GROCERY RETAIL STORE REVENUE PREDICTION FOR STORE LOCATION EVALUATION USING SPATIAL INTERACTION MODELS

by

Müge Sandıkçıoğlu

A Thesis Submitted to the
Graduate School of Engineering
in Partial Fulfillment of the Requirements for
the Degree of

Master of Science
in
Industrial Engineering

Koç University
August 2008

Koç University

Graduate School of Sciences and Engineering

This is to certify that I have examined this copy of a master's thesis by

Müge Sandıkçıoğlu

and have found that it is complete and satisfactory in all respects,

and that any and all revisions required by the final

examining committee have been made.

Committee Members:

Prof. Serpil Sayın

Associate Prof. Özden Gür Ali

Prof. Selçuk Karabatı

Associate Prof. Serdar Sayman

Assistant Prof. Sibel Salman

Date:

ÖZET

MAĞAZA YER SEÇİMİ İÇİN MEKANSAL ETKİLEŞİM MODELLERİ İLE SÜPERMARKET CİRO TAHMİNİ

Çalışmamızın amacı mekansal etkilesim modellerini kullanarak potansiyel mağazaların ciro tahmini yoluyla supermarket yer seçimi kararlarına destek olmaktır. Bu çalışmada amacımız kaliteli musteri harcama payi ve magaza ciro tahminleri veren esnek bir mekansal etkilesim modeli oluşturmaktır. Model oluşturmada, keşif analizinde elde ettiğimiz sonuçlar kullanılmaktadır. Keşif analizinde uzaklık, rekabet, mağaza satış alanı ve mağaza formatı gibi önemli olduğunu düşündüğümüz değişkenlerin etkileri makro ve müşteri analizi ile araştırılmaktadır. Keşif analizine göre, büyük mağazalar uzak bölgelerden daha fazla müşteri çekmektedir ve uzaklık genel olarak müşteri mağaza seçimini etkileyen önemli bir faktördür. Buna paralel olarak oluşturduğumuz modelde uzaklık bir değişken olarak kullanılmış ve uzaklık etki parametresi mağaza satış alanının bir fonksiyonu olarak belirlenmiştir. Bu modeli degisken uzaklik etki parametresi modeli olarak adlandirdik. Sonuçlarımıza göre desigken uzaklık etki parametresi modeli uzaklik ve magaza buyuklugu arasindaki iliskiyi daha iyi tanimladigi ve bunun sonucunda elde edilen musteri harcama payi tahminlerinin daha kesin oldugu gorulmustur. Bu model mağaza çekiciliği parametreleri eklendiginde ve belli bir aralik icinde kontrol edilen satis tahmin kisiti ile beraber optimizasyon yoluyla cozuldugunde Huff (1963), MCI (1974) ve Competing destinations (1988) modellerinden gercek veriler uzerinde esneklik, tahmin ve genelleme kalitesi kriterlerinde daha iyi sonuc vermistir. Bu tez çalışmasında aynı zamanda posta kodu bazında müşteri harcamalarına ve müşterilerin değişik perakende zincirlerindeki harcamalarına ihtiyaç duymayan bir mağaza ciro tahmini metodu önerilmiştir.

ABSTRACT

GROCERY RETAIL STORE REVENUE PREDICTION FOR STORE LOCATION EVALUATION USING SPATIAL INTERACTION MODELS

The goal of this thesis study is to predict revenue of a potential grocery store to support the store location decision using spatial interaction models. In this study we focus on creating a flexible spatial interaction model which provides accurate customer share prediction and store sales prediction results. Basis of our spatial interaction model is insights obtained in the exploratory analysis which we examine factors such as distance, competition, store format through macro and customer level analysis. According to our findings, bigger stores attract more customers from distant areas and distance is a significant variable affecting customer store choice behavior. Based on this finding, we add distance variable to our model and we represent the distance decay value as a function of store size. We name this model as variable distance decay model. We observe that this representation of the distance decay value provides improved customer share prediction results. In this study we also add store attractiveness values to the variable distance decay model and estimate it with constrained optimization including a sales prediction constraint which can be controlled with upperbound and lowerbound values. The new model combined with the estimation procedure becomes superior to Huff (1963), MCI (1974) and Competing destinations (1988) models in terms of flexibility, accuracy and robustness on real customer data. We also propose a method of store revenue prediction using spatial interaction models and customer loyalty card data when customer panel data and zip code level customer spending are not available.

ACKNOWLEDGEMENTS

I would like to thank to my supervisors Serpil Sayın and Özden Gür Ali who shared their experience and expertise with me and mentored me during this thesis study. I would also like to thank them for encouraging me to pursue this degree. I would like to thank Selçuk Karabatı, Serdar Sayman and Sibel Salman who kindly accepted to be a part of this thesis committee. I would like to thank to Hüseyin Aktuğ, Serhat Uğur and Varol Hıdır from Migros for sharing their knowledge and time. I would like to acknowledge Migros for their financial support and willingness to share their resources for this thesis study. I would also like to acknowledge TÜBİTAK, for their generous financial support.

I would like to thank Başak and Figen for their companionship and Selim for his friendship.

I would like to thank my parents Osman and Meltem Sandıkçıoğlu for supporting me in every aspect of life and encouraging me to pursue my dreams. I would also like to thank Hakan for his love and continuing support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

NOMENCLATURE

j	store
i	customer (may be individual or a small residential area)
k	variable
SS_j	sales of store *j*
p_{ij}	real share of spending of customer *i* in store *j*
$\hat{p}_{ij}$	predicted share of spending of customer *i* in store *j*
s_i	grocery budget of customer *j*
w_j	store size of store *j*
d_{ij}	distance between customer *i* and store *j*
β_D	distance decay parameter
X_{kij}	the kth variable describing store *j* for customer *i*
δ_k	coefficient or sensitivity parameter of *k*th variable
Z_{kij}	is the zeta-transformed *k*th variable
St_k	standard deviation of the *k*th variable.
CL	choice set of customers selected in the analysis
ψ	sensitivity parameter of likelihood function
O	number of stores
M	number of customers
m_j	market share of store *j*
b_i	buying power of customer *i*
A_j	attractiveness of store *j*
l	lowerbound value

u	upperbound value
c	a constant in the distance decay function
λ	store size sensitivity parameter in distance decay function
h	lifestyle group
H_i	predicted number of people living in residential area i
Ω	immediate area radius
t	year of the store opening
Q_t	yearly buying power or spending increase
N_j	the new demand increase parameter and
G_j	the area generalization parameter.
$\mathrm{V}_{\mathrm{t+1}}$, $\mathrm{V}_{\mathrm{t-1}}$	are adjusted vicinity sales after and before the store opening.
R	vicinity area surrounding the new store
K_1 and K_2	independent variables for prediction N
L_{t}	level value for estimation of Q_t
T_{t}	trend value for estimation of Q_t
η_1, η_2	values between 0 and 1 in level and trend estimation
ϑ_1, ϑ_2	coefficients of independent variables in estimation of N
α	a value between 0 and 1
U	the selected area

Chapter 1

INTRODUCTION

1.1. Background

The site selection problem can be defined as the process of selecting the optimal location in terms predetermined criteria considering the relevant constraints which depend on the type of the facility to be located. It is a very common and a versatile problem which almost all businesses face. The nature of the site selection problem changes depending on the operational use of the facility to be located. Site selection is especially different for revenue generating facilities such as retail and non-generating facilities like manufacturing and logistics. Thus, specialized approaches exist for each occasion. In this thesis study, we are interested in the revenue generation aspect, specifically for grocery stores.

A non-revenue generating facility can be defined as the facility which is not used directly by the customers for exchange of services. Logistics facilities, warehouses, office buildings and production facilities can be examples of non-revenue generating facilities. Site selection decision is infrequent for non-revenue generating facilities and often considered geographical area for candidate locations is very wide even international.

Different from the discussed ones, some facilities are directly used by customers for exchange of services such as bank branches, restaurants and retail stores. These kinds of facilities are the main sales channel for the companies using them and they directly generate revenue. Location decision is fairly frequent for revenue generating facilities especially for chains. For example, Turkish grocery retailer Migros opened 140 new stores in Turkey during 2008. The frequency of the location decision increases the need for standardized procedures and methods to deal with the site selection problem for revenue generating facilities.

In this thesis study, we focus on site selection in grocery retailing. Site selection in grocery retail facilities is customer driven. As a result, one should assume a structure for customer shopping behavior in order to predict the success of the candidate location. Researchers generally use spatial interaction models for this purpose. Spatial interaction models typically generate customers' share of spending in a facility by proportioning that facility's utility to sum of utilities of all facilities. Most commonly used spatial interaction models are Huff (1963), MCI (1974) and Competing destinations (1988).

1.2 Motivation and Objective

Revenue of a candidate store is an obvious criterion for grocery store location evaluation. We focus on spatial interaction models for grocery store revenue estimation in this thesis study because of their practicality. All spatial interaction models can be used for store revenue estimation by generating customer shares of spending. However, their performances vary due to differences in estimation procedures and model specifications.

Our main goal in this study is to create a flexible spatial interaction model which can predict revenue of a new grocery retail store and identify the potential customers with greater accuracy. Identifying the potential customers is possible by predicting customer shares in the candidate store. It is achievable to obtain the exact revenue of a candidate store by predicting customer shares of spending perfectly. However, all customers are different and by identifying general patterns in customer grocery shopping behavior it is not feasible to obtain perfect customer share predictions. Generally, there is a trade off between candidate store revenue estimation and customer share of spending prediction. This is because, the first problem considers the amount of customer spending and the second doesn't. Thus, for the first problem customers who spend more are more emphasized where in the second all customers are equally important. Accurate store sales predictions can improve the profitability of the company by reducing the risk of opening an unsuccessful store. Identifying the customers of a potential store with greater accuracy may be valuable for product portfolio and pricing decisions. Spatial interaction models do not necessarily perform in an acceptable range in both objectives. A flexible spatial interaction model can focus both on general store sales estimation and individual customer share of spending prediction performances. Current models focus only on the general sales prediction performance like Drezner and Drezner (2002) while others focus on the individual customer share of spending prediction performance like Huff (1963), MCI (1974) and Competing destinations (1988).

1.3 Modeling perspective

Our main objective in this thesis study is to create an accurate and flexible spatial interaction model for grocery store sales prediction. In order to pursue this goal we try to understand the customer shopping behavior and identify the crucial factors. We use the loyalty card data provided by Migros and panel data is not available to us. For this reason, we have to eliminate factors such as pricing, brand and reputation. We use the identified factors and obtained insights in spatial interaction model construction. We choose to use constrained optimization instead of log-transform or maximum likelihood estimation as a tool to calibrate the new spatial interaction model. Constrained optimization provides us the flexibility advantage. With constrained optimization we can minimize the individual customer share prediction error while controlling the general sales prediction performance.

In order to capture store specific effects we also include a store attractiveness value as a parameter to our model. In addition we explore the interaction between the distance decay value and the store size and use or findings in the new spatial interaction model.

Zip code level grocery spending data is not available to us for this study. However, customer loyalty card data including customer grocery spending from a grocery retail chain and customer address data coded in a geographical information system are available. We propose a new method to obtain the revenue of a grocery retail store using the spatial interaction models and loyalty card data when panel data and zip code level spending data is not available. We provide a template for a decision support system which standardizes the procedure and increases the efficiency of usage of the proposed method.

Our research contributes to the grocery store revenue prediction and site selection literature in terms of spatial interaction model specification by providing insights about the distance decay value and store attractiveness. Our research is different from the models in the literature because it minimizes the customer share prediction error while improving the store sales prediction performance. We obtain this result by introducing an explicit sales prediction constraint. Also this study provides a method to obtain the store sales when panel or survey data is not available which is not covered in grocery store sales prediction literature.

1.4 Outline of the thesis

Outline of this thesis is explained briefly in this section.

Chapter 2 provides a literature survey about site evaluation methods used for store revenue prediction and spatial decision support systems. Our main focus in the literature survey is spatial interaction models.

Exploratory analysis is covered in Chapter 3. The purpose of the exploratory analysis is mainly to identify the important factors influencing customer store choice behavior. Identifying these factors is important as grocery retail site selection is customer driven. These factors and other insights about the grocery retailing are then used in spatial interaction model construction in Chapter 4. Effects of a new store on existing stores are explored in this section in two levels: customer and the store level. In customer level analysis, we use several data mining models to predict the shares of spending of the customers after a new store opening and compare them with Huff (1963) model. Also in customer level analysis, we try to understand the relationship between customer spending and distance. In macro level analysis, we try to understand the effects of a store opening on existing store sales to gather insights about competition as well as sales patterns of grocery retailing. Our main conclusions in this section are that distance, store size, store format and competition are important factors in customer store choice behavior. Moreover, bigger stores are less affected from distance. We also conclude that spatial interaction models are robust and provide superior results to data mining models.

In Chapter 4 we propose a new spatial interaction model using the insights obtained from Chapter 3. We also compare its performance with models from the literature such as Huff (1963), MCI (1974), Competing destinations (1988) on a numerical example. In Chapter 4, we provide the motivations for the new model and details of the estimation procedure. There are two main motivations for the new model. First, there is a relationship between the store size and the distance decay value. Distance decay value is the power of distance variable and it determines the effect of distance on customer store spending. Second, store attractiveness values as used by Drezner and Drezner (2002) can be used for store sales prediction. We test the performances of the models in terms of mean squared error of individual customer shares and store sales prediction on the loyalty card data provided by Migros. We observe that the new model performs significantly better than models in the literature in both individual customer share and general store sales prediction.

Chapter 5 describes how spatial interaction models can be used to predict the store revenue using a particular example. Chapter 5 covers mainly two situations. In the first one, consumer panel data is available where in the second one only loyalty card data from a grocery retail chain is available. Obtaining the sales of a store in the first situation is easier the researcher can generalize to the store sales using the zip code level panel data and the population of each zip code. We propose a sales generalization method for the second situation. We first generalize the cannibalization amount to the sales obtained from the immediate area customers. Then we generalize the sales obtained from the immediate area to the general store sales. In Chapter 5 we also provide a numerical example where our proposed spatial interaction model in Chapter 4 is used to predict new store sales. In the end of the Chapter 5, an outline for a decision support system based on GIS using our proposed spatial interaction model and generalization method is provided.

Finally, Chapter 6 covers the conclusion and the discussion. The limitations of our model are also discussed. Further research areas are covered throughout the chapter.

Chapter 2

LITERATURE REVIEW

A retail store's location is a critical factor determining that store's success. Consequently, store sales for a potential location can be the criterion for retail site selection. Some site evaluation methods scan the geographical areas to find candidate store locations based on predetermined criteria. These methods are advantageous to discover new opportunities. However, the selected locations may not be available for a new store. For this reason, typically decision makers select the candidate locations and decide which location to select by reviewing their relative predicted success. Site evaluation models which require decision maker to select the candidate locations can be classified as subjective, analogue and analytical methods. Analytical methods can be further examined under the titles of: spatial interaction, optimization and regression models. In this thesis study we focus on analytical methods which require researcher to select the candidate locations, especially spatial interaction models.

2.1 Subjective and Analogue methods

According to Clarkson et al. (1996) almost all retail companies use subjective methods like checklists and analogue methods in addition to the regular financial analysis. Using location specialists who personally visit the potential facility location and decide whether it is a good location based on a checklist and experience is very common in retail site selection. Checklists allow specialists to assess a location on predetermined criteria

(Clarke, 1998). Checklists don't provide sales prediction results, but support subjective decision making. These methods may be somewhat successful depending on the experience of the specialist. Predetermined criteria may be subjective sometimes, such as convenience. Therefore, checklist and similar methods lack the ability to generate consistent results.

Analogue method is developed by Applebaum (1966). Analogue method basically profiles existing store locations and compares new locations to the existing ones to predict the new store sales (Mendes and Themido, 2004). Analogue method is somewhat subjective as it relies on expert opinion in profiling procedure.

2.2 Analytical methods

In addition to the subjective methods, analytical methods for retail site selection also exist, such as regression, spatial interaction and optimization models. Compared to regression and optimization models, spatial interaction models are more tailored for retail store revenue prediction. Spatial interaction models can be used individually for store revenue prediction. They can also be used as components of optimization models for store site selection. Once the store revenue prediction model is constructed either by using only the spatial interaction model or combining it with an optimization model, it can be used as a component of a decision support system. A spatial decision support system for site evaluation would increase the efficiency of site selection decisions.

2.2.1 Regression based models

Regression based models contribute to site evaluation by outputting store sales predictions or customer share predictions. Sales predictions can be obtained using various variables such as distance to the customers, population density, competition and store size. Some retail companies use regression based models in addition to the subjective and analogue models for location assessment (Clarkson et al., 1996). Regression models are advantageous as they are practical, intuitive and flexible in terms of variable specification. However, Clarkson et al. (1996) also point that it is often misused by companies, ignoring sample size requirements. In this thesis study we will not deal with spatial regression. More detailed information about spatial regression, its estimation procedures and techniques to deal with spatial autocorrelation is present in Dubin et al (1999) study.

2.2.2 Optimization models

Optimization models are widely used in location selection and are often used with spatial interaction models. The objective functions in optimization models are different for different types of facilities. For grocery retail facilities, one should maximize the revenue while minimizing the cost. In grocery retail site selection, revenue is mostly determined with customer store selection behaviour. There are different representations of this function and different assumptions about the customer store selection behaviour in the literature. For example, Colome et al. (2003) represent the problem for grocery retailing as a maximum capture model (MAXCAP). In other words, their objective function is to maximize the number of customers captured. According to their representation a facility should satisfy a minimum demand value in order to be opened. As all optimization models should do in this area of research, Colome et al. (2003) assume a structure for customer behaviour for demand distribution. They use the Huff (1963) model for customer store patronization probability generation. According to their model, if the probability generated by the Huff (1963) model exceeds a stochastic threshold then the customer is assigned to the particular store. Craig and Ghosh (1986) represent the same problem for general service companies with an objective function of maximizing total accessibility to the customers. They recommend usage of logit or conjoint analysis for determining customer behaviour. They also recommend values of attributes which may differ from operation to operation to be obtained with consumer surveys. In another study Craig and Ghosh (1983) try to maximize general grocery retail store chain profits by selecting multiple store locations. Number of stores is predetermined. They interpret the output of the MCI (1974) model as the share of customer spending. This thesis study focuses on spatial interaction models and their applications. However, different approaches of site selection exist independent from spatial interaction models thanks to flexibility of optimization models. For example Sakashita (2000) uses optimization for determining convenience store location and size, using competition information. Wong and Yang (1999) uses a continuous equilibrium modelling approach where customers try to minimize a generalized cost function while choosing a store. The generalized cost function considers both cost of distance and the cost of basket (Wong and Yang, 1999).

2.2.3 Spatial interaction models

Spatial interaction models are standardized formulas for predicting customer's share of spending in a particular store. They can be used to estimate the revenue of a potential store. A retail store's sales is equal to sum of customer grocery spending multiplied by customer share of spending (Orpana and Lampinen, 2003).

$$SS_j = \sum_i p_{ij} s_i \qquad (2.1)$$

Where in Equation (2.1) SS_j is sales of store *j*, p_{ij} is customer *i*'s share in store *j*. s_i is the customer *i*'s grocery budget. Often decision makers try to predict a potential store's sales for a future period. For this reason, s_i values should be adjusted for the future period which potential store sales are tried to be estimated.

In most situations, individual customer data such as grocery spending and distance is not available to the researchers. In such situations, researchers often use clusters of customers for retail store revenue prediction. For example, researchers divide locations into grids or just use natural clusters such as zip codes. In this case, *i* in the Equation (2.1) stands for the community *i* and revenue of the store sales is obtained by summing up the sales generated in different communities. In countries such as United States and Japan, zip code level buying power data is often available making the procedure more practical.

Spatial interaction models contribute to grocery store revenue prediction by estimating the p_{ij} values. Spatial interaction models predict customer share in a store by calculating the store's utility relative to the others.

The first model offered by Huff (1963) modeled utility as a function of distance and store size. Let p_{ij} denote the probability of customer *i* visiting store *j*. Then, according to Huff (1963),

$$\hat{p}_{ij} = \frac{\dfrac{w_j}{d_{ij}^{\beta_D}}}{\sum_l \dfrac{w_l}{d_{il}^{\beta_D}}} \tag{2.2}$$

where w_j denotes the store size of the retail location, d_{ij} is the traveling cost of customer *i* to store *j* such as customers' distance to the store , β_D is the distance decay parameter, and the index *l* runs over all stores.

The Huff (1963) model takes only store size and distance to the account as variables to predict customer share in a particular store. In their study, Drezner and Drezner (2002) try to identify additional factors by replacing w_j in the Huff (1963) model by a store specific attractiveness variable in an example with shopping centers. Drezner and Drezner (2002) infer optimal attractiveness values from the Huff (1963) model using buying power index and retail center sales. They estimate the attractiveness values using direct optimization minimizing the difference between predicted and real market share of a store. They also conduct a survey exploring the customers' shopping center choices. The attractiveness scores are later compared to the survey results. According to their findings, for shopping centers survey results and inferred attractiveness scores are consistent. An important implication of this finding is that attractiveness scores are interpretable and can be represented as a function of several variables. Drezner and Drezner (2002) model is covered in more detail in Chapter 4.

