Competitive Intelligence Analytics for Operations Managers

Müge Tekin

TESI DOCTORAL UPF / 2018

DIRECTOR DE LA TESI
Prof. Kalyan T. Talluri

DEPARTMENT OF ECONOMICS AND BUSINESS

Acknowledgements

"Hi Muge, we are excited to have you here. Looking forward to your arrival." Receiving this email from Prof. Kalyan Talluri, the so called *father* of revenue management who had agreed to be my advisor, marked one of my happiest moments. It was the first time I was leaving Turkey to live abroad. I was born and raised in a small town, moved to Istanbul for high school; Barcelona, however, has been a completely different city for me, with its many cultures, new experiences and vivid colors.

Aside from those personal challenges, when I started my PhD I found myself among a different academic community in the Economics department. It took me some time to adapt to this new reality. Thanks to my advisor's wise, calm and understanding nature, I was able to find my own path. I am very grateful for all his time, insightful ideas and guidance. Doing research with him has always been an intense learning experience. I appreciate working on real-life problems. I am very much indebted to him for all the opportunities he provided.

Prof. Robin Hogarth, a true exemplary teacher, was also an important figure in my trajectory. His devotion to academic life and energy always inspired me. The enthusiasm which he brought to his classes made them interesting and fun.

This work has been partially supported by Grant ECO2013-41131-R from the Spanish Ministry of Economy and Competitiveness. The participation to summer school in Zaragoza Logistics Center and international conferences were useful to interact with the research community in my field. I also would like to thank Prof. Mihalis Markakis for this opportunity.

I benefited from my visits to Imperial College immensely. The third floor of business school has been welcoming and supportive. I enjoyed our talks with Esma.

At UPF, we shared good times with my office mates and the lunch breaks were fun with Eda. Thanks also to Marta and Laura for being patient with my administrative problems.

I have been very fortunate to meet Pedro when I arrived to Barcelona. I will always remember his curious, eye-opening spirit and the good times we shared. Many thanks for all his patience, continuous support, help and being my best companion.

Finally, to my parents Emel and Yaşar, who are still not much conscious about what I have been doing but whose drive has always helped me to succeed, I want to dedicate my thesis.

Abstract

This thesis incorporates three studies that revolve around competitive intelligence of a firm, available data sources and improvement of operations through data-informed decision-making. The studies cover a range of topics in the field of operations research: revenue management, capital budgeting decisions and social learning. Chapter 1 presents a new econometric method for estimating customer choice model parameters based on competitor intelligence data. Chapter 2 combines public data on social review platforms with demographic and geographic data to inform facility location and product design decisions in service based industries such as restaurants and hotels. Chapter 3 investigates the effect of social reputation platforms on customer behavior and assesses whether these platforms transmit the general customer opinion. Overall, the studies provide novel theoretical reflections and practical ways through which businesses can implement competitive intelligence to add value to their operations.

Resumen

Esta tesis está compuesta por tres estudios que giran en torno a la inteligencia competitiva de una firma, las fuentes de datos disponibles y la mejora de las operaciones a través de la toma informada de decisiones. Los estudios cubren una variedad de temas en el campo de la investigación de operaciones: administración de ingresos, decisiones sobre el presupuesto de capital y aprendizaje social. El Capítulo 1 presenta un nuevo método econométrico para estimar los parámetros de modelos de toma de decisiones del consumidor basado en base a datos de inteligencia competitiva. El Capítulo 2 combina datos públicos, disponibles en las plataformas de reseña social, con datos demográficos y geográficos para informar la ubicación de instalaciones y las decisiones de diseño de productos en industrias basadas en servicios, como restaurantes y hoteles. El Capítulo 3 investiga el efecto de las plataformas de reseña social en el comportamiento del consumidor y evalúa si estas plataformas transmiten la opinión general de los clientes. En términos generales, los estudios proporcionan nuevas reflexiones teóricas y formas prácticas a través de las cuales las empresas pueden implementar inteligencia competitiva para agregar valor a sus operaciones.

The Landscape of Competitive Intelligence (CI)

> *"If you know the enemy and know yourself, you need not fear the result of a hundred battles. If you know yourself but not the enemy, for every victory gained you will also suffer a defeat. If you know neither the enemy nor yourself, you will succumb in every battle."*
>
> – Sun Tzu, *The Art of War*

Firms increasingly understand the importance of determining the *right* price to charge their customers, identifying the *right* location to operate and deciding the *right* design elements for their goods and services. Despite the existence of this generalized notion, one is left to wonder how those decisions should be made.

Surely, in-depth knowledge of one's firm operations is necessary, but is it sufficient? If you were a manager of a firm, wouldn't you also observe activities of your rivals' in the market along with customer behavior (say understand customer tastes) and take all these items into consideration?

In this thesis, we investigate those questions from an operational point of view in an integrated manner—understanding customer behavior, firm operations together with competitors'—to make better tactical and strategic decisions. In the age of digitalization, the consequences of wrong decisions can be very severe (e.g. Marks & Spencer suffers from it's Internet presence and sharper competition this creates (Wood and Butler, 2018)).

Recently, the notion of *Competitive Intelligence* (CI) is becoming popular among industry practitioners. It refers to the collection and analysis of competitor and market (competitive environment) data to inform optimal decision-making and broader firm strategies.

What exactly is CI? Let's illustrate with a story: Ali is a teenage boy selling a chocolate ice-cream for 3 euros per cup on a beach during summer holiday. He receives around 50 customers daily, each day getting closer to his dream of saving sufficient money to go camping with his friends. All is fine until an ice-cream shop opens nearby and Ali sees his customers going away. Desperate, he disguises himself and watches what happens in the newly opened shop. He realizes that around 20 varieties of ice-creams are being sold at 2 euros per cup, and also takes note of the various ice-creams customers buy the most. This knowledge can help him attract customers back. The next day, Ali can bring a few of the most desired ice-cream types to sell and re-price them at 2 euros or less per cup.

While this story illustrates the nature of a competitive environment and the usefulness

of observing competitor operations, today's complex business and customer environment is much more complicated.

First, there are many factors involved in purchasing decisions such as price, quality, location, and flexibility. Firms need to identify how (and how much) each factor affects customer behavior in real-life situations where all these factors come together and interact (making it hard to distinguish each factor's real impact).

Second, like Ali went to watch the new ice-cream shop operations, firms also "stalk" their competitors to some extent for gathering information on competitor prices as well as several other benchmarking measures (e.g. sales, profit, costs, budget, customer satisfaction). However, data on competitor is not so easily obtained. Some proprietary information sources on competitor, as we will mention later in detail, are available but unfortunately they provide only partial information, often only in aggregated forms.

Third, data is useful but never enough: it needs to be transformed into knowledge that can provide insights about how to improve firm operations. Lounamaa and March (1987) pointed to a similar argument: "[the] central dilemma in modern organization theory and operations research is the mismatch between analytical capabilities of human institutions and the complexity of the environment in which they function." Today, thirty-one years after that statement was written, this dilemma seems to have grown exponentially.

In *High Output Management*, Grove (2015) claims complex systems are black-boxes and an insight is like a window cut into the side of the black box that "sheds light" on what's going on inside. The operation of a business is a complex system for the reasons mentioned above. Not only are the factors involving firm operations and markets extremely complex, they are also a part of a dynamic and self-transforming process. In this sense, the insight gathered through CI practices can be used to improve business operations.

So, CI is the act of collecting and analyzing actionable information about competitors and the marketplace to form a business strategy. Its aim is to learn everything there is to know about the competitive environment outside to make the best possible decisions about how to run a business. Figure 1 illustrates the different elements involved in CI.

The term Competitive Intelligence first appeared in the U.S. during the 1970s. When CI practices first came to light, there were discussions over their legality, equating CI with industrial espionage. However, CI practices by definition are limited to legal means of information gathering and are considered a legal business practice, unlike the more sinister and illegal corporate espionage practices.

Figure 1: Elements of CI
Source: www.biakelsey.com

Despite being commonly associated with competitor analysis, CI requires considering the complex relations between products, customers, competitors—the whole range of market dynamics in short. We should emphasize that the focus is on the external environment of firm operations, considered as a multidimensional, complex and relational space.

In the foreword of Murphy (2007)'s *Competitive Intelligence: Gathering, Analyzing and Putting it to Work*, Michael Ridpath also states the importance of this form of intelligence:

> In business, as in war, politics or games, you cannot operate effectively unless you know what everyone else is doing. A general can never win a battle if he doesn't know the location and strength of his enemy. Political elections are about choice, how can a party provide the most attractive manifesto if it doesn't know the policies to which it will be compared? Imagine playing a game of chess as black when the white pieces are invisible. It would be a short game.

CI can help a firm predict future events (e.g. opportunities, threats) and inform quantitative and qualitative decision making, improving the chances of success in a competitive environment. Figure 2 shows the process of CI.

The tremendous rise in the size and availability of data has allowed firms to easily access more information. Firms continuously collect information from several data sources with the

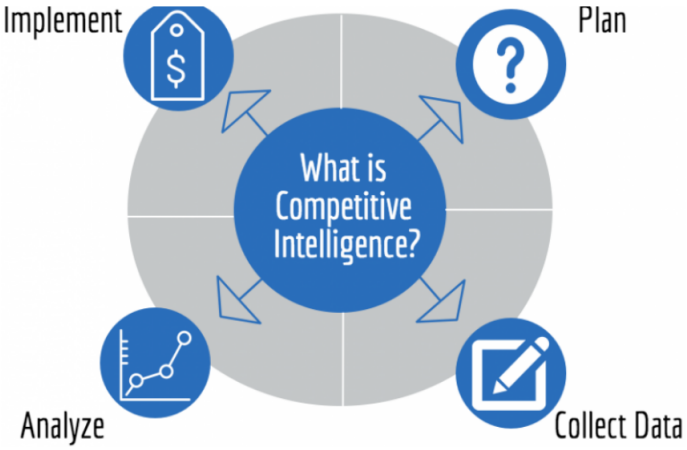


Figure 2: CI process
Source: www.engage3.com

hope of using it for CI practices. Henceforth, we will explain how data is typically gathered, highlighting several data sources, and discussing the limitations.

The data collection efforts in CI started with manual detective-style work (Deutsch, 1990; Sreenivasan, 1998). This procedure imitates the actions of our character, Ali; going to the field, observing what the competitors and customers do, taking notes of their actions. Sometimes it can be performed through surveys, field experiments, focus groups or interviews and the observations are recorded. As these forms allow direct interaction with the environment, they are trustworthy and effective. However, they can be very time-consuming and costly.

Subsequently, the data gathering process transformed into modernized, automated ways (Isaac and Lohr, 2017; Isaac, 2017). Web scraping and web application programming interfaces (APIs) are increasingly getting widespread in both industry and academia[1]. Web scraping (known also as web harvesting or web data extraction) is a technique for extracting information available in web pages. Web scraping can be done manually but it is a tedious and repetitive task. That process can be automated using a bot or web crawler, which involves fetching a web page (web crawling) and then extracting the information. Essentially, the idea is to parse, search, or reformat page content, copy the data into a spreadsheet or database and store it for later usage.

Companies such as Amazon and Google provide web scraping tools free of charge. There

[1]McLean and Samavi (2015) mention scraping accounts for 18% of site visitors and 23% of all Internet traffic in 2013.

are also open source, collaborative platforms such as Scrapy. However, sometimes the web page links and structure may change and the automated crawling and parsing process has to be re-coded again. Consequently, web scraping can be difficult and time-consuming also. For these reasons, there are also paid services such as PromptCloud that provide maintenance of data feeds and training.

Web scraping provides up-to-date, accurate information if it is done repeatedly, on a real time basis. In addition, some online services provide APIs, often used by developers to obtain third party data for their own software application or web service. Nevertheless, the legality of using data from company websites still remains as unsettled area of law. Caution has to be exercised not to violate copyright of the website's parent company. All in all, despite the significant benefits web scraping and APIs offer, the usage can be restrictive due to the legality issues.

Social media platforms, more specifically consumer review websites (e.g. Facebook, Twitter, Tripadvisor, Yelp, Glass Door) provide large data sets to learn competitor outcomes and also customer tastes. The taste aspect is especially challenging to understand, often requiring one to delve manually into customer reviews for learning customer sentiments towards products and competition.

Social media data can also reveal information on how many people follow the competitor vs. our firm in social media, how good/bad the reviews are for the competitor vs. our firm, what is competitors' marketing strategy and which demographic groups they target. In addition it can provide partial information on customer tastes.

Thus, data, freely available online, may have abundant applications in business operations. Among many others, monitoring competitor prices is essential for price comparison in the market; gathering demographic information provides population density, age and income levels useful in demand prediction for a firm considering to open a new facility.

In addition to above, there are also automated data-feeds almost universally subscribed by firms in several industry sectors to obtain secondary (external) information on competition. For instance, Smith Travel Research (STR), an American company founded in 1985, tracks demand and supply data for the hotel industry.

Initially, the company developed a Census Database consisting of names, addresses and phone numbers. Later, the founder, Mr. Smith, was contacted by the Holiday Inn hotel chain with a request for market share reports. Based on this, in 1988 the company created the first Smith Travel Accommodations Report (STAR) — which provides performance data on Av-

Date	DOW	Occupancy		ADR		RevPAR	
		My Prop	Comp Set	My Prop	Comp Set	My Prop	Comp Set
01/01/2015	Thu	73.8	80.2	214.97	294.95	158.57	236.54
01/02/2015	Fri	83.4	76.9	185.37	264.68	154.63	203.59
01/03/2015	Sat	83.2	66.9	193.51	249.54	160.94	167.06
01/04/2015	Sun	54.7	46.3	193.53	253.53	105.87	117.39
01/05/2015	Mon	60.6	43.3	197.58	263.45	119.82	114.09
01/06/2015	Tue	67.3	44.4	197.15	255.31	132.74	113.23
01/07/2015	Wed	66.3	48.8	193.01	255.38	128.04	124.66
01/08/2015	Thu	70.0	48.5	205.94	249.67	144.26	121.18
01/09/2015	Fri	69.8	46.5	190.92	227.09	133.26	105.63
01/10/2015	Sat	84.4	49.1	172.61	218.93	145.69	107.48
01/11/2015	Sun	52.7	43.7	181.06	230.80	95.46	100.75
01/12/2015	Mon	83.7	60.3	193.42	251.00	161.82	151.23

Figure 3: Daily performance data provided by STR in hotel industry

erage Daily Revenue (ADR), occupancy levels and Revenue per Available Room (RevPAR) against a self-selected competitive set (see, Figure 3). Nowadays, the benchmark report can be created quite flexibly: for instance each firm participating in the programme can select up to four competitive sets, each set consisting of several hotels. STR also provides guidance on the creation and evaluation of the competitive sets based on comparison measures of price, occupancy, revenue, distance, year of establishment and room count.

This information sharing mechanism works on a mutual participatory information-sharing basis. Hotel owners (both chain headquarters and independent ones) share their performance information and, in turn, every month they receive a detailed benchmarking report at a daily, weekly and monthly level. Today, more than 50,000 hotels participate in the STR programme, amounting to a total of over 6.8 million rooms.

Various examples of this kind of secondary data sources exist. Pegasus Solutions also operates in the hotel industry and gather similar forms of data as STR; Aena in Spain provides volume of passengers traveling at the airports to the airline industry; Context for Electronic Records Management (ERM) gives aggregated sales by region or period as chosen by the firm for the retail and supply chain industry; FCA provides Product Sales Data (PSD) in the mortgage industry.

Google also collects detailed information on businesses such as restaurants, bars and shopping centers to predict demand approximately. The “Popular Times" feature of Google Maps is launched in time for Black Friday event so that customers could screen the crowd intensity in the shopping centers on a real-time basis. Google uses GPS location signals and wireless connections to track the location of the customers who have Google Maps installed on their

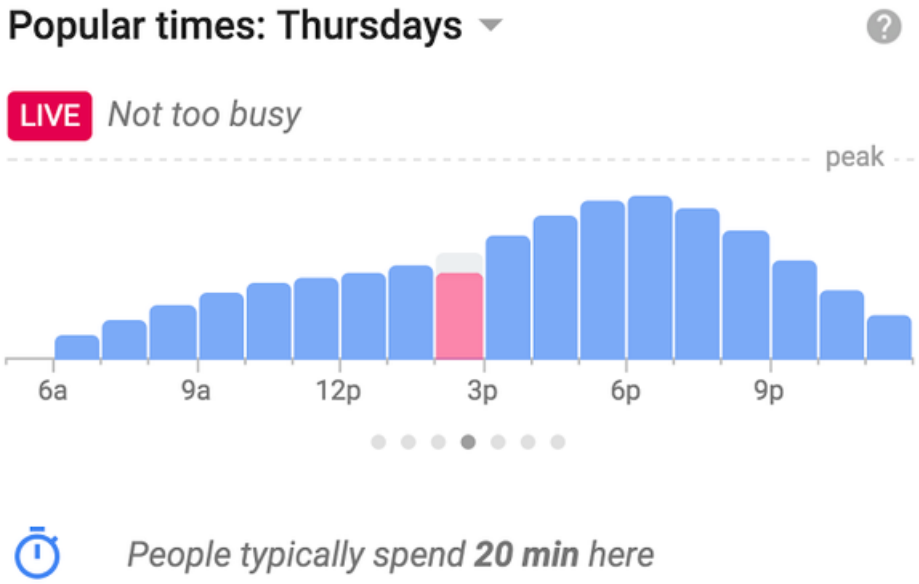


Figure 4: Popular Times feature of Google Maps

mobile devices. This data, anonymously gathered, is accumulated over weeks, months and years. The Popular Times feature is designed to inform how busy a facility is during different times of the day, even on a real-time basis and predict how long an average user would stay at a certain facility (Figure 4).

These third party sources offer performance measures to understand the external market conditions and benchmark own firm results against the competitor. All this external information, which is expected to be gathered under legal and ethical terms, is often combined with internal information and then stored in the corporations' data warehouses.

Nowadays, many firms gather this form of data and it needs to be turned into something useful and actionable. After analyzing the benchmarking reports and public sources, a strategy could be formed to maintain our firm's competitive advantages and strengthen its weak aspects. Hence, benefits can be gained by combining the intelligence information with strategic decisions.

Although CI could be quite beneficial for firms, it has been revealed that only half of the companies actually use the data operationally (Gilad and Fuld, 2016). In general, CI is not being utilized to the full extent but managers often use CI data primarily for benchmarking and performance evaluation purposes. Many firms seek consulting services for getting insights from data with the goal of adding value to their operations (Kamal, 2012; Ramakrish-

nan, 2017).

A site selection consultant interviewed by Phelps and Wood (2018) explains their role on helping a business to make a location decision:

> The industry really developed because there was a black box…Information at that time was what we really traded on. Then, since the Internet, information is available to anybody. So really we are just trading on knowledge on how to do these projects and how to do them without making a mistake… The reason they bring a professional in is to draw out what is really important to the success of a particular facility. Really what we do is to help guide them through the process.

This illustrates that the decision making problem is not only about data availability but also how to generate useful knowledge from it.

In the context of digital technology, one of the few companies that helps integrating CI in business operations is App Annie. Founded in San Francisco in 2010, it helps mobile app developers build an app and increase the chances of its discovery among roughly 1.6 million apps on Google Play and 1.5 million on the Apple App Store. The company offers analytics on app store, advertising, eBook store (free), market estimation and intelligence (paid) to help businesses get informed on possible marketing and investment strategies. For instance, it provides competitor data to app developers and offers assistance in strategic decision making such as what are the best keywords for tagging the app or the best periods to run a campaign.

CI is also becoming a professionalized field of specialization within the business sector. This is illustrated by the existence of *Strategic and Competitive Intelligence Professionals* (SCIP), a global nonprofit community that gathers experts from industry, academia and government to support firms on emerging issues and conduct CI research.

One of the reasons for poor adaptation of CI practices is that business models often rely on "detailed" data. Sales and price data of each firm is essential to analyze customer behavior in detail. However, those third party sources do not give raw competitor data to their clients but only reports with aggregated data. For instance, STR does not provide sales information, but aggregates it in the form of occupancy levels at each hotel. Hence, the main difficulty is how to exploit this limited (aggregated) information intelligently to obtain a more detailed understanding of competition, firm operations, and contribute to better decision-making.

In addition, using social media platforms to infer customer tastes is always a challenge. For example, are customer reviews true indicators of customer tastes within a given service based

industry (e.g. restaurant, hotel)? More precisely, can some other factors (e.g. price) play a role in customer review behavior so that the taste is not the sole criterion? If so, how can firms account for it? In a similar manner, are customer reviews representative of the general customer opinion? Running a survey, focus group or market research can disentangle the effects of each factor, but these are time consuming and costly. Even then, the market is dynamic, how can we rely on them after a period of time? All the hassle would have to be repeated.

In this thesis, we address these challenges to make strategic operational decisions. Given the nature of the problem, it is very likely that the status quo will remain unchanged forever. Our approaches illustrate the benefits CI can provide in several areas of operations management: revenue management, capital budget decisions and social learning.

In the first chapter we study customer behavior models to understand the importance of different product features such as price, location and schedule on purchase decisions. In a competitive market, firms need to assess the market, competition and optimize their operations accordingly. So, it is essential for firms to identify what type of customers come to the market and purchase the products. Going one step further, firms should know the latent needs of the customers so that they can adapt their strategies dynamically; use it to change the prices constantly with reference to the market prices and use this to offer better customer service.

Specifically, we consider the following generic estimation problem: A firm in a competitive market sells a single product over a finite selling season. It can only observe its own purchases. It wishes to estimate market-size and price-sensitivity, but based only on own data they are impossible to estimate. However in many revenue management and retail settings there is competition and firms can easily retrieve competitor price information. This by itself is not sufficient for identification either, but what if firms have access to aggregate competitor demand information, as possible in many industries (such as the STR report in the hotel industry). We exploit this to estimate not just market-size and price-sensitivity but also a competitor attractiveness factor.

The second chapter is on operational use of public data for facility location decisions in service based industries (e.g. restaurants, hotels) taking current and future competition in mind. Recently there has been considerable research activity on retail facility location decisions based on optimization models (Rogers, 1984; Birkin et al., 2002). We address two challenges in this stream of research: estimation of the demand model and incorporation of future competition. We demonstrate our findings for a restaurant chain that makes a location and design

decision: where to locate a new restaurant, of what type, and in which price range along with its other design features.

Site selection and demand forecasting research has also been important for superstore allocations (Wood and Browne, 2007) — carrying similar characteristics to that of the restaurant business. Firstly, it is taste-based. Secondly, there is a considerable fixed cost associated in the location decision. Therefore, accuracy in site selection is crucial (Wrigley, 1996). Thirdly, the success of the retail facility depends on expenditure levels of consumers, so it requires geo-demographic modeling.

To account for all these multi-dimensional aspects, we combine Yelp review data sets, with demographic, geographic and restaurant inspection data to build a model of demand. Lacking competitor demand information, we consider external sources such as OpenTable and volume of customer reviews provided by Yelp to calibrate our demand model. In this structural model we account for customer tastes, unobservable quality aspects, demographics, and size to estimate how the market will shift or grow once new competing restaurants locate nearby.

Subsequent to the estimation phase, we feed the estimates into an optimization framework and provide a tractable model that incorporates competitive effects to maximize the long-term profit.

Lastly, in the third chapter we investigate online reputation mechanisms. Customer decisions are usually thought to be affected by two factors: price and quality. Most of the times, price is discrete and known in advance. In other words, online marketing enables customers to learn about prices of goods faster and easier. However, consumers still lack full information on quality of products. Certainly, this information is more difficult to obtain than prices and it remains ambiguous prior to purchase. Online reviews reveal partial information on experience-related product attributes such as quality. Therefore, there is a great need to understand online reviews better.

In this study, we collaborate with a hotel chain in U.K. Firstly, we analyze the impact of online customer review ratings on sales using customer bookings data. Later, we ask the following question of the data: are reviews affected by price or value or neither, i.e. once customers pay to stay, do they treat what they had paid as a sunk cost and review based on only their experience, or are their reviews affected by the price they pay? Finally, we cross-validate our findings with a survey and investigate deeper on whether online reviews transmit the general customer opinion.

All in all, exploiting CI data seems critical in an extremely competitive environment. In

this thesis, we combine methodologies from engineering, econometrics, machine learning and optimization to empower the usage of data to gain a competitive edge. Our work shows there are significant opportunities to improve the models. We present real-world examples for the three topics and demonstrate how we tackle those challenges specifically for each of them.

Contents

List of Figures

List of Tables

1

EXPLOITING COMPETITOR INTELLIGENCE DATA FOR DEMAND ESTIMATION

Estimating a suitable model of demand and forecasting is central to revenue management and pricing. Recent research efforts have focused on estimating behavioral choice models of demand based on transactional data, addressing in particular the difficulty posed by not being able to observe no-purchases in most commercial situations.

This study presents a new econometric method for estimation based on—widely available—competitor intelligence data and also addresses two important and difficult gaps in this stream of research: (1) estimating with competitor effects (2) estimating when the firm sells a single product. We note that the aggregated competitor intelligence data is commonly available in many industries; prime example being the STR report in the hotel industry.

With access to such data, we show that the two problems we mentioned can in fact be solved together—marginal information can be used to estimate price-sensitivity along with a competitor location attractiveness parameter, even for the case where each firm sells a single product. A noteworthy complication in the real-world is that even competitors' initial capacity is uncertain because of private group sales, and we address this issue also via instrumental variables. We present numerical and recovery performance on both synthetic data as well as on real hotel bookings data.

1.1 Introduction

Revenue management (RM) operations typically proceed in the following sequence (i) estimation of the parameters of the model (ii) forecasting demand on a day-to-day basis and (iii) optimization of the prices or allocations. In this sequence estimation and forecasting are crucial steps as managers can compare the model output with what they observed (as opposed to comparing recommendations with an unknown and unobservable optimal price). Indeed in many RM departments forecasting accuracy gives management confidence in using the system's price recommendations, so for any implementation to be accepted it plays an oversized role in gaining credibility for the system. In contrast, "optimal" prices are more difficult to pin down and verify, at least in the short term, and play a smaller psychological role.

In addition to managerial confidence in the system, inferred choice parameter estimates are used as an input for the optimization of the controls. Biased or unstable estimates can result in significantly less revenue predictions, and eventually are very likely to cause serious problems in the implementation. Despite this, most of the early work in RM deals with optimization, assuming accurate customer choice parameters. Estimation of consumer choice behavior is a recent active research area initiated by Talluri and Van Ryzin (2004) for the single-leg RM in a methodologically complete way.

A fundamental problem in the estimation of RM models in practice is that one cannot observe customers who consider our products but do not purchase—either because the product has become unavailable, or the price is beyond their willingness-to-pay for the product. This unobservability can happen for instance if the product is sold via third parties, or in a physical retail store. In such situations the firm has a record only of purchases and the problem of estimating the market size becomes very difficult, especially so if the market size varies over time; the market size and price-sensitivity effects get confounded and it becomes impossible to separate them.

In this study we address the following complicating issues of RM estimation without observing no-purchases:

1. The manager setting prices may have some signals or knowledge of the future demands and be setting prices accordingly but the estimation procedure does not have access to this information. So for instance, the market size for a particular day might be high due to special local event, and the prices are high for that reason—demand will also be high and this throws off the estimation, even leading it to wrong signs of the parameters.

This problem is especially acute in RM as the analysts change prices after observing an initial period of demand, which in many cases is a strong signal of the market size for a specific date.

2. Competitor effects such as their price, brand and location differences can be influencing sales. While competitor prices are usually public information and can be incorporated relatively easily into the estimation, the complicating factor is that we can almost never observe their sales. In many settings firms also have private sales[1] which has two implications—(i) we are not sure of their initial capacities, (ii) a high competitor price may signal either a belief in high demand or a large number of private sales.

3. The network nature of the products, such as flight itineraries or three-night hotel stay starting on a specific date, create dependencies across the inventories as well as sales. So even if the demands for each individual product are assumed to be independent of other demands, estimation has to be done at the network level adding to both computational and estimation complexity.

The above-mentioned estimation challenges are insuperable unless we have access to some form of new data, and this is precisely what we exploit in this paper. It so turns out that there exist automated data-feeds in several industries to obtain information on competition. A prime example of this is the STR report that gives performance data on price (ADR), occupancy levels and revenue-per-available-room (RevPAR) against a competitive set in the hotel industry (Figure 1.1). The occupancy data aggregates all the multi-night stay demands sold at different prices along with the competitors' private sales.

Similarly, in the airline industry in many countries the airport authority resells aggregated data on airport traffic, by airline and market e.g. Aena in Spain. In the retail electronics industry for instance, private companies (such as Context in the U.K.) deal with such aggregated competitor sales information. Recently, Google also launched "Popular Times" feature to inform how busy a facility is on a real-time basis. This competitor intelligence data contains marginal aggregated information on competitor sales.

It is important to note this competitor intelligence data is not a new exotic one-of-a-kind data source but nearly universally subscribed to in these industries. Till now, however, they

[1] In this study we define private sales as sales that arise to the firm through a private channel, either as part of a long-term contract or negotiated sales.

Date	DOW	Occupancy		ADR		RevPAR	
		My Prop	Comp Set	My Prop	Comp Set	My Prop	Comp Set
01/01/2015	Thu	73.8	80.2	214.97	294.95	158.57	236.54
01/02/2015	Fri	83.4	76.9	185.37	264.68	154.63	203.59
01/03/2015	Sat	83.2	66.9	193.51	249.54	160.94	167.06
01/04/2015	Sun	54.7	46.3	193.53	253.53	105.87	117.39
01/05/2015	Mon	60.6	43.3	197.58	263.45	119.82	114.09
01/06/2015	Tue	67.3	44.4	197.15	255.31	132.74	113.23
01/07/2015	Wed	66.3	48.8	193.01	255.38	128.04	124.66
01/08/2015	Thu	70.0	48.5	205.94	249.67	144.26	121.18
01/09/2015	Fri	69.8	46.5	190.92	227.09	133.26	105.63
01/10/2015	Sat	84.4	49.1	172.61	218.93	145.69	107.48
01/11/2015	Sun	52.7	43.7	181.06	230.80	95.46	100.75
01/12/2015	Mon	83.7	60.3	193.42	251.00	161.82	151.23

Figure 1.1: Daily performance data provided by STR

have found little operational use, but examined instead primarily for bench-marking and performance evaluation of managers. We hope our research will encourage firms in these industries to make operational use of the data to better estimate and forecast their demands.

Now, despite this new source of operational data, it is not at all obvious how the data can make the estimation problem any easier. Surprisingly, we show that the very difficulties we mentioned above actually help identify the model (up to a small number of candidate solutions as we explain later) and obtain accurate estimates where all other methods fail. This is the case even if the firm sells a single product so even estimating the price-sensitivity parameter is problematic.

Specifically, the network effects create dependencies in the sales random variables even if the underlying market size random variable are independent. So under a very weak assumption that the demands for days that are sufficiently far apart (such as demands for Monday stays vs. Thursday stays) are independent of each other, we can pin down the parameters with moment conditions based only on sales. For the market size parameter, we do not get complete identification but only a partial one, up to a small number of candidates. To avoid choosing amongst these, we propose a new objective function for the estimation that is based on the insight we develop from our analysis and show that it works extremely well on synthetic and real data sets.

Our contribution is threefold. First, we exploit aggregated information on competitor sales to develop an enriched demand estimation model that accounts for unobservable location and incorporates network effects. Second, we use own private bookings as instrumental variables to correct for endogeneity bias and obtain an accurate measure of price elasticity even

in the troublesome and till now unresolved case in which the firm sells a single product. Third, we develop a novel way to estimate market size based on an independence assumption only exploiting the dependence created by the network nature of the products.

In the next section, we review the literature. In § 1.3 we describe the model and motivate our study. In § 1.4 we present our estimation methodology. In § 1.5 we develop the best-response pricing model to generate synthetic data. In § 1.6 we illustrate the performance on both synthetic and real hotel bookings data. Finally, in § 1.7 we summarize our findings.

1.2 Literature review

RM is based on models of demand that one can estimate operationally. Understanding customer preferences and purchase decisions is the key element to estimate demand. Developments in technology, especially online shopping via the Internet results in a trove of data on customer behavior. For this reason discrete-choice models, specifically the Multinomial Logit (MNL) model, are often used as they are parsimonious and relatively easy to estimate and operationalize the trade-offs among product attributes.

A typical assumption in older marketing and economic literature is that market size is known *apriori*. Nevo (2001), Besanko et al. (1998) and Berry (1994) are some examples of exogenous market assumption in the marketing literature. This is unrealistic in the RM context (as well as in many marketing situations) as the firms do not have data on those who purchased from competitor or did not purchase at all. A number of recent research papers (Talluri and Van Ryzin, 2004; Talluri, 2009; Vulcano et al., 2012; Newman et al., 2014) have recognized the importance of this problem.

Ratliff et al. (2008) follow a heuristic method based on demand mass balance equations of Andersson (1998) to estimate demand, spill and recapture across multiple flights and fare classes. Vulcano et al. (2012) propose to estimate choice probabilities based on primary (or first-choice) demand, that basically captures demand if all products were offered. Li et al. (2014b) provide evidence of strategic consumers and examine revenue implications via counterfactual analysis.

Talluri and Van Ryzin (2004) use the expectation-maximization (EM) algorithm to the incomplete-data maximum-likelihood estimation model and provide an exact analysis of the capacity allocation problem for a single-leg, multiple-fare RM problem. This allows them to simultaneously estimate arrival rates and parameters of a MNL choice model. Vulcano et al.

(2010) implement the generic EM framework described in Talluri and Van Ryzin (2004) for an airline market and assess revenue improvements of around $1\% - 5\%$.

Among the demand untruncation/uncensoring methods, the EM algorithm is the most commonly used statistical technique when the data is missing. Kök and Fisher (2007) generalize EM algorithm including a substitution matrix in the context of assortment planning. However, all these procedures suffer from a fundamental unidentifiability problem (pointed out in Talluri and Van Ryzin (2004) as well as Vulcano et al. (2010)), which essentially means there are a continuum of values —purchase probability and market size — that gives the same sales amount. However, those procedures will only find one such pair and the implications on pricing strategy might be completely different. Moreover, EM tends to be slow and we only know that it converges to a local minimum (Wu, 1983), negating our confidence in its statistical properties.

In order to deal with the indeterminacy problem, Talluri (2009) develops a risk-ratio based heuristic estimation method based on convex risk criteria. In the first step, conditioned on observed purchase data, customer choice parameters are estimated (e.g. price elasticity), which is consistent and tractable for the MNL model. In the second step, using a risk-ratio criteria, the total number of arrivals is estimated. The method is theoretically appealing and numerical simulations give promising results. In a similar vein, Newman et al. (2014) propose two-step estimation method, assuming the market consists of two or more products of own firm (thus observable) and no-purchases. The arrivals occur in accordance with a homogeneous Poisson process. The first step is the same as Talluri (2009). In the second step, total arrival rate is calculated using the variations in observed sales over time directly from the log-likelihood function. However, this step is not generally a concave optimization problem and so it is not possible to ensure identifiability (therefore the consistency) of the market size (equivalently, no-purchase) parameter.

