Proceedings of the 2008 Winter Simulation Conference
S. J. Mason, R. R. Hill, L. Mönch, O. Rose, T. Jefferson, J. W. Fowler eds.

SIMULATION BASED SALES FORECASTING ON RETAIL SMALL STORES

Hai Rong Lv
Xin Xin Bai
Wen Jun Yin
Jin Dong

IBM China Research Lab
Zuanshi Build, Zhongguancun Software park, Beijing 100084, CHINA,

ABSTRACT

As the competition becomes more and more intense, many retail small store chain operators are eager to know how to evaluate new store locations quantitatively to support a scientific business development decision, e.g. what will be the potential sales of new stores? How will a new store influence other existing stores? In this paper, we propose a novel framework that can predict the competitions between new store and competing stores. First, we examine the special characteristics of customers' demand on retail small stores and develop several models to estimate different customers' demand; and then model the relationship between customers and multiple stores, such as route model, user behavior model; finally, we compute the store sales via simulation. A real case on the micro-competition analysis in a China city for one of the famous retail small store chain operators shows that our method is more practical and accurate than other methods.

1 PROBLEM INTRODUCTION

Sales forecasting studies have increasing business application in opening new stores or renovating existent ones. Because of the large investment for store facilities, the intensive competition, and other factors, most retailers have tries various methods to evaluating store sales potential as accurately as possible. Some supermarkets used the integrated analog approach to forecasting the sales of new stores and analyzing variations in the performances of existing stores. Some leading retailers subdivided "trade area" into primary, secondary and fringe areas with their own staffs and professional consulting companies by empirical studies. As a result, they could measure each store's market share roughly and allocate resources to achieve niche marketing effectively. The above work mainly focused on stores like supermarket, however small stores have some specific characteristics. The demand for small stores varies at different occasions and there exists overlap among occasions too. Besides, the demand is non-rigid and with attenuation due to distance.

There has been a long history in the field of sales-forecasting research (William Applebaum 1965). The most popular theory is Reilly's law of retail gravitation, which states that two cities attract trade from an intermediate town in the vicinity of the breaking point approximately in direct proportion to the populations of the two cities and in inverse proportion to the squares of the distances from these two cities to the intermediate town (Reilly 1931). According to the Reilly's Law, professor Huff (1964) presented an alternative model, where the probability of a consumer traveling to a particular store *s* has positive relationship to the relative size of *s* and negative one to the relative travel time.

Gravity model is easy to use, but not coincident with the reality. Allen F. Jung (1959) stated that when the size of each city exceeds several hundred thousand people, perhaps it makes little difference to the consumer which city is patronized. Louis P. Bucklin (1967) took 15 weeks to interview 249 families and found out some data which showed the old stores with small size had a higher ratio of market share to square feet than the larger ones. In conclusion, drawing power is likely to be concave function of store size.

Although different models employed different variables for predicting sales, distance from areas of money generation to destinations like supermarket has been centered on. The followers improved gravity model by replacing distance with driving time or other variables (Brunner 1968, O. Gonzalez-Benito 2005, P. Greistorfer 2006, Michael Nwogugu 2006). Brunner (1968) obtained much information from the city of Toledo and suggested that about three-fourths of each center' shoppers resided within the15-minutes driving range.

978-1-4244-2708-6/08/$25.00 ©2008 IEEE

However, all exiting sales forecasting models have one or more of the following limitations:

- Low accuracy of market potential evaluation: market potential is relative with population characteristics that are determined by census and survey data. However, population with different characteristics has different consuming behaviors, which are not taken into consideration in most models.
- Over simplification of route: most models erroneously assume that the distance from each community to the store is constant and remains constant for each customer, for all trips to the store, and for all time, but most people do not work in their immediate communities and most people shop before going to work, during lunch, after work, or weekends.
- Over simplification of competition: competition is usually ignored or over simplified, which cause serious deviation of sales prediction from real value.
- Ignorance of randomness: both market potential and route have some random characteristics, which are not implemented properly in most models.

