction of the vectors is more important than their magnitude; unlike the Euclidean distance, it is robust to scale differences (Kryszkiewicz 2014). The metric is commonly used in text analysis and recommendation systems. However, it ignores vector magnitude, which may be crucial depending on the context.

This study addresses land lots, which are essentially tied to spatial relationships. Consequently, the use of Euclidean distance is necessary. However, land lots can differ significantly in attributes, such as price and size, necessitating a scale-robust metric, such as cosine similarity. Therefore, both metrics need to be utilized in the analysis, as they provide complementary perspectives on the relationships between land lots.

2.2. Use of similarity metrics in property valuation

Similarity metrics play a central role in property valuation, particularly in the sales comparison approach, which relies on identifying and adjusting comparable properties. Early studies have emphasized qualitative and quantitative techniques to operationalize similarities in valuation practices. Barańska (2009) proposed a two-step valuation framework in which comparable properties are first selected based on similarity metrics, followed by value estimation using those comparables. A key insight of this work is that valuation accuracy does not necessarily improve with an increasing number of attributes. Instead, focusing on a limited set of important characteristics can lead to more robust price predictions. This finding underscores the significance of attribute selection in sales comparison approach.

Building on traditional comparison methods, De Ruggiero and Salvo (2011) advanced the adjustment grid method by formalizing a similarity metric and introducing a reliability metric. Their approach quantifies not only how similar comparable properties are to the subject property, but also how reliable each comparable is within the valuation sample. By identifying anomalies and reducing their influence, these similarity and reliability metrics contribute to more consistent price estimates. This study highlights the reduction of subjectivity in valuation through a systematic similarity assessment.

Recent research has reflected a methodological shift toward data-driven similarity metrics. Li et al. (2023) employed neural networks to compute the similarity between geospatial elements, including linear features (e.g., rivers and roads) and polygonal features (e.g., land parcels and buildings). This research used Euclidean distance as its primary similarity metric. Its significance lies in how neural networks significantly reduced the need for manual feature engineering and data format conversion, thereby streamlining the comparison process.

Extending this trajectory, Renigier-Biłozor and Janowski (2024) proposed a human–machine synergy framework that conceptualizes property similarity across heterogeneous data types, including numerical attributes, textual descriptions, and images. By incorporating convolutional neural networks for visual data and large language models for text, their study broadens the notion of similarity beyond traditional tabular data. Collectively, these studies illustrate the evolution from rule-based similarity measures to AI-assisted approaches that enhance the scope and precision of property valuation.

As demonstrated in prior studies, the selection of sales comparables is a crucial initial step in determining property value. Currently, this selection is performed manually, introducing potential practitioner bias. Real estate platforms typically recommend properties to clients based on filters such as price and construction year, which could inadvertently exclude clients' preferences during the initial screening. This study reduces subjectivity and improves efficiency in the real estate sector by leveraging similarity metrics.

2.3. Conversion of textual data to numerical representations

Similarity metrics can only be applied to numerical data. It is difficult to apply them to textual data (qualitative data), such as customer reviews and property descriptions, because of the absence of a predefined format. Thus, similarity metrics have been primarily used for numerical data in outlier detection (Radovanović, Nanopoulos, Ivanović 2014) and time series analysis (Silva et al. 2018).

However, with the increasing popularity of large language models (LLMs), techniques for converting textual data into numerical representations have advanced rapidly. Numerous conversion models, that is, embedding models, are currently available. Embedding models transform high-dimensional textual data into low-dimensional vectors, thereby capturing semantic meanings and relationships (Li et al. 2020). In natural language processing, words or sentences are converted into dense vectors by these models that dramatically facilitate mathematical operations by mapping discrete data (such as words) into continuous vector spaces (Yang et al. 2014).

Term frequency–inverse document frequency (TF-IDF) is a statistical measure used to evaluate the importance of a word in a document relative to a collection of documents. It is one of the earliest models for textual representation that focuses on the frequency of terms while reducing the weight of commonly used words (Aizawa 2003, Ramos 2003). Following TF-IDF, word embedding models such as Word2Vec have emerged; they represent words in a continuous vector space and capture the semantic similarities between words by placing similar words closer together in the vector space. Word2Vec uses neural networks to learn word associations, thereby significantly improving the quality of text representations (Mikolov et al. 2013). More recently, transformers such as bidirectional encoder representations from transformers (BERT) and generative pre-trained transformers (GPT), have revolutionized natural language processing by using self-attention mechanisms to capture contextual relationships in texts more effectively (Rodrawangpai, Daungjaiboon 2022; Vaswani et al. 2017). Transformers can process entire sentences simultaneously, allowing for a better understanding and generation of the human language.

Textual data in the real estate industry, such as property descriptions in listings, offer valuable information, but have been underutilized because of the absence of structured data representations. This study converts descriptions of geographical locations into numerical representations using an embedding model, making them readily available for similarity metrics. Therefore, this study contributes to real estate literature by incorporating textual data to calculate property similarities, an approach rarely seen in the literature.