Another complicating factor that leads to biased estimates is endogeneity. Endogeneity arises when regressors are correlated with the error term. Unobserved product attributes (e.g. quality or reputation), marketing related practices (e.g. promotions, coupons) can be correlated with price, so this is a considerable issue in price-elasticity estimation. Fisher et al. (2017) focuses on competition based dynamic pricing in retail sector. They use own and competitor stock-outs as a valid source of variation to the consumer choice set to add extra moment conditions when competitor sales is not available. They use randomized prices in a field experiment setting to address the price endogeneity issue. Based on accurate measures of consumer choice

parameters, they solve the best response pricing algorithm. However, field experiments are not always feasible, especially when the conditions are changing. Moreover, not many firms can generate enough data using field experiment. In this regard, using instruments to deal with endogeneity is quite efficient—if we can find the appropriate instruments.

As our domain is hotel industry, the estimation has to be done integrating network effects. This is another complication of our study as competitor sales are not available at day and length-of-stay level but rather provided in the form of occupancies. There is another stream of research called "Network Tomography" (NT) that also deals with estimation based on marginal aggregated data. NT problem infers the node-to-node traffic intensity between all source-destination pairs of nodes from repeated measurements of traffic flow on the links of a network. Tebaldi and West (1998) use Bayesian inference and Vardi (1996) uses method of moments based approach to infer node-to-node road traffic. In § 1.3, we describe the connection of our study to NT in more detail.

An early influential work in the econometrics literature that focuses on demand estimation of discrete-choice models for differentiated products in an oligopoly market is by Berry et al. (1995) (BLP). BLP gained popularity due to the use of aggregate data, heterogeneity in its demand model and the subsequent estimation of price elasticities accounting for endogeneity. As mentioned by (Berry et al., 1995), the endogeneity problem is especially challenging in the case of nonlinear problems. Therefore, BLP first linearize the choice model and then use a contraction mapping embedded in an IV estimation. BLP however assumes the market-size is known—apart from the aggregate sales data of competitors— and has recently come under criticism for (i) its use of equilibrium conditions (Ackerberg et al., 2007), (ii) the possibility of converging to the wrong parameters, if at all it converges (Knittel and Metaxoglou, 2014) and (iii) its exceptionally slow numerical performance (Su and Judd, 2012). The key differences of our study from BLP and Newman et al. (2014) is given in Table 1.1.

In the next section, we introduce the problem, provide an overview of discrete choice models, and motivate our study in relation to both estimation and NT literature.

1.3 Model problem statement and motivation

We briefly discuss the problem and describe the choice model to show how consumers choose among differentiated products. Later, we motivate our study referring it to both estimation and NT literature.

BLP, 1995	Newman et al., 2014	Our study
m markets, j products (e.g. each model-year for automobiles)	Own firm products (e.g. different types of hotel rooms)	Own firm and aggregated competitor products (e.g. competitor occupancy)
Individual level-choice data (or aggregate and individual)	At least 2 products	One product for each firm
Heterogeneity effects		Network effects BLP association; market is days, product is LOS
Assumption on the market size (no estimation of β_0)	No competitor; it is assumed to belong to no-purchases	Competitor and no-purchases are treated separately
Endogeneity; GMM method, fixed-point alg., IV variables		Endogeneity; GMM method, IV variables

Table 1.1: Detailed comparison of similar references

1.3.1 Setting

While our proposed method is applicable to airlines or rental cars or any industry which practices network RM, to make the problem specific and to be in sync with the subsequent dataset, we describe the problem and set the notation in the context of a hotel.

Hotels sell room stays starting on a specific date of multiple durations (called length-of-stays (LOS), with most ranging from one to five nights). For each starting date and LOS, the hotel may be selling multiple products—either physically different such as room types, or virtual RM products differentiated by sale restrictions. Alternately—and the most difficult for estimation purposes—the firm could be selling a single room type and single RM product but just be changing the prices of these products over time. In the rest of the study, when we say product, it refers to a specific starting date and a LOS.

The hotel also faces competition, from a set of neighboring hotels. For simplicity we assume there is a single competitor[2]. The hotels compete but the attractiveness of the competitor is distinct, so customers may prefer our firm even if our price is higher. We aim to estimate this attractiveness from the data, i.e. the customer's perception of it reflected in their actions.

To keep notation simple we assume each firm sells a single product (for each LOS). Multiple products actually make the estimation problem easier as the price-sensitivity parameter can be estimated conditioned on the purchases. So we deal with the more challenging version of the problem.

We denote own hotel by the superscript o, and the competition by c, and the no-purchase

[2] If the firm is dealing with a larger competitor set, one can average or otherwise create a reference competitor

option by 0. So for each day and LOS, there are at most $\mathcal{N} = 3$ options for the customer.

Sales are of two types, private or general public. Private sales occur through long corporate relationships (eg: airlines negotiate rooms for their crew) or the sales and marketing department (for instance, a negotiated corporate retreat or a wedding) and their primary feature for our purposes is that they happen independently for our hotel and the competition's. The presence of private sales significantly complicates the estimation of choice parameters. We let G^n_{dl} be the random variable where n indexes the three choices o, c or 0, d is the starting day of the stay, and l is the LOS. We assume private sales occur before the public sales.

Public sales are retail sales from individuals who consider both our product and the competitor and either decide to purchase one of the options or decide not to purchase and exit the market (no-purchasers). We let S^n_{dl} be the random variable for public sales for hotel n. The sales observed at any hotel is the sum of its private and public sales. Public sales are assumed to be driven based on a customer choice model that incorporates price and location attractiveness.

A consumer on day d to stay for l nights obtains utility, U^n_{dl}, from purchasing the product from hotel n ($n = o, c$ or 0, the outside option)

$$U^n_{dl} = \beta_p p^n_{dl} + \beta_a + \epsilon^n_{dl},$$

where β_p is the price elasticity parameter, p^n_{dl} is respective hotel's *average* prices on product (d, l) —also known as average daily rate. β_a is the parameter that measures the relative location attractiveness of competitor hotel compared to ours. We can only identify the differences across these hotels' attractiveness, so we normalize our hotel location attractiveness to 0. Specifically, the case of leaving the market empty handed gives the utility

$$U^0_{dl} = -\beta_0 + \epsilon^0_{dl}$$

where β_0 is the parameter to measure the weight of the outside option. We denote by $\boldsymbol{\beta} = (\beta_p, \beta_a, \beta_0)$ unknown vector of parameters. It is also possible to enhance the model including demographics and some other observable characteristics, but we keep it clean for clarity.

Customers are utility maximizers, they choose an option with the most possible utility among their choice alternatives. We use the MNL choice model (Ben-Akiva and Lerman, 1985), by far the most frequently used choice model in implementations, to describe a cus-

	Public sales					Private sales					Occupancy
Date	LOS=1	LOS=2	LOS=3	LOS=4	LOS=5	LOS=1	LOS=2	LOS=3	LOS=4	LOS=5	
01/01/15	25	19	17	9	4	2	1	3	1	1	82
01/02/15	51	50	21	7	8	1	0	0	1	0	194
01/03/15	92	23	7	6	5	1	0	0	2	0	258
01/04/15	22	18	10	2	6	10	0	0	0	0	163
01/05/15	30	25	14	16	5	1	1	0	11	0	180
01/06/15	36	24	15	4	10	10	0	0	1	0	211

Figure 1.2: Sales transactions for our hotel

tomer's choice behavior. The MNL model is based on the assumption that ϵ's are i.i.d. Gumbel (or double-exponential) with mean zero and scale parameter one for all the alternatives in the choice set. Under this assumption, the choice probability for option n is given as

$$\mathcal{P}^o_{dl}(\boldsymbol{\beta}) = \frac{e^{\beta_p p^o_{dl}}}{e^{-\beta_0} + e^{\beta_p p^o_{dl}} + e^{\beta_p p^c_{dl} + \beta_a}}, \tag{1.1}$$

and for the competition

$$\mathcal{P}^c_{dl}(\boldsymbol{\beta}) = \frac{e^{\beta_p p^c_{dl} + \beta_a}}{e^{-\beta_0} + e^{\beta_p p^o_{dl}} + e^{\beta_p p^c_{dl} + \beta_a}}. \tag{1.2}$$

Note that $\sum_{n \in \mathcal{N}} \mathcal{P}^n_{dl}(\boldsymbol{\beta}) = 1$ holds $\forall d, l$.

Now the structure we impose is that market size comes from a specific random process for each day-of-week to stay for $l = 1, \ldots, L$ nights, specifically can be denoted by $M_{\omega_d l}$ where ω is a day-of-week $\omega = 1 \ldots 7$. Then, public sales at any hotel n on date d to stay for l nights can be obtained in a market-share model by multiplying total arrivals to the market with choice probability of the respective hotel

$$S^n_{dl} = M_{\omega_d l} \mathcal{P}^n_{dl}(\boldsymbol{\beta}). \tag{1.3}$$

We observe our own sales, s^o_{dl}, but we do not have information on competitor sales, $\tilde{s}^c_{dl}$.

The challenge in estimation of the parameters, $\boldsymbol{\beta} = (\beta_0, \beta_p, \beta_a)$ is the lack of sales data on competitor side and unobservable no-purchases.

1.3.2 Data

We have all the sales transactions for our firm; public, s^o_{dl}, and private, g^o_{dl}. A sample is shown in Figure 1.2. Of the competitor, we can observe the prices of each product—indeed, at every

point in time, as they are public—but we do not have data on the realization of either their private sales, $\tilde{g}^c_{dl}$ nor the realization of their public sales, $\tilde{s}^c_{dl}$. Consequently do not know their capacities for public sales either at the beginning of the public sales horizon.

The marginal competitor data that we mentioned earlier aggregates the competitor demands of both public as well as private ones for the different LOS. So, we have access to only the daily occupancy levels of the competitor, y^c_d, which satisfies the following system of equations

$$y^c_d = \sum_{d'=d-L}^{d} \sum_{l=d-d'+1}^{L} \underbrace{s^c_{d'l}}_{\text{Public}} + \underbrace{g^c_{d'l}}_{\text{Private}}. \tag{1.4}$$

This brings two issues upfront: aggregation of product instances by LOS and the randomness in remaining capacity of competitor. The first one creates limitation of data, NT methods have been used to deal with aggregation of data, as we will see in the next section. The latter one is problematic due to unpredictable nature of private sales, and its effect on prices. Standard estimation techniques of customer choice literature fails because of the unobservable private sales as well as endogeneity. Next, we discuss the relevant literature along with its deficiency to show the necessity of this research.

1.3.3 Estimation problem summary

We summarize now the estimation problem. Our goal is to estimate the parameters $\boldsymbol{\beta} = (\beta_0, \beta_p, \beta_a)$ as well as the parameters of the distribution generating $M_{\omega_d l}$, specifically $M_{\omega l}$ which we assume for concreteness to be a normal distribution. The only data we have is a complete picture of our and competitor prices, our sales, and "occupancy" of the competitor in the history. We assume the managers set prices knowing something about the market sizes and shocks specific to a day d.

Notice that if we estimate $\boldsymbol{\beta} = (\beta_0, \beta_p, \beta_a)$, as we know our own sales, we can invert the Equation 1.3 to obtain estimates of $M_{\omega_d l}$. So the problem is essentially to estimate $\boldsymbol{\beta} = (\beta_0, \beta_p, \beta_a)$.

1.3.4 Current approaches

As mentioned in the introduction, the problem of estimating from marginals comes up in engineering applications (transportation and communication networks), where it goes by the

name of network tomography. In this section we describe their ideas along with the difficulties in solving it via the EM algorithm and the traditional method used in RM. We mentioned the BLP procedure in the introduction but we do not use it as we do not model customer heterogeneity. So we do not describe BLP here. In any case, as we see in the Markov Chain Monte Carlo (MCMC) procedure of Tebaldi and West (1998) below, such methods are too slow for operational use.

a Network Tomography (NT)

Our problem of estimating from marginal data is reminiscent of estimating origin-destination (O-D) flows based on link-level traffic. In traffic systems, directly measuring the O-D matrix is often not feasible, but link traffic measurements are relatively easy to obtain. The similarity to our application is that competitor occupancies data y is akin to link-level measurement and we want to estimate the O-D itinerary (stay start, LOS pair) sales s. The relation is given by the linear form

$$\boldsymbol{y} = A(\boldsymbol{s} + \boldsymbol{g}) \tag{1.5}$$

where $\boldsymbol{y}$ is occupancy vector, $\boldsymbol{s}$ is public sales vector, and $\boldsymbol{g}$ is private sales vector. A denotes $r \times c$ routing matrix, where the rows, $r = D$, corresponds to days, and the columns, c, corresponds to LOS's starting on each day.

Entries of A are 1 or 0 depending on whether that stay-night instance is crossed or not. For instance, suppose we consider 3-day period and a maximum of 2 LOS, the following relation holds:

$$\begin{bmatrix} y_{d_1} \\ y_{d_2} \\ y_{d_3} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} s_{d_1,l_1} + g_{d_1,l_1} \\ s_{d_2,l_1} + g_{d_2,l_1} \\ s_{d_3,l_1} + g_{d_3,l_1} \\ s_{d_1,l_2} + g_{d_1,l_2} \\ s_{d_2,l_2} + g_{d_2,l_2} \\ s_{d_3,l_2} + g_{d_3,l_2} \end{bmatrix} \tag{1.6}$$

We arrange the A matrix sequentially for each day of arrival such that 1-night customers, 2-night customers etc. are represented. So, for instance, the first column of A is day 1 customers who stayed 1-night, the second column is day 2 customers who stayed 1-night and so on. Considering the characteristics of A matrix in the hotel instance, it has the consecutive 1's property for columns. Therefore, it is a totally unimodular matrix, however it is not invertible.

So, there should exist many feasible solutions to this system of equations since, in general, r<c. This implies the system $y = A(s + g)$ is under-determined. Specifically, for hotel network the maximum number of columns of A matrix can be $c = D \times L$, where L is the maximum LOS.

b EM based algorithms

There have been several approaches to solve this problem in NT literature using the EM algorithm. Firstly, it can be solved using maximum likelihood procedure. The incomplete data likelihood function can be written as follows:

$$\begin{aligned} L(\boldsymbol{\beta}, \boldsymbol{M}) =& P((s_{dl} + g_{dl}) \in \mathcal{Y}_d | \boldsymbol{\beta}, \boldsymbol{M}) \\ =& \prod_{d=1}^{D} \prod_{l=1}^{L} \sum_{(s_{dl}+g_{dl}) \in \mathcal{Y}_d} \frac{\exp^{-\boldsymbol{M}\mathcal{P}_{dl}(\boldsymbol{\beta})} (\boldsymbol{M}\mathcal{P}_{dl}(\boldsymbol{\beta}))^{s_{dl}+g_{dl}}}{(s_{dl} + g_{dl})!} \end{aligned} \tag{1.7}$$

The likelihood function given in (1.7) is hopelessly difficult to maximize. In fact, it is even difficult to evaluate as one has to sum over all feasible solutions of a linear system (Vardi, 1996). The EM algorithm suffers from the same issue; expectation step is computationally impossible in addition EM leads to terrible solutions, even can converge to the local minimum (Vanderbei and Iannone, 1994). Similarly, Vardi (1996) shows that depending on the starting point, the algorithm may converge to a non-MLE point.

To mitigate the difficulty in implementing the EM algorithm, Vardi (1996) deals with the same problem based on method of moments under the Poisson model, which is basically equating the sample's first and second moments to their theoretical values to obtain linear system of equations to estimate the parameters. Although, this idea works well for engineering problems that allows repeated measurements, the need for a large number of samples makes it useless in business cases. Indeed, specifically, in case of hotel data it is even worse since we have only a sample of one for the whole network. Airlines do have more samples but they are still not adequate for these approaches to work properly. Tebaldi and West (1998) use Bayesian approach with iterated simulations methods or MCMC implementation using the link counts from a single measurement interval. Yet, it is extremely slow since it runs MCMC to convergence at each step of the inner loop.

To summarize, these methods are not satisfactory since they do not give an accurate estimate of the parameters when using small samples in practical cases (usually a single season's

Parameter	True Values	Mean Estimates	Std. Dev. Estimates	Percent Error in mean	Coeff. of Variation
β_{king}	0	—	—	—	—
β_{suite}	0.4	0.39929	0.00949	0.18	0.024
β_p	-0.01719	-0.01717	0.00012	0.12	0.007
β_0	-1.3	-1.29858	0.09368	0.11	0.072
M (per day)	40	39.94145	0.65632	0.15	0.016

Table 1.2: Average results of (Newman et al., 2014)'s algorithm over numerous runs when there is no competitor in the market. Our product purchases are 0.10.

worth of data). We use the methods of Vardi (1996) and Tebaldi and West (1998) as a benchmark and test their performances on both synthetic and real data set.

c RM approaches ignoring competition

Due to today's abundance of data, ignoring competition in the choice set (or, considering it as belonging to no-purchase outcomes) is quite naive and may lead to biased estimates. We carry out a very simple empirical study closely related to Newman et al. (2014) in order to reveal the value of incorporating competitor information in the estimation. This is a simplified version of Newman et al. (2014)'s simulation in terms of product types. Following their construct, we assume there are at least 2 products of our firm. The simulations are applied to 50 distinct data sets, each set consisting of a sample size of 1000 to guarantee empirical unbiasedness. In the first case, there is no competitor in the market as in Newman et al. (2014), so we replicate their simulation results (Table 3 in Newman et al. (2014)). Accordingly, the estimates are quite accurate (Table 1.2) [3]. In the second case, we apply the same procedure to a market where competitor exists. Although, the mean estimates for product-level attributes (e.g. room type, price) can be revealed accurately using the variations between our products, the mean estimates for the no-purchase weight, β_0, and the mean arrival rate, M, are obviously biased. Also, standard deviation of those parameters and their coefficient of variation are quite high (Table 1.3). Especially, standard deviation of market size parameter is notable. This is because M is comparatively larger than β_0 in absolute terms and the bias gets amplified. Hence, this motivates us to exploit the marginal data on competitor sales in order to reveal accurate estimates of the parameters and develop enriched demand model.

[3] M is essentially the average market size parameter in Newman et al. (2014), denoted by λ.

Parameter	True Values	Mean Estimates	Std. Dev. Estimates	Percent Error in Mean	Coeff. of Variation
β_{king}	0	—	—	—	—
β_{suite}	0.4	0.39797	0.01032	0.51	0.026
β_p	-0.01719	-0.01717	0.00013	0.12	0.008
β_0	-1.3	-0.56944	1.08614	56.20	1.907
M (per day)	40	50.89057	63.38431	27.23	1.246

Table 1.3: Average results of (Newman et al., 2014)'s algorithm over numerous runs when our product purchases are 0.10, competitor product purchases are 0.60.

1.4 Estimation methodology

Our strategy for estimation is first to estimate the price-sensitivity parameter β_p and the location parameter β_a by a set of moment conditions. In this part, we discuss how marginal information on competitor sales (competitor occupancy) can be used as a source of identification to identify the price and location sensitivity parameters.

In the second part, we fix those parameters and estimate the market-size parameter β_0 using a novel idea (at least in the RM area) based only on independence of market arrivals.

In our model, we investigate price and location effect on sales relying information only on competitor occupancy, our sales and price differences. We use method of moments framework to demonstrate how we exploit the relation between competitor occupancy and our sales making use of logit function form.

Observed competitor occupancy includes both public and private sales aggregated at a daily-level. Public sales to our hotel on date d to stay for l nights are given by

$$s^o_{dl} = M_{\omega_d l} \frac{e^{\beta_p p^o_{dl}}}{e^{-\beta_0} + e^{\beta_p p^o_{dl}} + e^{\beta_p p^c_{dl} + \beta_a}}. \tag{1.8}$$

Inverting (1.8), we obtain total arrivals

$$M_{\omega_d l} = s^o_{dl} \frac{e^{-\beta_0} + e^{\beta_p p^o_{dl}} + e^{\beta_p p^c_{dl} + \beta_a}}{e^{\beta_p p^o_{dl}}}. \tag{1.9}$$

On the competitor side we have

$$s^c_{dl} = M_{\omega_d l} \frac{e^{\beta_p p^c_{dl} + \beta_a}}{e^{-\beta_0} + e^{\beta_p p^o_{dl}} + e^{\beta_p p^c_{dl} + \beta_a}}, \tag{1.10}$$

then substituting (1.9) into (1.10), we can eliminate $M_{\omega_d l}$ and obtain competitor public sales in terms of β_p, β_a parameters only (that is, no β_0).

$$s^c_{dl} = e^{\beta_a} s^o_{dl} e^{\beta_p \Delta p_{dl}}, \tag{1.11}$$

where $\Delta p_{dl} = p^c_{dl} - p^o_{dl}$ is the price difference between competitor and our hotel. Let $\hat{s}_{dl}(\beta_p) = s^o_{dl} e^{\beta_p \Delta p_{dl}}$ and $\hat{\boldsymbol{s}}(\beta_p)$ be the vector of these values arranged in the same order as the columns of the matrix A.

The presence of private sales significantly complicates the estimation of customer choice parameters. These private sales are negotiated individually by each hotel and can come for multiple units of capacity. They are not necessarily sold at a discount—the negotiations can center on utilization guarantees such as ability to cancel without penalty, conference rooms, dining etc. The customer demand is a combination of public and private sales. We insert competitor public sales (1.11) and competitor group sales into occupancy–sales equation (1.5), $y = As$, to obtain customer demand model

$$y^c_d = \underbrace{e^{\beta_a} A \hat{\boldsymbol{s}}(\beta_p)}_{\text{Public}} + \underbrace{y^{G^c}_d}_{\text{Private}} + \tilde{\epsilon}^c_d \tag{1.12}$$

where $y^{G^c}_d = A G^c_{dl}$ is competitor private occupancy and $\tilde{\epsilon}^c_d = A \epsilon^c_{dl}$ is the error term. The distribution of ϵ^c_{dl} is given by $\epsilon^c_{dl} \sim N(0, \sigma^2 = M_{\omega_d l} \mathcal{P}^c_{dl})$. Once we scale the $\hat{\boldsymbol{s}}(\beta_p)$ with a location factor we obtain total competitor public sales.

Before we present our approach, for reference sake, we use a simple non-linear regression method to estimate the parameters. It works reasonably well on clean synthetic data (if we ignore the existence of private sales) but fails on real-world data due to endogeneity.

1.4.1 Naive non-linear least squares (NLS) regression

For any vector $[\cdot]$, we denote the jth element by $[\cdot]_j$.

We observe our private sales, G^o_{dl}, but we have no information on private arrivals of the competitor, G^c_{dl}. So as a naive approach we assume a misspecified demand function which omits the presence of private sales

$$y^c_d = e^{\beta_a} [A \hat{\boldsymbol{s}}(\beta_p)]_d + \tilde{u}^c_d, \tag{1.13}$$

where $\tilde{u}_d^c = y_d^{G^c} + \tilde{\epsilon}_d^c$.

If we run a regression directly on (1.13), the term β_a factors out and its value is arbitrary. For example, if we make this small, then it appears as if occupancy of the competitor is high while its prices relative to us is high and we can get a positive sign on the price parameter. So, as one easy way to get around this problem, we suggest estimating via the occupancy ratios given for competitor hotel between consequent days. We proceed to estimate the parameters β_p and β_a in two-steps.

First step: Using the occupancy ratio between consequent stay nights d for competitor hotel we can write

$$\frac{y_d^c}{y_{d+1}^c} = \frac{e^{\beta_a}[A\hat{\boldsymbol{s}}(\beta_p)]_d}{e^{\beta_a}[A\hat{\boldsymbol{s}}(\beta_p)]_{d+1}},$$

hence β_a cancels out and in the first step we obtain an estimate of β_p. From Equation (1.13), we can identify β_p based on the variation of difference of prices and competitor occupancy.

Second step: We insert $\hat{\beta}_p$ into (1.13) to estimate β_a using

$$y_d^c = e^{\beta_a}[A\hat{\boldsymbol{s}}(\hat{\beta}_p)]_d + \tilde{u}_d^c. \tag{1.14}$$

We can identify β_p and β_a using this procedure as long as the price difference between our and competitor hotel is not constant along the horizon.

$s_{dl}^c = e^{\beta_a} A s_{dl}^o e^{\beta_p \Delta p_{dl}} = A s_{dl}^o e^{\beta_a + \beta_p \Delta p_{dl}}$ is convex in the parameters follows from the convexity of the exponential function and the linearity of the operations.

Proposition 1. *Competitor public sales, s_{dl}^c, is convex in parameters β_p and β_a.*

We estimate β_p and β_a to minimize the least square errors based on the system of non-linear equations given only competitor occupancy, y_d^c. We use nonlinear least squares (NLS) method. The error terms' variance in (1.14) is $M_{\omega_{dl}} \mathcal{P}_{dl}^c$, so the error terms cause heteroskedasticity. We report heteroskedasticity corrected (Huber-White) standard errors in the numerical analysis to account for non-constant variance (for instance in Table 1.5 in brackets under the means of the parameters).

a Price endogeneity due to model misspecification

Managers set prices based on some knowledge of the markets and demand for specific days. Data does not necessarily reflect that knowledge. In our case we do not observe competitor

private sales and it for sure is correlated with competitor prices as we know there is limited inventory.

Specifically, once we omit $y_d^{G^c}$, it is absorbed in $\tilde{\epsilon}_d^c$ and what we are actually estimating is

$$y_d^c = e^{\beta_a}[A\hat{\boldsymbol{s}}(\beta_p)]_d + \tilde{u}_d^c.$$

One strongly suspects that there is a correlation between $y_d^{G^c}$ and the observed price difference. The regression equation (1.11) has an independent variable, Δp_{dl}, that is correlated with the error term. This means Δp_{dl} is an *endogenous* variable.

This is a classical problem in econometrics, known as the *endogeneity* problem. In case of endogeneity, it is well known that naive estimation methods such as NLS give biased and inconsistent estimates of the parameters (β_p, β_a).

b First moment condition based on IV

Instrumental variables (IVs) are the standard tools to deal with the endogeneity issue. As the demand model is nonlinear and heteroskedasticity is present, we use an IV estimation technique of Generalized Method of Moments (GMM) framework to address the endogeneity problem.

A valid IV should satisfy two conditions. We should find a variable that is

- correlated with the endogenous variable,
- uncorrelated with the errors.

Hypothesis 1. *Our own private sales are correlated to our public sales (or, our prices) due to capacity restrictions/dynamic pricing, but uncorrelated to competitor's private sales.*

We propose using our private occupancy, $y_d^{G^o}$, as an instrument in (1.12). First of all, there is no reason to believe that our private sales are correlated with competitor private sales. The reason is that private sales (at least in the hotel context) are based on corporate relationships and private negotiations. Hence, our private sales $y_d^{G^o}$ are independent from the competitor private sales $y_d^{G^c}$ and also error term $\tilde{\epsilon}_d^c$ is independent of all others.

We next argue that our private sales are highly correlated with the observed price differences, Δp_{dl}. For example, if our own private sales are high, since this reduces our capacity,

in our best response function, based on the fixed competitor price we will increase our own price.

Let V_d denote our private occupancy (mean normalized). Then, GMM imposes that the set of K orthogonality conditions are satisfied

$$g_d(\beta_p|\beta_a) = E[V_d\tilde{\epsilon}_d^c] = 0,$$

where $g_d(\beta_p|\beta_a) = \sum_d V_d(y_d^c - e^{\beta_a}[A\hat{\boldsymbol{s}}(\beta_p)]_d) - \sum_d V_d y_d^{G^c}$. Due to IV being independent of competitor private sales, the last term cancels out and we have

$$g(\beta_p|\beta_a) = \sum_d V_d y_d^c - e^{\beta_a} \sum_d V_d[A\hat{\boldsymbol{s}}(\beta_p)]_d.$$

To obtain the GMM estimator β_p, we solve

$$\underset{\beta_p}{\text{minimize}} \quad G(\beta_p), \tag{1.15}$$

where $G(\beta_p) = g(\beta_p|\beta_a)^T g(\beta_p|\beta_a)$. Numerically, we illustrate the quasi-convexity of $G(\beta_p)$ in Figure 1.3.

c Second moment condition based on CV

We derive the second moment condition based on the following assumption.

Assumption 1. *Coefficient of variation (CV) of competitor private occupancy is the same as CV of our private occupancy (observed).*

However, the A matrix prohibits us from using any such moment condition in a simple way. Next, we show why this is an issue.

Example 1. Λ_1, Λ_2, Λ_3 *are three Gaussian processes for private sales of LOS* $1, 2, 3$ *respectively. Then,*

$$y_2^c - e^{\beta_a} \sum_{d'=2-L}^{2} \sum_{l=2-d'+1}^{L} \hat{s}_{d'l}(\beta_p) = \lambda_{12} + \lambda_{13} + \lambda_{21} + \lambda_{22} + \lambda_{23}$$

$$y_3^c - e^{\beta_a} \sum_{d'=3-L}^{3} \sum_{l=3-d'+1}^{L} \hat{s}_{d'l}(\beta_p) = \lambda_{31} + \lambda_{22} + \lambda_{32} + \lambda_{13} + \lambda_{23} + \lambda_{33}.$$

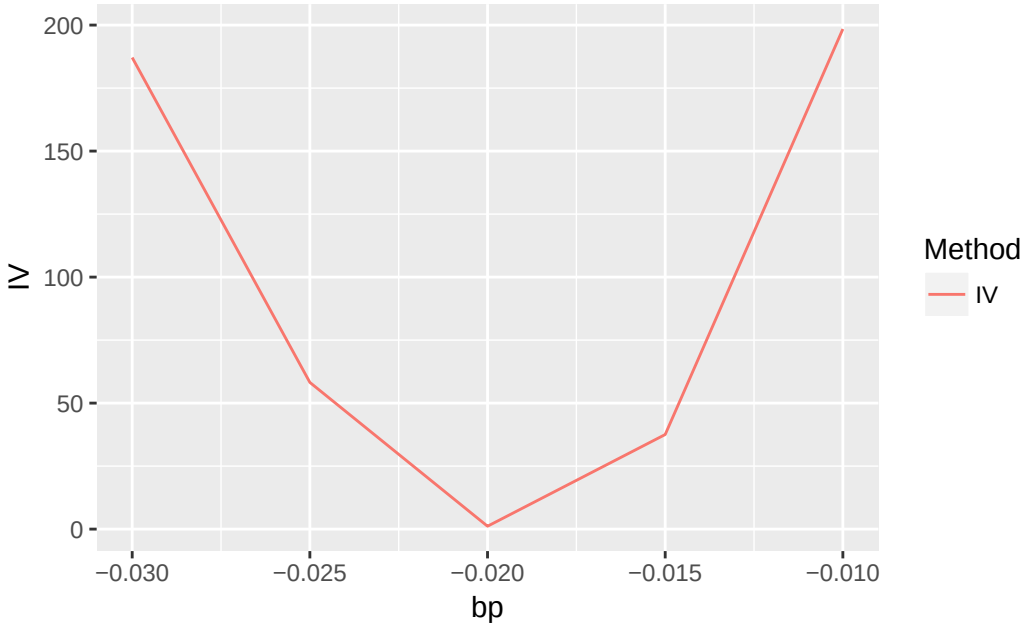


Figure 1.3: A graph of of $G(\beta_p)$ for a fixed value of β_a

The variance of right-hand-side becomes a complicated function of λ's (unknown competitor groups) and unless A is square and invertible (not the case in our problem) there is no way to solve it.

Our approach to tackle this problem is based on the independence of aggregate private sales for days that are max-LOS days apart.

Example 2. *Let's consider max LOS=3 case, we can take* $y_2^c - e^{\beta_a} \sum_{d'=2-L}^{2} \sum_{l=2-d'+1}^{L} \hat{s}_{d'l}(\beta_p)$ *and* $y_5^c - e^{\beta_a} \sum_{d'=5-L}^{5} \sum_{l=5-d'+1}^{L} \hat{s}_{d'l}(\beta_p)$ *pairs as they have no random variables in common on the right-hand-side, they are independent so that we can carry out CV calculations.*

To estimate β_a, we impose a moment condition that CV of competitor private occupancy is equal to CV of our private occupancy. As we mentioned before, we consider only the days that are max-LOS days apart. So, our objective is to minimize the difference of CV functions between our and competitor private occupancy for any day sample j where $d \equiv j \mod L$.

$$\underset{\beta_a}{\text{minimize}} \quad \|c_j(\beta_a)\|_2^2 \tag{1.16}$$

where $c_j(\beta_a) = (\underbrace{\text{CV}(y_j^c - e^{\beta_a}[A\hat{\boldsymbol{s}}(\hat{\beta}_p))]_j}_{\text{CV}_j(\beta_a|\beta_p)} - \underbrace{\text{CV}(y_j^{G^o})}_{\text{Constant}})^2$.

Lemma 1. *CV estimator is asymptotically unbiased and consistent as noted in Sharma and Krishna (1994).*

Proposition 2. *For a given β_p, $CV_j(\beta_a|\beta_p)$ is quasi-convex function of β_a with the minimum at true β_a.*

Proof. CV function is given by $\frac{\sqrt{\text{Var}[\boldsymbol{y}^{G^c}(\beta_a)]}}{\text{E}[\boldsymbol{y}^{G^c}(\beta_a)]}$ where the inside term is

$$\boldsymbol{y}^{G^c}(\beta_a) = \boldsymbol{y}^c - e^{\beta_a} A\hat{\boldsymbol{s}}(\hat{\beta}_p).$$

Now, we investigate the structural properties of numerator and denominator terms separately.

