Effects of Spatial Agglomeration on Relative Net Profitability: a prospective look at Seoul’s retail market sector using machine learning

Steven (Jaehun) Song

Seoul National University

Seoul, South Korea

Abstract

This research examines the relationship between clusters of common retail stores and the profitability of the stores in such clusters. Methods in this paper demonstrate that an analysis of clusters with machine learning can prove that spatial agglomeration of retail stores in Seoul has a statistically significant effect on the relative popularity, and thus the relative net profitability of the stores (Singh 2011).The tools used for the analysis are descriptive statistics, Welch’s unequal variance t-test, density-based spatial clustering of applications with noise to find outliers, and K-means algorithm. The findings of this research show that there is a clear difference in the popularity of stores within clusters and the popularity of stores in outlier districts with a statistical significance level of $p < 0.05$.

1 Introduction

1.1 Background

With more than 16 million tourists visiting in 2019 alone (Kang 2019), Seoul, a sprawling metropolis located at the heart of South Korea, is home to several popular attractions for both foreigners and Koreans alike. Walking down one such street, one might also notice a common but peculiar phenomenon concerning the stores around him or her. It would seem as if cafes, restaurants, or competing retail stores were generally gathered in a particular area, and less so in other areas. This economic phenomenon, called spatial agglomeration, is explained by Hoteling’s Model of Spatial Competition and game theory (Haan 2012, Pasidis 2013). In other words, stores cluster together because market competition drives suppliers to choose suboptimal locations to open their stores. Thus, these clusters become a Nash equilibrium.

This phenomenon can be easily visualized today as market competition in the city led to the congregation of markets in certain neighborhoods. In turn, agglomeration of stores caused these neighborhoods to become famous for a set of goods (i.e. Mapo for its galbi [Korean beef ribs], Itaewon for foreign goods, Dongdaemun for apparel, etc.) (Padykula 2019). Such localized specialization defines much of Seoul’s commerce today, especially among the country’s younger populace.

1.2 Problem

The existence of such clusters within competing companies begs the question whether such companies earn more on average than their counterparts that are less clustered together. Particularly, stakeholders looking to open new businesses will often dispute which part of a city they should open in. The rise of localities in Seoul has made this decision process harder since entrepreneurs now have to decide whether to open business in a fiercely competitive neighborhood or risk defaulting in a less famous location. Thus, this paper will use Seoul’s retail markets to assess whether clustering markets holds an inherent economic advantage over individual markets.

2 Data

2.1 Data Acquisition

The data necessary for this research was gathered using three API’s and an open data source. Specifically, the data from Seoul Open Data Plaza provided the *names* of subway stations in Seoul. This was done with the assumption that most economically active neighborhoods are located near stations, since that is where the vast majority of foot traffic takes place. Furthermore, coordinates of these stations were provided by Kakao Geo-Locator API. Nearby venues centered around the stations were provided by Foursquare Places API. Finally, a popularity indicator for each venue using the number of blog posts per search query was provided by Naver Search API.

2.2 Data Cleaning

Locational data from Seoul Open Data Plaza provided the data basis for the purpose of this research. After downloading the names of subway stations, each name was joined with its corresponding coordinates from Kakao Geo-Locator API. Then, using Foursquare Places API, I gathered data on venues nearby the stations using the coordinates from the previous data frame. Specifically, I collected data with constraints of at most 100 venues within a range of 500 meters within a centroid located at these stations. The resulting data frame was *left joined* to the station table.

Moreover, Foursquare API provides a *filter* that allows the developer to look for specific categories of venues. For the purpose of this research, I used four types of filters to split the data into four types of stores. The first filter was a holistic one that returned *all* stores nearby that provided **food**. The second filter returned all types of **restaurants** including but not limited to: Asian restaurants, buffets, barbeques, food courts, European restaurants, steakhouses, theme restaurants, etc. The third filter returned all types of **cafes** nearby, including bubble tea cafes, tea rooms, pet cafes, coffee places, etc. The fourth filter returned all types of **late-night joints** excluding bars, including fried chicken places, wings joint, pizzerias, etc. The four types of data were saved into four respective data frames.

Afterwards, I used a Naver Search API to find the number of blog posts corresponding to a search query containing the *name* of the store, the *location* of the store based on nearby stations, and the *type* of the store. Each query returned a json file containing a list of all blogs related to the query. This research only utilized the length of the json file, hence the number of total blog posts per query, to determine the relative popularity of the store. Future research might consider streamlining the query for higher accuracy and using natural language processing (NLP) to assess the sentiment of the blogs and use a corresponding score as part of the data.