3. Data and method

3.1. Study area and dataset

Gwangjin-gu, one of 25 districts in Seoul, was selected as the study area. Gwangjin-gu is a typical district in Seoul with regard to land price levels and resident incomes. Figure 1 shows the study area and locations of the land lots.

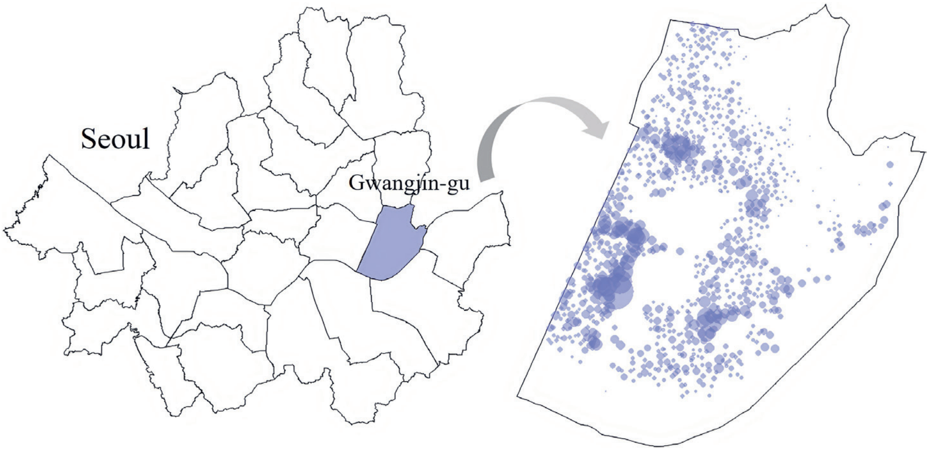

Fig. 1 – Study area and locations of land lots. The circle sizes are proportional to their unit prices.

The circle sizes in the figure are proportional to the unit prices of the land lots. Expensive land lots are concentrated in the western part of the study area, which corresponds to the main boulevard. The sparsely populated areas correspond to preserved and mountainous areas.

The land lots shown in Figure 1 are benchmark lots. The benchmark lot data were obtained from the Ministry of Land, Infrastructure, and Transport (MOLIT).[1] MOLIT conducts annual surveys of these benchmark lots, making the results publicly available for stakeholder use in transactions and investments, as well as for monitoring real estate market trends. The price data were taken from 2022 surveys by licensed property appraisers. While the data encompass price information and various land characteristics such as lot area and zoning, they lack information on the distance to subway stations, which is important for property valuation and recommendation. This information was collected separately[2] and added to the dataset. Table 2 shows the summary statistics of the 1,014 benchmark land lots.

Based on the median values from the table, a typical lot in Gwangjin-gu can be described as a land lot with an area of 180.0 m^2 and probable price of 2,600,000 KRW/m^2 (approximately 1,900 USD/m^2). This lot is most likely located in the second residential zone, adjacent to a 7-meter-wide road. It is situated 466.9 and 581.4 meters from the nearest subway station and railway, respectively.

The dataset includes the description of the benchmark lots' geographical locations. Table 3 presents examples of these descriptions. As shown in the table, it provides valuable information about the accessibility and connectivity of a lot,

[1] The data are publicly available at https://www.data.go.kr/data/15004246/fileData.do.

[2] Subway station data are available at https://data.kric.go.kr/rips/M_01_01/detail.do?id=32.

Table 2 – Summary statistics (n = 1,014)

	Min.	Median	Mean	Max.
Unit price (KRW/ m²)	20,500	2,600,000	3,120,000	21,300,000
Site area (m²)	69.0	180.0	584.2	69,292.0
Road width (m)	0 (no road)	7	11	32
Distance to station (m)	27.2	466.9	476.4	1,458.8
Distance to railway (m)	5.5	581.4	486.7	640.0
Zoning (number of lots, proportion)	open space zone (14, 1.4%), 1st residential zone (118, 11.6%), 2nd residential zone (590, 58.2%), 3rd residential zone (168, 16.6%), quasi-residential zone (85, 8.4%), commercial zone (39, 3.8%)			

Table 3 – Examples of the description of geographical location

Lot #	Description of geographical location
6	Near the southern side of the National Hospital.
37	Near the northern side of Junggok Traditional Market.
429	Near the northern side of the Gwangjin Post Office.
845	On the southwest side of Sejong University.
997	On the northern side of Konkuk University Station (Subway line 2).

which can influence its price. For example, the description of land lot 997 suggests that it is easily accessible from the main transportation hub (Konkuk University Station). Therefore, excluding this information could result in an incomplete comparison of land characteristics. This study adds "description of geographical location" to the list of comparisons by converting it into numerical form.