Firstly, let's concentrate on the numerator term. Using the definition of variance, one can write $\text{Var}[\boldsymbol{y}^{G^c}(\beta_a)] = \text{E}[(\boldsymbol{y}^{G^c}(\beta_a))^2] - (\text{E}[\boldsymbol{y}^{G^c}(\beta_a)])^2$. Expanding this out and gathering the common terms together, we have

$$\begin{aligned} h(\beta_a) = \text{Var}[\boldsymbol{y}^{G^c}(\beta_a)] =& \sum_j (y_j{}^c)^2 - (\sum_j y_j{}^c)^2 \\ &- 2e^{\beta_a}(\sum_j y_j{}^c [A\hat{\boldsymbol{s}}(\hat{\beta}_p)]_j - \sum_j y_j{}^c \sum_j [A\hat{\boldsymbol{s}}(\hat{\beta}_p)]_j) \\ &+ e^{2\beta_a} \sum_j [A\hat{\boldsymbol{s}}(\hat{\beta}_p)]_j^2 - (\sum_j [A\hat{\boldsymbol{s}}(\hat{\beta}_p)]_j)^2. \end{aligned}$$

Now, it is easy to see that we have a quadratic polynomial form

$$h(\beta_a) = e^{2\beta_a}\text{Var}[A\hat{\boldsymbol{s}}(\hat{\beta}_p)] - 2e^{\beta_a}\text{Cov}[\boldsymbol{y}^c, A\hat{\boldsymbol{s}}(\hat{\beta}_p)] + \text{Var}[\boldsymbol{y}^c] \geq 0.$$

Notice that $\text{Var}[A\hat{\boldsymbol{s}}(\hat{\beta}_p)] > 0$ so the quadratic polynomial has a positive leading coefficient. Then, we let the numerator function to be denoted by

$$\begin{aligned} v(\beta_a) =& \sqrt{\text{Var}[\boldsymbol{y}^{G^c}(\beta_a)]} \\ =& \sqrt{e^{2\beta_a}\text{Var}[A\hat{\boldsymbol{s}}(\hat{\beta}_p)] - 2e^{\beta_a}\text{Cov}[\boldsymbol{y}^c, A\hat{\boldsymbol{s}}(\hat{\beta}_p)] + \text{Var}[\boldsymbol{y}^c]}. \end{aligned}$$

We check for the first derivative

$$v'(\beta_a) = \frac{h'(\beta_a)}{2\sqrt{h(\beta_a)}} = \frac{2e^{\beta_a}(e^{\beta_a}\mathrm{Var}[A\hat{\boldsymbol{s}}(\hat{\beta}_p)] - \mathrm{Cov}[\boldsymbol{y}^c, A\hat{\boldsymbol{s}}(\hat{\beta}_p)])}{2\sqrt{h(\beta_a)}}.$$

Accordingly, if

$$\beta_a = \log\Big(\frac{\mathrm{Cov}[\boldsymbol{y}^c, A\hat{\boldsymbol{s}}(\hat{\beta}_p)]}{\mathrm{Var}[A\hat{\boldsymbol{s}}(\hat{\beta}_p)]}\Big),$$

$v'(\beta_a) = 0$, if

$$\beta_a > \log\Big(\frac{\mathrm{Cov}[\boldsymbol{y}^c, A\hat{\boldsymbol{s}}(\hat{\beta}_p)]}{\mathrm{Var}[A\hat{\boldsymbol{s}}(\hat{\beta}_p)]}\Big),$$

$v'(\beta_a) > 0$ meaning $v(\beta_a)$ is increasing in β_a and if

$$\beta_a < \log\Big(\frac{\mathrm{Cov}[\boldsymbol{y}^c, A\hat{\boldsymbol{s}}(\hat{\beta}_p)]}{\mathrm{Var}[A\hat{\boldsymbol{s}}(\hat{\beta}_p)]}\Big),$$

$v'(\beta_a) < 0$ meaning $v(\beta_a)$ is decreasing in β_a. This implies that $v(\beta_a)$ is quasi-convex in β_a.

Secondly, we consider the denominator, $f(\beta_a) = \mathrm{E}[\boldsymbol{y}^{G^c}(\beta_a)]$, which is of affine form. The first derivative is

$$f'(\beta_a) = -e^{\beta_a}\mathrm{E}[A\hat{\boldsymbol{s}}(\hat{\beta}_p)] < 0$$

and the second derivative is

$$f''(\beta_a) = -e^{\beta_a}\mathrm{E}[A\hat{\boldsymbol{s}}(\hat{\beta}_p)] < 0$$

so that $f(\beta_a)$ is a decreasing and concave function.

Finally, as Boyd and Vandenberghe (2004) indicates on page 102, the composition of a quasi-convex function with an affine function (nondecreasing) yields a quasi-convex function. $f(\beta_a)$ is decreasing, then $\frac{1}{f(\beta_a)}$ is increasing so that CV function being the composition of quasi-convex and nondecreasing functions is quasi-convex. The numerical illustration for quasi-convexity of CV function can be seen in Figure 1.4 and Figure 1.5. □

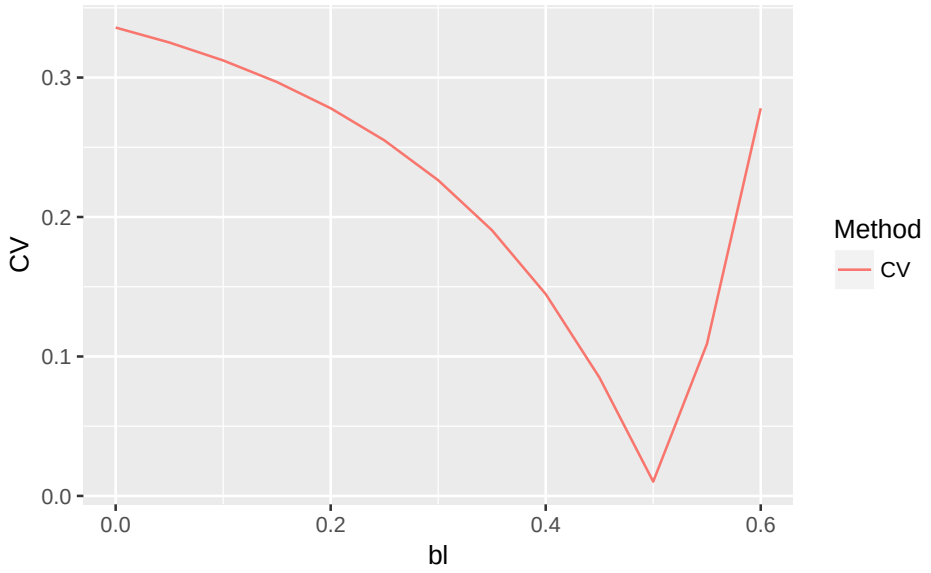


Figure 1.4: $\|c_j(\beta_a)\|_2^2$ as a function of location parameter when true $\beta_a = 0.5$

d Estimation algorithm

A necessary condition for identification of β_p, β_a via the GMM framework is the order condition, which states that number of instruments, K, is greater than or equal to the number of explanatory variables, E, that is $K \geq E$. In our case, we have an exactly identified model because $K = E = 1$.

To obtain the GMM estimator, $\hat{\beta}_p, \hat{\beta}_a$, we solve the system of two moment conditions.[4] However the solution is not so simple as these are highly non-linear equations. Exploiting the quasi-convexity, we solve the system by an alternating optimization—find an estimator β_p fixing a value of β_a and vice-versa, till the values converge.

1.4.2 Estimation of market size

Once we estimate price sensitivity and location parameters, we proceed to estimate the no-purchase parameter for the estimated $\hat{\beta}_p$ and $\hat{\beta}_a$. We have absolutely no information on the outside option so we have to make some assumptions. We make the following reasonable and minimal assumption on the demand-generating-process:

[4] When the model is just identified, regardless of weighting matrix, GMM reduces to the standard IV estimator. Moreover, the choice of the weight matrix does not affect the asymptotic distribution of the estimator for the just identified case (see Hayashi (2000)).

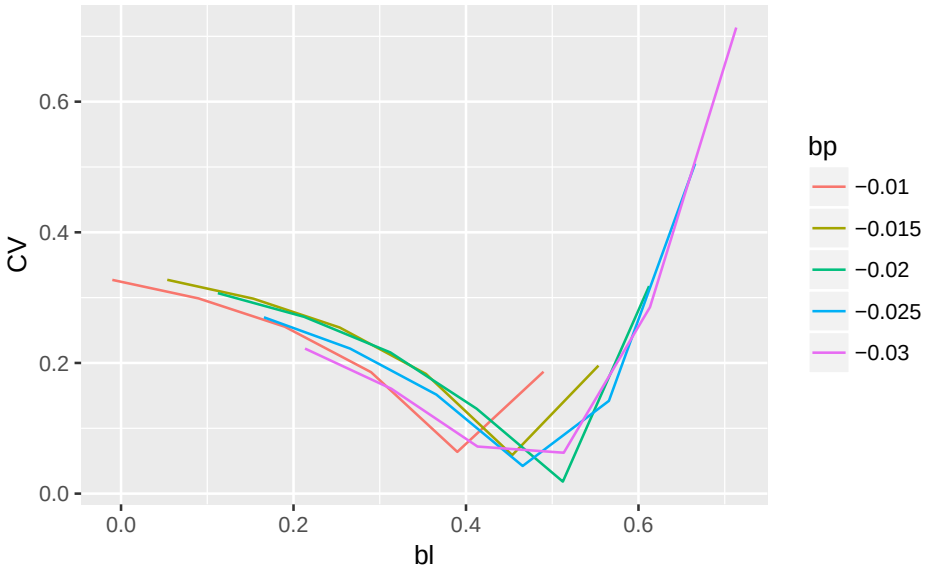


Figure 1.5: $\|c_j(\beta_a)\|_2^2$ function given several β_p parameters

Assumption 2. *Market for each day-of-week come from a specific process (the distribution might change for each day-of-week) and markets for days that are sufficiently far apart are independent of each other (eg: Sunday market is independent of Wednesday market).*

We let ω represent day-of-week (Mon, Tue, etc) and assume for simplicity that the demand for LOS is then divided by a stationary multinomial distribution. This demand is split into LOS where demand for LOS l has a mean of π_l, $\sum_{l=1}^{L} \pi_l = 1$.

Let $M_\omega = \{M_{\omega l}\}$ represent a time-series of (unobserved) market-size for day-of-week ($\omega = 1, 2, ..., 7$) and instance $l, l = 1, \ldots, L$. Likewise $\{s^o_{\omega l}\}$ is the public sales for our hotel (observed) for day-of-week ω. The series are linked by

$$M_{\omega l} = s^p_{\omega l}[e^{-\beta_0} + e^{\beta_p p^o_{\omega l}} + e^{\beta_p p^c_{\omega l} + \beta_a}] \quad \forall w \in W \tag{1.17}$$

where $s^p_{\omega l} = \frac{s^o_{\omega l}}{e^{\beta_p p^o_{\omega l}}}$. Based on the standard MNL model, we denote the (LOS-wise) average of the right-hand-side by

$$M_\omega = \frac{1}{L} \sum_{l=1}^{L} s^p_{\omega l}[e^{-\beta_0} + e^{\beta_p p^o_{\omega l}} + e^{\beta_p p^c_{\omega l} + \beta_a}] \quad \forall w \in W$$

Then, we assume M_ω and $M_{\omega+k}$ (say $k = 3$) are independent. However, public sales s_ω

and $s_{\omega+k}$ are not necessarily independent as network effects and capacity constraints create dependencies for the sales. Moreover, our public sales is an observed phenomenon, so based on the correlation of our public sales we can threshold out a value of k if this is close to 0.

Proposition 3. *If M_ω and $M_{\omega+k}$ are independent and $|\rho(s_\omega, s_{\omega+k})| \neq 0$, β_0 is partially identified as one of two values (roots of a quadratic equation). Further conditions to pin them down are that it is a real number with $e^{-\beta_0} > 0$.*

Proof. We make no assumptions on the M_ω's except they are independent series. Their distribution might change for each ω. We take the co-variance operator on both sides (defined purely algebraically for a series, and not as an expectation) between the two series $\omega = 1, 2$, to get

$$Cov(M_1, M_2) = \frac{1}{L}\sum_{l=1}^{L}(s^p_{1l}[e^{-\beta_0} + z_{1l}] - \bar{M}_1)(s^p_{2l}[e^{-\beta_0} + z_{2l}] - \bar{M}_2)$$

where $z_{\omega l} = e^{\beta_p p^o_{\omega l}} + e^{\beta_p p^c_{\omega l} + \beta_a}$ and $\bar{M}_w$ being the (unknown) average market size for day-of-week ω. This expression can be written as

$$\begin{aligned} Cov(M_1, M_2) =& Cov(e^{-\beta_0}s^p_1 + s^p_1 z_1, e^{-\beta_0}s^p_2 + s^p_2 z_2) \\ =& e^{-2\beta_0}Cov(s^p_1, s^p_2) + e^{-\beta_0}Cov(s^p_1, s^p_2 z_2) + e^{-\beta_0}Cov(s^p_1 z_1, s^p_2) \\ &+ Cov(s^p_1 z_1, s^p_2 z_2). \end{aligned} \tag{1.18}$$

Solving the quadratic equation, we obtain 2 roots:

$$\begin{aligned} \beta_0 = -\log\Big(&\frac{-(Cov(s^p_1, s^p_2 z_2) + Cov(s^p_1 z_1, s^p_2))}{2Cov(s^p_1, s^p_2)} \\ &\pm \frac{\sqrt{(Cov(s^p_1, s^p_2 z_2) + Cov(s^p_1 z_1))^2 - 4Cov(s^p_1, s^p_2)Cov(s^p_1 z_1, s^p_2 z_2)}}{2Cov(s^p_1, s^p_2)}\Big) \end{aligned}$$

Now, let us suppose that our procedure (any procedure) based on a moment condition satisfying this equation comes up with $\beta_0 + \tilde{\beta}_0$. The questions on identifiability would be — does our procedure always set $\tilde{\beta}_0 = 0$? Can any $\tilde{\beta}_0 \in \Re$ satisfy this equation? If neither, is there a possibility that we find a unique $\tilde{\beta}_0$, but it is $\neq 0$?

Suppose we are solving the moment condition $Cov(M_1, M_2) = 0$ (unobservable) by trying to find the β_0 that satisfies the right-hand-side (observable, but with unknown parameter

β_0) set to 0. Do we recover the true β_0 or some other $\hat{\beta}_0 \neq \beta_0$? Let us represent the estimated $\hat{\beta}_0 = \beta_0 + \tilde{\beta}_0$. We solve the following empirical equation and for sufficiently large sample of days, say it is close to 0 for all samples.

$$\frac{1}{L}\sum_{l=1}^{L}(s_{1l}^p[e^{-\beta_0} + \tilde{x} + z_{1l}] - (\tilde{x}\bar{s}_1^p + \bar{M}_1))(s_{2l}^p[e^{-\beta_0} + \tilde{x} + z_{2l}] - (\tilde{x}\bar{s}_2^p + \bar{M}_2)) = 0. \quad (1.19)$$

Then we get

$$Cov(M_1, M_2) + \tilde{x}^2 Cov(s_1^p, s_2^p) + \tilde{x}Cov(s_1^p, M_2) + \tilde{x}Cov(s_2^p, M_1) = 0.$$

Now, since the first term is 0 by independence assumption, $\tilde{x}$ has to be

$$\tilde{x} = -\frac{(Cov(s_1^p, M_2) + Cov(s_2^p, M_1))}{Cov(s_1^p, s_2^p)}$$

This tells us that if $Cov(s_1^p, s_2^p) = 0$, pretty much any $\tilde{\beta}_0$ would satisfy the moment condition. The bias amount in the parameter estimate is given by $e^{-(\beta_0+\tilde{\beta}_0)} = e^{-\beta_0} + \tilde{x}$. Then, making the logarithmic transformation leads to

$$\tilde{\beta}_0 = -\log(\frac{\tilde{x}}{e^{-\beta_0}} + 1).$$

□

Although, we make minimal assumptions on the market-size distributions, it still leaves open the question: are the (unobserved) markets truly independent, if so which pairs? This can never be fully verified. As a practical matter, it is also not clear how to choose k when each k gives a different solution. We illustrate this with an example.

Example 3. *Generated synthetic data correlations of s^p and M can be seen in Figure 1.6 and 1.7. Then, solving the quadratic equation for each pair of days in a week, we obtain a triangular matrix that proposes a set of candidates for β_0. NRR stands for no real root, and X is when one of the roots of a quadratic equation is negative and eliminated due to the logarithm.*

As can be seen in this example, each pair of days in a week results in a different β_0. So, the question is which one is the true β_0. Based on the sparsity matrix assumption meaning that most of the market correlations are 0, we develop a new criteria to find β_0. We propose to find

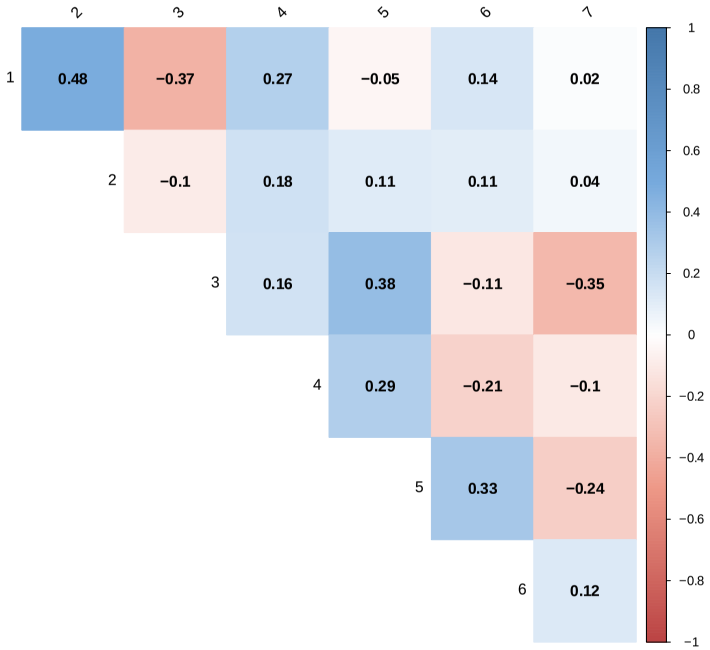


Figure 1.6: Correlation of s^p

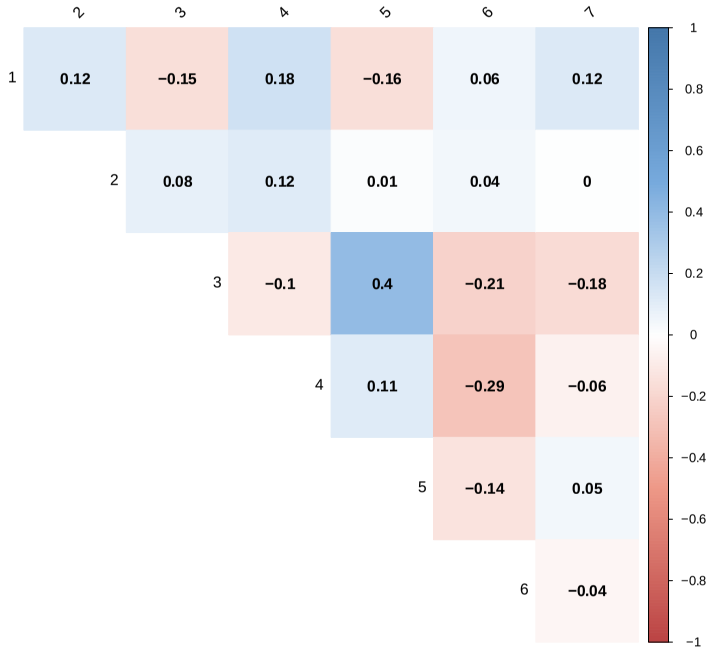


Figure 1.7: Correlation of market size

β_0^1							β_0^2						
NRR	3.123	X	NRR	X	3.156	NRR	NRR	X	3.468	NRR	X	X	NRR
	NRR	3.745	4.339	2.952	3.224	2.326		NRR	1.944	X	X	5.033	2.875
		NRR	2.01	X	X	X			NRR	X	X	4.228	3.312
			NRR	NRR	X	NRR				NRR	NRR	4.566	NRR
				NRR	2.327	4.263					NRR	X	2.186
					NRR	2.542						NRR	X
						NRR							NRR

Table 1.4: Independence method estimates, $\hat{\beta}_p = -0.02$, $\hat{\beta}_a = 0.5$, $\hat{\beta}_0 = 3$.

β_0 that minimizes the 0-norm of the (predicted) correlation matrix. Therefore, we develop approximate solutions as follows:

- Find β_0 that minimizes the *sum* of 1-norms of the (upper-triangular) correlation matrix,
- Find β_0 that minimizes the *medians* of 1-norms of the (upper-triangular) correlation matrix.

We compare performances of each method on a synthetic and real world data setting in the numerical analysis section. Moreover, one can develop more advanced spectral criteria via experimenting the market characteristics.

Next, we provide the basis for the equilibrium model and the best-response pricing model. We use this to generate synthetic data prices to test the recovery of the true parameters.

1.5 Equilibrium and best-response pricing model

Once we obtain the choice behavior parameters, we can optimize prices for our hotel based on competitor prices to maximize total revenue subject to the capacity constraint.

Available competitor data enables us to model the equilibrium price response of competitors. Even though pricing under MNL has received a lot of attention, we have not found the exact results that fit our model, so in this section we derive the equilibrium pricing conditions, indeed we show first that such an equilibrium exists, when each player has their own beliefs about the true parameters. So essentially, the only assumption we make is that players optimize their prices. While such equations cannot be used to estimate the true parameters, they will be useful for us to generate the prices in simulations to test the recovery of the true parameters.

1.5.1 Single-product, single-leg case

We know our hotel's remaining capacity but not the competitor's remaining capacity (due to limited occupancy data and unobserved competitor private sales). In our model, there is a subtle assumption on timing. We assume private sales are negotiated before public sales. Also, public sales occur based on a choice model if there is available capacity. We do not model overflow to the competitor if we run out of capacity, as this requires a dynamic model (see Martínez-de Albéniz and Talluri (2011) for such analysis and related references).

We assume that the prices are determined endogenously as a function of remaining capacities as well as competitor pricing, in a Nash equilibrium condition (full information) with each party having their own estimates of parameters. This is the standard logit equilibrium pricing but with competition, and capacities.

Let our price be p^o, and competitor price be p^c. Our hotel capacity is c^o and competitor capacity is c^c respectively, and hold beliefs (with the super-scripts) about the parameters $\boldsymbol{\beta}^o$ and $\boldsymbol{\beta}^c$ respectively (without the super-scripts, they are the true parameters). We define $S^o(p^o|p^c, \boldsymbol{\beta}^o)$ as the Gaussian random variable representing own unconstrained demand (and vice versa for the competitor) and $\bar{S}^o(p^o|p^c, \boldsymbol{\beta}^o)$ as demand that is constrained by own capacity, that can be written as

$$\bar{S}^o(p^o) = S^o(p^o) - [S^o(p^o) - c^o]^+.$$

Suppressing the fixed parameters, our expected sales is given by MNL choice model (given the beliefs on the parameters):

$$s^o(p^o) = E[S^o(p^o)] = M\frac{e^{\beta_p p^o}}{e^{-\beta_0} + e^{\beta_p p^o} + e^{\beta_p p^c + \beta_a}}. \tag{1.20}$$

Then, our firm's best reaction function, given competitor price, p^c, is

$$\max_{p^o} p^o \bar{s}^o(p^o).$$

The same problem can be expressed as the following constrained nonlinear optimization problem:

$$\begin{aligned} \underset{p^o}{\text{maximize}} \quad & R^o(p^o) \\ \text{subject to} \quad & 0 \leq s^o(p^o) \leq c^o. \end{aligned} \tag{1.21}$$

where $R^o(p^o) = p^o s^o(p^o)$.

a Concavity of revenue function

As is well-studied, the MNL revenue function is not concave in prices (Hanson and Martin, 1996) but in quantities (Dong et al., 2009; Song and Xue, 2007; Li and Huh, 2011). So we transform the problem by first defining the inverse function for price as a function of the constrained demand $p(\bar{s})$ as the price that achieves the constrained demand $\bar{s}$. Such an inverse function is well-defined as $\bar{s}(p)$ is a strictly decreasing function of p (for both our as well as competitor, for fixed prices of the competitor). The transformed objective function can be written as

$$R^o(s^o) = p^o(s^o)s^o. \tag{1.22}$$

Proposition 4. *(Dong et al., 2009) The MNL revenue function is concave in sales.*

1.5.2 Multiple-products, network case

In this section, we derive the best-response prices of our firm for a network environment. As shown in previous section, the single product revenue function under MNL is concave in sales. Then, the network RM problem under competition is the concave optimization problem

$$\begin{aligned} \max_{s^o_{dl}} \quad & \sum_{(d,l)} R^o(s^o_{dl}) \\ \text{s.t.} \quad & As^o_{dl} \leq c^o_d \\ & 0 \leq s^o_{dl} \leq M_{dl} \end{aligned} \tag{1.23}$$

where $R^o(s^o) = s^o_{dl} p^o(s^o_{dl})$ and $p^o(s^o_{dl})$ is a function of competitor prices, p^c. The problem described in (1.23) is known as the Generalized Nash equilibrium problem (GNEP), which is an equilibrium problem that the feasible sets depend on the other player's decisions. Specifically, in our case, it consists of 2 players, our and competitor hotel. We denote by $\mathbf{s}$ the vector of decision variables; $\boldsymbol{s}= (\boldsymbol{s^o}, \boldsymbol{s^c})$. Each player has an objective function of maximizing revenue for its own firm based on both its own variables, $\boldsymbol{s^o}$ as well as the variables of other players, $\boldsymbol{s^c}$. Also, the feasible set of each player depends on the rival player's strategy.

1.5.3 Existence of the Nash equilibrium

Now, we derive the necessary conditions for the existence of Nash equilibrium. A point $\bar{s}$ is called a Nash equilibrium if no player can increase its revenue by changing their strategy to any other feasible point.

Lemma 2. *The objective function given in (1.23) for fixed p^c is concave in s_{dl}.*

Proof. The argument follows from Proposition 4. The sum of concave functions is concave. So, $\sum_{(d,l)} R^o(s^o_{dl})$ is concave. □

The GNEP is called jointly convex if in addition to the objective function, the feasible set is a closed and convex set. The objective of our problem is to maximize the revenue. To be concise with GNEP literature, we can multiply it by -1 and write it as min $-R^o(s^o)$ so that we obtain a convex objective function. In addition, the feasible region of the optimization problem is given by linear constraints, therefore it is convex. Then specifically GNEP is jointly convex.

Theorem 1. *For every player, the objective function given in (1.23) for fixed p^c is concave and the feasible region is given by linear constraints. Then, a generalized Nash equilibrium exists.*

Proof. As stated by Vives (2001) a Nash equilibrium exists if the strategy sets are non-empty convex and compact, and the payoff to the firm is continuous in the actions of all firms and quasi-concave in its own action. The firms can set a sale quantity restricted by the capacities, so $s = [0, c]$ which is compact set. Also, as the feasible region is given by linear constraints, it is a nonempty convex set. Finally, the revenue function is continuous in s and according to Lemma 2 it is concave in s_{dl}. Therefore, a Nash equilibrium exists. □

1.6 Numerical results

In this section, we present a comparison of methods on both synthetic and real-world hotel data sets. Synthetic data enables us to concretely understand how well the methods perform on a known demand system since we know the true underlying model parameters. On the other hand, the real-world setting is necessary to test the real performance of our method. Also, real data gives us clues to understand if there are any limitations into our methodology.

1.6.1 Synthetic data

We use best-responce pricing mechanism to generate the synthetic data setting. We generate the private sales from the Normal distribution for both our and competitor hotel. Then, once we obtain the remaining capacities, we derive public sales and the respective prices of both our and competitor firm based on a Nash equilibrium setting. In order to solve the GNEP, we follow a simple heuristic, called Nonlinear Gauss-Seidel, that is quite popular among practitioners. The steps of the algorithm are described as follows:

- Step 0: Choose a starting point $s^0 = (s^{0,o}, s^{0,c})$ and set k:=0
- Step 1: If s^k satisfies a suitable termination criterion: STOP (i.e. $|s^k - s^{k+1}| < 0.01$)
- Step 2: For our hotel, taking $n = o$, compute a solution $s^{k+1,o}$ to the nonlinear constrained optimization problem given in (1.23).
- Step 3: Similarly, for the competitor hotel, taking $n = c$, compute a solution $s^{k+1,c}$ to Equation (1.23). Set $s^{k+1} := (s^{k+1,o}, s^{k+1,c})$, $k \leftarrow k + 1$ and go to Step 1.

We conduct the simulation with 50 weeks horizon (equivalently 350 days) and a maximum of 2-LOS. We take $M_\omega = [100, 300, 160, 90, 110, 200, 150]$ and $\pi_l = [0.4, 0.6]$. We generate market arrivals from the Normal distribution with $N(M_\omega, 3\sqrt{M_\omega})$ along time horizon (50 weeks), distribute it proportionally with π_l so that we find the daily LOS (itinerary) arrivals. Later, we generate the private sales for both our and competitor hotel, we take $\lambda = 80$. We use the Normal distribution with $N(\lambda, 3\sqrt{\lambda})$. Afterwards, based on the remaining capacities, we derive the public sales and its prices with MNL probabilities based on a Nash equilibrium setting. In the results, we present four methods and compare their performances.

- NLS: We use naive NLS regression without accounting for the existence of private sales
- IV: We consider the existence of private sales and use IV technique to deal with endogeneity problem based on GMM. In the estimation of β_0 step, we also try variations based on minimizing the sum/median of 1-norms of the (upper triangular) correlation matrix, named as IV (sum)/IV(median).

Morever, we compare these method with NT based methods.

- TW: We use MCMC based method of Tebaldi and West (1998). We apply TW with conjugate gamma prior parameters $\alpha = 1$ and $\beta = 0$.

- Vardi: We use method-of-moments based method of Vardi (1996). Actually, Vardi is not directly applicable to our problem setting since there exists only one sample demand for the whole network (even if the arrivals were homogeneous, due to the change in prices, the generated sales are unique). So, we apply Vardi method based on an assumption that sales of each day of week repeats itself. Accordingly, we have 50 replications.

NT methods can untruncate sales from aggregated occupancy data. So they can predict competitor sales directly. To estimate β_p and β_a parameters, we adopt TW and Vardi methods as follows

- Using NT idea, we first untruncate competitor sales, so we predict s^c_{dl}, from competitor occupancy, y^c_d.

- Similar to NLS method, we ignore the existence of private sales and we regress $log(s^c_{dl}) = \beta_a + log(s^o_{dl} e^{\beta_p \Delta_{dl}}) + \epsilon$ to estimate β_p and β_a.

Later, once we obtain location and price-sensitivity parameter estimates using all four of the methods, we apply independence method to estimate β_0 parameter. We present synthetic data results considering several scenarios. Throughout the numerical analysis, the Huber-White standard errors at 95% are reported under the mean of parameter estimates in parenthesis to account for heteroskedasticity.

a True parameters until convergence

In this part, we assume both our and competitor hotel have true knowledge of customer choice parameters. We run the best-response pricing algorithm until convergence. The estimation results are given in Table 1.5 and Table 1.6. We validate the accuracy of the estimation through the comparison of predictions with observed competitor public occupancy, y^c_s, and market occupancy, y^M. We observe that IV method gives accurate estimates. NLS, TW, Vardi methods do not consider the existence of competitor private sales, so price-sensitivity parameter obtained using those methods are upward biased. Computationally, TW takes considerably long due to MCMC method integration. Vardi clearly performs the worst among all.

Parameter	True values	IV (sum)	IV (med.)	NLS	T&W	Vardi
β_p	-0.025	-0.023	-0.023	-0.012	-0.017	-0.001
		(0)	(0)	(0.002)	(0.002)	(0.01)
β_a	0.35	0.356	0.356	0.645	0.751	0.856
				(0.006)	(0.016)	(0.066)
β_0	1	1.265	1.435	0.4	0.31	5
Validation check						
y_s^c	187.66	190.86	190.86	265.5	290.37	344.77
y^M	617.2	525.94	496.52	632.17	755.05	492.69
$U(p^o\|p^c = 67)$	55.74	61.73	62.75	97.45	73.2	1753.93
$C(p^o\|p^c = 67)$	64.18	61.73	62.75	97.45	73.2	1753.93

Table 1.5: Synthetic data results for 50 weeks horizon, max-LOS=2 network, $\beta_p = -0.025$, $\beta_a = 0.35$, $\beta_0 = 1$.

Parameter	True values	IV (sum)	IV (med.)	NLS	T&W	Vardi
β_p	-0.03	-0.028	-0.028	-0.015	-0.02	-0.002
		(0)	(0)	(0.001)	(0.002)	(0.015)
β_a	0.25	0.264	0.264	0.571	0.653	0.205
				(0.005)	(0.012)	(0.094)
β_0	1.8	1.945	2.195	-2	-0.04	4.325
Validation check						
y_s^c	200.83	204.03	204.03	280.23	302.66	197.42
y^M	617.2	557.47	514.94	3945.86	1188.43	361.03
$U(p^o\|p^c = 71)$	54.06	57.37	58.82	70.54	60.43	710.02
$C(p^o\|p^c = 71)$	76.49	70.95	68.42	77.51	74.76	710.02

Table 1.6: Synthetic data results for 50 weeks horizon, max-LOS=2 network, $\beta_p = -0.03$, $\beta_a = 0.25$, $\beta_0 = 1.8$.

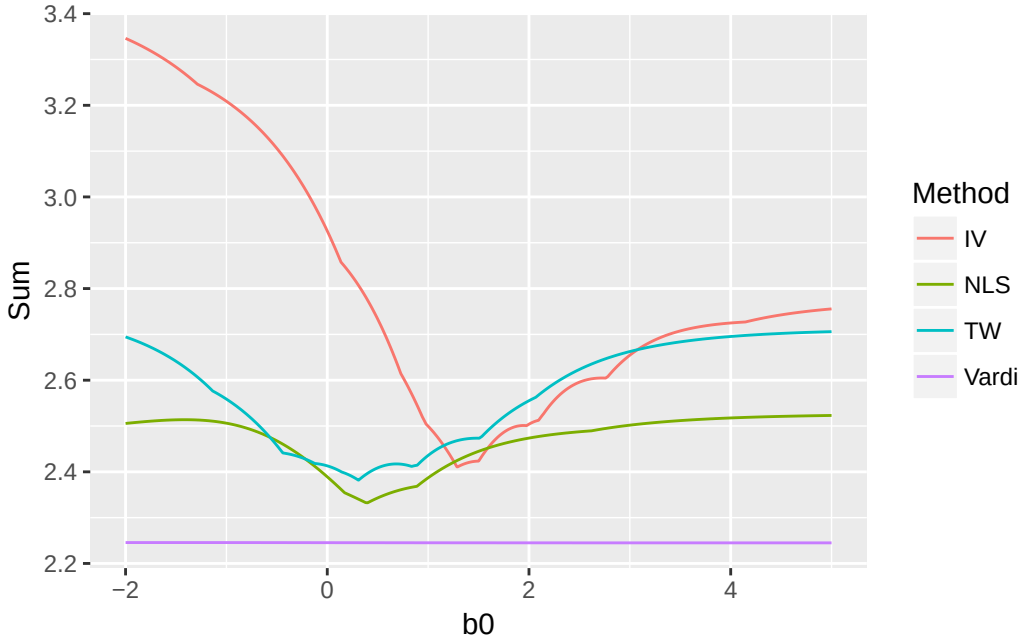


Figure 1.8: Independence method based on sum criteria to estimate $\beta_0 = 1$

b True parameters but half-way to convergence

Here, we assume both of the hotels know true customer choice parameters but while they each solve the best reaction function they stop at half-way of the iterations to convergence (maximum number of iterations until convergence is 8-10). We want to test how the estimation results would change under this scenario. Based on the results given in Table 1.7 and 1.8, we verify that IV method estimates are quite robust whereas NLS, TW and Vardi estimates are not reliable. TW and Vardi can be incapable of estimation or sometimes may result in positive price-sensitivity parameter, so we do not report these methods.

c Wrong information on parameters

In this part, the consumer demand occurs according to a set of true parameters, but both our and the competitor hotel have erroneous estimates of the parameters and solve the best-response pricing according to the erroneous estimates. We create the erroneous estimates by randomly perturbing the parameters by 20%, 30%, 40% up and down for both our and competitor hotel. As can be seen in Tables 1.9, 1.10, 1.11, IV method estimates are again quite robust.

