Search well and be wise: A machine learning approach to search for a profitable location

Dr. Shuihua Han
Department of Management Science, School of Management,
Xiamen University, Xiamen, China
Email: hansh@xmu.edu.cn

Ms. Xinyun Jia
Department of Management Science, School of Management,
Xiamen University, Xiamen, China
Email: 243283738@qq.com

Ms. Xinming Chen
Department of Management Science, School of Management,
Xiamen University, Xiamen, China
Email: chenxm8967@qq.com

Dr. Shivam Gupta
Department of Information Systems, Supply Chain Management & Decision Support,
NEOMA Business School, 59 Rue Pierre Taittinger, 51100 Reims, France
Email: shivam.gupta@neoma-bs.fr

Dr. Ajay Kumar*
EMLYON Business School, France
Email: akumar@em-lyon.com
(*: Corresponding Author)

Dr. Zhibin Lin
Durham University Business School, Durham University,
Mill Hill Lane, Durham DH1 3LB, United Kingdom
Email: zhibin.lin@durham.ac.uk

Search well and be wise: A machine learning approach to search for a profitable location

Abstract

A good location is critical for the sales performance of a hotel or a restaurant. This study proposes a machine learning-based model for location selection that focuses on the estimation of sales potential of a prospective site and helps to overcome the lack of historical data for the prospective site and the subjective criteria used in conventional models. The proposed model involves three major steps. First, we use an attribute selection algorithm to identify the key factors that contribute to the profitability of a specific location. Second, we evaluate the similarity between the candidate site and the existing stores by using an improved grey comprehensive evaluation method. Finally, we use a kernel regression model to predict the sales potential of the candidate site. A case study of a well-known international restaurant chain is used to illustrate the application of the proposed data-driven model. The results indicate that our proposed model helps to accurately select the most profitable locations.

Keywords: Location selection; Sales prediction; Multi-criteria decision making; Machine learning; Improved Grey comprehensive evaluation

1. Introduction

Selecting a right location that can generate profits is of strategic importance for the success of a restaurant (Chen & Tsai, 2016; Yang, Roehl, & Huang, 2017) or a hotel (Yang, Mao, & Tang, 2018), particularly in today's turbulent business environment (Godinho, Phillips, & Moutinho, 2018; Phillips & Moutinho, 2014). Many chain restaurants such as Chili's, BJ's and Cheesecake Factory are struggling to survive, despite that consumers love eating out regardless of the poor economic situations (Peltz, 2017). Many chain restaurants are closing their stores in poor performing locations, as in the case of Cheesecake Factory (Carreau, 2019). Over 1000 Subway restaurants closed in both 2018 and 2019, and due to the double blow of low sales per location and COVID-19 pandemic, it is said that it has closed as many as 2200 to 2400 restaurants in 2020 (Jonathan, 2021). Starbucks announced that it would close 500 underperforming restaurants and add 850 openings in 2021, aiming to have more profitable locations to improve margins (Jonathan, 2020).

Previous studies have examined hospitality services' location selection patterns and how location determinants shape these patterns (Chen &Tsai, 2016; Pillsbury, 1987; Yang, Roehl, & Huang, 2017; Smith, 1983). Various location models have been proposed and widely adopted in the industry such as gravity model, Huff model, analog model (Yang Luo, & Law, 2014), and in recent years, the models have been more advanced and sophisticated, for example the use of multi-criteria decision-

making (MCDM) methods in location selection (Chang and Hsieh, 2014; Temur, 2016; Velasquez & Hester, 2013). However, most of existing models require the construction of a qualitative evaluation hierarchy or grading locations' features such as analytic hierarchy process (Tzeng et al, 2002) and rough set theory (Chen & Tsai, 2016), which could lead to invalid or biased results. Hospitality mangers expect the selected location to generate substantial revenue, and an accurate sales prediction for the new outlet is thus fundamental (Merino & Ramirez, 2016). A few studies have made location selection decisions by predicting the future sales revenue of the candidate location (e.g. Chen &Tsai, 2016; Ting et al. 2018; Zeng & Tang, 2019). Despite the numerous studies on sales prediction or forecasting (Kourentzes & Petropoulos, 2016; Palmer, Montano, & Sesé 2006), most of them rely on historical sales data, which are unavailable for a new outlet, and as such they are unable to provide an accurate estimation of its sales potential.

Recent advance in computer science, particularly machine learning enables more accurate performance prediction for hospitability services (Law et al., 2019). Adopting a non-parametric technique, we propose a new, integrated, machine learning-based method in this study to accurately estimate the sales potential of new locations, based on a number of location-related determinants of sales performance. Our location selection target is to select the candidate location with highest estimated sales as the best candidate store. Unlike existing location selection methods, we use sales data from the current branches, and obtain sales estimates based on these

branches' similarity to the candidate location. This integrated multi-criteria method consists of three main steps: attribute selection, similarity measurement, and sales potential evaluation. Our approach has three unique features:

a) Kernel regression is used to predict sales, which takes into account the distinctions between different analog stores.

b) An improved grey comprehensive evaluation method is proposed to measure the similarity between an analog store and the candidate location, and then to provide an unbiased weight for the kernel regression model.

c) A more accurate multi-criteria hierarchy is designed for demographic factors, based on characteristics of the business.

To illustrate the superior performance of our proposed method, we conducted a case study of a well-known fast food restaurant chain in China. Based on a selection of 24 analog stores, we used leave-one-out cross-validation to demonstrate that this method can accurately estimate the sales potential of prospective locations. Our empirical illustration demonstrates that our proposed method can help hospitality managers to evaluate a candidate location's potential performance, make operational decisions for each single branch, and assess the overall development of their entire business operation.

The remainder of this paper is organized as follows. In Section 2, we review previous research on location selection. In Section 3, we propose the integrated multi-criteria method for location selection based on machine learning. In Section 4, we use

an empirical study of a fast food chain to demonstrate the validity of the proposed approach and discuss its potential field of application. Section 5 concludes the paper and proposes possible directions for future research.

2. Literature review

2.1 *Location selection theories*

Among the classic location theories are the central place theory (Christaller, 1933), the spatial interaction theory (Reilly, 1931) and the principle of minimum differentiation (Hotelling, 1929). Central place theory, proposed by Christaller (1933), assumes that consumers shop at the nearest place that meet their needs. Road condition and geographic factors are also significant determinants (Kuo, Chi, & Kao, 2002; Chen & Tsai, 2016). As transportation and technology advance, the importance of distance decreases. Researchers are paying more attention to the impact of customers' behavior and the presence of competitors (Karande & Lombard, 2005; Dasci & Laporte, 2005). Spatial interaction theory, developed by Reilly (1931), assumes that customers would trade off the attractiveness of alternative shopping areas against the obstacle effect of distance. It highlights three major factors: demand, retail attractiveness, and market accessibility (Nakaya et al., 2007). Nakaya et al. (2007) adopt spatial interaction models for location analysis with a focus on the consumer behavior by lifestyle group. Recently, Aksoy and Ozbuk (2017) further confirmed that accessibility and convenience of the location are a priority when

customers choose a hotel.

The principle of minimum differentiation, proposed by Hotelling (1929), emphasizes the relationship between stores and competitors, and states that the agglomeration effect plays an important role in stores' location pattern. Agglomeration increases the attractiveness of whole area and decreases the search cost for customers. This effect also exists among hotels and restaurants (Prayag, Landré, & Ryan, 2012; Yang, Wong, & Wang, 2012). Based on these theories, several researchers have summarized the factors for location selection in the hospitality industry. Aksoy and Ozbu (2017) identified three determinants for evaluating hotel location: accessibility, regional development, and tourist attractions. Yang, Roehl, and Huang (2017) found that the location determinants are traffic, population, and market geodemographic. Demographic factors were investigated on a family basis; for example, household income and composition etc., were considered (Yang, Roehl, & Huang, 2017). Different trade area types represent different demographic compositions and purchasing behaviors, which could influence the future operation of new stores. When choosing a new site, managers must take into account the demographic characteristics of their core target customers as well as the general population in the surrounding area (Yang, Roehl, & Huang, 2017).