3.2. Methodology

Nine variables are used as comparison components: unit price, site area, road width, distance to station, distance to railway, geographical coordinates X and Y, zoning, and description of geographical location.[3] All variables are numerical data

[3] In the Korean real estate market, factors such as price, road width, and zoning are crucial for land valuation and were included in this study. Attributes such as distance to transportation hubs were included based on available data. While additional factors like school district reputation and foot traffic could enhance valuation accuracy, they were omitted due to data limitations. The selection of these nine attributes balances valuation practices with data availability, providing a robust framework for analyzing property similarities within the scope of this study.

except for zoning and description of geographical location: zoning is categorical, and description of geographical location is textual. To facilitate the calculations, zoning is encoded based on the price levels in each zone as follows: open space zone is coded as 1, 1st residential zone as 2, 2nd residential zone as 3, 3rd residential zone as 4, quasi-residential zone as 5, and commercial zone as 6. This coding generally aligns with average price levels in each zone.

BERT is used to convert the textual data (description of geographical location). BERT is a transformer-based model designed to pre-train deep bidirectional representations by considering both the left and right contexts in all layers (Devlin et al. 2019). Although early embedding models like Word2Vec and TF-IDF have their merits, particularly in scenarios with simpler text data, BERT's ability to capture contextual and semantic information makes it better suited for this study. BERT ensures that textual data is meaningfully incorporated into similarity metrics. Unlike earlier models, which may struggle with rare words or polysemy, BERT can effectively handle these challenges by leveraging contextual embeddings. BERT has several variants, and a relevant variant is selected considering the performance and computational burden. The BERT variant selected in this study has 12 layers (transformer blocks), 768 hidden units per layer, and 12 attention heads.[4] Owing to the 768 hidden units per layer, the output embedding vector has 768 dimensions, resulting in a 1,014 × 768 matrix.

The 768-dimensional embedding vector can be directly fed into the calculation for similarity metrics. However, to enhance computational efficiency, the embedding vector is reduced to smaller dimensions using principal component analysis (PCA). PCA statistically converts a large number of dimensions into smaller dimensions that continue to contain most of the larger sets' information (Jolliffe 2002). By reviewing the variance explained by each principal component, the number of principal components is set to 10, which explains 65.1% of total variance.

Euclidean distance and cosine similarity are then applied to the nine variables to compute the similarity values among the land lots.

4. Results

4.1. Specifications of the reference land lot

To calculate the similarity values between the data points, a reference data point needs to be first established. The reference point serves as a standard against

[4] Specifically, the BERT-base-uncased version is used for analysis. The "base" version of BERT has a total of 110 million parameters, and "uncased" means that the text is converted to lowercase before tokenization. This version balances performance and computational efficiency.

Table 4 – Specifications of the reference land lot

Unit price (KRW/ m^2)	Site area (m^2)	Road width (m)	Distance to station (m)
2,600,000	180.0	7	466.9
Distance to railway (m)	Zoning	Geographical coordinate X	Geographical coordinate Y
581.4	2nd residential zone	207,635.142	449,023.161
Description of geographical location			
[–0.1339, –0.7720, –0.0329, 0.0001, 0.0134, –0.1375, 0.0387, –0.0308, 0.0418, –0.0333]			

which similarity values for all other data points are computed. Table 4 presents the specifications of a reference lot in Gwangjin-gu using the median values for each variable. For example, the unit price in the table represents the median price of the 1,014 land lots included in this study. The most frequently observed description was selected to describe the geographical location of the reference land lot. We use this imaginary lot as the reference land lot, which serves as the standard for comparison.

4.2. Comparison of the reference land lot with comparables

Table 5 shows the results when Euclidean distance is employed for comparison: five most similar and dissimilar comparables. In Euclidean distance, a value close to zero means that the two vectors being compared are very similar, while the larger the value, the more different the vectors.

The most similar comparables chosen are close to the reference lot for all the attributes presented in the table. In particular, the categorical variable zoning is identical for both the reference lot and five comparables, all of which are in the 2nd residential zone. Conversely, the most dissimilar comparables differ significantly from the reference lot for most of the attributes shown in the table.

Figure 2 illustrates the locations of the comparables presented in Table 5. The five most similar comparables are positioned close to the reference lot, whereas the most dissimilar comparables are located at a relatively greater distance. This outcome was expected because geographical coordinates X and Y were employed as a component of the comparison. Using geographical coordinates to calculate the Euclidean distance ensures geographical proximity to the reference lot.

Table 6 shows the results when cosine similarity is employed for comparison: five most similar and dissimilar comparables. In cosine similarity, a value close to one means that the two vectors being compared are very similar, whereas a value around zero indicates that the vectors are dissimilar. Cosine similarity can have negative values, and a value close to –1 indicates that the vectors are opposites.