To deal with the above problems, we propose a new approach for sales forecasting based on optimization and simulation techniques. Our main contribution lies on four points:

- We take customer differences into consideration, including age, income, employment status, occasion (living, working, shopping, etc). Customers in different segments have different shopping habit, whose contribution to store sales are different.
- An optimization model to evaluate different customer segments' market potential (demand) is proposed to integrate multiple data sources to improve the accuracy.
- Random characteristics, such as route choice, store choice, demand variation, are modeled via simulation techniques, more detailed and accurate than previous approaches.
- All data and algorithms are integrated into one unified GIS platform, which automates the sales forecasting process and ensures the accuracy and objectivity of the results.

This paper is organized as follows. In section 2, we introduced the model structure and its details. A real case is presented in Section 3. Finally, section 4 presents the conclusion and future work.

2 MODEL INTRODUCTION

As shown in Figure 1, the framework of the proposed sales forecasting system is built upon GIS platform, which mainly consists of the following parts:

(1) A customer segmentation model to identify customers with similar consuming habits;

(2) An optimization model (Demand Evaluation) to evaluate the customers' demand according to multiple data source;

(3) A demand generator to determine who will consume and how much they will consume;

(4) A route model to evaluate the connectivity and convenience of the road;

(5) A behavior model to simulate how customers travel to the store and how they choose among competing stores;

(6) The analysis results including predicted sales of new stores and their influence on competing stores will be visualized in an intuition way via the GIS platform.

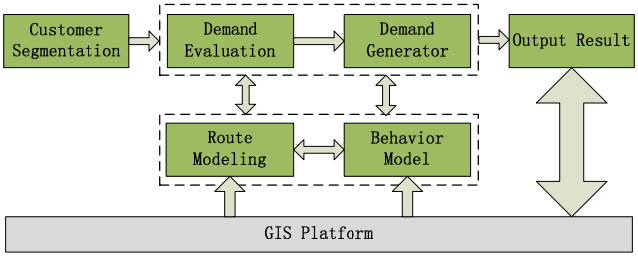

Figure 1: Sales forecasting framework

2.1 Customer Segmentation Model

Customer segmentation is the process of dividing customers into different groups on the basis of common attributes. In most application, customer segmentation is accomplished by defining numerical attributes which describe a customer's value based on economical and demographic considerations. It means that customers in the same segment usually have the same age group, same shopping habit, etc. Cluster algorithms are usually used in order to discover groups of customers with similar attributes (Shin H. 2004). Segmentation methods based on clustering require us to carefully select the attributes of customers. Since in our system, the purpose of customer segmentation is for sales prediction and customers' shopping behavior depend on where they stay, we take the geographical factors into consideration. For example, customers in a residential area and customers in an office building belong to two distinct segments, even if their demographic attributes are the same.

2.2 Demand Evaluation

It's very difficult to evaluate the demand D of customers, in other words, how much will a customer spend on a specific store each day? D is dependent to the probability of frequency and expenditure per time. The difficulties come from the several aspects. First, D is relative to the customer segment. Second, D is non-rigid and has attenuation with route/distance. Third, there exists overlap among customer segments, that means a customer may belongs to different customer segments when he is in different geographical occasions and his total expenditure may have overlap among different customer segments. Finally, demand of a customer segment is not a constant, it subjects to a certain distribution, which is assumed and proved to be normal distribution.

In order to clarify the customer demand for stores, however, we take geographical factors (occasion) into consideration, just as section 2.1 has pointed out. Figure 2 shows a simplified version of our solution, in which other attributes (e.g. age, income, etc.) are ignored and only the geographical element a customer belongs to is taken into consideration for customer segmentation. Sometimes, we have statistics on sales contribution of grids (physical area, e.g. 100*100 meters) around existing stores from on-site survey and we can link geographic elements to grids. Of course, the economic and demographic attributes of the geographic elements can also be provided by survey. For example, S_3 is the sales contribution of the facilities (Facility 3 and 4) in Grid3 to store. u means the utility from the facility to the store. However, the demand attenuation can be represented by a function $f(u, T(F))$, where $T(F)$ is the type of the geographical Facility F, e.g. commercial/residential. Followed are some situations:

Facility 1 is very close to the store, so $u = 0$ and $f(u, T(Facility_1)) = 1$. We can write the relationship as:

$$S_1 = A(Facility_1) \times D(T(Facility_1)) \quad (1)$$

Facility 2 is in Grid 2, with the utility u_2 to Store, so:

$$S_2 = A(Facility_2) \times D(T(Facility_2)) \times f(u, T(Facility_2)) \quad (2)$$