Market condition should be taken into consideration when designing location selection model. Retail businesses emphasize customer behavior and focus on

patronage behavior prediction. This results in the creation of a great number of mathematical models with the patronage probability being used as a quantified measurement (Li, Y., & Liu, L. 2012). Retail customers are mostly local residents, while hotel customers do not come from the neighborhood, and prefer to stay in a hotel in an area where various services are available (Weaver, 1993). Hence, not only hotels should choose the location near commercial area, public transportation, they also need to consider the provision of the overall distribution blueprint for the region. Researches showed, availability of public service and products including banks, tourism attractions (Shoval & Cohen-Hatab, 2001; Shoval, 2006) play an important role in location selection for hotels. Compared with the hotel and retail industries, the restaurant industry is more diverse with various customer segments, as food represents various cultures and ethnicities. The type of restaurants varies from cafés such as Starbucks and fast food outlets such as McDonald's to full-service luxury restaurants, meeting the different demands of the potential customers in the neighborhood. Hence, not only the type of restaurant would have a significant impact on the location selection (Smith, 1983), the demographical factors of the location are also of paramount importance (Yang, Roehl, & Huang, 2017).

2.2 *Location selection models*

The conventional location models include checklist approach, proximal area model, gravity model, Huff model, analog model, and the recent ones include multi-

criteria decision-making methods (Kuo, Chi, & Kao, 2002). Checklist method is an examination of a list of location factors that are used to determine the revenues and costs of stores in different locations. Proximal area model is based on the sales in similar location to predict the sales at a prospective store. The gravity model uses data of the population and distance to determine trade area boundaries. Huff model considers a number of variables including product variety, store size, travel time and customer preferences (Huff, 1964). Analog model, first proposed by Applebaum (1966), attempts to estimate sales potential of new store by examining similar stores. Analog approach has been well received and widely used by practitioners, often combined with other analytical techniques such as statistical modeling, cluster analysis and decision-tree analysis (Kuo, Chi, & Kao, 2002). Moreover, regression model is often used to predict potential store sales by using predictor variables such as population size, average income, the number of households, direct competitors, and traffic patterns. In the case of hotel location selection, in addition to the agglomeration and the multi-dimensional models, Yang Luo, and Law (2014) reviewed the tourist-historic city model and the mono-centric model. Moreover, the authors further examined several empirical and operational models and call for the adoption more sophisticated models and the use of Geographic Information System (GIS). In fact, GIS has been widely in retail location decision for many years (Tayman & Pol, 1995). It has been combined with machine learning big data analytics in recent years. For example, Wang, Tsai and Lin (2016) used GIS data and retail location theory through

spatial-temporal analysis for selecting bike sharing sites in Taipei. Using geospatial analytics based on data in Malaysia, Ting et al. (2018) showed that sales performance can be estimated and optimal site identified. Rohani and Chua (2018) proposed a model using data from Google map for location analytics to decide an optimal retail site. Using data mining techniques, Zeng and Tang (2019) extracted data of geographic features, commercial area and human movement using several machine learning models to estimate the popularity of a retail site.

Many of the recent location selection models are fairly sophisticated by using MCDM methods, such as multi-attribute utility method, analytic hierarchy process and fuzzy set theory, involving multivariate statistical techniques and mathematical programming models (Velasquez & Hester, 2013). For example, Tzeng et al. (2002) helped a restaurant in Taipei evaluate new locations using analytic hierarchy process to build a location evaluation hierarchy. Chou, Hsu, and Chen (2008) selected 21 determinants for international hotels' location selection, and used Fuzzy MCDM to decide the weight of each criterion. Chang and Hsieh (2014) used the TOPSIS method to handle the multi-criteria such as store, rent cost, site features, and crowds to determine optimal site. Temur (2016) took into consideration of high uncertainty of the decision criteria and proposed a model with "cloud based design optimization" to deal the uncertainty and aid mangers to make site selection decision.

These MCDM methods consider a limited number of criteria, and they need

decision makers to establish a qualitative hierarchy, which is largely subjective and the results are prone to be biased. It would be more reliable by using methods such as the combination of the Huff model with Monte Carlo simulation, as suggested by Merino and Ramirez-Nafarrate (2016). However, due to the limitations of the Monte Carlo simulation method, the model might not be suitable for a dynamic market that features rapid development and intense competition. Chen and Tsai (2016) constructed a data-mining framework to support location selection, applying rough set theory to location selection for restaurant chains and predicting future sales. However, these methods suffer from inadequate location selection indicators and require large scale of sample data, thus have limited practical use.

2.3 *Machine learning models*

Thanks to the recent advancement of computer science and information technology, location selection studies based on multi-criteria decision making methods have attempted data-mining tools and machine learning to improve sales prediction accuracy. Machine learning techniques, which integrate artificial intelligence systems, seek to effectively perform a specific task without using explicit instructions, relying on patterns and inference instead. Empirical studies using machine learning commonly have two main phases: training model and test model performance (Xiao, Xiao, Lu, & Wang, 2014). Machine learning involves many basic techniques such as artificial neural networks (ANNs), support vector machines

(SVMs), and random forests (RFs) (Henrique, Sobreiro, & Kimura, 2019).

Machine learning, as a key “new age technologies”(Kumar, Ramachandran & Kumar, 2020), have been used in a wide variety of applications, such as email filtering, and computer vision. In recent years, increasing attention has been paid to using machine learning techniques in the sectors of tourism and hospitality: the analysis of online reviews (Martinez-Torres & Toral, 2019; Taecharungroj & Mathayomchan, 2019; Xiang, Du, Ma, & Fan, 2017), the analysis photos and thus studying tourists’ behavior (Deng & Li, 2018; Giglio, Bertacchini, Bilotta, & Pantano, 2019; Zhang, Chen, & Li, 2019), the prediction of tourist arrival volumes (Law, Li, Fong, & Han, 2019; Rice, Park, Pan, & Newman, 2019; Sun, Wei, Tsui, & Wang, 2019), the analysis of determinants of tourism spending (Brida, Lanzilotta, Moreno, & Santiñaque, 2018), and the location selection for a hotel (Yang, Tang, Luo, & Law, 2015) or a retail site (Ting et al. 2018; Zeng & Tang, 2019). These aforementioned studies have considerably illustrated the potential application of machine learning, and thus have greatly inspired us to apply machine learning techniques to further improve location selection methods for hospitality services.

3. Proposed method

Different locations have different commercial environments that can have a significant impact on the store’s sales performance. It is critical to combine all the relevant information from these branches to evaluate the sales potential of a candidate

location. If the candidate location is similar to an analog store, it is likely to have similar sales to this analog store (Han et al., 2014). Thus, the key point of the sales potential evaluation is to measure the similarity between two locations and weight different locations based on their similarity. We therefore propose an integrated multi-criteria method that combines an improved grey comprehensive evaluation method with kernel regression. Improved grey comprehensive evaluation method can help to obtain the similarity of analog stores to the candidate location, and the results of kernel regression are based on the difference between two elements, as such it is suitable for its adoption in the proposed method.

Our proposed method consists of three steps. First, we choose analog stores similar to the candidate location from other branches, and use the Correlation-based feature subset selection (CfsSubsetEval) algorithm (Hall, 1998) and Best-First search method for attribute selection to pick out the criteria with stronger influence on the sales performance of the locations. Second, we use Improved grey comprehensive evaluation method to obtain the similarity of each analog store to the candidate location. Finally, we obtain the sales potential of the candidate location using kernel regression. Managers can make a location selection decision by comparing the estimated sales potential with their expectation. The methodological framework is shown in Figure 1.