Parameter	True values	IV (sum)	IV (med.)	NLS (sum)	NLS (med.)
β_p	-0.03	-0.027	-0.027	-0.009	-0.009
		(0.001)	(0.001)	(0.001)	(0.001)
β_a	0.25	0.244	0.244	0.517	0.517
				(0.005)	(0.005)
β_0	1.8	1.51	1.815	-2	-2
Validation check					
y_s^c	192.41	190.03	190.03	272.88	272.88
y^M	617.2	613.36	547.03	2695.63	2695.63
$U(p^o\|p^c = 57)$	51.41	55.15	56.67	119.37	119.37
$C(p^o\|p^c = 57)$	69.13	66.29	61.94	119.37	119.37

Table 1.7: Robustness check, $\beta_p = -0.03$, $\beta_a = 0.25$, $\beta_0 = 1.8$.

Parameter	True values	IV (sum)	IV (med.)	NLS (sum)	NLS (med.)
β_p	-0.025	-0.021	-0.021	-0.005	-0.005
		(0.002)	(0.002)	(0.001)	(0.001)
β_a	0.35	0.316	0.316	0.559	0.559
				(0.005)	(0.005)
β_0	1	0.285	0.755	-2	-2
Validation check					
y_s^c	182.21	181.1	181.1	261.11	261.11
y^M	617.2	766.9	605.61	1981.31	1981.31
$U(p^o\|p^c = 64)$	55.33	60.55	63.68	205.96	205.96
$C(p^o\|p^c = 64)$	62.84	61.71	63.68	205.96	205.96

Table 1.8: Robustness check, $\beta_p = -0.025$, $\beta_a = 0.35$, $\beta_0 = 1$.

		+20%			-20%		
Parameter	True val.	IV (sum)	IV (med.)	NLS	IV (sum)	IV (med.)	NLS
β_p	-0.025	-0.022	-0.022	-0.012	-0.022	-0.022	-0.012
		(0)	(0)	(0.002)	(0)	(0)	(0.002)
β_a	0.35	0.354	0.354	0.615	0.366	0.366	0.693
				(0.005)			(0.006)
β_0	1	1.235	0.785	0.135	1.1	1.23	0.61
Val. check							
y_s^c	197.28	200.98	200.98	275.17	172.92	172.92	246.95
y^M	617.2	534.79	634.07	718.89	536.09	507.44	543.85
$U(p^o\|p^c)$	55.07	62.13	59.47	96.54	63.21	64.16	103.98
$C(p^o\|p^c)$	62.18	62.13	59.47	96.54	63.21	64.16	103.98

Table 1.9: 20% perturbation in the true value of the parameters.

		+30%			-30%		
Parameter	True val.	IV (sum)	IV (med.)	NLS	IV (sum)	IV (med.)	NLS
β_p	-0.03	-0.028	-0.028	-0.011	-0.028	-0.028	-0.017
		(0)	(0)	(0.002)	(0)	(0)	(0.001)
β_a	0.25	0.266	0.266	0.531	0.274	0.274	0.637
				(0.005)			(0.006)
β_0	1.8	2.035	2.41	-0.2	1.845	1.935	-0.03
Val. check							
y_s^c	208.48	212.4	212.4	287.95	182.05	182.05	257.47
y^M	617.2	543.88	494.87	917.69	563.75	542.51	970.83
$U(p^o\|p^c)$	53.12	57.49	59.26	102.02	58.86	59.55	71.41
$C(p^o\|p^c)$	74.2	68.69	65.72	102.02	75.59	74.47	71.41

Table 1.10: 30% perturbation in the true value of the parameters.

		+40%			-40%		
Parameter	True val.	IV (sum)	IV (med.)	NLS	IV (sum)	IV (med.)	NLS
β_p	-0.02	-0.019	-0.019	-0.009	-0.02	-0.02	-0.014
		(0)	(0)	(0.001)	(0)	(0)	(0.001)
β_a	0.5	0.5	0.5	0.634	0.51	0.51	0.801
				(0.004)			(0.005)
β_0	3	3.21	3.035	0.3	3.155	3.465	0.615
Val. check							
y_s^c	219.59	222.71	222.71	299.58	199.37	199.37	275.34
y^M	617.2	555.47	583.93	999.63	562.96	502.44	1303.51
$U(p^o\|p^c)$	101.82	106.19	104.64	138.84	112.48	116.28	98.34
$C(p^o\|p^c)$	161.56	153.74	156.91	161.5	167.46	157.89	169.09

Table 1.11: 40% perturbation in the true value of the parameters.

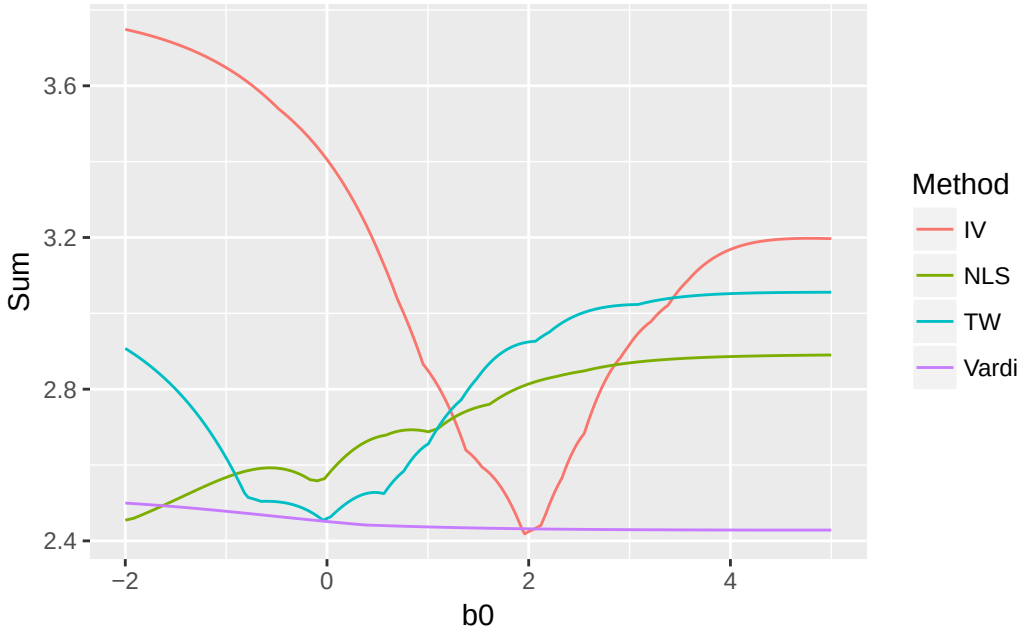


Figure 1.9: Independence method based on sum criteria to estimate $\beta_0 = 1.8$

1.6.2 Real-world hotel data sets

In this section, we consider the real-world data from a set of hotels. We collaborate with a hotel chain in U.K. and test our estimation method on the real sales data. As seasonality might be a problem, we keep the horizon shorter than synthetic data and select a period of between 10 to 20 weeks. Firstly, we obtain the summary statistics, as presented in Appendix 1.A.

The estimation results of the real-world hotel data sets are given in Tables 1.12, 1.13, 1.14. IV method gives a negative price-sensitivity whereas NLS gives positive. This is due to endogeneity problem. Also, we present the validation check results on competitor public and market size occupancy. All but one seems to accurately capture the correct market dynamics, and hence the right parameter estimates. However, we are not able to estimate the choice parameters for one of the hotels (Table 1.19 in Appendix 1.B). This is because when our and competitor hotel price difference does not vary along time horizon, there is no way to capture price-sensitivity parameter. The prices for this set of hotels can be seen through Figure 1.11 in Appendix 1.B.

Moreover, we can derive implications of customer behavior and hotel characteristics based on the parameter estimates. For instance, Hotel V customers are the most price sensitive of

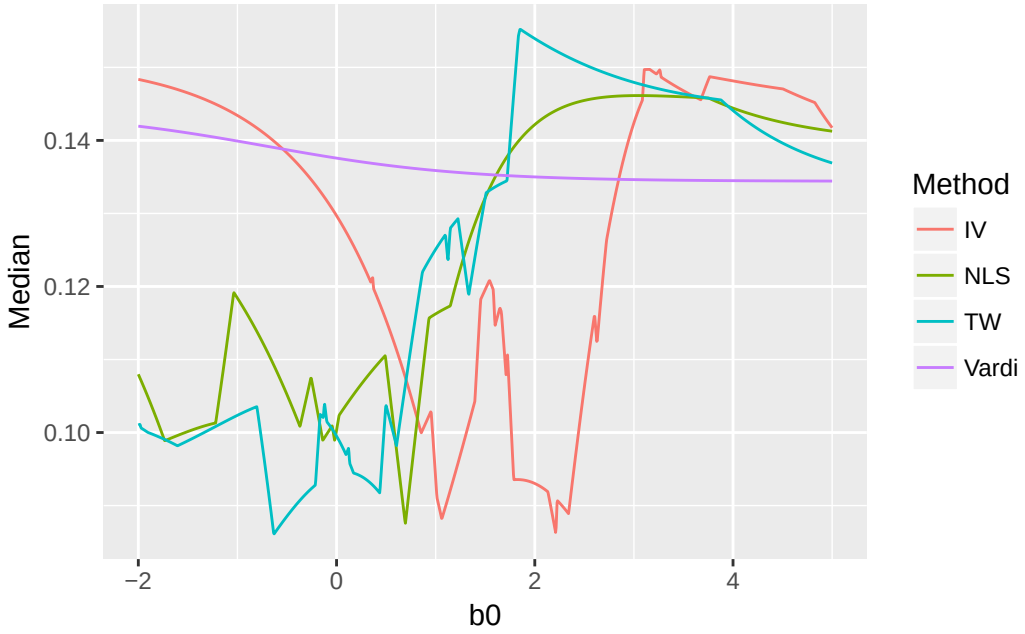


Figure 1.10: Independence method based on median criteria to estimate $\beta_0 = 1.8$.

all as $\beta_p = -0.019$ is the largest in absolute terms among them. Hotel V is considerably less attractive compared to its competitors. So, hotel V managers may try to increase the hotel's attractiveness by organizing an event/a conference or promoting leisure facilities (e.g. gym, spa, pool). Furthermore, we study Hotel M across two distinct periods in time (M1 in winter and M2 in summer). We observe that customers' price sensitivity is similar, but Hotel M seems to be more attractive during winter than summer. Thus, Hotel M can increase the prices slightly during winter without losing revenue.

1.7 Conclusion

What is the value perceived by customers for my product vs. the competitor's? For instance, a hotel may have a better aspect, brand and location attractiveness than a local competitor in the eyes of the customer. How does the customer perceive this difference and what additional margin can I charge or discount because of these differences? Estimation of customer choice parameters plays a crucial role for any RM setting involving uncertain demand. Biased estimates can result in significant problems as the parameters are used as an input for RM optimization procedures. To charge the right price in a certain market, one needs to know

Parameter	IV (sum)	IV (med.)	NLS	IV (med.)
β_p	-0.009	-0.009	-0.003	
	(0)	(0)	(0.002)	
β_a	-0.1	-0.1	-0.077	
			(0.013)	
β_0	8	4.34	0.6	
Validation check				
y_s^c	157.53	157.53	194.45	145.62
y^M	386.74	416.89	671.50	399.98
$U(p^o\|p^c = 275)$	253.13	242.89	464.75	242.20
$C(p^o\|p^c = 275)$	253.13	242.89	464.75	242.20

Table 1.12: Hotel M1 estimates over 20 weeks, starting mid-February 2015 and cross-validation of parameter estimates to a period beginning February 2016.

Parameter	IV (sum)	IV (med.)	NLS	IV(med.)
β_p	-0.011	-0.011	-0.004	
	(0)	(0)	(0.001)	
β_a	0.2	0.2	0.195	
			(0.024)	
β_0	1.32	3.78	-2	
Validation check				
y_s^c	143.21	143.21	199.38	166.80
y^M	1387.25	437.51	5043.71	471.31
$U(p^o\|p^c = 326)$	152.92	228.43	254.30	220.64
$C(p^o\|p^c = 326)$	297.38	270.45	257.73	247.31

Table 1.13: Hotel M2 estimates over 20 weeks, starting mid-July 2015 and cross-validation of parameter estimates for 10 weeks (starting June 2015).

Parameter	IV (sum)	IV (med.)	NLS	IV
β_p	-0.019	-0.019	0.009	
	(0.008)	(0.008)	(0.01)	
β_a	0.5	0.5	-0.283	
			(0.018)	
β_0	2.58	2.58	6	
Validation check				
y_s^c	106.92	106.92	127.14	103.79
y^M	290.98	290.98	250.04	350.47
$U(p^o\|p^c = 130)$	93.90	93.90	76751.97	102.76
$C(p^o\|p^c = 130)$	97.58	97.58	111.27	122.74

Table 1.14: Hotel V estimates over 20 weeks, starting mid-December 2015 and cross-validation of parameter estimates during 10 weeks (starting Sept. 2015).

the true customer parameters and account for competitor actions. The challenge arises due to the severe limitation in observing competitors' sales, and also unobservable product and customer characteristics.

In this study, we exploit competitor intelligence data and develop an enriched demand estimation model. We present a new econometric method for estimation. In this stream, we address two important issues (i) estimating with competitor effects (ii) estimating when the firm sells a single product. Further complications in the real world are competitor initial capacity, capacity constrained sales, network effects and choice parameter estimation when price is correlated to unobserved demand shocks. Surprisingly some of these complications actually help in identifying the parameters of the model.

We deal with a common problem peripheral to many practical situations: endogeneity of prices. We show that once a good set of instruments is found, IV procedure is quite reliable In our setting, we use own private bookings as IV to correct for the bias and obtain an accurate measure of price elasticity. We also address endogeneity of market size. This problem gets even more difficult in an industry such as hotel since the market size varies considerably even on a daily basis. We develop a novel method to estimate the market-size based on independence assumption.

We survey related methods from NT and RM, compare the performance of our method with them. The results verify robustness of our estimation methods tested on both synthetic

and real-world data sets.

Appendices

1.A Descriptive statistics of hotel data sets

	Mean	Median	Std. Dev.	Max	Min
y_s^o	229.70	232.50	47.85	357	113
y_g^o	47.84	37	33.24	165	7
y^o	275.92	283	62.31	406	82
y^c	200.03	201.77	22.94	242.12	142.69
p^o	245.57	240.34	32.10	359.41	192.20
p^c	275.94	270.49	24.84	336.81	223.16

Table 1.15: Hotel M statistics, starting mid-February 2015

	Mean	Median	Std. Dev.	Max	Min
y_s^o	206.41	200.50	48.53	335	108
y_g^o	65.46	50.50	52.17	248	1
y^o	275.92	283	62.31	406	82
y^c	213.73	215.30	20.20	250.22	156.86
p^o	269.87	265.07	34.51	423.61	217.03
p^c	326.59	325.53	20.09	408.94	286.59

Table 1.16: Hotel M statistics, starting mid-July 2015

	Mean	Median	Std. Dev.	Max	Min
y_s^o	123.38	126	36.62	205	26
y_g^o	20.68	12	22.12	113	0
y^o	159.92	168	40.18	230	25
y^c	132.47	134.65	29.49	190.51	54.10
p^o	95.72	93.41	15.03	139.98	66.77
p^c	130.42	128.66	12.95	184.68	104.93

Table 1.17: Hotel V statistics, starting mid-December 2015

	Mean	Median	Std. Dev.	Max	Min
y_s^o	86.64	85	23.38	142	23
y_g^o	40.77	39	26.79	120	0
y^o	133.15	139	29.92	229	35
y^c	217.75	229.91	46.01	275.72	101.02
p^o	138.55	141.69	40.48	292.94	46.26
p^c	154.37	154.13	47.89	289.71	74.27

Table 1.18: Hotel N statistics

1.B Hotel N statistics and estimates

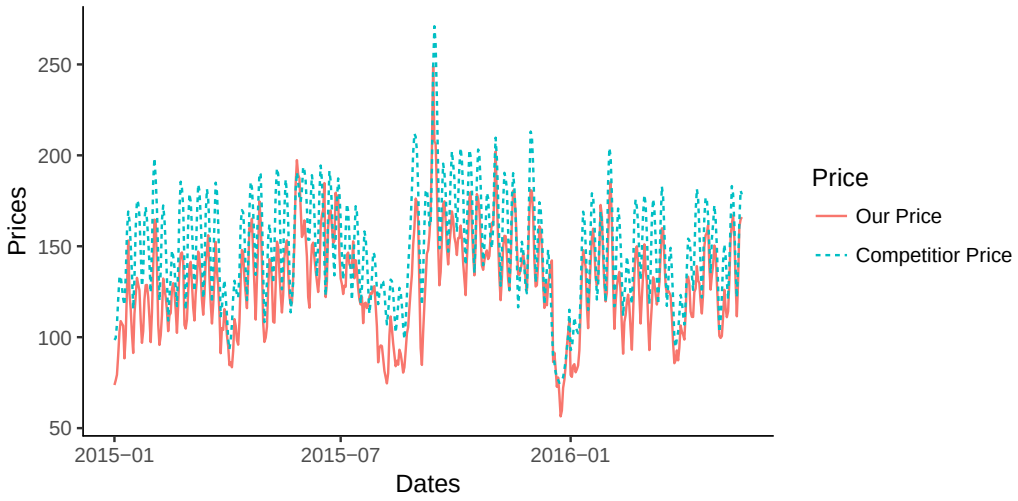


Figure 1.11: Average price pattern of Hotel N and its competitor.

Parameter	IV (sum)	IV (median)	NLS (median)
β_p	0	0	0
	(NaN)	(NaN)	(0.002)
β_a	0.48	0.48	0.872
			(0.022)
β_0	6	6	-2
Validation check			
y_s^c	139.421	139.421	206.44
y^M	225.906	225.906	927.715
$U(p^o\|p^c = 154)$	157917725.195	157917725.195	25265286.035
$C(p^o\|p^c = 154)$	157917725.195	157917725.195	7041.311

Table 1.19: Hotel N estimates over 20 weeks.

2

FACILITY LOCATION DECISIONS FROM PUBLIC DATA

Over the past few years, a significant number of optimization models for the location decision of retail facilities based on spatial customer-choice behavior have been proposed in the Operations Research (OR) literature. The practical implementations of those theoretical advancements for site selection and design, however, have been limited. Especially, small businesses do not have the time, budgets and expertise to invest in the data and analysis. Location planning is usually based on availability of sites (Pioch and Byrom, 2004), and common sense and intuition (Hernandez and Bennison, 2000). The obstacle in implementing the aforementioned advancements is twofold. First, the resources required to perform those analysis are often out of reach: both relevant data and the computational capability of integrating large data sets into modeling tend to be unavailable (Moutinho et al., 1993). Second, theoretical research is often unable to drive practical insights (Wood and Reynolds, 2012).

Specifically, location models require estimating how demand expands and shifts if a new facility is located in a hypothetical location. This is a difficult task as the firm (i) does not observe demand for the existing competitor retail facilities, and yet, (ii) needs to estimate a structural model of how demand will change as a function of multiple variables such as location, price and design of the planned facility.

For instance, in the hospitality industry higher quality is usually associated with higher prices. Thus, even when competitor demand is known but more detailed information is missing, it is hard (even sometimes impossible) to distinguish the effects of price and quality on

customer purchase decisions.

The focus of our study is a restaurant industry, in which econometric analysis is particularly challenging since unobservable factors such as taste, quality and value determine to a large extent the success of an establishment. We make use of the data that circulates on Internet web sites and social media platforms (e.g. Yelp), along with other demographic and geographic data, to tackle the challenges surrounding econometric modeling and estimation.

We propose an estimation procedure based on a latent-factor model with side-information incorporating features (e.g. good for groups, takes reservation) to explain user review ratings, and a spatial choice model to predict demand. We integrate the parameters obtained from demand and rating estimation in an optimization framework that takes into account potential future competitor entry.

Our integrated approach is novel among the existing competitive facility-location models. Due to the large capital expenditure and long term profit stream (spreading over many years) involved in locating a retail facility, and the impossibility of correcting the decision, the location should be defensible against competition, merger and acquisition activity (Poole et al., 2006).

To validate our model, we take a representative city (Las Vegas) and predict restaurant demands according to our model and public data, and then validate it by taking a sample of observed demands scraped from Google's new "Popular Times" features (that is based on mobile phone locations). The ratings and demand predictions of our model turn out be remarkably accurate, tested out-of-sample as well as against the alternate prediction made by Google.

2.1 Introduction

The facility-location problem is one of the central problems in Operations Research (OR) and Computer Science (CS). The vast majority of the existing work however focuses on optimization rather than estimation. Estimation of the parameters of the facility location model is not a big concern when the facility is intended to satisfy internal demand, such as the location of warehouses serving retail demand points; data on costs and demand is internally available and the estimation part of the models is often an afterthought.

In contrast to such facility-location applications, the recent introduction of competitive facility-location models brings the estimation task to the fore. Here the firm is considering

opening one or more retail facilities (stores, restaurants etc.) in a city and wishes to determine "optimal" locations for them. Customers choose to patronize firms based on reputation, prices, distance and the quality of the firm. The location of a new facility both expands the size of the demand in that market, and takes away as well a portion of the current demand from the existing incumbent facilities of competitors. A structural model is usually developed to estimate how the demand grows and shifts. However, to estimate such a model we inevitably need to know the sales of the firms in the area and this brings us to the central problem with these models–such data is almost never available to a firm.

In this study we link the estimation problem with CI, performed at scale, based only on public data. Our application area is service based industry like restaurants and hotels, for which customer recommendations are available and can be used to infer operational decisions such as facility location and design.

When needing insight into future customers, say for product-design decisions, firms usually resort to focus groups and survey-based research. There are some profound problems with this approach, especially in our context. First, as with all surveys, we need to get hold of a truly random and *representative* sample. For services which are "taste"-dependent like restaurants, this is not an easy task to do with any degree of conviction. Second, the time and expense involved in a proper market-research study is simply prohibitive for most small businesses. Finally, in the context of facility design, the product itself will take shape some time in the future, and there are very few instances where the firm can create a prototype and test out the ideas on a focus group.

Data-based competitive-intelligence (CI) is a promising alternative, especially when the data is readily and freely available, and this is what we propose. There is some effort involved in *how* to do it, collecting the data and estimating the model, but once established is quick and nearly cost-free. To prove our point, we base our model's estimation entirely on free publicly available data.

Our model predicts demand based only on publicly available data quite accurately. We corroborate our demand estimation by scraping observed demand for a sample of restaurants from the new Google feature called "Popular Times"[1]. To give a preview of the results, we demonstrate the relation between the predicted demand by our model and observed demand

[1]Such scraped data cannot be used directly as it is prohibited by Google; we manually collect the data on a small sample for validation purposes.

from Google for a particular city (Las Vegas) in Figure 2.1[2].

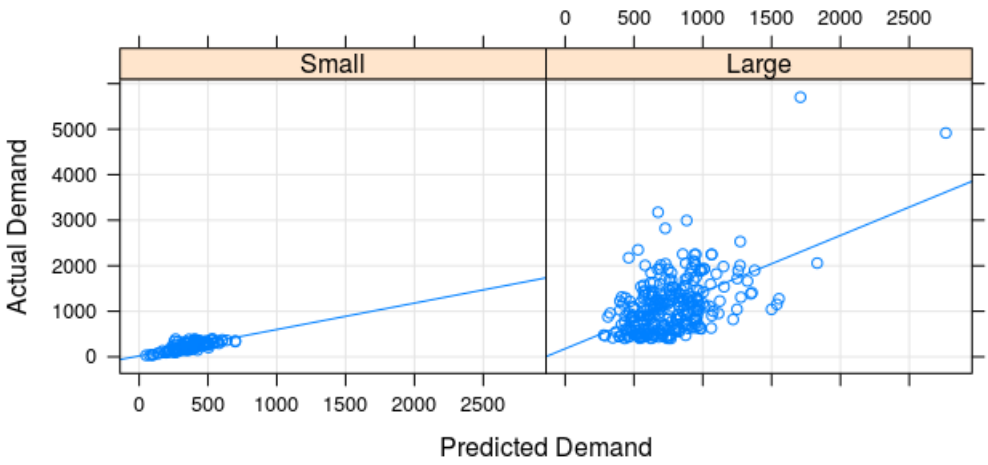


Figure 2.1: Demand comparison based on groups of small and large businesses.

A second contribution of this study is in the modeling of a firm's decision-making criteria. While the firm knows that incumbents cannot change their locations, it is typically aware that new entrants might appear in the future once it opens the facility and its profit projections will then go awry. As location involves a large fixed-cost, a site that appears profitable based on current competitive scenario may become unprofitable if a set of new entrants appear. So firms should consider not only the attractiveness of a location based on current competitive landscape but also the threat of future entry. A second-best location may be preferable to the most profitable one if it is less vulnerable to new entrants. We therefore take the objective as finding a good (profitable) but also viable (safe) location based on a long time-frame. We believe that such "future-proof" decision-making reflects how firms decide on locations, rather than on myopic or even equilibrium models.

Our approach works in three steps. In the first step we develop a customer choice model to predict demand, one of the inputs of which is the reputation of the facility, given by user ratings. In the second step we build a prediction model to explain ratings for a facility at a given location with certain features (e.g. cuisine, design and price for a restaurant) using a latent-factor model (similar to the famous Netflix ratings model) further enhanced by instrumental

[2]The classification is based on observed daily demand; small/large restaurants' daily demand is less/more than 400

variables. In the third step, we feed these estimates as an input to an optimization model to decide on the best location and product features with a long-term profit objective.

In the next section, we review the literature on related research. In § 2.3 we develop demand and rating prediction models and show the results on a representative city in § 2.4. In § 2.5 we describe the long term profit optimization model and its approximations. Finally, in § 2.6 we conclude with a summary of challenges and results.

2.2 Literature review

Our study is closely related to three streams of literature: (i) optimal competitive facility location models and their parameter estimation, (ii) data analytics, (iii) market entry and exit models.

The first stream considers spatial and product-design dimensions of a new facility. The classical facility location problem in OR (and also CS) literature deals with locating warehouses or factories with the aim of minimizing fixed cost and transportation costs. The main features of this problem are common to different contexts since it relies on a spatial abstraction: the candidate facility locations are contained within a metric space, either discrete (i.e. a network) or continuous. The optimal decision is given by an optimization function based on a criteria selected by the decision maker. Earlier works usually assume that customer choices are only affected by average travel distance or travel time (used interchangeably). Proximity is desirable for facilities such as emergency centers, schools and fire-stations whereas for facilities such as nuclear power plants and landfills, accessibility is undesirable. Based on those contextwise considerations, the location attractiveness function is formed.

The seminal work in this line is by Hotelling (1929) who studies equilibrium conditions in the case of two homogeneous facilities located on a line segment (e.g. two ice cream sellers along a beach strip). Later, Hakimi (1983) generalizes that reasoning to a network of competitors. However, once facilities differ in the total bundle of benefits, facility attractiveness levels ought to be incorporated into models (Drezner, 1994). In the case of service based industries like restaurants customers consider price, product features and quality (or reputation of the restaurant) along with proximity. Demand of the facilities are estimated using customer choice models, which integrates subjective customer “tastes" or preferences. Hence, facilities compete to maximize their profit with respect to different attractiveness criteria. These problems are known as competitive location-allocation or maximum capture problems. Indeed,

when the invested budget is fixed, maximizing profit becomes equivalent to maximizing the market share (see McFadden (1974), Ben-Akiva and Lerman (1985), Lerman (1985) on market share attraction models in the marketing literature).

Another approach to estimate market share is using gravity models originated by Huff (1964). Consumers are attracted to each facility as a function of its attractiveness and distance. Huff proposed that the probabilities are formed proportional to retail facility's size (floor area) and inversely proportional to a power of the distance to it. A complete framework to all competitive location models is that market share depends on the relationship between buying power (demand), distance and attractiveness of the facility. Ben-Akiva and Watanatada (1981) develops a continuous spatial choice logit model and demonstrates an application to urbanized area travel demand. Achabal et al. (1982) study multi-facility location problems in discrete space. Aboolian et al. (2007a), Plastria and Carrizosa (2004) extend space dimension by considering simultaneous optimization of location and design decisions. The reader can refer to Eiselt et al. (1993), Berman et al. (2009) and Ghosh et al. (1995) for a detailed overview of these problems.

Most of the above papers primarily concentrate on modeling and optimization, but relegate estimation of the models (and ways to acquire the data) to a minor role. Nakanishi and Cooper (1974) is one of the few papers, that concentrates primarily on estimating the demand at a single new facility without an optimization component. Our study contributes to demand estimation literature based on empirical methods, enhancing it with taste and quality aspects. While these factors are important, they cannot be observed, so their incorporation into models is challenging. Moreover, we tackle this challenge in a competition-aware framework using social media information.

A recent development with big-data is the prevalence of customer reviews and comments on social-media. Social media platforms not only allow us to gather information on tastes, but also serve as data to estimate the influence of such signals in customer choices. Among several others, Chevalier and Mayzlin (2006) argue that consumers use online consumer reviews as a signal of product quality. Similarly Luca (2011) finds that customer reviews affect demand for experience goods using Yelp.com reviews and restaurant data from the Washington State Department of Revenue. Moreover, Cui et al. (2017) empirically shows that aggregating social media information in sales forecast improves firms' operations decisions. Our study also contributes to this stream of research by showing how online reviews can be used to estimate demand and inform a location decision. Our application area is competitive facility location, a

relatively new branch of facility location optimization models, where customers choose which facility to patronize, and the firm has to make a location decision. In this retail facility-location context data-based "competitive intelligence" can be especially valuable as firms do not have even their own observed sales to go by when they are entering a new market.

Businesses have long indulged in CI, gathering information about competitors. Over time, such efforts have moved from manual detective-style work (Deutsch, 1990; Sreenivasan, 1998) to more modern incarnations consisting of scraping web-sites, collecting data via apps, or even poking into consumer e-mails (Isaac and Lohr, 2017; Isaac, 2017). Such practices often occupy a grey area of legality, but given the value of having key information on competitors' activities, most companies practice it to some extent or the other. For instance hotel, airline and retail firms usually subscribe to competitor price-lists, supplied by data or market-research firms which scrape publicly posted prices. The use of data-driven OR techniques, econometric modeling and machine-learning algorithms is a recent and quite powerful trend in many OM industries. Some examples of these approaches are: Caro and Gallien (2012) for the optimization of clearance prices, Fisher and Vaidyanathan (2014) in assortment optimization, Ferreira et al. (2015) in demand estimation and price optimization, Glaeser et al. (2017) in retail location decisions.

The third stream of literature deals with models of entry and exit. Most competitive location models in the literature decide on the optimal location of the new facilities by maximizing the current market share (or profits) without considering future evolutions of the market. The standard spatial interaction model, for instance, assumes total market size is fixed (inelastic demand assumption) and ignores the market expansion effect, which is the increase of total customer demand in a certain region once new facilities are located in that region, and cannibalization effect, which occurs when new facilities take away some of the demand from pre-existing facilities.

Establishing a facility involves a large fixed cost and the expectation is that once opened the facility stays profitable for a long period of time. Research efforts focused on identifying the factors that are critical to the success and failure. Parsa et al. (2005) present a comprehensive framework from managerial, economic and marketing perspectives on the reasons why restaurants fail. Zhang and Luo (2016) indicate the volume and valence of online reviews as the strong predictors of restaurant survival. For extensive literature survey on strategic facility location, we refer the reader to Owen and Daskin (1998). Here we focus on the role a future competition's entry/exit decisions plays on our present location decisions.

Berman and Krass (2002) introduce non-constant expenditure functions to capture both market expansion and cannibalization effects. However, it results in computationally challenging optimization problems and therefore, Berman and Krass (2002) analyze the monopolistic case only. Aboolian et al. (2007b) extend the analysis to concave demand case. Pancras et al. (2012) develop a demand model to account for latent goodwill dynamics, location endogeneity and spatial competition of proximate retail outlets. Farahani et al. (2014) and Drezner (2014) are examples of the survey papers on future competition.

The competition-aware investment literature typically assumes two competitors are aware of each other and it has a game-theoretical or competitive equilibrium type of modeling. Preemption is used as the main strategic tool, and optimal timing of investment is decided. Our main modeling contribution in the optimization part is location decision incorporating fixed-costs and anticipating that new facilities will be opened in the future. Such anticipation is different from traditional competitive models, in the sense there is no game or strategic response involved (at least not on the location decision). Future profits, however, are affected by other firms' decisions, and investing a large fixed cost without taking future location decisions and potential competition into considerations would be foolish. Drezner and Drezner (1998) and Plastria and Vanhaverbeke (2008) in the Location Science area, and Lambrecht and Perraudin (2003) in Economics, and Ghosh and Craig (1983) in Marketing (which uses a competitive equilibrium model) are some prominent examples of such modeling. We differ from the former in not assuming there will be a single competitor and an eventual split in the market share; rather, over the horizon of interest, a series of new firms may open up and the probability of each opening is proportional to the profits at that stage.

In Industrial Organization field, Pakes et al. (2007) develop estimation strategy exploiting information asymmetries to transit from commonly used two period setting to dynamic models of entry and exit. Similarly, Bajari et al. (2007) and Aguirregabiria and Mira (2007) are the theoretical papers on estimation of dynamic games. Dunne et al. (2013) and Collard-Wexler (2013) are some examples of empirical applications for investment models.

Trigeorgis (1991) investigates the impact of competition on the optimal timing of project initiation using an options analogy. Similarly, Lambrecht and Perraudin (2003), citing Walmart as an example, considers the investment decision of a firm when there is a first-mover advantage by being able to completely block out a single future potential competitor. This would be a special case of our model where the fixed costs of opening a facility is greater than the future expected revenues for a competitor once we locate a facility in that area; however

we deal with a much more complicated and feature-rich probabilistic model of how future competition evolves, on a larger spatial and longer time scales.

All in all, our study fills an important gap by combining estimation of customer choice models and optimization of facility location and design to maximize the long term profit considering future evolutions of the market based on customer tastes, demographics and size. In the next section, we develop the demand and rating prediction model.

2.3 Demand and ratings model

In this section we describe a spatial-choice model of demand at a proposed facility. Discrete choice models are natural—at least amongst relatively tractable options—for modeling customers making independent decisions based on product features. They are very popular and are widely used in transportation, industrial organization, marketing and revenue management literature.

We model the customer choice behavior using Multinomial logit (MNL) framework, the most commonly used random utility model in economics and marketing (the estimation of which in realistic conditions being the subject of our previous chapter). MNL functions are often used in facility location literature to derive the market share of the facilities (De Palma et al. (1989), Benati (1999), Marianov et al. (2008) and Haase and Müller (2014)).