In Grid 3, there exist two facilities, so:

$$S_3 = f(u, T(Facility_3)) \times \sum_{i=3}^{4} A(Facility_i) \times D(T(Facility_i)) \quad (3)$$

However, sometimes we only know the total sale S of a store, so:

$$S = \sum_i A(Facility_i) D(T(Facility_i)) f(u, T(Facility_i)) \quad (4)$$

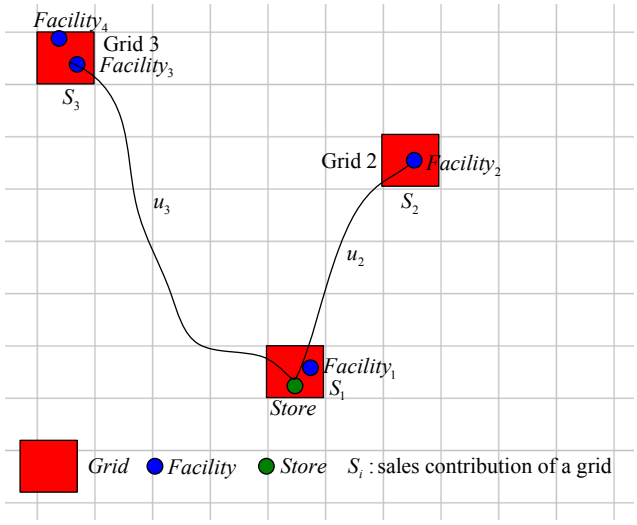

Figure 2: Sales contribution via on-site survey

We also have some other channels to acquire demand evaluations. Finally, we use an optimization model to integrate all these channels to work out the original unit demand for different customer segments by evaluating their means and variances.

2.3 Demand Generator

As stated in section 2.1, even for a same customer segment, customer demand is a random variable. According to the mean and variance of customer segment's demand, we can generate the probability of customers' wish for going to store and how much they will spend there. If a customer wants to go to a store, he will evaluate the stores' utilities and then decide if he insist or give up the trip by probability. Of course, he also needs to make a choice among the competing stores.

2.4 Route Modeling

Numerous analysis are being carried out to increase customer insight in the retail industry, e.g. 'who are target customers?', 'what are they buying?', etc. However, there is one question being neglected: How do the customers get to the store? It's really an important problem that the customer-store route can greatly influence customer's consuming behavior. Modeling the route is a complex task, which requires incorporating many factors including customer profile (customer segment, time preference, etc.), and environmental data (road profile like width, traffic, bus station, direction etc., and other info like trade area along roads, etc.).

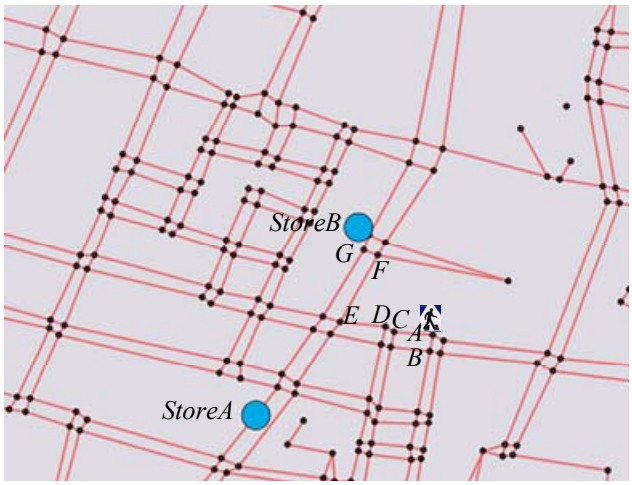

Figure 3: Route modeling

As shown in Figure 3, we use some basic route elements to represent the roads, such as sidewalk (AB), crossover (FG), road side (AC), etc. Each route has some basic attributes, e.g. location, direction, length, convenience for walking, bike, or bus, etc., and some other advanced attributes. Hence, we can measure the comprehensive route utility from a customer to a store according to the route elements and vehicles the customer selects.

