부동산학연구 제30집 제4호, 2024. 12, pp. 143~162 http://dx.doi.org/10.19172/KREAA.30.4.8

Journal of the Korea Real Estate Analysts Association
Vol.30, No.4 2024. 12, pp. 143~162

Regionalization of Retail Market Using Real Estate Transaction Data

부동산 거래정보를 이용한 리테일 상권 설정

민 성 훈 (Min, Seonghun)*

< Abstract >

Establishing appropriate spatial districts is critical in understanding the retail market because market data is surveyed based on them. In general, retail districts have been established from the perspective of usage such as the demand of customers and the supply of retailers. Conversely, this study focused on the transaction of real estate and drew implications by comparing these two approaches. This study analysed statistics of real estate transaction in Seoul using the spatial fuzzy C-means algorithm which is a spatial clustering technique based on machine learning and compared the results with the generally accepted retail districts stemmed from the perspective of usage. The comparison indicates that: First, the transaction-based approach distinguished core and neighbouring retail districts successfully, and the locations of the core retail districts were similar to those of the usage-based approach. Therefore, retail districts could be reliably established using transaction data. Second, despite the similarity in the locations, significant differences were found between the boundaries of the retail districts according to the two approaches. Furthermore, the spatial changes over time, such as the expansion or contraction of retail districts, were identified from the differences. Third, the increase in the prices of retail districts focusing on the transaction was significantly larger than that of retail districts focusing on the usage. This difference indicates that statistical bias in the retail market can occur depending on whether the regionalization is based on usage or transaction.

Keyword : Retail Districts; Regionalization; Machine Learning; Spatial Fuzzy C-Means; Price Index

I. Introduction

Retail is an essential function that constitutes a city and a sector that occupies a large part of the real estate market. The retail related spatial scope can be defined from two perspectives: the central place and the hinterland. The central place refers to areas where retail properties are gathered, while the hinterland refers to surrounding areas. In general, the former is called retail district or commercial district, and the latter is called market area, trade area, or catchment area. This study focuses on the former.

Numerous institutions have been investigating

* Regular Member of the Society, Professor in the Department of Urban Planning and Real Estate, The University of Suwon, smin@suwon.ac.kr

and publishing the conditions of retail districts. However, the statistics and analytical results have varied depending on how the boundaries of retail districts were established. Therefore, information providers on retail districts strive to reflect the market accurately by establishing retail districts that consider both supply aspects, such as the number of stores, and demand aspects, such as foot traffic.

However, this approach is limited in focusing only on the usage aspect of real estate and overlooks the transaction aspect. For a comprehensive understanding of the retail market, not only customers and retailers but also investors should be considered.

As the value of real estate is based on the degree of activation of the retail market, retail districts established based on usage and those based on transaction cannot be fundamentally different. However, the boundaries of retail districts based on the two approaches may not be commensurate for the following reasons. First, there may be a time difference between the activation of a retail district and the reflection of such activation in real estate transaction. Second, since real estate transaction reflects the expectation of investors, the boundaries based on price and transaction volume may differ from the boundaries where customers and retailers are currently gathered.

The purpose of this study is to demonstrate the necessity of incorporating transaction aspect along with usage aspect when defining retail districts. To verify this, the study establishes retail districts based solely on transaction data and examines how they differ from traditional usage-based retail districts, as well as whether these differences offer meaningful implications. If the results are favorable, it would suggest that a balanced approach, integrating both perspectives, should be adopted.

Consequently, spatial clustering is performed using real estate transaction data published by the Ministry of Land, Infrastructure, and Transport (MOLIT). Subsequently, the results are compared with the retail districts focusing on usage, which were established by the Korea Real Estate Board (REB). Regarding region, this study targets Seoul, the capital of the Republic of Korea, and regarding period, this study uses data from 2006, when MOLIT began to publish real estate transaction data, to 2023.

The specific contents of the analysis are in three parts: First, core retail districts in Seoul are derived using transaction data, which are reviewed to identify whether the number and locations match the general perception. Second, the boundaries of the core retail districts are compared with those of the REB retail districts. Meanwhile, this study reviews whether spatial clustering using transaction data captured changes such as the expansion or contraction of retail districts. Third, price indices are prepared for retail districts focusing on the transaction and usage aspects and tested for differences. If significant differences are found between the two price indices, this study would imply that the inappropriate methods of regionalization can cause bias in the statistics of retail market.

This paper is structured as follows. Sections 2 and 3 examine previous studies and statistical cases addressing retail districts, respectively. Through this process, this study identifies the core retail districts of Seoul that focus on usage. Section 4 introduces this study's analysis, data, spatial unit, and analysis model. Section 5 presents the results of establishing retail districts using transaction data and compares them with retail districts regarding the usage identified in Section 3. Finally, Section 6 summarizes the contents of the study and its implications.

II. Literature Review

Studies on retail have developed, centered on the hinterland, following Hotelling's (1929) proposal of Location Theory that maximized market share using the linear city concept. Most classic studies on spatial competition do the same, ranging across Reilly's (1931) Law of Retail Gravitation, which demonstrated that the retail district's ability to gather customers is proportional to the population of the hinterland and inversely proportional to the distance; Converse's (1949) Breaking Point Formula which found boundaries of the hinterland considering the attractiveness of shopping malls, variety of products, and accessibility of customers; Huff's (1963, 1964) Probability Model which introduced probabilistic factors into the Gravitation Model; and Christaller's (1966) central Place Theory which differentiated the range of customer reach according to the nature of the product.

The origins of studies on retail districts, which is this study's subject of interest, lie in seeking the basic units necessary to calculate the appropriate size of the central business district (CBD) in urban planning (Woodbury, 1928). This necessity prompted Hartman's (1950) study which sought the geographic pattern of CBD through its relationship with the hinterland and was developed into a study by Vance, Jr. to explore the specific boundaries of CBD (Vance Jr., 1955). Since then, studies on retail districts have been developed in various forms, including those focusing on social characteristics such as population (Briggs, 1974) and performing a space syntax analysis using the degrees of social and physical gathering (Şıkoğlu et al., 2020). The development of technology resulted in attempts to identify CBD by capturing the emission of light from night lighting (Jie at.al., 2023).

Recently, spatial clustering algorithms that impose spatial constraints on cluster analysis have been gaining attention. Spatial clustering has developed into various types, including partition-based, hierarchy-based, grid-based, and density-based models (Neethu and Surendran, 2013, pp.15-24). Among these, the partition-based model simultaneously clusters all entities subject to analysis. This design contrasts with the hierarchy-based model, which gradually groups or divides all entities. In the partition-based model, each entity is assigned to a single cluster, and every cluster must include one or more entities. This characteristic is evident in the term “partition” in the model's name. However, instead of strict division, fuzzy or soft techniques have recently been developing, where all entities are assigned to multiple clusters. Then, the final cluster is determined by comparing the strength of its belonging to each cluster. An example of this is the spatial fuzzy C-means algorithm.

An example of implementing spatial clustering using the spatial fuzzy C-means algorithm can be found in Gelb and Apparicio (2021). Their study used ten environmental and social variables and found the spatial fuzzy C-means algorithm excellent for spatial clustering. Negative environmental indicators, such as pollution; positive environmental indicators, such as plants; and social indicators, such as population structure and unemployment, were used as input variables. The analysis method they adopted was extended into a study by Wu (2022) for the zoning of the 2020 United States presidential election results. In addition to the general variables known to affect endorsed candidates, such as income and education level, he included variables such as mask wearing and mortality rates resulting from COVID-19. An example of spatial clustering for Seoul can be found in Min (2023). He used the spatial fuzzy C-means algorithm to establish office districts in Seoul. He reported that, upon conducting spatial clustering using price, transaction volume, road condition, and building size as input variables, core office districts were distinguishable from other areas, and their boundaries differed from general perceptions based on administrative districts, albeit with differences in degree depending on the district.