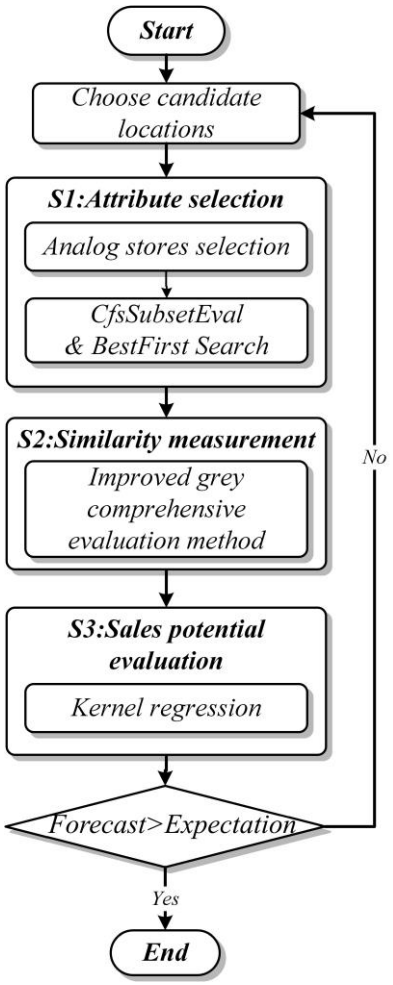

Figure 1. The methodological framework

3.1 *Attribute selection*

We initially selected 40 attributes to be examined in our empirical model, including 26 that are based on a review of previous studies and 14 that are derived from interviews with decision makers at the restaurant chain. Figure 2 shows a concept framework of criteria classification, where we classify the criteria into seven categories: basic information, fixed population, floating population, transportation, store features, competitors, and historical sales. The detailed evaluation hierarchy is shown in Table A1 in Appendix A.

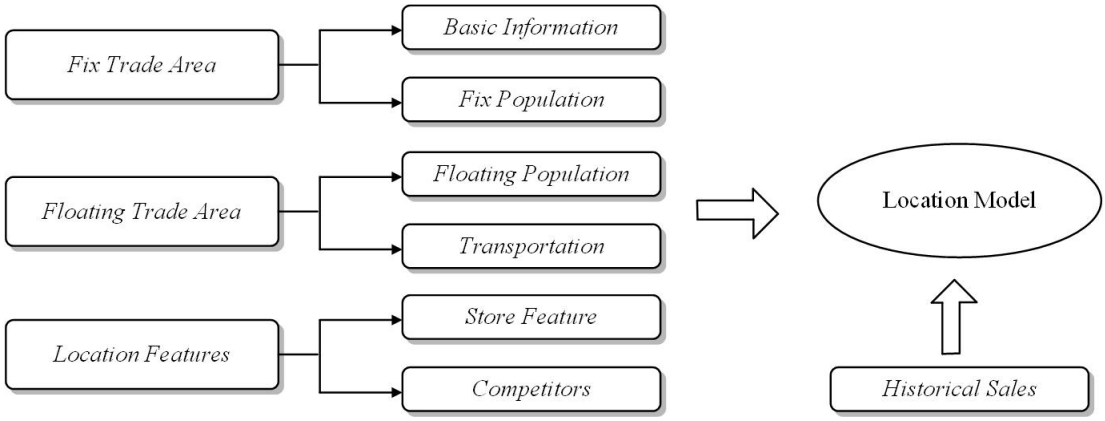

Figure 2. The conceptual framework of criteria classification

We classify the fixed population as residents, office workers, and students to enrich the demographic factors, and use transportation and pedestrian flow to describe the floating population, thus matching the characteristics of the restaurant industry. Correspondingly, We captured the demographic characteristics of potential customers by classifying trade areas into five different types: residental areas, working areas, school areas, commercial areas, and mass transportation areas. In largely residential trade areas, most potential customers will be residents; in largely commercial areas, most sales will come from shoppers. Furthermore, different types of customers have different purchasing behaviors and consumption levels, which will influence sales.

Based on this classification, we classify the store types into five types according to the type of trade area where they are located. Among them, residental type, school type and office type are fixed population stores, while commercial type and mass transportation type belong to floating population stores. Accordingly, we select the analog stores with the same store types as the candidate location to collect data. In this way, we ensure all the stores investigated are comparable.

Although all 40 attributes are identified, there is no need to consider all attributes in this method. Too many attributes leads to redundancy or inaccuracy. Therefore, in this study, we select attributes based on the actual data of analog stores and the candidate location, and remove less influential attributes. In this way, we improve the efficiency of the subsequent estimation steps and increase the accuracy of the results.

3.2 *Similarity measurement*

After determining the appropriate attributes, we obtain the similarity of analog stores to the candidate location, and consider all the influential attributes, this is because using the kernel regression method for predicting the sales of the candidate location requires the evaluation of the similarity between the candidate location and the existing branches. To achieve this, improved grey comprehensive evaluation method is introduced into our proposed method. This method sorts the system by analyzing the correlation of elements to an optimal value. Usually, we would choose the maximum or minimum value of a criterion to be the optimal one (Deng, 1989). However, because we want to know how similar the analog stores are to the candidate location, we improve this method by regarding the candidate location as optimal. Thus, the similarity value of the candidate location would be 1, the highest, and the value of analog stores would represent their similarity to the candidate location.

To avoid biases brought about by subjective weighting, the entropy method (Shannon, 1948) is used. This method can objectively determine the weights of

criteria by obtaining the information entropy of each one. The lower the information entropy, the greater the impact of this criterion on the results.

Let n be the number of attributes, m be the number of the analog stores, Let P denote the similarity value vector, $P=[p_0, p_1, ..., p_m]$. Then the similarity value of the analog store i can be calculated using Eq. 1:

$$p_i = \sum_{k=1}^{n} \xi_k^i \cdot \omega_k \qquad (1)$$

where ξ_k^i represents the correlation coefficient between attribute k of store i and of the candidate location ($k = 1, 2, ..., n$, $i = 0, 1, ..., m$), and $i = 0$ represent the candidate location. Let w_k represents the weight of attribute k . The improved grey comprehensive evaluation method process are shown in Appendix B.

P represents the similarity of all stores to the candidate location. Because all the stores are compared with the candidate location, it is reasonable that the similarity value of the candidate location would be 1. Using this method, we can comprehensively evaluate the overall condition of each location to obtain the similarity of analog stores to the candidate location, to prepare for the final prediction.

3.3 *Sales potential evaluation*

The evaluation of potential sales performance of the candidate location is based on the results from similarity measurement. Specifically, the sales potential is the weighted average of the sales of analog stores last year; an analog store with greater similarity should be given greater weight, which means that it has a greater impact on

predicted sales. Note that in kernel regression, the weight of each element is decided by its distance from the element to be estimated. As a non-parametric technique, kernel regression is used to estimate the conditional expectation of a random variable; it can be adapted to various hospitality businesses without large samples or fitting any distribution pattern.

In this model, sales of the candidate location Q_0 are obtained based on the sales Q_i of analog store i and its similarity value P_i . Based on the kernel regression model presented by Nadaraya (1965), the model is constructed as shown in Eq. 2:

$$Q_0 = \hat{f}(x) = \sum_{i=1}^{n} \frac{K(\frac{p_0 - p_i}{h})Q_i}{\sum_{i=1}^{n} K(\frac{p_o - p_i}{h})} \tag{2}$$

where K is the kernel function and h $(h>0)$ represents the bandwidth. The bigger h is, the smaller the variance and the larger the deviation. Moreover, function K is usually specifically formed as a Gaussian function (Eq. 3):

$$K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} \tag{3}$$

It is clear that in Eq. 2, the weights of analog stores are determined by both the bandwidth h and the differences between the similarity values of analog stores and the candidate location, as shown in Eq. 4:

$$w_i = \frac{K(\frac{p_0 - p_i}{h})}{\sum_{i=1}^{n} K(\frac{p_0 - p_i}{h})} = \frac{K(\frac{d_i}{h})}{\sum_{i=1}^{n} K(\frac{d_i}{h})} \tag{4}$$

From Eq. 4, we can see that there is a negative correlation between w_i and the

difference d_i. That is to say, the more similar the analog store is to the candidate location, the greater the weight it is given.