2.5 Behavior Model

Customer behavior model describes how customers behave when they are doing business with the store network. According to customer behavior theory in marketing science (Hawkins, Del I. 1997), customer shows rich behavior patterns, and the term "customer behavior" also covers many aspects, such as habit, preference, selecting, tolerance, etc. In this paper we only consider three major aspects, including demand generation, demand attenuation and store selection, that will affect store sales prediction.

***Demand generation*:** The customers are distributed in certain regions of the map and staying in different geographical elements (such as residential areas, office buildings, supermarkets, etc.). How to generate demands has been discussed in section 2.2. However, it should be pointed out again that the demand generated here is only "original demand". In other words, if the store is just aside the customer, the "real consumption" in the store from the customer will be equal to the "original demand"; otherwise, the "real consumption" may be less than the "original demand" on average sense, for that the customer may give up the consumption in the store due to long distance.

***Demand attenuation*:** As stated above, the "real consumption" may be less than the "original demand", which is called this "demand attenuation". Demand attenuation characteristic is relative with customer segment, for example, customers in residential areas and those in official buildings have different attenuation characteristics. Demand attenuation characteristic can not be described by an analytical formula, but we can learn the attenuation curve $f(u,T(F))$ from samples according to the relationship of demand and store sales, which have been described in section 2.2.

***Store selection*:** A store selection model considering spatial information can be defined by Monte Carlo simulation (Aaker 1971; David 1972). The probability of chosing each store is computed as functions of the discriminant scores. As shown in Figure 4, if a customer decide to choose store A, he then has two routes for choice: route 1 and route 2. If he choose route 2, he will meet store B on the way, so he might change his idea and choose store B for consumption. Similar things occur in various scenarios. The function of the discriminant scores can be set by empirical study.

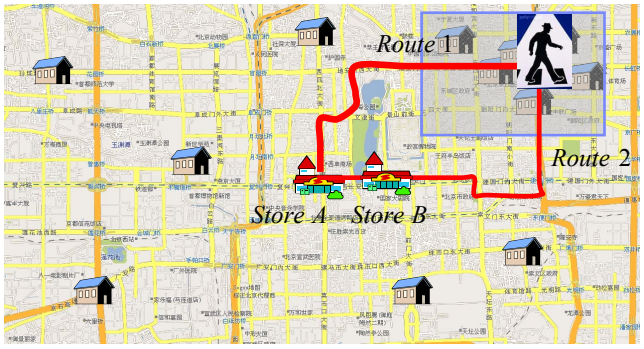

Figure 4: Store selection simulation

2.6 GIS Platform

GIS platform offers spatial information of the objects in the whole region such as streets, rivers, buildings, etc. All the data in the GIS platform are organized into geographic layers, which can be queried for more comprehensive analysis (Ming Xie 2007). In the proposed sales prediction system, GIS platform acts an important role as the data provider and result visualizer. The route model and behavior model are determined through the related layers of the GIS data. And the prediction results can be more intuitively presented into the map through GIS viewer.

3 CASE STUDY

In this section, we report a case study of an application of this integrated framework. The case study is drawn from a routine business research project that we recently conducted.

The client in the case study was a famous specialty small store chain operator in the world with two important needs. First, it desired a reliable method of forecasting the sales of new stores and, second, it wanted a better under-

standing of the reasons for variations in the performances of existing stores while planning a new store around. We select a tie-3 China city for the case study. The chain operated dozens of small stores and thus had a sufficient sample data for a model-building approach to forecasting sales. Moreover, its main competitors are also taken into consideration.

According to session 2.3, the demand of each customer segment is a random variable. There is considerable variability of the predicted sales across the replications and Figure 5 shows a histogram of the sales of store 1 for the 400 runs. The average sales per replication is 2.690*10^4 RMB/day.

Table 1 shows the predicted results of five stores. The column "Predicted Sale" is the average sales for the 400 runs and the "actual sale" is the real sale value (new store or existing store). We can see that the error between them is very small, which prove the effectiveness of the proposed method in this paper. It must be pointed out that all the predicted sales are computed based on only outer geographical data.