Choi et al. (2021) examined and identified commercial districts. They captured the city center, Gangnam, Mapo, and Yeongdeungpo commercial districts by performing a kernel density analysis on retail businesses' closure and survival data from 2014 to 2016. They identified the expansion and contraction trends of each commercial district. Additionally, Oh et al. (2022) analyzed the changes in the boundaries and hierarchy of commercial districts in Seoul from 2000 to 2019 through kernel density analysis and hydrological modelling. They divided Seoul into five zones comprising the city center, southeast, northeast, southwest, and northwest, compared the analysis results with that of the "2030 Seoul Living Area Plan," and argued that the commercial districts in Seoul were growing centered on the city center and southeast zone and that the southeast zone, in particular, was functioning as a city center beyond a city subcenter.

Meanwhile, Kwon and Jeon (2022) examined the changes before and after the COVID-19 pandemic using data on foot traffic and revenues from 2015 to 2020 from "Our Village Store Commercial District Analysis Service" provided by the Seoul Metropolitan Government. They analysed the spatial changes in commercial districts using local indicators of spatial association and analysed the degrees of changes by the panel regression model. They argued that significant spatial changes were observed before and after the COVID-19 pandemic and that the degrees of such changes varied depending on the commercial district and business type.

Therefore, studies on retail districts began from urban planning needs and developed toward analysing social phenomena, progressing from models using macroeconomic variables toward those using microscopic data such as the characteristics of individual buildings and the operational performance of retailers. Despite these developments, most studies have only focused on usages, such as retailers and customers, while overlooking transaction aspects, such as real estate prices. Although Min (2023) attempted to cluster spaces by inputting variables such as prices and transaction volumes, the subject was limited to offices.

Retail function supply and demand lead to real estate supply and demand, which is reflected in retail real estate prices. This mechanism operates more quickly when the capital market is efficient. Therefore, performance can improve if spatial clustering is implemented by inputting both usage and transaction data. This study differs from previous studies in that it analyzes retail districts using real estate transaction data, which has been insufficiently covered in previous studies. Moreover, this research focuses on investors, unlike previous ones, which focused on retailers and customers.

III. Retail Districts in Seoul

An example of the establishment of retail districts in Seoul can be found in the REB. The REB has been conducting the "Survey of Commercial Real Estate Rental Trends" every quarter and announces the results since 2002, which include four types of real estate: office, medium and large shopping centers, small stores, and strata-type shopping centers. Three of these, excluding office, are retail. REB surveys by establishing districts according to the characteristics of the market though strict procedures.

When it comes to characteristics, the REB website states that it “reviews the spatial distribution of areas with a high concentration of shopping centers through population analysis, and comprehensively considers a collection of opinions from professional REB investigators, media issues, and the degree of activation of retail districts.” This statement effectively represents the usage approach since it focuses on retailers and customers.

Regarding the procedure, according to an interview with the REB representative, since this statistic is a National Approved Statistic under the Statistics Act, any changes to the districts require thorough research on the rationale and impact, followed by consultation with the Statistics Korea. Therefore, it is difficult to modify the districts on a quarterly or yearly basis, and such changes are made irregularly only when the necessity is recognized. During this process, both internal and external experts of the REB closely examine the aforementioned usage characteristics of the market.

In addition, unlike other service providers, REB precisely discloses the boundaries of each commercial district on a map. Therefore, this study establishes core retail districts using real estate transaction data and compares them with the boundaries delimited by the REB. Among the retail districts of the REB, 61 met the criteria of medium and large shopping centers (gross floor area exceeding 330㎡), which best suited this study’s purpose.

The cores among the 61 retail districts are identified from market reports by real estate service agencies who are publishing statistics for institutional investors. This study utilizes four reports: Cushman and Wakefield (2023), Savills (2023), Avison Young (2019), and CBRE (2022).

Cushman and Wakefield (2023) define Myeong-dong, Gangnam, Hongik University, Garosu-gil, Hannam/Itaewon, and Cheongdam as the six major retail districts. They include Yeouido in case the focus was on shopping malls and arcades.

Savills (2023) mentions Myeong-dong, Itaewon, Garosu-gil, Gangnam Station, and Hongik University as major retail districts and mentions Yeonnam/Mangwon, Yongsan/Samgakji, Anguk, Seongsu, and Apgujeong/Cheongdam as emerging retail districts. In a report published in 2019,

Avison Young (2019) classifies Gangnam Station, Myeong-dong, Jongno, Yeouido, Hongik University/Hapjeong, Itaewon/Hannam, Sinsa/Apgujeong, and Konkuk University as the eight major retail districts in Seoul.

CBRE (2022) defines Garosu-gil, Hongik University entrance, Apgujeong Rodeo Street, Nonhyeon Station, Jongno 3(sam)-ga Station, Seoul National University of Education Station, Yeoksam Station, Yeouido Station, Jamsil Saenae Station, and Sillim Station as the ten major retail districts in Seoul. Table 1 presents the summary of these districts.

<Table 1> Core Retail Districts in Seoul

C&W	Savillls	AY	CBRE
		Jongno	Jongno 3(sam)-ga
Myeongdong	Myeongdong	Myeongdong	
Hannam Itaewon	Itaewon	Itaewon Hannam	
		Konkuk University	
Hongik University	Hongik University	Hongik University Hapjeong	Hongik University
Yeouido		Yeouido	Yeouido Station
Garosugil	Garosugil	Sinsa Apgujeong	Garosugil Apgujeong
Gangnam	Gangnam Station	Gangnam Station	

In summary, Seoul can be divided into seven core retail districts. This result is drawn from selecting areas designated as core retail districts by two or more agencies. The seven core retail districts are Jongno, Myeong-dong, Itaewon/Hannam, Hongik University, Yeouido, Garosu-gil, and Gangnam Station. Among these, six retail districts, excluding Yeouido, are listed in the 61 retail districts noted by the REB; the REB lists Yeouido as a collective shopping center district.

IV. Materials and methods

1. Analysis Data

This study utilizes real estate transaction statistics published by MOLIT. The data is divided into residential and commercial real estate, with commercial real estate further subdivided into office, retail, hotel, and industrial properties. Each category includes both building-level transactions and transactions for individual units within buildings. MOLIT provides brief information for each transaction, such as the approximate location, land and building area, and transaction price. The data has been published with national coverage since January 2006 and is publicly accessible and can be freely downloaded from the website.

Unlike residential or office properties, retail real estate exhibits significant differences across submarkets. Therefore, when defining retail districts, it may be more effective to select a submarket that aligns with the specific objective.

Accordingly, this study focuses only on building-level transactions for properties of a certain size from the MOLIT data. The minimum size criterion is set as land area of at least 100㎡ and building area of at least 300㎡, which aligns with the sampling standards used in the KB Real Estate Small and Medium-Sized Building Price Index. As a result, this study excludes undersized and strata properties, thereby focusing specifically on retail real estate that meets the defined criteria, rather than the entire population. The subject of analysis concerning space was Seoul. Its duration was 207 months, ending in March 2023.