Once the other parameters are determined, it is crucial to determine the bandwidth h. Based on least squares cross-validation, which is suitable for small samples, the best bandwidth should minimize the mean squared error (MSE), shown in Eq. 5:

$$MES(h)=\frac{1}{n}\sum_{i=1}^{n}(\hat{f}_h(x_i)-y_i)^2 \tag{5}$$

where

$$\hat{f}_h(x_i)=\frac{\sum_{i=1,\,j\neq i}^{n} K(\frac{p_0-p_i}{h})y_i}{\sum_{i=1,\,j\neq i}^{n} K(\frac{p_o-p_i}{h})} \tag{6}$$

To find the optimal bandwidth, the hybrid bandwidth selection methodology (Silverman, 1986) is adopted. According to this method, the best bandwidth is between $[0.25h_0,1.5h_0]$. h_0 is calculated using Eq. 7:

$$h_0=(\frac{4}{3n})^{\frac{1}{5}}\sigma \tag{7}$$

where σ is the standard deviation of p_i, and n is the number of analog stores.

After we determine the bandwidth, the sales estimation of the candidate location can be calculated using Eq. 2. In this way, we can obtain a precise evaluation of the sales potential of this new location. When this evaluation is combined with cost considerations, decision makers can determine whether to select this location for a

new business.

4. Empirical investigation

This section presents an empirical study of a world-renowned fast food chain that planned to expand its business in Yibin, Sichuan Province, China, and wanted to ensure the candidate location could bring in high profits. The chain had more than 2,400 stores scattered across China in 2016 and continued to grow. The chain currently used gravity model (Anderson et al. 2010) is adopted for location section.

The candidate location was located on the first floor of a shopping plaza. This plaza had an area of about 120,000 square meters; it featured over 800 parking spaces. In addition, it was on a street corner with high accessibility and convenient transportation. The core trade area was mainly upscale residential, and thus potential customers were mostly residents with high incomes and consumption levels. In terms of the presence of competitors, there were five restaurants in the plaza and more in the core trade area. Figure 3 shows the map of this core trade area. In general, this location was highly suitable, and thus it was a good candidate for evaluation of its sales potential.

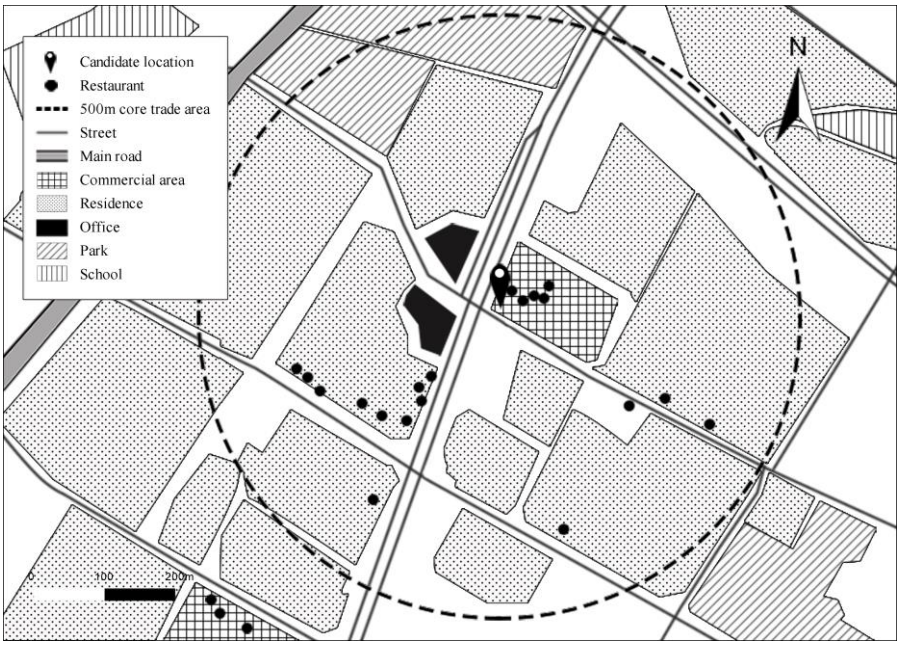

Figure 3. The map of the candidate location

4.1 ***Data collection***

Since the trade area was mostly residential, the candidate location can be classified as residential type, and the original analog stores with residental type are selected among the same chain store. It should be noted that the higher the similarity with the candidate store, the greater the impact on the sales prediction. Therefore, by calculating the similarity of the original analog stores with the candidate location store, we select the top 24 stores with a similarity of more than 30% as the final analog stores for analysis. All criteria were investigated using either questionnaires or the company's internal database. Demographic factors were collected from the neighborhood committee. For pedestrian volume, we selected several spots with large pedestrian volume and recorded the volume every 15 minutes during the shopping plaza's business hours.

4.2 ***Attribute selection***

To identify factors with the greatest influence on sales, we used the CfsSubsetEval evaluator and Best-First search in Weka. As a result, nine attributes

were selected: per capita annual disposable income in the trade area (A1), urban per capita annual disposable income (A2), population size of the trade area (A3), number of residents in the trade area (A4), number of residents within 500 meters (A5), distance to the nearest bus stop (A6), number of entrants (A7), number of seats in the restaurant (A8), number of seats at the counter (A9). Note that the number of residents was included, which means that residents had a greater impact on sales in this location. This confirmed the correctness of our classification. The corresponding criteria values of the candidate location and analog stores are shown in Table 1.

Table 1. Corresponding criteria values of stores

	A1	A2	A3	A4	A5	A6	A7	A8	A9
The candidate location	20000	23392	116000	65000	18000	400	1	90	70
1	25000	23392	98900	54700	2000	300	2	113	66
2	18659	23392	106000	56000	4000	600	2	90	0
3	21000	23392	144040	124500	9450	400	1	90	110
4	18788	14430	4200500	17799	1550	30	3	270	270
5	23152	14430	241050	51530	4100	50	1	167	131
6	13755	14430	235300	49503	3200	80	3	136	136
7	18788	14430	187300	17799	1700	25	1	208	208
8	24705	14430	125000	75330	5100	35	1	130	24
9	19830	23392	125020	90250	6725	500	2	90	55
10	20906	18911	173525	53765	4050	325	2	129	66
11	21682	18911	115500	65665	4550	318	2	110	12

12	22076	18911	192545	88015	6775	225	1	129	121
13	22853	18911	134520	99915	7275	218	1	110	67
14	23929	14430	183025	63430	4600	43	1	149	78
15	21894	18911	2149700	36250	1775	165	3	192	168
16	24076	18911	169975	53115	3050	175	2	140	99
17	19378	18911	167100	52102	2600	190	3	125	101
18	21894	18911	143100	36250	1850	163	2	161	137
19	20970	14430	2220775	34665	2825	40	2	219	201
20	16272	14430	2217900	33651	2375	55	3	203	203
21	18788	14430	2193900	17799	1625	28	2	239	239
22	18454	14430	238175	50517	3650	65	2	152	134
23	20970	14430	214175	34665	2900	38	1	188	170
24	16272	14430	211300	33651	2450	53	2	172	172

4.3 ***Similarity measurement***

After attribute selection, the entropy method was used to decide the weights of each attribute. The information entropy and weight of each selected attribute can be calculated using Eq. B2 and Eq. B3, as shown in Table 2. Information entropy can be used to decide whether the attributes are effective in identifying distinctions among analog stores. From Table 2, we can see that all nine attributes were effective.

Table 2. Information entropy and weight of each criteria

Criteria	Information entropy	Weight
A1	0.96	0.03
A2	0.75	0.17

A3	0.62	0.27
A4	0.90	0.07
A5	0.86	0.10
A6	0.81	0.13
A7	0.84	0.11
A8	0.89	0.07
A9	0.93	0.05

Improved grey comprehensive evaluation method was then adopted to determine the similarity of analog stores to the candidate location. After calculating correlation coefficients using Eq. B1, we can obtain similarity values using Eq. 1, which are displayed in Table 3. The results indicate that 19 out of 24 analog stores had a similarity value higher than 0.5. Store No. 3 had the highest similarity to the candidate location, with a similarity value of 0.894; Store No. 4 was the most dissimilar, with a similarity value of 0.380.

