Received 11 March 2019
Revised 10 April 2019
Accepted 25 April 2019

Intelligent site selection for bricks-and-mortar stores

Dongdong Ge
Research Institute for Interdisciplinary Sciences and Shanghai Institute of International Finance and Economics, Shanghai University of Finance and Economics, Shanghai, China
Luhui Hu
Yonghui Superstores Co. Ltd, Shanghai, China
Bo Jiang
Research Institute for Interdisciplinary Sciences, Shanghai University of Finance and Economics, Shanghai, China
Guangjun Su
Cardinal Operations, Shanghai, China and Department of Management Science and Engineering, Stanford University, Stanford, California, USA, and
Xiaole Wu
School of Management, Fudan University, Shanghai, China

Abstract

Purpose – The purpose of this paper is to achieve intelligent superstore site selection. Yonghui Superstores partnered with Cardinal Operations to incorporate a tremendous amount of site-related information (e.g. points of interest, population density and features, distribution of competitors, transportation, commercial ecosystem, existing own-store network) into its store site optimization.
Design/methodology/approach – This paper showcases the integration of regression, optimization and machine learning approaches in site selection, which has proven practical and effective.
Findings – The result was the development of the "Yonghui Intelligent Site Selection System" that includes three modules: business district scoring, intelligent site engine and precision sales forecasting. The application of this system helps to significantly reduce the labor force required to visit and investigate all potential sites, circumvent the pitfalls associated with possibly biased experience or intuition-based decision making and achieve the same population coverage as competitors while needing only half the number of stores as its competitors.
Originality/value – To our knowledge, this project is among the first to integrate regression, optimization and machine learning approaches in site selection. There is innovation in optimization techniques.

Keywords Optimization, Site selection, Convenience store, Data-driven intelligent decision

Paper type Case study

1. Introduction

Yonghui Superstores Co., Ltd, founded in 1998, is characterized by its expertise in managing fresh food. Currently, it operates three types of stores: Yonghui Bravo, Yonghui Super Species and Yonghui Life. Yonghui Bravo is akin to a traditional supermarket, while the other two store types follow the so-called new retailing style and consumers can

Modern Supply Chain Research and Applications
Vol. 1 No. 1, 2019
pp. 88-102
Emerald Publishing Limited
2631-3871
DOI 10.1108/MSCRA-03-2019-0010

purchase both online and offline. Yonghui Life is a convenience store that sells fresh produce, and is much smaller in size than Bravo. A Yonghui Super Species store combines the features of a Yonghui Life store and a restaurant that mainly provides fresh food. Yonghui's goal is to design the network of all types of Yonghui stores in a city to offer convenience for most residents. To this end, opening a number of Yonghui Life stores would be the most economical and efficient. When the project began, Yonghui owned 101 Yonghui Life stores in Shanghai and was planning to open more to cover a larger population. Site selection is among the most important decisions that determine the profitability of a store. Traditional site selection in the industry is heavily reliant on a tremendous amount of work from the site selection team to collect all potential regions' relevant information. However, overall, such information might be fragmented. Furthermore, site decisions can be highly dependent on the team's subjective judgment. This is referred to as manual site selection, which is inefficient, and the performance of a selected location is difficult to forecast. As information technology has advanced, substantial amounts of location-related information have been digitalized, and some is available from third-party information providers. Data-based site selection is the second approach to the location decision and involves how to best combine the external macro and location-related data with a firm's internal data to select an ideal location. It may also use competitors' information and integrate certain business logic to improve the decision. Albeit much more efficient, data-based site decision still relies much on human judgment during the selection process. The second approach can be further upgraded by applying cutting-edge optimization and machine learning techniques, and empowering machines with decision-making capability. This will result in the third approach to site selection, which we refer to as intelligent site selection. In its collaboration with Cardinal Operations, Yonghui seeks to develop an intelligent site selection system.

In the retail industry, easily implementable models such as gravity models (e.g. Buckner, 1998) and multiple regression models (e.g. Breheny, 1988) are widely employed in site selection. Interested readers are referred to Rogers (2007) and Ladle *et al.* (2009) for surveys of more practical methods of site selection. In terms of general principles, Buckner (1998) identified three major trends and the advanced statistical techniques used in retail site selection and Rogers (2007) proposed some key elements of retail site selection, including competition, transportation convenience, economic development and popularity.

Some researchers (e.g. Berman and Krass, 2002) studied the optimal location problem as a covering problem if maximizing the population to be served is the only objective. A more related problem is the so-called facility location problem (FLP), where the demand of each client must be served by one or more open facilities and the total cost of opening facilities and for serving demand is minimized. FLP is a well-known NP-hard problem in combinatorial optimization. Therefore, the existing literature mainly focused on designing efficient approximation algorithms (e.g. Shmoys *et al.*, 1997; Zhang *et al.*, 2005; Aggarwal *et al.*, 2010; Li, 2013). In particular, an algorithm is said to be a (polynomial) ρ-approximation algorithm for a minimization problem if the algorithm runs in polynomial time and outputs a solution that has a cost at most $\rho > 1$ times the minimal cost for any instance of the problem, where ρ is called the approximation ratio of the algorithm. The currently known best approximation ratio for unconstrained FLP is 1.488 (Li, 2013) and Sviridenko (2002) proved that the ratio cannot be better than 1.463 unless $P = NP$. For capacitated FLP, the best known ratios are three with uniform capacity (Aggarwal *et al.*, 2010) and $5.83+\epsilon$ with nonuniform capacity (Zhang *et al.*, 2005).