A more comprehensive model proposed by Nakanishi and Cooper (1974) is known as the Multiplicative Competitor Interaction (MCI) model. Their model is estimated on log-transformed variables. Ordinary Least Squares and Generalized Least Squares are the methods they propose to be used for parameter estimation. Ghosh et al. (1984) compare different structures for spatial interaction models including the MCI (1974) model and estimate the models with both GLS and OLS regression. MCI (1974) model represents patronization probability as:

$$\hat{p}_{ij} = \frac{\prod_{k=1}^{q} x_{kij}^{\delta_k}}{\sum_{l=1}^{m} \prod_{k=1}^{q} x_{kil}^{\delta_k}} \qquad (2.3)$$

where X_{kij} is the k^{th} variable describing store *j* for customer *i*, and δ_k is the sensitivity parameter for variable *k*. Variables may include distance to the store, competition, store size.

Nakanishi and Cooper (1983) also try usage of zeta-transformed variables. These variables are used to include nominal, ratio and interval data to the MCI (1974) model (Cliquet, 1995). Zeta transformed variables are represented in relative terms as described in Equation (2.4):

$$Z_{kij} = \frac{X_{kij} - \overline{X}_{kij}}{St_k} \tag{2.4}$$

where Z_{kij} is the zeta-transformed k^{th} *v*ariable, X_{kij} is the real value of the k^{th} variable, $\overline{X}_{kij}$ is the mean of k^{th} variable, and St_k is the standard deviation of the k^{th} variable. Usage of zeta-transformed variables increases the applicability of the MCI (1974) model as it allows usage of nominal, ratio and interval data.

Black, Ostlund and Westbrook (1985) compare performances of different forms of MCI (1974) model: a basic model with a distance measure, a more comprehensive form with marketing strategy variables and distance, a scale-value model which includes a composite variable for store attractiveness parameter and finally an outlet specific model in which sensitivity parameters of strategy variables are estimated individually for stores. They add store characteristics such as discount pricing, advertising and reputation to the model as marketing strategy variables. According to their findings, distance, store size, advertising, reputation and discount pricing are significant variables. Moreover, 2nd , 3rd and 4th models provide similar results in terms of prediction accuracy which is significantly compared to 1st model .

The multinomial logit model is also widely used for share prediction (Fotheringham, 1988). The multinomial logit model describes the share of spending of customer *j* in store *i* as in the Equation (2.5) and is very comprehensive like the MCI (1974) model in terms of variable specification.

$$\hat{p}_{ij} = \frac{e^{\sum_k X_{kij}\delta_k}}{\sum_j e^{\sum_k X_{kij}\delta_k}} \qquad (2.5)$$

where δ_k is the coefficient of variable *k*. The multinomial logit model has the independence from the irrelevant alternatives property. Fotheringham (1988) describes independence from the irrelevant alternatives property as customer's probability of patronizing a store is independent of the other alternatives so that when a new alternative is added to the existing choices the predicted share for a store has to be in proportion to its original share. However, in retail this property doesn't hold as customer may choose the store which is in close proximity to other stores for comparison purposes (Fotheringham, 1988).

A generalization of the multinomial logit model, the mother logit model is covered in the study of Borges et al. (1991) study. Where, in the mother logit model utility of a store both depends on the attributes of that particular store and attributes of other stores in the choice set. The mother logit model simply adds additional constants for correction of misspecifications caused by independence from irrelevant alternatives property of multinomial logit model (Borges et al., 1991). Borges et al. (1991) study indicates that the mother logit model perform slightly better than the multinomial logit model which they think as non-significant from a managerial perspective. However, they recommend researchers to evaluate the significance of additional constants in other words, the cross-effects. If the cross effects are non- significant, researchers can use the multinomial logit model, which means misspecifications caused by the independence from irrelevant alternatives property are negligible (Borges et al., 1991). A side finding of their study is that competition effect is stronger among the stores in the same hierarchical level.

The nested logit model is an extension of the multinomial logit model and it has a hierarchical nature. It describes the store choice probability as in Equation (2.6).

$$p_{ij} = \begin{cases} \dfrac{e^{\sum_k X_{kik}}}{\sum_j e^{\sum_k X_{kij}}} & \forall j' \in CL \\ 0 & \forall j' \notin CL \end{cases} \tag{2.6}$$

where CL is the choice set of the customer. Fotheringham (1988) says that consumers do not necessarily evaluate all competing stores because of time and other limitations. He also says that customers first form a choice set and evaluate the alternatives in it.

Benito, Gallego, Rayes (2006) use geo-demographic segmentation and the nested logit model together. They focus on geo-demographic characterization of discount stores, hypermarkets and regular stores. The factors they consider include level of business activity in the area, employment, family characteristics, education and population density. They estimate their model parameters with maximum likelihood estimation. According to their findings, supermarkets are mostly patronized by highly educated people where hypermarkets and discount stores are selected by people having lower education levels

Fotheringham (1988) proposes a special kind of nested logit model known as the competing destinations model. Customer store choice probability according to the competing destinations model (1988) is described as in Equation (2.7). The competing destinations model (1988) is different from the nested logit model because the probability of visiting a store is determined by both the utility of a particular store and the probability of that store to be in the choice set. The probability of a store to be in the choice set is defined with a likelihood function which is often a function of distance.

$$\hat{p}_{ij'} = \frac{e^{\sum_k X_{kij}\delta_k} L(j' \in CL)}{\sum_j e^{\sum_k X_{kij}\delta_k} L(j \in CL)} \tag{2.7}$$

where $L(j \in CL)$ is likelihood of store *j* to be in the choice of selected customers. Different likelihood functions are proposed in different studies like Borges and Timmermans (1987), Fotheringham (1983), Meyer and Eagle (1982). These functions try to represent the dissimilarity of the alternatives in terms of distance or other variables (Erymann, 1995). The competing destinations model doesn’t have the independence from irrelevant alternatives property (Fotheringham, 1988).

Nakaya et al (2007) use micro simulation and the competing destinations model (1988) together in their study. They group customers into lifestyle segments and estimate parameters of the competing destinations model (1988) according to each customer group. Micro-simulation basically simulates the shopping behavior of customer groups. Their research findings suggest that micro-simulation can successfully used for data generation. Moreover, customer groups actually have different attractiveness functions. For example, price and quality turned out to be an important factor for families where store size is equally significant. According to Nakaya et al. (2007) store size should be considered together with variables such as product portfolio and pricing.

All of the models discussed above are widely known and applied in the literature. However, a complete comparison of the above models in grocery retailing is not available. Distance is the default variable of the spatial interaction models. We believe that distance variable's effect may vary in each store especially depending on the store size. None of the spatial interaction models introduced in this section address to this issue. We see the relationship between the effect of distance and store size variables as a research opportunity. We also believe that Drezner and Drezner's (2002) research is worthwhile to further examine in grocery retail setting and can be used to improve sales prediction performances.

2.3 Spatial decision support systems

A geographical information system (GIS) is a combination of a digital map and a database in which users can store visual and text data. The availability of geographical information systems allows usage of more analytical and complex methods for site selection as spatial, customer and time data can be stored in the same environment (Longley, 2004; Byrom et al., 2001; Hernandez, 2005). Also, geographical information systems allow decision makers to see distribution of their customers. Geographical information systems can become spatial decision support systems for store site selection and sales prediction if they are enhanced with analytical site evaluation tools such as spatial interaction models.

Clarke et al. (1995), describe a decision support system's characteristics as, ability to support decisions in unstructured and structured problems, ability to provide effectiveness, efficiency and practicality in decisions and ability to integrate data and analytical methods. If the GIS system is integrated with spatial interaction models they can support retail site selection decisions by providing sales predictions for the candidate location. A spatial decision support system supporting location decisions would obviously be more practical, efficient and effective compared to using analogue and subjective methods which most companies do (Clarkson, 1996).

There are four main components of a decision support system: The decision model, interface, analysis module and the database management system (Clarke et al., 1995). The interface and the database management system already exist in the GIS system. A spatial interaction model if integrated with GIS can easily be used as decision model component of a spatial decision support system. Analysis model should be integrated to the GIS system according to the needs of the users.

Klosterman and Xie (1997) use GIS, spreadsheets and spatial interaction models together in their proposed decision support system for site selection of grocery retail stores. They use a version of the Huff (1963) model as their spatial interaction model. Their model has both store size sensitivity and distance decay parameters. For parameter selection, the program generates sum of squared error values for different sensitivity parameters. The sum of squared error values are calculated on the previous sales data. Their spatial decision support system provides sales predictions as output.

Usage of spatial interaction models and GIS is covered in great detail in Batty et al. (1996), Fotheringham et al. (2000) and Fotheringham et al. (1994).

Chapter 3

EXPLORATORY ANALYSIS

3.1 Introduction

The purpose of this chapter is to identify important variables of customer store choice behavior. Identified variables and insights obtained from this chapter are later used for spatial interaction model construction in Chapter 4. Another goal of the chapter is to explore the relationship between the distance and customer store choice behavior. In this chapter we examine effect of new store openings on existing stores on macro and customer level. Macro level analysis should provide us insights about competition in addition to sales patterns in grocery retail. Customer level analysis is used to test several variables with data mining tools and identify important ones. Also, in this chapter we compare share prediction performances of data mining models with Huff (1963) model. This evaluation comparing the data mining tools with a conventional spatial interaction model led to our decision to continue our research with spatial interaction models.

3.2 Macro level effects of a new store opening

3.2.1 Experiment design and data

This analysis aims to explore the effects of a new store opening on existing store sales and can provide us insights about competition variable. This analysis should also provide us insights about the sales patterns in grocery retailing. For this analysis a metropolitan, mixed residential and business area in Istanbul is chosen. There are two new store openings and two existing stores in the area. Of the stores provided in Table 3.1, 3 and 4 are the new stores and their effects on existing stores; 1 and 2 are examined in macro level in this section.

Store	Brand	Format	Store size (m2)	Opening Year
1	Migros	MMM	2200	1991
2	Migros	M	420	1998
3	Migros	MMM	3440	2005
4	Tansaş	Midi	507	2004

Table 3.1: Attributes of stores included in the analysis

Attributes of the stores included in the analysis are provided in Table 3.1. First three stores in Table 3.1 are of the Migros brand and last store is Tansaş brand. Both Tansaş and Migros are grocery retail chains. Tansaş company is acquired by Migros company in 2005. For Migros stores, MMM is the store format with the highest product variety and M is the store with the lowest product variety. Midi is the medium product variety store of Tansaş retail chain. Store sizes of all four stores are also provided in Table 3.1. By store size, we refer to area of the store excluding the storage spaces. Distances of the stores to each other are provided in Figure 3.1.

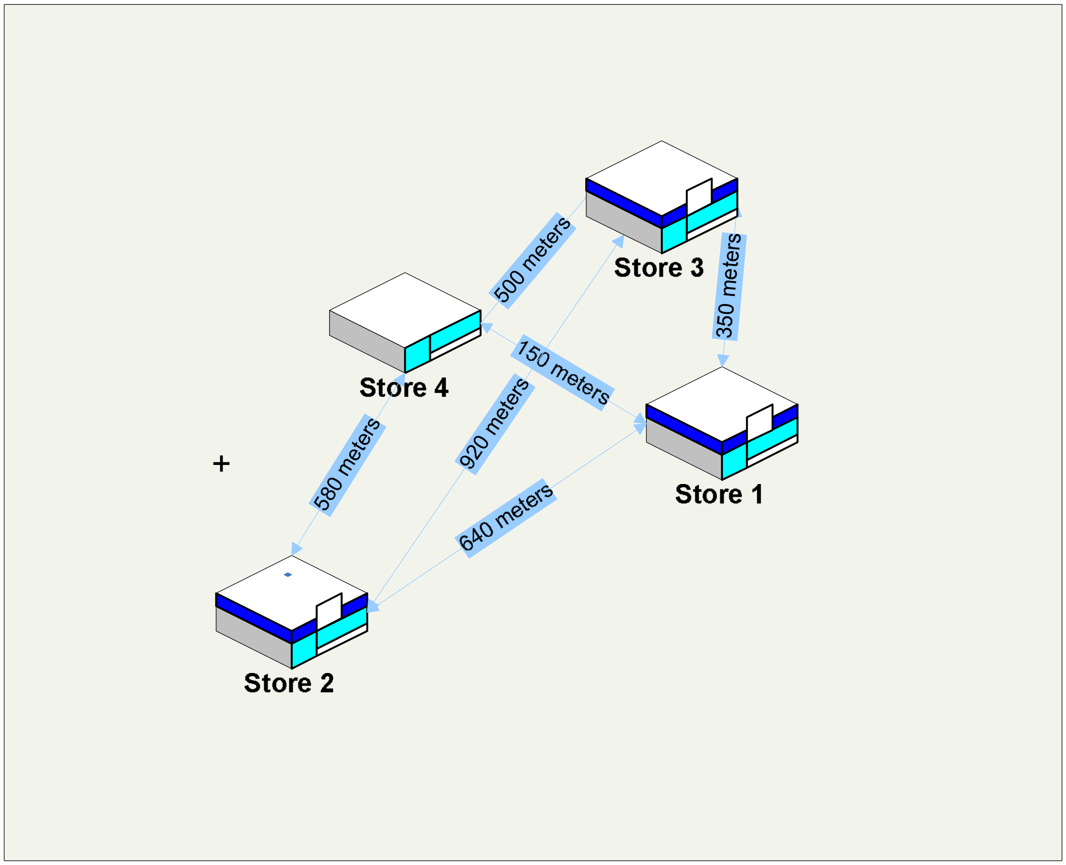


Figure 3.1: Locations of the stores and their distances to each other.

In the macro level analysis we examine the sales trends of existing stores: 1 and 2 in 3 periods to understand the effects of new store openings. As the control series, we also examine the Migros İstanbul sales excluding the new store openings after 2003 in three periods.

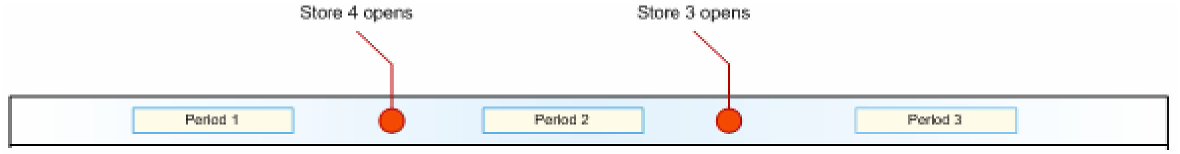


Figure 3.2: Macro level analysis periods.

Period 1 in Figure 3.2 is the period before the Store 4's opening. Length of this period is 21 months. Period 2's length is 11 months and it is defined as after the store 4 opening and before the store 3 opening. Finally, period 3 is after the both store openings and its length is 13 months.

In macro level analysis, data provided in Table 3.2 is used.

	DATA
1	Monthly sales of the stores 1,2,3 and 4 in periods 1, 2 and 3.
3	Monthly sales of Migros İstanbul stores excluding the stores opened after 2003

Table 3.2: Data used in macro analysis

3.2.2 Macro level effects of a new store opening

We use the chain sales from stores in Istanbul as a control series for the selected area stores that indicate the sales trends that could be expected had there not been any changes in the environment. The stores which opened after 2003 are excluded from this dataset in order to differentiate between the effect of time and additional stores. Through the period we examine store chain sales are in growth in Istanbul. Furthermore, the data is highly seasonal as expected. Sales make a peak in January and a decrease in July. Istanbul sales can be examined in Figure 3.3.

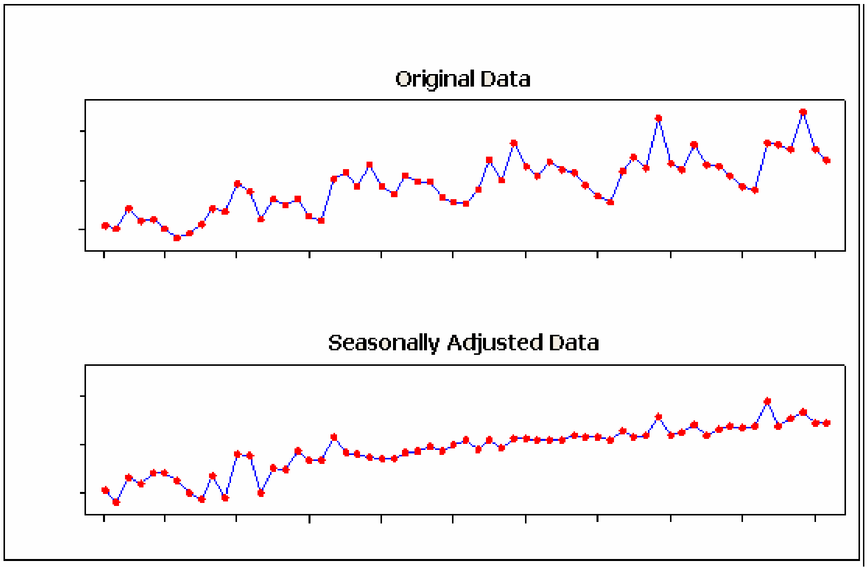


Figure 3.3: Migros İstanbul sales excluding new stores.

According to Figure 3.3 Istanbul sales have a positive trend through the period we examine. The Figure 3.3 provides us the grocery retail sales in İstanbul excluding the new store openings in periods 1,2 and 3.

Before starting the trend analysis we deseaosonalize the data to eliminate the monthly seasonality effects using multiplicative component analysis (see e.g Hanke and WİChern, 1988) with Minitab version 14. Deseasonalized data is available in Figure 3.3 After deseonalizing the data we examine the trend in 3 periods to see whether it is significant. We compare the trend to store trend as a control set in later sections.

We run separate linear regression analyses against time for the deseasonalized Istanbul chain sales excluding the new store openings for the three periods. We want to examine whether trend changed significantly in the periods defined in Figure 3.2. In the regression analysis, independent variable is time (month) where dependent variable is the deseasonalized sales. Regression analysis significance results and regression equations are provided in Table 3.3 for Migros İstanbul sales excluding new store openings.

Period	R-square	Regression equation	P-value for time
1	0.48	intercept+418*time	0
2	0.63	intercept+284*time	0.0003
3	0.40	intercept+125*time	0.02

Table 3.3: Regression analysis for trend Migros Istanbul sales excluding the new stores

The regression analysis results indicate significant positive trend at α=0.01 with decent R-square values for all three periods. We see that sales increases for three periods. If we observe a trend change in the store level analysis, we can say that this change is not due to sector related factors.

3.2.3 Store level analysis

In order to see the effects of new store openings or competition on existing stores sales, we examine the sales trend of existing stores and question whether a significant change occurred concurrent with the store openings in the area and compare these changes to the control series sales trends.

Store 1

Before closely examining three periods for store 1, we deseasonalize the data as we did in the general Migros Istanbul sales data. After removing the season effects from the data we can study closer the trends in three periods. Via regression analysis we can see whether a significant change occurred in store 1 sales trend after the new store openings. Independent variable is time and dependent variable is deseasonalized store 1 sales. Regression analysis results for three periods are available in Table 3.4.

Period	R-square	Regression equation	P-value for time
1	0.83	intercept+12*time	0
2	0	intercept+146*time	0.96
3	0.41	intercept-3*time	0.01

Table 3.4: Regression analysis results for store 1 for three periods

If we closely examine Table 3.4 we will see that in period 1 store 1 has significant sales increase. In period 2, growth of store 1sales stops. The trend, based on the regression analysis, is not significant. Moreover, in period 3 there is a significant negative trend. Sales of store 1 start to decrease in this period. Comparing to the steadily increasing sales of the chain excluding the new store openings in Istanbul overall, the flat and decreasing trends suggests that, opening of the competitor caused the focal store's sales increase to stop, while the opening of the larger chain store caused the focal store sales to decrease.

Store 2

After removing the seasonality from the data we run the regression analysis for three periods on store 2. Regression analysis results are provided in Table 3.5.

Period	R-square	Regression equation	P-value for time
1	0.23	intercept+0.7*time	0.02
2	0.08	intercept+0.7*time	0.4
3	0.49	intercept+1.7*time	0.01

Table 3.5: Regression analysis results for Store 2 for 3 periods.

Results of store 2's regression analysis are quite interesting. In the first period sales increases by time and regression analysis is significant although R-square value is low. In the second period sales stops increasing, as regression analysis is not significant. Finally, in the last period we observe that sales start to increase again. From the regression analysis we can conclude that store 4 affects store 2 negatively. However, store 3's opening doesn't affect store 2. Examining the Migros İstanbul sales trend analysis and store based analysis we can say that store 1 sales trends are not parallel to general sales trends. However, store 2 sales trends are more similar to the general sales trends.

The opening of the competitor store affects both stores negatively by inhibiting their growth while the sales of the control stores continued to increase. On the other hand, the store 3 opening has different effects on the two existing stores in the area. We can think of two reasons for the difference of the results between store 2 and store 1. The first reason is, store 3 is distant to store 2 compared to store 1. The second reason is, store 2 is a small store where store 1 is a big store with a store size of 3440 square-meters. So the theory is customers may have loyalty to store formats. It is important to note that store 1 which is more affected from store 3's opening is a big store like store 3. The opening of a large store that is not very close may have increased brand loyalty and hence traffic in the small store for convenience shopping, while the closer medium size store may have seen its sales cannibalized by the large store. To summarize the insights about competition effects: Generally, competition affects store sales negatively. However, this effect may vary according to distance between the stores or types of stores.

3.3 Customer level analysis

Goal of this analysis is to gain insights about customer shopping behavior and identify the essential variables. In customer level analysis, under the title of preliminary analysis (Section 3.3.2) we first examine the relationship between sales generated in the immediate area of the store and store size. Secondly, under the same title we examine the relationship between customer distance to a store and customer spending.