Table 3. Similarity values of analog stores

No.	Similarity value	No.	Similarity value
1	0.803	13	0.753
2	0.796	14	0.697
3	0.894	15	0.493
4	0.380	16	0.671

5	0.662	17	0.670
6	0.593	18	0.651
7	0.632	19	0.463
8	0.691	20	0.439
9	0.826	21	0.446
10	0.724	22	0.616
11	0.734	23	0.648
12	0.733	24	0.589

4.4 *Sales potential evaluation*

Once similarity values are obtained, least squares cross-validation can be used to search the best bandwidth. The standard deviation of similarity values of analog stores was 0.13; thus, $h_0 = 0.03$ according to Eq. 7. The best bandwidth was therefore in the range $[0.008, 0.045]$. We then iterated through this interval with a step size of 0.001. It is clear that MSE was minimized when the bandwidth was 0.034. Under this bandwidth, the weight of each analog store can be obtained using Eq. 4. Finally, Eq. 2 can be used to predict the sales potential of the candidate location to be RMB￥5868994.83 per year.

Based on quantified sales potential, decision makers can compare different candidate locations and obtain a better understanding of their strengths. Decision makers should take costs into account, including rent, utilities, and wages, when

estimating profits. However, compared with sales, cost information is easier to obtain. If the results of the evaluation do not meet the company's expectations, decision makers should select another candidate location and repeat the whole procedure.

4.5 *Performance evaluation*

We applied leave-one-out cross-validation to the 24 analog stores. That is, we estimated the sales potential of one location by treating the other 23 stores as analog stores. We compared our method with the regression model as well as the gravity model (Anderson et al. 2010), as shown in Figure 4. As Figure 4 shows, the proposed method was better able to measure the sales potential of candidate locations, thus providing a stronger basis for location selection decisions. Besides, we provide the Mean Square Error (MSE) and the Mean Absolute Percentage Error(MAPE) metrics to quantitatively prove the superiority of the proposed method in sales prediction. Table 4 shows that the MSE and the MAPE of the proposed method(3.3257E+12, 10.7%) is lower than that of the gravity model(7.17857E+12, 22.5%), which indicates the proposed method can more accurately predict the actual sales potential than the gravity model.

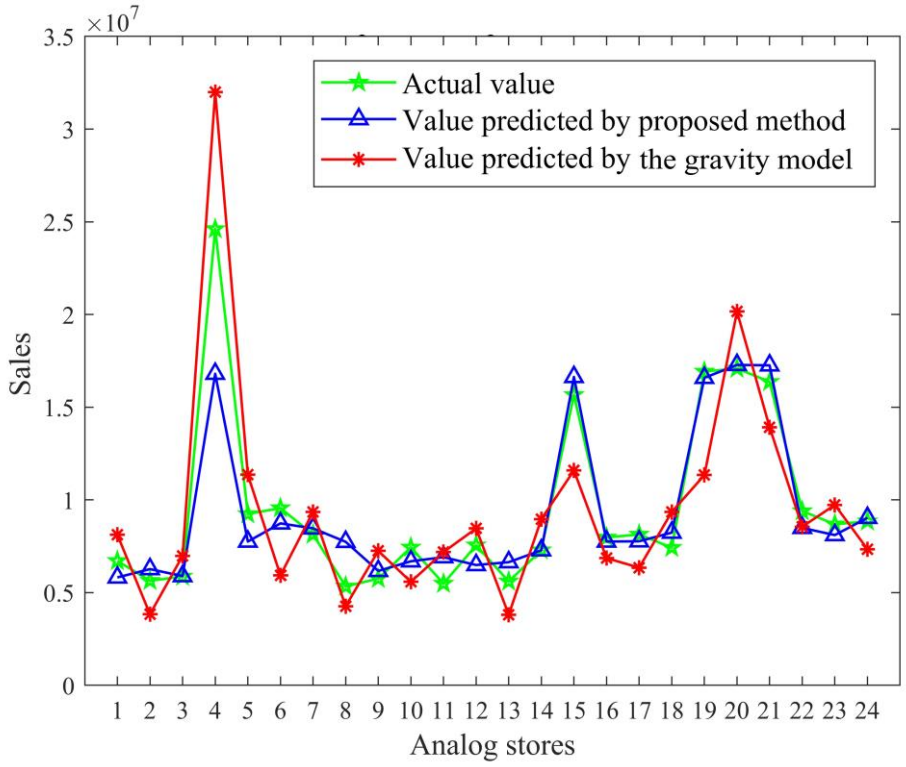

Figure 4. Comparison between the results of the proposed method and the gravity model

Table 4. Prediction evaluation metrics and Error Statistics

	MSE	MAPE	Error statistics	
			Mean	Std. deviation
The proposed method	3.3257E+12	10.7%	1023933.483	1541514.722
The gravity model	7.17857E+12	22.5%	2192012.175	1573802.763

To further analyze the performance of the proposed method and the gravity model, the Wilcoxon signed rank test were conducted to determine whether there any significant differences in the errors of each method. Table 4 shows the mean and standard deviations of the errors generated by the proposed method and the gravity model, while Table 5 displays the results of the Wilcoxon signed rank test. Based on

Table 4 and Table 5 below can be seen that the mean and standard deviation of the proposed method(M=1023933.483, SD=1541514.722) are lower than those of the gravity model(M=2192012.175, SD=1573802.763). Moreover, the test results indicated that the proposed method generates significantly lower errors than the gravity model (Z=-3.600, p<0.05).

Table 5. Wilcoxon Signed Rank test on the difference in the errors of the proposed method and the gravity model

	The proposed method vs. The gravity model
Z	-3.600[b]
Asymptotic. Sig. (2-tailed)	0.000

[b.] *Based on negative ranks.*

4.6 ***Survey analysis***

In June 2015, the new store was open at the candidate location. To verify the robustness of our method and whether this location met our expectations, we conducted the survey of passers-by at the new store using questionnaires. The questionnaire contained four parts: personal information (gender, age, occupation, residence and monthly income), actions before (transportation, time spent, last stop), actions after (transportation, next stop), the reason for being here. The survey took place from June 10 to 11, 2016, and 400 questionnaires were collected.

From the survey, we can identify where our customers come from. As shown in Figure 5, the store location was at the junction of South Shore West District and South

Shore East District, which led to almost half of customers coming from these two districts. Besides, some of customers come from Upstream North District and Downstream North District, which were nearby urban areas. The distribution of customers' residence shows that the residents might be the majority of the customers for these restaurants. Hence, the sales generated from the residents could have a stronger impact on store performance than other types of customers, such as office workers and students.

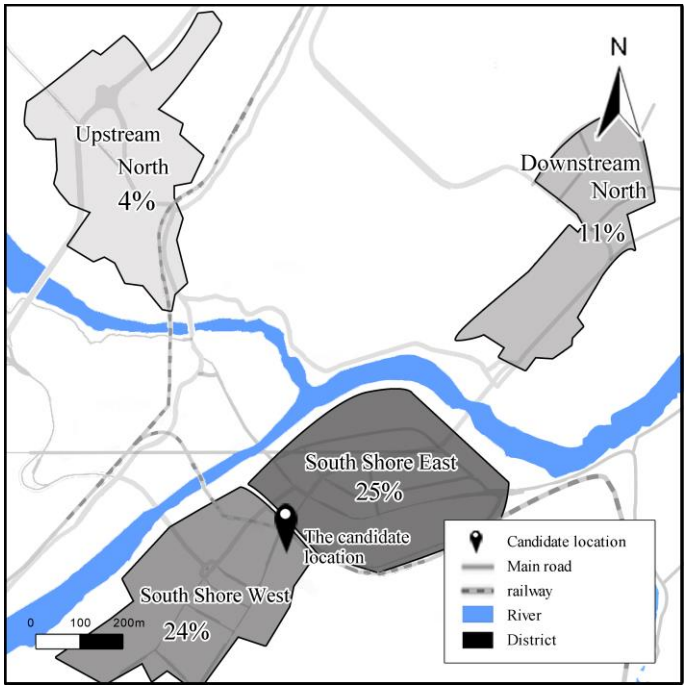

Figure 5. The distribution of customers' residence

A1 and A2 are under the category Basic Information and show the income of customers might have a strong influence on the sales of restaurants. In the survey, we found the average monthly income of customers investigated were under ￥4000, which is less than the urban average income of ￥4163 in 2016 (Obtained from 2016 Yibin Government Report). 36% customers had a monthly income of between ￥2500

and ￥5000, and 30% customers earned less than ￥2500 per month. This result indicated that this new store is attractive to customers who have low income.