Demand is modeled as a population of potential customers who make a choice amongst available alternatives, in this case to visit one of the retail facilities, or choose not to take any action at all, the no-purchase option. The choices are affected by the location of the facilities and the customer (via the distance between the two), the individual's idiosyncratic tastes as well as the features of the facility. Note that many of the features, such as quality and the individual's tastes may be unobservable which complicates estimation. In our context, a further complication is the need to predict how customers would perceive and rate the new facility as a function of its features and their tastes. We deploy a latent-factor model reminiscent of recommendation-engine technology, embedded in the aggregate choice model.

2.3.1 Spatial-choice model of aggregate demand

Facilities can be of T different types (e.g.: cuisine types) and are located over a region divided into G rectangular grids. Each facility also has additional features such as being suitable for

groups or kids, taking reservation, having delivery service etc..

Customer behavior is modeled as a two-stage process: First, a customer located in a grid g chooses a facility of some type t based on features (including the rating that others have given) and its distance from where they are located. Next, a fraction of the customers leave a rating for that establishment based on their own idiosyncratic tastes and quality (both unobservable to us), and (observable) features of the facility. This leads to an observable ratings distribution for that facility, usually on a scale of 1 to 5.

Consequently, we also estimate our model in two stages. First, we model ratings as a function of demographics, features and unobservable latent factors such as tastes and quality. Then, we model demand for the establishment as a function of the obtained ratings and the location of the facility. This is done for two reasons: One, because the choice of the facility and the ratings happen at two distinct decision points in time in the customer journey so the decision on facility choice is based on a set of existing ratings given by others (that is ratings form expectations of quality). Two, even though the underlying factors driving demand and ratings may be the same and are deeply confounded, predicting rating first (for the proposed facility) and *then* demand leads to a tractable model. The estimate of the ratings encapsulate many of the unobservable elements (quality, service, idiosyncratic tastes) in practice, that makes the prediction of demand more robust.

Customers are utility maximizers and utility of each alternative is a function of observed and unknown (random) components, ϵ. The primary drivers of utility are rating of the establishment r_i—that encapsulates both the unobservable quality as well as price—and the distance between the customer located in grid g and the facility i, d_{ig}. The utility of a facility i for a customer located at grid point g is given by $U_{ig} = V_{ig} + \epsilon_{ig}$ where V_{ig} is the deterministic part and ϵ_{ig} a random component, that, for its parsimonious, tractable properties, we assume to be i.i.d. with a Gumbel distribution.

Our model of the deterministic component coincides with what is typically done in competitive facility location (Berman et al., 2009), except we introduce a ratings variable which serves as a proxy for value namely,

$$V_{ig} = \alpha R_i + \beta d_{ig}$$

where $R_i = S_i - \gamma p_i$, with S_i represents random (so modeled as it is unobservable and idiosyncratic) quality and p_i is the price. The random variable R_i then is the rating distribu-

tion of facility i. For simplicity we take $V_{ig} = \alpha r_i + \beta d_{ig}$, i.e., based on the average rating $r_i = s_i - \gamma p_i$, where $r_i = E[R_i]$ and $s_i = E[S_i]$. The parameters α and β are estimated from data. We do not estimate γ as in the data we neither have information on quality nor a very accurate price (only a price "category" of the facility).

The probability that a customer located at grid g chooses restaurant i is defined by the well-known expression of the multinomial-logit (MNL) model (as a consequence of the independence and extreme-value distributional assumptions):

$$Pr_{ig} = \frac{e^{V_{ig}}}{\sum_{i \in \mathcal{C}} e^{V_{ig}}}.$$

The reader can view this as a market-share prediction when the consideration set $\mathcal{C}$ is the same for all the customer population.

Although neither value nor idiosyncratic tastes may be observable we enhance the demand model with demographic information as follows. We obtain demographic information from U.S. Census Bureau (2014) so that we incorporate income into demand estimation by multiplying the number of people residing at each grid with the mean income level of the corresponding zip code to obtain buying power. Let the buying power of customers residing at grid g be denoted by $M_g,\ g \in G$.

We assume that a fraction μ_i of the visitors leave reviews and demand at any restaurant i is given by

$$n_i = \mu_i \sum_g M_g Pr_{ig}. \tag{2.1}$$

Summing this across all restaurants in the city, we obtain total number of reviews

$$n = \sum_i n_i = \mu_i (\sum_g M_g \sum_i Pr_{ig}) \tag{2.2}$$

Now, since we do not observe demands at the facilities, but assume a constant proportion who leave feedback, we can at least get a share of the total patrons of all the facilities by dividing (2.1) to (2.2), we obtain

$$\frac{n_i}{n} = \frac{\sum_g M_g Pr_{ig}}{\sum_g M_g \sum_i Pr_{ig}} = \frac{\sum_g M_g Pr_{ig}}{\sum_g M_g}. \tag{2.3}$$

	OpenTable Booking
(Intercept)	18.834**
	(3.903)
Yelp Reviews	0.038***
	(0.002)
R^2	0.635
Adj. R^2	0.626
Num. obs.	41
RMSE	26.757

Note. Robust standard errors are in parentheses. $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 2.1: The relation between Yelp number of reviews and OpenTable bookings.

Value and tastes (and hence the ratings) may be explained by a number of observable features of the facility (such as parking, views, ambience etc.) and in the next section we develop a latent-factor model with side information to explain ratings, and hence partially predict the rating of a future hypothetical facility with certain design features.

Estimation of the constant scale As we mentioned, without competitor demand information it is impossible to estimate parameters such as μ_i that would have allowed us to predict demand from reviews. However, we often have partial information that can help us gauge our estimations and obtain a reality-check. To calibrate market share information into demand, we can use a sampling of such data sources.

For instance, for the restaurant industry, we have data published by a reservation site called OpenTable. Another source is Google with its "Popular Times" feature. We use the former to calibrate and the latter to validate the results of our model. To calibrate our model, we collect a sample of booking data from (OpenTable, 2016) and verify that the correlation between number of reviews and demand is quite high, for this sample it is 75.05%. The relation is shown in Figure 2.2.

We are aware that the observed demand at the Opentable platform alone is not complete as reservations can be made by phone-calls or customers walk-in without a reservation, so we show in § 2.4.3 the performance of booking estimates with respect to realizations of bookings with a more complete demand information from Google.

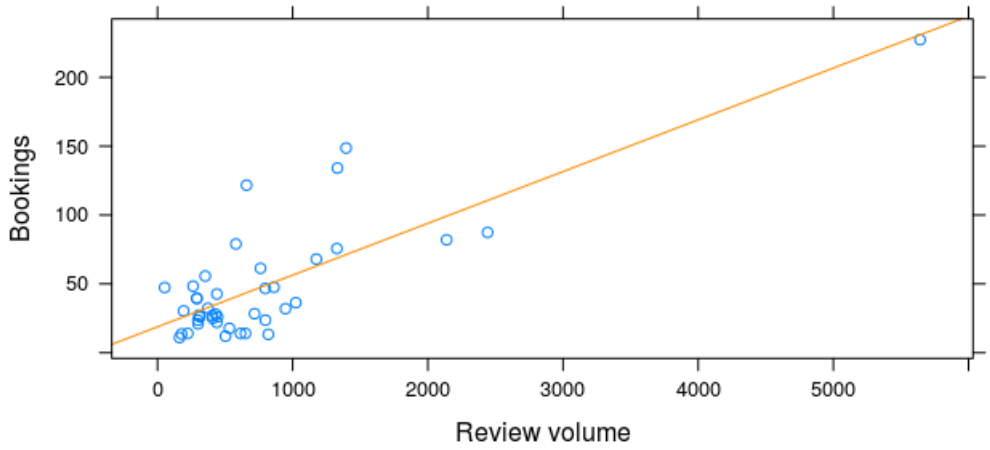


Figure 2.2: The relation between Yelp reviews and OpenTable bookings .

2.3.2 Conventional model for ratings prediction

We need to accurately predict the ratings of a new facility, and we do so by explaining the ratings of existing facilities as a function of their features as well as the characteristics of the regions—more accurately, the customers in those regions—where they are located. The spatial resolution for our model is zip-code level as demographic information is usually available only at this level (density however is usually available at a finer level).

One may be curious how a conventional model where we regress ratings using dummy variables for zip-codes as well as cuisines and features performs to estimate demand. It so turns out that traditional regressions methods were unable to tease out the latent factors and give some obviously erroneous estimates for the co-efficients.

In this part, we apply the conventional ordinary least squares (OLS) regression using dummy variables for zip, cuisine and features of a restaurant. The rating of a restaurant is given by

$$r_i = \zeta_0 + \boldsymbol{\zeta}^T y_i + \boldsymbol{\delta}^T z_i + \boldsymbol{\gamma}^T c_i + \epsilon$$

where the parameters are ζ_0 (intercept), $\boldsymbol{\zeta}$'s (covariate weight vector on features), δ and γ are, respectively, the coefficients of zip and cuisines. Note that $\boldsymbol{\zeta}$ is common to all facility types. Among design features, we also include price variable on a scaled basis, specifically 0.25 for \$, 0.50 for \$\$, 0.75 for \$\$\$ and 1 for \$\$\$\$, to estimate its effect on ratings.

The estimated coefficients are given in "No correction" line of Table 2.8. However, price coefficient turns out to be 0.054. Accordingly, as the prices increase rating should increase, too! The estimate is certainly biased. Later, in §2.3.4, we argue about the reason of the problem and suggest a plausible way to correct it.

2.3.3 Latent factor model (LFM) with side-information for ratings prediction

In this part we adopt a latent-factor model in style, commonly used in recommendation engines, with the mathematical problems behind it going under the name of Incomplete Matrix Completion problems or Matrix Factorization methods. These methods came into prominence because of the Netflix prize (Koren et al., 2009). We enhance the model with side-information.

Each facility i of cuisine-type c has some facility-specific information (e.g: for restaurants, taking reservations, capability to serve for groups etc.) that are captured in y_i. We propose a ratings model that makes a prediction of the rating distribution of a facility given as inputs the planned features and the latent factors between the location (at a zip code level where the facility is planned to be located) and the cuisine.

The rating distribution model of a facility i is given by

$$r_i \sim Poisson(\zeta_0 + \boldsymbol{\zeta}^T y_i + \boldsymbol{u}_c^T \boldsymbol{v}_z).$$

The latent factors are $\boldsymbol{v}_z$ and $\boldsymbol{u}_c$'s. The latent factor vector $\boldsymbol{v}_z$ is zipcode-specific and $\boldsymbol{u}_c$ is cuisine-specific, and both are of the same dimension (we take a dimension of k=2 latent factors). The aim is to find the optimal parameters to minimize a certain form of loss function. In our case, to learn the latent factors and unknown coefficients of facility features and zip code demographics we minimize the regularized squared error on the set of known ratings

$$\min_{\zeta_0, \boldsymbol{\zeta}, \boldsymbol{u}_c, \boldsymbol{v}_z} L$$

where

$$L = \sum_{i \in \kappa} (r_i - \zeta_0 - \boldsymbol{\zeta}^T y_i - \boldsymbol{u}_c^T \boldsymbol{v}_z)^2 + \underbrace{\lambda(||\boldsymbol{u}_c||^2 + ||\boldsymbol{v}_z||^2 + ||\boldsymbol{\zeta}||^2)}_{\text{Regularization}}$$

and κ is the training set for which r_i is known, λ is the regularization term to avoid overfit. This problem is not convex in general, however over the last few years the alternating gradient-descent method has proven to perform well among the recommender systems community.

To give a preview of the latent factor model, we briefly describe the out-of-sample performance for our restaurant data. We first randomly split the data into training and testing data set with 3:1 ratio. Then, we use the training data to build regression based models and we test the performance of models on the testing data set. We apply this procedure at the restaurant level with the addition of design features that are binary coded (1 if the feature exists, 0 otherwise). Estimated coefficient weights of different features on rating and root mean square errors (RMSE) are given in Table 2.6 and Table 2.7. Accordingly, LFM with biases and features model performs the best as both the test and training data RMSE is the lowest. Moreover, the price coefficient being -0.051 ("No correction" line) in Table 2.7 is notable as it is nearly zero and we analyze this further with instrumental variables in §2.3.4.

2.3.4 Endogeneity

Now for both the prediction models of §2.3.2 and §2.3.3 if we wish to estimate rating, and indirectly the demand, as a function of price, we run into a problem. If we take ratings as a proxy for quality and do a straightforward regression say, it turns out that the higher the price point, the higher the rating. This is a typical occurrence in econometrics, as we do not correct for unobservable quality of the restaurants—simply put, higher-priced restaurants may just be giving more *value*, defined as quality minus price (or alternately quality divided by price), and we cannot observe quality.

As we mentioned before, if we use a standard logic to apply either LFM of §2.3.3 or the conventional dummy variable regression of §2.3.2, we run into a problem: the price coefficients are around 0 and in the conventional regression case it is even worse, it comes out positive. Both of the results are given in "No correction" line of Table 2.7 and Table 2.8, respectively. This is of course the problem of endogeneity in the regression—regressors being correlated with the error term because of omitted variables. Using instrumental variables (IV) is the most popular way to correct for it . . . if we can find valid and strong instruments.

We use income and age information of the zip codes as the instruments, as there is no reason to believe they are correlated with the quality of the restaurant, and we can expect correlation with price—high-end restaurants are more likely to be present in zip-codes where

the income levels are high; similarly for age.

For both the latent-factor model of §2.3.3 and the conventional regression model of §2.3.2, we use the two-stage least squares (2SLS) technique. Firstly, we illustrate it in the case of conventional regression model. The procedure consists of two-steps.

First step: We use OLS to estimate the linear projection of the endogenous explanatory variable, specifically price, on the instruments income and age and the rest of exogenous variables, in our case, binary coded features of the restaurant.

$$price = \theta_0 + \boldsymbol{\theta}^T \begin{bmatrix} \text{features} \\ \text{zip dummy} \\ \text{cuisine dummy} \\ \text{income} \\ \text{age} \end{bmatrix} + \epsilon.$$

Using the estimated coefficients in the first step, we obtain the estimated values, $\widehat{price}$.

Second step: We use OLS to regress ratings on $\widehat{price}$ and the rest of exogenous variables,

$$\text{rating} = \zeta_0 + \boldsymbol{\zeta}^T \begin{bmatrix} \text{features} \\ \text{zip dummy} \\ \text{cuisine dummy} \\ \widehat{\text{price}} \end{bmatrix} + \epsilon.$$

In the latent model case, we do a variation of 2SLS as follows: The first step is identical, however in the second step instead of using OLS, we use the LFM.

Using income and age demographical information corrects endogeneity bias and we get meaningful coefficient estimates. As we show in "IV correction" line of Table 2.7, price coefficient turns out to be -0.511. However, conventional dummy variable model is not able to capture the noise as shown in "IV correction" line of Table 2.7, the price coefficient is 4.237. So, we rely on LFM based with IV correction.

Also, we perform diagnostics test to validate our choice of instruments (Table 2.9). According to weak instrument test, we reject the null so that our instruments are not weak. In the case of Hausmann, we again reject the null so we can claim OLS and IV estimates are not similar, and endogeneity was a problem in our data set. Lastly, we do not reject the Sargan test meaning that we do not have any proof that the instruments are invalid.

2.4 Estimation for the case of restaurants

We chose the restaurant industry to apply our model, and in the later §2.5 apply facility-location optimization to choose the right location, type and design.

As we mentioned earlier the restaurant industry is particularly challenging to estimate as the unobservable demand itself is a function of non-quantifiable measures of quality and tastes. Moreover, in our particular context, we do not have precise information of price either, except a crude classification into price bands (typically with symbols such as $$). It is also an industry where data-based CI would be the most useful, as finding a representative sample to conduct a survey or market research is a nearly impossible task. In dense tourism-oriented cities (like the representative city Las Vegas), there is also an issue with transient population that does not show up in demographic population data.

2.4.1 Public data for restaurants

In this section we describe the public data for the restaurant industry. To be specific, in our estimation we pick one representative city, Las Vegas, NV with 3124 restaurants.

We collect restaurant data from Yelp Dataset Challenge (Yelp, 2016) provided by Yelp.com. Yelp, founded in 2004, is a website that customers can write online reviews and give star ratings upon visiting a restaurant. As of 2016, it has 135 million monthly visitors and 95 million reviews. The data set contains business (e.g. cuisine type, price range—number of dollars, 1$ to 4$, location (latitude/longitude coordinates, zip code) and review (e.g. number of reviews, average review rating, distribution of ratings) data. It also provides qualitative information on restaurant attributes such as takes reservation, delivery and parking service as shown in Figure 2.9. For our representative city Las Vegas, NV, this data has 426743 customer reviews for restaurants spread over 38 zip codes. There are 31 cuisines.

We obtain free and public spatial consumer distribution from NASA (2016) that provides population density given at a very fine grid level. The grid dimensions are 0.927 km in height and 0.752 km in width and there are a total of 1419 grids in our representative city.

In a city such as Las Vegas, a significant percentage of demand in certain locations is made of transient tourist population staying in hotels. Data from Factual (2016) gives exact geo-location of hotels along with the number of rooms. According to U.S. Hotel occupancy rate statistics Statista (2016), the hotels are 63% occupied, on average. So, we multiply this rate with the number of rooms to find the average tourist population at hotels. For instance local

Variable	Mean	Median	Std. Dev.	Max	Min	Count
Reviews						
Star Rating	3.61	3.5	0.6	5	1.5	1,928
Count	237.34	119	393.63	5,642	26	1,928
Demographics						
Local Pop.	1,328.85	821.19	1,480.12	10,292	0	1,526
Tourist Pop.	413.43	138.6	647.22	4,147.29	12.6	250
Income	66,032.45	66,785.44	17,595.50	104,708.00	33,417.09	39
Cost						
Rent ($/sq ft)	7.65	1.96	12.13	52.51	0.1	39
Restaurant						
Capacity	109.61	84	81.06	390	6	221
Size (sq ft)	1,252.14	1,080	726.5	4,500	50	221

Table 2.2: Summary statistics of public data

population of Las Vegas is 2027828, after accounting for transient population, it increases to 2097671 averaged over a year.

We obtain demographic information from the U.S. Census Bureau. Income distribution especially will serve us as a key instrumental variable to estimate price-sensitivity.

In addition we use some other sources (such as OpenTable and Craigslist and Google) indirectly, to calibrate and validate our estimation procedures. As we mentioned in §2.3.1, not all parameters of our model are estimable without observing demand. Our objective is to estimate up to a single non-estimable scale factor, and make inferences on this parameter either through domain knowledge or observable factors. The public data sources are summarized in Table 2.3 and summary statistics data is provided in Table 2.2.

We present the results from our demand model estimation in Table 2.4. Accordingly, we observe that rating coefficient is positive (0.496), so star ratings have a positive effect on the utility of dining at a particular restaurant. However, distance coefficient is negative (-0.071), so distance has a negative effect. These elasticity estimates are in line with the ones reported in literature (see, for example Pancras et al. (2012)).

2.4.2 Results for our representative city

In this section we describe the results for our representative city Las Vegas, NV. Recall that our goal is to explain ratings as a function of some causal factors, so that we can use such

Source	Information/Field
Yelp Dataset Challange	Total number of reviews (proxy for demand), user rating distribution, restaurant location (lat., long.), average star rating of a restaurant, cuisine type, price range (1-4 $),
NASA, SEDAC (2015)	Gridded population count by lat-long., 0.927 km in height, 0.752 km in width
Factual API	Hotel location (lat. long.) and number of rooms
Google Maps API	Retrieves zip codes from lat-long. coordinates
OpenTable	Daily bookings for a sample of restaurants, used to correlate with number of reviews
Craigslist	Commercial rent prices at zip code level, ($/sqft)
United States Census Bureau American Fact Finder (2014)	Population: age, sex, race, households: total households, family households, income: household income, families, married-couple families, nonfamily households
Popular times feature, Google Maps	Restaurant occupancy in percentage
Restaurant seat and size (square footage)	City of Las Vegas Building Department

Table 2.3: A listing of all our public data sources we used to develop the model estimation

	Demand Ratio $(\frac{n_i}{n})$
α	0.415^{***}
Avg. Rating	(0.064)
β	-0.081^{**}
Distance	(0.024)
Observations	1928

Note. Robust standard errors are in parentheses. $^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$.

Table 2.4: Demand estimation results

estimation to gain insight as well as inputs to an optimization model for facility location. We describe the results for both the latent factor model of §2.3.3 as well as the more conventional model of §2.3.2.

We place the ratings data from Yelp in a matrix of 38 zip codes and 31 cuisine types. Each cell in the matrix can have multiple restaurants and moreover, a rating is in fact a distribution given by fraction who rates the facility as 1, as 2 etc. Each restaurant also has additional features, the so-called "side-information" in latent-factor models, that in our case is any combination from 3 design options (no design, groups for groups, takes reservation option).

The total number of unique zip-cuisine-design pairs in our restaurant data set is 1102. So, sparsity level is $1 - \frac{1100}{38 \cdot 31 \cdot 4} = 76.7\%$. We use Root Mean Squared Error (RMSE) to evaluate

accuracy of predicted ratings:

$$\text{RMSE} = \sqrt{\frac{\sum(r_i - \hat{r}_i)^2}{T}}$$

where T is the number of observation, r_i is the actual rating for restaurant i and $\hat{r}_i$ is the predicted rating. As the first cut results, we compare actual and mean of predicted ratings given by the rate of Poisson distribution on a sample of cuisines (Table 2.5). The predicted ratings are quite close to actual values. We also test rating model estimates with and without side information. Overall, the model performs quite well, we observe relatively low RMSE after convergence; the errors are slightly lower once we include features (Table 2.6). Also, addition of features enhances our knowledge about each features' contribution on ratings.

Cuisine	Actual Rating	Predicted Rating
American	3.558	3.539
Asian	3.500	3.672
Bakeries	4.000	3.998
Barbeque	3.116	3.575
Buffet	3.000	3.500
Italian	3.600	3.570
Japanese	4.250	3.846
Mexican	3.833	3.586
Pizza	3.750	3.523
Steak	3.937	3.876
Sushi	3.500	3.889

Table 2.5: Comparison of the average predicted and actual ratings at zip code 89110 for some of the cuisines. $r \sim \zeta_0 + u_c + u_z + u_c^T v_z$.

	Plain LFM	Enriched LFM with Biases	LFM with Biases and Features
Model (r_i)	$. \sim u_c^T v_z$	$. \sim \zeta_0 + u_c + v_z + u_c^T v_z$	$. \sim \zeta_0 + u_c + v_z + u_c^T v_z + \boldsymbol{\zeta} y_i$
Test RMSE	0.582	0.564	0.560
Train RMSE	0.551	0.544	0.539

Table 2.6: Comparison of several model performances. LFM (plain), LFM with biases, LFM with biases and features

	Constant Term	Good for Groups	Takes Reservations	Standardized Price	RMSE Train (Test)
No correction	$\zeta_0 = 2.874$	$\zeta_1 = 0.255$	$\zeta_2 = 0.288$	$\zeta_3 = -0.051$	0.650 (0.701)
IV correction	$\zeta_0 = 3.022$	$\zeta_1 = 0.264$	$\zeta_2 = 0.365$	$\zeta_3 = -0.511$	0.645 (0.701)

Table 2.7: LFM estimates without and with IV correction.

	Constant Term	Good for Groups	Takes Reservations	Standardized Price	RMSE Train (Test)
No correction	$\zeta_0 = 3.360$	$\zeta_1 = 0.243$	$\zeta_2 = 0.181$	$\zeta_3 = 0.054$	0.631 (0.693)
IV correction	$\zeta_0 = 1.651$	$\zeta_1 = 0.298$	$\zeta_2 = -0.376$	$\zeta_3 = 4.237$	0.793 (0.877)

Table 2.8: Conventional regression estimates without and with IV correction.

2.4.3 Validation from Google's "Popular Times"

In this part, we measure the quality of our models to show how well our model estimates of star ratings and bookings predict the future real-world evaluations of the market. We conduct the analysis at restaurant level. Realized star ratings are already available at Yelp. So, we perform holdout sample validation by estimating the ratings model on a randomly chosen training set and test the estimates on test (or validation) set. We added each explanatory variable (restaurant features) of ratings one at a time and compared the test errors.

The linear relationship between actual and predicted ratings is given in Figure 2.3, correlation coefficient is 0.444. Then, we compare the ratings based on groups of star ratings categorizing them to low, medium and high as shown in Figure 2.4. However, demand is not readily available. Remember that in our model we use review volume as a proxy on booking. To obtain realized bookings, we gather seat information for a small subset of restaurants in Las Vegas from the Building and Safety Department of Nevada (Las Vegas City Hall, 2014). We also obtain the occupied percentages using popular times feature of Google Maps at a restaurant level. We cross these two data sets to obtain bookings with an assumption on the average time spent at a restaurant. We assume price range category is positively correlated with average time spent. The demand estimation model did not incorporate the capacity as we lack this information for all set of restaurants. Yet, we need to integrate the capacities to validate the demand predictions. The predicted constrained demand is given by $E[D|D < C]$. The following steps outline our constrained demand calculation:

- Take the unconstrained demand estimation, $\hat{d}$ and distribute it according to the popular times curve at a hourly basis.

Diagnostics	df1	df2	statistic	p-value
Weak instruments	2	3089	42.812	<2e-16***
Wu-Hausman	1	3089	5.396	0.0203*
Sargan	1	NA	1.906	0.1675

Note. $^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$.

Table 2.9: Diagnostics tests on instrumental variables.

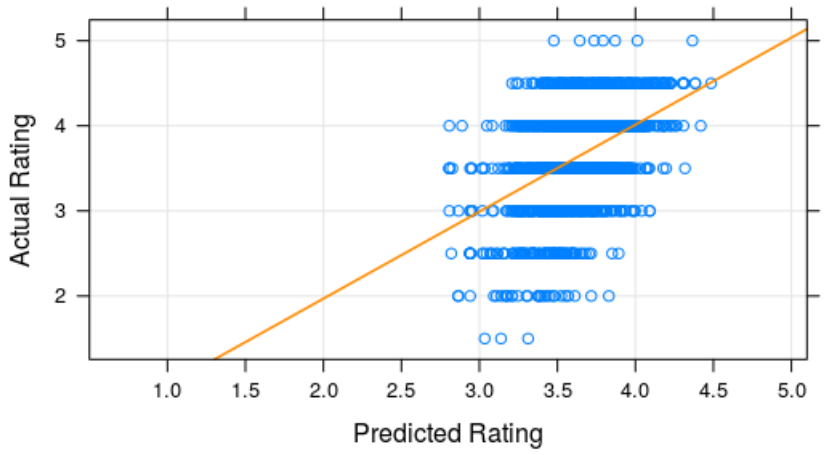


Figure 2.3: Rating comparison of the latent-factor method applied on test set.

- For each restaurant, calculate hourly capacity $x = \frac{\text{Total seats}}{\text{Average time spent}}$.
- Predict hourly demand using $\min\{\hat{d}, x\}$ and sum it over the open hours to obtain the daily bookings.
- Repeat the procedure for each day, and take the mean to find the average daily constrained demand.

We demonstrate the relation between actual and predicted bookings by groups of small (actual bookings are less than 400) and large businesses (actual bookings are more than 400) as given in Figure 2.1. We also carry out the regression based on small and large businesses (Table 2.11 and Table 2.12). The performance of demand model is remarkably accurate. Especially, in the case of small businesses, we obtain $R^2 = 0.591$. However for both cases we underestimate the demand. This can be because we do not have actual bookings information and we use

	Dependent variable:
	Actual Rating
Predicted Rating	1.022^{***}
	(0.054)
Constant	-0.074
	(0.196)
Observations	482
R^2	0.198
Adjusted R^2	0.197
Residual Std. Error	0.539 (df = 480)
F Statistic	355.652^{***} (df = 1; 480)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 2.10: Linear relationship between actual vs. predicted rating

review volume as a proxy on bookings. So, we are omitting the presence of repeat customers in the demand estimation model, as they are unlikely to write online reviews after the first time visit.

2.5 Facility-location optimization with entry

The estimation results of §2.4 give valuable insights to a firm on location, customer preferences and the effects of price and design features. Note that all this is without possessing any direct data on demand at even one location.

In this section we incorporate the estimation into an optimization framework to find the best possible location, facility-type and features for a firm contemplating opening a new retail facility. One novel feature of our framework is that we incorporate the threat of future entry of competitors.[3]

We believe competitive entry is a very important consideration that is currently missing in most retail facility-location models—location involves large fixed costs, and once a facility is

[3] The model can be potentially extended to incorporate future exits also, but we leave it for future research.

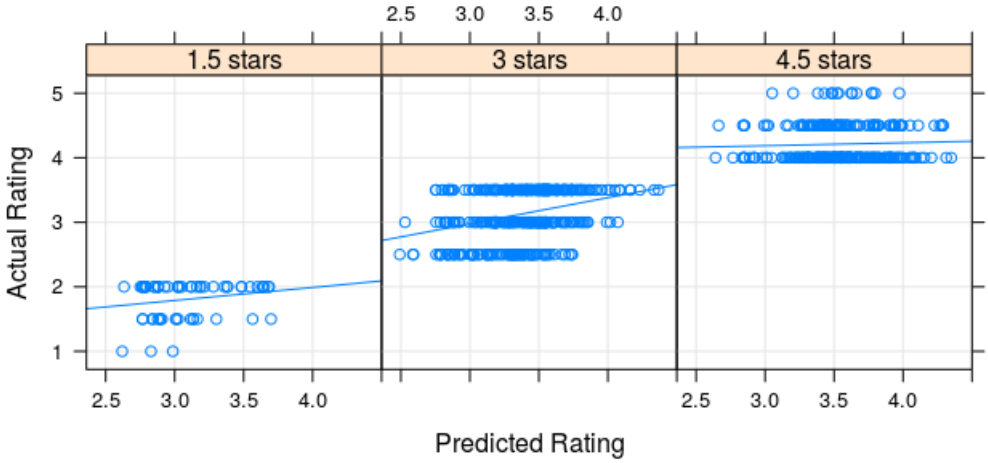


Figure 2.4: Rating comparison of latent-factor method applied on test set based on groups of rating categories.

situated it is usually not easy to move the business[4]. The general expectation is that the investment will be recovered from a steady future profit stream; however such profit calculations can go awry if competitors locate nearby. It is very unlikely that any rational firm would not be aware of these risks and take them into account in its facility-location decision. Therefore any model that ignores the risk of competitor entry would leave out an important consideration.

In this section, we model the probability of an entry in a period as proportional to the profitability of the location and facility-type. We assume the size of the population and its characteristics to remain stable; while demographics do change over time, it is usually on a longer time-scale and much harder to forecast, so we do not model changes in the population, or its tastes.

We assume that all the grids in the network are candidates for location of the new facility. To reduce notation, we let a facility of *characteristic* x represent a combination of location (grid g), type c and price-point p, (g, c, p), with $x = 1, \ldots, N$. We let all possible characteristics as $\mathcal{X}$ which includes the $\emptyset$. To make a facility-location decision, we need to have some estimates of costs. These costs are typically of two different types, fixed (independent of demand) and variable (linear in demand). The fixed costs can be further classified as one-time

[4]The classical model of Hotelling gives ice-cream vendors as an example, which are mobile and therefore location is a strategic dimension; however most facility location decisions in Operations Research are for facilities that cannot be moved at all, like factories or retail outlets.

	Dependent variable:
	Actual Demand
Predicted Demand	0.580^{***}
	(0.046)
Constant	20.415
	(17.613)
Observations	112
R^2	0.591
Adjusted R^2	0.588
Residual Std. Error	62.977 (df = 110)
F Statistic	159.219^{***} (df = 1; 110)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 2.11: Actual vs. predicted rating fit for small sized restaurants.

cost to set up the facility at x as F_x (say construction or renovation or goodwill cost), and an operating fixed cost (such as rent, employee costs), f_x. We include the dependence on characteristic x as costs are likely to depend on location, the type as well as price-category one wants to compete in.

Our state space is a vector $\boldsymbol{n}$ whose elements are the *number* of facilities of characteristic $x = g, c, p$. We denote $\boldsymbol{n}^t$ if we want to specify the vector at time t. We call this the facility *profile* at time t. We use the indicator function $\mathbb{1}_{[x]}$ to represent a vector with a 1 in the position corresponding to x. If $\boldsymbol{n}$ is the current profile and if a new one facility opens with feature set x, we update the vector as $\boldsymbol{n} + \mathbb{1}_{[x]}$ to represent the new profile (so if we are in period t, $\boldsymbol{n}^t = \boldsymbol{n}^{t-1} + \mathbb{1}_{[x]}$).

We let ρ be the discount factor (determined exogenously, say 5%) that represents time-value of money. We model an infinite time-horizon dynamic program, and we assume that in each period there is at most one or no entry from a new competitor.

We let $r(\mathbb{1}_{[x]}, \boldsymbol{n} + \mathbb{1}_{[x]})$ be the per-period profit (assumed stationary) if we locate x given the current competitive profile is $\boldsymbol{n}$ and let $R^t(\mathbb{1}_{[x]}, \boldsymbol{n}+\mathbb{1}_{[x]})$ as the discounted expected profit from period t onwards (subject to some law of evolution of the profile $\boldsymbol{n}+\mathbb{1}_{[x]}$). The objective

	Dependent variable:
	Actual Demand
Predicted Demand	1.244^{***}
	(0.121)
Constant	178.246^{*}
	(98.429)
Observations	274
R^2	0.281
Adjusted R^2	0.278
Residual Std. Error	548.915 (df = 272)
F Statistic	106.376^{***} (df = 1; 272)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 2.12: Actual vs. predicted rating fit for large sized restaurants.

function to determine x is

$$\max_{x} \sum_{t=0}^{\infty} \rho^t E[r(\mathbb{1}_{[x]}, \boldsymbol{n}^t)] \tag{2.4}$$

with the expectation over $\boldsymbol{n}^t$ evolving according to a law that we specify shortly. But we first need to define a few other quantities.

The optimization problem can be stated in the form of a dynamic program as

$$\max_{x} R^0(\mathbb{1}_{[x]}, \boldsymbol{n}^0)$$

where $R^t(\mathbb{1}_{[x]}, \boldsymbol{n}^t) = r(\mathbb{1}_{[x]}, \boldsymbol{n}^t) + \mathrm{E}\{\rho R^{t+1}(\mathbb{1}_{[x]}, \boldsymbol{n}^{t+1})\}$ with the state in $t+1$, $\boldsymbol{n}^{t+1}$ being a random vector that evolves from $\boldsymbol{n}^t$, with the probabilities depending on the state $\boldsymbol{n}^t$.

Now these recursions are impossible to solve for any but trivial state-evolutionary laws as the state-space explodes, even for the single-location case (recall that facility type and price categories are also our design variables). To gain some insight, we formulate approximations based on some assumption and derive analytical results for a simple case.

Let $p(\mathbb{1}_{[x']}, \boldsymbol{n})$ denote the probability that a new entrant will open a facility with characteristic x' given the state is $\boldsymbol{n}$.