Although this data includes numerous transactions over an extended period, they have quite a few typos and do not separately identify transactions under exceptional circumstances, such as urgent trades and trades between stakeholders. Therefore, this data analysis required strict preprocessing and removal of outliers.

Data preprocessing was performed as follows: First, cases where descriptions of zoning or subzoning were absent were removed. Next, cases where information regarding road width was omitted were also removed. Then, cases in which the construction completion year was omitted, other descriptions were erroneous, or building was too old (pre 1955 construction) were also excluded from the analysis. Finally, cases of partial share transactions were removed because their share prices were lower than general ones, so the disparity may have caused bias in spatial clustering.

Outlier removal was performed referring to land and building area and price. Regarding area, cases <100㎡ for land and <330㎡ for buildings are removed since the price per unit area may significantly differ from the average price level. Cases where the value of the building area divided by the land area did not fall within the 0.5-20 range are also removed. If the value is ≤0.5, the land may be too large compared to the building and, therefore, could be considered a land transaction. Conversely, if the value is ≥20, the land may be too small compared to the building and, therefore, could be a partially traded site. The range from 0.5 to 20 is established considering the legal upper limit of the floor area ratio according to each zoning in Seoul. Regarding transaction price, cases deviating from three times the interquartile range (IQR) yearly are removed based on the natural logarithmic value of land and building unit prices. Along the way, owing to this study's extended analysis period, from January 2006 to March 2023, all transaction prices are converted into values using Seoul's monthly inflation rate.

Following data preprocessing and outlier removal, 112,031 transacted cases were retained (see Table 2). They comprise 18,592 buildings and 93,439 building units. This study only analyzed 18,592 buildings with singular ownership, removing ownership type bias.

<Table 2> Number of Transactions

Process	Number of transactions
Raw Data	261,290
After Rreprocessing	208,636
After Outlier Removal	112,031
After Units Removal	18,592

2. Spatial Unit of Analysis

The width of the spatial unit of analysis must be established to perform spatial clustering. The narrower the spatial unit, the more precise the establishment of retail districts. However, this step requires the obtainment of precise information regarding the location of each transacted case, and the analysis may result in multiple spatial units with insufficient or no transacted cases. Conversely, obtaining data is easy when a broad spatial unit is chosen; however, the boundaries of the retail district may not be precise.

Candidates for the spatial unit include administrative and national basic districts (the latter of which is the reference for postcode assignment) and lots. Administrative districts are difficult to use because their areas are extensive compared to general retail districts. Lots are difficult to use because MOLIT does not disclose the exact addresses of transacted cases. By contrast, national basic districts are suitable as spatial units for retail districts. They are smaller in area than administrative districts and homogeneous. Currently, Seoul comprises 5,665 national basic districts. This study establishes retail districts in units of national basic districts.

Allocating each transacted case to a single national basic district was complex as MOLIT discloses only the first digit of the lot number and the road name for the location of the transacted case. This study performs this task by combining these two pieces of information and matching them with a lot number road name address conversion code provided by Korea Post. Consequently, this study allocates 18,592 transacted cases to corresponding national basic districts and spatially clustered 5,665 national basic districts according to real estate prices and transaction volumes. Therefore, the location of each transacted case for calculating the spatial weight matrix became the central point of the corresponding national basic district.

3. Analysis Model

This study uses the spatial fuzzy C-means algorithm (SFCM), a machine learning-based spatial clustering technique, for the following reasons. First, the data provided by the MOLIT does not include precise location for individual transactions. Therefore, grid-based models, which require the assignment of individual transactions to researcher-defined grids, could not be employed. Second, for the same reason, density-based models, which rely on the density of clustered data points, were deemed inappropriate. Third, given the objective of this study, which is to explore the boundaries of commercial districts, partition-based models were considered more suitable than hierarchical models which primarily focus on forming hierarchical structures. Lastly, among partition-based models, fuzzy C-means clustering was deemed more appropriate for delineating boundaries compared to dichotomous models such as K-means clustering, as it allows for the estimation of the probability that each national basic district belongs to multiple clusters.

The basic fuzzy C-means algorithm minimizes the objective function that quantifies the heterogeneity between each entity and cluster center. Thus, the entity belonging to each cluster and cluster center are repeatedly replaced until the objective function is minimized. The objective function in the basic fuzzy C-means algorithm is as follows.

$$J=\sum_{i=1}^{N}+\sum_{k=1}^{C}u_{ik}^{m}d(x_i,c_k)^2$$

Here, J represents the objective function, N is the number of data points, C is the number of clusters, x_i represents the data point, c_k is the cluster center, $u_{m,ik}$ is the membership degree, m is the fuzzification parameter, and $d(x_i, c_k)$ denotes the distance between a data point and a cluster center. The basic fuzzy C-means algorithm initializes cluster centers and membership degrees first, and iteratively updates these values to minimize the spatially weighted objective function

$$u_{ik}^{m}=\frac{1}{\sum_{j=1}^{C}\left(\frac{d(x_{i,}c_k)}{d(x_{i,}c_j)}\right)^{\frac{2}{m-1}}}$$

$$C_k=\frac{\sum_{i=1}^{N}u_{ik}^{m}x_i}{\sum_{i=1}^{N}u_{ik}^{m}}$$

The SFCM extends the basic fuzzy C-means algorithm by incorporating spatial information, leading to results that better reflect spatial continuity. The spatially weighted objective function adds a term to consider the influence of neighboring data points.

$$\begin{aligned}J_s=&\sum_{i=1}^{N}\sum_{k=1}^{C}u_{ik}^{m}\,d_s(x_{i,}c_k)^2\\&+\alpha\sum_{i=1}^{N}\sum_{j\in N_i}\sum_{k=1}^{C}u_{ik}^{m}d_s(x_{j,}c_k)^2\end{aligned}$$

$$u_{ik}^{m}=\frac{1}{\sum_{j=1}^{C}\left(\frac{d_s(x_{i,}c_k)}{d_s(x_{i,}c_j)}\right)^{\frac{2}{m-1}}}$$

$$d_s(x_i,c_k)=d(x_i,c_k)+\alpha\sum_{j\in N_i}d(x_j,c_k)$$

Here, a (alpha) is a parameter that controls the weight of spatial influence, N_i is the set of neighbors for data point x_i, and the rest of the terms are defined as in the basic fuzzy C-means algorithm. This adjustment integrates the influence of neighboring points to promote spatial consistency.

Techniques that apply spatial constraints to the basic fuzzy C-means algorithm include not only the SFCM, but also the spatial generalized fuzzy C-means algorithm (SGFCM). SGFCM further extends SFCM by using a generalized spatial weight matrix to better capture spatial relationships and interactions, offering greater flexibility in modeling complex spatial structures.

The analysis model of this research is referenced from Min (2023), however, it differs from previous studies in that only two variables are inputted, the real estate price and transaction volume, which are pure transaction data, and that a small spatial unit of analysis is established considering the characteristics of retail districts with sensitive boundaries.