Table 5 – Comparables selected by Euclidean distance

Land lot	Unit price (KRW/m²)	Site area (m²)	Road width (m)	Dist. to Stn. (m)	Dist. to Rw. (m)	Zoning	Similarity value
Reference	2,600,000	180.0	7	466.9	581.4	2nd resi.	0.00
Most similar comparables	3,040,000	155.7	6	569.4	614.8	2nd resi.	0.64
	2,490,000	185.8	6	587.6	564.7	2nd resi.	0.69
	2,110,000	191.7	6	476.7	562.3	2nd resi.	0.71
	2,140,000	145.5	7	500.0	603.6	2nd resi.	0.77
	2,270,000	113.4	6	572.1	599.1	2nd resi.	0.78
Most dissimilar comparables	5,050,000	35,977.0	32	182.4	2.1	quasi-resi.	11.07
	21,300,000	889.9	32	91.7	30.3	commercial	12.41
	28,500	52,407.0	0	1,038.4	639.9	open space	14.75
	35,500	57,960.0	10	887.7	587.4	open space	16.10
	20,500	69,292.0	0	693.3	588.6	open space	19.08

Note: Dist. to Stn., Dist. to Rw., 2nd resi., and quasi-resi. indicate the distance to station, distance to railway, 2nd residential zone, and quasi-residential zone, respectively. Geographical coordinates *X* and *Y*, and description of geographical location are not shown for brevity.

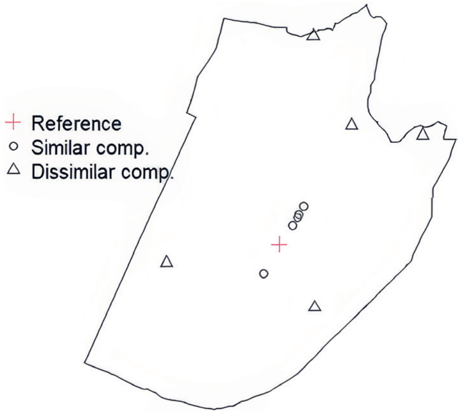

Fig. 2 – Locations of comparables selected by Euclidean distance

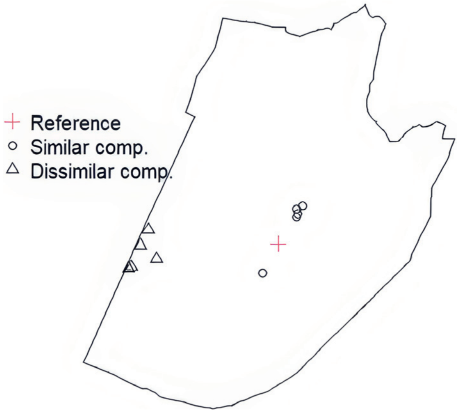

Fig. 3 – Locations of comparables selected by cosine similarity

In short, cosine similarity ranges from –1 to 1, and the larger the value, the more similar the vectors.

As with Euclidean distance, the most similar comparables exhibit attributes similar to the reference lot, as illustrated in the table. The stars in the last column indicate comparables that were also detected using Euclidean distance. The most dissimilar comparables are notably different from the reference lot for most attributes shown in the table, similar to Euclidean distance. However, dissimilar

Table 6 – Comparables selected by cosine similarity

Land lot	Unit price (KRW/m²)	Site area (m²)	Road width (m)	Dist. to Stn. (m)	Dist. to Rw. (m)	Zoning	Similarity value
Reference	2,600,000	180.0	7	466.9	581.4	2nd resi.	1.000
Most similar comparables	1,740,000	245.0	6	540.1	626.5	2nd resi.	0.885
	2,140,000	145.5	7	500.0	603.6	2nd resi.	*0.882
	2,110,000	191.7	6	476.7	562.3	2nd resi.	*0.881
	2,270,000	113.4	6	572.1	599.1	2nd resi.	*0.877
	3,040,000	155.7	6	569.4	614.8	2nd resi.	*0.870
Most dissimilar comparables	5,550,000	762.7	32	551.6	314.4	3rd resi.	–0.753
	5,560,000	916.7	28	268.2	5.8	3rd resi.	–0.766
	4,620,000	211.9	25	510.2	328.9	3rd resi.	–0.801
	5,330,000	375.5	32	502.2	16.7	3rd resi.	–0.813
	4,200,000	209.6	28	621.3	282.1	3rd resi.	–0.822

Note: Dist. to Stn., Dist. to Rw., 2nd resi., and 3rd resi. indicate the distance to station, distance to railway, 2nd residential zone, and 3rd residential zone, respectively. Geographical coordinates *X* and *Y*, and description of geographical location are not shown for brevity.

comparables appear to share common attributes. For instance, their unit prices range from 4,000,000 to 5,500,000 KRW, and they are all in the 3rd residential zone.

Figure 3 shows the locations of the comparables listed in Table 6. As with Euclidean distance, the five most similar comparables are situated close to the reference lot, while the most dissimilar comparables are further west of the reference lot.