From macro level analysis results we know that new store openings affected both store 1 and 2. This effect can be further explored from a different perspective: customer behavior. Examining customer behavior is necessary to understand the factors determining a stores success. We try to pursue this goal in predictive analysis in Section 3.3.3, predictive analysis. In predictive analysis, we try to predict customer share of spending after the new store opening with data mining models and Huff (1963) model. We also compare the customer share of spending prediction performances of data mining models and Huff (1963) model.

3.3.1 Data

In customer level analysis of the two new store openings we study store 3 due to limitations on the customer level data in store 4. Attributes of store 3 is provided in Table 3.1. For customer level analysis, we select the customers located in a circle originating from the store 3 with a radius of 1000 meters with the GIS of Migros. 1000 meters is chosen for practical purposes only. Goal of this analysis is not to predict store sales but to gather insights about the grocery retailing. Radius of the circle does not play a significant role in this experiment.

An illustration of the area which we select the customers from is presented in Figure 3.4. As can be seen from the Figure 3.4 the area also includes the stores introduced in Table 3.1. Store 4 also exists in the selected area. However, we don't have customer level data for Store 4. For this reason, store 4 is not included to this analysis.

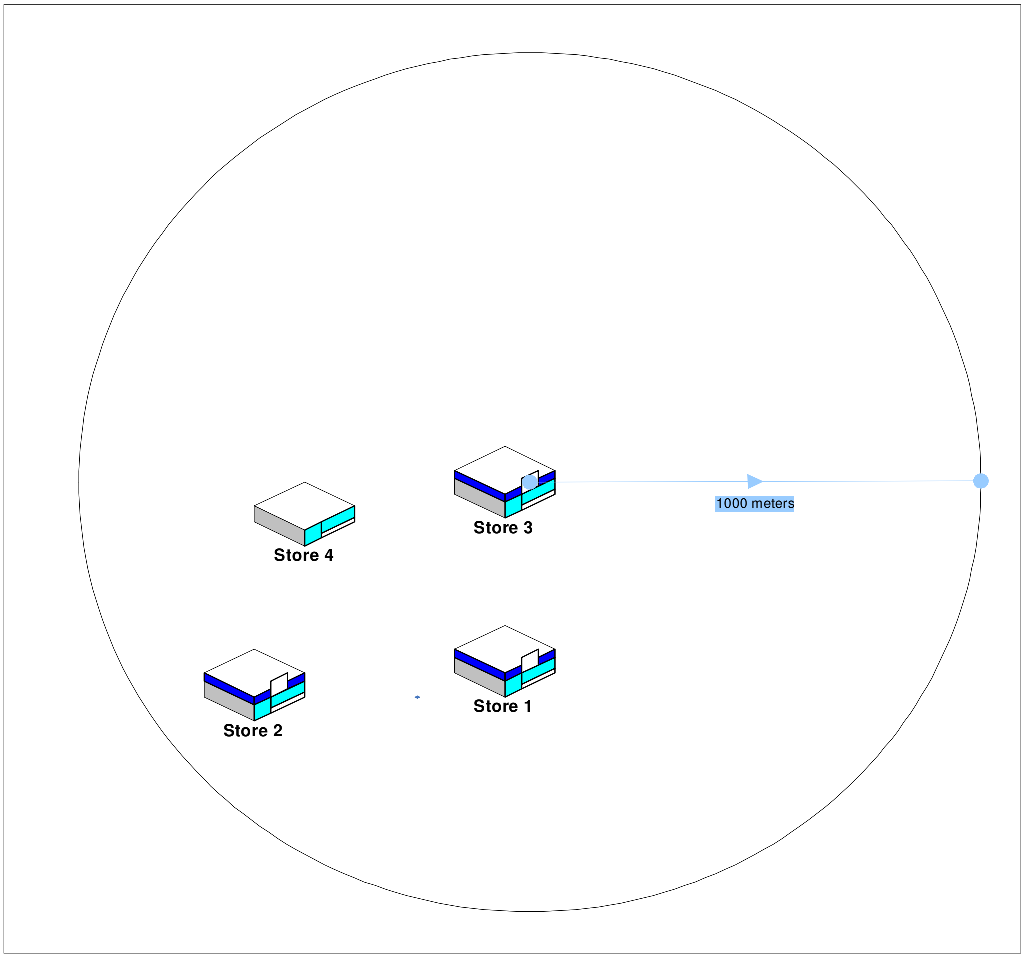


Figure 3.4: Selected area

3.3.2 Preliminary analysis

Loyalty card data for selected customers are available for this analysis. The data we use in the preliminary analysis is provided in Table 3.6.

	DATA
1	Monthly sales of stores 1, 2 and 3 between 2003 and 2007
2	Monthly spending of customers in the selected area in 3 stores between 2003 and 2007
3	Driving distances of customers in the selected area to 3 stores.

Table 3.6: Data used in preliminary analysis

19,947 Migros customers live in the selected area. Distribution of the customers according to number of stores they shop in the selected area is provided in Table 3.7.

Number of stores	Number of people
0	13314
1	4582
2	1841
3	210

Table 3.7: Distribution of customers according to stores they shop in the selected area

For example, according to Table 3.7, 210 people shop from store 1, 2 and 3. Moreover, 13314 customers don't shop from the selected area at all. Out of these 13314 customers, 10649 customers are not active in any store of the chain. This number corresponds to 53% of customers in our sample. It is important to note that we define active customers as people who spend more than 100 YTL on average in three years. Moreover, active customers spend 51% of their Migros spending in Şişli area.

Percentage of sales generated in the immediate area versus store size

As discussed in the previous sections, for this analysis we select the customers located in the circle originating from store 3 with radius of 1000 meters. We call this area the immediate area of the store, immediate area radius is defined by the researcher. Catchment area also known as trade area is defined as the area where measurable amount of the store sales come from (Melaniphy, 1992). Immediate area concept is similar to the catchment area but it is more restricted to the vicinity of the store. The ratio of spending of customers in our sample to total store sales is provided below in Table 3.8. As can be expected, for a small store such as store 2, a higher percentage of sales come from the area.

Store	Percentage of sales coming from Şişli area
1	0.14
2	0.40
3	0.06

Table 3.8: Percentage of sales generated in the selected area

The values on Table 3.8 are calculated on the average of yearly purchase of customers divided by yearly sales of a particular store. The goal of this calculation is to gather insight about the relationship between the sales generated in the immediate area and store size. Radius of the immediate area does not have a significant effect on this analysis as all stores are affected from the radius and we are evaluating the percentage of sales generated in the area relative to the other stores.

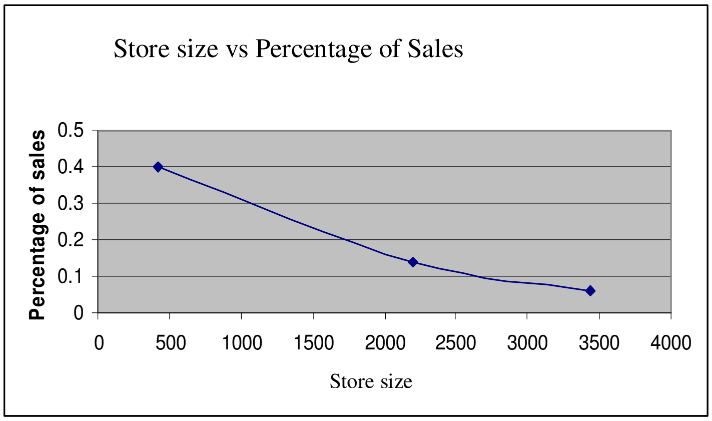


Figure 3.5: Percentage of sales generated in the area versus store size

When we compare percentage of sales coming from the selected area by examining Figure 3.5 we see that a bigger percentage of small store sales are generated in the selected area. We can conclude that bigger stores attract more customers from distant areas. From the above finding, we may infer that bigger stores have bigger catchment areas.

Numbers in Table 3.7 underestimate the actual sales that come from the area since we have only the loyalty card data and only around 45% of the loyalty card customers are coded in geographical information system of Migros**.** In addition, loyalty card data captures 80% of the customer spending on average.

Effect of distance on customer spending

In order to form an idea about the effect of distance on customer store patronization behavior we graphed average customer spending on our dataset with distance for stores 1,2 and 3 in Figures 3.6, 3.7 and 3.8 respectively. The spending values provided in Figures 3.6, 3.7 and 3.8 are average of yearly spending values for customers in the specified distance to the store.

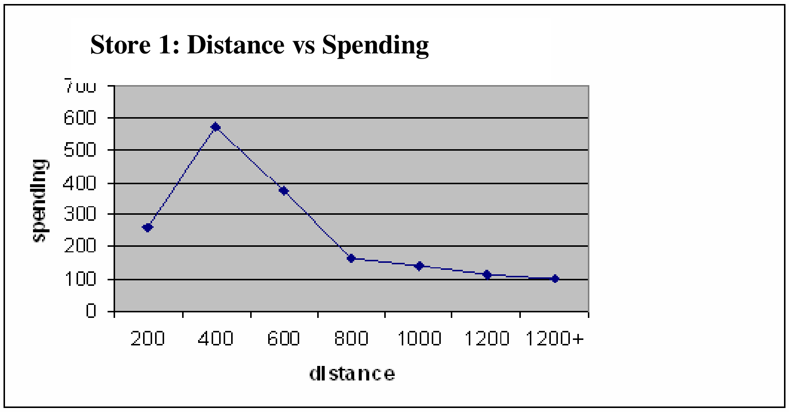


Figure 3.6: Customers' Store 1 Spending and Distance

Figure 3.6 shows a non-smooth decreasing relationship between customer spending and customer distance to the store. The interruption in the decreasing pattern may be the result of competition. store 4 is closely located to store 1 causing spending of customers in close proximity to the store 1 to decrease.

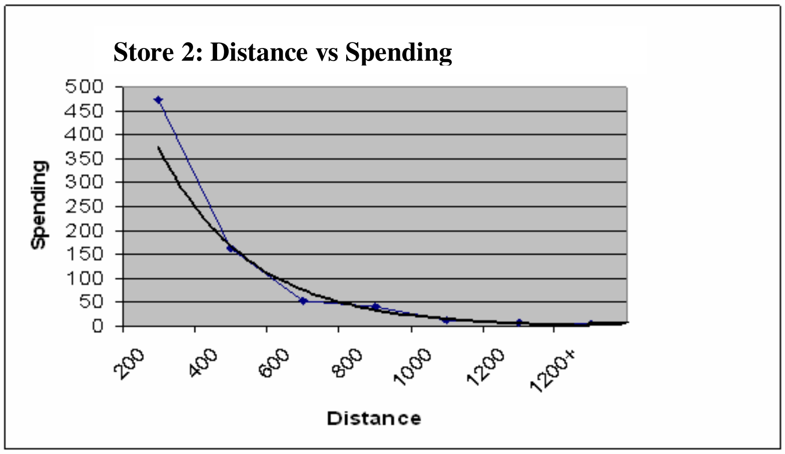


Figure 3.7: Customers' Store 2 Spending and Distance

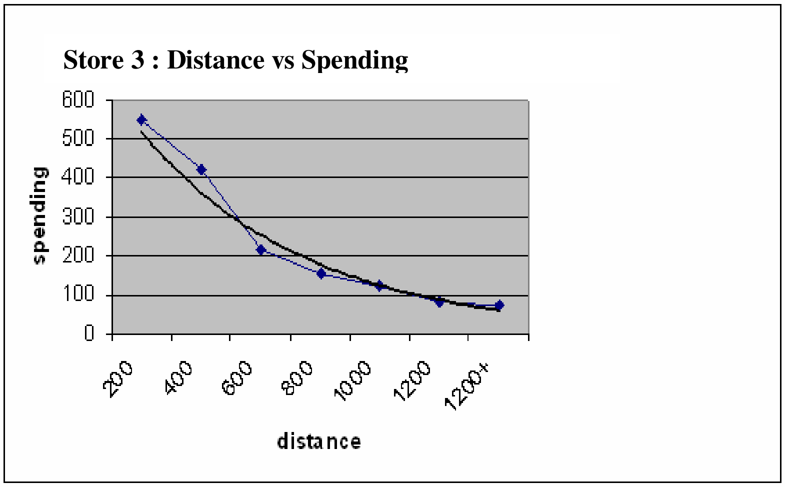


Figure 3.8: Customers' Store 3 Spending-Distance

Examining the Figure 3.6, 3.7 and 3.8 we can conclude that customer spending decreases as distance increases. If we study these graphs closer, we see that store 1's graph is a little different from store 2 and 3's plots. For store 1 there is an interruption in spending decrease at 400 meters. This may be result of its neighbor store 4. Moreover, plot of store 3 is very much similar to store 1's graph. One difference is that the graph of store 2 is steeper. So, as distance increases spending decreases with a higher rate for store 2. This result is not shocking considering the format differences between stores. Store 3 is a 3M store, its store area is around 3000 square meters and is located in a busy shopping mall where store 2 is a M store with a 407 square meters store area. Although we can see the nice trend in store 3's distance-spending plot we should keep in mind that this graph only includes customers who live in Şişli area. Given that store 3 is in a shopping mall and visited by people all around İstanbul, we may not see this nice curved trend if we include all customers of store 3.

Main insights of this section are: Distance effects customer grocery spending negatively. Distance's effect may differ according to store size of the store. The decreasing trend between the spending and distance may be interrupted as a consequence of competition.

3.3.3 Predictive analysis: Customer share prediction for variable identification

In this section we try to determine significant variables affecting customer store patronization behavior via data-mining and compare performances of data mining models. We also compare data mining models with Huff (1963) model. This analysis is similar to macro analysis in its nature. We adopt a before-after type analysis for customer share prediction. Different from macro level analysis we are using customer level data and we are only examining store 3's opening. We use customers located in the immediate area of store 3 with radius of 1000 meters. An illustration of the area is provided in Figure 3.4. We know customer share of spending distribution among the existing stores: store 1 and 2 and we are trying to predict the customer shares in 3 stores after store 3's opening. We also know customer shares after store 3's opening but we are using this information for performance evaluation. Share is defined as, customer spending in a particular store in the selected area divided by the total customer spending in the stores in the selected area. For example, customer share before store 3 opening for store 1 is defined as,

$$\text{Before the new store opening share for store 1} = \frac{Store\ 1}{Store\ 1 + Store\ 2}$$

$$\text{After the new store opening share for store 1} = \frac{Store\ 1}{Store\ 1 + Store\ 2 + Store\ 3}$$

For store 3, before the new store opening share is accepted as 0.

To summarize, using the data before store 3's (period 1 and 2) opening we try to predict customer store shares after the store 3's opening (period 3). Dataset is prepared according to customer-store pairs. Therefore, we have 3 instances for each customer corresponding to the three stores in the area.

Input variables for the data mining models are: customers' share of spending before the new store opening, customers' distance to the store, store size and store format. Before the new store opening share is calculated on the total spending of the customer in periods 1 and 2, this is a 32 month period. The dependent variable we are trying to predict: after the new store opening share is calculated on the total spending in period 3 which in a 13 month period. Distance values are driving distances to the stores. Store format is the Migros store format it can take values of: M, MM and MMM. Where M is the store format with lowest product variety and MMM is the store with the highest product variety. For this analysis we only use M and MMM, as we don't have a MM store in our experiment. The dependent variable is the particular store's share of the customer's spending after the new store opening.

Variables	Type	Dependent/Independent
Distance to the store	Numeric-continous	I
Share before the store opening	Numeric-continous	I
Store format= M	Binary	I
Store format=MMM	Binary	I
Store size	Numeric-continous	I
Share after the store opening	Numeric-continous	D

Table 3.9: List of variables

For model comparison purposes, we run models in Table 3.10 using variables in Table 3.9 on our data. We use variables presented in Table 3.9. Models are constructed on the dataset consisting of 8787 instances. The test set includes 5814 instances. Error values reported in Table 3.10 are calculated on the test set. Models are run on Weka version 3. 4.

Model	**Mean Absolute Error**
Linear Regression	0.31
Regression tree (M5P)	0.29
Decision table	0.30
Bagging	0.30
Huff model	0.26

Table 3.10: Data mining models performance results for share prediction

According to Table 3.10 the Huff (1963) model introduced in Chapter 2 in Equation (2.2) gives better prediction results compared to data mining methods. The Huff (1963) model uses distance and store size as variables. According to the Huff (1963) model utility of a store is equal to store size divided by a power of the distance.

$$\hat{p}_{ij} = \frac{\dfrac{w_j}{d_{ij}^{\beta_D}}}{\sum_l \dfrac{w_l}{d_{il}^{\beta_D}}}$$

where w_j denotes the store size of the retail location, d_{ij} is the customers' distance to the store , β_D is the distance decay parameter, and the index *l* runs over all stores.

Store shares are obtained by proportioning a stores utility to sum of utilities of stores in the area. We use 2 as the distance decay value for this analysis.

As Huff (1963) model performs better than data mining models which use more information, we can conclude that store size and distance are important variables and explain most of the variance in customer store choice behavior from this result. We can also say that distance has a negative effect on customer store choice behavior where store size has a positive effect. We can also conclude that spatial interaction models are very effective in customer share prediction. We think that improvement comes from the special structure of the Huff (1963) model which implicitly takes competition into account by proportioning that stores utility to sum of utilities of all stores.

Among the data mining models Quinlan's M5P regression tree (1992) works better than simple linear regression and other data mining tools provided in Table 3.10 on our data. More detailed information about the models provided in Table 3.10 can be found in Weka version 3.4. For interpretation and simplicity purposes we can say that regression tree is the best method for share prediction among other data mining models tried. Regression tree and rules are provided in Appendix A. According to the regression tree, distance, share before the new store opening, store size and M as a store format are important variables creating the branches of the regression tree. Examining the regression tree we see that generally distance has a negative impact on after the new store opening share. Thus, higher the distance to the store a person's probability of switching stores is higher. Also, if the distance to the new store is lower probability of customers patronizing that store is higher. It is important to note that for a M store distance variable has negative coefficients for all branches of the tree. We can hypothesize that smaller stores are more affected from the distance variable. As expected, customers share before the new store opening in a particular store is high it is less probable for that customer to switch stores.

3.4 Conclusion

The goal of this chapter is to obtain insights about customer store choice behavior and identify important factors. We pursue this goal by exploring the effects of a new store opening on the existing stores and the customers of the existing stores on the case of a 1000 meters radius area originating from store 3. The area is residential and business mix within Istanbul. In Section 3.3 we observe that a competitor store opening affected both the stores in the area negatively. On the other hand, the opening of a large store of the same chain decreases the closer midsize store sales appreciably, while the farther small store sales rather increased. Based on this observation, we conclude that when a new store opens it may affect similar stores negatively but not less similar stores. Another hypothesis to explain this behavior is that the distance is the overriding determinant of the adverse effect. Moreover, this may be because customers develop type based loyalty. In Section 3.3 we also observe that retail sales have a trend over time and is subject to seasonality.

In Section 3.3 we explore the factors affecting customer store choice behavior. We conclude that generally distance is an important factor. In Section 3.3.2 we observe a negative relationship between distance and customer spending for all stores. Different from stores 2 and 3, this trend is interrupted with a peak in the case of store 1. We hypothesize that, the interruption in the trend is caused by competition in the area. Another observation about the relationship between the distance and spending is that as distance increases spending decreases more rapidly for store 2 compared to store 3. We believe that this is the effect of store size. In Section 3.3.2 we also observe that bigger percentage of store sales come from the immediate area for smaller stores. We hypothesize that as store size increases customer are less affected from distance.

In Section 3.3.3 we observe that Huff (1963) model works better than data mining tools. Examining the Huff (1963) model and data mining results, we conclude that distance and store size are important factors. If a customer is closely located to a existing store his probability of switching stores is less. If a customer is closely located to a competing store his probability of switching to a new store is higher. In addition, a store with store format M is more affected from distance. This may be the result of the store size effect. In this section we also observe that store size affects the customer store share positively. We see that store size has a positive effect on customer shopping behavior. Moreover, we see that customers' previous shopping behavior is significant in determining after the new store opening shares.

Finally, based on our findings in this chapter, we focus on distance, store size, customers before the new store opening share and competition as variables in the next chapters. Moreover in the following chapters, we explore the relationship between the spending decrease caused by distance and store size. In addition to these, we also want to examine area related factors which we ignored in this chapter such as workplace density and population density. Also, because of the seasonality of the retail sales data in the following chapters we will complete our analyses on yearly spending data. As Huff (1963) model gives better prediction results than data mining tools we confirm our decision to continue with spatial interaction models to our research.

Chapter 4

SPATIAL INTERACTION MODELS IN RETAIL SITE SELECTION

4.1 Overview

In Chapter 3, the exploratory analysis we conclude that distance, store size, competition are among the important factors to be considered for modeling consumer store choice behavior. One of our side findings is that bigger stores attract more customers from abroad. Hence, they are less affected from distance. In Chapter 3 we observe that spatial interaction models are quite robust and even work better than less structured methods such as regression trees for customer share of spending prediction. Based on the last finding we dedicate this chapter to customer share of spending prediction with spatial interaction models. In the light of the previous findings we try to build a precise and flexible spatial interaction model.

Spatial interaction models generate customer store shares of spending as outputs. Generated outputs can be used in grocery store revenue prediction which is our ultimate goal. In this chapter we examine the most widely used spatial interaction models. These models are studied thoroughly for their customer share and store sales prediction performances. We also propose a new spatial interaction model for customer share prediction. Properties of the related optimization problems for parameter calibration are also examined.