The importance of the spatial term in the objective function of the SFCM is controlled through the spatial weight parameter (alpha). The value of alpha is greater than 0, and the importance of the adjacent entity in the objective function increases as the value increases. No spatial constraint has been applied if alpha is 0, each entity and adjacent entity are identically weighted if alpha is 1, and the adjacent entity is weighted twice as much as each entity if alpha is 2. In general, the optimal alpha is derived through simulation. The degree to which each entity overlaps over multiple clusters is controlled through the fuzzification exponent (m); 1.5 is generally applied to m. The scaling exponent (beta), which controls the importance of the distance between each entity and the cluster center, is added to the SGFCM. This strategy increases the flexibility of the objective function. The value of beta ranges from 0 to 1, and, similar to alpha, its optimal value is determined by simulation

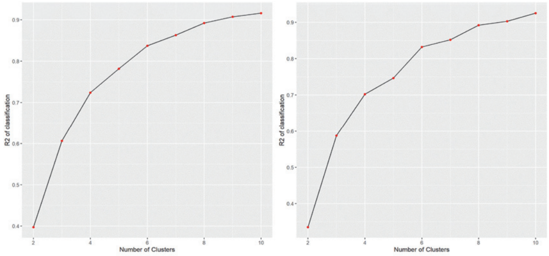
V. Results

1. Spatial Clustering

This study performs spatial clustering of retail in Seoul by inputting price, transaction volume, and the spatial information of transacted cases. Regarding price, the price per unit area (KRW/㎡) adjusted by the inflation rate is used, and regarding transaction volume, two measures, transacted amount (TAM) and transacted area (TAR), are applied. The analysis technique uses two algorithms, the SFCM and SGFCM. Subsequently, this study adopts the one with the best performance in spatial clustering among the four models.

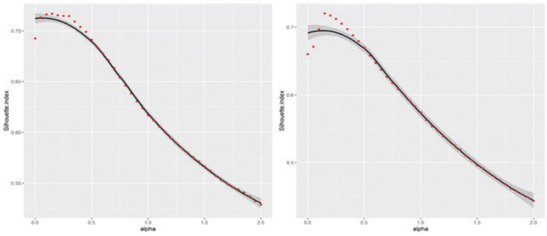
The number of clusters is first determined for spatial clustering. This study used K-means clustering with two variables, price and transaction volume, from two to ten clusters. The number at which the increase in R2 slows was identified. Figure 1 shows how the slope of the curve slowed at three clusters using TAM and TAR as transaction volumes. Therefore, the subsequent analysis was performed with three clusters.

<Figure 1> Number of Clusters
(left: TAM, right: TAR)



The optimal alpha value as a hyperparameter for the SFCM was selected by comparing the performance upon clustering with increments of 0.05 from 0 to 2. The performance comparison is based on the silhouette (the larger, the better), which measures how clearly each cluster is divided. The fuzzy index m is set to 1.5, the threshold intensity with which clustering is reserved is set to 0.5, and the maximum number of repetitions is set to 500. Figure 2 presents the result of repeatedly conducting SFCM. It can be seen that the silhouette peaks at an alpha of 0.15 in both models using TAM and TAR as transaction volume.

<Figure 2> Alpha of SFCM
(left: TAM, right: TAR)



To implement the SGFCM, beta and alpha must be determined simultaneously. In this study, the silhouette is compared within the same interval as aforementioned for alpha and compared with increments of 0.05 from 0 to 1 for beta. The results are presented in Figure 3. Regarding the model using TAM as transaction volume, the silhouette is the highest at an alpha of 0.3 and a beta of 0.9. Regarding the model using TAR as transaction volume, the silhouette is the highest at an alpha of 0.5 and a beta of 0.9.

<Figure 3> Alpha and Beta of SGFCM
(left: TAM, right: TAR)

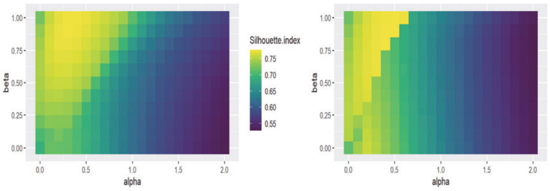
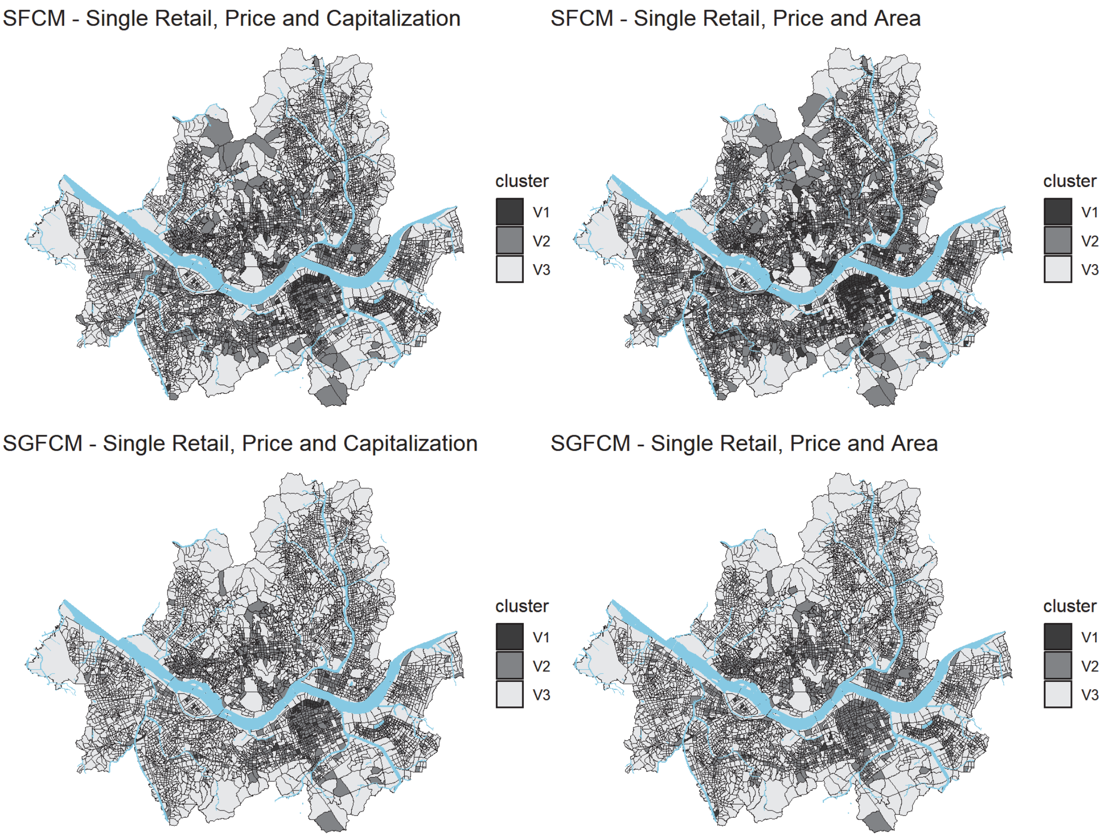


Figure 4 presents the results of spatial clustering using the established hyperparameters. In the central region of Seoul, major retail districts form V1, which differs from the

<Figure 4> Results of Spatial Clustering using SFCM and SGFCM


SFCM - Single Retail, Price and Capitalization

SFCM - Single Retail, Price and Area

SGFCM - Single Retail, Price and Capitalization

SGFCM - Single Retail, Price and Area

surrounding area. The neighboring retail districts are scattered broadly, forming V2 near the surrounding residential area. Concurrently, the remaining V3 resembles the background without forming a consistent zone. However, as the unit of analysis is small, recognizing the difference in model performance with the naked eye is difficult.

Model comparison is conducted using various indicators that demonstrate the clustering performance. This study compares the performance through consistency, an indicator of consistent cluster setting, and Moran's I, which presents the degree of spatial autocorrelation within each cluster. Moran's I is calculated for each cluster, among which this study compares V1, which is significant as a retail district. Consequently, considering indicators, the SFCM using TAR as transaction volume was found to be best and selected as the final model(Table 3).