Recently, Karamshuk *et al.* (2013), Wu *et al.* (1991) and Xu *et al.* (2016) used data out of information technology to help select the store locations. For examples, Roig-Tierno *et al.* (2013) and Karamshuk *et al.* (2013) used geographic information system (GIS) and consumption information to solve the problem of retail store placement. Wang *et al.* (2016)

took advantage of user-generated reviews to construct predictive features for assessing the attractiveness of candidate locations. Wu *et al.* (1991), Xu *et al.* (2016) and Rohani and Chua (2018) showed that mining map query data are also helpful in site selection. Moreover, Yilmaz *et al.* (2017) proposed an optimal location predictor which accepts partial information about customer locations and returns a location for the new facility.

Building upon existing work, Yonghui and Cardinal Operations aim to select potential sites for hundreds of Yonghui Life stores in Shanghai by integrating methodologies from statistics, optimization and machine learning. There are two major challenges in this project. The first is limited existing store data. In Shanghai, there are only data from 101 Yonghui stores available for prediction and most Yonghui consumers purchase offline, which makes it difficult to characterize their purchasing behavior. Therefore, the data-driven approaches in the recent literature are usually infeasible. The second challenge is to take a network perspective on the problem. Note that one of our objectives is to maximize the total population covered by all Yonghui stores. The site candidates are scattered throughout the city. The procedure of ranking all candidates and then selecting a few best ones is not applicable because it is very likely that some population will be covered by several stores in this procedure. As a result, we need to conduct a network optimization given all site candidates, which results in a challenging large scale mixed integer programming. In this paper, we detail how we address these challenges and develop an intelligent site selection system.

2. Solution framework

The solution for intelligent site selection consists of three major modules: business district scoring (BDS), intelligent site engine (ISE) and precision sales forecasting (PSF). Figure 1 provides an overview of the solution and its three modules. The BDS module is designed to make the initial assessment of each region's potential as a store site. This involves two major steps: first, by using the data of the 101 existing Yonghui Life stores across Shanghai and the city's location-based big data, we can establish, through regression models, the relationship between a store's key performance indicators (KPIs) and the surrounding factors such as points of interest (POI), population flow and population characteristics (such as gender, income, housing, consumption level and education background). KPI usually refers to sales but can also be the number of visitors during the expansion stage under competition. Second, for each new region without a Yonghui Life store, we apply this relationship and the new region's data to estimate the sales potential of the region. The outputs of the BDS module are scores for potential regions based on KPIs, which serve

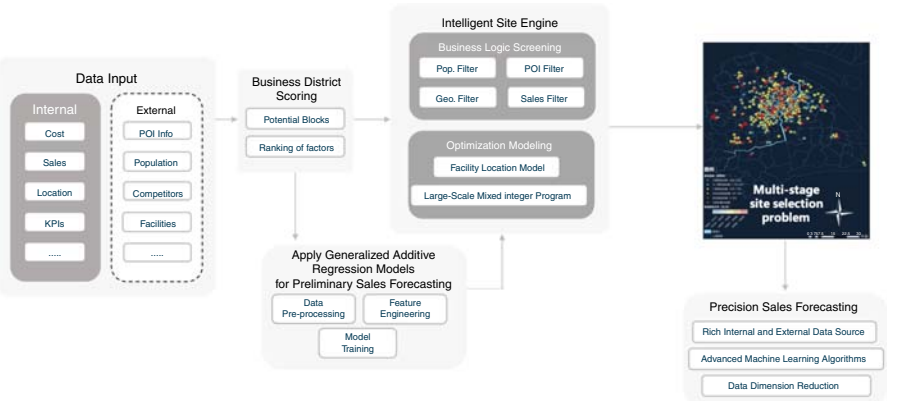

Figure 1.
An overview of the intelligent site selection system

as inputs of the ISE. In addition, the ISE, a semi-autonomous optimization model, needs an explicit relationship between the KPIs and surrounding factors. Therefore, in the BDS, we cannot apply some advanced machine learning algorithms, like random forest and neural network, for preliminary sales forecasting, because such approaches cannot deliver an explicit relationship on how the surrounding factors affect the KPIs. Thus, we build a simple parametric regression model to predict the sales performance based on some basic and significant features.

Population coverage and sales are two major measures used in site selection. These two measures are not always consistent, especially during the expansion stage. In that case, should we cluster stores in regions with high sales or spread them to achieve wider population coverage? By applying multi-objective integer programming, the ISE module derives an optimal set of regions that balance the two objectives. There are also certain constraints considered in the ISE module. For example, the distance between two stores should not be shorter than 500 meters. As the offline stores support both the online and offline systems, the store network design is crucial for Yonghui's ecosystem. Given limited resources and budget, the ISE module also prioritizes the regions to expand the store network.

The regions suggested by the ISE module need to be checked by the retail property development team. The retail developers visit each suggested region and identify potential sites within the region to locate a store. The PSF, as an additional tool to help with the developers and managers, mainly focuses on the prediction accuracy rather than an explicit characterization of the effects of factors on sales. The sales forecast conducted in the PSF module includes greater detail than the assessment of each region's potential in the BDS module. Using a specific address, the PSF module can take into account more detailed information such as the distance to major streets, store size and the surrounding business environment. All of the factors considered amount to over 100 dimensions of data. In addition, we can also use more sophisticated machine learning algorithms, such as neural network and random forest, in order to pursue higher prediction accuracy.

3. Three modules

3.1 Business district scoring

Data quality and dimensions directly influence the accuracy of model predictions. The first step of the BDS module is to establish a city database that includes five data streams: transportation convenience, consumption profile, commercial ecosystem, population flow and competitor profile. The details of these five streams are shown in Table I. We choose these five data streams after extensive trial analysis and back-and-forth discussions with Yonghui's team of developers. In addition to these five data streams, there are some other dimensions we thought might influence the sales of a store, but they turn out to be inessential after the analysis, such as average housing price or number of primary schools within a region. On the other hand, data availability is a constraint. For example, population flow at the entrance of a store should be an important dimension, but such data is not available at third-party data providers.