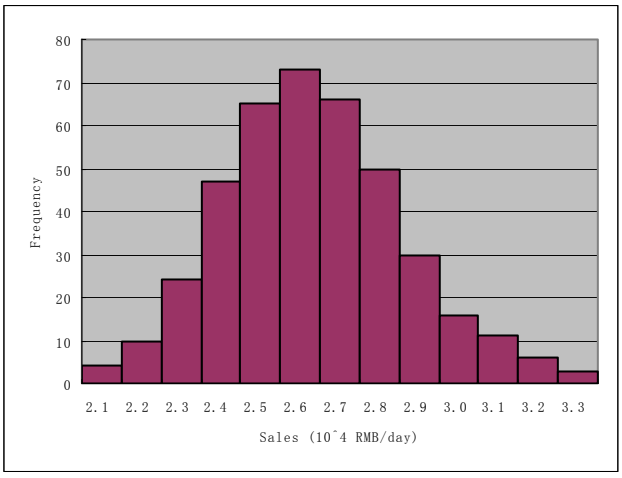

Figure 5: Histogram of sales for store 1 (the x-axis values are the center values in each interval)

Table 1: Predicted Sales

Store ID	Predicted Sale (10^4 RMB/day)	Actual Sale (10^4 RMB/day)	Error (%)
1	2.690	2.672	0.7
2	3.072	2.977	3.1
3	6.084	6.023	1.0
4	2.289	2.421	5.5
5	2.131	2.143	0.5

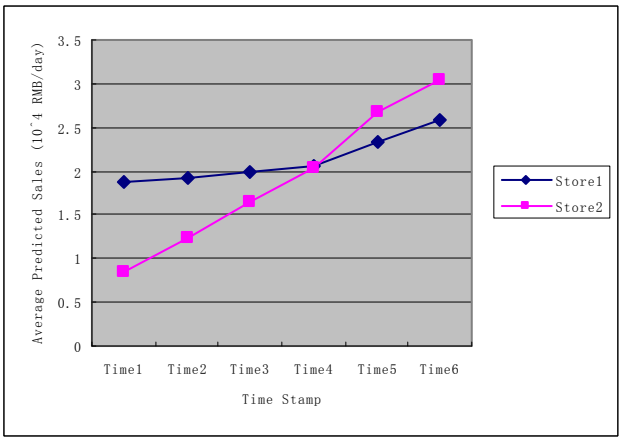

Figure 6: Sales trends of different stores

Figure 6 shows the sales of two stores at different time stamps. We can see that, the sales keep changing as the time going on, because that the outer environment and population keep changing. Store 1 had relative vantage to store 2 at time stamp 1, but more and more facilities were built around store 2 and the gap between the two stores are becoming smaller. Finally, at time stamp 5, the situation is converse.

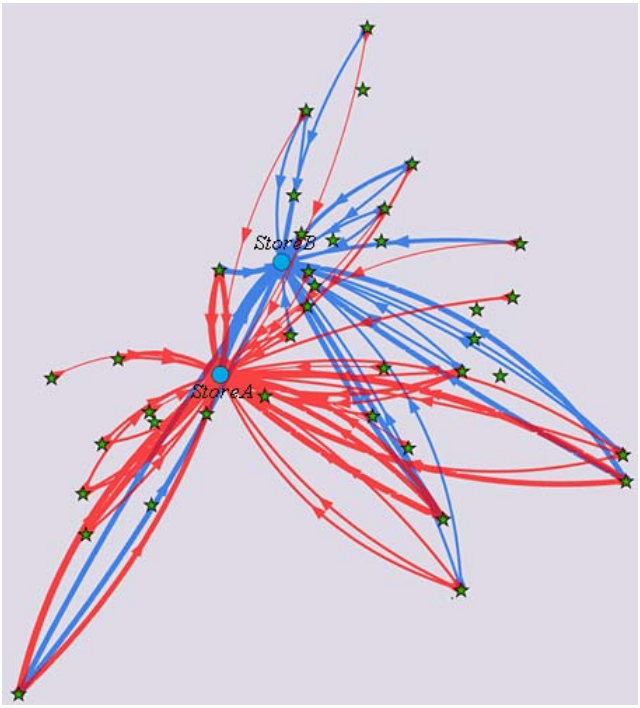

Figure 7: Micro-competition scenario

Figure 7 demonstrates the microanalysis result of a competitive scenario. Customers in different geographic elements have different selection probabilities for the two

stores, where the thickness degree of the lines represent the sales contribution to the corresponding stores.