Table 4 presents the descriptive statistics of the logarithmic values of unit prices for the three clusters. As expected, the price level of V1 is much higher compared to the other clusters.

<Table 3> Performance of spatial clustering

Classification	SFCM		SGFCM	
	TAM	TAR	TAM	TAR
Consistency	0.6732	0.7146	0.5593	0.5070
Moran's I V1	0.2608	0.3662	0.2238	0.0394
Moran's I V2	0.2992	0.2397	0.4326	0.4984
Moran's I V3	0.3637	0.3112	0.4720	0.5087

<Table 4> Log price of each cluster

Cluster	V1	V2	V3
mean	2.00	0.48	-0.68
median	1.69	0.34	-0.77
sd	1.69	0.48	0.23
min	-0.14	-0.30	-0.77
max	16.05	2.55	0.09
count	453	2272	2940

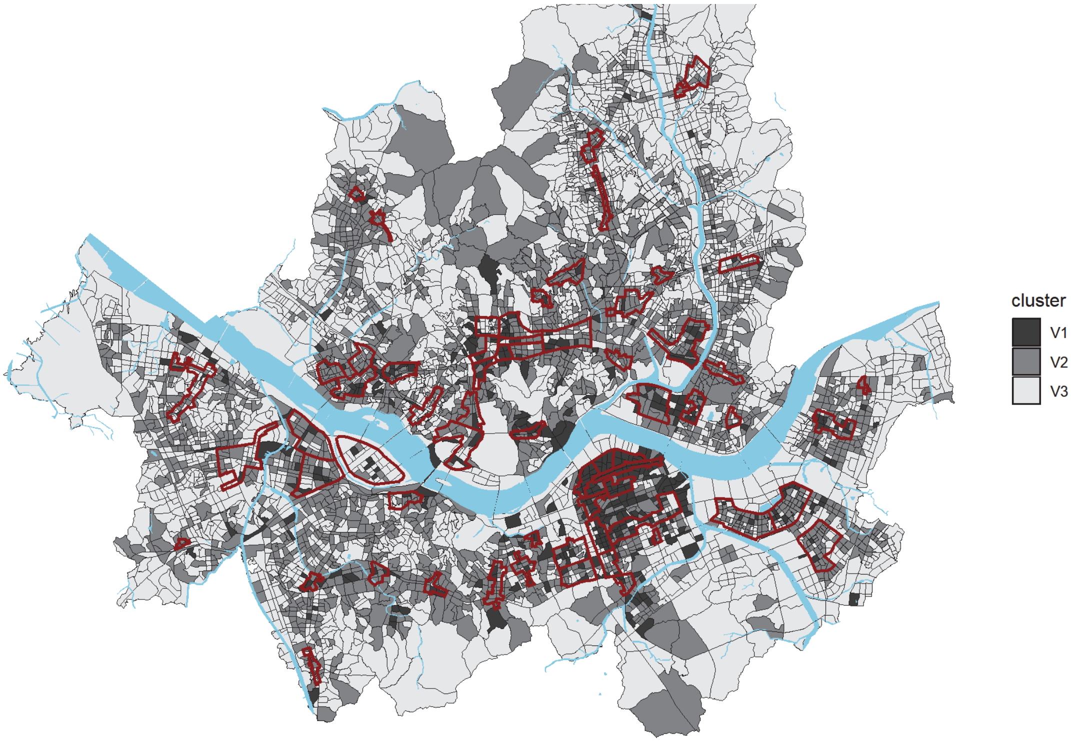
Figure 5 presents the result of magnifying the map of the final model and marking the boundaries of the 61 retail districts established by the REB. Figure 6 is a magnified version of Figure 5, focusing on the major areas of Seoul to enhance readability.

The distribution of V1 generally coincides with the distribution of the REB retail districts across Seoul. Although the REB retail districts do not include all of the national basic districts classified as V1, those that deviate from the districts are few, while the number of the REB retail districts not including any of the national basic districts classified as V1 are very few. In addition, no cases are identified in which the REB established a retail district with only national basic districts classified as V3. Although this comparison is not a statistical test, it affirms that spatial clustering using transaction data is similar to the establishment of retail districts considering usage.

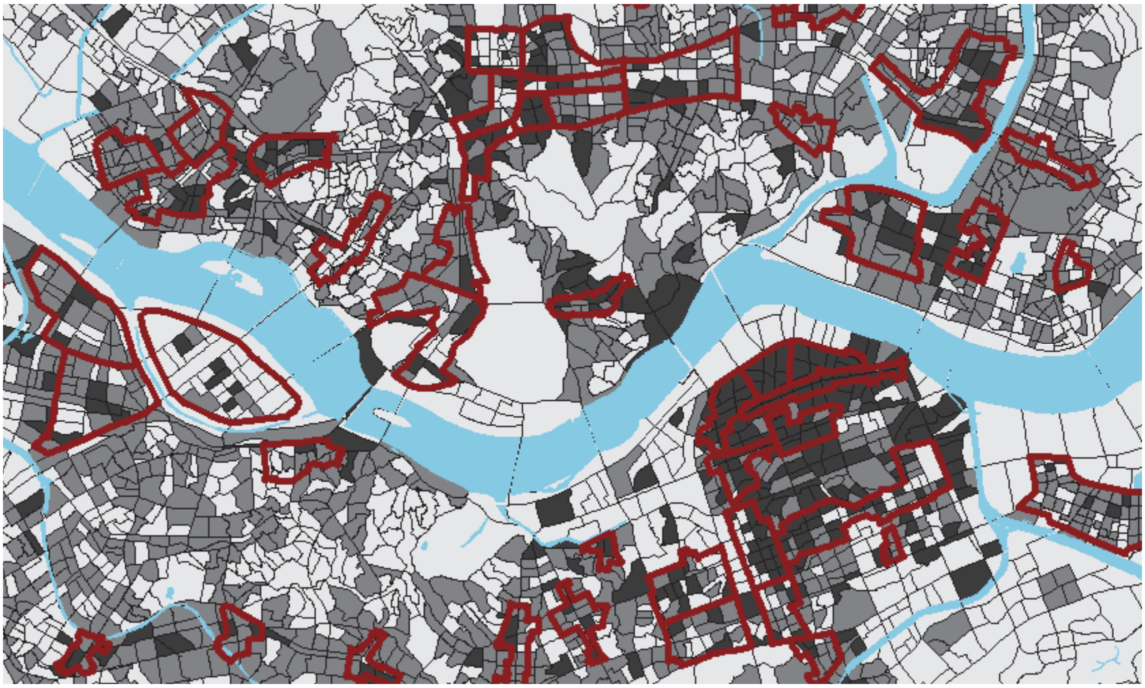
2. Comparison of Boundaries

With reference to the spatial clustering results, Figure 7 presents a detailed illustration of the seven core retail districts of Seoul identified in Section 3. The Jongno and Myeong-dong retail districts are geographically near each other. Therefore, they are presented within one map, where the location in the north of the map is Jongno, and the location in the south is Myeong-dong. As demonstrated in the figure, V1 and V2 are distributed more broadly than the boundaries of the REB retail districts for all retail districts except Yeouido.