4.3. Evaluating the performance of Euclidean distance and cosine similarity

Relying on a single reference land lot is insufficient for a definitive evaluation of the effectiveness of similarity metrics. This study compares the performance of Euclidean distance and cosine similarity in land price estimation using the k-nearest neighbor (KNN) method. The process involves two steps. First, the prices of land lots are estimated by selecting the k most comparable lots using Euclidean distance calculations and averaging their prices. Second, the same procedure is repeated using cosine similarity calculations. Subsequently, the performance of both metrics is evaluated using the mean absolute percentage error (MAPE) and coefficient of dispersion (COD), which are widely recognized criteria for assessing valuation accuracy (IAAO 2013; Numan, Yusoff 2024).

The MAPE represents the average percentage difference between the predicted and observed prices, with lower values indicating higher accuracy. Although no

Table 7 – Performance comparison of KNN and regression models on the test dataset

Evaluation criterion	KNN with Euclidean distance	KNN with cosine similarity	Regression model
MAPE	18.5	17.1	25.9
COD	19.1	17.3	26.3

universal threshold exists for MAPE, Numan and Yusoff (2024) reported that values between 10% and 20% are generally considered acceptable for practical application. COD assesses the consistency of price estimates, with a lower COD signifying more reliable valuation. According to IAAO (2013) guidelines, COD values below 20% and 25% are acceptable for urban and rural land lots, respectively. The formulas for MAPE and COD are as follows:

$$MAPE = \frac{1}{n}\sum\left|\frac{prediced\ price - observed\ price}{observed\ price}\right| \times 100 \tag{3}$$

$$COD = \frac{1}{n}\sum\frac{|ratio - median\ ratio|}{median\ ratio} \times 100, where\ ratio = \frac{prediced\ price}{observed\ price} \tag{4}$$

The dataset of 1,014 land lots is divided into training (70%, 709 lots) and test (30%, 305 lots) data. The performance of the two similarity metrics is assessed using the test dataset. For comparison, a standard regression model is also applied, which uses the unit price of each land lot as the dependent variable and the comparison components from the similarity calculations as independent variables.[5] The results of the KNN and regression models for the test dataset are presented in Table 7.[6]

The results indicate that KNN with cosine similarity outperforms KNN with Euclidean distance, while the standard regression model performs the least effectively. To visually confirm these findings, Figure 4 illustrates the goodness-of-fit of the three models. The similarity-based models (Euclidean distance and cosine similarity) show predicted prices that closely align with the observed prices, whereas the standard regression model exhibits a greater deviation from the observed values.

While the relatively small sample size of 305 land lots prevents a definitive conclusion about which similarity metric is more effective for land price estimation, the results suggest that the similarity-based approach outperforms traditional regression models. Moreover, the performance of this approach meets industry standards, as evidenced by criteria such as the MAPE and COD.

[5] They include site area, road width, distance to station, distance to railway, geographical coordinates X and Y, zoning, and description of geographical location.

[6] For KNN, the parameter k was set to 5 based on performance assessments during KNN training.