2.3 Feature Engineering

After data cleaning, there were a total of 5642 entries and 93 columns representing the data. However, most of the columns were significant to the data. Thus, some of the techniques of feature engineering used to select significant features were one-hot encoding, grouping, imputation, and handling outliers.

Imputation was required to provide numbers to a small minority of queries that did not return any blog posts when searched for. In this case, I assumed that this phenomenon implied that the venue was either very new or it was not popular enough to garner any such blog reviews. In any case, since this would have a detrimental effect on the popularity of the venue in question, I imputed 0 as the number of blog reviews instead of using the mean number of blog reviews.

One-hot encoding was used to categorize each venue into a specific *type* of store in a binary format so that I could use machine learning algorithms to analyze the data.

Grouping algorithms were used to *tidy* the data for streamlined machine learning algorithms.

Finally, a visual representation of the data showed that there were a few outliers in the number of blog reviews a certain store received. Initially I considered using logarithmic transformation, but the bimodal aspect of the distribution made this impractical. Thus, I removed all entries with blog reviews that exceeded the interquartile range, allowing for a more normal distribution of the data.

The resulting data frame looks as follows:

Neighborhood	Address	Category	Reviews
Station Name, Station longitude, Station latitude	Name of Venue, Address of Venue, Longitude, Latitude	One hot encoded value of category	Number of posts each venue received

Table 1. Feature selection from data cleaning

3 Method

3.1 K-Means Clustering the Data

Since the purpose of this research is to see the effect of clustering on a store's ability to earn money, I start clustering the venues based on their retail type and their locations relative to each other and the distance between the stores and the stations. Since we have the coordinates of the stores and the stations, I simply take the Euclidean distance between the entries for the duration of the *k*-means algorithm.

However, before the data can be effectively clustered, it is imperative that the optimal *k* number of clusters be found for each of the four data frames. I have done this by implementing an *elbow test* where a visual representation of the squared distance loss empirically shows the optimal amount of *k* required for each data frame. In order to do so, the data is Min-Max Scaled, transformed, and put through a series of tests to plot the squared loss.

The result of the experiment is as follows:

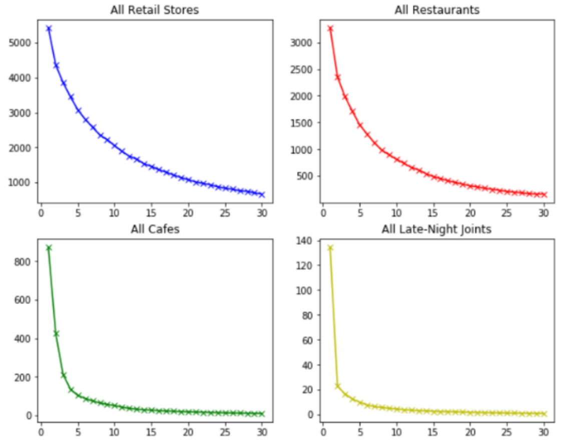
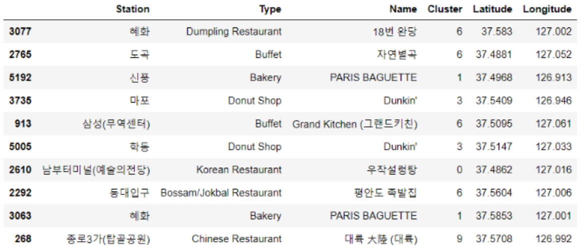


Figure 1. Elbow Test for Finding Optimal *k*

The test empirically gives the optimal *k* for each model. In this case, the optimal *k* for the data frames are approximately 15, 15, 5, 2 for all venues, restaurant venues, café venues, and late-night joint venues respectively. Thus, the *k*-means model for each data frame were trained with these *k* values.

Sampling the data gives a data frame with relevant features extracted that looks like as follows:



	Station	Type	Name	Cluster	Latitude	Longitude
3077	혜화	Dumpling Restaurant	18번 완당	6	37.583	127.002
2765	도곡	Buffet	자연별곡	6	37.4881	127.052
5192	신풍	Bakery	PARIS BAGUETTE	1	37.4968	126.913
3735	마포	Donut Shop	Dunkin'	3	37.5409	126.946
913	삼성(무역센터)	Buffet	Grand Kitchen (그랜드키친)	6	37.5095	127.061
5005	학동	Donut Shop	Dunkin'	3	37.5147	127.033
2610	남부터미널(예술의전당)	Korean Restaurant	우작설렁탕	0	37.4862	127.016
2292	동대입구	Bossam/Jokbal Restaurant	평안도 족발집	6	37.5604	127.006
3063	혜화	Bakery	PARIS BAGUETTE	1	37.5853	127.001
268	종로3가(탑골공원)	Chinese Restaurant	대륙 大陸 (대륙)	9	37.5708	126.992

Table 2. K-Means Cluster Representation of Data

3.2 Finding Outliers With DBSCAN

The next step in assessing the effects of spatial agglomeration on relative profit is to determine what types of stores are categorized as clustered, and what types of stores are categorized as outliers, or not determined to be part of a spatially agglomerated cluster.