Data stream	Key dimensions
Transportation convenience	Subways, Bus stations, Entrance/exit of highway
Consumption profile	Surrounding housing prices, rents and age
Commercial ecosystem	Commercial complex/facilities, per capita commercial area
Population flow	Daytime/nighttime population flow, population profile
Competitors profile	Nearby convenience stores, supermarkets, fresh markets

Table I. Key dimensions of each data stream

For site selection, we first need to define a basic unit of area for evaluation. Consider Shanghai as an example. The city is divided into more than 100,000 blocks that are 250 × 250 meters. A block is the smallest unit for site selection and is assigned the associated city data including the five data streams. Among these over 100,000 blocks, the BDS module suggests the blocks with the greatest potential for store sites.

The main task for the BDS module is preliminary sales forecasting. There were 101 Yonghui Life stores in the city when the project began. Each store is located in a single block as defined above. Based on these 101 blocks' associated data of the five streams and the corresponding stores' historical sales, we use a generalized additive regression model to learn how each environmental factor influences a block's sales. The relationship is then applied to forecast the potential sales of the blocks without a Yonghui store based on those blocks' associated data. In the Yonghui Intelligent Site Selection System, each block's preliminary forecasted sales is referred to as each block's business district score. Figure 2 illustrates the logic of the preliminary sales forecasting.

To analyze the impact of environmental factors on sales, we first considered linear regression models, but their predictive power was not satisfactory. Therefore, we turned to a generalized additive model. There are two challenges: First, we have 101 existing stores, so the sample size is not large; second, there are over one hundred potential regressors. Since some variables are counted based on the distance to a store, say 200, 300 or 500 meters, these variables may be highly collinear. In such instances, we use a stepwise approach to select the most relevant potential regressors.

After specifying the regressors, the generalized additive model results in a function f that describes how each factor influences potential sales. This function will be used in the second module, the ISE, which accounts not only for sales but also population coverage. In addition, the BDS module also results in a ranking of all blocks in terms of potential sales. Based on this ranking, the second module can select a pool of candidate blocks in its optimization by applying certain rules, such as the blocks with sales exceeding a certain level.

3.2 Intelligent site engine

The ISE returns a set of blocks for site selection by solving an optimization problem. Its execution relies on the BDS module in two aspects. First, the BDS module helps to eliminate blocks with small populations, including the most remote areas where Yonghui stores could not survive, from our decision set. Second, as mentioned above, the BDS module provides the ISE module with a mapping f describing the relationship between the potential sales and selected environmental factors of a block, which is used in the optimization problem (P)

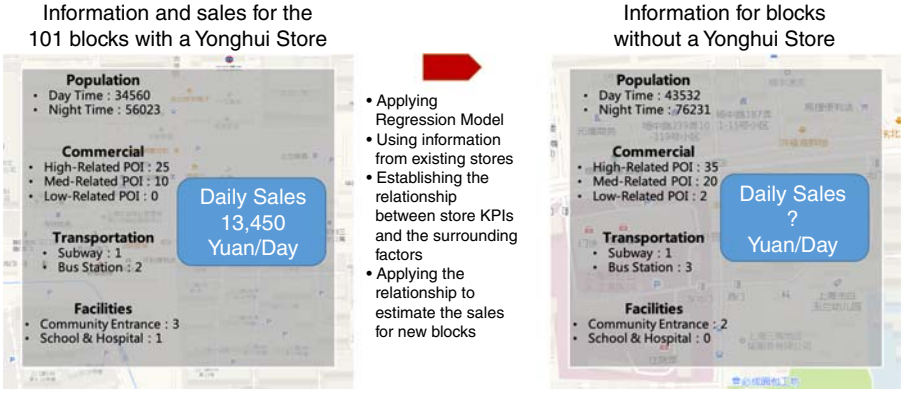

Figure 2.
Logic of preliminary sales forecasting

below as the first part of the objective to represent the sales of all Yonghui stores. The specific form of f could affect the complexity of solving the resulting optimization problem.

Given the data and the outputs of the BDS module, we model the site selection problem as a 0–1 integer programming (P), where the objective entails two parts: the profit earned by all the Yonghui stores and the total population served by Yonghui stores. γ_1 and γ_2 in the objective function are the weights of the two objectives, respectively, imposed by the decision maker. The relative importance of the two objectives depends on the development stage and the firm's strategic preference or various practical considerations. To determine the weight for each objective, we have gone through an iterative process with the development team. We try different combinations of weights, and for each combination we solve the optimization problem and output a set of blocks for site selection; the development team reviews the set of blocks to judge whether the optimization outcome makes sense; if not they propose a new combination of weights and we solve the optimization problem again. The process terminates until we find a pair of weights that lead to most reasonable optimization outcome:

$$\max_{x_j, y_i^r} \quad \gamma_1 \left(\sum_{j \in J} f\left(\left[\sum_{i \in I_j^r} Q_{p,i} y_i^r \right]^{r \in R}, Q_{c,j}, Q_{o,j}, Q_{t,j}, Q_{e,j} \right) - c_j \right) x_j + \gamma_2 \sum_{i \in I} Q_{p,i} \sum_{r \in R} y_i^r$$

$$\text{s.t.} \quad \sum_{j \in J} x_j \leqslant K, \tag{1}$$

$$(P) \quad y_i^r \leqslant \sum_{j \in J_i^r} x_j \quad \forall i \in I \quad r \in R, \tag{2}$$

$$\sum_{r \in R} y_i^r \leqslant 1 \quad \forall i \in I, \tag{3}$$

$$x_j \in \{0, 1\} \quad \forall j \in J, \tag{4}$$

$$y_i^r \in \{0, 1\} \quad \forall i \in I \quad \forall r \in R. \tag{5}$$

The model parameters are defined as follows:

- I: set of blocks to be served by Yonghui stores;
- J: set of candidate blocks for site selection;
- K: number of Yonghui stores to open;
- $Q_{p,i}$: population of block i;
- $Q_{c,j}$: competitors profile at candidate block j;
- $Q_{o,j}$: consumption profile at candidate block j;
- $Q_{t,j}$: transportation convenience at candidate block j;
- $Q_{e,j}$: commercial ecosystem at candidate block j;
- c_j: operational cost of running a store at candidate block j;
- d_{ij}: distance between served block i and candidate block j;

- $R = \{1, 2, \ldots, \bar{r}\}$: index set of circles around a block;
- $[\cdots]^{r \in R}$: a vector with each component having index r and $r \in R$;
- $I_j^r = \{i \in I \,|\, (r-1)d \leqslant d_{ij} \leqslant rd\}$: set of served blocks at rth circle around candidate block j; and
- $J_i^r = [j \in J \,|\, (r-1)d \leqslant d_{ij} \leqslant rd]$: set of candidate blocks at rth circle around served block i.

The decision variable $x_j = 1$ indicates that candidate block j is selected as the site of a Yonghui store; $y_i^r = 1$ means that block i is served by a Yonghui store at the rth circle around it. R represents the index set of circles around a block and the circles are indexed according to the radius in an increasing order. Constraint (3) indicates that block i is served by at most one Yonghui store at some circle r, while constraint (2) means that block i is not served by any Yonghui store at the rth circle if there is no store open at this circle.

Problem (P) is a huge 0–1 programming problem given a city such as Shanghai, which is quite challenging even for the most powerful commercial optimization solver, Gurobi. Recall that the BDS module can eliminate blocks with small populations, which will significantly reduce the number of decision variables in our problem. To further reduce the problem scale, we exclude some invalid blocks such as those in a park or river from the candidate set. In our modeling, we do not introduce the binary variable for every block i and candidate block j. Instead, all are aggregated into one variable y_i^r if every candidate block j is in the rth circle around block i. After implementing these techniques, the problem size is reduced significantly but is nevertheless excessively large to solve efficiently. Therefore, we further apply some decomposition techniques to the reduced problem and solve a few medium-sized problems. In particular, we divide the decision variables into several blocks of variables according to their geometric features. Then we cyclically solve a block variable problem while fixing the others until the improvement becomes marginal.

Finally, note that a block suggested by the ISE is a square of 250 × 250 meters, which is too large for site selection. Therefore, the exact location to open a store has to be further explored and identified by the PSF module.

3.3 Precision sales forecasting

Given the potential blocks suggested by the ISE module, a team of developers is assigned to search for a list of candidate spots for new store sites around the suggested blocks. This search process is guided primarily by a handbook compiled by the retail property development committee in which they classify criteria into a "must" list and a "want" list that include, for example, high traffic volume, maximum street frontage, accessibility and proximity to other draws such as restaurants and grocery stores. Most of these criteria can be quantified, but how can they be integrated into a scientific and systematic store site selection process?

Sales forecasting plays a critical role in the success of store site selection, as such predictions are the main reference for the committee that makes the final accept/reject decision. Now the key questions reduce to the following: What are the key site selection criteria and how are they related to the sales forecasting? When the project began, the committee had developed a basic scoring and ranking method, which gauged new retail stores' sales performance into several levels. However, this method suffered from poor predictive accuracy, so much of the site selection decision was still made based on the committee's instincts and past experience. Our aim here is to revolutionize the scoring and ranking methodology with cutting-edge artificial intelligence methods to improve predictive power.

In fact, the BDS module also includes a sales forecasting model, but it differs from the PSF in terms of concentration. The BDS module focuses primarily on evaluating the potential market size of given business districts from a macro perspective, while the PSF module focuses on predicting the sales of a specific store in a more detailed and precise way. Furthermore, since our aim is to derive an explicit relationship between sales and environmental factors (i.e. a function f) that can be used in the ISE module, we are restricted to using parametric regression models in the BDS module; thus, the predictive power could be further improved. In the PSF module, by incorporating extensive store-specific data and using advanced machine learning algorithms, we can forecast sales performance with a higher level of accuracy. The details are described below.

The sales performance of a retail convenience store is highly related to both the catchment area characteristics (e.g. population, competition, transportation convenience) and store-specific characteristics (e.g. store size, visibility, accessibility). In most cases, the store characteristics are more pivotal than the catchment area characteristics. Some characteristics are difficult to quantify, such as traffic volume in front of a store, visibility and accessibility. Therefore, we introduce a scoring scheme to subjectively quantify these micro characteristics of each store. Overall, the input data for the PSF module are comprehensive and the details of these data are shown in Table II.