First, we assume that a steady-state solution exists (i.e. we assume

$$R(\cdot,\cdot) = \lim_{t\to\infty} R^t(\cdot,\cdot)$$

as well defined), we obtain the steady-state optimality equations

$$R(\mathbb{1}_{[x]}, \boldsymbol{n} + \mathbb{1}_{[x]}) = r(\mathbb{1}_{[x]}, \boldsymbol{n} + \mathbb{1}_{[x]}) + \rho \sum_{x'\in\mathcal{X}} p(\mathbb{1}_{[x']}, \boldsymbol{n} + \mathbb{1}_{[x]}) R(\mathbb{1}_{[x]}, \boldsymbol{n} + \mathbb{1}_{[x]} + \mathbb{1}_{[x']}).$$

Note that as $\emptyset \in \mathcal{X}$, so there is the possibility that no new facility opens in a period.

The probability of entry $p(\mathbb{1}_{[x']}, \boldsymbol{n} + \mathbb{1}_{[x']})$ given the state is modeled as follows. A firm decides to enter at characteristic x' depending on the relative long-term profit, defined as the (discounted) operating profit minus the initial investment to establish at x', $F(x')$.

Then, the MNL type formula is

$$p(\mathbb{1}_{[x']}, \boldsymbol{n} + \mathbb{1}_{[x]}) = \frac{e^{[R(\mathbb{1}_{[x']}, \boldsymbol{n}+\mathbb{1}_{[x]}+\mathbb{1}_{[x']}) - F_{x'}]}}{M + \sum_{\mathbb{1}_{[x'']}\in\mathcal{X}} e^{[R(\mathbb{1}_{[x'']}, \boldsymbol{n}+\mathbb{1}_{[x]}+\mathbb{1}_{[x'']}) - F_{x''}]}} \tag{2.5}$$

M is a parameter that controls the rate at which *any* new facility opens in a particular period (which depends on our choice of the duration of each period). We will calibrate M based on the observed rate of new facility openings.

So the recursion for expected long-term revenue if we locate with characteristic x is

$$R(\mathbb{1}_{[x]}, \boldsymbol{n} + \mathbb{1}_{[x]}) = r(\mathbb{1}_{[x]}, \boldsymbol{n} + \mathbb{1}_{[x]}) + \rho \left[\underbrace{\frac{M R(\mathbb{1}_{[x]}, \boldsymbol{n} + \mathbb{1}_{[x]})}{M + \sum_{x''\in\mathcal{X}} e^{[R(\mathbb{1}_{[x'']}, \boldsymbol{n}+\mathbb{1}_{[x]}+\mathbb{1}_{[x'']}) - F_{x''}]}}}_{\text{No firm opens anywhere}} + \sum_{x'\in\mathcal{X}} \frac{e^{[R(\mathbb{1}_{[x']}, \boldsymbol{n}+\mathbb{1}_{[x]}+\mathbb{1}_{[x']}) - F_{x'}]} R(\mathbb{1}_{[x]}, \boldsymbol{n} + \mathbb{1}_{[x]} + \mathbb{1}_{[x']})}{M + \sum_{x''\in\mathcal{X}} e^{[R(\mathbb{1}_{[x'']}, \boldsymbol{n}+\mathbb{1}_{[x]}+\mathbb{1}_{[x'']}) - F_{x''}]}} \right] \quad \forall x, \boldsymbol{n}$$

The objective is to find x given an initial state $\boldsymbol{n}$ that maximizes the long-term expected profit. However the recursion for $R(\cdot,\cdot)$ is challenging to solve exactly. We consider approximate solutions in the next section.

2.5.1 Approximate solution

We use approximate dynamic programming techniques to solve the problem that suffers from the curse of dimensionality. We need to solve the Bellman's equation and compute the value function $R(x)$ but the number of states are too large to evaluate so we have to rely on approximate dynamic programming ideas to gains some insight or tractability. First, we consider a simpler transformation of $R(\cdot, \cdot)$ by some algebra.

$$\sum_{x'' \in \mathcal{X}} e^{[R(\mathbb{1}_{[x'']}, \boldsymbol{n}+\mathbb{1}_{[x]}+\mathbb{1}_{[x'']})-F_{x''}]} \big[R(\mathbb{1}_{[x]}, \boldsymbol{n}+\mathbb{1}_{[x]}) - r(\mathbb{1}_{[x]}, \boldsymbol{n}+\mathbb{1}_{[x]}) - \rho R(\mathbb{1}_{[x]}, \boldsymbol{n}+\mathbb{1}_{[x]}+\mathbb{1}_{[x'']})\big] = M\left[(\rho-1)R(\mathbb{1}_{[x]}, \boldsymbol{n}+\mathbb{1}_{[x]}) + r(\mathbb{1}_{[x]}, \boldsymbol{n}+\mathbb{1}_{[x]})\right] \quad \forall x, \boldsymbol{n} \quad (2.6)$$

If the revenue we expect from each period is r, the NPV with a discount rate of ρ is given by $R = \alpha r$ where $\alpha = \frac{1}{1-\rho}$. This calculation basically assumes a constant revenue in each period. In a similar spirit, our approximations are based on the following premise: state-dependent long-term revenue is a function of per-period revenue functions.

$$R(\mathbb{1}_{[x]}, \boldsymbol{n}) = \alpha(\mathbb{1}_{[x]}, \boldsymbol{n}) r(\mathbb{1}_{[x]}, \boldsymbol{n}).$$

Now different assumptions on the functional form of $\alpha(\mathbb{1}_{[x]}, \boldsymbol{n})$ lead to different approximations.

In the following we assume the per-period revenue function can be generated by an oracle to abstract away its precise calculations based on employee payments and facility rental values.

2.5.2 Constant approximation

This is the simplest approximation, substituting $\alpha(\mathbb{1}_{[x]}, \boldsymbol{n}) = \alpha_x$. The equations simplify to a single equation that we solve by combining with an estimate of M from observed market

data. Equation (2.6) becomes

$$\sum_{x'' \in \mathcal{X}} e^{[\alpha_{x''} r(\mathbb{1}_{[x'']}, \boldsymbol{n} + \mathbb{1}_{[x]} + \mathbb{1}_{[x'']}) - F_{x''}]} [(\alpha_x - 1) r(\mathbb{1}_{[x]}, \boldsymbol{n} + \mathbb{1}_{[x]}) - \rho \alpha_x r(\mathbb{1}_{[x]}, \boldsymbol{n} + \mathbb{1}_{[x]} + \mathbb{1}_{[x'']})]$$
$$= M \left[((\rho - 1)\alpha_x + 1) r(\mathbb{1}_{[x]}, \boldsymbol{n} + \mathbb{1}_{[x]}) \right] \quad \forall x \quad (2.7)$$

We wish to estimate M and α_x for each location x based on observed data.

Equation (2.7) gives us a value of α_x as a function of M, so once we fix M, we have values of α_x for all x. M controls the rate at which new facilities open up vis-à-vis expectations of profitability.

We estimate M from the observed evolution of the facilities in the region that best fits our model. We look at the number of time periods where there were no openings of new firms. If s out of T periods do not have any openings (and the time intervals are small enough so at most one opening per period), we should obtain (here $\boldsymbol{n}_t$ is the set of firms open in period t)

$$\frac{s}{T} = \sum_{t=1}^{T} \frac{M}{M + \sum_{x'' \in \mathcal{X}} e^{[\alpha_{x''} r(\mathbb{1}_{[x'']}, \boldsymbol{n}_t + \mathbb{1}_{[x'']}) - F_{x''}]}} \quad (2.8)$$

We solve for $M \geq 0$ and α_x, using the set of equations (2.7) and (2.8).

With any threat of future competition we expect that our long-term revenues will fall. We show that this is indeed true whenever $r(\mathbb{1}_{[x]}, \boldsymbol{n} + \mathbb{1}_{[x]} + \mathbb{1}_{[x'']}) \leq r(\mathbb{1}_{[x]}, \boldsymbol{n} + \mathbb{1}_{[x]})$.

Proposition 5. *Assuming* $0 < r(\mathbb{1}_{[x]}, \boldsymbol{n} + \mathbb{1}_{[x]} + \mathbb{1}_{[x'']}) \leq r(\mathbb{1}_{[x]}, \boldsymbol{n} + \mathbb{1}_{[x]})$, *for all* x, $\alpha_x \leq \frac{1}{1-\rho}$.

Proof. The proof is by contradiction. If $\alpha_x > \frac{1}{1-\rho}$, the right hand side of Equation (2.7) would be negative (note that $\alpha, \rho > 0$ and $\rho < 1$). Let's consider the left hand side. The exponential term is positive, so we focus on the second term of the left hand side. As $r(\mathbb{1}_{[x]}, \boldsymbol{n} + \mathbb{1}_{[x]}) > 0$,

$$(\alpha_x - 1) r(\mathbb{1}_{[x]}, \boldsymbol{n} + \mathbb{1}_{[x]}) - \rho \alpha_x r(\mathbb{1}_{[x]}, \boldsymbol{n} + \mathbb{1}_{[x]} + \mathbb{1}_{[x'']})$$

can be rewritten as

$$r(\mathbb{1}_{[x]}, \boldsymbol{n} + \mathbb{1}_{[x]})[(\alpha_x - 1) - \rho\alpha_x \frac{r(\mathbb{1}_{[x]}, \boldsymbol{n} + \mathbb{1}_{[x]} + \mathbb{1}_{[x'']})}{r(\mathbb{1}_{[x]}, \boldsymbol{n} + \mathbb{1}_{[x]})}] \geq r(\mathbb{1}_{[x]}, \boldsymbol{n} + \mathbb{1}_{[x]})[(\alpha_x - 1) - \rho\alpha_x].$$

But $\alpha_x > \frac{1}{1-\rho}$ would imply $(\alpha_x - 1) - \rho\alpha_x > 0$, and this contradicts with the right hand side of Equation (2.7) being negative so α_x's that satisfy the set of equations for any $M \geq 0$ should satisfy $\alpha_x \leq \frac{1}{1-\rho}$. □

Example with two locations We illustrate the problem when there are only two potential locations, say 1 and 2. If we consider locating a facility now in location 1, then in the next period a competitor might locate at 1 or 2 or perhaps no one will locate at all. We denote the per period revenues for our location at 1 as $r(\mathbb{1}_{[1]}, \boldsymbol{n} + \mathbb{1}_{[1]}), r(1, \boldsymbol{n} + \mathbb{1}_{[1]} + \mathbb{1}_{[2]}), r(1, \boldsymbol{n} + \mathbb{1}_{[1]} + \mathbb{1}_{[1]})$ for the cases where no additional facilities are located in the next period (neither at 1 nor 2), a facility is located at 2 and a facility is located at 1 respectively. The equations then become:

$$\begin{aligned} &e^{[\alpha_1 r(\mathbb{1}_{[1]}, \boldsymbol{n} + \mathbb{1}_{[1]} + \mathbb{1}_{[1]}) - F_1]}[(\alpha_1 - 1) r(\mathbb{1}_{[1]}, \boldsymbol{n} + \mathbb{1}_{[1]}) - \rho\alpha_1 r(\mathbb{1}_{[1]}, \boldsymbol{n} + \mathbb{1}_{[1]} + \mathbb{1}_{[1]})] + \\ &\quad e^{[\alpha_2 r(\mathbb{1}_{[2]}, \boldsymbol{n} + \mathbb{1}_{[1]} + \mathbb{1}_{[2]}) - F_2]}[(\alpha_1 - 1) r(\mathbb{1}_{[1]}, \boldsymbol{n} + \mathbb{1}_{[1]}) - \rho\alpha_1 r(\mathbb{1}_{[1]}, \boldsymbol{n} + \mathbb{1}_{[1]} + \mathbb{1}_{[2]})] \\ &\qquad = M[((\rho - 1)\alpha_1 + 1) r(\mathbb{1}_{[1]}, \boldsymbol{n} + \mathbb{1}_{[1]})] \end{aligned} \quad (2.9)$$

and

$$\begin{aligned} &e^{[\alpha_1 r(\mathbb{1}_{[1]}, \boldsymbol{n} + \mathbb{1}_{[2]} + \mathbb{1}_{[1]}) - F_1]}[(\alpha_2 - 1) r(\mathbb{1}_{[2]}, \boldsymbol{n} + \mathbb{1}_{[2]}) - \rho\alpha_2 r(\mathbb{1}_{[2]}, \boldsymbol{n} + \mathbb{1}_{[2]} + \mathbb{1}_{[1]})] + \\ &\quad e^{[\alpha_2 r(\mathbb{1}_{[2]}, \boldsymbol{n} + \mathbb{1}_{[2]} + \mathbb{1}_{[2]}) - F_2]}[(\alpha_2 - 1) r(\mathbb{1}_{[2]}, \boldsymbol{n} + \mathbb{1}_{[2]}) - \rho\alpha_2 r(\mathbb{1}_{[2]}, \boldsymbol{n} + \mathbb{1}_{[2]} + \mathbb{1}_{[2]})] \\ &\qquad = M[((\rho - 1)\alpha_2 + 1) r(\mathbb{1}_{[2]}, \boldsymbol{n} + \mathbb{1}_{[2]})] \end{aligned} \quad (2.10)$$

We concentrate on Equation (2.9), rewriting it as (and viewing both sides of the equality as

functions of α_1 and α_2):

$$\begin{aligned}
&-e^{[\alpha_2 r(\mathbb{1}_{[2]}, \boldsymbol{n}+\mathbb{1}_{[1]}+\mathbb{1}_{[2]})-F_2]} r(\mathbb{1}_{[1]}, \boldsymbol{n}+\mathbb{1}_{[1]})+ \\
&e^{[\alpha_1 r(\mathbb{1}_{[1]}, \boldsymbol{n}+\mathbb{1}_{[1]}+\mathbb{1}_{[1]})-F_1]}[\alpha_1(r(\mathbb{1}_{[1]}, \boldsymbol{n}+\mathbb{1}_{[1]})- \\
&\rho r(\mathbb{1}_{[1]}, \boldsymbol{n}+\mathbb{1}_{[1]}+\mathbb{1}_{[1]})) - r(\mathbb{1}_{[1]}, \boldsymbol{n}+\mathbb{1}_{[1]})] \\
+e^{[\alpha_2 r(\mathbb{1}_{[2]}, \boldsymbol{n}+\mathbb{1}_{[1]}+\mathbb{1}_{[2]})-F_2]} &\alpha_1[r(\mathbb{1}_{[1]}, \boldsymbol{n}+\mathbb{1}_{[1]}) - \rho r(1, \boldsymbol{n}+\mathbb{1}_{[1]}+\mathbb{1}_{[2]})] \\
&= Mr(\mathbb{1}_{[1]}, \boldsymbol{n}+\mathbb{1}_{[1]}) + M[((\rho-1)\alpha_1) r(\mathbb{1}_{[1]}, \boldsymbol{n}+\mathbb{1}_{[1]})] \qquad (2.11)
\end{aligned}$$

Notice now that the right hand side is a linear equation with a positive intercept and a negative slope (as $\rho < 1$). While the left hand side has a negative intercept and the second term is an increasing function of α_1 in the range where $\alpha_1(r(\mathbb{1}_{[1]}, \boldsymbol{n}+\mathbb{1}_{[1]}) - \rho r(\mathbb{1}_{[1]}, \boldsymbol{n}+\mathbb{1}_{[1]}+\mathbb{1}_{[1]})) > r(\mathbb{1}_{[1]}, \boldsymbol{n}+\mathbb{1}_{[1]})$, and the third term is an increasing function of α_1 (from our assumption that $\rho < 1, r(\mathbb{1}_{[1]}, \boldsymbol{n}+\mathbb{1}_{[1]}) > r(\mathbb{1}_{[1]}, \boldsymbol{n}+\mathbb{1}_{[1]}+\mathbb{1}_{[1]}), r(\mathbb{1}_{[1]}, \boldsymbol{n}+\mathbb{1}_{[1]}) > r(\mathbb{1}_{[1]}, \boldsymbol{n}+\mathbb{1}_{[1]}+\mathbb{1}_{[2]})$.) So for any value $M \geq 0$, the two curves should intersect for any value of α_2 in the range

$$\alpha_1 > \frac{r(\mathbb{1}_{[1]}, \boldsymbol{n}+\mathbb{1}_{[1]})}{r(\mathbb{1}_{[1]}, \boldsymbol{n}+\mathbb{1}_{[1]}) - \rho r(\mathbb{1}_{[1]}, \boldsymbol{n}+\mathbb{1}_{[1]}+\mathbb{1}_{[1]})} = \frac{1}{1-\rho\frac{r(\mathbb{1}_{[1]}, \boldsymbol{n}+\mathbb{1}_{[1]}+\mathbb{1}_{[1]})}{r(\mathbb{1}_{[1]}, \boldsymbol{n}+\mathbb{1}_{[1]})}}.$$

Since $\rho < 1$ if the drop in daily revenue for 1 with any new opening is sufficiently high, this gives a fairly large range for α_1. Similarly for α_2.

2.5.3 Single characteristic analysis

To gain further insight, we focus on a single location with a single type of firm, and we obtain the following simple equation:

$$\begin{aligned}
e^{[\alpha_x r(\mathbb{1}_{[x]}, \boldsymbol{n}_T+2\mathbb{1}_{[x]})-F_x]}&[(\alpha_x-1) r(\mathbb{1}_{[x]}, \boldsymbol{n}_T+\mathbb{1}_{[x]}) - \rho\alpha_x r(\mathbb{1}_{[x]}, \boldsymbol{n}_T+2\mathbb{1}_{[x]})] \qquad (2.12) \\
&= M[((\rho-1)\alpha_x+1) r(\mathbb{1}_{[x]}, \boldsymbol{n}_T+\mathbb{1}_{[x]})]
\end{aligned}$$

$$\frac{s}{T} = \sum_{t=1}^{T} \frac{M}{M + e^{[\alpha_{x''} r(\mathbb{1}_{[x'']}, \boldsymbol{n}+\mathbb{1}_{[x'']})-F_{x''}]}} \qquad (2.13)$$

We can solve the system of equations described in the previous section by the nonlinear least squares method. The objective is given by

$$min_{\alpha_x, M} L(\alpha_x, M) \tag{2.14}$$

where

$$L(\alpha_x, M) = [g(\alpha_x) - f(\alpha_x, M)]^2 + [\frac{s}{T} - y(\alpha_x, M)]^2 \tag{2.15}$$

in which

$$g(\alpha_x) = e^{[\alpha_x r(\mathbb{1}_{[x]}, \boldsymbol{n}_T + 2\mathbb{1}_{[x]}) - F_x]}[(\alpha_x - 1)r(\mathbb{1}_{[x]}, \boldsymbol{n}_T + \mathbb{1}_{[x]}) - \rho\alpha_x r(\mathbb{1}_{[x]}, \boldsymbol{n}_T + 2\mathbb{1}_{[x]})]$$

$$f(\alpha_x, M) = M[((\rho - 1)\alpha_x + 1)r(\mathbb{1}_{[x]}, \boldsymbol{n}_T + \mathbb{1}_{[x]})]$$

and

$$y(\alpha_x, M) = \sum_{t=1}^{T} \frac{M}{M + e^{[\alpha_{x''} r(\mathbb{1}_{[x'']}, \boldsymbol{n}_t + \mathbb{1}_{[x'']}) - F_{x''}]}}.$$

For a simpler notation we denote the revenue without competitor threat by

$$r = r(\mathbb{1}_{[x]}, n_T + \mathbb{1}_{[x]})$$

and the revenue under competitor threat by $r' = r(\mathbb{1}_{[x]}, n_T + 2\mathbb{1}_{[x]})$.

Proposition 6. *$g(\alpha_x)$ is an increasing and convex function when $\alpha_x > \frac{1}{1 - \rho\frac{r'}{r}}$.*

Proof. Note that this condition corresponds to the one derived in *Example with two locations*. The first derivative is given by

$$\frac{dg(\alpha_x)}{\alpha_x} = r' e^{[\alpha_x r' - F_x]}[(\alpha_x - 1)r - \rho\alpha_x r'] + (r - \rho r')e^{[\alpha_x r' - F_x]}.$$

The second derivative can be obtained as

$$\frac{d^2 g(\alpha_x)}{{\alpha_x}^2} = (r')^2 e^{[\alpha_x r' - F_x]}[(\alpha_x - 1)r - \rho\alpha_x r'] + 2r'(r - \rho r')e^{[\alpha_x r' - F_x]}.$$

The second term is always positive but depending on the first term sign, $g(\alpha_x)$ may end up

non-convex, so we need to put some restrictions on the first term. When

$$(\alpha_x - 1)r > \rho\alpha_x r',$$

we can state that $\frac{d^2 g(\alpha_x)}{{\alpha_x}^2} > 0$ so that $g(\alpha_x)$ is convex. Rewriting this condition we obtain

$$\alpha_x > \frac{r}{r - \rho r'} = \frac{1}{1 - \rho\frac{r'}{r}}.$$

Under this condition we see that $\frac{dg(\alpha_x)}{d\alpha_x} > 0$ so that $g(\alpha_x)$ is increasing, too. □

Proposition 7. *$f(\alpha_x, M)$ is a linear function of both variables; it is increasing in M and decreasing in α_x.*

Proof. First, let's look at $f(\alpha_x, M)$ along M.

$$\frac{df(\alpha_x, M)}{dM} = ((\rho - 1)\alpha_x + 1)r > 0$$

since $\alpha_x < \frac{1}{1-\rho}$. Therefore, $f(\alpha_x, M)$ increases with M. Secondly, we can look at $f(\alpha_x, M)$ along α_x.

$$\frac{df(\alpha_x, M)}{d\alpha_x} = M(\rho - 1)r < 0$$

since $\rho < 1$. Thus, $f(\alpha_x, M)$ decreases with α_x. □

Therefore, for any value of $M \geq 0$, the two curves should intersect for any value of α_x in the range where $\frac{1}{1-\rho\frac{r'}{r}} < \alpha_x < \frac{1}{1-\rho}$. The upper bound on α is quite meaningful. Remember that if there were no future threats (in other words, the revenue in each period is r) the objective of our firm is $\sum_{t=0}^{\infty} \rho^t r = \frac{r}{1-\rho}$. Moreover, if the drop in daily revenue with any new opening is sufficiently high, this will give a fairly large range for α.

Corollary 1. *Considering the extreme cases, when competitor threat is very strong r' might be zero, in which case $\alpha_x > 1$ or when it is very weak, then $r' \approx r$ and $\alpha_x = \frac{1}{1-\rho}$.*

2.5.4 Numerical illustration

We apply the constant approximation method to find the best location for a specific cuisine type (e.g. American) in a certain zip code. In this example, we assume the business owner

wants to open a restaurant of price range $ and 5000 sq-ft size.

We consider the reviews for all the restaurants of this specific type in Yelp. We sort the restaurants according to the first review they received. We also calculate the horizon T by taking the difference between the latest and the first review made among these restaurant. Based on this, we calculate the ratio $\frac{s}{T}$ from data. We use daily compounding with an interest rate of 5%. So, the discount factor is given by $\rho = \frac{1}{1+\frac{0.05}{365}}$. The optimization results are presented in Table 2.13, the rate of entry is found to be $M = 2410$.

Zip	Latitude	Longitude	r	r'	F	a	R	Profitable?
89102	-115.204	36.154	374	327	24179	33	12504	0
89102	-115.196	36.154	380	332	24179	37	13970	0
89102	-115.188	36.154	379	332	24179	55	20781	0
89102	-115.179	36.154	373	326	24179	62	23265	0
89102	-115.171	36.154	363	317	24179	26	9371	0
89102	-115.162	36.154	350	306	24179	57	19899	0
89102	-115.204	36.146	385	337	24179	34	13171	0
89102	-115.196	36.146	391	342	24179	45	17737	0
89102	-115.188	36.146	390	341	24179	19	7303	0
89102	-115.179	36.146	383	335	24179	23	8743	0
89102	-115.171	36.146	370	324	24179	66	24360	1
89102	-115.162	36.146	356	311	24179	36	12949	0
89102	-115.204	36.138	388	340	24179	58	22438	0
89102	-115.196	36.138	393	344	24179	63	24606	1
89102	-115.188	36.138	389	341	24179	70	27351	1
89102	-115.179	36.138	380	333	24179	54	20394	0
89102	-115.204	36.129	376	329	24179	42	15895	0
89102	-115.196	36.129	380	333	24179	29	11038	0

Table 2.13: Profitability estimation for American cuisine with price range 1 and hypothetical size of 5000 sq-ft.

2.6 Conclusion

In this study our goal is to operationalize competitive retail-location models based on customer choice. Traditional market-research based on focus groups or surveys is difficult as finding representative samples is not easy, and the product is difficult to visualize as it is a facility location that will be appearing some time in the future. So we believe data-based CI is the

most promising alternative.

Nevertheless, the challenges are many. First and foremost we normally do not have access to competitors' demand information. So simple structural models cannot be calibrated. We show that by crossing many public data sets on density, demographics, income, ratings and with some minor calibration of a single parameter from a sample of scraped data, we can very accurately predict ratings as well as demand. We demonstrate our findings for the restaurant industry.

In addition, we change the objective of facility location to revenue accumulated over a period of time. Here, the threat of future competitor entry is a major factor. To our knowledge, we are the first ones to even formulate and attempt to solve the optimization problem with this realistic criteria. Our work shows there are significant opportunities to improve the models and their estimation in the area of facility location.

As a remark, we address these challenges for a small restaurant at a small scale. The methods and arguments have the potential to be extended to bigger firms with larger amounts of data. But, our aim is to develop an operational viewpoint on it. To demonstrate our point, we make use of free public data.

There can be further extensions to our model. For instance, loyalty cards are being used extensively among customers. This data can strategically be gauged to understand customer behavior. Furthermore, viability of proposed store locations has to tested whether they are convenient locations in terms of store visibility, pedestrian footfall and car parking availability etc..

Appendices

2.A Restaurants statistics from Yelp.com

Price Range	Number of Restaurants	Average Review Count
1	1532	64.80
2	1398	203.86
3	136	456.96
4	58	465.01

Table 2.14: Average demand by price range.

Stars	Number of Restaurants	Average Review Count
1	11	7.54
2	222	28.52
3	913	73.54
4	1607	203.23
5	371	197.39

Table 2.15: Average demand by review rating.

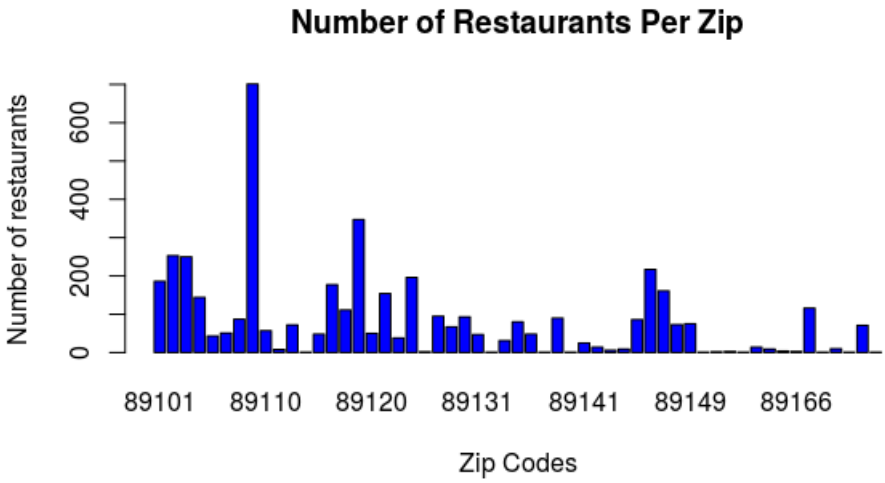


Figure 2.5: Number of restaurants by zip code.

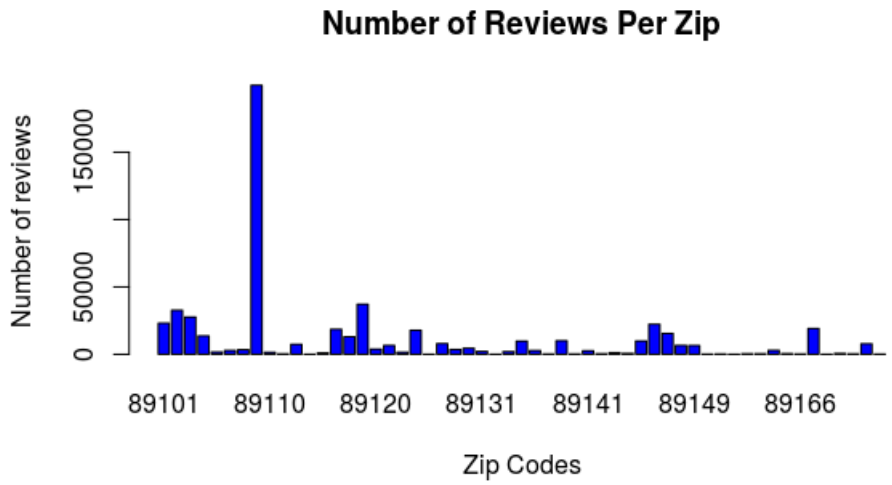


Figure 2.6: Number of review counts by zip code.

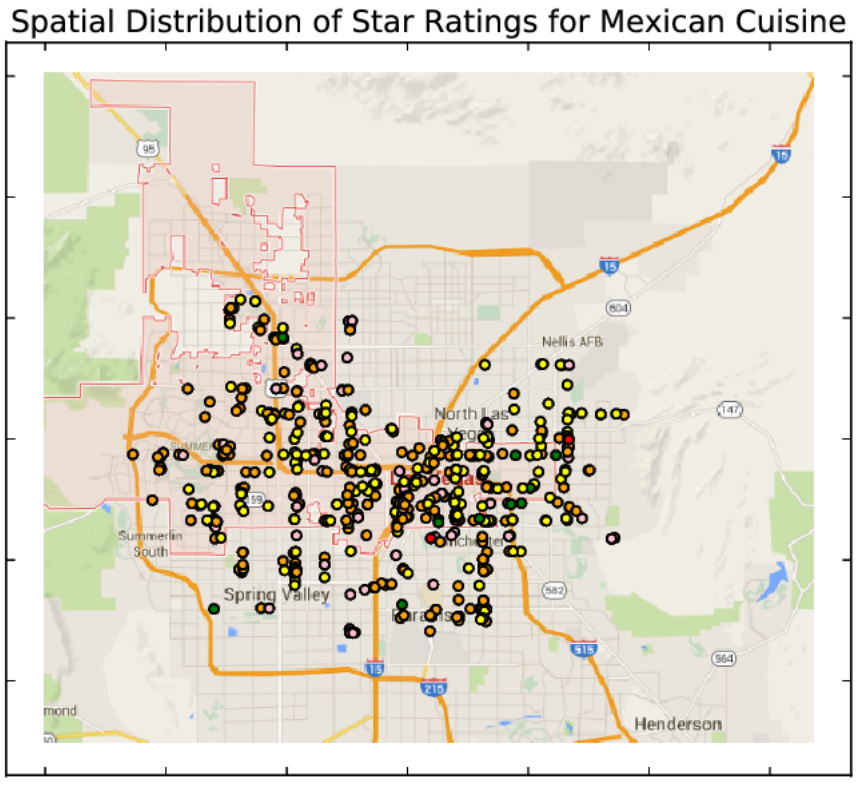


Figure 2.7: Color coded star ratings: ● 1-1.5 ● 2-2.5 ● 3-3.5 ● 4-4.5 ● 5.

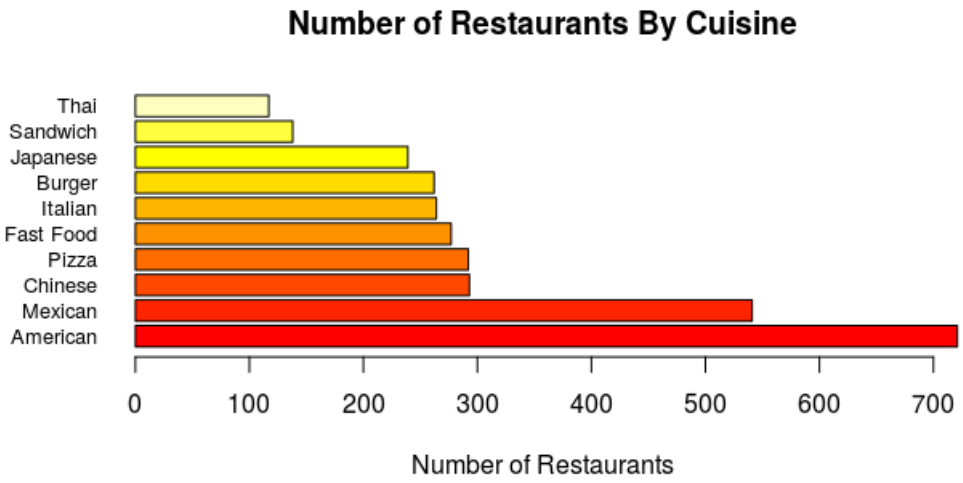


Figure 2.8: Most common cuisine types in Las Vegas.

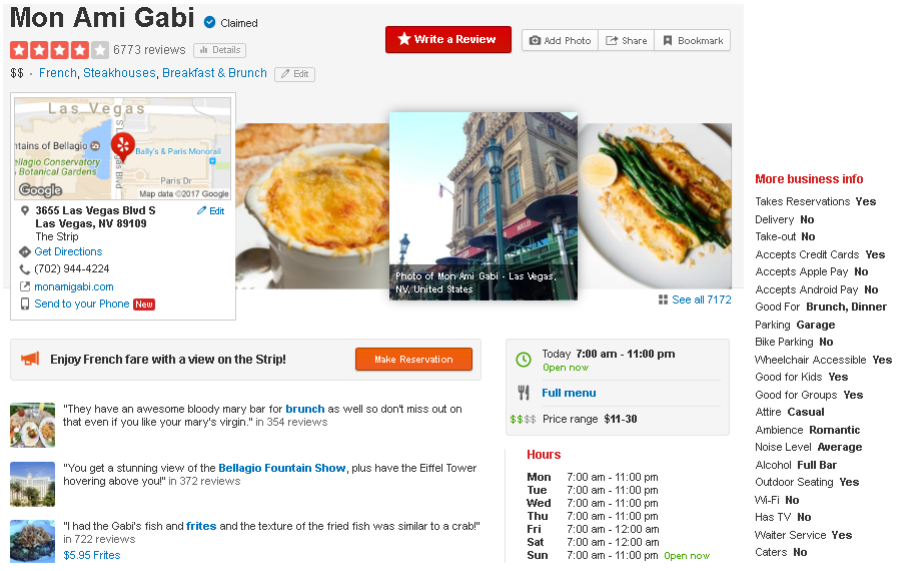
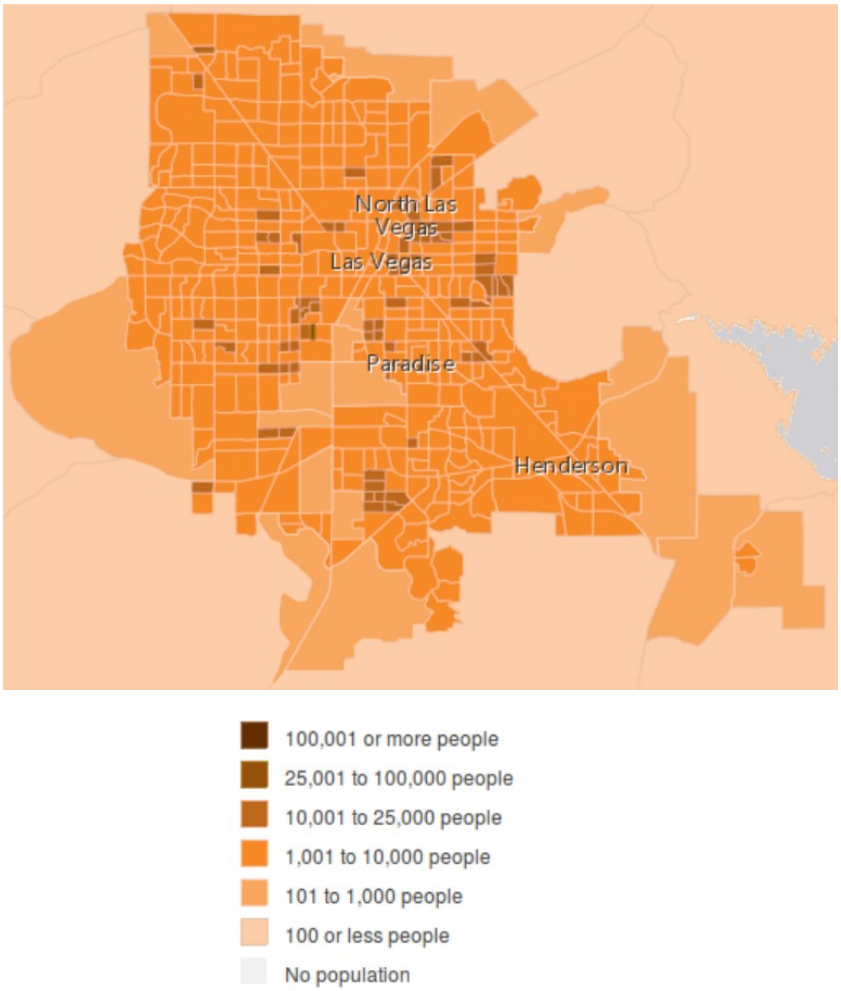


Figure 2.9: Example of a restaurant shown in Yelp's website.