This can be done by utilizing a density based spatial clustering of applications with noise (DBSCAN) machine learning algorithm to determine which groups of venues are determined to be either within a cluster, or without a cluster.

First, it is imperative to decide upon an optimal epsilon, or maximum distance threshold for two points to be considered as part of a cluster. This is done with a DMDBSCAN algorithm (Elbatta 2012) with a pseudocode to find the optimal epsilon as follows:

Purpose	*To find suitable values of eps*
Input	*Data set of size n*
Output	*Eps for each varied density*

Procedure	*for i*
	  for j = 1 to n
	    d(i, j) // find distance (i to j)
	    find minimum values of distances to nearest 3
	  end for
	end for
	sort distances ascending and plot
	EPS corresponds to critical change in curves

Figure 2. Pseudocode DMDBSCAN algorithm (Elbatta 2012)

Using this algorithm, I can visually check which values of epsilon would be optimal using the Euclidean distance between points using its coordinates. The plot created to determine the best epsilon result as follows:

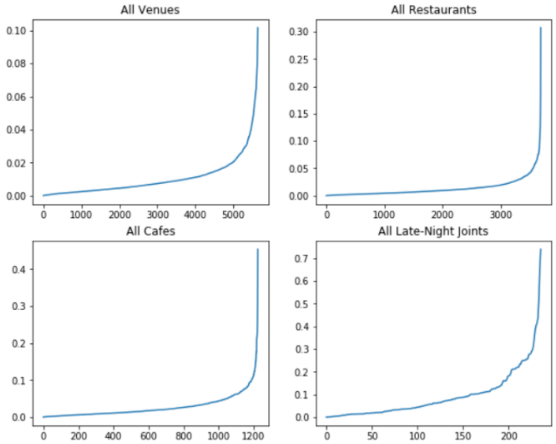


Figure 3. DBMSCAN Plots for Optimal EPS

Thus, empirical evidence shows that the optimal eps is equal to 0.04, 0.05, 0.1 and 0.2 for all venues, all restaurants, all cafes, and all late-night joints respectively. The minimum samples required in each data frame is loosely based upon the number of optimal clusters from the *k*-means clustering algorithm. Thus, the minimum samples required is set to 10, 10, 3, 2 respectively.

Training the models with these parameters give each data frame a label containing either a -1 or a positive integer based on whether it is categorized as an outlier or part of a cluster. Using the python folium library, a map that visually shows these clusters and outliers can be illustrated.