A6 is under the category Public Transportation, indicating that the customers in this area prefer the restaurants with the bus stops nearby. This can be supported by the survey: more people came to and leave the store location by bus. Meanwhile, we noticed that customers also preferred cycling and walking, suggesting the store has great accessibility.

A7, A8 and A9 are under the category Store Features, indicating that the number of entrants and the number seats have a significant effect on the attractiveness of restaurants in this area.

5. Discussion and conclusions

5.1 *Theoretical contributions*

Selecting the right locations is of strategic importance for a hospitability company to succeed in the increasingly fierce competition. In this study, we proposed an integrated location selection method to overcome the shortcomings of extant location selection methods such as incomplete measurements and less robust methodologies. In a real case study, we demonstrated the model's new capabilities by applying the model to a well-known international restaurant chain to facilitate its location selection. The results show that our proposed method offers a more accurate evaluation of new locations' sales potential and thus contribute to success of the store's performance.

Ultimately, this study investigated the consistency between the actual demographic characteristics obtained by a questionnaire and the demographic characteristics assumed by the model situation to further verify the importance of the selected attributes.

This research contributed to the current knowledge of restaurant location selection in several ways. First, we used kernel regression based on the data from other existing analog branches to predict the sales potential of the candidate location. In contrast to the commonly used forecasting methods that rely on historical sales data, our proposed methods can help to solve the problem for a potential store that has no historical data. Meanwhile, given that the service and operation strategies of each branch in the same chain company are basically identical, using data from other branches can guarantee consistency among internal factors and thus ensures the accuracy of estimation, which makes the proposed method adaptable to various types of hospitality services.

Secondly, our proposed machine-learning model combines improved grey comprehensive method with kernel regression. An improved grey comprehensive evaluation method is proposed to measure the similarity between an analog store and the candidate location, and then to provide an unbiased weight for the kernel regression model. As a result, our proposed model avoids the biases caused by subjective weighting and scoring.

Finally, we classify the fixed population as residents, office workers, and students to enrich the demographic factors, and use transportation and pedestrian flow to describe the floating population, thus matching the characteristics of the restaurant industry. Moreover, we classify the location into five types according to the demographic characteristics of potential customers: living areas, working areas, school areas, commercial areas, and mass transportation areas, then select branches similar to the candidate location type and obtain a more accurate estimation. Considering that demographics of the residents are a prominent factor for the restaurant industry, this new classification contributes the greater accuracy of estimation.

5.2 *Managerial implications*

Our model also provides several important management insights to restaurant investors and managers. First of all, through the machine learning method proposed in this paper, managers can supervise and control the operation of new stores. Managers can take into account the changing external environment (such as people flow, number of competitors, etc.) to reassess the existing branches regularly, and adjust their operation strategies according to the analysis results. Secondly, the prediction model can be used to establish the growth evaluation system of restaurants. By using this model, we can take the existing open annual report data of restaurants with good operation as the training set, evaluate and compare the growth potential of start-ups. Finally, the theoretical model can still be applied to chain enterprises in other

industries by selecting different evaluation criteria according to the characteristics of other industries. Specifically, for example, the hotel industry pays more attention to traffic factors (such as convenience of public facilities and public safety, good road network, traffic convenience, etc.)”

5.3 *Limitation and future research*

Several limitations of our study should be highlighted. First, we only focus on the revenue of individual stores, without considering the impact of the new store’s opening on other stores in the same chain. A location selected as optimal may not be optimal to the chain as a whole once stores open, because different branches might cannibalize each other’s customers (Ghosh & Craig, 1983). Further research could include the estimation of the changes of sales of the neighboring stores in the same chain to consider the effect of cannibalization on sales prediction. Second, new trend, such as new taste, new delivery mode and new consumption mode may affect the sales of different candidate location. Although the new trend can be identified by questionnaire, it can be further taken into account in our model in the future. Finally, future research could continue applying new technologies in computer science to improve our proposed location selection method.

References

Aksoy, S., & Ozbuk, M. Y. (2017). Multiple criteria decision making in hotel location: Does it relate to postpurchase consumer evaluations?. *Tourism Management Perspectives*, *22*, 73-81.

Anderson, S. J., Volker, J. X., & Phillips, M. D. (2010). Converse's Breaking-Point Model Revised. *Journal of Management and Marketing Research*, 3, 1.

Applebaum, W. (1966). Methods for determining store trade areas, market penetration, and potential sales. *Journal of Marketing Research, 3*(2), 127-141

Assaf, A. G., Josiassen, A., & Agbola, F. W. (2015). Attracting international hotels: Locational factors that matter most. *Tourism Management*, 47, 329-340.

Brida, J. G., Lanzilotta, B., Moreno, L., & Santiñaque, F. (2018). A non-linear approximation to the distribution of total expenditure distribution of cruise tourists in Uruguay. *Tourism Management*, 69, 62-68.

Carreau, G. (2019). The full list of restaurant chains that are closing in 2019. Available online at https:// twentytwowords.com/ the-full-list- of-restaurant- chains-that-are-closing-in-2019/ https://www.rd.com/food/fun/fast-food-chain-closing-more-locations/accessed on May 06, 2019.

Chang, H. J., & Hsieh, C. M. (2014). A TOPSIS model for chain store location selection. *Review of Intgerative Business and Economics Research*, 4(1), 410-416.

Chen, L. F., & Tsai, C. T. (2016). Data mining framework based on rough set theory to improve location selection decisions: A case study of a restaurant chain. *Tourism Management*, 53, 197-206.

Chou, T. Y., Hsu, C. L., & Chen, M. C. (2008). A fuzzy multi-criteria decision model for international tourist hotels location selection. *International Journal of Hospitality Management*, *27*(2), 293-301.

Christaller, W. (1933). *Central Places in Southern Germany*. In C. W. Baskin (Ed.). translated by C. Baskin, 1966, Englewood Cliffs, NJ: Prentice-Hall.

Dasci, A., & Laporte, G. (2005). A continuous model for multistore competitive location. *Operations Research*, *53*(2), 263-280.

Deng, J. (1989). Introduction to grey system theory. *Journal of Grey system*, *1*(1), 1-24.

Deng, N., & Li, X. (2018). Feeling a destination through the "right" photos: A machine learning model for DMOs' photo selection. *Tourism Management*, 65, 267-278.

Erbıyık, H., Özcan, S., & Karaboğa, K. (2012). Retail store location selection problem with multiple analytical hierarchy process of decision making an application in Turkey. *Procedia-Social and Behavioral Sciences*, 58, 1405-1414.

Ghosh, A., & Craig, C. S. (1983). Formulating retail location strategy in a changing environment. *Journal of Marketing*, *47*(3), 56-68.

Giglio, S. , Pantano, E. , Bilotta, E. , & Melewar, T. C. . (2019). Branding luxury hotels: evidence from the analysis of consumers' "big" visual data on tripadvisor. *Journal of Business Research*. https://doi.org/10.1016/j.jbusres.2019.10.053

Godinho, P., Phillips, P., & Moutinho, L. (2018). Hotel location when competitors may react: A game-theoretic gravitational model. *Tourism Management*, 69, 384-396.

Hall, M.A. (1998). Correlation-Based Feature Subset Selection for Machine Learning, *University of Waikato*.

Han, S. , Ye, Y. , Fu, X. , & Chen, Z. (2014). Category role aided market segmentation approach to convenience store chain category management. *Decision Support Systems*, 57, 296-308.

Henrique, B. M., Sobreiro, V. A., & Kimura, H. (2019). Literature review: Machine learning techniques applied to financial market prediction. *Expert Systems with Applications*, 124, 226-251.

Hotelling, H. (1929). Stability in competition. *Economic Journal*, *39*(153), 41.

Ho, H. P., Chang, C. T., & Ku, C. Y. (2013). On the location selection problem using analytic hierarchy process and multi-choice goal programming. *International Journal of Systems Science*, *44*(1), 94-108.