4.2 Spatial Interaction Models

4.2.1 The Huff (1963) Model

The Huff (1963) model is the one of the earliest and most widely used spatial interaction models. It defines utility of a grocery retail store as store size inversely proportional with a power of the customer distance to the store. The power of the distance is called the distance decay. The Huff (1963) model leaves estimation of distance decay parameter to researcher. It is calibrated in each analysis, and it is not store specific. It often takes a value around 2 for grocery retail facilities.

Let p_{ij} denote the probability of customer *i* visiting store *j*. Then, according to Huff (1963),

$$p_{ij} = \frac{\dfrac{w_j}{d_{ij}^{\beta_D}}}{\sum_l \dfrac{w_l}{d_{il}^{\beta_D}}}$$

where w_j denotes the store store size d_{ij} is the traveling cost of customer *i* to store *j*, β_D is the distance decay parameter, and the index *l* runs over all stores. Our variable is β_D.

According to Haines et al. (1972), in the literature Huff (1963) model parameters are generally estimated via Huff-Blue procedure. The Huff-Blue procedure uses a Fibonacci search approach over an interval of potential β_D values. The method tries to find a distance decay parameter yielding an error value which is lower than a predetermined value. According to Haines et al. (1972) if the method can not find an error value lower than the predetermined value, it reports the best error value. Haines et al. (1972) propose usage of maximum likelihood for estimation of the distance decay parameter. They obtain the parameter by first taking the logarithm of the likelihood function and then differentiating it. They use direct search to find the parameters maximizing the likelihood function. Consequently, their maximum likelihood estimates are not necessarily the minimum variance estimators or found estimates are not necessarily the global optimal solutions (Haines et al., 1972). According to Haines et al. (1972) chances of finding the global optima depends on the usage of direct search. The model parameter can also be estimated by optimization, minimizing the MSE. The optimization model proposed for parameter calibration of Huff (1963) model is provided in M.1. The variable in M.1 is the β_D value.

(M.1)

Minimize,

$$MSE = \frac{1}{OM} \sum_i \sum_j (p_{ij} - \hat{p}_{ij})^2 \tag{4.1}$$

Subject to,

$$p_{ij} = \frac{\dfrac{w_j}{d_{ij}^{\beta_D}}}{\sum_l \dfrac{w_l}{d_{il}^{\beta_D}}}$$

Where M is the number of customers and O is the number of stores. If Model 1 is convex then we can be sure that found solution is the global minimum. However, according to the Definition C.1, in Appendix C the Huff (1963) model's error function defined by Model 1 is not quasiconvex or not convex. Counterexample for quasiconvexity of Huff (1963) model error function is provided in Appendix C, Table C.1. Thus, we may not converge to global optima.

Although the error function of the Huff (1963) model is not convex or quasiconvex, it is quite stable and we obtain consistent results when we re-run the model with different starting points on numerical examples. Because we have only two dimensions we can easily check the quality of our local optimum. We can graph the distance decay values with the error values to see the shape of the function. As none of the models proposed in the literature can provide global optimum we can use the direct optimization model provided in (M.1) because of its practicality.

4.2.2 The MCI (1974) Model

The MCI (1974) model is proposed by Nakanishi and Cooper. It is a widely used spatial interaction model both in site selection and market share prediction. It mainly differs from Huff (1963) model in terms of the variables used to predict the store share. The MCI (1974) is a general model that allows the researcher to define the explanatory variables. The MCI (1974) model is provided below.

$$\hat{p}_{ij} = \frac{\prod_{k=1} x_{kij}^{\delta_k}}{\sum_{l=1}^{m} \prod_{k=1} x_{kil}^{\delta_k}}$$

where X_{kij} is the k^{th} variable describing store *j* for customer *i*, and δ_k is the sensitivity parameter for variable *k*.

δ_k are estimated by log-transforming the variables and applying OLS regression. The regression equation then becomes,

$$\log(\frac{p_{ij}}{\tilde{p}_i}) = \sum_k \delta_k \log(\frac{x_{kij}}{\tilde{x}_{ki}}) \tag{4.2}$$

where, $\tilde{p}_i$ is the geometric mean of p_{ij}'s and $\tilde{x}_{ki}$ is the geometric mean of x_{kij}'s.

Final customer share predictions can be obtained using Equation (4.3). This transformation is referred to as the inverse-log transform (Nakanishi and Cooper, 1982). Alternatively, the researchers can use the original MCI (1974) presented above (Equation (2.3)) for obtaining share values, this procedure is called log-transform method (Nakanishi and Cooper, 1974)

$$\hat{p}_{ij} = \frac{e^{\sum_k \delta_k \log(x_{kij} / \bar{x}_{ijk})}}{\sum_j e^{\sum_k \delta_k * \log(x_{kij} / \bar{x}_{ijk})}} \tag{4.3}$$

4.2.3 The Competing Destinations (1988) Model

Fotheringham proposes a special kind of nested logit model known as competing destinations (1988) model. He suggests that consumers do not necessarily evaluate all competing stores because of time and other limitations. According to Fotheringham (1988) customers first choose a cluster of stores and then choose their main store within the selected cluster. The model can be specified as such,

$$\hat{p}_{ij'} = \frac{e^{\sum_k X_{kij} \delta_k} L(j' \in CL)}{\sum_j e^{\sum_k X_{kij} \delta_k} L(j \in CL)}$$

where $L(j \in CL)$ is likelihood of store *j* to be in the selected cluster. Likelihood can be defined by a store's similarity or dissimilarity to other stores. There are variations of proposed functions for both approaches. To give an example of the similarity approach, Borgers and Timmermans (1987) proposes the function provided below:

$$L(j' \in CL) = (\frac{1}{O-1} \sum_j d_{j'j})^{\psi} \tag{4.4}$$

Where ψ is a sensitivity parameter for similarity of distance factor (Fotheringham, 1988) Fotheringham (1988) suggests least squares approach of Nakanishi and Cooper (1974) for parameter calibration.

4.2.4 A New Spatial Interaction Model with Variable Distance Decay and Store Attractiveness Values

In this chapter our ultimate goal is to create an accurate and flexible spatial interaction model based on our findings in exploratory analysis covered in Chapter 3. In Chapter 3 we conclude that distance, store size, competition, customer share before the new store opening and store format are important variables. Our main observations are that distance effects customer share negatively. Store size and customer share before the new store opening effects the customer share positively. For store format, we can conclude that M, which is a small and low product variety store, is more affected from the distance. Also we observe that competition effects store sales negatively. However, this effect may not be significant for all stores. In our proposed model we decide to use, distance and store size as variables. Moreover, we use customer share before the new store opening for parameter calibration. We don't use the store format because a smaller store will automatically carry low variety of products. Competition is included to our model indirectly as spatial interaction models estimate customer's store share by proportioning its utility to utilities of other stores. Competition is also included to our model implicitly with the free attractiveness variable.

Motivation 1: The distance decay value and the store size

All of the models discussed previously assume that distance decay value is same for all stores included in an analysis. However, bigger stores could be less affected from distance. In Chapter 3, we observe that store 2 which is the smaller store is more affected from distance compared to store 1 and 3 in terms of individual customer share. Also, in Chapter 3 we observe that as distance increases customer spending decrease more rapidly for store 2 compared to store 3. In addition, we see that a smaller percentage of big store sales is generated in the immediate area. The Huff (1963) model tries to capture this affect by using store size as an attractiveness measure. However, a different approach is to represent this effect by adjusting the distance decay values according to store size. Orpana and Lampinen (2003) observe that distance decay parameters are smaller for big stores such as hypermarkets, but they do not build a model using this observation.

By graphing store sizes and distance decay values we can see the effect described in the previous paragraph more clearly. In order to do this, we select 10 stores in various sizes from Beşiktaş area, Istanbul. We also select the customers stored in the GIS system of Migros who are located in Beşiktaş area. Firstly, we find the customer shares in each store.

$$p_{ij} = \frac{s_{ij}}{\sum_{j \in U} s_{ij}} \tag{4.5}$$

where U is the selected area, Beşiktaş

Customer distances to the stores are already available to us. Secondly, we graph customer distances to the stores and customer shares for each store. We accomplish this by converting the distance values to discrete such as, less than 200 meters, between 200 and 400 meters. Finally we fit a line to the each graph. By doing this we obtain the distance decay values for the stores. Distance decay value is accepted the power of distance in the power function represented with the fitted line. Figure 4.1 illustrates the relationship between share and distance for a store located in Beşiktaş region.

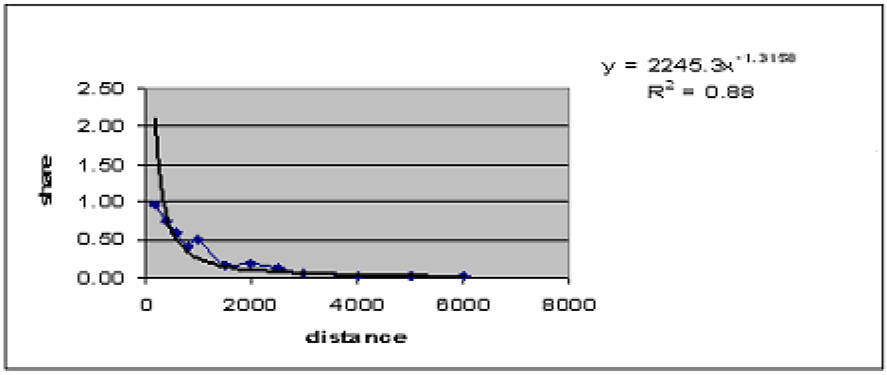


Figure 4.1: Share- distance graph for a Migros store

After finding the distance decay values of the stores via described method, we graph the distance decay values with store sizes. We obtain Figure 4.2 as a result of this process.

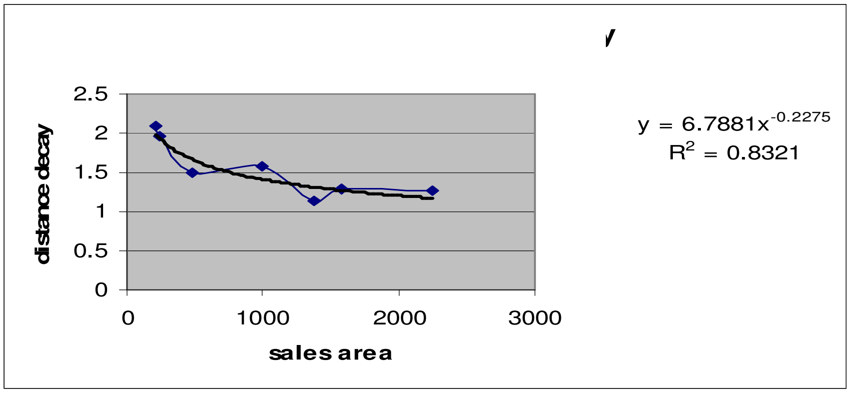


Figure 4.2: Distance decay values obtained in the previous step are graphed with sizes of the stores.

Examining Figure 4.2, where we plot the distance decay values against the store sizes we can conclude that distance decay values can be represented as a power function which decreases as store size increases. Based on our observation of Figure 4.2, we decide to represent the distance decay value as a function of store size in our model.

Motivation 2: Drezner and Drezner (2002) model

Our second motivation for a new model is Drezner and Drezner (2002) study. Drezner and Drezner (2002) replaced w_j (store size) of the Huff (1963) model with a store specific attractiveness variable. Then they inferred the real attractiveness values of the shopping centers using the shopping center sales data and the buying power index. They tried different distance decay functions: $d^2{}_{ij}$, $d^{2.5}{}_{ij}$ and $e^{1.705 d_{ij}^{0.409}}$. They then compared the calculated attractiveness scores with the survey data. According to their findings attractiveness scores obtained are parallel to the survey results, thus are interpretable. Drezner and Drezner (2002) model infers the attractiveness scores with the following model:

(M.2)

Minimize,

$$(\sum_j (m_j - \sum_i b_i p_{ij})^2 \tag{4.5}$$

Subject to,

$$p_{ij} = \frac{\dfrac{A_j}{F(d_{ij})}}{\sum_j \dfrac{A_j}{F(d_{ij})}} \tag{4.6}$$

$$A_1 = 1 \tag{4.7}$$

Where, A_j is the attractiveness value of store j and $F(d_{ij})$ is a function of distance.

Drezner an Drezner (2002) define market share of a retail facility as,

$$m'_j = \sum_i b_i p_{ij} \qquad (4.8)$$

Where, b_i is the fraction of buying power at community i to total buying power of all communities.

According to Drezner and Drezner (2002) this model can be solved analytically and either has infinitely many solutions or no solution. This is because there are n homogenous equations with n variables in this problem. By homogenous we refer to the situation, when both the numerator and denominator is multiplied with a constant the obtained results don't change. So, it is possible to obtain infinitely many solutions to the problem if there exist a solution. Variables are store attractiveness values and parameters of the distance function are specified before. Obtaining an error value of 0 in Model 2, means that we obtained the analytical solution. Constraint specified in Equation (4.7) is included to avoid infinitely many solutions.

In their study, Drezner and Drezner (2002) observe that obtained attractiveness scores are not sensitive to distance decay functions. They obtained error value of 0 for all distance decay values. So, they found the analytical solution in each case.

Drezner and Drezner (2002) use their model for exploratory research on shopping center data. We find Drezner and Drezner (2002) model valuable in terms of capturing the store specific factors in grocery retailing. Drezner and Drezner (2002) model can also be used for customer share prediction and store sales estimation instead of exploratory purposes. It should especially provide exceptional sales prediction performances as Drezner and Drezner (2002) calibrate their model by minimizing the general error. Because we are interested in customer specific share prediction performance as well as the sales prediction performance we want to add a store specific attractiveness measure to our proposed model. We can improve the Drezner and Drezner (2002) model by changing the objective function to minimize customer share prediction error and introducing a flexible sales prediction constraint to control the sales prediction performance.

The method

Motivated by the observation about the relationship between the distance decay value and the store size we can represent the distance decay value for a particular store *j* as a function of its size. The distance decay represented as a function of store size captures the store size effects. Therefore, we don't need to use store size as an attractiveness parameter. Inspired by Drezner and Drezner's (2002) research, we may still keep the attractiveness as a parameter to be calibrated to capture the factors other than distance and store size that might influence customer store choice.

This leads to the model:

$$\hat{p}_{ij} = \frac{A_j d_{ij}^{\theta_j}}{\sum_l A_l d_{il}^{\theta_l}} \tag{4.9}$$

A_j is the attractiveness of store *j* and θ_j is the distance decay parameter which is a function of store size.

The distance decay function can be represented as a power function.

$$\theta_j = cw_j^{\lambda} \tag{4.10}$$

where w_j is the store size. Equation (4.10) represents distance decay value as a power function of the store size where Motivation for Equation (4.10) is provided in Figure 4.2.

The constant c and the power of store size λ are the parameters of the distance decay function. So, the new model lets researcher to calibrate A_j, c and λ.

For direct estimation the optimization model becomes,

(M.3)

Minimize, $$MSE = \frac{1}{OM} \sum_i \sum_j (p_{ij} - \hat{p}_{ij})^2$$

Subject to, $$\hat{p}_{ij} = \frac{A_j d_{ij}^{\theta_j}}{\sum_l A_l d_{il}^{\theta_l}}$$

$$\theta_j = cw_j^{\lambda}$$

$$l \le \frac{\sum_i (s_i \hat{p}_{ij})}{\sum_i (s_i p_{ij})} \le u \tag{4.11}$$

$$A_j \ge 0 \quad \forall \quad j \tag{4.12}$$

$$A_1 = 1 \tag{4.13}$$

$$\lambda < 0 \tag{4.14}$$

where Equation (4.11) controls the general error. Constraint (4.12) guarantees that the customer share predictions are positive. Constraint (4.13) is similar to Drezner and Drezner

(2002) constraint and included to avoid infinitely many solutions. Constraint 4.14 is parallel to our findings in Chapter 3 and 4 that the store size's effect on distance decay value is negative. Model 3 reduces to a similar problem to the Drezner and Drezner (2002) model's when both l and u are equal to 1 in the sense of both models try to estimate store attractiveness values giving the real store sales. This model is different from Drezner and Drezner (2002) model in terms of distance decay value specification. More importantly, this model controls both the individual customer errors and the general error where Drezner and Drezner (2002) minimize only the general error. Also, this model is superior to Drezner and Drezner (2002) model because lowerbound and upperbound values are specified by the researcher. Hence, it provides improved flexibility and better individual error results.

We observe that direct estimation of A_j, c and λ via optimization is very difficult. When we try to solve the model with direct estimation for local optima, Excel solver can not solve the problem due to numerical instability.

To find a solution to the problem we can continue with a different approach. First we find the distance decay function parameters by assuming A_j=1 for all *j*. At this step we are excluding the Equation (4.11) or setting a wide upper bound- lower bound range to make sure the problem doesn't become infeasible. Then, in the second step we find the optimal attractiveness values given the distance decay and including Equation (4.11). Goal of this differentiation is to stabilize and simplify the error function so that we can find a result.

Then, in our two step approach, Step 1 problem will reduce to,

(M.4)

Minimize, $$MSE = \frac{1}{OM}\sum_i \sum_j (p_{ij} - \hat{p}_{ij})^2$$

Subject to, $$\hat{p}_{ij} = \frac{A_j d_{ij}^{\theta_j}}{\sum_l A_l d_{il}^{\theta_l}}$$

$$\theta_j = cw_j^{\lambda}$$

$$A_j = 1 \quad \forall \quad j \tag{4.15}$$

$$\lambda < 0$$

First step optimization problem is different from the previous problem in terms of Equation (4.15). Assuming all stores are equally attractive, we are trying to estimate *c* and *λ*.

We should examine the Step 1 model for quasiconvexity to make sure that the obtained solution is the global optimal solution. According to Definition C.1 and the counter-example provided in Appendix C, in Table C.2 step 1 problem is not quasiconvex or convex. Although the function is not quasiconvex or convex, we observe on our example that when we run the optimization model with different starting points we converge to same solutions. The first step optimization problem (M.4) is more stable compared to optimization model 3 (M.3).

Once we obtain the distance decay values, in step 2 the problem will become,

(M.5)

Minimize
$$MSE = \frac{1}{OM}\sum_i \sum_j \ (p_{ij} - \hat{p}_{ij})^2$$

Subject to
$$\hat{p}_{ij} = \frac{A_j d_{ij}^{\theta_j}}{\sum_l A_l d_{il}^{\theta_l}}$$

$$l \le \frac{\sum_i \ (s_i \ \hat{p}_{ij})}{\sum_i (s_i \ p_{ij})} \le u$$

$$A_1 = 1, \ \ A_j \ge 0 \ \ \forall \ \ j$$

where l is a lower bound and u is an upper bound value and θ_j is constant at this point. In this model distance decay value is a constant and variables are the store attractiveness values. The sales prediction constraint aims to control general sales prediction error while the objective function controls the individual customer errors. In this step the problem reduces to a linear fractional problem. Sales prediction constraint Equation (4.11), Equation (4.12) and (4.13) are also included to this model.

4.3 Comparison of Spatial Interaction Model Performances

4.3.1 Experiment and Data

In this section we want to evaluate the models discussed in the previous section which include: Huff (1963), MCI (1974), Competing destinations (1988) and Variable distance decay model with store attractiveness model proposed by us. Goal of the evaluation is to compare the performances of the new model and the conventional spatial interaction models. This experiment doesn't include a new store opening. For this analysis, we are not using catchment area or immediate area concepts as we don't have a new store for sales prediction. There are two evaluation criteria for the discussed models. First one is customer share prediction error for the test set which is calculated as,

$$MSE = \frac{1}{OM} \sum_i \sum_j (p_{ij} - \hat{p}_{ij})^2$$

The second one is the sales prediction ratio and it is calculated as,

$$\frac{\sum_i \sum_{j \in U} s_{ij} \hat{p}_{ij}}{\sum_i \sum_{j \in U} s_i p_{ij}} \tag{4.16}$$

Where, s_{ij} is the grocery spending of customer *i* in store *j*. *U* here stands for the area we selected. $\hat{p}_{ij}$ is the predicted share of customer *i* in store *j*, p_{ij} is the real share of customer *i* in store *j*. Sales prediction ratio in this example is used for only performance evaluation purposes. When sales prediction ratio is used in a context with a new store opening say at time *t*, the researcher should use the predicted grocery spending, s_i for the time he tries to predict revenue for example, time *t+1*. This is because, the only information available to the researcher is the spending up to time *t-1* and assuming customer spending will not change at time *t+1* is unrealistic. Our experiment doesn't include time dimension. In our experiment we know the s_i values for our training set and we assume to know the s_i values for the test set as we want to eliminate the error caused by this estimation.

For this experiment we choose Beşiktaş area in Istanbul. 10 stores in various sizes located in the area are included to the analysis. Data is very heterogonous in terms of the stores used in this analysis and the spatial distribution of customers. Beşiktaş neighborhood in the city of Istanbul is selected for this analysis that includes ten stores that range in size between 200 to 2500 square meters. All stores are of the same brand, so service quality and prices are homogenous. However, product assortments are quite different as a result of the difference in store size. Moreover, one of the big stores is located in a shopping mall.

Loyalty card customers of Migros Company who are stored in the GIS system of Migros are selected for this analysis. The number of the selected customers is 6942. Of these customers 4615 are in the training set and 2327 are in the test set. The data regarding individual customer spending in selected Migros stores and total Migros spending between the years 2003-2007 are provided to us by Migros. Customers' total grocery budgets are not available, since data includes only Migros spending. Real shares of spending of the customers are calculated on the total customer spending through 2003 and 2007.