For this panel data set, we adopt artificial intelligence models to determine the relationship, as shown in Equation (6), between store sales and a range of characteristics, e.g., store characteristics, catchment area characteristics, seasonality and trend of sales. Once a model is established, it can then be used to forecast the sales of a proposed store by substituting values of the independent variables:

$$\text{Sales} \sim \text{Store characteristics} + \text{Catchment area characteristics} + \text{Sales seasonality} + \text{Sales trend} + \ldots \quad (6)$$

Based on the preliminary data analysis, we find that the five measures of store dimensions have strong predictive power for store sales, and the Pearson correlations between sales and each measure are shown in Figure 3. It consists of five scatter plots and each has a fitted regression line. The horizontal axis represents 1–5 measurement scale, and the vertical axis represents monthly sales. The preliminary data analysis, illustrated in Figure 3, shows that these five subjective measurements are indeed strongly related

Category	Attribute	Description
Catchment area	Population	Daytime and nighttime population within 200\300\500\1,000 m of the store
	Transportation	Number of bus stations and subway stations within 200\300\500\1,000 m of the store
	Convenience	
	Commercial Ecosystem	Categorize POI into three different types and count the number for each type
	Open Date	Open date of the store
Store	Area	Area of the store
	Size of Facade	Size of the store's facade
	Traffic Volume	Measurement (1–5)[a] of the traffic volume in front of the store
	Accessibility	Measurement (1–5) of the accessibility to the store
Scoring	Visibility	Measurement (1–5) of the visibility of the store
	Competition	Measurement (1–5) of competition dynamics near the store
	Business dynamics	Measurement (1–5) of business dynamics near the store

Note: [a]Measurement scale 1–5, 1 is very negative to sales and 5 is very positive to sales

Table II. Input data for the PSF module

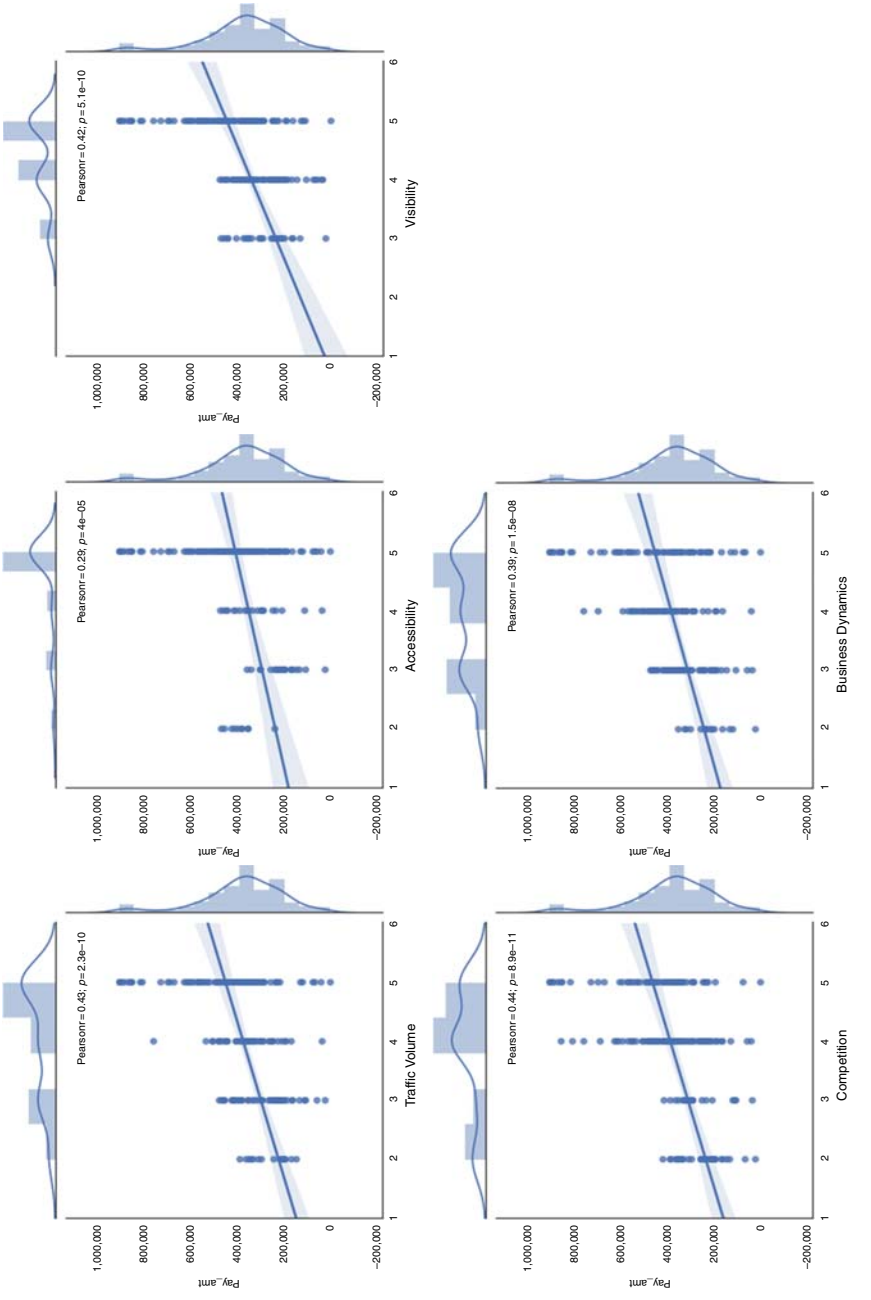

Figure 3.
Relationship between sales and the five measurements

to the sales performance. Therefore, incorporating such subjective micro characteristics is necessary in the PSF module.

In addition, the catchment area's characteristics for different distance ranges are highly correlated with each other, as illustrated by the correlation matrix in Figure 4. This matrix represents the pairwise correlation of the catchment area characteristics. A dark block indicates that the correlation coefficient is close to +1, whereas a light block indicates that it is close to 0. The matrix shows that all the characteristics of the catchment area are positively correlated with each other. Multicollinearity is a phenomenon in which the coefficient estimates of the multiple regression may change erratically in response to small changes in the model or the data. From Figure 4, it is evident that some sets of variables are strongly correlated to each other, and can be linearly predicted from each other, such as the set {DISTlow, r200low, r300low and r500low}, where DISTlow denotes the minimum distance between a store and a low-frequency-type facility and rx00low denotes the number of low-frequency-type facilities within × hundred meters of a store. Therefore, it is necessary to introduce the principal component analysis (PCA) to minimize the effect of multicollinearity.