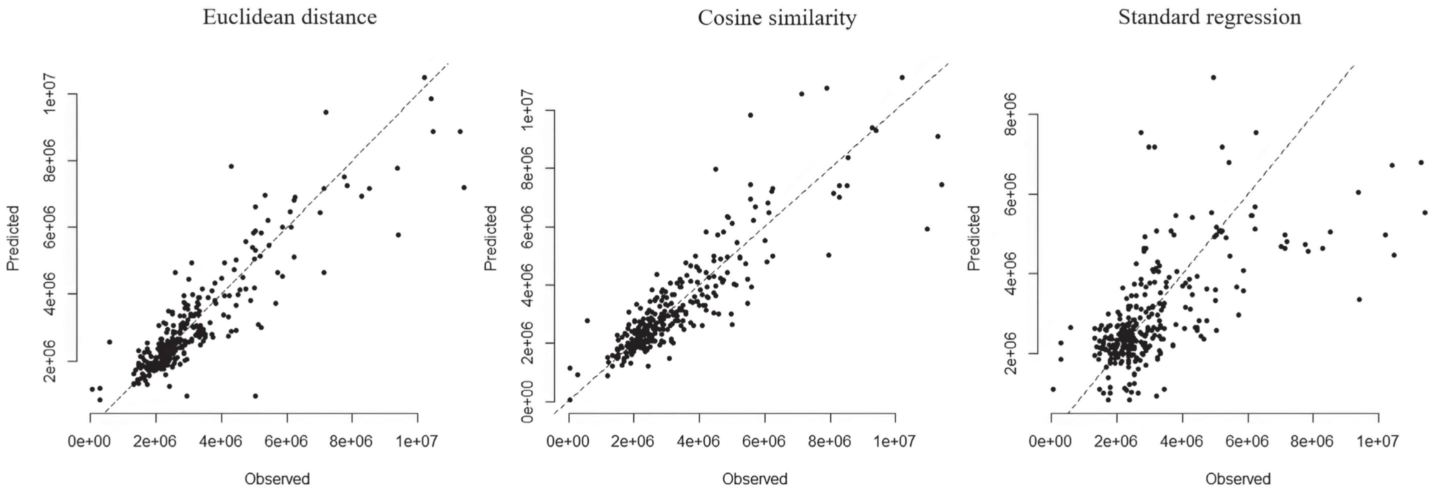

Fig. 4 – Goodness-of-fit for KNN and regression models on the test dataset

5. Discussion

The most similar comparables selected by the two similarity metrics are almost identical, with four of the five overlapping (Tables 5 and 6). However, the most dissimilar comparables identified by the two metrics are distinct, with no overlap. When using Euclidean distance, the most dissimilar comparables are significantly different from the reference lot for various attributes. This difference is more apparent in geographical coordinates. In Figure 2, dissimilar comparables are located far from the reference land lot and scattered across the study area. This is because Euclidean distance penalizes large differences both in magnitude and direction, leading to a dispersed distribution of dissimilar comparables.

By contrast, cosine similarity measures the cosine of the angle between two vectors, focusing solely on direction, not magnitude. The most dissimilar comparables selected by cosine similarity (as with Euclidean distance) differ from the reference lot for various attributes. However, the cosine similarity resulted in a clustered set of dissimilar comparables. As shown in Table 6, they share three common attributes: price range (4,000,000–5,500,000 KRW/m^2), road width (25–32 m), and zoning (3rd residential zone). This pattern is more apparent in the geographical coordinates. As shown in Figure 3, the comparables are clustered in one corner of the study area. The comparables that are directionally different from the reference lot are considered dissimilar even if they are not necessarily the farthest in terms of actual magnitude. This led to dissimilar comparables being grouped into a particular area with similar directional deviations from the reference lot. These findings offer insight into the behavioral differences between the two metrics and highlight the importance of metric selection in spatial analyses.

Beyond the analysis based on a single reference land lot, this study rigorously evaluated the performance of a similarity-based approach using MAPE and COD. The similarity-based approach surpasses the traditional regression model and aligns with industry benchmarks. This empirical evidence reinforces the superiority of instance-based valuation over model-based valuation. Model-based valuation, such as linear regression, relies on learning explicit functions or parameters during training, which may introduce limitations owing to the underlying assumptions. However, instance-based valuation, such as KNN, operates by directly comparing new instances to stored examples, avoiding the need for rigid assumptions (e.g., linearity or homogeneity). Their flexibility makes them particularly robust in scenarios where data relationships are complex. Thus, this study advocates for the broader adoption of instance-based valuation in mass appraisal and automated valuation systems, particularly in contexts where land prices are influenced by intricate and localized factors.

The application of similarity metrics extends beyond valuation to the enhancement of property recommendation systems. Current real estate platforms typically employ sequential filter-based searches, in which users incrementally apply criteria such as price or size. This approach risks the premature exclusion of properties that may align with user preferences. By leveraging similarity scores to present candidates holistically – considering all attributes simultaneously – the likelihood of overlooking ideal properties in the early search stages is significantly reduced. This shift could improve user satisfaction and the efficiency of property matching.

We calculated the similarity values using nine comparison components: unit price, site area, road width, distance to station, distance to railway, geographical coordinates X and Y, zoning, and description of geographical location. Each component was assigned an equal weight to calculate its similarity value. However, the weights can be adjusted for each component to integrate domain-specific knowledge. For example, in South Korean land valuation, road width and zoning are typically considered critical factors in determining land prices; hence, these two components may be given higher weights than the others. Therefore, for Euclidean distance, Equation (1) can be modified as follows:

$$d_w(\boldsymbol{a}, \boldsymbol{b}) = \sqrt{w_1(a_1 - b_1)^2 + w_2(a_2 - b_2)^2 + \cdots + \ w_n(a_n - b_n)^2} \tag{5}$$

where denotes the weight assigned to each component. Weights can be derived from expert surveys, Bayesian estimation, or correlations between attributes and prices. This adaptability enables practitioners to tailor similarity metrics to sector-specific requirements and enhance the accuracy in diverse applications.

This study bridges two critical gaps in the literature. First, it provides a comprehensive analysis of both similar and dissimilar comparables, offering a nuanced understanding of the influence of different metrics on comparable selection. Second, it rigorously evaluates the performance of instance-based valuation against industry standards, highlighting their potential as a viable alternative to traditional model-based valuation. By addressing these understudied aspects, this study paves the way for more refined valuation and recommendation systems.

6. Conclusion

Although similarity metrics have significant potential, their application to the real estate sector has been largely overlooked. This study shows that these metrics can be leveraged for property valuation and recommendation. By analyzing 1,014 benchmark land lots in Gwangjin-gu, Seoul, and using nine key land characteristics (including geographical descriptions converted into numerical representations via an embedding model), we found that Euclidean distance and cosine similarity metrics yielded nearly identical results for the most similar comparables. This suggests that both the metrics can be effectively employed in the real estate industry. However, the most dissimilar comparables identified by these metrics differed.