4.3.2 Estimated parameters and details

1. **Huff (1963) model**

Huff 1: Original Huff (1963) model: We use the Huff (1963) model specified in Equation 2.2. Distance sensitivity parameter is calibrated via optimization. The Huff (1963) Model error function is not convex but when we graph the function we can see the minimum MSE point for a given range of distance decay values**.** We estimated the optimal β_D as 1.7. We can confirm this minimum point when the distance decay value ranges between -10 and 10. According to our literature survey this value should be around 2. Finding a better solution outside our specified range is highly unlikely. Therefore we are using β_D = 1.7 for our share estimation process.

Huff 2: Huff (1963) model without calibration. The distance decay value is accepted as 2.

2. MCI (1973) model

MCI 1: MCI (1973) model with all variables. We use MCI (1973) model specified in Equation (2.3). We include distance, store size, population density, workplace density, Migros competition and other competition as variables. Distance, store size, competition are identified as important variables in Chapter 3. In this chapter we also want to include area related factors: population and workplace density.

- **Distance** represents, customers' driving distance to a particular store.
- **Population density** is given by the number of people living in the neighbourhood where the selected store is located divided by the area of the neighbourhood.
- **Workplace density** is the percentage of office buildings within a circle of 1500 meters radius around the store.
- **Migros competition** is the number of Migros stores located in the circle with radius of 3000 meters originating from the particular store.
- **Other competition** is the number of competitor stores owned by other grocery store chains located in the circle with radius of 3000 meters originating from the particular store.

We estimated the parameters as described in Section 4.4**.** Estimated parameters (δ_k) are as follows: -1.75, 1.77, -0.29, -4.31, 0.25, 0.35 for distance, store size, population density and workplace density, Migros competition, other competition respectively.

MCI 2: MCI (1973) model with limited variables In Chapter 3 we observed that Huff (1963) model is quite robust and give better solutions than data mining models including more variables. Store size and distance are variables of the Huff (1963) model. So, we use only distance and store size as variables. δ_k values are, -1.84, 1.43 for distance and store size.

3. Competing destinations model

Fotheringham (1988) recommended usage of Nakanishi and Cooper (1974) transformation for estimation of variables. We interpreted this transformation of Equation (2.7) as provided in Equation (4.16) and estimated the coefficients with OLS regression.

$$\log(\frac{p_{ij}}{\tilde{p}_{ij}}) = \delta_k x_{kij} + \psi \log(\frac{d_{jj'}}{\tilde{d}_{jj}}) \qquad (4.16)$$

CD 1: Competing destinations (1988) model with all variables We use distance, store size, population density, workplace density, Migros competition and other competition as variables. δ_k are estimated as, -0.00046 for distance, 0.000811 for other competition, 0.0243 for Migros competition, 2.756 for workplace density, -0.028 for population density, 0.000649 for store size. Likelihood sensitivity parameter ψ is estimated as 1.38.

CD 2: Competing destinations (1988) model with limited variables We use distance and store size as variables and obtain the coefficients as ψ: 0.87 and δ_k are -0.00035 and 0.00094 for distance and store size.

4. Variable distance decay model with store attractiveness values

VD 1: Variable distance decay model with sales prediction constraint

In the first step problem we obtain the following parameters: c: -3.05, λ: -0.08

We confirm that minimum error value is obtained at this point by graphing the error values**.** We also run the model multiple times from random starting points and observe that the error value converges to the same point.

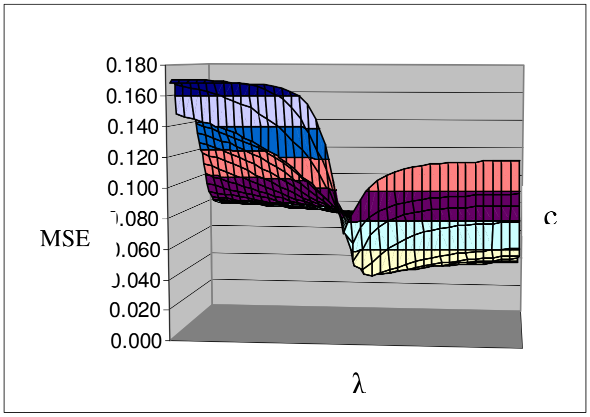
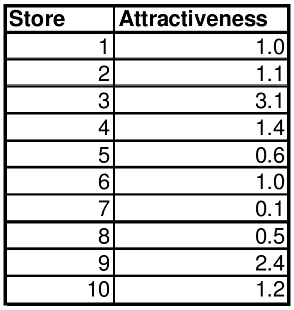


Figure 4.3: Error function of variable distance decay model

The error function is undefined for large values of the distance decay value. Because distance values are high, when the distance decay value gets larger, the error values becomes undefined due to division with 0 while proportioning utilities of stores. This property narrows down the range of solutions that we are searching for the global optima. Thus, it becomes easier to graph the error function of the model versus parameters of the distance decay value. In Step 2, sales prediction constraint (Equation (4.11)) is set as, Lower bound: 1, Upper bound: 1, which amounts to requiring that the model is calibrated to produce the actual store sales. Following attractiveness values provided in Table 4.1 are obtained:



Store	Attractiveness
1	1.0
2	1.1
3	3.1
4	1.4
5	0.6
6	1.0
7	0.1
8	0.5
9	2.4
10	1.2

Table 4.1: Attractiveness values for variable distance decay model 1

Parameters are estimated with two-step optimization procedure which is described in the previous section.

VD 2: Variable distance decay model without the sales prediction constraint.

The sales prediction constraint defined by Equation (4.11) controls the general sales prediction performance. However, there is a tradeoff between the general sales prediction and individual customer share prediction performances. Without the sales prediction constraint better individual error values can be obtained. In order to see the tradeoff between general sales prediction and the individual share prediction performances we run the alternative model without the sales prediction constraint.

In step 1 we find the following c and λ values: c: -3.05, λ: -0.08

Store	Attractiveness
1	0.66
2	0.81
3	0.06
4	0.46
5	1.57
6	0.46
7	1.00
8	0.64
9	1.69
10	0.68

Table 4.2: Attractiveness values obtained without the sales prediction constraint

In step 2 we find the attractiveness values provided in Table 4.2.

4.3.3 Error results

Table 4.3 displays the average mean squared error results for test set and the training set. Parameters are calibrated on the training set which includes 4615 customers. First row of the Table 4.3 provides the training set MSE results. The second row of the table provides the test set MSE results. The test set consists of 2327 customers.

	Huff 1	Huff 2	MCI 1	MCI 2	CD 1	CD 2	VD 1	VD 2
Models	1.1	1.2	2.1	2.2	3.1	3.2	4.1	4.2
MSE train	0.0412	0.0415	0.0442	0.0416	0.0517	0.0553	0.0413	0.0400
MSE test	0.0413	0.0415	0.0466	0.0412	0.0524	0.0559	0.0406	0.0392

Table 4.3: Error results for models

According to paired t-test statistics both Model 4.1’s and 4.2’s error values are significantly lower than models 1.1, 1.2, 2.1, 2.2, 3.1 and 3.2 at alpha=0.05 level. T-test statistics are provided in the Appendix B in more detail. We can conclude that the variable distance decay model is superior to other models in terms of predicting individual store shares. Moreover, examining Table 4.3 we observe that both variable distance decay models have better error values in test set than in training set. Thus they have generalizing ability. It is interesting to see that Model 2.1 which includes more variables than Model 2.2 has a worse error value. We may think that additional variables included in this analysis are irrelevant to our dependent variable. We may also conclude that interaction between the variables cause misspecifications in the model. Coefficients of the model estimated for competition are positive where we expect them to be negative. The correlation results between the attractiveness values and the competition variables are parallel to this expectation.

Table 4.4 provides us the sales prediction ratios for individual stores for different models on the test set.

	Huff 1	Huff 2	MCI 1	MCI 2	CD 1	CD 2	VD 1	VD 2
Stores	**1.1**	**1.2**	**2.1**	**2.2**	**3.1**	**3.2**	**4.1**	**4.2**
1	1.10	1.10	1.26	1.11	0.9	0.93	0.97	1.13
2	0.87	1.01	0.47	0.59	1.3	0.78	0.80	0.85
3	1.22	1.21	1.53	1.17	0.5	0.55	0.94	0.95
4	0.78	0.80	0.49	0.63	0.8	0.51	1.07	1.04
5	1.02	0.87	0.14	0.71	1.5	1.19	0.97	0.55
6	1.08	0.97	0.76	1.13	1.8	2.08	0.95	1.39
7	1.13	1.16	1.36	1.18	0.8	0.94	1.09	0.97
8	0.41	0.46	0.24	0.21	0.9	0.63	0.82	0.66
9	0.25	0.27	0.18	0.71	1.0	0.64	1.02	0.79
10	14.77	13.15	3.20	6.920	27.7	24.93	0.87	1.08
Average Sales Prediction Ratio excluding store 10	0.87	0.87	0.71	0.83	1.05	0.92	0.96	0.93
Average Sales Prediction Ratio	2.26	2.10	0.96	1.44	3.71	3.32	0.95	0.94

Table 4.4: Individual store sales prediction ratios for different models

Ideally, sales prediction ratio should be equal to 1. When the sales prediction ratio is equal to 1 for a store it means that the model predicted the sales of that store perfectly. When we examine Table 4.4 we see that Model 4.1 has an exceptional performance due to the explicit constraint defined in Equation (4.11). Although Model 4.2 doesn't have a sales prediction constraint it performs reasonably well in comparison to other models. Among other models, on average MCI (1973) model performs well. We can conclude that Models 4.1 and 4.2 are quite robust. The percentage of sales predicted for store 10 is quite high for each model except for Models 4.1 and 4.2. The reason for this is the fact that store 10 is close to the corner of the selected region and it is possibly not in the choice set of the customers selected for this study. It is important to note that because we estimate individual store attractiveness score for each store, in Models 4.1 or 4.2 we can capture store 10's extreme situation. Average sales prediction ratio without store 10 is also provided in Table 4.4. We observe that, excluding store 10 competing destinations model provide accurate sales prediction results.

4.3.4 Attractiveness values

In this section we will try to interpret Model 4.1's attractiveness values which are obtained with the sales prediction constraint.

	Attractiveness
Competitor overall	-0.82
Competitor discount	-0.81
Migros discount	-0.68
Migros regular	-0.61
Population density	-0.52
Workplace density	-0.17

Table 4.5: Correlation results of attractiveness values with selected variables.

Table 4.5 displays the correlation of attractiveness values with selected variables. Population density is the number of people living in the neighbourhood of the store divided by neighbourhood area. Workplace density is the percentage of workplaces among Migros customers around the store. There are four variables for competition: competitor discount, regular and Migros discount and regular. Discount stands for discount stores which have lower prices and smaller store sizes generally. Regular stores have bigger store size and prices of the goods are higher compared to discount stores. Competition is measured as such: stores located in a circle originating from the store with radius of 3000 are selected. Sum of the store sizes of the selected stores are included in the analysis as the competition values. According to Table 4.5 attractiveness values are strongly negatively correlated with competition. The attractiveness values appear to be negatively correlated with population density too. We observed that population density is strongly negatively correlated with competition. Thus, negative correlation between population density and attractiveness values may be the result of the relationship between population density and competition.

4.3.5 Discussion

To summarize the general results, although objective functions of Models 4.1and 4.2 are not convex, we obtain consistent estimates. It is important to note that, found solutions are not guaranteed to be the global optimal solution. However, Model 4.1's and 4.2's error results are significantly lower than other models error results. We can say that models with distance decay values as a function of store size provide better error results. We can also conclude that, attractiveness values left to be calibrated provide better store specific sales prediction results. One can use Model 4.1 to predict revenue of a potential store. One obstacle in this situation is to predict attractiveness values. Because attractiveness values are relative numbers and estimated in a small area generally users will not have enough data points to estimate attractiveness values with traditional methods. However, in Section 4.3.4 we observed that attractiveness values are negatively correlated with competition and one can use this insight to determine an attractiveness value for the potential store. In this chapter a situation with a new store opening is not considered. Chapter 5 will thoroughly cover the revenue prediction of a new store and propose a decision support system for retail store.

Chapter 5

PREDICTING STORE REVENUE USING SPATIAL INTERACTION MODELS

5.1 Overview

We dedicate Chapter 4 to the spatial interaction models and after our evaluations we conclude that the newly proposed variable distance decay model with store attractiveness values proposed by us performs better than conventional spatial interaction models. The objective of this chapter is to predict a potential grocery retail store's sales using spatial interaction models especially the newly proposed one. Output of a spatial interaction model is the customer's share of spending in that particular store. Moreover, output of the spatial interaction model can also be used for store revenue prediction. Store revenue prediction is relatively simple if consumer panel data of grocery spending and consumer location information is available. Store revenue prediction in this case is covered in Section 5.2. In most cases panel data is not available to decision makers. We propose a different methodology for retail store revenue prediction using loyalty card data and GIS in Section 5.3.1 and present a numerical example in Section 5.3.2. It is important to note that if the decision maker is using a spatial interaction model without calibrating the parameters, there is no need to differentiate between panel data available and not available case. Panel data is needed for parameter calibration in spatial interaction models. Finally, in Section 5.3.3 we provide a template for a spatial decision support system for store revenue prediction using the proposed methodology in Section 5.3.2.

5.2 Customer panel data available

When the customer panel data on customer grocery spending and location, and competitor information such as store size are available, store revenue prediction is a relatively easy problem. The sum of customers' share predictions obtained from spatial interaction models multiplied by their grocery budget gives us the total revenue of a store. However, enumerating all customers is not possible. Not all customers may have loyalty cards. Most researchers cluster the customers and continue the research with new instances which are clusters of customers. These clusters are often naturally formed, like zip codes. In some countries like US, average income, spending data and population figures are available at the zip code level. In case of availability of the panel data, store sales can be represented as a function of average grocery spending in the residential area, number of people living in the residential area and the residential area's share of grocery spending in that store. For example, Orpana and Lampinen (2002) represent store sales as,

$$SS_j = \sum_i p_{ij} s_i$$

Where in the above equation SS_j is sales of store *j*, p_{ij} residential area *i*'s share in store *j*. s_i is the residential area *i*'s grocery budget.

A sophisticated version of the method with lifestyle segments is covered in Nakaya et al.(2007) study. Nakaya et al. (2007) describe a store's sales as,

$$SS_j = \sum_h \sum_i s_i^h H_i^h p_{ij}^h \qquad (5.1)$$

Where SS_j is the sales of store *j, i* in this model stands for a small residential area like zip code, H_i^h is the predicted number of people from lifestyle group *h* living in residential area *i*, s_i^h is the expected grocery spending in residential area *i* for *h* lifestyle group and p_{ij}^h is the share of spending generated by the spatial interaction model for store *j*. Nakaya et al (2007) use a multinomial logit framework with store attractiveness values to generate p_{ij}^h values. Implicit assumption of Nakaya et al. (2007) and Orpana and Lampinen (2002) models is, obtained clusters consist of customers with homogenous grocery spending. This assumption is reasonable for only Nakaya et al. (2007) given they use life style segments. Nakaya et al. (2007) use survey data combined with household expenditure data which is created with micro-simulation.

5.3 Panel data not available

Model 5.1 is a general method and requires panel data to work. Panel data requested for the methodology described in Section 5.2 is rarely publicly available and often costly to obtain. In addition, in the previous section both models use residential areas instead of individual customers. Often, zip code level detailed data is not available. However, most companies have loyalty cards which hold individual customer spending information at store and customer level as well as the customer address data. In more detail, most companies have company owned store spending of customers instead of the general grocery expenditure and the panel data. Thus, a method using the information on hand to predict sales of a potential store would be valuable for grocery companies. The method we propose in the next section requires availability of customer distances to the stores, customer spending in existing company owned stores and existing company owned store sales as well as competition related data. Customer address data can be used to calculate customer distances to the stores. Competition related data such as number of competitor stores in the area can be easily obtained through simple observation.

5.3.1 Methodology

Suppose that a grocery retail chain plans to open a new store. Also, suppose that zip code level customer spending and panel data are not available to the company. In this section we propose a new method of grocery store sales prediction using loyalty card and GIS data. In order to obtain the potential store sales value using Equation (2.1) or Equation (5.1) we have to enumerate all customers. However, enumeration is not possible for three reasons. First reason is the new store attracts new customers who didn't shop from our retail chain in the past. The second and third reasons are not all customers hold loyalty card data and not all customers are stored in the GIS.

Instead of enumerating customers, the method we propose in this section uses the immediate area concept introduced in Chapter 3. Immediate area concept is similar to catchment area concept which is also introduced in Chapter 3 but it limits the selected area to a close proximity area. The idea is to predict the store sales in the immediate area and then generalize to the total sales value. We use the immediate area concept in order to limit the customers and competition included in the analysis. Limiting the customers to the immediate area is advantageous because as distance of the customers to the stores increase randomness in the store choice behavior also increase. For example, a customer can visit a distant store while passing from that location by chance. Spatial interaction models require competition data to work as they predict customers' share of spending in a store by proportioning store utility to sum of utilities of all stores in the customer's choice set. Moreover, without the immediate area concept competition data needed for the analysis would increase dramatically almost capturing all stores in the city. In order to limit the competition included in the analysis we introduce a vicinity concept dependent on the immediate area radius. We include the competition located in the vicinity of the new store to the analysis. Vicinity radius should be double of the immediate area radius. The idea is to capture the competitor stores at which are least as close to the customers as the new store.

Although using the immediate area and the vicinity concepts simplifies the problem, there are still challenges due to time related factors and absence of panel data. Suppose at the time *t-1* the decision maker plans to open a new store at time *t*. The decision maker knows the customer spending in his company's stores at time *t-1*. He considers predicting the sales of the candidate store at *t+1*. The decision maker is interested in the store sales at time *t+1* because of the time required for stabilization of the grocery retail store sales. When we examine store openings from the past data we observe that stores' tend to under-perform in the first year of their opening. Therefore, we have to adjust current *(t-1)* spending of customers for time *t+1*. The second challenge is the fact that the decision maker doesn't have the panel data so he doesn't know the customers' spending in the competitor stores. Hence, if he uses the Equation (2.1) or Equation (5.1) for store sales prediction, he ignores the sales captured from the competitor stores owned by other retail chains. For this, reason in the new method we have to have a new demand adjustment parameter.

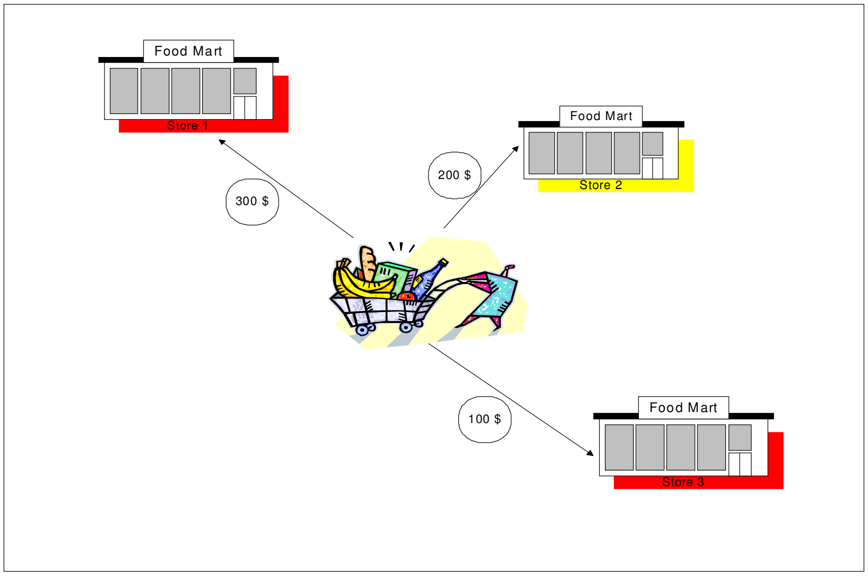


Figure 5.1: Before the new store opening (t-1)

For example, in Figure 5.1 store 1 and store 3 are company owned stores, where store 2 is the competitor store. The decision maker can only see customer spending in store 1 and store 3.