2.B Demographics

Population density distribution of Las Vegas.



Income Range	Number of People
Low <25 K	500818
Middle 25 K – 100 K	1239421
High >100 K	368573.1

Table 2.16: Income distribution of Las Vegas city.

2.C Conventions

Yelp Symbol	Actual Value
$	< 10
$$	11-30
$$$	31-60
$$$$	> 61

Table 2.17: Yelp Price Range Conversion in $.

3

ANALYSIS OF ONLINE REPUTATION MECHANISMS

Online customer reviews have become increasingly important for consumer decision making. One of the most prominent examples is the hotel industry where consumer reviews on web sites such as TripAdvisor, Expedia, Bookings.com and Hotels.com play a critical role in consumers' choice of a hotel. The hotel industry deserves a critical analysis for the following reasons.

First, the demand for a hotel is a mix of repeat customers and first-time customers, where the former naturally have nearly full information on the product so the purchase decision is expected to depend on their past experience rather than online review ratings.

Second, even among first-time customers, there is a mix of different segments (generally labeled leisure, business and groups) who have different booking patterns and might have different sensitivity to reviews and prices.

Third, there are many unobservable aspects of a hotel that can mitigate negative or positive reviews: while the location and price of the hotels can be pinpointed exactly, the importance of the location and prices to a customer is in general an unknown and unobservable. Moreover, as the hotel is an experiential product, the opinions may vary dramatically due to idiosyncratic and unobservable tastes of the reviewers.

Finally, hotel demand is highly stochastic and hotels follow revenue management practices varying their prices based on a number of factors such as day-of-week, occupancy levels as well as competitor prices. The varying demand pattern makes the analysis even more difficult.

In this study, we use a data set of 8 hotels of the same (star) category located in a cluster in a medium-sized town to disentangle the impact of customer reviews from a number of different factors that might also be affecting demand (e.g. location effect). Our first goal is to understand whether purchasers are influenced by customer opinions at the time of booking, and if so, the level of influence on those who come early vs. those who come later. Later, we investigate whether customer reviews are representative of the general customer opinion through a survey study. Finally, we also ask the following question of the data: are reviews affected by price or value or neither, i.e. once customers pay to stay, do they treat what they had paid as a sunk cost and review based on only their experience, or are their reviews affected by the price they had paid?

3.1 Introduction

Consumers constantly decide which item(s) to purchase among several alternatives. Price and quality of items are thought to be the main factors driving purchase decisions. Price is usually transparent and clear enough to obtain, but information on quality is more difficult to obtain and it often remains ambiguous prior to purchase.

So quality is rarely observable and the determinants of quality has been obscure as there is no objective measure of quality. Especially, for products that are bought sporadically or only once, and products purchased at a distance without viewing the product (such as a stay at a hotel), consumers have very limited direct information on quality and often rely on advice and tips from others. For a long time professional reviewers played such a role, but their scope and reach in large dispersed markets was limited[1]. Their role has been partially usurped by user-generated-content on the internet — specifically, reviews of products and service posted online by consumers who had experience with the product[2]. This new source on the one hand offers a richer and more varied set of reviewers with a significantly wider coverage of products, but on the other hand brings into the equation idiosyncratic and unobservable tastes and standards of the (anonymous) reviewers. The subject of this study is the real impact of the reviews on consumer and firm behavior.

There have been considerable research on reviews and their impact on demand over the

[1] Travel guide books for instance cover only a small fraction of the hotels or restaurants in a city.

[2] According to the study Phocuswright (2011), two-thirds of vacation travelers consult online reviews before booking.

last ten years. Those works track the growing importance of reviews in consumer decision-making. Studies have appeared quantifying the impact of reviews on the sales of products such as books, DVDs, video games, movies as well as hotels. Our focus is exclusively on the last category, the hotel industry. Moreover, we differ from previous studies in our goal of studying the impact of reviews at a more detailed level, as well as quantifying the reverse direction, the impact of firm pricing behavior on reviews; this latter aspect of reviews has been less examined in the literature, and in the hotel category, we know of only one other study.

The hotel product category shares some similarities with other products where reviews play an important role in the sales process. It is an experience good; is bought in advance of usage and usually without any direct visual evaluation. It is however significantly more expensive than a book or a movie and is often a part of a long-awaited vacation or celebration with considerable importance attached to the experience, and the experience itself lasts over many days with little recourse to an alternative. Moreover, while books, movies and video games are discretionary items, with no real necessity, a hotel, conditional on a decision to take a trip, is a mandatory purchase. As a consequence customers spend a considerable time going through reviews, which explains the market valuation of a firm like Tripadvisor ($8.4 billion in 2017) which is not much more than a web site that collects and organizes hotel reviews, written entirely by anonymous individuals.

Hoteliers are very keen on understanding the role reviews play for customers in choosing their hotel over an alternative. They are well aware that customer reviews have taken an importance on par with pricing and location. Of the two factors (pricing and reputation) on which the hotelier has some control, reputation is clearly the more challenging and there is a great need to understand it better. But the picture is complicated for the reasons mentioned earlier. Mainly, the difficulty rises due to hotel demand being highly stochastic. Hotels follow revenue management practices varying their prices based on a number of factors such as day-of-week, occupancy levels as well as competitor prices. Moreover, hotel product is an experience-product so customers may have idiosyncratic tastes. Lastly, demand comprise of a wide range of customer types (e.g. business/leisure, the first time/repeat) who have different booking patterns and might have different sensitivity to reviews. All these factors complicate the analysis of online reviews and it becomes a challenge to untangle the impact of customer reviews on purchase decisions from other factors such as price and location and map it to the different customer segments.

Some recent studies question the ability of online reviews to reflect the general customer

opinion. The consequences of customer reviews not being representative of true opinion may be substantial. A very good manager may be penalized or even fired when the customers are actually happy, but we are unaware of it because they do not leave reviews. A hotel slowly losing a high-value customer segment may still be doing well in ratings (as that segment does not write reviews) and one is unaware of it. There are many exogenous factors (competitor prices, our prices, competitor ratings etc.) that confound this issue.

In this study, we use a data set of 8 hotels located in a cluster in a medium-sized town to disentangle the impact of customer reviews from a number of different factors that might also be affecting demand. We use a set of customer opinions obtained from various websites namely Booking, Expedia, Hotels.com, Orbitz, Priceline, Tripadvisor and Yelp. We create a unique data set by first crossing detailed booking information for a hotel with the occupancy levels and average daily rates of all its competitors, and then with customer reviews for all the hotels during that period.

Our econometric analysis has three broad goals. Our first goal is to understand whether purchasers are influenced by customer opinions at the time of booking, and if so, the level of influence on those who come early vs. those who come later; business vs. leisure. We find for instance that early purchasers are influenced by review ratings more than later purchasers which might be useful for hotel managers to know. Moreover, the impact of customer review ratings is even stronger than the impact of price of a hotel for the very early purchasers group. We correct for problems of endogeneity due to unobservable location importance and heterogeneity of customer types in estimating our models. Our second goal is to reveal how representative reviews are of the true quality[3] of a hotel through a survey study. We also want to reveal the influence of customer reviews for repeat customers vs. new customers. Lastly, we answer the following question of the data: are reviews affected by price, or value, or neither, i.e. once customers pay to stay, do they treat what they had paid as a sunk cost and review based on only their experience, or are their reviews affected by the price they pay? We show evidence that value plays a significant role in how customers review a hotel, where value is defined as the price relative to competitor prices in that area; so one possible inference is that as customers have decided on a stay in the area, price is relative to the other offerings rather than an absolute number.

[3] True quality can never be measured to a whole extent but what we mean by true quality in this study is a measure gathered by collecting customer opinions based on a randomized sample. We use true quality and general customer opinion interchangeably in the remaining of the study.

In the next section, we survey the relevant literature. In §3.3 we examine the impact of reviews on demand. §3.4 we reveal how representative reviews are of the general customer opinion. In §3.5 we analyze the relation between reviews and value. Finally, in §2.6 we conclude with a summary of the results.

3.2 Literature review

This literature review covers the articles most relevant to our study: the impact of reviews on purchase decisions, whether reviews are biased, motivation to write a review, and the relation between prices and reviews. We ignore a vast amount of literature on social learning (as we consider only anonymous reviewers) and studies that have a more machine learning flavor based on text and sentiment analysis.

3.2.1 Online reviews' impact on sales

Prior to purchase, customers are not informed about the true valuation of products. This creates uncertainty in deciding which product to choose. Customers consult friends or other peoples' opinions who experienced the product before to form beliefs about its quality and decide whether a product is worth buying at a certain price or not. Consumer learning literature studies to understand the way consumers form and update their beliefs of products and most use a natural Bayesian framework (Roberts and Urban, 1988; Erdem and Keane, 1996; Ackerberg, 2003; Narayanan et al., 2005; Zhang, 2010; Acemoglu et al., 2017).

Before the Internet, word-of-mouth marketing (WOM) meant a consumer spreading the (hopefully positive) experience amongst friends and relatives. Marketing researchers have long studied WOM and there is a significant literature on it. Internet changed WOM in two distinct ways: one, social network sites expanded and made efficient (and observable) the spreading of this word; second, specialized review sites and e-commerce sites enabled a reach beyond friends and family.

Dellarocas (2003) is one of the early papers on internet and WOM. Up to a certain extent, reputation mechanisms (e.g. Yelp, TripAdvisor, Expedia) reveal information on experience-product attributes such as quality, reliability that can only be observed after the purchase. This certainly helps customers to make more informed and efficient decisions. Bickart and Schindler (2001) show that reviews on internet forums and bulletin boards have a greater im-

pact than marketer-created information. Resnick and Zeckhauser (2002) find that sellers with better customer ratings are more likely to sell their products. In an online experiment, Senecal and Nantel (2004) indicate that subjects who read product reviews select the products twice as often as the ones who do not read product reviews. Besides, they indicate that recommendations for experiential products are much trusted. Tellis and Johnson (2007) state that product ratings provide valid source of quality information in strategic and financial terms. Zhao et al. (2013) focus on experiential products and study the effect of reviews on purchases using a Bayesian model, integrating both product quality and review credibility. In the context of restaurants, Cai et al. (2009) show that customers tend to order the most popular dishes if they have this information. Anderson and Magruder (2012) study the link between online customer reviews with the popularity of restaurants. They analyze the restaurant ratings of Yelp given in half unit increments. They conclude that an extra half-star on Yelp enables restaurants to sell out 19 percentage points more frequently.

While it is mostly supported that online customer reviews affect the market in a positive way, the results in the literature are not consistent. Some of the research supports a hypothesis that online reviews significantly increase sales whereas others challenge this finding and claim that online customer reviews may also have a perverse effect. The inconsistency might be mainly due to the following aspects.

First, it may be caused by focusing on different aspects of online reviews, such as reviews' persuasive effect and awareness effect. The awareness effect is that reviews enable customers to recognize products through dispersion and select the products for their choice set. On the other hand, persuasive effect is related to the assessment of product quality, which may influence customers' purchase decisions. Godes and Mayzlin (2004) focus on the effect of dispersion of WOM for TV shows. They illustrate that dispersion of WOM can be an early indicator of a product's success, while volume (number of reviews) is responsible in later periods. Their study however cannot separate the effects of the quality of TV shows from the ratings of the shows. Liu (2006) studies WOM data from the Yahoo! Movies and indicate that during pre-release valence (average numerical rating) is important but after the movie is launched message volume of the previous week becomes the best predictor of product sales. Dellarocas et al. (2007) use a modified Bass diffusion model in their analysis of reviews' effect in forecasting movie revenue and they indicate that the valence plays more important role in predicting future movie revenues than other factors considered. Though being the same research context of movie industry, a few studies fail to derive any significant relation between

WOM and revenues. For instance, Neelamegham and Chintagunta (1999) conduct an empirical study and they find no effect between WOM and weekly revenues.

Second, WOM is taken as an exogenous factor in many studies. The review information has both a direct effect on purchase via revealing quality and an indirect impact of fostering customers to gather more information about the same or similar products. Just like purchase decisions, learning from external sources and processing this knowledge require deliberate actions, and both should be treated as endogenous. However, consumer learning studies from external quality signals treat those signals as exogenous to simplify analysis such as advertisement (Ackerberg, 2003), price (Erdem et al., 2008), or product review (Zhao et al., 2013; Erdem and Keane, 1996). Therefore, there is an inherent difficulty in establishing causality between WOM and product sales: sales could be high as a result of the product's quality that eventually increases the online customer review ratings. There are several studies that bring out this aspect. For instance, Van den Bulte and Lilien (2001) show that in Coleman et al. (1996) the influence of WOM on physicians' adoption decision is overestimated because of lack of control for the drug companies' marketing performances. Thus, the endogenous character of online reviews is mostly ignored. Many studies are conducted in a cross-sectional setting that cannot differentiate whether sales change due to differences in product quality or the effect of WOM.

Chintagunta et al. (2012) controls for endogeneity using a rational expectation model to reveal drug qualities. A study by Chevalier and Mayzlin (2006) shows that WOM has a causal impact on consumer purchasing behavior at two online retail sites. They use book sales data from Amazon.com and BarnesandNoble.com to examine the relationship between the customer reviews and firm sales while controlling for other factors of book sales. The authors can better establish causality effect of customer reviews by comparing the sales for a given book across two booksellers and using difference-in-differences approach. They state that better valence of books' reviews leads to higher sales. Besides, they indicate that verbal data has an impact beyond numeric ratings. They also show that the effect of negative reviews is larger than the effect of its positive counterpart. Zhang and Dellarocas (2006) provide similar results in the case of movies. Additionally, Eliashberg and Shugan (1997) indicate that online reviews are predictors rather than influencers of product sales. In an empirical study, Chen et al. (2004) use book reviews in Amazon.com and find that recommendations are positively related to sales and the impact is larger for less popular products. However, they find no significant impact of consumer ratings on sales. Duan et al. (2008) employ an endogenous model of

Yahoo! Movies reviews on box-office sales. They state that the volume but not the valence has an impact on revenues. Online ratings do not have persuasiveness effect but only awareness. Moreover, bad ratings have no sales impact or sometimes even positive impact on sales due to awareness and careful consideration.

Lastly, recent studies question the ability of online reviews to reflect the real quality of a product. The credibility of an online review is important to spread more accurate information. Leaving aside the large issue of fake reviews and credibility, representativeness is a main concern. Reviewers are not sampled randomly from the user population. In his study, Anderson (1998) states that extremely satisfied and extremely dissatisfied customers are more likely to post product reviews. There can be external interventions to the reviews; Kuksov and Xie (2010) advise firms to follow product augmentation (frills) strategy in early periods in order to enhance demand and therefore improve product rating in later periods. There is also an issue about the possible risk of manipulation of online reviews. Favorable/unfavorable reviews may be artificially inflated by biased interested groups to increase/decrease sales. Dellarocas (2006) and Mayzlin (2006) state the possibility that firms may write online reviews on behalf of the products, in a way they manipulate the reviews, with the aim of increasing awareness. Nevertheless, a few works claim the opposite. Clemons et al. (2006) demonstrate that the variance of ratings and the strength of the most positive quartile of reviews significantly affect beer sales. A study by Banerjee and Fudenberg (2004) suggests that the reporting bias—tendency to post extreme ratings rather than average ones, does not lessen the effect of WOM for perfect social learning. We emphasize that manipulation is different than the inherent bias of reviews and we discuss the relevant literature on the representativeness in detail next.

3.2.2 Are online reviews biased?

A statistic (such as the average rating) or a distribution is representative only if we sample randomly from the population. i.e., each customer leaves a review with equal probability. However, as the reviews are left on a voluntary basis, and writing a review in itself involves some time-cost, it is not at all clear if the sample is random and representative. Online customer review literature mentions *under-reporting* and *self-selection* interrelated concepts as the main factors to shape the distribution of online ratings.

Under-reporting bias is explained by satisfaction literature (Anderson and Sullivan (1993)). It suggests that consumers with extreme opinions are more likely to post a review than cus-

tomers with moderate opinions. This results in extreme (high or low) ratings. Hu et al. (2006) show empirically that the distribution of numeric ratings posted is bimodal where most of the reviews are positive and only a few are negative. Thus, there is some variance on the valence of reviews.

Self-selection bias derives from utility theory that asserts consumers with high perceived quality make purchase decisions and so have a chance to write a review. These biases change the presumably normal distribution of online customer review ratings into J-shaped. Hu et al. (2006) Li and Hitt (2008) and Nevskaya (2012) identify self-selection biases in online customer reviews. Hu et al. (2006) and Admati and Pfleiderer (2004) report the existence of bimodal (or J-shaped) ratings distribution and they conclude that the average of ratings may not reveal true quality of products. Moreover, there are various sources of individual related factors independent from product quality that can influence ratings such as social influences, cultural differences, dynamics of reviews but they are not the main concern of our study.

All in all, findings suggest that product ratings do not accurately assess quality and so there are several concerns about commercial website ratings. As a consequence, Sun (2012) suggests that other distributional characteristics of reviews such as "skewness" other than the average may affect consumer decisions. To correct for the bias, it can be important to know the factors that motivate customers to leave reviews on online platforms.

3.2.3 Motivation to write a review and overcoming biases

Writing a review involves some time and cost and very little is known about what motivates customers to write online reviews. Are they first time or repeat customer? Is there disconfirmation between the review and the experience? These questions are important to investigate to reveal the real value of reviews. Yoo and Gretzel (2008) find the motives as helping a travel service provider, concerns for other consumers, and needs for enjoyment/positive self-enhancement. Toubia and Stephen (2013) question the motivation consumers contribute content on Twitter. They model each user's decision in each time period as a multinomial-choice and estimate each user's utility function using a dynamic discrete choice model. They find that image related utility is larger for most of users than intrinsic type.

In relation to this, it is also important to understand "Do customers rate pure quality?" There is extensive research on how customers' perceived quality (subjective judgment) of products is shaped. As mentioned in Mitra and Golder (2006) prior expectations based on one's

own and others' experiences, brand reputation, price and advertising matter. *Reference comparison* idea is studied by Boulding et al. (1993), Cronin et al. (1992), Oliver (1980), Parasuraman et al. (1985), Zeithaml (1988). Similarly, Ho et al. (2013) analyze customer posting behavior based on quality disconfirmation. Anderson and Sullivan (1993) link satisfaction to retention in utility framework. Moreover, as Prospect Theory suggests losses loom larger than gains and so expectations being higher than the realized quality is found to have a greater impact on satisfaction and retention than expectations being lower than realized quality. It is also common for some consumers to rate *value* (quality for the money), which leads to penalizing the more expensive products. On the other hand some consumers rate product quality alone. So, this results in heterogeneous consumer tastes (De Langhe et al., 2015).

There is a significant research to mitigate the biases. Nevskaya (2012) suggests using the distribution of tastes and price sensitivities of consumers to reveal the true quality of a hotel. Askalidis et al. (2017) use email prompted reviewers to mitigate the biases confronted in online reviews such as social influences and selection biases. Jurca and Faltings (2009) suggest reward schemes to correct for the bias and dishonest reporting in online reviews. Besbes and Scarsini (2018) compare the information transmission of reviews under two scenarios: one, when customers observe all the past reviews, two customers only observe the average statistic of the past reviews. They find that even limited statistics of past reviews communicate strong information content to future customers as long as customers are minimally sophisticated.

A number of studies analyze whether the reviews influence each other, whether they change over time and their influence over time, and how their impact varies over different product categories. Since these issues are not the main subject of our study, we do not cite the literature on these questions.

3.2.4 Price-reviews relation

The impact of price on reviews has also taken considerable attention in online review literature. Papanastasiou et al. (2015) study a monopolist's pricing and inventory decisions. They consider heterogeneous customers with both perfect (Bayesian) and imperfect social learning mechanisms. They find that prices affecting revenues directly, also moderates social learning process, contributing to its content as well as its amount. Liu (2010) analyzes relationship between product's price and consumer's rating of price-discounted value (price/value or subjective quality) of the product. He finds that the heterogeneity in price/value relationship is in-

fluenced by different market conditions, brand name and manufacturer's warranty. Jing (2011) focuses on durable goods and analyzes the impacts of social learning on the dynamic pricing in a two-period monopoly to distinctively identify timing of purchases; early (uninformed) or late (informed) buyers. He studies optimal pricing strategy, the adoption equilibrium, and the advantages of social learning on profit and welfare. Abrate et al. (2011) use hedonic price functions to explain the relationship between quality signals and prices. Yu et al. (2015) analyze the impact of customer reviews on a firm's dynamic pricing strategy. A firm can control its revenue and the review information through its initial sales amount. They assume that the reviews are endogenously generated and study the impact of reviews on profits, prices, and sales.

In an experienced good industry, Dubois and Nauges (2010) follow a structural empirical approach, based on Olley and Pakes (1996), to disentangle the effect of experts' grades from the effect of true quality (unobserved) on the pricing of wines.

Previous studies show that consumers are inclined to rely more on WOM when purchasing experience goods versus search goods (Senecal and Nantel, 2004). So, WOM study is especially important for experienced based products such as hotels. The reasons are following; being an experience product, the quality cannot be determined until consumption and also during the consumption, consumers are possibly involved more with the product which increases the chances of WOM being generated after consumption. Moreover, the hotel product has unique features. For instance, compared to books and movies, hotel is more like a perishable product but hotels have longer life cycle. Also, books and movie recommendations can be obtained through friends or media but for hotels the information is more limited which makes consulting to WOM more essential when customers want to book a hotel.

Overall, it can be clearly seen that customer reviews are a complex phenomenon and require a detailed, elaborate analysis to reveal their real value. The studies closest to ours as they concentrate on the hotel industry or relation between prices, reviews and quality are due to Anderson (2012) who studies the impact of reviews on hotel performance, specifically, its revenue and occupancy, and Li and Hitt (2010), Li et al. (2014a) who investigate the reverse direction, the role of prices on consumer reviews. Wang et al. (2015) study the effects of WOM on room price and hotel star rating. They find that for hotels receiving positive/negative WOM, their online sales performance is less/more likely to be influenced by room price and star rating. Ye et al. (2014) state that price positively affects perceived quality and negatively affects perceived value. Moreover, for higher starred hotels, the effect of price on perceived quality is

higher. Lastly, this effect is significant for business travelers but insignificant for leisure travelers.

3.3 Impact of reviews on segments

In this section we examine the impact of reviews on demand. The complicating factor is price, the hotel's, as well as the competitors' at the time of booking. Moreover hotel customers are quite heterogenous (the basis for many hotels' RM), and the impact of reviews on their purchase behavior can be uneven. Our intention is to separate the effects of several customer segments so that hotels can have a better idea on how to manage pricing for the different segments.

3.3.1 Data description

Our analysis is based on online customer reviews and sales for a hotel (hotel X), located in a medium sized town in the U.S. To estimate the influence of reviews at the time of booking, we merge two independent data sources. Our first data is booking data of hotel X over nearly two years. It consists of over twenty-four thousand entries containing information about booking dates, stay start dates and stay end dates. The data set also indicates whether the reservation is made for group or individual stays. Furthermore, it contained information about cancellations, which we exclude from the analysis. After this process, the working data consists of over eighteen thousand entries.

Second, we have a data set that consists of online customer review ratings for X as well as all its competitor set, during the period under analysis. It consists of a complete set of customer opinions obtained from various third-party websites, also called online travel agencies, namely Booking, Expedia, Hotels.com, Orbitz, Priceline, Tripadvisor and Yelp. All the review sites require the customer to assign an evaluation score of their experience (along various dimensions such as cleanliness, service etc) between $0-5$ or $0-10$. The scale varies by review site. These numbers are gathered from the online travel agencies by a third-party web site which combines all the measures into a single one (Review Score, a weighted moving average) and we use this aggregated measure of online customer review ratings in our analysis.

Our aim is to understand whether purchasers are influenced by customer review ratings at the time of booking. Moreover, we assess the impact of the time interval between the date

of purchase and the date of intended arrival on this relation.

3.3.2 Impact of reviews on sales

The demand pattern may vary widely due to seasonality, group stays and some other effects. We build our model to account for variability in demand. Thus, in order to observe the actual impact of review score on the sales pattern, booking distributions should be normalized for each stay date considering. For this purpose, we first compute the overall demand for each initial stay date, intended date of arrival. Then, we calculate the frequency of bookings for each specific stay start date, distributed along the time interval between booking and stay start dates. Finally, we take the proportion of frequency of bookings over demand for each stay start date and we find the booking fractions.

We thus create a table that summarizes booking information as n-day-prior booking fractions, which can then be related to review score at the time of booking. Sample data of our analysis can be seen in Table 3.7 in the Appendix 3.A. For the sake of clarity, we exemplify the process for this data. In this sample, the total number of bookings for stays starting on 05-04-12 is 41, 3 of which were booked on 07-02-12, 6 on 08-02-12, 1 on 02-03-12, etc. Then, we calculate booking fractions such as the 58-day prior booking fraction, which is $3/41$; the 57-day prior booking fraction, $6/41$; the 34-day prior booking fraction, $1/41$, and so on. This procedure enables us to account for demand variety across time. Then, correlations can be calculated for 58-day prior bookings and review score of the respective booking day, which is $3/41$ against 79.2; as for the 57-day prior bookings it is $6/41$ against 80.8 etc.

After computing the n-day-prior distribution of booking fractions for each stay start date, to get more stability we can aggregate them in booking antecedence intervals in relation to the stay start date. We decided to group them into four bins aggregating $0-5$ day prior bookings, $6-18$ day prior bookings, $19-55$ day prior bookings, and $56-597$ day prior bookings. We analyze these four bins using the following logarithmic regression, where b indexes the bin number:

$$\log(\text{Bookings Fraction}_{bt}) = \alpha_b + \beta_b \log(\text{Review Score}_{bt}) + \epsilon_{bt}. \tag{3.1}$$

We use log-log regression model. We summarize the results of regression model in Table 3.1.

We see that when bookings are made $56-597$ days prior to the initial stay date, customer

	bin [0-5]	bin [6-18]	bin [19-55]	bin [56-]
β_b	0.343	0.560	0.825^{*}	1.791^{***}
Review Score	(0.550)	(0.371)	(0.457)	(0.682)
α_b	-4.167^{*}	-6.082^{*}	-7.398^{*}	-11.474^{*}
Constant	(2.453)	(1.657)	(2.039)	(3.032)
Observations	1,687	1,664	1,670	1,149
R^2	0.0002	0.001	0.002	0.006
Adjusted R^2	-0.0004	-0.001	-0.001	-0.005
F Statistic	0.389	2.272	3.263^{*}	6.889^{***}
	(df = 1; 1685)	(df = 1; 1662)	(df = 1; 1668)	(df = 1; 1147)

Note. Robust standard errors are in parentheses. $^{***}p < 0.001, ^{**}p < 0.01, ^{*}p < 0.05$

Table 3.1: Regression output of log-log model for the different bins

review ratings become significantly correlated with fraction of bookings at 0.001 level. Yet, R-squared value is 0.00597. This is very low; only 0.59% of the variation of bookings fraction can be explained by review scores through this regression model. Thus, despite the noise, review score values are correlated with the fraction of bookings. $\beta = 1.79$; this means that one percent change in review score corresponds to 1.79 percent increase in the bookings fraction, on average. One possible reason for having such a low goodness-of-fit value is that we are using only one explanatory variable in our model, which is review score. However, there are other variables such as price of a particular day that clearly influences the sales more than customer review ratings and in our model we are not taking them into consideration.

The other bin considers the bookings that are made $19 - 55$ days prior to initial stay date. Review score is correlated with the fraction of bookings at 0.05 significance level. R-squared value decreases to 0.00195. Overall, from these results we can conclude that as the time of booking gets closer to the intended date of arrival, the impact of customer review ratings on sales decreases since β decreases. Yet, for the bins of $0 - 5$ and $6 - 18$ days prior bookings, review score is not significant. Moreover, goodness-of-fit decreases slightly.

Overall, this model tells us that for each initial stay date the probability of a visitor having booked on a specific date increases as that dates associated review score increases. Moreover, we can state that early purchasers are influenced by review ratings more than later purchasers. To our knowledge, this is an interesting new finding about customer reviews impact for the

hotel industry.

3.3.3 Including competition

Up to now, we only focused on the data from hotel X and the review scores, and did not consider its competitors. Yet, in real life the prices of competing hotels in the neighborhood plays an important role in customer purchase decisions; For instance, A might be getting very good reviews, but if its price is too much above a competing property B, many customers might choose B.

We model the customer purchase process as follows: Each customer has a probability p of choosing our hotel X, and with probability $(1 - p)$ the customer chooses to stay at that hotel's competitors. So, if there are M customers in the market for a certain stay-date, then on average, the number of people who choose our hotel would be Mp. This is a market-share model then, that we estimate based on aggregate occupancy and average price data for each stay-date that we obtain from a different source. There is a large amount of literature on econometric approaches to forecast market shares of a company including its competitors (Leeflang et al., 2000). Since market shares are bounded within $[0, 1]$ interval, in our analysis we use a generalized linear model, specifically, a logit regression model.

For the market share analysis, we cross our previous two data sets with X's competitors data on a daily basis. The competitor set is defined by hotel X and based on intimate knowledge of their customers' tastes as well as the location and category of the competitors. We take the total market size on a daily basis as the sum of the total demands[4].

A small sample of our data displaying occupancy percentages, prices and market size information can be seen in Table 3.2. ADR stands for average daily rate for that stay date. Competitor occupancy is the average occupancy, across all the hotels in the competitor set—this is the level at which we have competitor data. We also know the total number of rooms of the competitor set, by which we obtain the exact number who stayed in the competitor hotel on this date. Of course as customers can stay for multiple nights, we need to separate occupancy from bookings. Lacking precise information on this, we take a fraction of the total number staying on a certain date as those starting their stay on that date, and this fraction we obtain from hotel X.

[4] The true market size is no doubt bigger, as some people might choose to stay home or go outside, but for our estimation purposes this is a reasonable approximation

	Sa	Su	Mo	Tu	We	Th	Fr	Sa	Su
	1	2	3	4	5	6	7	8	9
X Occup.	83.9	25.8	30.3	49.7	45.8	32.3	45.2	46.5	20
Comp. Occup.	47.1	26.1	60.5	76.5	70.2	48.8	51.1	51.1	32.5
X ADR	67.2	71.6	78.1	78.3	79.2	74.6	79.4	77.8	72.4
Comp. ADR	93.3	98.3	101.2	103.4	101.7	99.2	93.2	96.3	100.3
X Demand	140.1	43.1	50.6	83	76.5	53.9	75.4	77.6	33.4
Comp. Demand	405	224	520	657	603	419	439	439	279
Market size	545.1	267.1	570.6	740	679.5	472.9	514.4	516.6	312.4

Table 3.2: Benchmark measures of Hotel X and its competitors

We now describe MNL model of choice behavior to estimate hotel X's share of the market. The market share model is given by the logistic distribution function form for each stay date d summed over all purchases made over booking period t

$$Pr_d^o = \sum_t \frac{\exp(\beta_p^o \text{Price}_d^o + \beta_r^o \text{Review Score}_t^o)}{\exp(\beta_p^o \text{Price}_d^o + \beta_r^o \text{Review Score}_t^o) + \exp(\beta_p^c \text{Price}_d^c + \beta_r^c \text{Review Score}_t^c)} \tag{3.2}$$

where Pr_d^o denotes the market share for our hotel against its competitors to start stay on date d, and is obtained from own as well as competitor occupancy data and is simply the proportion of all bookings made during time t to start the stay on date d divided by the total market size for the same date d. Then, the relation with our sales is given by $\frac{s_d^o}{Pr_d^o} = M_d$. Here, we assume the market on date d is the sum of all the customers who start their stay on date d. We could carry out a more concrete analysis based on network tomography findings and untruncate the market size to its LOS itineraries.

The microeconomic model behind (3.2) is that customers have a linear utility based on price and quality, and quality is replaced by a proxy variable, the review score of the hotel. The competitor is assumed to include the outside option. Since we have the final market share for the specific stay date d, we estimate the parameters from the observed market share.

A sample of data prepared to analyze this regression model in R can be found in Table 3.8 in the Appendix 3.B. It consists of information on the date of arrival, number of days between the booking date and arrival date, number of bookings the hotel X had on any day, the number of bookings its competitors had on any day, review score of X for each booking date and lastly its price at the time of stay. The whole data set consists of 1228 entries. Similar as before, we separate the whole data set into several bins of the n-day prior bookings. We can vary the co-

efficients in (3.2) to be dependent on the time of booking as before, and the result of regression is given in Table 3.3.