Hoch, S. J., Kim, B. D., Montgomery, A. L., & Rossi, P. E. (1995). Determinants of

store-level price elasticity. *Journal of marketing Research*, *32*(1), 17-29.

Huff, D. L. (1964). Defining and Estimating a Trading Area. *Journal of Marketing*, *28*(3):34-38.

Jonathan, M. (2020). Starbucks will close another 100 U.S. locations. Available online at https://www.restaurantbusinessonline.com/financing/starbucks-will-close-another-100-us-locations/ accessed on Oct 30, 2020.

Jonathan, M. (2021). Subway continued closing locations last year. Restaurant Business. Available online at https://www.restaurantbusinessonline.com/financing/subway-continued-closing-locations-last-year/ accessed on Jan 19, 2021.

Karande, K., & Lombard, J. R. (2005). Location strategies of broad-line retailers: an empirical investigation. *Journal of Business Research*, *58*(5), 687-695.

Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence*, *97*(1-2), 273-324.

Kourentzes, N., & Petropoulos, F. (2016). Predictioning with multivariate temporal aggregation: The case of promotional modelling. *International Journal of Production Economics*,181, 145-153.

Kumar, V. & Ramachandran, Divya & Kumar, Binay. (2020). Influence of new-age technologies on marketing: A research agenda. *Journal of Business Research.* https://doi.org/10.1016/j.jbusres.2020.01.007

Kuo, R. J., Chi, S. C., & Kao, S. S. (2002). A decision support system for selecting convenience store location through integration of fuzzy AHP and artificial neural network. *Computers in Industry, 47*(2), 199-214.

Law, R., Li, G., Fong, D. K. C., & Han, X. (2019). Tourism demand forecasting: A deep learning approach. *Annals of Tourism Research*, 75, 410-423.

Liu, X., Singh, P. V., & Srinivasan, K. (2016). A structured analysis of unstructured big data by leveraging cloud computing. *Marketing Science, 35*(3), 363-388.

Li, Y., & Liu, L. (2012). Assessing the impact of retail location on store performance: A comparison of Wal-Mart and Kmart stores in Cincinnati. *Applied Geography*, *32*(2), 591-600.

Martinez-Torres, M. R., & Toral, S. L. (2019). A machine learning approach for the identification of the deceptive reviews in the hospitality sector using unique attributes and sentiment orientation. *Tourism Management*, 75, 393-403.

Melaniphy, J. C. (1992). *Restaurant and Fast Food Site Selection*. John Wiley & Sons, Inc..

Merino, M., & Ramirez-Nafarrate, A. (2016). Estimation of retail sales under competitive location in Mexico. *Journal of Business Research*, *69*(2), 445-451.

Nadaraya, E. A. (1965). On non-parametric estimates of density functions and regression curves. *Theory of Probability & Its Applications*, *10*(1), 186-190.

Nakaya, T., Fotheringham, A. S., Hanaoka, K., Clarke, G., Ballas, D., & Yano, K.

(2007). Combining microsimulation and spatial interaction models for retail location analysis. *Journal of Geographical Systems*, *9*(4), 345-369.

Palmer, A., Montano, J. J., & Sesé, A. (2006). Designing an artificial neural network for predictioning tourism time series. *Tourism Management*, *27*(5), 781-790.

Peltz, J. (2017). Americans still love eating out. So why are restaurants like Chili's, BJ's and Cheesecake Factory struggling? Los Angeles Times. Available online at: https://www.latimes.com/business/la-fi-agenda-casual-dining-20170918-story.html accessed on May 06 2019.

Phillips, P., & Moutinho, L. (2014). Critical review of strategic planning research in hospitality and tourism. *Annals of Tourism Research*, 48, 96-120.

Pillsbury, R. (1987). From Hamburger Alley to Hedgerose Heights: Toward a model of restaurant location dynamics. *The Professional Geographer*, *39*(3), 326-344.

Prayag, G., Landré, M., & Ryan, C. (2012). Restaurant location in Hamilton, New Zealand: clustering patterns from 1996 to 2008. *International Journal of Contemporary Hospitality Management*, *24*(3), 430-450.

Reilly, W. J. (1931). *The Law of Retail Gravitation*. New York: WJ Reilly.

Rice, W. L., Park, S. Y., Pan, B., & Newman, P. (2019). Predictioning campground demand in US national parks. *Annals of Tourism research*, 75, 424-438.

Rohani, A. M. B. M., & Chua, F. F. (2018, May). Location Analytics for Optimal Business Retail Site Selection. In *International Conference on Computational*

Science and Its Applications (pp. 392-405). Springer, Cham.

Sánchez-Franco, M.J., Navarro-García, A., Rondán-Cataluña, F.J. (2019). A naive bayes strategy for classifying customer satisfaction: a study based on online reviews of hospitality services. *Journal of Business Research, 101*(8), 499-506.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*(3), 379-423.

Shoval, N., & Cohen-Hattab, K. (2001). Urban hotel development patterns in the face of political shifts. *Annals of Tourism Research*, 28(4), 908-925.

Shoval, N. (2006). The geography of hotels in cities: An empirical validation of a forgotten model. *Tourism Geographies*, 8(1), 56-75.

Silverman B W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall.

Smith, S. L. (1983). Restaurants and dining out: geography of a tourism business. *Annals of Tourism Research*, *10*(4), 515-549.

Sun, S. L., Wei, Y. J., Tsui, K. L., & Wang, S. Y. (2019). Predictioning tourist arrivals with machine learning and internet search index. *Tourism Management*, 70, 1-10.

Taecharungroj, V., & Mathayomchan, B. (2019). Analysing TripAdvisor reviews of tourist attractions in Phuket, Thailand. *Tourism Management*, 75, 550-568.

Tayman, J., & Pol, L. (1995). Retail site selection and geographic information systems. *Journal of Applied Business Research*, 11(2), 46.

Temur, G. T. (2016). A novel multi attribute decision making approach for location decision under high uncertainty. *Applied Soft Computing*, 40, 674-682.

Timor, M., & Sipahi, S. (2005). Fast-food restaurant site selection factor evaluation by the analytic hierarchy process. *The Business Review, Cambridge*, *4*(1), 161-167.

Timoshenko, A., & Hauser, J. R. (2019). Identifying customer needs from user-generated content. *Marketing Science*, *38*(1), 1-20.

Ting, C. Y., Ho, C. C., Yee, H. J., & Matsah, W. R. (2018). Geospatial analytics in retail site selection and sales prediction. *Big data*, *6*(1), 42-52.

Tzeng, G. H., Teng, M. H., Chen, J. J., & Opricovic, S. (2002). Multicriteria selection for a restaurant location in Taipei. *International Journal of Hospitality Management*, *21*(2), 171-187.

Velasquez, M., & Hester, P. T. (2013). An analysis of multi-criteria decision making methods. *International Journal of Operations Research*, *10*(2), 56-66.

Wang, J., Tsai, C. H., & Lin, P. C. (2016). Applying spatial-temporal analysis and retail location theory to public bikes site selection in Taipei. *Transportation Research Part A: Policy and Practice*, 94, 45-61.

Weaver, D. B. (1993). Model of urban tourism for small Caribbean islands. *Geographical Review*, 134-140.

Xiang, Z., Du, Q., Ma, Y., & Fan, W. (2017). A comparative analysis of major online review platforms: Implications for social media analytics in hospitality and

tourism. *Tourism Management*, 58, 51-65.

Xiao, Y., Xiao, J., Lu, F., & Wang, S. (2014). Ensemble ANNs-PSO-GA approach for day-ahead stock e-exchange prices predictioning. *International Journal of Computational Intelligence Systems*, *7*(2), 272-290.

Yang, Y., Luo, H., & Law, R. (2014). Theoretical, empirical, and operational models in hotel location research. *International Journal of Hospitality Management*, 36, 209-220.

Yang, Y., Wong, K. K., & Wang, T. (2012). How do hotels choose their location? Evidence from hotels in Beijing. *International Journal of Hospitality Management*, *31*(3), 675-685.

Yang, Y., Tang, J., Luo, H., & Law, R. (2015). Hotel location evaluation: A combination of machine learning tools and web GIS. *International Journal of Hospitality Management*, 47, 14-24.