After the step of feature engineering, we considered many different approaches, such as quantile regression forests, gradient boosting decision tree and residual neural network. We ultimately decided to employ quantile regression forests due to the best predictive performance. We use the 14 most recently opened stores as the test data, and the mean absolute percentage error (APE) is 15.03 percent, which is considerably more accurate than the original scoring method, and the detailed performance data are reported in Table III.

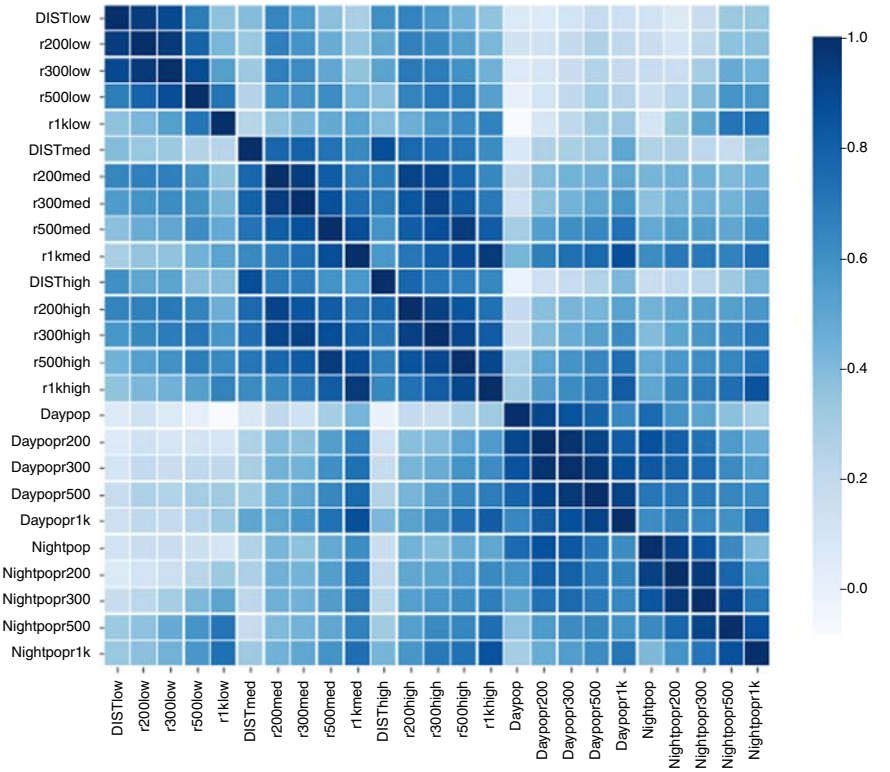

Figure 4.
Correlations matrix of catchment area characteristics

shop_id	actual_sales	predict_sales	APE (%)
9DAP	251,875	245,817	2.4
9DAZ	448,136	378,755	15.5
9DBC	280,922	286,619	2.0
9DBD	175,057	250,618	43.2
9DBE	259,036	305,302	17.9
9DBS	275,714	241,131	12.5
9DBW	237,398	249,989	5.3
9DBX	418,593	405,742	3.1
9DAD	325,934	250,626	23.1
9DBZ	300,390	250,029	16.8
9DC1	231,725	240,319	3.7
9DC2	252,247	240,481	4.7
9DC6	475,695	348,201	26.8
9DCD	171,802	198,796	15.7
9DCE	486,483	238,282	51.0
9DCH	241,366	229,157	5.1
9DCI	465,806	418,068	10.2
9DCJ	238,018	228,313	4.1
9DCT	337,497	238,164	29.4
9DCZ	258,013	236,746	8.2

Table III.
Sales prediction performance

4. Validation and discussion

When selecting a new retail site, profitability is the key performance index that must be evaluated by the development committee. Since opening a new store takes time and is costly, it is impossible for us to assess our model performance by opening a few new stores according to our suggestions and waiting for three months to monitor their revenue performance. Through collaboration with executives and developers in Yonghui, we constructed two feasible and repeatable evaluation approaches. One is hold-out validation, and the other is manual review.

Hold-out validation is a concept that we borrowed from machine learning and essentially divides the data set into a "training" and a "testing" set. In our case, the training set includes the initial 101 stores and is what the model is trained on; the testing set is used to check how well our model performs on unseen data. Fortunately, Yonghui opened 28 additional stores while the project was in progress, and these new stores can serve as testing data for model evaluation.

In total, we suggested over 500 new retail sites. For each of the 28 newly opened stores, we choose the nearest among these over 500 sites and compute the distance between the two. We say the suggested site hits the target if it is in the same block as one of the 28 stores. In fact, 4 of these 28 stores are called invalid stores because they are so close to the existing 101 stores in the training set that our model would not recommend any store nearby because doing so would violate the minimum distance constraint. For the remaining 24 stores, we find that 13 suggested stores exactly hit the target, and 83.3 percent of them have the nearest suggested store at most 250 meters away. The detailed performance is also shown in Table IV, where the distance in terms of number of blocks is illustrated in Figure 5. In the figure, the dark grey blocks are one block away from the central store, and the light grey blocks are two blocks away. In other words, at least 80 percent of our recommendations would have been acceptable to experienced developers and executives.