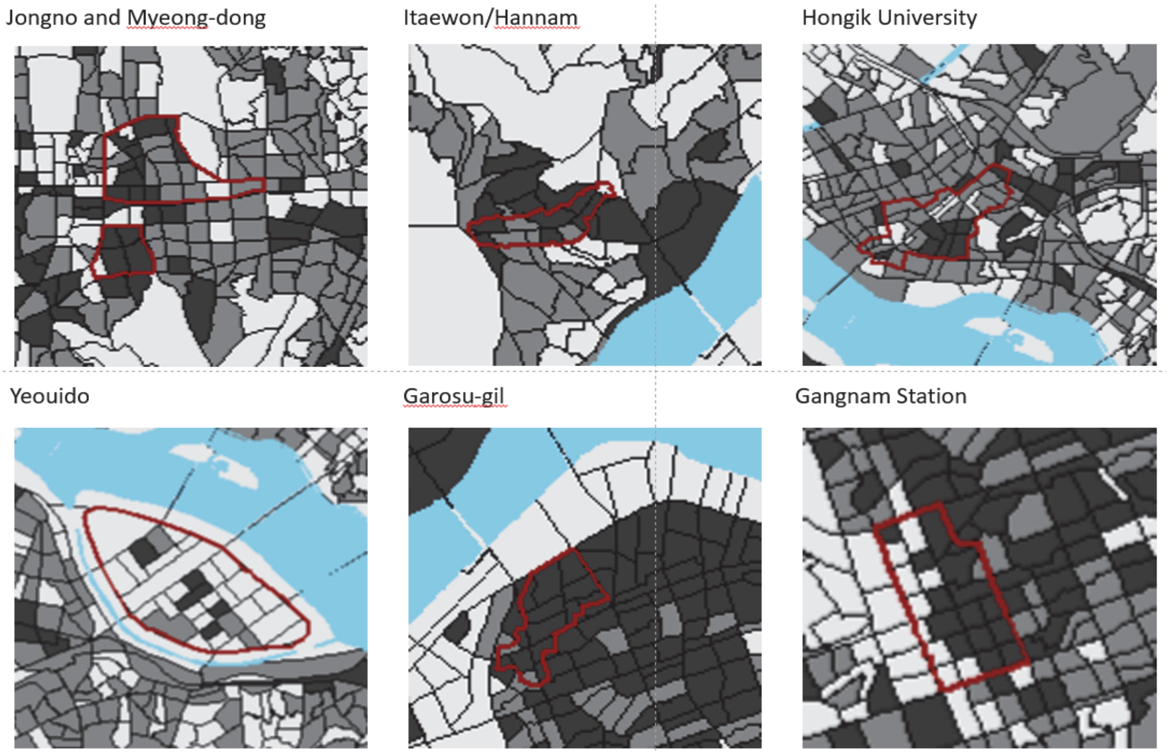
<Figure 5> Result of Spatial Clustering by the Final Model (Seoul)



<Figure 6> Result of Spatial Clustering by the Final Model (Central Area)



<Figure 7> Boundary comparison of the seven core retail districts



First, Jongno and Myeong-dong form two separate retail districts even in spatial clustering using transaction data. The degree of activation appears to vary in that the west side of the REB retail district for Jongno is included in V1. However, the east side is included in V2, and in particular, the east end appears disconnected from the west. Regarding Myeong-dong, most REB retail districts belong to V1. Therefore, the two approaches do not differ significantly. Spatial clustering using transaction data demonstrates that the Myeong-dong retail district is expanding to the south.

Second, regarding Itaewon/Hannam, a significant difference is found between the two approaches. The REB established a linear retail district in the east to west direction along Itaewon-ro, a street, and the spatial clustering result demonstrates that the retail district had been broadly expanded to the south and north. This pattern could be due to the revitalization of retail in nearby areas and the increased expectations for the redevelopment of the southern riverside.

Third, regarding Hongik University, the spatial clustering result is contracted relative to the REB boundaries. Although only the Hapjeong area in the south is classified as V1, most areas, including Hongik University Station, are classified as V2, indicating that no significant difference was found between the inside and outside of the boundary.

Fourth, regarding Yeouido, the REB establishes the entire area surrounded by the river as a single retail district; however, the location where transactions are active is limited to the central region. This central region is where main roads and subway stations converge. Among the seven core retail districts, Yeouido demonstrates the most significant difference between the two approaches.

Fifth, Garosu-gil occupies a part of the vast V1 area. It is the largest retail district in Seoul and includes Garosu-gil and numerous other regions (see Figure 5). The broad distribution of V1 in the eastern and southern sides of Garosu-gil indicates no expansion of the Garosu-gil retail district but instead the division of the extensive retail district into several subdistricts by the REB. Most of the area inside the Garosu-gil retail district is classified under V1.

Sixth, similarly, Gangnam Station is included in the same extensive retail district as Garosu-gil. Garosu-gil forms the northwest boundary, while Gangnam Station forms the southwest boundary. Therefore, concluding that the V1 area distributed along the north and east sides of Gangnam Station results from an expansion of the Gangnam Station retail district is difficult. Gangnam Station differed from Garosu-gil in that the side to the east of Gangnam-daero, a street crossing through the north and south, was active while the side to the west was not. Therefore, related only to the transaction aspect, considering the sides to the east and west of Gangnam Station to be a single retail district is challenging.

As the analysis used unsupervised machine learning, tasks such as significance tests could not be performed. However, by aggregating the analysis, the following exploratory results are obtained. First, from a macroscopic perspective, results similar to retail districts focusing on usage were obtained by performing spatial clustering using real estate transaction data. Second, the expansion and contraction of the retail district and the directions of changes are identified through spatial clustering using real estate transaction data.

3. Comparison of Price Indices

To demonstrate that even minor boundary differences can lead to significant variations in retail real estate statistics, this study constructs separate price indices for transaction cases located in national basic zones classified as V1 by SFCM and for those located within retail districts designated by REB, and then tests whether the differences are statistically significant.

Since the boundaries of the REB retail districts are not identical to those of the national basic districts, the selected transaction cases focusing on usage were completed in the national basic districts where more than 50% of the area belongs to the REB retail districts.

Of the 18,592 transactions, 5,783 were completed within national basic districts classified as V1 by the SFCM, and 6,176 were completed within the REB retail districts. Since many transaction cases overlap, the trajectories of the price indices are likely to be similar. However, if significant differences in the price indices arise from the non-overlapping transaction cases, it can be concluded that greater caution should be exercised when setting the boundaries of retail districts.

The price indices are prepared using the hedonic price model to control differences in the characteristics of real estate. The dependent variable is the logarithmic value of the price per unit area and the independent variables are as follows.

The time dummy variables are essential independent variables for constructing price indices. In this study, dummy variables were incorporated on a quarterly basis, using the first quarter of 2006 as the reference point.

As variables representing location, Gu dummy variables corresponding to the 25 administrative districts of Seoul and adjacent road width dummy variables deviding roads into four categories with 8m, 12m, and 25m were included.

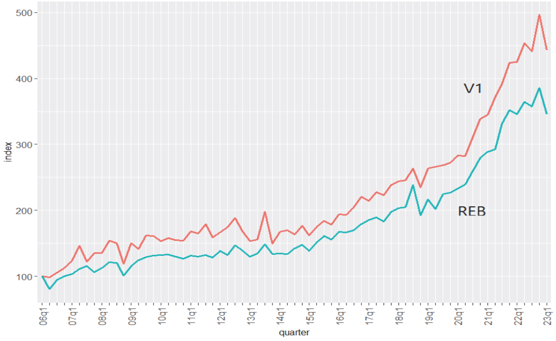
As variables reflecting the regulative use of lands and buildings, zoning and building use dummy variables were included.