Yang, Y., Roehl, W. S., & Huang, J. H. (2017). Understanding and projecting the restaurant scape: The influence of neighborhood sociodemographic characteristics on restaurant location. *International Journal of Hospitality Management*, 67, 33-45.

Yang, Y., Mao, Z. X., & Tang, J. Y. (2018). Understanding guest satisfaction with urban hotel location. *Journal of Travel Research*, 57, 243-259.

Zeng, J., & Tang, B. (2019, May). Mining heterogeneous urban data for retail store

placement. In *Proceedings of the ACM Turing Celebration Conference-China* (p. 53). ACM.

Appendix A

Table A1：Evaluation hierarchical structure for restaurant location selection

Category	Attribute	Description	Sources
Basic information	Disposable household income in the trade area	Annual disposable Income per household in the trade area	Assaf, Josiassen, & Agbola (2015)
	Urban disposable household income	Urban annual disposable Income per household	Assaf, Josiassen, & Agbola (2015)
	Population size of the trade area	Population size of the trade area	Central place theory
	Population age distribution	Age distribution of population in the trade area	Melaniphy (1992)
	Rating of trade area	The overall rating of sales potential of the trade area	Managerial experience
	Housing prices	The housing price level of the trade area	Land use model
Fixed population	Number of residents in trade area	The number of residents living in the trade area	Central place theory
	Number of residents within 500m	The number of residents living within 500m around the store	Managerial experience
	Sales(%) from residents	The percentage of sales from residents in the trade area	Managerial experience
	Number of office workers in trade area	The number of office workers working in the trade area	Managerial experience
	Number of office workers within 500m	The number of office workers working within 500m around the store	Managerial experience
	Sales(%) from office workers	The percentage of sales from office workers in the trade area	Managerial experience
	Number of students in trade area	The number of students studying in the trade area	Managerial experience
	Number of students within 500m	The number of students studying within 500m around the store	Managerial experience
	Sales(%) from students	The percentage of sales from students in the trade area	Managerial experience
Public transportation	Number of subway lines	The number of subway lines passing through the trade area	Central place theory Gravity model

	Number of bus stops nearby		The number of bus stops within 100m around the store	Central place theory Gravity model
	The frequency of bus routes nearby		The frequency of bus routes within 100m around the store	Central place theory Gravity model
	Nearness to bus stop		The distance between the store and the closest bus stop	Kuo, Chi, & Kao (2002)
Competitors	Number of competitors		The number of competitors within 500m around the store	Spatial competition model
Store features	Visibility	Closeness to sidewalk	The distance to the sidewalk	Management experience
		Sign visibility	The visibility of the sign of the store	Timor & Sipahi (2005)
		Store visibility	The visibility of the stores, including the advertising signs	Kuo, Chi, & Kao (2002)
	Accessibility	Number of entrances	Number of entrances of the store	Managerial experience
		Store Accessibility	Whether the store is easy to find in the building	Pillsbury (1987) Yang, Roehl, & Huang (2017）
	Area & seats	Store size	The size of the store	Ho, Chang, & Ku (2013)
		Number of floor levels	The number of floor levels of the store	Managerial experience
		Checkout counter floor level	The floor level which the checkout counter is on	Managerial experience
		Number of seats inside	The number of seats inside the store	Managerial experience
		Number of seats outside	The number of seats outside the store	Managerial experience
		Number of seats on checkout counter floor	The number of seats on the floor level which the checkout counter is on	Managerial experience
	Parking area	Own parking capacity	The number of parking spaces owned by the store	Kuo, Chi, & Kao (2002)
		Public parking capacity	The number of available public parking spaces within 50m around the store	Erbıyık. H. Özcan & Karaboğa. K. (2012).
		Convenience of parking lot	Whether it is convenient to find and use the parking lot	Central place model Gravity model
Historical sales	Per customer transaction		Average purchase of per customer per transaction in the store	Hoch et al. (1995)
	Total sales		Sales of the store last year	Managerial experience
Floating population	Pedestrian Volume		The number of pedestrians passing by the store	Kuo, Chi, & Kao (2002)
	Pedestrian capturing rate		The percentage of pedestrians purchasing in the store	Retail chain management theory
	Traffic flow		The number of vehicles passing by the store	Kuo, Chi, & Kao (2002)
	Vehicle capturing rate		The percentage of motorists purchasing in the	Retail chain management

		store	theory

Appendix B

Similarity measurement

Let n be the number of attributes, m be the number of the analog stores, j_k^i be the value of attribute k of store i ($k=1,2,...,n$, $i=0,1,...,m$), and $i=0$ represent the candidate location. Let p_i be the evaluation value of store i . The improved grey comprehensive evaluation method process is as follows.

1） Decide the optimal set

Define matrix D as follows:

$$D=\begin{bmatrix} j_1^0 & j_2^0 & \cdots & j_n^0 \\ j_1^1 & j_2^1 & \cdots & j_n^1 \\ \cdots & \cdots & \cdots & \cdots \\ j_1^m & j_2^m & \cdots & j_n^m \end{bmatrix}$$

Because we consider the candidate location as the optimal store, we let $j_k^*=j_k^0$, and add the vector $\left[j_0^*, j_1^*,..., j_n^* \right]$ to the first row of the matrix D. Then we name this new matrix D^*:

$$D^*=\begin{bmatrix} j_1^* & j_2^* & \cdots & j_n^* \\ j_1^0 & j_2^0 & \cdots & j_n^0 \\ \cdots & \cdots & \cdots & \cdots \\ j_1^m & j_2^m & \cdots & j_n^m \end{bmatrix}=\begin{bmatrix} j_1^0 & j_2^0 & \cdots & j_n^0 \\ j_1^0 & j_2^0 & \cdots & j_n^0 \\ \cdots & \cdots & \cdots & \cdots \\ j_1^m & j_2^m & \cdots & j_n^m \end{bmatrix}$$

2） Normalize

In the matrix D^*, dividing j_k^i by j_k^*, we can map D^* to the interval (0,1) and get the matrix C :

$$C=\begin{bmatrix} C_1^* & C_2^* & ... & C_n^* \\ C_1^0 & C_2^0 & ... & C_n^0 \\ ... & ... & ... & ... \\ C_1^m & C_2^m & ... & C_n^m \end{bmatrix}$$

3） Calculate the correlation coefficient $\xi_i(k)$

We set the vector $[C_1^*, C_2^*, ..., C_n^*]$ as the reference vector, and $[C_1^i, C_2^i, ..., C_n^i]$ as comparison vector. Then we use the relative analysis method to obtain the correlation coefficient between attribute k of store i and of the candidate location, according to Eq. B1:

$$\xi_k^i = \frac{\min\limits_i \min\limits_k \left|C_k^* - C_k^i\right| + \rho \max\limits_i \max\limits_k \left|C_k^* - C_k^i\right|}{\left|C_k^* - C_k^i\right| + \rho \max\limits_i \max\limits_k \left|C_k^* - C_k^i\right|} \tag{B1}$$

where $\rho \in [0,1]$ is the resolution ratio, usually set as 0.5. The introduction of ρ is meant to reduce the impact of extremum value.

4） Obtain the weight matrix using the entropy method

We then use the entropy method to obtain weight matrix $W = (w_1, w_2, ..., w_n)$. w_k represents the weight of attribute k (shown in Eq. B2):

$$w_k = \frac{1 - e_k}{\sum_{k=1}^{n} (1 - e_k)} \tag{B2}$$

where e_k represents the information entropy of attribute k (shown in Eq. B3).

$$e_k = -\frac{1}{\ln(m)} \sum_{i=1}^{m} \frac{x_{ik}}{\sum_{i=1}^{m} x_{ik}} \ln\left(\frac{x_{ik}}{\sum_{i=1}^{m} x_{ik}}\right) \tag{B3}$$

5） Calculate the similarity value

Let P denote the similarity value vector, $P=[p_0, p_1, ..., p_m]$. Then the similarity

value of the analog store i can be calculated using Eq. B4:

$$p_i = \sum_{k=1}^{n} \xi_k^i \cdot \omega_{\mathrm{k}} \tag{B4}$$