In principle, hold-out validation is easy and computationally efficient, but it is not significantly convincing since our data set is not large enough. Therefore, we developed a second method of validation, called manual review. A team of experienced developers was involved in the manual review process and asked to scrutinize our recommendations

one by one and indicate whether a suggested block was suitable. After a tedious and labor-intensive review process, we compiled the validation result with respect to Shanghai's ring road[1], namely for locations within and outside of it, and by district. Overall, the KPI, which is the percentage of "Hit" and "Nearby" as defined in footnotes a and b, is 82.2 percent, i.e., 82.2 percent of our recommendations were accepted by a team of experienced developers. From the validation result with respect to the ring road, as shown in Table V, we conclude that our model performs slightly better in the suburbs than in the central part of Shanghai. After brainstorming and in-depth qualitative analysis with Yonghui's executives and developers, we developed several potential explanations:

- in the central part of Shanghai, there are many areas with old houses, which do not meet the minimum operational requirement of a site, although all other criteria are satisfied, suchas population, traffic and business dynamics;

Scenario	Count	Cumulative percentage
Same block	13	54.1
1 block away	7	83.3
2 block away	4	100
⩾ 3 block away	0	100
Invalid	4	/

Table IV.
Hold-out validation result

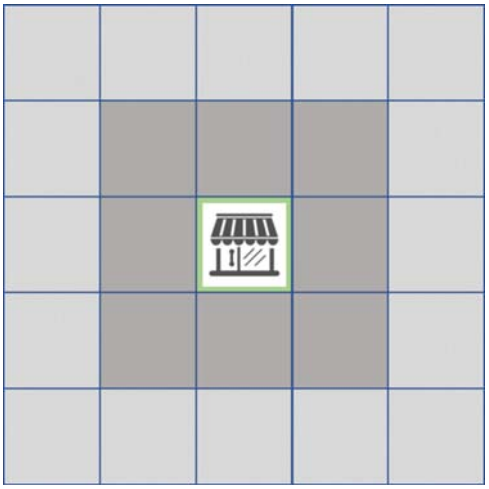

Figure 5.
Distance in terms of number of blocks

Ring	Hit[a]	Nearby[b]	Miss[c]	Invalid Case	% (Hit + Nearby) W/O Invalid
Within outer ring	262	5	65	11	80.4
Out of outer ring	144	4	25	8	85.5
Grand total	406	9	90	19	82.2

Notes: [a]We refer to the case that a suggested block is accepted by the reviewers as "Hit." [b]We refer to the case that a suggested block is declined but there is at least one block nearby accepted by the reviewers as "Nearby." [c]We refer to the case that neither the suggested block nor a block nearby is accepted by the reviewers as "Miss"

Table V.
Validation result from manual review (by ring road)

- business dynamics in the central part of Shanghai are substantially more complex, and in some cases, we cannot capture all the key criteria for site selection through our existing data source; and
- the scattered development pattern of the suburban area of Shanghai makes site selection somewhat easier than in the central part of Shanghai.

The validation result by district, as shown in Table VI, strengthens our prior conclusion that our model performs better in developing areas (e.g. Jiading, Minhang) than in developed areas (e.g. Huangpu, Hongkou and Yangpu).

After the detailed validation by the team of experienced retail property developers, Yonghui acknowledged that our intelligent approach would facilitate the process of site selection and improve selection performance, especially in new cities with very few experienced developers. The key feature of our approach is its ability to quantify the effect of important environmental factors on sales while adopting a network perspective in site optimization."

5. Conclusion

Some industry practitioners believe that 70 percent of a store's sales depend on the selected location. Traditional retail site selection relies primarily on manual search efforts by the development team according to their expertise in various regions' traffic volume, accessibility, competition and other characteristics. This might perform well when locating several stores within a city. However, when selecting hundreds of retail stores in an effort to cover most of a city's population, the human brain's scale limitations can be a constraint. Furthermore, there is no way to evaluate the performance of a selected site since we cannot obtain a conceptual upper bound by taking all potential regions into account. Potential regions may differ in some key dimensions. For instance, one region has high traffic flow but unfortunately also high competition with many existing stores, while another region has moderate traffic flow but little competition. Which region is more attractive as a potential store site? Without a model that quantifies each dimension's effect on sales, developers have to make a decision based on their subjective judgment, and no one knows whether the selected site will perform better than the alternative.

This paper describes how Yonghui partnered with Cardinal Operations to use regression, optimization and machine learning techniques to address such common issues in site selection and suggest approximately 500 potential blocks of 250 × 250 square meters for Yonghui to locate Yonghui Life convenience stores. The result of this collaboration is the

Table VI.
Validation result from manual review (by district)

District	Hit	Nearby	Miss	Invalid case	% (Hit + Nearby) W/O Invalid Case
Jiading	27	0	2	0	93.1
Boshan	35	0	10	1	77.8
Xuhui	29	1	8	1	78.9
Putuo	23	1	5	0	82.8
Yangpu	19	0	8	2	70.4
Songjiang	30	0	5	0	85.7
Pudong	113	4	32	8	78.5
Hongkou	12	0	4	2	75.0
Changning	20	1	0	0	100.0
Minhang	53	0	2	1	96.4
Qingpu	12	0	4	1	75.0
Jing'an	23	2	3	3	89.3
Huangpu	10	0	7	0	58.8
Grand total	406	9	90	19	82.2

Intelligent Site Selection System that includes three modules. In the BDS module, by using the existing 101 Yonghui Life stores' sales data, we calibrate, using a generalized additive regression model, the key dimensions' effects on sales. Such relationships can be used to assess the potential sales of any potential region. The short-term sales should not be the unique criterion to consider, and population coverage is another criterion for long-run development. Then in the ISE module, we integrate these two criteria, together with some practical constraints, and use integer programming to select the top regions. Each region (or block) is of 250 × 250 square meters and in each region, there might be several candidate sites proposed by developers after visiting the suggested region. Which site should Yonghui ultimately select? We need a more precise sales forecasting model that also takes the potential stores' information and more detailed micro-environment data into consideration. This is executed in a machine learning model in the PSF module. The site with the highest forecasted sales will enter the final pool of sites for Yonghui Life stores.