For other characteristics, land area(㎡), total floor area(㎡), and elapsed years since the completion of the building were included in the form of continuous variables.

Appendix 1 presents the OLS implementation results based on the two approaches. Most input variables are significant for both approaches, indicating that the locational and physical characteristics are well controlled. The adjusted R2 of each of the two models was high at 62.88% and 65.67%, respectively. The time dummy variables representing quarters was omitted from the table.

Figure 7 presents the results of preparing price indices using the coefficient values of the time dummy variables. Across the entire period, the price index of the region classified as V1 in the SFCM is higher than that of the REB retail district. Furthermore, its volatility is greater. This result is natural because V1 differs from other areas concerning price and transaction volume. The two indices, which started at 100 at the beginning of 2006, increased to 442.73 and 345.64 in the first quarter of 2023. Table 5 presents the basic statistics of the two indices.

<Figure 8> Price Indices



<Table 5> Basic Statistics of Indices

Region	Min	Mean	Median	Max	SD
V1	98.25	216.18	176.79	496.74	95.88
REB	80.31	178.24	147.10	385.75	76.52

The test of differences between the indices is performed using two methods: the parametric paired t-test and the nonparametric Kolmogorov-Smirnov test (KS test). Given this study’s extended analysis period, the quarterly preparation of the index produced 69 time points. Therefore, applying these methods was judged to be reasonable. The results of the equality of variance test showed that equality was present at the 5% significance level with F = 1.57 and p-value = 0.06. Therefore, a paired t-test assuming equal variance was performed, and the difference between the indices was significant, with t = 14.84 and p-value < 0.001. Next, the KS test was conducted with an exact two sample test. Similar to the paired t-test, the difference between the indices was found to be significant with D = 0.38 and p-value < 0.001. These results confirm significant differences in price indices depending on how the boundaries of the retail districts are established. In particular, whether focusing on usage or transaction is a crucial source of differences between indices can be determined.

VI. Conclusion

Establishing appropriate retail districts is essential for understanding retail markets. This study conducted various analyses to demonstrate the necessity of considering not only usage but also transaction data when defining retail districts, and obtained the following results. First, retail districts can be reliably established by spatial clustering using transaction data. Second, differences were found in the boundaries of core retail districts using the two approaches. Third, there was a significant difference between the two price indices derived from usage and transaction-based districts. These findings provide the following implications.

First, setting appropriate boundaries for retail districts is crucial when constructing real estate statistics. As demonstrated in this study, even minor boundary differences can result in significant variations in statistical outcomes, even if the retail districts are defined in similar locations and numbers. This phenomenon is likely to be more pronounced in the retail market. In particular, using boundaries defined solely from a usage perspective is not suitable for compiling some statistics, such as price indices, because it fails to adequately reflect the transactions occurring around the area.

To address this issue, two alternatives can be considered. The first is to establish retail districts that balance both usage and transaction perspectives. The second is to use usage-based boundaries when compiling usage-related statistics, such as visitor counts, and transaction-based boundaries for transaction-related statistics, such as price indices. However, the latter approach may result in lower spatial consistency between different statistics, making it inconvenient when analyzing multiple statistics simultaneously.

Second, retail districts derived from transaction data can provide important insights when adding new retail districts or adjusting existing ones. As confirmed in this study, the boundaries of retail districts defined solely by transaction data can offer valuable information regarding the expansion and contraction of these districts.

As mentioned above, it is desirable to comprehensively and rigorously incorporate both usage and transaction data when defining retail districts. Separately, when there is a need to monitor changes in retail districts, spatial clustering using only transaction data can be a useful approach, as it can be conducted using publicly available statistics provided by MOLIT without requiring additional surveys.

Third, in the context of the retail market, where the size of districts is small and boundary volatility is high, spatial clustering using small spatial units, such as national basic districts, can be a useful method for defining retail boundaries. Unlike residential or office markets, larger administrative units, such as Gu or Dong, are not suitable for capturing the fine boundaries of retail districts. As shown in this study, the spatial fuzzy C-means algorithm is particularly effective in performing this function.

However, national basic districts are not the only option. Better results could potentially be achieved using more detailed and homogeneous grids or by calculating the densities of transaction cases. This would require precise address information for each transaction. However, due to legal restrictions related to personal data protection, MOLIT is currently unable to provide such information. If this data becomes available in the future, it would be possible to identify the optimal model through a comparison of various algorithms.

Although this study demonstrated the necessity of establishing retail districts using both usage and transaction data, it had some limitations. First, owing to the nature of unsupervised machine learning algorithms, differences between the two approaches based on usage and transaction could not be statistically tested. This study conducted an indirect test by comparing price indices, and more diverse comparisons are required to increase the robustness of the analysis. Second, this study analyzed data from 69 quarters in Seoul. The analysis should be expanded to various cities and periods to generalize the obtained results. In particular, with respect to the period, this study utilized the entire period for which data was available. However, in order to ensure the model's utility, further research is required to analyze how retail districts have evolved over time and to determine the optimal length of historical data needed for defining current retail districts.

논문접수일 : 2024년 9월 19일
논문심사일 : 2024년 10월 8일
게재확정일 : 2024년 10월 23일

References

1. Kwon, D. W. and Jeon, J. S., “The COVID-19 Pandemic Impact on the Seoul Retail Market: A Spatial Perspective,” Journal of the Korea Real Estate Analysts Association, Vol. 28 No. 3 2022. 9, pp. 25-44

2.. Min, S., “Regionalization of Seoul Office Market Using Machine Learning Algorithm for Spatial Clustering Based on Transaction Data,” Korea Real Estate Review, Vol. 33 No. 3, 2023, pp. 31-52

3. Oh, Y. K., Lee, B. K., and Lee, S. K., “Comparative Analysis of the Hierarchy of Market Areas and Central Place System in Seoul,” Journal of The Korean Cadastre Information Association, Vol. 24 No. 1, 2022, pp. 65-76

4. Choi, E. J., Cheon, S. H., and Lee. S. G., “An Analysis of Spatial Changes in Commercial Districts using Survival-Exit Dynamics of Commercial Businesses in Seoul, Korea,” Journal of The Korean Cadastre Information Association, Vol. 37 No. 4, 2021, pp. 3-19

5. Briggs R., “A model to relate the size of the central business district to the population of a city,” Geogr Anal, Vol. 6 No. 3, 1974, pp. 209-312

6. Christaller W., “Central Places in Southern Germany,” Prentice-Hall, 1966

7. Converse P., “New laws of retail gravitation,” J Mark, Vol. 14, 1949, pp. 379-384

8. Gelb J. and Apparicio P., “Apport de la classification floue c-means spatiale eng´eographie: Essai de taxinomie socio-r´esidentielle et environnementale `a lyon,” Cybergeo, 2021

9. Hartman G. W., “The central business district-a study in urban geography,” EconGeogr, Vol. 26 No. 4, 1950, pp. 237-244

10. Hotelling H., “Stability in competition,” Econ J, Vol. 39 No. 153, 1929, pp. 41-57

11. Huff D., “A probabilistic analysis of shopping center trade areas,” Land Econ, Vol. 39 No. 1, 1963, pp. 81-90

12. Huff D., “Defining and estimating a trade area,” J Mark, Vol. 28, 1964, pp.34-38

13. Jie N. et al, “A new method for identifying the central businessdistricts with nighttime light radiance and angular effects,” Remote Sens, Vol. 15 No. 1, 2023, p. 239