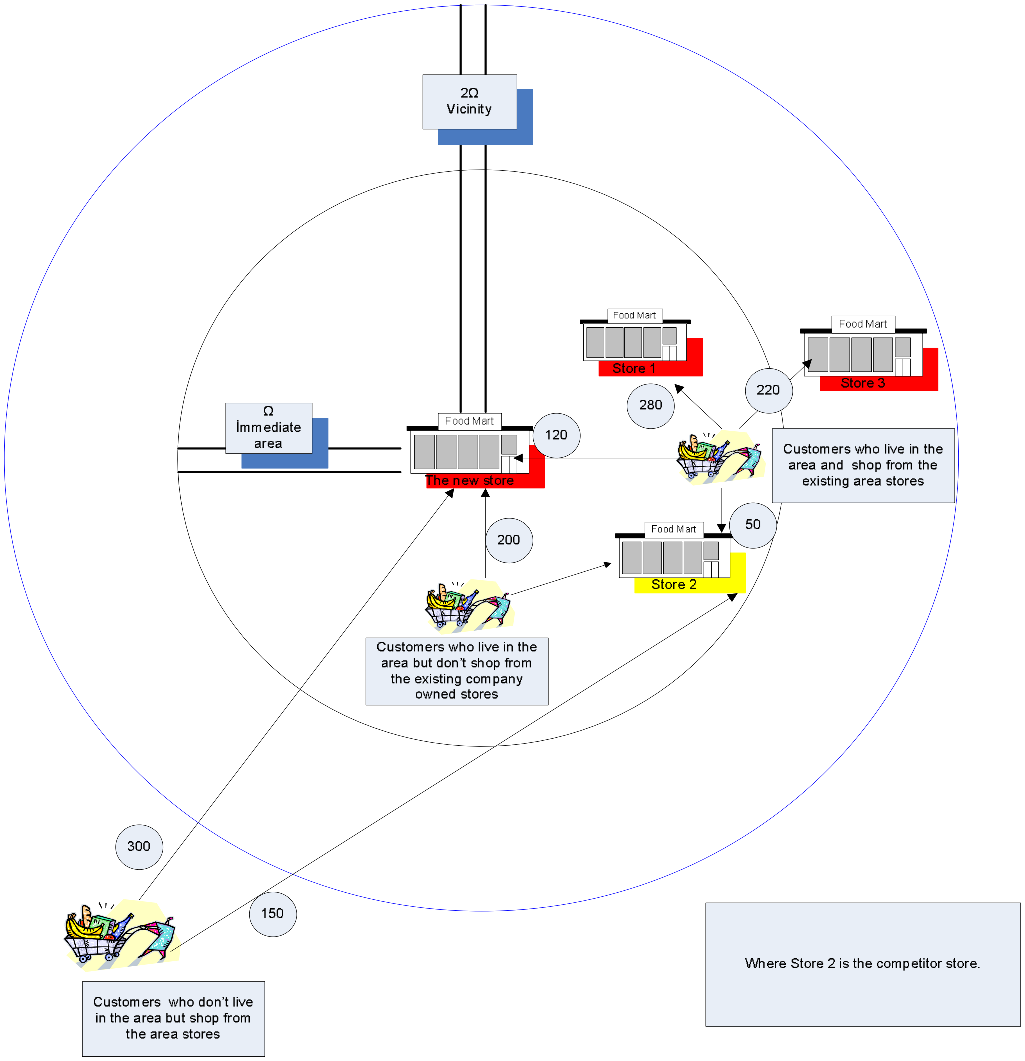

2Ω
Vicinity
Ω
İmmediate
area
Food Mart
Store 1
Food Mart
Store 3
Food Mart
The new store
Food Mart
Store 2
280
220
120
50
200
300
150
Customers who live in the area and shop from the existing area stores
Customers who live in the area but don't shop from the existing company owned stores
Customers who don't live in the area but shop from the area stores
Where Store 2 is the competitor store.

Figure 5.2: After the new store opens. (t+1)

Figure 5.2 illustrates the dynamics of a new store opening. The center of the two concentric circles is the new store location, where the inner circle defines the immediate area and the outer circle defines the vicinity. Immediate area and vicinity concepts are used for practical concerns.

When a new store opens its sales come from mainly three set of customers:

1) **Customers who live in the immediate area and shop from the company owned existing stores.** These customers may also shop from the competitor stores. The customers shift their grocery spending from existing company owned and competitor stores to the new store.

2) **Customers who live in the immediate area and don't shop from the company owned existing stores.** Previously, these customers only shopped from the competitor stores. When the new store opens, they shift their spending from the competitor store to new company owned store.

3) **Customers who live outside the immediate area but shop from the immediate area**.

Sales, captured from company owned stores in the vicinity which is also known as cannibalization amount can be predicted with spatial interaction models as customer spending data for company owned stores is available. The cannibalization amount in the vicinity is the part of sales coming from the first set of customers. This amount is equal to $\sum_i (\sum_{j \in R} s_{ij} p_{ij}) Q_t Q_{t+1}$ where s_i is the grocery spending in company owned stores at *t-1*. *R* is the vicinity area. p_{ij} is the share of customer spending and estimated by spatial interaction models. If we use $\sum_i (\sum_{j \in R} s_{ij} p_{ij})$ as the vicinity cannibalization amount we would assume that customer grocery spending doesn’t increase through time. However, in Chapter 3 macro analysis we observe in aggregate level grocery sales increase with time. For this reason, we multiply the term $\sum_i(\sum_{j\in R} s_{ij}p_{ij})$ with sales growth parameters for *t* and *t+1*.

In order to obtain the sales generated in the immediate area we should adjust cannibalization amount with sales captured from the competitor stores, represented by the factor N_j. This is the sales coming from the second set of customers and first set of customers. We represent this factor by adding a new demand increase parameter. Sales generated in the immediate area becomes, $\sum_i(\sum_{j\in R} s_{ji}p_{ij})Q_tQ_{t+1}N_j$. The N_j value is not present in Equation (5.1) because Equation (5.1) includes competitor stores to the spatial interaction model as panel data is available.

Finally, sales coming from the immediate area should be adjusted with the percentage of sales generated in the immediate area, G_j, to obtain the total sales. This final adjustment is included to capture the sales of third set of customers.

Then the store sales become,

$$SS_j = \frac{\sum_i(\sum_{j\in R} s_{ij}p_{ij})Q_tQ_{t+1}N_j}{G_j} \qquad (5.2)$$

where $\sum_i(\sum_{j\in R} s_{ij}p_{ij})$ is the sales coming from the immediate or the vicinity cannibalization amount in terms of customer spending in *t-1*, Q_t and Q_{t+1} represent the sales growth due to economic conditions for *t-1* to *t* and *t* to *t+1*, N_j is the new demand increase parameter and G_j is the percentage of sales generated in the area. Implicit assumptions of our model are,

- When a new store opens, it generates extra demand from the customers in the immediate area beyond what is observed in the chain stores in the area. This representation assumes that this extra demand is proportional to the demand of the customers in the immediate area that the new store cannibalizes from other company owned immediate area stores observable in the data. The sales growth due to economic conditions such as spending power increase is represented with Q_t and Q_{t+1}.

- The new demand increase parameter is represented with N_j. We need this new demand generation parameter because we are only including the existing company owned stores in the area to the analysis. Therefore, if we don't include this parameter we would ignore the sales captured from non-company owned competitor stores.

- Finally we assume that store sales obtained from the immediate area customers is proportional to general store sales. This assumption is represented with G_j, percentage of sales generated in the area.

Classification of store revenue prediction models are provided in Chapter 2. Our proposed model I uses spatial interaction model as a basis for store revenue prediction. We use analogue approach in parameter specification for G_j percentage of sales generated in the area and A_j the store attractiveness value of the new store.

5.3.2 A numerical example

In this section we test the proposed method in the previous section on a numerical for store revenue prediction. For model performance testing we choose store 3 which is a Migros store and located in Bakırköy, Istanbul. Stores 1, 2, 4, 5 and 6 are the competitor Migros stores which are the existing stores located in the vicinity of store 3. Radius of the vicinity is defined as 3000 meters and immediate area is defined as 1500 meters for this example for practical purposes. Store 3 opened in time *t* . The goal of this experiment is to estimate Store 3 's sales in *t+1* using the data in *t-1*. Details of the stores are provided in the table below. All stores except store 3, existed before time *t*.

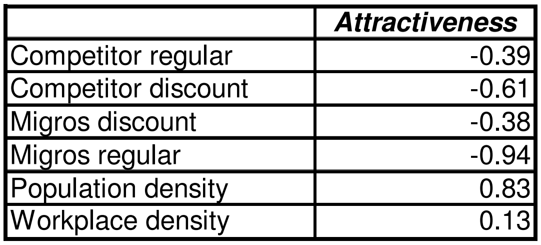
Store	Store Size	Attractiveness	Format
1	360	80.94	M
2	3000	0.63	MMM
3	2340	N/A	MMM
4	3485	1.00	MMM
5	4000	0.45	MMM
6	2603	0.47	MMM

Table 5.1: Format and store size of the stores included in the analysis

The spatial interaction model chosen for this analysis is model representing the distance decay function as a power function which is proposed and covered thoroughly in Chapter 4. We calibrate the model using the customer spending and distance data in *t-1*. We use the sales prediction constraint in calibration process and set lowerbound value as 0.95 and upperbound value as 1.05. Table 5.1 reports the calculated attractiveness values of the stores before Store 3's opening. We estimate the c and λ parameter as, 6.63 and -0.14 respectively.

Estimation of the parameters:

In order to estimate store 3 sales we need the attractiveness value of store 3. Predicting store 3 attractiveness is not very simple because there are few data points. Therefore, using different scenarios for store 3 attractiveness may be the best solution to problem.



	Attractiveness
Competitor regular	-0.39
Competitor discount	-0.61
Migros discount	-0.38
Migros regular	-0.94
Population density	0.83
Workplace density	0.13

Table 5.2: Correlation of attractiveness values with selected variables.

According to Table 5.2 attractiveness values are strongly negatively correlated with competition. Competition values are defined as the sum of sizes of competitor stores located in the vicinity of the store. Population density is defined as population of the neighborhood which the store is located in, divided by the area of the neighborhood. Workplace density is percentage of workplaces among all customers located in the immediate area. It is important to note that in this example the correlation between the population density and attractiveness values is positive. This finding is intuitive but conflicts with the finding in Chapter 4 attractiveness values.

In order to continue our analysis we have to specify an attractiveness value for the new store. Competition values of the Stores 3 and 6 are close to each other. Based on this information we may assume that attractiveness of these two stores are also similar. However, to be on the safe side for store 3 attractiveness we will try the following attractiveness values: 0.38, 0.64, 0.89. Mean attractiveness value for existing stores is 0.64 and standard deviation is 0.25. Attractiveness values to be tested are selected using mean and the standard deviation. For example, 0.38 is a standard deviation lower than the mean.

For estimation of the $\boldsymbol{Q_t}$ parameter we propose to use the predicted same store sales, by applying exponential smoothing (Holts method) on all similar same store sales time series data, and taking the ratio of the forecasted sales to current sales. Store sales stabilize approximately one years after the store opening. One should be careful to include only stores with stabilized sales to the analysis. Q_t is defined as,

$$Q_t = \frac{\sum_j SS_{jt}}{\sum_j SS_{jt-1}} \tag{5.3}$$

where SS_{jy} is the store sales at time t for store j. The smoothing parameters η_1 and η_2 can be estimated by examining the MAE values on historical forecasts.

Where, $Q_t > 1$ if sales growth occurred at year t

$0 \leq Q_t \leq 1$ is sales growth didn't occur at year t

$$\hat{Q}_t = L_{t-1} + T_{t-1} \tag{5.4}$$

Where L_t is the level and T_t is the trend value. Level and the trend value is estimated as:

$$L_t = \eta_1 Q_t + (1-\eta_1)L_{t-1} \tag{5.5}$$

$$T_t = \eta_2 (Q_t - Q_{t-1}) + (1-\eta_2)T_{t-1} \tag{5.6}$$

In our example, Q_t's are estimated as follows: $Q_{2003} = 1.14$ and $Q_{2004} = 1.12$. We use 0.2 and 0.1 as η_1 .and η_2 parameters.

N_j is the parameter for new demand generation in the immediate area due to new store opening. We propose to use regression analysis to model the new demand generation in previous store openings. The explanatory variables in this regression are the percent increase in the total store area in the vicinity due to the new store opening and the sales per area before the store opening, representing the increase in retail area, and the sales potential. We identified 18 store opening situations where vicinity stores were available in the last 10 years and fit the regression, and used its equation to estimate the N_j

value for store 3.

$$N_j : \frac{\mathrm{V}_{\mathrm{t+1}}}{\mathrm{V}_{\mathrm{t-1}}} \tag{5.7}$$

Where $\mathrm{V}_{\mathrm{t+1}}$ and $\mathrm{V}_{\mathrm{t-1}}$ are adjusted vicinity sales after and before the store opening. Vicinity sales values are calculated from general store sales data.

$$\mathrm{V}_{\mathrm{t-1}} = \frac{\sum_{j\in R} SS_{j_{t-1}}}{\sum_{j} SS_{j_{t-1}}} \tag{5.8}$$

where R is vicinity area and SS_{jt} is the store sales.

$$\mathrm{V}_{\mathrm{t+1}} = \frac{\sum_{j\in R} SS_{j_{t+1}}}{\sum_{j} SS_{j_{t+1}}} \tag{5.9}$$

Both $\mathrm{V}_{\mathrm{t+1}}$ and $\mathrm{V}_{\mathrm{t-1}}$ are adjusted with sum of same store sales to eliminate time related effects. We should also say that sum of same store sales doesn’t include any new store openings.

We propose regression for estimation of new demand parameters. Regression equation can be estimated on past data, examining past store openings in different regions.

Regression equation for estimation of new demand parameter is provided in Equation (5.10).

$$\hat{N}_j = \vartheta_1 K_1 + \vartheta_2 K_2 \tag{5.10}$$

Independent variables for predicting N_j are K_1 and K_2. Alternatively researchers can use the independent variables by taking their logarithms depending on the R-square values.
These variables are defined in Equation (5.11) and (5.12) respectively. w_j is the size of the store *j*.

$$K_1 = \frac{\sum_{j \in R} w_{j_{t+1}}}{\sum_{j \in R} w_{j_{t-1}}} \tag{5.11}$$

where K_1 is the store size increase in the vicinity due to new store opening.

$$K_2 = \frac{V_{j_{t-1}}}{\sum_{j \in R} w_{j_{t-1}}} \tag{5.12}$$

where K_2 captures the area related effects such as customer grocery spending intensity.

We estimate the equation for N_j as,

$$N_j = 0.33 K_1 \text{ -}0.24 \log K_2 \tag{5.13}$$

N_j is estimated on a sample of 18 stores. R-square value for the above equation is 0.91. *N* value for store 3 is estimated as 1.53 with the above equation. Different combinations of independent variables (K_1, K_2) are used for estimation and the combination with K_1 and log(K_2) is selected because they provide a better R-square value.

G_j value for store 3 is estimated subjectively with analogue in this example. G_j values for existing stores are basically the ratio of their sales coming from the immediate area and their total sales.

Store	G
1	0.05
2	0.04
3	N/A
4	0.21
5	0.07
6	0.24

Table 5.3: G_j values for existing stores.

Examining G_j values in Table 5.3 we can say that, area generalization parameter for store 3 should be close to stores 6 and 4's. These two stores area closer to the center of the area selected. Moreover, as store 3 is smaller than 4 and 6 we expect a higher G_j value. Our expectation is parallel to the finding in Chapter 3. Bigger percentage of sales in generated in the immediate area for smaller stores. So, we use 0.25 for store 3 area generalization parameter. We use 1500 meters for immediate area radius. However, researchers can use larger radius values to decrease the sales prediction error caused by the area generalization parameter.

Sales prediction results:

Store	ABTC=0.38	ABTC=0.64	ABTC= 0.89	Average
1	0.63	0.60	0.57	0.60
2	0.62	0.60	0.58	0.60
3	0.96	1.31	1.59	1.29
4	0.73	0.68	0.64	0.68
5	0.72	0.67	0.63	0.67
6	0.73	0.70	0.67	0.70

Table 5.4: Model results for different attractiveness values

Table 5.4 shows the ratio of predicted overall sales to real overall sales for different attractiveness values of store 3. According to the results, predicted sales highly depend on attractiveness value of the potential store. Based on our observation about competition values of store 3 and store 6 we would predict attractiveness value of the store 3 to be in the first half of the table which is highlighted. According to the Table 5.4 attractiveness value of store 3 should be between 0.38 and 0.64.

	ABTC=0.38	ABTC=0.64	ABTC= 0.89	Real
1	0.07	0.06	0.06	0.07
2	0.32	0.32	0.31	0.28
3	0.09	0.12	0.15	0.06
4	0.23	0.22	0.21	0.25
5	0.19	0.18	0.17	0.19
6	0.10	0.10	0.10	0.14

Table 5.5: Predicted market shares of the stores with different values of store 3 attractiveness

	Huff	VDD
Deviation from real sales	0.25	0.12

Table 5.6: Comparison of prediction results of Huff and the Variable Distance Decay model

Table 5.6 provides the prediction results of the Huff (1963) model where distance decay value is 1.92 (optimal distance decay). When we compare the Huff (1963) model results with the alternative model's results where attractiveness value is equal to 0.6375 we obtain better MSE results for the alternative model. 0.6375 level is the midpoint of the Table 5.4. MSE value is deviation of predicted sales from the real sales in terms sales prediction ratio. It is calculated as,

$$MSE = \left(\frac{\frac{\hat{S}S_j}{SS_j} - 1}{O} \right)^2 \qquad (5.13)$$

where, O is the number of stores

The variable distance decay model is superior to the Huff (1963) model in two aspects: capturing the area related effects, utilizing the information we have about the existing stores before the store opening. Alternative model predicts the existing store sales after the new store opening better than Huff (1963) model. Moreover, by using the attractiveness values obtained, area specific effects can be captured by the alternative model. A limitation of the alternative model is prediction of new store attractiveness value which the model results highly depend on.

5.3.3 A decision support system based on the model 5.3.1

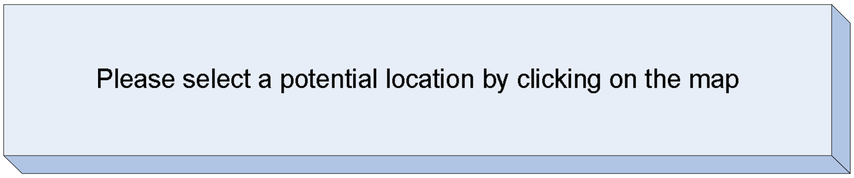
The method described in previous sections can be standardized to be used in a spatial decision support system. This section gives the details of the decision support system. The new system will be used for potential store revenue prediction purposes. The decision support system will be based on the GIS system. Parameter calibration can be handled in Microsoft Excel Solver by the GIS system automatically, or it can offer parameter values that can be overridden by the user. The following model which has also been described in detail in Section 5.3.1 as Equation (5.2) will be used for sales prediction of the potential stores.

$$SS_j = \frac{\sum_i (\sum_{j\in R} s_{ji} p_{ij}) Q_t Q_{t+1} N_j}{G_j}$$

$$\hat{p}_{ij} = \frac{A_j d_{ij}^{\theta_j}}{\sum_l A_l d_{il}^{\theta_l}}$$

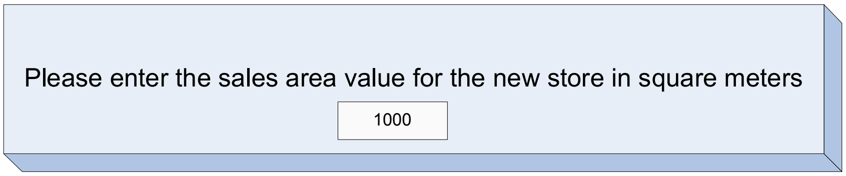
Of all the parameters above, d_{ij}, s_i and w_j data are stored in the GIS system. A_j, c and λ are estimated with the two step estimation procedure described in Chapter 4. Q_t, Q_{t+1}, N_j and G_j are estimated separately using the data in the GIS system as described in Section 5.3.1.

First step for decision makers for store revenue prediction is to select the analysis tool from the GIS menu. Then decision maker should select a potential location for the store and then enter a store size value for the potential store.



Please select a potential location by clicking on the map

Figure 5.3: User selects the location for the new store

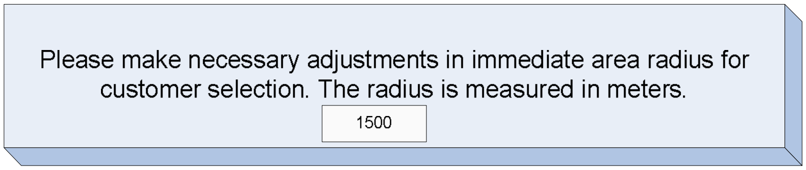


Please enter the sales area value for the new store in square meters

1000

Figure 5.4: User enters the size of the new store

The model we are using for revenue prediction first predicts the vicinity cannibalization amount and then generalizes this amount the total sales. In order to predict the cannibalization amount we need the company owned stores around the potential store. For this purpose the system finds the coordinates of the selected location and includes the company owned stores in the vicinity area, where radius is defined as the double of the immediate area radius which is defined by the user. Then these stores are listed for the decision maker to make adjustments on them. Firstly, we have to point out the customers to be included in the analysis. The default option is to select customers in the immediate area where radius is defined as 1500 meters. The radius of the immediate area is can be changed by the user. After the immediate area definition, stores to be included in the analysis can be selected using the vicinity concept.



Please make necessary adjustments in immediate area radius for customer selection. The radius is measured in meters.