Our findings indicate that similarity metrics can enhance property valuation and recommendation by reducing the subjectivity involved in selecting sales comparables and minimizing the risk of missing a user's preferred property during the initial search. Furthermore, these metrics can be tailored to align with domain-specific knowledge by assigning different weights to each land attribute, thereby increasing their relevance across various real estate sectors.

The dataset used consists of 1,014 benchmark land lots, which may not be sufficiently large to capture the full diversity of real estate characteristics. A larger dataset could provide more generalizable results. Additionally, the focus on a single district, Gwangjin-gu in Seoul, limits the applicability of the findings to other regions with potentially different market dynamics. Future research should consider expanding the dataset size and geographical scope to include diverse urban and rural areas, thereby enhancing the generalizability of similarity metrics in the real estate industry.

References

AIZAWA, A. (2003): An information-theoretic perspective of TF-IDF measures. Information Processing & Management, 39, 1, 45–65. https://doi.org/10.1016/S0306-4573(02)00021-3

ALBONE, A. (2024): Building customer and product networks with cosine similarity in graph analytics for deep customer insight. Engineering, MAthematics and Computer Science Journal, 6, 3, 215–218. https://doi.org/10.21512/emacsjournal.v6i3.11693

BARAŃSKA, A. (2009): Qualitative and quantitative methods for assessing the similarity of real estate. Value in the process of real estate management and land administration. Towarzystwo Naukowe Nieruchomości, Olsztyn, 31–42.

DE RUGGIERO, M., SALVO, F. (2011): Misure di similarità negli adjustment grid methods. Aestimum, 58, 1, 47–58.

DEVLIN, J., CHANG, M.W., LEE, K., TOUTANOVA, K. (2019): BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

FAISAL, M., ZAMZAMI, E.M. (2020): Comparative analysis of inter-centroid k-means performance using Euclidean distance, Canberra distance and Manhattan distance. Journal of Physics: Conference Series, 1566, 1, pp. 012112). IOP Publishing. https://doi.org/10.1088/1742–6596/1566/1/012112

FAUZAN, R., LABIB, M.I.A., JOHANNIS, J.O.T., NOOR, S. (2022): Semantic similarity of Indonesian sentences using natural language processing and cosine similarity. In 2022 4th International Conference on Cybernetics and Intelligent System, IEEE, pp. 1–5. https://doi.org/10.1109/ICORIS56080.2022.10031416

GUNATHILAKA, T.M.A.U., MANAGE, P.D., ZHANG, J., LI, Y., KELLY, W. (2025): Addressing sparse data challenges in recommendation systems: A systematic review of rating estimation using sparse rating data and profile enrichment techniques. Intelligent Systems with Applications, 200474. https://doi.org/10.1016/j.iswa.2024.200474

HUANG, R., CUI, C., SUN, W., TOWEY, D. (2020). Poster: Is Euclidean distance the best distance measurement for adaptive random testing? In 2020 IEEE 13th International Conference on Software Testing, Validation and VerificationIEEE, pp. 406–409. https://doi.org/10.1109/ICST46399.2020.00049

IAAO (2013): Standard on ratio studies. Kansas City, MO: International Association of Assessing Officers.

JOLLIFFE, I.T. (2002): Principal component analysis for special types of data, New York: Springer, 338–372.

KOSUB, S. (2019): A note on the triangle inequality for the Jaccard distance. Pattern Recognition Letters, 120, 36–38. https://doi.org/10.1016/j.patrec.2018.12.007

KRYSZKIEWICZ, M. (2014): The Cosine similarity in terms of the Euclidean distance. In Encyclopedia of Business Analytics and Optimization. IGI Global, 2498–2508. https://doi.org/10.4018/978-1-4666-5202-6.ch223

KUSWARDANA, D.A., PRASETYA, D.A., TRIMONO, T., DIYASA, I.G.S.M., AWANG, W.S. W. (2025): Customer transaction clustering with k-prototype algorithm using Euclidean-Hamming distance and elbow method. International Journal of Advances in Data and Information Systems, 6, 2, 259–275. https://doi.org/10.59395/ijadis.v6i2.1381

LAHITANI, A.R., PERMANASARI, A.E., SETIAWAN, N.A. (2016): Cosine similarity to determine similarity measure: Study case in online essay assessment. In 2016 4th International Conference on Cyber and IT Service Management, IEEE, 1–6. https://doi.org/10.1109/CITSM.2016.7577578

LEGENDRE, P., LEGENDRE, L. (2012): Ecological resemblance. In Developments in Environmental Modelling, 24, 265–335. https://doi.org/10.1016/B978-0-444-53868-0.50007-1