4 CONCLUSION

This article demonstrates how optimization and simulation techniques can be adopted in retail store sales prediction and micro-competition analysis. Based on GIS platform, we develop a more elaborate analysis framework than traditional gravity model. Simulation technique provides the means of directly evaluating the impact of these factors. Besides the sales prediction of new stores, this framework can also provide an explicit estimate of the impact of new stores on existing competing stores, which is of more value for new site evaluation work of business development people. Real case has proved the effectiveness of the framework.

In our framework, we deal with the stores in the same way and ignore their differences. However, multiple factors, including visibility, store size, service level, environment, etc. can influence customers' behavior greatly. Therefore, how to build a quantitative model to evaluate the stores' influence on the customers will be an interesting work.

In current framework, we perform the sales prediction on given candidate sites of new stores. In real case, how to select those candidate sites is also a complex analysis work. Therefore, another interesting extension of the framework would be incorporating sales prediction into automated site location.

REFERENCES

Aaker David A., J. Morgan Jones, 1971. Modeling store choice behavior. *Journal of Marketing Research*, 8: 38-42

Allen F. Jung, 1959. Is Reilly's law of retail gravitation always true? *Journal of Marketing*, 24: 62-63.

Cheng Binpo, 2007. *McDonalds and KFC: the key success secret*. China Economy Publishing House.

Cheng Guang, 2004. *KFC strategy*. China Enterprise Management Publishing House.

David B. MacKay, 1972. A micro-analytic approach to store location analysis. *Journal of Marketing Research*, 9: 134-140

David L. Huff, 1964. Defining and estimating a trading area. *Journal of Marketing*, 28: 34-38.

Hawkins, Del I.; Best, Roger J.; Coney, Kenneth A., 1997. Comsumer Behavior: Building Marketing Strategy (7th edition), McGraw Hill

Michael Nwogugu, 2006. Site selection in the US retailing industry. *Applied Mathematics and Computation*, 182: 1725-1734

Ming Xie, Wei Wang, Wenjun Yin, Jin Dong, 2007. iFAO-SIMO: a spatial-simulation based facility network optimization framework. *Proceedings of the 2007 Winter Simulation Conference (WSC 2007)*

O. Gonzalez-Benito, J. Gonzalez-Benito, 2005. The role of geo-demographic segmentation in retail location strategy. *International Journal of Market Research*, 47: 295-305

P. Greistorfer, C. Rego, 2006. A simple filter-and-fan approach to the facility location problem. *Computers & Operations Research*, 33: 2590-2597

Shin H., Sohn S., 2004. Segmentation of stock trading customers according to potential value. *Expert Systems with Applications*, 27(1): 27–33

William Applebaum, 1965. Can store location research be a science? *Economic Geography*, 41: 234-237.

William J. Reilly, 1931. The law of retail gravitation. New York: published by the author

AUTHOR BIOGRAPHIES

HAIRONG LV received the B.E. in 2002 and Ph.D. in 2007, both in Department of Automation, Tsinghua University, China. Currently, he is working as a researcher in IBM China Research Laboratory. His research interests include pattern recognition, machine learning and business intelligence. His email address is <lvhr@cn.ibm.com>.

XINXIN BAI received her Master's in Tsinghua University from China in 2005 and then joined IBM China Research Laboratory. Currently, her research interests include Business Analytics and Optimization, Data Mining and Business Intelligence, Retail solutions. Her email address is <baixx@cn.ibm.com>.

WENJUN YIN received his Ph.D. in Tsinghua University from China in 2004 and then joined IBM China Research Laboratory. Currently, his research interests include Supply Chain Management and Logistics, Business Analytics and Optimization, Data Mining and Business Intelligence. His email address is <yinwenj@cn.ibm.com>.

JIN DONG is the Manager of Supply Chain Management and Logistics Research in IBM China Research Laboratory. He received his Ph.D. in Tsinghua University from China in 2001. Before joined IBM, he was the Research Assistant Professor in Industrial Engineering Department of Arizona State University in USA. His email address is <dongjin@cn.ibm.com>.