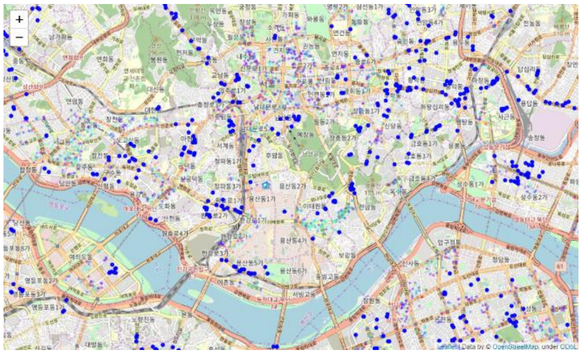


Figure 4. Outliers in All Venues

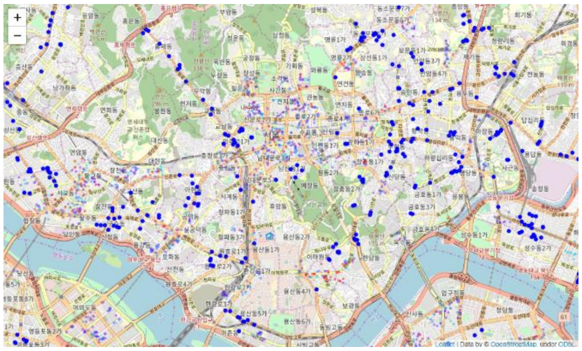


Figure 5. Outliers in Korean Restaurants

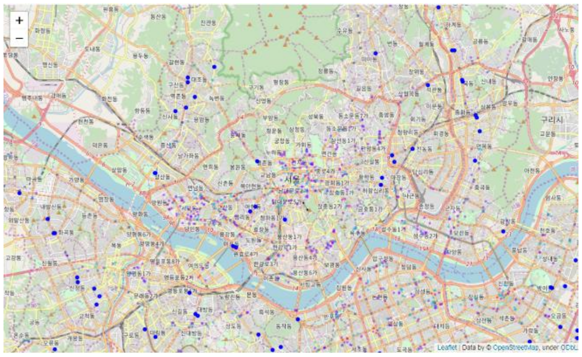


Figure 6. Outliers in all Cafés

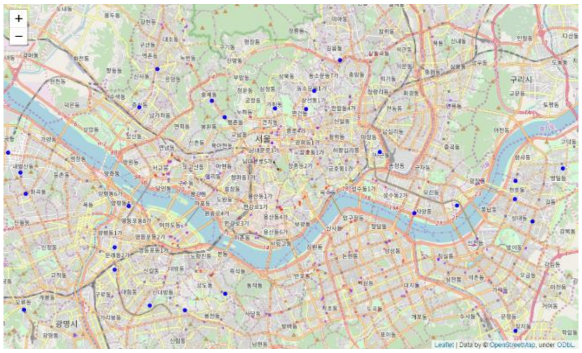


Figure 7. Outliers in all Late-Night Joints

Here, spatial agglomeration is visualized by the multicolor points (color determined by k cluster) within clusters, while the outliers are visualized as a blue dot.

3.3 Statistical Inference

The relationship between the number of reviews a venue receives and its status as being part of a cluster or not can be seen using a box plot for each of the data.

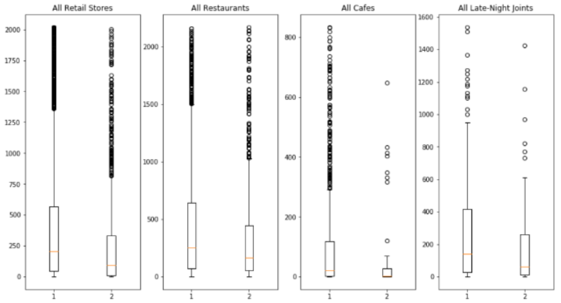


Figure 8. Box Plot of Blog Reviews Per Venue

Seeing that the orange line represents the mean number of blogs, the box plot empirically shows that the number of blogs that a store that is spatially agglomerated (or within a cluster) is generally *higher* than the number of blogs that an outlying venue receives.

In order to see if this difference is statistically significant, I used Welch's unequal variance t-test. This is because the number of samples in each split data was unbalanced, as was the variance. Then, statistical analysis with a one-tailed t-test was used.

The null hypothesis $H_0\colon \mu_1 - \mu_0 \leq 0$ was set to determine if the difference in means were either the same, or if the mean number of reviews of the clustered venues were actually *lower* than the number of reviews in outlying venues. Consequently, the alternative hypothesis became $H_{\alpha}: \mu_1 - \mu_0 > 0$

The t-statistic was calculated using the following equation: $t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}$

The degrees of freedom were also calculated using $v \approx \frac{\left(\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}\right)^2}{\frac{s_1^4}{N_1^2 v_1} + \frac{s_2^4}{N_2^2 v_2}}$ where $v_j = N_j - 1$ is the degrees of freedom associated with the jth variance estimate.

Running the t-test and estimating the average t statistic, p value and the average difference had the following results:

t	p-value	Average diff.[1]
10.185	5.39181e-24	0.50729
6.472	1.20080e-10	0.33406
3.691	0.0003589	1.12421
0.359	0.72119	0.09841

Table 3. Stat Scores for Respective Data Frames

Furthermore, the value counts for each cluster in the data frame containing **all** types of venues that had a p-value below 0.05 included the following:

Type of Venue	Average diff in mean
Bakery	1.46261
Korean Restaurant	0.42768
Café	1.27759
BBQ Joint	0.34261
Vietnamese, Bunsik, Regular[2], Asian, Burgers, Indian, Burrito, etc.	0.33557
Italian Restaurant	0.61667
Japanese Restaurant	1.06998
Chinese Restaurant	0.86983
Seafood Restaurant	1.23311
Fast Food Restaurant	0.55152

4 Results

4.1 Result Analysis

The statistical analysis shows that there is a statistically significant difference between the popularity of markets (determined by the number of blog reviews) and the degree in which they are clustered together. This is increasingly evident in certain types of stores identified by a k-means algorithm that grouped these stores together by common characteristics. Specifically, the results showed that the types of stores that were most affected by spatial agglomeration were as follows:

Data containing *all* retail stores
Korean Restaurants, Bakeries, Cafe's, Vietnamese Restaurants, Bunsik, Asian Restaurant, Burger Joints, Japanese Restaurant, Fried Chicken Joints with an average popularity margin of 50.729%.
Data containing *only* restaurants
Japanese Restaurants, Korean Restaurants, BBQ Joints, Bunsik, (unnamed) Restaurants, Seafood Restaurants, Vietnamese Restaurants with an average popularity margin of 33.406%
Data containing *only* cafes
Cafés, bakeries and donut shops with an average popularity margin of 112.421%

Furthermore, although there was little evidence to reject the null concerning late-night joints, there was ample evidence and a strong correlation to reject the null concerning all other types of venues.

4.2 Discussion

This research provides strong evidence that there is correlation between spatial agglomeration and the respective store's popularity. Thus, the result of this paper should help stakeholders or entrepreneurs who are looking to open new businesses in Seoul determine in which locations they should do so in order to garner the most popularity, and as such, the most profit.

In other words, it is clear from this study that entrepreneurs looking for new places to open their stores should, on principle, find locations in, or nearby, popular neighborhoods where many other stores are clustered together. This is especially true for those who are looking to open a store that coincide with one of the following:

[1] Average difference in mean is determined by $(\mu_1 - \mu_0)/\mu_1$ where μ_1 is the mean number of reviews a clustered venue received.

[2] Regular restaurants are grouped by several venues that sell food yet are not categorized. The number of "regular" is 65, has a p-value of 0.003 and an average difference in mean of 1.68008

Japanese Restaurants, Korean Restaurants, BBQ Joints, Bunsik, (unnamed) Restaurants, Seafood Restaurants, Vietnamese Restaurants, Cafes in general, and Bakeries

5 Conclusion

5.1 Further Research

Further research on this topic may include the influence of real estate price, the net profit of local stores, and the average duration a stores stay open in an area before defaulting as additional parameters to consider before one decides upon a location to open their stores. Furthermore, this research can be refined by implementing an NLP algorithm that determines whether a blog post about a store is considered *positive* or *negative* and use the net number of *positive* blog posts as a secondary parameter. In this case, one would be able to assess the overall social perspective of stores in a general locality and use this as an additional parameter for consideration.

Another recommendation for further research is to use temporal data to find trends in popularity based on spatial agglomeration, and perhaps even create a predictive model that will show *when* the optimal time to open a new store in a particular location.

5.2 Concluding Remarks

In this study, I analyzed the relationship between the common phenomenon of stores clustering together and whether such clustering actually holds an inherent economic benefit compared to stores that do not participate in such clustering. Over the course of this study, I identified the clusters that existed within Seoul, the types of restaurants contained within these clusters, and the relationship between location and the popularity that the store might garner. The results of this research proved that certain types of clustered stores are more popular on average than outlying stores of the same type.

6 References

- Jac de Haan. (2012), “Why do competitors open their stores next to one another?”, Retrieved from: https://www.ted.com/talks/jac_de_haan_why_do_competitors_open_their_stores_next_to_one_another?language=en

- Pasidis I.N. (2013), “Spatial Competition vs Spatial Agglomeration”, Spatial, Transport and Environmental Economics MSc. University of Amsterdam

- Ben Gardiner, et al. (2010), “Does Spatial Agglomeration Increase National Growth?”, Journal of Economic Geography. Pp.1-28

- Anthony Heath, et al. (2005), "By Popular Demand: The Effect of Public Opinion on Income Inequality", Comparative Sociology

- Sanjeet Singh, et al. (2011), "Causal Effect of Advertisement on Profit and Sales", SSRN Electronic Journa, Advertisments and Firm Performance

- Paul Farris, David Reibstein (1979), "How Prices, Ad Expenditures, and Profits are Linked", Harvard Business Review

- Jessica Padykula, (2019), "10 Must-visit Neighborhoods in Seoul", Retrieved from:https://www.tripsavvy.com/guide-to-the-neighborhoods-of-seoul-4147397

- Kang Seung-woo, (Dec 24, 2019), "Number of foreign tourists to hit record high in 2019", The Korea Times

- Nadiah Rahmah, (2013), "Determination of Optimal Epsilon (Eps) Value on DBSCAN Algorithm to Clustering Data on Peatland Hotspots in Sumatra", Earth and Environmental Scienes