1500

Figure 5.5: User determines the immediate area criteria for customer selection

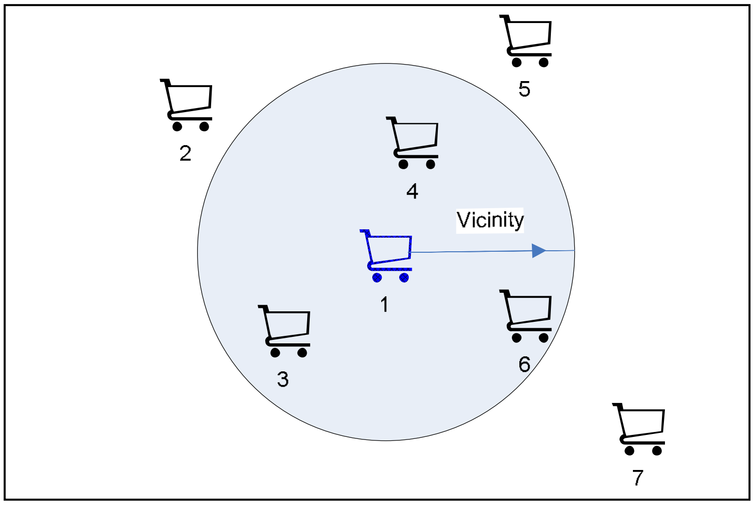
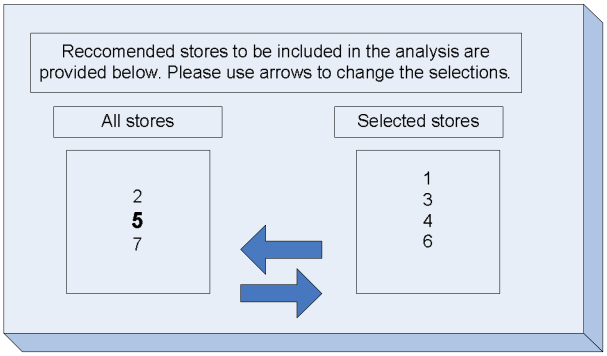


Figure 5.6: Company owned competitor store selection

Lets assume that 1 is the new store.



Reccomended stores to be included in the analysis are provided below. Please use arrows to change the selections.

All stores

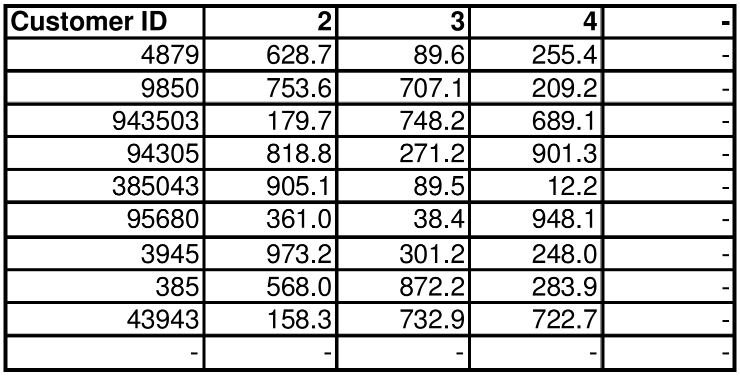
2
5
7

Selected stores

1
3
4
6

Figure 5.7: User sees the selected competitor stores to be included in the analysis and can add or remove stores from the list.

Once the customers to be included in the analysis are selected on GIS, we can obtain the distance of customers to selected stores from a distance table stored in GIS.

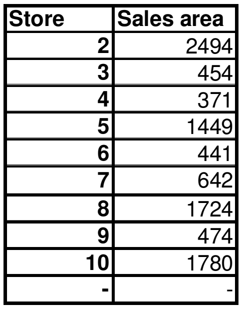


Customer ID	2	3	4	-
4879	628.7	89.6	255.4	-
9850	753.6	707.1	209.2	-
943503	179.7	748.2	689.1	-
94305	818.8	271.2	901.3	-
385043	905.1	89.5	12.2	-
95680	361.0	38.4	948.1	-
3945	973.2	301.2	248.0	-
385	568.0	872.2	283.9	-
43943	158.3	732.9	722.7	-
-	-	-	-	-

Table 5.7: Example picture illustrating the distance of customers to existing stores.

Customer distances to the new store are not readily available on the GIS. Thus, the customer distances to the potential store are calculated using the coordinates.

Like customers' distances to the existing stores, store sizes of the existing stores are readily available in the GIS database and can easily retrieved from the database for selected stores. Store size of the potential store is entered by the user.



Store	Sales area
2	2494
3	454
4	371
5	1449
6	441
7	642
8	1724
9	474
10	1780
-	-

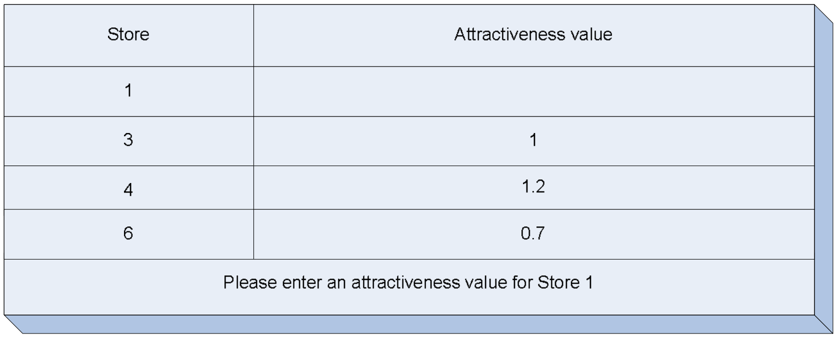
Table 5.8: Table illustrating the existing store sizes.

For data calibration purposes we need the customer spending data in the previous year. Customer spending data is available for existing stores. Data calibration is completed on the existing stores.

Customer ID	2	3	4	-
4879	150.2	76.1	186.5	-
9850	419.7	458.2	233.5	-
943503	478.7	151.2	477.1	-
94305	131.5	137.8	179.5	-
385043	232.6	346.8	13.6	-
95680	198.5	444.9	108.2	-
3945	396.5	354.7	308.8	-
385	38.4	296.4	270.8	-
43943	372.2	48.1	252.0	-
-	-	-	-	-

Table 5.9: Customer spending data for all existing stores. Selected customer spending data for selected stores is retrieved from this table.

Now that the all data is available for data calibration the system can determine the parameters. System runs the two step optimization procedure described in Chapter 4 and estimates the *c, λ* and A_j variables. Excel Solver can be used for this procedure. Calibration process is handled by the system integrating Excel and GIS system. Data needed for the analysis will be exported to Excel. The user will not see the calibration process. A_j variables are displayed to user for existing stores. A_j variable for the new store should be determined by the user. Values of *c, λ* and A_j are determined in Excel as described in Chapter 4.



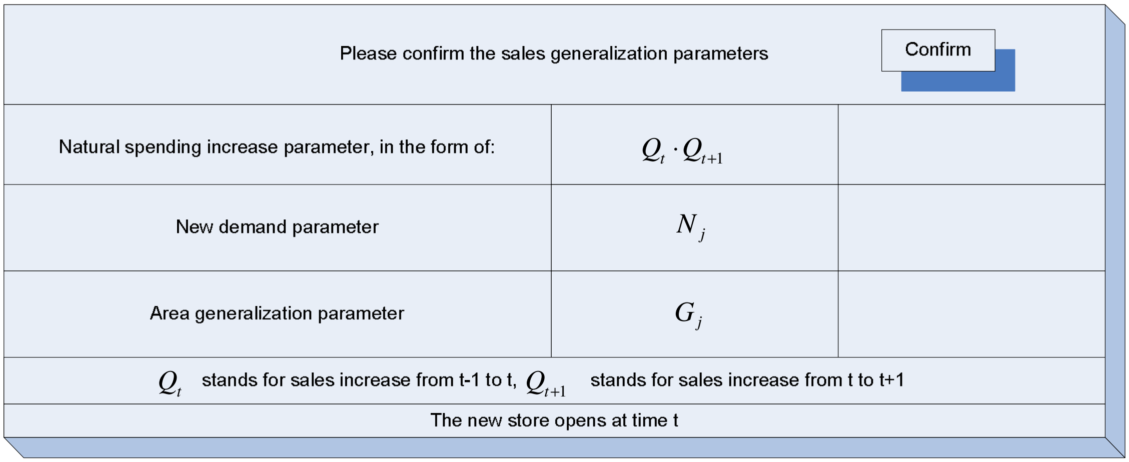
Store	Attractiveness value
1	
3	1
4	1.2
6	0.7

Please enter an attractiveness value for Store 1

Figure 5.8: User enters the values for the new store attractiveness parameter

A_j value for the new store is not readily available in the GIS system. User can enter a value based on his judgment. In addition the system will report the sales result for the following attractiveness values for the new store: Mean attractiveness score of the selected stores, one standard deviation of existing store attractiveness scores added or subtracted from the mean and two standard deviations of existing store attractiveness scores added or subtracted from the mean.

Now we obtained the sales coming from the area selected in terms of previous spending of customers. We should adjust the value for new demand generation, natural spending increase and the sales coming from out of the area selected. *N* parameter is estimated using Excel's regression tool, similarly *Q* parameters are calibrated on Excel automatically as described in the previous sections. Data needed for the analysis is stored in the GIS system and exported to the Excel before the analysis. The user doesn't see the calibration process. After the calibration, the system will display the values to the user and let the user make necessary adjustments. *G* parameter is specified by the user.



Please confirm the sales generalization parameters

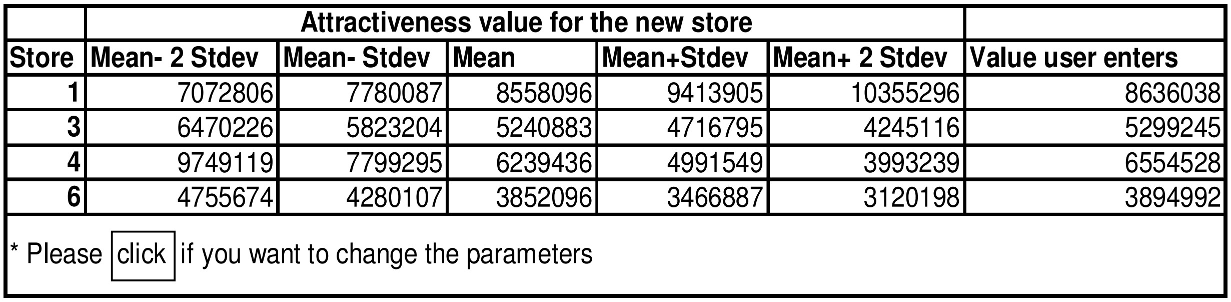
Natural spending increase parameter, in the form of:	$Q_t \cdot Q_{t+1}$	
New demand parameter	N_j	
Area generalization parameter	G_j	

Q_t stands for sales increase from t-1 to t, Q_{t+1} stands for sales increase from t to t+1

The new store opens at time t

Figure 5.9: User confirms the values for spending increase, new demand and area generalization parameter.

After the user confirms the relevant parameters, GIS system provides the sales predictions for different attractiveness values of the new store.



	Attractiveness value for the new store					
Store	Mean- 2 Stdev	Mean- Stdev	Mean	Mean+Stdev	Mean+ 2 Stdev	Value user enters
1	7072806	7780087	8558096	9413905	10355296	8636038
3	6470226	5823204	5240883	4716795	4245116	5299245
4	9749119	7799295	6239436	4991549	3993239	6554528
6	4755674	4280107	3852096	3466887	3120198	3894992

* Please click if you want to change the parameters

Figure 5.10: Displays the sales predictions of selected stores for different attractiveness values of the new store. User can change the parameters and go back to user screen 6 by clicking the button.

Figure 5.10 provides the sales predictions of the stores included to the analysis. The sales predictions change for different attractiveness values of the new store. Six attractiveness values of the new store are covered in this screen: two pessimistic scenarios, two optimistic scenarios an average case scenario and a user determined scenario. Mean value is the average of attractiveness values of existing stores selected and standard deviation is the standard deviation of existing stores. User can go back to Figure 5.9 to change the sales generalization parameters.

5.4 Discussion

Spatial interaction models can be used for store sales revenue prediction. If the panel data and zip code level detailed data are available sales revenue prediction for a new store is relatively simple. If the data available is limited to the customer loyalty card data, decision makers can use the method proposed in Section 5.3.1 which includes sales prediction in the immediate area using company owned stores and then generalization to the total sales using various parameters. According to the numerical example in Section 5.3.2 the method proposed in Section 5.3.1 can be used with the alternative spatial interaction model proposed in Chapter 4. This combination generates reasonable results if the attractiveness value is selected appropriately in comparison to the Huff (1963) model. The new method proposed in Section 5.3.1 can be integrated with the GIS system to serve as a decision support system.

Chapter 6

CONCLUSION

6.1 Overview

The main goal of this thesis research is to predict revenue of a potential grocery retail store with greater accuracy compared to the models in the literature. This information is valuable in the grocery store site selection context. Grocery retail store sales are customer driven. Thus, in order to predict revenue of a grocery store we should understand how customers patronize grocery retail stores. Spatial interaction models are analytical models which portray customer store choice behavior. Spatial interaction models are very structured, practical and provide reasonably good estimates. For these reasons, we focus on creating a new flexible spatial interaction model which provides more accurate estimates compared to the existing ones. Flexibility is used in terms of variable and constraint specification and a flexible model can perform better in multiple objectives such as providing more accurate customer share predictions and store sales estimations.

This research is different from the other studies in the literature in terms of spatial interaction model specification, estimation, objective function and sales generalization method. Our spatial interaction model contains a new distance deterrence function. In this new function we define the distance decay value as a function of store size. We observe that this representation of the distance decay value provides better customer share prediction results compared to Huff (1963) model. In order to capture site specific effects we introduce store attractiveness values. Site specific attractiveness values are also used in Drezner and Drezner's (2002) research on shopping center data for exploratory research. We use the spatial interaction model with the variable distance decay and site specific attractiveness values for grocery store sales prediction.

We estimate our model with optimization which is fairly uncommon in spatial interaction model calibration except for the Drezner and Drezner (2002) study. Most of the spatial interaction models are calibrated with the log-transform approach of Nakanishi and Cooper (1974) like MCI (1974) and Competing destinations (1988). Log-transform approach is quite practical but not flexible in terms of constraint and objective function specification. Usage of optimization provides us greater flexibility in terms of objective function specification. Different from previous studies like, MCI (1974), Competing destinations (1988), Drezner and Drezner (2002) we minimize customer store share prediction error while controlling for the general sales prediction performance of the model. We obtain this result by introducing a sales prediction constraint to the model. We observe that the new method provides better customer share prediction and store sales prediction results compared to the discussed models. Finally this thesis covers a method of sales prediction with spatial interaction models using loyalty card data when the panel or survey data is absent. This situation is not covered in the literature.

6.2 Discussion of findings

In this thesis research we study the factors influencing the customer shopping behavior in Chapter 3. In Chapter 4 we propose a new spatial interaction model with store attractiveness values and variable distance decay. Chapter 4 also includes empirical evaluation of the model on existing stores for customer store selection and empirical evaluation of the model to predict customer store share and store sales. In Chapter 5 we propose a method for generalizing the store sales obtained from the immediate area to the real store sales. Moreover, in Chapter 5 we also provide a decision support system that incorporates our share prediction method. We provide our main findings, conclusions and limitations of our research throughout this chapter.

In Chapter 3, in macro level analysis we conclude that distance between the stores, store size, format and competition are among the important factors. A new store opening affects a distant store less compared to a close store. Moreover, if the format of the new store is different probability of the present store sales to decrease is less. A managerial implication of this finding is that a grocery retail store chain should open a new store distant to the company owned and competitor stores. Moreover, company may consider choosing a different format for the new store if it should open the new store close to the company owned stores. Also, we observe that grocery retail sales are highly seasonal. Thus, in order to reduce this affect companies may focus on increasing the demand in the off-season period by promotion campaigns.

In the customer level analysis, we observe that as customer distance to the store increases, customer average yearly spending decreases. We also observe this trend may be interrupted because of the competition around the store as we see in the store 1 case. Managerial implication of this finding is that grocery store chains should prioritize areas with high population density and low competition while selecting a location. In customer level analysis, we see that bigger store sales are generated in a larger area. Large stores are less affected from distance effects. We may say that, for low density residential areas companies should open bigger stores to attract more customers from abroad.

According to our findings in Chapter 4, the proposed spatial interaction model provides significantly better mean squared error results than other spatial interaction models including Huff (1963), MCI (1974), Competing destinations (1988) in both with and without the explicit sales prediction constraint cases. We observe that explicit sales prediction constraint decreases the performance of the new model in individual share prediction but increases the model performance in sales prediction. Moreover, in both cases the new spatial interaction model provides better sales prediction ratios compared to the existing models. It is advantageous to set the upperbound and lowerbound range wider if the customer share prediction performance is more important for the decision maker. Predicting customer shares with greater accuracy is valuable for format and pricing decisions. For example, the company can identify the potential stores' clientele. Examining the potential customers' shopping behavior decision makers can determine the appropriate format or product portfolio.

In Chapter 4 we observe that as store size increases, distance decay value decreases. Our conclusion is, bigger stores are less affected from distance. This finding is parallel to the insights in Chapter 3.

We also examine the store attractiveness values obtained in Chapter 4 and Chapter 5 with correlation analysis. A finding of this analysis is the negative correlation between the attractiveness values and the competition. An implication of this final finding is that competition affects store sales negatively as expected. This finding is parallel to the insights obtained in Chapter 3. In Chapter 3 we also observe that as distance to the store increases customer spending decreases. Contradicting with this observation, we see a negative correlation between the population density and customer share in Chapter 4. It is important to note that there is a significant positive correlation between the population density and the competition. Therefore, high population density areas also attract competitor store chains. In contrast to the finding in Chapter 4 we observe a positive correlation between attractiveness values and population density. We can conclude that, competition's effect on attractiveness values depend on the intensity of the population density. It is crucial to note that the area selected in Chapter 4 has a higher population density compared to the area in Chapter 5. If population density is intense enough possibility of its effect on attractiveness to be positive is higher. So, it is crucial for grocery retail companies to follow trends in residential areas. It is advantageous for grocery retailers to be the first entrants in high population density areas. Companies may invest in predicting high population density areas of near future. Also, it is advantageous to open stores to the areas where population density is more intense as possibility of market saturation is lower in this case.

In Chapter 5 we propose a method to use the new spatial interaction model provided in Chapter 4 to predict retail store sales. The generalization method provides reasonably accurate results when the attractiveness value of the new store is predicted in an appropriate range.

6.3 Contributions

Our research contributes to retail store revenue prediction and site selection literature with a more accurate and flexible spatial interaction model and a flexible estimation procedure. We use optimization as a tool for spatial interaction model calibration. Our objective function is to minimize customer store share prediction error. In addition, we include a store sales prediction constraint which allows researcher to manage the tradeoff between the store sales prediction and individual customer error performance. Our research is valuable because it provides accurate store sales and customer share estimates. Accurate customer share estimations are valuable for store format and pricing decisions where accurate sales predictions are relevant to store location evaluation. Also, our research provides insights about the relationship between the distance decay value and the store size as well as insights about customer shopping behavior and grocery retailing. We conclude that when distance decay value is represented as a function of store size it provides better error results compared to conventional spatial interaction models such as Huff (1963), MCI (1974) and Competing destinations (1988). Finally, our research provides a new sales generalization method where survey and panel data is not available. This situation is not covered in the literature. The method we propose is intuitive and provides reasonably well estimates on numerical examples.

6.4 Limitations

An obvious limitation in this study is about the data available to us. We only have data from one retail chain so we have to ignore factors such as brand, pricing and reputation.

The new spatial interaction model proposed in Chapter 4 provides significantly better error results than most well known and widely used models in the literature on our available data. Although, the model provides superior results, estimation is problematic. Because the utility of a store is defined with double power function estimation with log-transform method (1974) is not possible. We estimate the function with a two-step optimization procedure. However, the obtained solution is not guaranteed to be the global optimal solution. Also, in the two step estimation procedure proposed in Chapter 4, step 1 is not convex or quasi-convex but yet the solution is estimated with optimization. Therefore, there is possibility to converge to a local optimal solution rather than the global optima. This problem can be solved by graphing the function and confirming the global optima as solution space is limited to a range because of the power function. This problem is common in spatial interaction model parameter estimation like Haines et al. (1972) study.

Another limitation of the new model is prediction of attractiveness of a new store. In our covered examples, because of the sample size limitations in the data we could not find a generalized formula for store attractiveness. We examined the correlation results and decided subjectively. In our correlation results, we observe that attractiveness scores were negatively correlated with competition. However, samples are too small for generalization of this rule.

In Chapter 5 we propose a rather intuitive method for sales generalization when panel or survey data is not available. The area generalization parameter is problematic in the proposed formula because of the unique structure of each location. We propose the analogue approach for estimation of area generalization parameter but this method is subjective and accuracy of the parameters estimated are limited to the experience of the decision maker. Our sales generalization method can predict retail sales when multiple company owned stores are available in the location even panel data is not available. However, sales prediction with our model is not possible if no company owned stores exist in the candidate location.

6.5 Further research

Global optimization or a more efficient way of estimation of the proposed model in Chapter 4 may be included to the further research topics. Our model includes store attractiveness values. These values are estimated via optimization using the past data for existing stores. For a new store estimating attractiveness is difficult due to data limitations. This problem can possibly be solved by covering an example with more stores. Examining more stores, researchers may propose an analytical formula for store attractiveness. Finally, in Chapter 5 a more objective method may be proposed for area generalization parameter estimation. Using an analytical formula for estimation of area generalization would pursue this goal.

BIBLIOGRAPHY

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

Batty, M., Longley, P.(1996). Spatial Analysis Modelling in a GIS Environment John Wiley & Sons, Inc

Bazaara,M., Sherali,H., Sheti,C.M(1993). Nonlinear programming. Theory and Algorithms John Wiley & Sons, Inc

Benito, O., Gallego, P., Reyes, C. (2006). Isolating geodemographic characterization of retail format choice from effects of spatial convenience. *Marketing Letters*, 18, 45-59.

Black, W., Ostlund, L., Westbrook, R. (1985). Spatial demand models and intraband concept. *Journal of Marketing*, 49, 3, 106

Borgers, A., Timmermans, H., (1987). Choice model specification, substitution and spatial structure effects: A simulation experiment. *Regional Science and Urban Economics,* 17, 29-47.