The above hybrid approaches to site selection have been applied and proven effective. In an industry with many competitors adopting optimization and artificial intelligence models to improve their decision making, it is difficult to imagine that a firm will be able to survive while relying on traditional approaches. In other words, in an era with massive amounts of information collected and digitalized, we believe that firms can excel by exploiting recent developments in intelligent decision-making approaches.

Acknowledgments

D. Ge is supported by the Program for Innovative Research Team of Shanghai University of Finance and Economics and Shanghai Institute of International Finance and Economics. B. Jiang is supported in part by the National Natural Science Foundation of China (Grant Nos 11771269 and 11831002) and the Program for Innovative Research Team of Shanghai University of Finance and Economics. X. Wu is supported by the National Natural Science Foundation of China (Grant Nos 71622001 and 71571047) and Shuguang Project of Shanghai Municipal Education Commission and Shanghai Education Development Foundation (Grant No. 15SG07). The authors are listed in alphabetical order.

Note

1. Yonghui divided Shanghai into two parts with respect to the Outer Ring Road. The central part of Shanghai is located within the Outer Ring Road, while the suburbs of Shanghai are outside the Outer Ring Road.

References

Aggarwal, A., Anand, L., Bansal, M., Garg, N., Gupta, N., Gupta, S. and Jain, S. (2010), "A 3-approximation for facility location with uniform capacities", *Proceedings of the 14th Integer Programming and Combinatorial Optimization*, pp. 149-162.

Berman, O. and Krass, D. (2002), "The generalized maximal covering location problem", *Computers Operations Research*, Vol. 29 No. 6, pp. 563-581.

Breheny, M.J. (1988), "Practical methods of retail location analysis: a review", in Wrigley, N. (Ed.), *Store Choice, Store Location and Market Analysis*, Routledge, London, pp. 39-86.

Buckner, R.W. (1998), *Site Selection: New Advancements in Methods and Technology*, Chain Store Publishing Corp, New York, NY.

Karamshuk, D., Noulas, A., Scellato, S., Nicosia, V. and Mascolo, C. (2013), "Geo-spotting: mining online location-based services for optimal retail store placement", *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM*, pp. 793-801.

Ladle, J.K., Stiller, D. and Stiller, D. (2009), "Retail site selection: a new, innovative model for retail development", *Cornell Real Estate Review*, Vol. 7 No. 1, pp. 1-26.

Li, S. (2013), "A 1.488 approximation algorithm for the uncapacitated facility location problem", *Information and Computation*, Vol. 222, pp. 45-58.

Rogers, D.S. (2007), "Retail location analysis in practice", *Research Review*, Vol. 14 No. 2, pp. 73-78.

Rohani, A.M.B.M. and Chua, F.F. (2018), "Location analytics for optimal business retail site selection", in Gervasi, O., Murgante, B., Misra, S., Stankova, E., Torre, C.M., Rocha, A.M.A.C., Taniar, D., Apduhan, B.O., Tarantino, E. and Ryu, Y. (Eds), *Computational Science and Its Applications – ICCSA 2018, Lecture Notes in Computer Science*, Vol. 10960, Springer Nature, AG, pp. 392-405, available at: https://link.springer.com/chapter/10.1007/978-3-319-95162-1_27#citeas

Roig-Tierno, N., Baviera-Puig, A., Buitrago-Vera, J. and Mas-Verdu, F. (2013), "The retail site location decision process using GIS and the analytical hierarchy process", *Applied Geography*, Vol. 40, pp. 191-198.

Shmoys, D.B., Tardos, E. and Aardal, K. (1997), "Approximation algorithms for facility location problems", *Proceedings of the 29th Annual ACM Symposium on Theory of Computing*, pp. 265-274.

Sviridenko, M. (2002), "An improved approximation algorithm for the metric uncapacitated facility location problem", *Proceedings of the 9th International Conference on Integer Programming and Combinatorial Optimization*, pp. 240-257.

Wang, F., Chen, L. and Pan, W. (2016), "Where to place your next restaurant? Optimal restaurant placement via leveraging user-generated reviews", *CIKM '16 Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pp. 2371-2376.

Wu, Z., Wu, H. and Zhang, T. (1991), "Predict user in-world activity via integration of map query and mobility trace", UrbComp'15, Sydney, August 10, available at: www.cs.uic.edu/~urbcomp2013/urbcomp2015/papers/User-In-World-Activity-Prediction_Wu.pdf

Xu, M., Wang, T., Wu, Z., Zhou, J., Li, J. and Wu, H. (2016), "Demand driven store site selection via multiple spatial-temporal data", *SIGSPATIAL/GIS*, Vol. 40 No. 10, pp. 1-40.

Yilmaz, E., Elbasi, S. and Ferhatosmanoglu, H. (2017), "Predicting optimal facility location without customer locations", *KDD' 17 Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 2121-2130.

Zhang, J., Chen, B. and Ye, Y. (2005), "A multiexchange local search algorithm for the capacitated facility location problem", *Mathematics of Operations Research*, Vol. 30 No. 2, pp. 389-403.

Corresponding author
Xiaole Wu can be contacted at: wuxiaole@fudan.edu.cn