LESKOVEC, J., RAJARAMAN, A., ULLMAN, J.D. (2020): Mining of massive data sets. Cambridge University Press. https://doi.org/10.1017/9781108684163

LI, B., ZHOU, H., HE, J., WANG, M., YANG, Y., LI, L. (2020): On the sentence embeddings from pre-trained language models. arXiv preprint arXiv:2011.05864. https://doi.org/10.18653/v1/2020.emnlp-main.733

LI, P., YAN, H., LU, X. (2023): A Siamese neural network for learning the similarity metrics of linear features. International Journal of Geographical Information Science, 37, 3, 684–711. https://doi.org/10.1080/13658816.2022.2143505

MERIGO, J.M., CASANOVAS, M. (2011): Induced aggregation operators in the Euclidean distance and its application in financial decision making. Expert Systems With Applications, 38, 6, 7603–7608. https://doi.org/10.1016/j.eswa.2010.12.103

MIKOLOV, T., CHEN, K., CORRADO, G., DEAN, J. (2013): Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

NUMAN, J.A., YUSOFF, I. M. (2024): Identifying the current status of real estate appraisal methods. Real Estate Management and Valuation, 32, 4, 12–27. https://doi.org/10.2478/remav-2024-0032

PATEL, K., PATEL, H.B. (2020). A state-of-the-art survey on recommendation system and prospective extensions. Computers and Electronics in Agriculture, 178, 105779. https://doi.org/10.1016/j.compag.2020.105779

PERMATA, R.P., ALIFAH, A.N., SANJAYA, I.M.W.A. (2025): Optimizing k-means clustering through distance metric simulation for strategic enrollment segmentation in private universities. CAUCHY: Jurnal Matematika Murni dan Aplikasi, 10, 2, 616–629. https://doi.org/10.18860/cauchy.v10i2.33089

RADOVANOVIĆ, M., NANOPOULOS, A., IVANOVIĆ, M. (2014): Reverse nearest neighbors in unsupervised distance-based outlier detection. IEEE Transactions on Knowledge and Data Engineering, 27, 5, 1369–1382. https://doi.org/10.1109/TKDE.2014.2365790

RAMOS, J. (2003): Using TF-IDF to determine word relevance in document queries. In Proceedings of the First Instructional Conference on Machine Learning, 242, 1, 29–48.

RENIGIER-BIŁOZOR, M., JANOWSKI, A. (2024): Human-machine synergy in real estate similarity concept. Real Estate Management and Valuation, 32, 2, 13–30. https://doi.org/10.2478/remav-2024-0010

RODRAWANGPAI, B., DAUNGJAIBOON, W. (2022): Improving text classification with transformers and layer normalization. Machine Learning With Applications, 10, 100403. https://doi.org/10.1016/j.mlwa.2022.100403

SILVA, D.F., GIUSTI, R., KEOGH, E., BATISTA, G.E. (2018): Speeding up similarity search under dynamic time warping by pruning unpromising alignments. Data Mining and Knowledge Discovery, 32, 988–1016. https://doi.org/10.1007/s10618-018-0557-y

SINGH, R.H., MAURYA, S., TRIPATHI, T., NARULA, T., SRIVASTAV, G. (2020): Movie recommendation system using cosine similarity and KNN. International Journal of Engineering and Advanced Technology, 9, 5, 556–559. https://doi.org/10.35940/ijeat.E9666.069520

SINGH, R., SINGH, S. (2021): Text similarity measures in news articles by vector space model using NLP. Journal of The Institution of Engineers (India): Series B, 102, 329–338. https://doi.org/10.1007/s40031-020-00501-5

THRUN, M.C. (2022): Exploiting distance-based structures in data using an explainable AI for stock picking. Information, 13, 2, 51. https://doi.org/10.3390/info13020051

UYANIK, B., ORMAN, G.K. (2023): A Manhattan distance-based hybrid recommendation system. International Journal of Applied Mathematics Electronics and Computers, 11, 1, 20–29. https://doi.org/10.18100/ijamec.1232090

VASWANI, A., SHAZEER, N., PARMAR, N., USZKOREIT, J., JONES, L., GOMEZ, A.N., POLOSUKHIN, I. (2017): Attention is all you need. Advances in Neural Information Processing Systems, 30.

WANG, J., DONG, Y. (2020): Measurement of text similarity: A survey. Information, 11, 9, 421. https://doi.org/10.3390/info11090421

YANG, B., YIH, W. T., HE, X., GAO, J., DENG, L. (2014): Embedding entities and relations for learning and inference in knowledge bases. arXiv preprint arXiv:1412.6575.

DATA AVAILABILITY

The data used in this study are publicly available and can be accessed at https://www.data.go.kr/data/15004246/fileData.do and https://data.kric.go.kr/rips/M_01_01/detail.do?id=32.

ORCID

CHANGRO LEE
https://orcid.org/0000-0002-7727-3168