Borges, A., Timmermans, H., Waerden, P. (1991). Mother Logit Analysis of Substitution Effects in Consumer Shopping Destination Choice. *Journal of Business Resources*, 23, 311-323.

Byrom, J., Hernandez, T., Bennison, D., Hooper, P. (2001). Exporing the geographical dimension in loyalty card data. *Marketing intelligence & planning*, 19, 3,162.

Clarke, G.(1998). Changing methods of location planning for retail companies. *Geojournal*, 45,289

Clarke, I., Rowley, J. (1995). A case for spatial-decision support systems in retail location planning. *International Journal of Retail& Distribution Management*, 23, 3, 4.

Clarkson, R., Clarke, C., Robinson, T. (1996). UK supermarket location assessment. *International Journal of Retail & Distributions Management,* 24, 6, 22.

Cliquet,G. (1995). Implementing a subjective MCI model: An application to the furniture market. *European Journal of Operations Research,* 8, 4, 279-291.

Colome,R., Lourenço H., Serra D. (2003). A new chance constrained maximum capture location problem. *Annals of Operations Research*, 122, 121.

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

Craig,S., Ghosh, A. (1986). An approach to determining optimal locations for new services. *Journal of Marketing Research,* 23, 4, 354-362.

Cooper, L., Nakanishi, M. (1974). Parameter estimation of a MCI model, least squares approach. *Journal of Marketing Research*, 11, 3, 303-311.

Cooper, L., Nakanishi, M. (1983). Standardizing variables in MCI model. *The journal of consumer research*, 10, 1, 96-108.

Drezner, Z., Drezner, T. (2002). Validating the gravity-based competitive location model using inferred attractiveness. *Annals of Operations Research*, 111, 1, 227.

Dubin, R., Pace, K., Thibodeau, T.(1999) Spatial autoregression techniques for real estate data. *Journal of Real Estate Literature*, 7, 79-95

Erymann, A. (1995). Consumers' Spatial Choice Behavior, Heidelberg: Physica-Verlag

Fotheringham, S. (1983). Some theoretical aspects of destination choice and their relevance to production-constrained gravity models. *Environment and Planning*, 15, 1121–1132

Fotheringham, S. (1988). Consumer store choice and choice set definition. *Marketing Science*, 7, 3, 299.

Fotherinham, S., Rogerson, P., National Center of Geographic Information and Analysis (1994) Spatial analysis and GIS, Applications in GIS, CRC Press.

Fotheringham ,S, Wegener, M. (2000). Spatial models and GIS, CRC Press.

Ghosh, A., Neslin, S., Shoemaker, R. (1984). A comparison of market share models and estimation procedures, *Journal of Marketing Research*, 21, 2, 202-210

Haines, G. Simon.L, Alexis, M. (1972). .Maximum likelihood estimation of central-city food tradind areas, *Journal of Marketing Research,* 9, 2, 154- 159

Hanke,R. Wichern,D. (1998). Business Forecasting, Englewood Cliffs, Prentice Hall

Hernandez, T. (2005), Visual decisions: Geo-visualization techniques within retail decision support. *Journal of Targeting, Measurement and Analysis for Marketing*, 13, 3, 209.

Huff, D. (1963). A probabilistic analysis of shopping center trade areas. *Land Economics,* 39, 81–90

Klosterman, E., Xie, Y. (1997). Retail impact analysis with loosely coupled GIS and a spreadsheet. *International Planning Studies,* 2, 175

Longley,P. (2004). Geographical information systems: on modeling and representation. *Progress in Human Geography*, 28, 108

Melaniphy, J. (1992) Restaurant and fast food site selection, John Wiley & Sons

Mendes, A. and Themido, I. (2004). Multi-outlet retail site location assessment. *International Transactions in Operational Research,* 11, 1-18

Meyer, R. J. and Eagle, T. C. (1982). Context-Induced Parameter Instability in a Disaggregate Stochastic Model of Store Choice. *Journal of Marketing Research,* 19, 62-71.

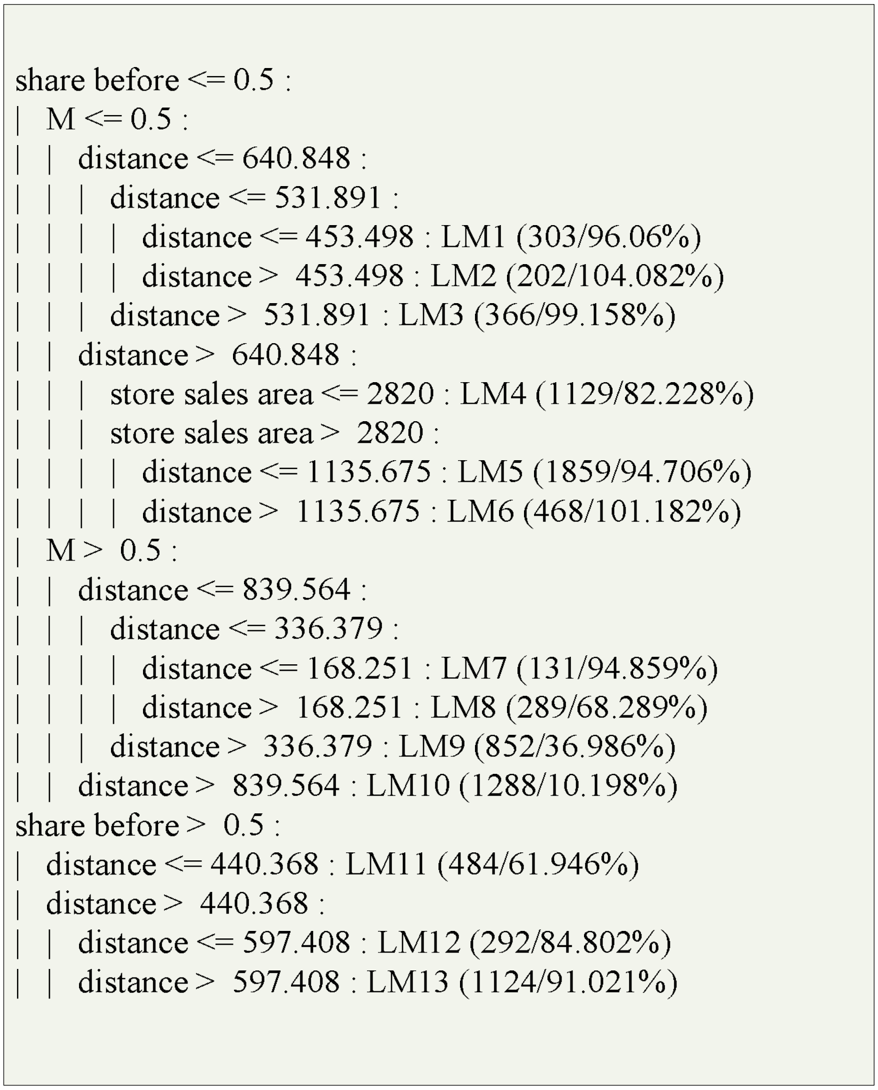
Nakaya, T., Fotheringham, S. , Hanaoka, K., Clarke, G., Ballas, D.,Yano, K. (2007). Combining micro simulation and spatial interaction models for retail location analysis. *Journal of Geographical Systems*, 9, 345-369.

Sakashita, N. (2000). An economic analysis of convenience-store location. *Urban studies*, 37, 3, 471-479.

Quinlan J. (1992). Learning with continuous classes. *Proceedings of the Australian Joint Conference on Artificial Intelligence*., 343--348. World Scientific, Singapore.

Wong,S, Yang, H. (1999). Determining market areas captured by competitive facilities: A continous equlibrium modelling approach. *Journal of Regional Science*, 39, 1, 51-72.

Appendix A



```
share before <= 0.5 :
|   M <= 0.5 :
|   |   distance <= 640.848 :
|   |   |   distance <= 531.891 :
|   |   |   |   distance <= 453.498 : LM1 (303/96.06%)
|   |   |   |   distance >  453.498 : LM2 (202/104.082%)
|   |   |   distance >  531.891 : LM3 (366/99.158%)
|   |   distance >  640.848 :
|   |   |   store sales area <= 2820 : LM4 (1129/82.228%)
|   |   |   store sales area >  2820 :
|   |   |   |   distance <= 1135.675 : LM5 (1859/94.706%)
|   |   |   |   distance >  1135.675 : LM6 (468/101.182%)
|   M >  0.5 :
|   |   distance <= 839.564 :
|   |   |   distance <= 336.379 :
|   |   |   |   distance <= 168.251 : LM7 (131/94.859%)
|   |   |   |   distance >  168.251 : LM8 (289/68.289%)
|   |   |   distance >  336.379 : LM9 (852/36.986%)
|   |   distance >  839.564 : LM10 (1288/10.198%)
share before >  0.5 :
|   distance <= 440.368 : LM11 (484/61.946%)
|   distance >  440.368 :
|   |   distance <= 597.408 : LM12 (292/84.802%)
|   |   distance >  597.408 : LM13 (1124/91.021%)
```

Figure.A Regression tree

```
LM num: 1
after share  =
        0.0004 * distance
        + 0.0022 * share
before
        + 0.0001 * store store
size
        + 0.1198

LM num: 2
after share  =
        + 0.0022 * share
before
        + 0.5749

LM num: 3
after share  =

        + 0.0022 * share
before
        + 0.0002 * store store
size
        - 0.0801

LM num: 4
after share  =
        + 0.0022 * share
before
        + 0.216

LM num: 5
after share  =
        -0.0001 * distance
        + 0.0022 * share
before
        + 0.3968

LM num: 6
after share  =
        + 0.0022 * share
before
        + 0.4073

LM num: 7
after share  =
        -0.0001 * distance
        + 0.0427 * share
before
        + 0.0003 * M
        + 0.4248

LM num: 8
after share  =
        -0.0001 * distance
        + 0.4936 * share
before
        + 0.0003 * M
        + 0.1704

LM num: 9
after share  =
        -0.0001 * distance
        + 0.0169 * share
before
        + 0.0003 * M
        + 0.0877

LM num: 10
after share  =
        + 0.0096 * share
before
        + 0.0003 * M
        + 0.0065

LM num: 11
after share  =
        -0.0003 * distance
        + 0.7238 * share
before
        + 0.0011 * M
        + 0.1815

LM num: 12
after share  =
        + 0.0178 * share
before
        + 0.0011 * M
        + 0.6722

LM num: 13
after share  =
        + 0.0102 * share
before
        + 0.0011 * M
        + 0.5883
```

Appendix B

t-Test: Paired Two Sample for Means

	4.1	*1.1*
Mean	0.040642	0.041324
Variance	0.001475	0.001593
Observations	2327	2327
Pearson Correlation	0.957883	
Hypothesized Mean Difference	0	
df	2326	
t Stat	-2.86711	
P(T<=t) one-tail	0.00209	
t Critical one-tail	1.645509	
P(T<=t) two-tail	0.00418	
t Critical two-tail	1.960984	

t-Test: Paired Two Sample for Means

	4.1	*1.2*
Mean	0.040642	0.041537
Variance	0.001475	0.001816
Observations	2327	2327
Pearson Correlation	0.959873	
Hypothesized Mean Difference	0	
df	2326	
t Stat	-3.53543	
P(T<=t) one-tail	0.000208	
t Critical one-tail	1.645509	
P(T<=t) two-tail	0.000415	
t Critical two-tail	1.960984	

t-Test: Paired Two Sample for Means

	4.1	*2.1*
Mean	0.040642	0.046584
Variance	0.001475	0.002093
Observations	2327	2327
Pearson Correlation	0.914032	
Hypothesized Mean Difference	0	
df	2326	
t Stat	-15.1873	
P(T<=t) one-tail	4.82E-50	
t Critical one-tail	1.645509	
P(T<=t) two-tail	9.64E-50	
t Critical two-tail	1.960984	

t-Test: Paired Two Sample for Means

	4.1	*2.2*
Mean	0.040642	0.041203
Variance	0.001475	0.001713
Observations	2327	2327
Pearson Correlation	0.962407	
Hypothesized Mean Difference	0	
df	2326	
t Stat	-2.38542	
P(T<=t) one-tail	0.00857	
t Critical one-tail	1.645509	
P(T<=t) two-tail	0.017139	
t Critical two-tail	1.960984	

t-Test: Paired Two Sample for Means

	4.1	*3.1*
Mean	0.040642	0.052357
Variance	0.001475	0.000642
Observations	2327	2327
Pearson Correlation	0.662176	
Hypothesized Mean Difference	0	
df	2326	
t Stat	-19.63702	
P(T<=t) one-tail	7.3E-80	
t Critical one-tail	1.645509	
P(T<=t) two-tail	1.46E-79	
t Critical two-tail	1.960984	

t-Test: Paired Two Sample for Means

	4.1	*3.2*
Mean	0.040642	0.05592
Variance	0.001475	0.000745
Observations	2327	2327
Pearson Correlation	0.563034	
Hypothesized Mean Difference	0	
df	2326	
t Stat	-22.85933	
P(T<=t) one-tail	8.4E-105	
t Critical one-tail	1.645509	
P(T<=t) two-tail	1.7E-104	
t Critical two-tail	1.960984	

Figure B.1 t-test results for significance Model 4.1

t-Test: Paired Two Sample for Means

	4.2	*1.1*
Mean	0.039183	0.041324
Variance	0.001441	0.001593
Observations	2327	2327
Pearson Correlation	0.957672	
Hypothesized Mean Difference	0	
df	2326	
t Stat	-8.98502	
P(T<=t) one-tail	2.61E-19	
t Critical one-tail	1.645509	
P(T<=t) two-tail	5.22E-19	
t Critical two-tail	1.960984	

t-Test: Paired Two Sample for Means

	4.2	*1.2*
Mean	0.039183	0.041537
Variance	0.001441	0.001816
Observations	2327	2327
Pearson Correlation	0.95592	
Hypothesized Mean Difference	0	
df	2326	
t Stat	-8.86085	
P(T<=t) one-tail	7.72E-19	
t Critical one-tail	1.645509	
P(T<=t) two-tail	1.54E-18	
t Critical two-tail	1.960984	

t-Test: Paired Two Sample for Means

	4.2	*2.1*
Mean	0.039183	0.046584
Variance	0.001441	0.002093
Observations	2327	2327
Pearson Correlation	0.88025	
Hypothesized Mean Difference	0	
df	2326	
t Stat	-16.3513	
P(T<=t) one-tail	2.84E-57	
t Critical one-tail	1.645509	
P(T<=t) two-tail	5.68E-57	
t Critical two-tail	1.960984	

t-Test: Paired Two Sample for Means

	4.2	*2.2*
Mean	0.039183	0.041203
Variance	0.001441	0.001713
Observations	2327	2327
Pearson Correlation	0.95877	
Hypothesized Mean Difference	0	
df	2326	
t Stat	-8.19682	
P(T<=t) one-tail	2.01E-16	
t Critical one-tail	1.645509	
P(T<=t) two-tail	4.03E-16	
t Critical two-tail	1.960984	

t-Test: Paired Two Sample for Means

	4.2	*3.1*
Mean	0.039183	0.05236
Variance	0.001441	0.00064
Observations	2327	2327
Pearson Correlation	0.693858	
Hypothesized Mean Difference	0	
df	2326	
t Stat	-23.2313	
P(T<=t) one-tail	7.7E-108	
t Critical one-tail	1.645509	
P(T<=t) two-tail	1.5E-107	
t Critical two-tail	1.960984	

t-Test: Paired Two Sample for Means

	4.2	*3.2*
Mean	0.039183	0.05592
Variance	0.001441	0.00074
Observations	2327	2327
Pearson Correlation	0.598613	
Hypothesized Mean Difference	0	
df	2326	
t Stat	-26.256	
P(T<=t) one-tail	1.3E-133	
t Critical one-tail	1.645509	
P(T<=t) two-tail	2.7E-133	
t Critical two-tail	1.960984	

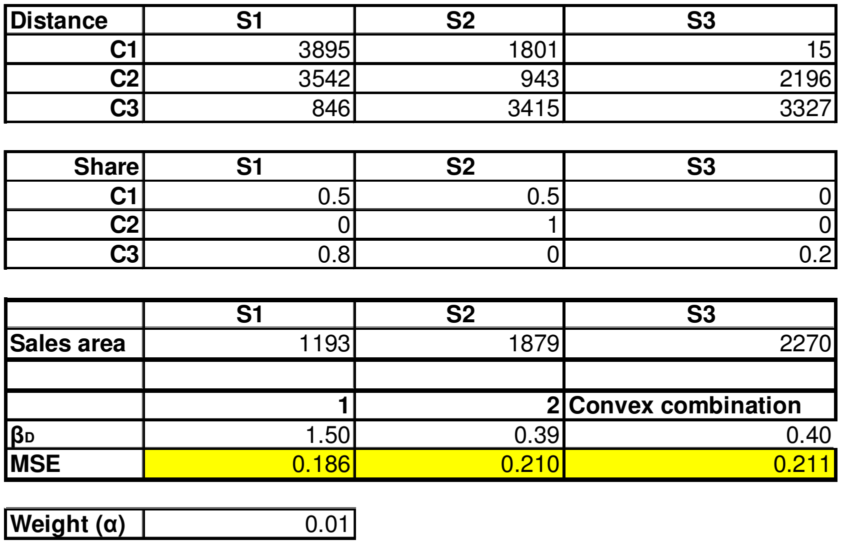
Figure B.2: t-test results for significance Model 4.2

Appendix C

Quasiconvexity

Definition C.1: if $f(\alpha[\beta_{D1}]+(1-\alpha)[\beta_{D2}]) \le \max(f[\beta_{D1}], f[\beta_{D2}])$ then f is quasi-convex. (Bazaara et al.,1993)

Counterexample for quasiconvexity Huff(1963) Model



Distance	S1	S2	S3
C1	3895	1801	15
C2	3542	943	2196
C3	846	3415	3327

Share	S1	S2	S3
C1	0.5	0.5	0
C2	0	1	0
C3	0.8	0	0.2

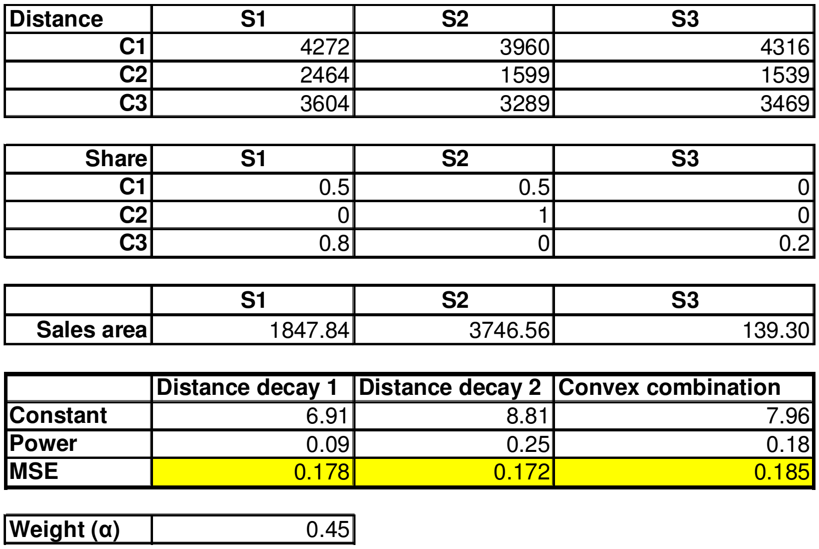
	S1	S2	S3
Sales area	1193	1879	2270
	1	2	Convex combination
βD	1.50	0.39	0.40
MSE	0.186	0.210	0.211

Weight (α)	0.01

Table C.1: Counterexample for quasi-convexity of the Huff (1963) model.

Table C.1 provides a counterexample on hypothetical data for quasi-convexity of the Huff (1963) Model. C1,C2 and C3 are customers and S1, S2 and S3 are stores. Distance, share and store size data are also provided in Table C.1. Weight (α) is accepted as 0.61. Examining Table 4.1 we observe that MSE error value for convex combination is higher than maximum error value of β_D's. Thus, Huff (1963) Model MSE function is not quasi-convex.

Counter example for quasiconvexity of Step 1



Distance	S1	S2	S3
C1	4272	3960	4316
C2	2464	1599	1539
C3	3604	3289	3469

Share	S1	S2	S3
C1	0.5	0.5	0
C2	0	1	0
C3	0.8	0	0.2

	S1	S2	S3
Sales area	1847.84	3746.56	139.30

	Distance decay 1	Distance decay 2	Convex combination
Constant	6.91	8.81	7.96
Power	0.09	0.25	0.18
MSE	0.178	0.172	0.185

Weight (α)	0.45

Table C.2: Counterexample quasiconvexity for Step 1

Table C.2 provides a counterexample for quasiconvexity of Step 1 MSE function on hypothetical data. C1, C2, C3 are customers and S1,S2 and S3 are stores. Used distance, store size and share values are provided in the above example and α is accepted as 0.45. According to Table 2 MSE value for convex combination of distance decay 1 and distance decay 2 is higher than the maximum MSE value for distance decay 1 and 2. Thus, we can conclude that the function is now quasi-convex.

VITA

Müge Sandıkçıoğlu was born in Ankara June 8, 1984. She graduated from TED Ankara College in 2001. She received her BS degree in Business Administration from Koç University, Istanbul. She joined Industrial Engineering department of Koç University as a teaching assistant in September 2006.