14. Neethu C. V., Surendran S (2013) “Review of spatial clustering methods“, Int J Inf Technol, Vol. 2 No. 3, 2013, pp. 15-24

15. Reilly W., “The Law of Retail Gravitation,” Knickerbocker Press, 1931

16. Sikoglu E., Kaya H., and Arslan H., “Identification of central business district (cbd)boundaries by space syntax analysis and the case of elazı˘g (turkey),” A–Z ITU JFac, Vol. 17 No. 3, 2020, pp. 115-125

17. Vance Jr. J. E., “Delimitation and analysis of the little rock central business district,” JAAS, Vol. 8, 1955, pp. 181-193

18. Woodbury C., “The size of retail business districts in the chicago metropolitanregion,” J Land Public Util Econ, Vol. 4 No. 1, 1928, pp. 85-91

19. Wu S., “Spatial fuzzy c-means clustering analysis of u.s. presidential election andcovid-19 related factors in the rustbelt states in 2020,” MDPI, Vol. 11 No. 8, 2022, p. 401

20. Avison Young, “Seoul Retail Market Report Q4 2019,” 2019

21. CBRE, “2022 Korea Real Estate Market Outlook Mid-Year Review,” 2022

22. Savills Korea, “2023 Korea Retail Market Outlook,” 2023

23. Cushman, Wakefield, “Seoul retail market report Q1 2023,” 2023

<국문요약>

부동산 거래정보를 이용한 리테일 상권 설정

민 성 훈 (Min, Seonghun)

리테일 시장을 이해하는데 있어서 적절한 공간적 범위를 설정하는 것은 매우 중요하다. 대부분 시장통계가 이러한 상권을 기반으로 조사 및 작성되기 때문이다. 일반적으로 리테일 상권은 고객의 수요와 소매업자의 공급과 같은 사용 관점에서 설정되어 왔다. 반면 본 연구는 부동산의 거래정보인 가격과 거래량에 초점을 맞추어 상권을 설정하고, 두 가지 접근방법을 비교하여 시사점을 도출하였다. 본 연구는 머신러닝에 기반한 공간 군집화 기법인 공간 퍼지 C-평균 알고리즘을 사용하여 서울의 리테일 부동산 거래통계를 분석하고, 이를 사용의 관점에서 설정된 상권과 비교하였다. 그 결과는 다음과 같다: 첫째, 거래기반 접근방법은 핵심 상권과 근린 상권을 성공적으로 구분하였으며, 핵심 상권의 위치 또한 사용기반 접근방법과 유사하였다. 따라서 거래정보만으로도 신뢰할 수 있는 상권 설정이 가능한 것을 알 수 있었다. 둘째, 위치의 유사성에도 불구하고 두 접근방법에 따른 상권의 경계에는 상당한 차이가 있었다. 또한, 이러한 차이로부터 상권의 확장 또는 축소와 같은 시간에 따른 공간적 변화를 식별할 수 있었다. 셋째, 거래기반 상권의 가격지수 상승이 사용기반 상권의 가격지수 상승보다 크게 나타났다. 이러한 차이는 상권을 어떤 방법으로 설정하는 가에 따라 시장통계에 편향이 발생할 수 있다는 것을 암시한다.

주 제 어: 상권, 권역설정, 머신러닝, 공간 퍼지 C-평균, 가격지수

Appendix 1

Regression analysis result based on the two approaches

Variable	SFCM			SGFCM		
	Estimate	t-value	p-value	Estimate	t-value	p-value
(Intercept)	15.1406	133.9451	0.0000	15.1729	138.1861	0.0000
gu2	-0.5898	-16.1459	0.0000	-0.6680	-15.2180	0.0000
gu3	-0.8649	-7.9505	0.0000	-1.0150	-24.3736	0.0000
gu4	-0.7569	-19.6413	0.0000	-1.0222	-29.3470	0.0000
gu5	-0.7582	-35.0865	0.0000	-0.7139	-25.7016	0.0000
gu6	-0.4366	-9.1679	0.0000	-0.6774	-21.4931	0.0000
gu7	-0.7140	-11.2018	0.0000	-0.9708	-14.8733	0.0000
gu8	-0.9346	-15.1882	0.0000	-1.1670	-10.1546	0.0000
gu9	-0.7314	-1.7235	0.0848	-1.0525	-22.5229	0.0000
gu10	-1.2626	-10.8474	0.0000			
gu11	-1.0408	-31.7978	0.0000	-1.0381	-43.5508	0.0000
gu12	-0.4139	-7.1148	0.0000	-0.5382	-10.8838	0.0000
gu13	-0.1756	-5.5478	0.0000	-0.3697	-15.8941	0.0000
gu14	-0.4254	-12.3698	0.0000	-0.4352	-12.9098	0.0000
gu15	-0.2733	-14.9317	0.0000	-0.3263	-16.4121	0.0000
gu16	-0.3268	-6.3833	0.0000	-0.5214	-12.9658	0.0000
gu17	-0.4597	-6.7312	0.0000	-0.6713	-14.6003	0.0000
gu18	-0.4576	-17.1406	0.0000	-0.5948	-31.0532	0.0000
gu19	-0.7079	-4.1078	0.0000	-0.6887	-14.5779	0.0000
gu20	-0.7540	-17.2888	0.0000	-0.9617	-30.2248	0.0000
gu21	-0.0590	-1.8627	0.0626	-0.4126	-12.6152	0.0000
gu22	-0.5481	-5.6354	0.0000	-0.7107	-13.5120	0.0000
gu23	-0.3020	-8.2139	0.0000	-0.5342	-17.8967	0.0000
gu24	-0.6062	-19.2587	0.0000	-0.6686	-22.3573	0.0000
gu25	-0.8636	-16.7632	0.0000	-1.0499	-18.8301	0.0000
zone2				-0.0204	-0.0693	0.9447
zone3	-0.0292	-0.4220	0.6731	0.1409	1.6105	0.1073
zone4	-0.2082	-3.2918	0.0010	0.0213	0.2658	0.7904
zone5	-0.1249	-1.9646	0.0495	0.0812	1.0101	0.3125
zone6	-0.0701	-1.0489	0.2943	0.2204	2.6620	0.0078
zone7	0.1942	1.7931	0.0730	0.2960	2.7941	0.0052
zone8	0.1421	2.1875	0.0287	0.2906	3.5785	0.0003
zone9	0.1710	0.8563	0.3918			
zone10	0.6410	6.5574	0.0000	0.8808	8.2852	0.0000
zone11	-0.1225	-1.5804	0.1141	0.1872	2.1430	0.0322
zone12	0.0814	0.1907	0.8488			
road2	0.1409	11.7128	0.0000	0.1558	14.0599	0.0000
road3	0.0277	2.3214	0.0203	0.0254	2.2719	0.0231
road4	0.0250	1.9015	0.0573	0.0138	1.1783	0.2387
use2	0.0021	0.1661	0.8681	0.0035	0.3110	0.7558
use3	0.1035	2.1778	0.0295	0.1151	2.7803	0.0054
area_land	0.0001	8.8509	0.0000	0.0002	12.0550	0.0000
area_bldg	0.0000	-13.1527	0.0000	0.0000	-13.2554	0.0000
age	0.0054	10.0418	0.0000	0.0047	9.6187	0.0000
A-R2		0.6288			0.6567	