Days Prior		(Intercept)	Review Score	Price
Bin[0-5]	Coefficients	-6.19	0.0394	-0.0262
	Pr(>Chi)		0.2246	$7.19e-08^{***}$
Bin[6-20]	Coefficients	-4.13	0.0055	-0.0266
	Pr(>Chi)		0.376	0.0006^{***}
Bin[21-55]	Coefficients	-6.6629	0.0657	-0.0574
	Pr(>Chi)		0.0046^{**}	<2.2e-16 ***
Bin[56-]	Coefficients	-19.23	0.188	-0.0322
	Pr(>Chi)		$< 2.2e-16^{***}$	$8.903e-07^{***}$

Table 3.3: Regression output including competitor for different bins

Due to the nature of logit regression, to interpret the parameter estimates, we will also take a look at the odds ratios of this model. The odds ratio is the change in the odds of an option being chosen given one unit change in the independent variable. In our case, this is the odds of the hotel room being booked and the odds are the probability of being chosen divided by the probability of not being chosen $p/(1-p)$. Odds ratios are given in Table 3.4.

	0-5 days bin	6-20 days bin	21-55 days bin	56-223 days bin
Review Score	1.0402	1.0055	1.0679^{**}	1.2068^{***}
Price	0.9741^{***}	0.9738^{***}	0.9441^{***}	0.9682^{***}

Table 3.4: Odds ratios

We observe that price always significantly affects the sales. However, the impact of review score is not significant for the recent purchasers, up to 20 days. As expected, we also see that as the hotel increases its price, the hotel's sales decreases since the odds ratios are less than one. On the other hand, the odds ratios of review score attribute are all greater than 1; therefore, as review score of a hotel increases, sales increases, too. For instance, if review score of the hotel increases by one unit, the odds of a room being sold increases by a factor of 1.0679 whereas if price of the hotel increases by one unit, the odds of a room being sold decreases by a factor of 0.9441. Moreover, we can also see that review score is more effective for early purchasers ($56-223$ days bin) than later purchasers ($21-55$ days bin). Lastly, the influence of review

score is stronger than the influence of price on sales for the very early purchasers ($56 - 223$ days bin). However, the residual deviances for those four bins are respectively $0, 0.24, 0, 0.02$. Thus, except for the $6 - 20$ days bin they are highly significant and we need a better model. One possible reason for the bad fit could be that 3-month data is not sufficient for the analysis.

3.4 How representative are reviews of the general customer opinion?

In this part, we wish to answer a fundamental question at the back of everyone's mind: How representative are the reviews of the true quality? There are a few studies that analyze several aspects of the quality. Parasuraman et al. (1985) reveal 10 determinants of consumer's perceptions of service quality: tangibles (look, appearance), reliability (keeping promises), responsiveness (prompt and willing), competence, access, courtesy, communication, credibility, security, understanding/knowing the customer. However, there is no absolute measure of quality and we have to rely on other measures.

Assuming that once a customer comes into the market he does not leave without purchasing a product, he faces the decision of selecting our hotel or one of its competitors. In order to reveal the true quality, we use information on repeat customers –those who stayed at our hotel before and come into the market again. We hypothesize that once customers stayed in a hotel, they experience a customer specific shock around the true quality value ($\bar{q} + q_i$) and hence, in their next visit, they will only consider competitor review ratings and price difference between our hotel and the competitors. For reasons of simplification, let's also assume for the moment that our prices are the same as the competition's. When a customer who stayed before at our hotel comes again to the market, the previous experience is the strongest reference for comparison with competitor ratings rather than the rating of our hotel. Accordingly, if $\bar{q} + q_i \geq r^c$, they will choose to stay in our hotel, otherwise they will go to the competitor. Note that we wish to answer this from the point-of-view of the hotelier (and not an external agency), so we have access to transactional data and some other partial market information. We have the following data

- Own bookings with detailed information of customers; booking date, arrival and departure date and price they paid

- Partial competitor information: price and customer ratings
- Market information (STR): Using techniques we developed based on network tomography, we estimate the market size (total population interested in a hotel in a certain day-location) for each day for each LOS.

Customers are assumed to make the purchase decision based on prices and reviews. We build a behavioral model where customers who have stayed at our hotel before (repeat customers) are not concerned about reviews of our hotel as they have already experienced the hotel and base their decision on the competitor's rating at the time of booking and the differences in prices. So we assume a functional form on the distribution of the probability of a purchase and estimate the parameters of the true quality distribution based on the number of repeat customers (adjusted by this quality distribution) matching a fraction of the market-size. We believe *repeat customers* have higher chances to experience the products/services and so they are more likely to observe the true quality of the products.

Hypothesis 2. *Commercial website ratings are not representative of the general customer opinion and there is a considerable bias.*

Hypothesis 3. *The percentage of repeat customers is indicative of true quality.*

We first illustrate a partial confirmation of repeat customer hypothesis in simulations to give valuable insight into the differences between the true quality and the review distributions. However, the best way to answer this question is doing a random sampling of the customers, and, overcoming the usual problems with surveys. However, surveys are costly and time-consuming and it is difficult to do it at every property. Moreover one would like to do this test periodically. Meanwhile a data-based procedure would allow us to scale it to any number of properties and can be done regularly as a control mechanism.

a Structural model

First, we want to understand the relationship between repeat customers' bookings and competitor ratings. Since the repeat customer bookings are highly skewed, we take the log of this variable and run the following linear regression model:

$$\log(R_t) = r_t^c + \epsilon_t$$

where R_t is our repeat customers who booked on day t and r_t^c is competitor rating at the time of booking t. Accordingly, one point increase in competitor ratings decreases the number of

	Dependent variable:
	Repeat customers
Competitor ratings	-0.212^{***}
	(0.052)
Constant	19.236^{***}
	(4.175)
Observations	237
R^2	0.066
Adjusted R^2	0.062
Residual Std. Error	0.567 (df = 235)
F Statistic	16.499^{***} (df = 1; 235)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 3.5: Repeat customers vs. competitor ratings

repeat customers in our hotel by about 21.04%.

Based on the assumption of our repeat customer behavior hypothesis, we obtain the histogram of true quality frequency, given in Figure 3.1. On the horizontal line we see the competitor ratings, grouped by 1 point intervals and the vertical line shows the fraction of repeat customers. In line with our expectation as the ratings of competitor improve, the number of repeat customers to our hotel decreases (excluding the first $78 - 79$ bar).

Next, we impose the following market-share model

$$R_{td} = f M_{td} Pr(q > r_t^c) \tag{3.3}$$

where R_{td} is our repeat customers who booked on day t to start their stay on day d, M_{td} is total market size excluding no-purchases and f is fraction of customers who are in the market and had already stayed at our place.

We do not know M_{td} but we have information on M_d. So, we need to sum over all book-

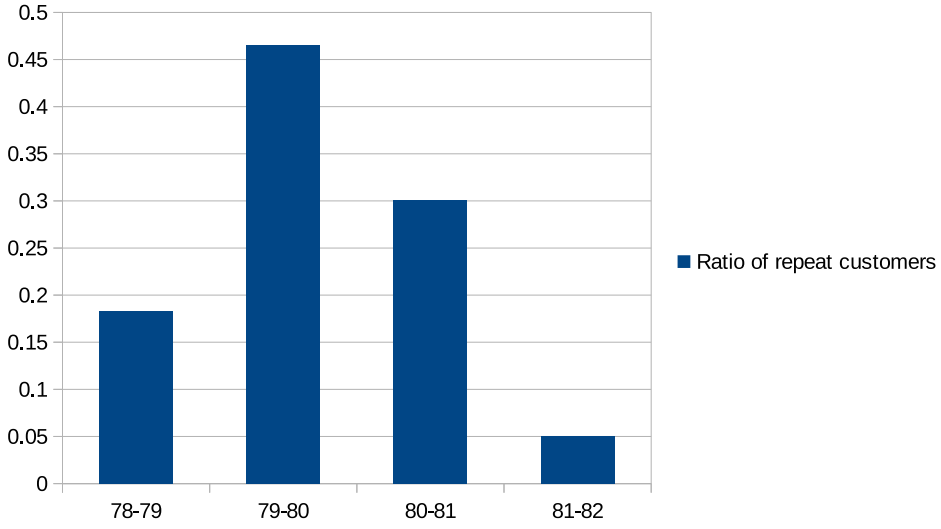


Figure 3.1: Our repeat customers' ratio as a function of competitor ratings.

ing days t for each stay start day d. Then we obtain

$$\sum_t \frac{R_{td}}{Pr(q > r_t^c)} = fM_d$$

We follow 2-step estimation strategy, in the first step we can get rid of the fraction f by taking the ratio of market size M between consequent stay start days

$$\frac{\sum_{t=1}^{d_i} \frac{R_{td_i}}{Pr(q>r_t^c)}}{\sum_{t=1}^{d_{i+1}} \frac{R_{td_{i+1}}}{Pr(q>r_t^c)}} = \frac{M_{d_i}}{M_{d_{i+1}}}$$

This reduces the problem into one unknown only and we can solve this system using non-linear least squares method to minimize the sum of square errors. We assume $Pr(q > r_t^c)$ is given by logistic distribution and

$$Pr(q > r_t^c) = \frac{\exp(\frac{-r_t^c+\mu}{s})}{1+\exp(\frac{-r_t^c+\mu}{s})}.$$

Applying this form to our hotel data, we find

$$\mu = 3.75 \quad s = 0.6.$$

In the second step, we can insert the value of those parameters into the original formulation to obtain

$$Y = fM_d$$

where

$$Y = \sum_t \frac{R_{td}}{\frac{\exp(\frac{-r_t^c+3.75}{0.6})}{1+\exp(\frac{-r_t^c+3.75}{0.6})}}.$$

Running a linear regression directly

$$Y = fM_d + \epsilon_d$$

we obtain $f = 0.747$. This means if market size changes by 1 point, we expect sum of the ratio of repeat customers to the probability of success (true quality being greater than competitor ratings) to increase by 74.7%.

$$\sum_d M_d = 8698.76 \quad \sum_t \sum_d R_{td} = 2409 \quad \sum_d M_d^o = 5513.$$

The ratio of sum of our repeat customers to sum of market size is 27.7%.

b Cross-validation: Survey

In order to validate the effectiveness of our econometric model, we cross-validate our findings with a survey. So we can estimate the true quality from primary data (surveys and experiments) and then test how the data-based methods compare to the direct findings. If the data-based methods are found to be robust and accurate, it provides a very valuable methodology for hoteliers to infer accurately the true quality opinion of its customer population, monitor it periodically, and make the correct decisions.

Specifically, we collaborate with a hotel chain in U.K. We have run an on-site survey across 2 hotels within the same chain based in U.K. over a month period. We keep the names anonymous as we agreed that it is a confidential information. We collect the survey data at the entry and exit to be able to compare how customer opinions are shaped during the stay experience. The survey design can be seen in Appendix 3.C.

Total number of responses collected is 64; of those 45 of them belong to H1 and 19 of them belong to H2. The first time customers are 34 in H1 and 12 in H2. The percentage of

	1	2	3	4	5
H1	0.767	0.023	0.07	0.023	0.116
H2	0.75	0	0.062	0.062	0.125

Table 3.6: The percentage of business (5) vs. leisure (1) customers.

business and leisure customers can be seen in Table 3.6. Most of the customers recognize the hotel on booking sites and search engines (Figure 3.2). Also, through correspondence analysis, we observe that location, price and loyalty are the main drivers of customers' choices. Among them location and price are closely associated to each other (Figure 3.3).

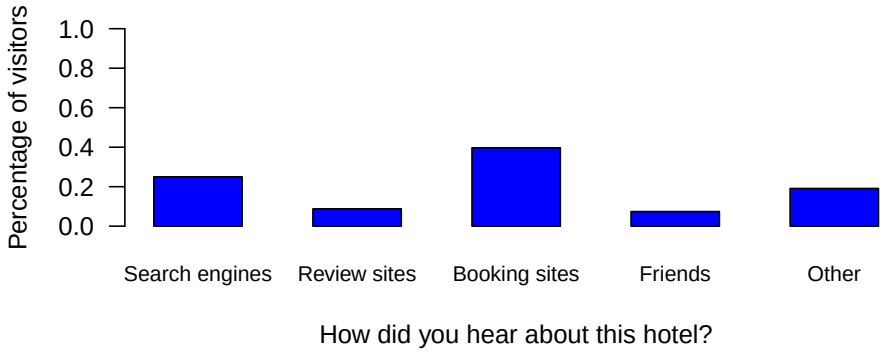


Figure 3.2: Several platforms that visitors hear about the hotel.

As a remark we are aware that the survey sample size is quite small. So, we should be cautious about the statistical tests interpretation. For instance, the usage of t-test in hypothesis testing requires the samples to be normally distributed. Nevertheless, in cases where necessary we use an adaptation of Student's t-test to account for heteroskedasticity (e.g. Welch t-test).

Is there a discrepancy between the prior expectation vs. observed experience of quality? No…

We apply paired Student's t-test with the null hypothesis being the means are the same, $H_0 : \mu_1 = \mu_2$. According to paired t-test results, t = -1.821, df = 63, p-value = 0.073. The p-value is greater than 0.05 then we can accept the hypothesis of equality of the averages. Therefore, there is no significant difference between the prior vs. posterior beliefs. 95% confidence interval is $(-0.262, 0.012)$. Sample mean of the difference is -0.125.

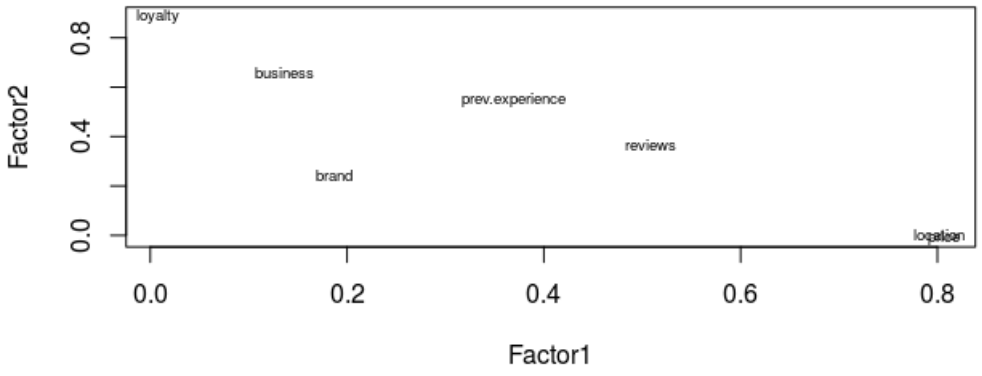


Figure 3.3: Hotel choice elements

Considering the price element, do customer valuations differ? Yes…

Firstly, we use paired t-test with the null hypothesis being the means are equal to each other, $H_0 : \mu_1 = \mu_2$. t = 4.943, df = 60, p-value = 0. Since the p-value is very low, we reject the null hypothesis. There is a strong evidence that customer valuations do differ. 95% confidence interval is $(0.312, 0.736)$. Sample mean of the differences is 0.524.

Furthermore, we test the valuation of quality being greater than value for money. We find t = 4.943, df = 60, p-value = 0. Since the p-value is very low, there is a strong evidence that customers rate quality higher than value. Sample mean of the differences is 0.524. 95% confidence interval is $(0.347, Inf)$.

Do online reviews play less important role for repeat vs. first time customers on purchase decisions? Yes…

We compare the effect of online reviews for the repeat vs. first time customers' purchase decisions. We use one sided t-test for independent groups with a hypothesis that the effect of online reviews for repeat customers is less than the first time customers. We find, t = -1.960, df = 24.405, p-value = 0.030. Since the p-value is less than 0.05, we reject the null hypothesis. There is a strong evidence that the effect of online reviews for repeat customers $(M = 3.058, s = 1.853)$ is less than the first time customers $(M = 4.043, s = 1.519)$. 95% confidence interval is $(-Inf, -0.126)$. Sample mean of the difference is -0.984.

Do repeat vs. first time customers' opinions (satisfaction levels) significantly differ? No…

We compare the means of two groups' ratings using t test. We apply the test to all hotels, we find t=-0.821, df = 32.13, p-value = 0.4175. We do not reject the null, so there is no sufficient evidence to conclude that the repeat ($M = 4.157$) and first time customer opinions ($M = 4.326$) significantly differ. 95% confidence interval is $(-0.586, 0.249)$ However, we observe that repeats rate it slightly lower than first time visitors. The mean of likeliness to come back to the hotel on the next visit is 3.925, which is quite close to repeat customers' satisfaction level.

Moreover, repeat customers seem to be more coherent about their replies. More precisely, the mean of satisfaction level and likeliness to come back to the hotel are quite close for repeat customers, sequentially $(4.157, 4.210)$ than the first time customers $(4.326, 3.812)$.

Thus, we find partial evidence to our Hypothesis 2 that states "The percentage of repeat customers is indicative of true quality." In order to truly validate this, it would be necessary to extend this survey study to a large number of hotels.

Are commercial website ratings representative of the general customer opinion? Yes…

We test whether there exists a discrepancy between customers' opinions (once we randomize the sample by running a survey design) and the online review ratings. We compare our collected sample with Tripadvisor ratings. In the survey the satisfaction level of H1 customers is $(M = 4.4, s = 0.78)$ and H2 customers is $(M = 3.94, s = 0.62)$. In the Tripadvisor sample, the satisfaction level of H1 customers is $(M = 4.21, s = 0.99, n = 3211)$ and H2 customers is $(M = 4.02, s = 1.08, n = 2118)$. We compare the means of these two independent groups using t test for each hotel. For H1, we find t = 0.563, df = 18.996, p-value = 0.579. We do not reject the null, so there is no sufficient evidence to conclude that the customer valuations significantly differ. 95% confidence interval is $(-0.220, 0.383)$. For H2, we find t = -1.565, df =46.025, p-value = 0.124. Once again, we do not have sufficient evidence to conclude that customer valuations in survey sample (randomized) and online platforms differ. 95% confidence interval is $(-0.421, 0.052)$.

Based on this, we reject Hypothesis 1 claiming that "Commercial website ratings are not representative of all the general customer opinion and there is a considerable bias".

3.5 Reviews and value

So far, we have considered the effect of reviews (and prices) on sales and whether the reviews reflect the true quality. But one can also ask if the reviews themselves are affected by price. If

so, this has implications on the pricing strategy of the firm. The firm may choose to price low (assuming low prices lead to higher reviews) so that customers continue to give higher ratings which in turn increases its market share.

The model is that each hotel has an inherent unobservable true quality Q. A customer i who stayed at the hotel experiences this quality plus an idiosyncratic 0-mean shock Q_i, so the quality he experiences (if he decides to stay, ex-post) is the realization of $Q+Q_i$, namely $q+q_i$. Now at the moment of purchase the customer is not aware of neither q nor q_i and bases his purchase decision on a review signal r, as a proxy to quality, and $E[Q|r]$ and the price p. After he stays at the hotel, he realizes the value of $q + q_i$ and bases his review on this realization and the price he paid (value).

So, we model review left by customer i as an increasing function $R(\frac{q+q_i}{p})$, that is higher value leads to better reviews. Ex-ante, his utility is given by $\beta_0 + \beta_1 p + \beta_2 E[Q|r]$, as in our generalized linear model.

Price given in the utility equation is endogenous, since it is very likely to be affected from both the reviews and unobservable quality, Q. Ex-post quality ($Q + Q_i$) is expected to be correlated with reviews.

This model, if correct, leads to an interesting problem of determining the optimal dynamic pricing strategy for the hotel. More generally, in a competitive setting, is there an equilibrium set of prices and reviews, or do reviews and prices cycle? We leave these theoretical questions to a different study, but here we test the hypothesis that indeed there is a positive correlation between review valence and *value* more than review and *price*.

In order to test the relations between variables, we apply cross-correlation function. One should focus on the negative lags to infer the effect of variable x on variable y. In Figure 3.4, we observe that as the price of our hotel increases, our ratings decrease. This effect is expected to occur after 20 days interval. Moreover, as the price of our hotel increases the relative rating of our hotel, which is defined as the proportion of X's rating to the competitor's, is again expected to decrease (Figure 3.5). We also check the relation between *value*— defined as the proportion of our price to the competitor — and relative ratings. Through Figure 3.6, we see that value is positively correlated to relative ratings. So, once customers have decided to stay in the area, price is a relative perception with respect to the other offerings rather than an absolute number. Moreover, this effect is expected to occur relatively later (around 30 days) than the effect of price.

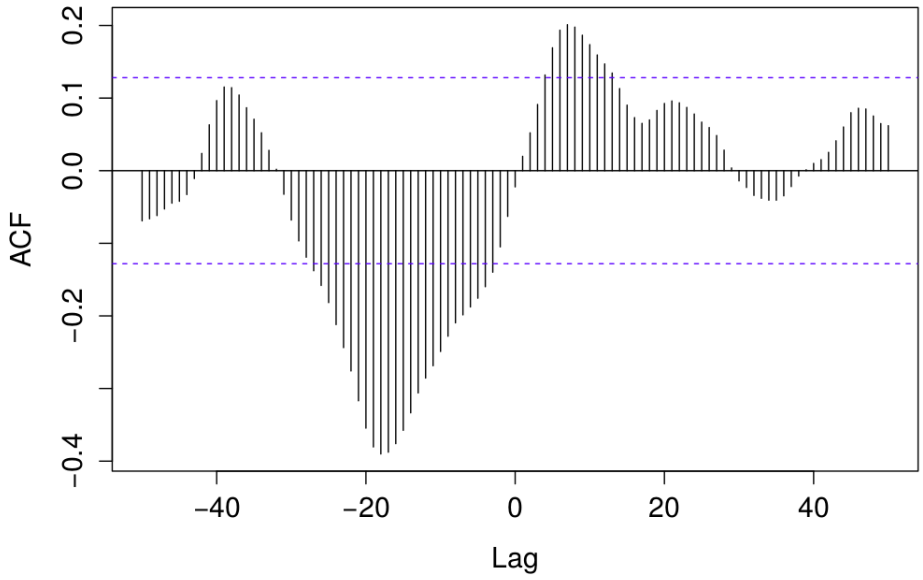


Figure 3.4: Cross correlation function for price and ratings of Hotel X.

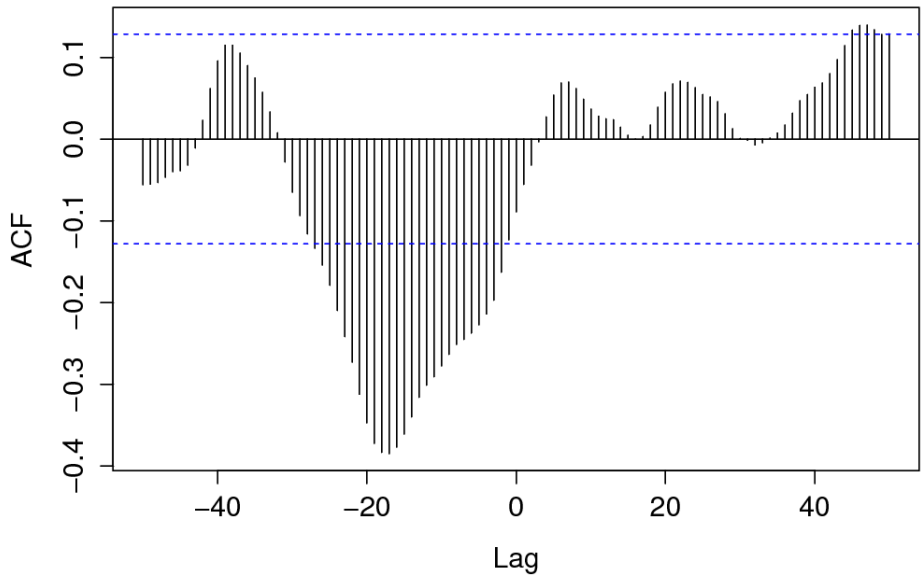


Figure 3.5: Cross correlation function for price of Hotel X and relative ratings of Hotel X.

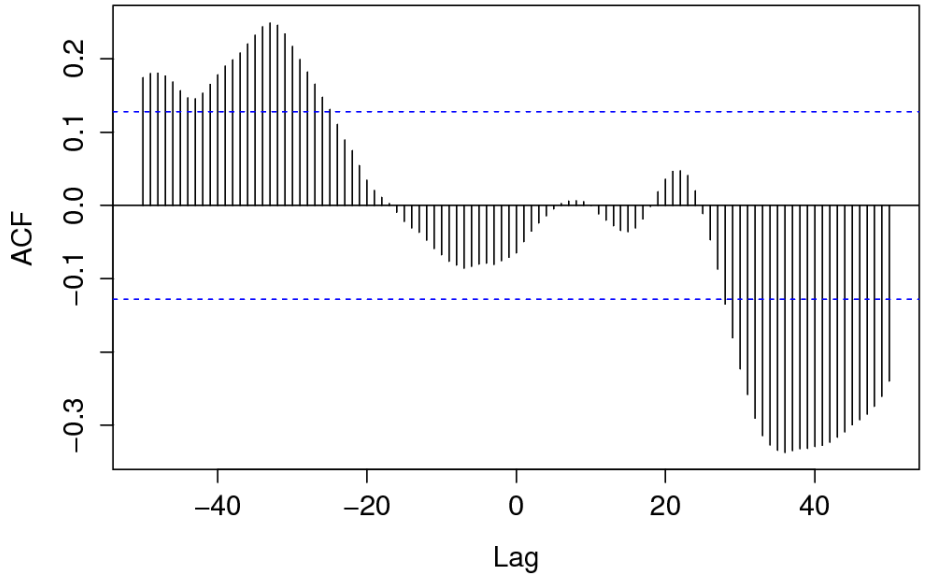


Figure 3.6: Cross correlation function for value of Hotel X and relative ratings of Hotel X.

3.6 Conclusions

In this project, we analyze the real value of reputation mechanisms. Social network platforms are a complex phenomenon and it requires a fine-grained analysis to understand the dynamics at a whole extent. Firstly, we find that the online customer reviews improve the information available about hotels through econometric models. The impact of this information is larger for the early purchasers. Moreover, price also significantly affect customers' decisions. As an interesting point, we find that for the very early purchasers group, the impact of reviews is stronger than the impact of price. Secondly, we cross-validate our findings with a survey study. Supporting our finding of the first section, we observe that price do affect customer valuations. We have no sufficient evidence to conclude there is discrepancy between customer reviews in online platforms and our collected survey sample. Moreover we find partial evidence to the hypothesis that the percentage of repeat customers is indicative of true quality. Finally, we empirically show there is a positive correlation between review valence and *value* more than review and *price*.

One of the limitations of this study is about the sample size of the survey. Since the survey is paper based, we do not have sufficiently large number of responses due to time limitations. This makes our conclusions less strong and therefore brings up the reliability issue. Moreover, it would be interesting to understand the relation between reviews and value in a complete

manner. One can theoretically study to understand whether is there an equilibrium set of prices and reviews in a competitive environment, or do reviews and prices cycle?

Appendices

3.A Sample data for log-log regression model

booking_date	days prior	stay_start	no_booked	Review score	no_ stayed	fractions
05-04-12	0	05-04-12	14	85.2	41	0.341463
04-04-12	1	05-04-12	2	84.2	41	0.04878
03-04-12	2	05-04-12	6	84.2	41	0.146341
02-04-12	3	05-04-12	1	84.2	41	0.02439
01-04-12	4	05-04-12	1	84.2	41	0.02439
28-03-12	8	05-04-12	1	82	41	0.02439
26-03-12	10	05-04-12	1	81.1	41	0.02439
23-03-12	13	05-04-12	2	82.4	41	0.04878
22-03-12	14	05-04-12	1	82.4	41	0.02439
13-03-12	23	05-04-12	2	76.4	41	0.04878
02-03-12	34	05-04-12	1	83.6	41	0.02439
08-02-12	57	05-04-12	6	80.8	41	0.146341
07-02-12	58	05-04-12	3	79.2	41	0.073171
06-04-12	0	06-04-12	23	85.2	74	0.310811
05-04-12	1	06-04-12	4	85.2	74	0.054054
04-04-12	2	06-04-12	10	84.2	74	0.135135
03-04-12	3	06-04-12	3	84.2	74	0.040541
02-04-12	4	06-04-12	5	84.2	74	0.067568
01-04-12	5	06-04-12	1	84.2	74	0.013514
30-03-12	7	06-04-12	2	84	74	0.027027
29-03-12	8	06-04-12	1	83.2	74	0.013514
28-03-12	9	06-04-12	1	82	74	0.013514
27-03-12	10	06-04-12	2	81.1	74	0.027027
25-03-12	12	06-04-12	1	81.1	74	0.013514
23-03-12	14	06-04-12	1	82.4	74	0.013514
22-03-12	15	06-04-12	1	82.4	74	0.013514
19-03-12	18	06-04-12	1	82	74	0.013514
17-03-12	20	06-04-12	1	79.6	74	0.013514
13-03-12	24	06-04-12	1	76.4	74	0.013514

Table 3.7: A sample of data used in log-log regression model.

3.B Sample data for logit regression model

Days prior	stay_start	bookings	M_d	Bookings_comp	Review score	ADR
0	01-12-12	13	545.0645	532	85.5	67.24
1	01-12-12	11	545.0645	534	86.7	67.24
2	01-12-12	3	545.0645	542	86.7	67.24
3	01-12-12	2	545.0645	543	86.7	67.24
4	01-12-12	5	545.0645	540	80.1	67.24
5	01-12-12	1	545.0645	544	78.9	67.24
6	01-12-12	3	545.0645	542	77.3	67.24
11	01-12-12	2	545.0645	543	80	67.24
12	01-12-12	1	545.0645	544	80.7	67.24
13	01-12-12	1	545.0645	544	83.1	67.24
16	01-12-12	2	545.0645	543	83.5	67.24
27	01-12-12	1	545.0645	544	89.6	67.24
33	01-12-12	6	545.0645	539	87.5	67.24
41	01-12-12	1	545.0645	544	86.3	67.24
46	01-12-12	1	545.0645	544	82.4	67.24
51	01-12-12	1	545.0645	544	82	67.24
125	01-12-12	4	545.0645	544	89.9	67.24
66	01-12-12	1	545.0645	541	87.1	67.24
0	02-12-12	6	267.0968	261	85.5	71.58
1	02-12-12	1	267.0968	266	85.5	71.58
2	02-12-12	4	267.0968	263	86.7	71.58
3	02-12-12	2	267.0968	265	86.7	71.58
4	02-12-12	3	267.0968	264	86.7	71.58
7	02-12-12	1	267.0968	266	77.3	71.58
12	02-12-12	1	267.0968	266	80	71.58
16	02-12-12	1	267.0968	266	83.5	71.58
19	02-12-12	1	267.0968	266	82.3	71.58
24	02-12-12	1	267.0968	266	80.6	71.58
26	02-12-12	1	267.0968	266	87.5	71.58
34	02-12-12	6	267.0968	261	87.5	71.58

Table 3.8: A sample of data used in logit regression model

3.C Survey

Code:

Dear Guest,
Thank you in advance for your help with the survey. We are researchers from Imperial College doing a research project with the permission of this hotel. We intend to study how customers make decision in real life. We assure this survey is completely confidential. i.e. your individual response will not be shared with anyone (even this hotel) in any form. We will be happy to share the results with you (email ID to send: _____). We really appreciate you taking the time to participate.

The survey consists of two parts:

1. This is the first part, to be filled when you check-in (or, as soon as possible)
2. The second part will be sent to your room and should be filled prior to your departure

The research is valuable only if we have your honest responses to both parts.

Thankfully yours: *Prof. Kalyan Talluri (Imperial College) & Muge Tekin (Phd student,UPF)*

---

1. During your visit how likely are you to be working? (e. g. meetings, study, conference etc.)
   ☐ 1 (very unlikely) ☐ 2 ☐ 3 ☐ 4 ☐ 5 (very likely)

2. How many times have you been in this city before?
   ☐ None ☐ 1 time ☐ 2 times ☐ 3 or more times

3. How many times have you been to this hotel before?
   ☐ None ☐ 1 time ☐ 2 times ☐ 3 or more times

4. How did you hear about this hotel?
   ☐ Search engines (Google, Bing etc.) ☐ Review sites (Tripadvisor, Yelp etc.)
   ☐ Booking sites (Booking, Expedia etc.) ☐ Friends ☐ Other

5. Please rate the following factors' role in your decision to choose this hotel…

	Not Important				*Very Important*	*Not Applicable*
▪ Location (neighbourhood level)	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5	☐
▪ Business (meeting room etc.)	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5	☐
▪ Online reviews	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5	☐
▪ Price	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5	☐
▪ Brand reputation	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5	☐
▪ Loyalty/Miles program	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5	☐
▪ Previous stay experience	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5	☐

6. How do you recall the rating of this hotel in online reviews?

Very Low				*Very High*	*Don't recall*
☐ 1	☐ 2	☐ 3	☐ 4	☐ 5	☐

7. How many other hotels have you seriously considered before reserving a room at this hotel?
   ☐ Only this one ☐ 1 other ☐ 2 others
   ☐ 3 or more others ☐ I didn't choose, someone else chose it for me

8. At this time, considering all the previous experiences and available information, your expectation regarding overall quality of your stay is
   ☐ 1 (very poor) ☐ 2 ☐ 3 ☐ 4 ☐ 5 (excellent)

Code:

Dear Guest,
Thank you for participating in the Entry survey. This is the Exit survey.
Please recall that the survey consists of two parts:

1. Previous one, you received at the check-in
2. This one, to be filled prior to your departure

The research is valuable only if we have your honest responses to both parts.

Thankfully yours: *Prof. Kalyan Talluri (Imperial College) & Muge Tekin (Phd student,UPF)*

---

1. Now that you have experienced the stay, your overall satisfaction is
   ☐ 1 (very poor) ☐ 2 ☐ 3 ☐ 4 ☐ 5 (excellent)

2. Please rate the following factors' role in your view of the overall "quality" of the stay

	Not Important				*Very Important*	*Not Applicable*
▪ Cleanliness	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5	☐
▪ Employees (helpfulness, courtesy)	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5	☐
▪ Room (size, features)	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5	☐
▪ Amenities (spa, gym etc.)	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5	☐
▪ Waiting times	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5	☐
▪ Price level	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5	☐

3. Considering comparable hotels, your opinion about the quality of this hotel relative to its price (value for money) is

very poor		*fair*		*excellent*	*I don't know the price*
☐ -2	☐- 1	☐ 0	☐ +1	☐+2	☐

4. How likely are you to recommend this hotel to a friend/colleague?
   ☐ 1 (very unlikely) ☐ 2 ☐ 3 ☐ 4 ☐ 5 (very likely)

5. How likely are you to post a review about this hotel on online review websites?
   ☐ 1 (very unlikely) ☐ 2 ☐ 3 ☐ 4 ☐ 5 (very likely)

6. Considering any hotel stay, please rate the following factor's role in motivating you to write a review on online review sites (e.g. Tripadvisor, Yelp)

	Not Important				*Very Important*
▪ A positive experience	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
▪ A negative experience	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
▪ A much better experience than I expected	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
▪ A much worse experience than I expected	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
▪ I want to help others to choose	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5

7. Your time estimate (roughly), if you were to write a review on online review sites is
   ☐ Little (< 5 min) ☐ Moderate (5-15 min) ☐ Longer (>15 min)

8. On your next visit to the city, how likely are you to come back to this hotel (with similar prices)?
   ☐ 1 (very unlikely) ☐ 2 ☐ 3 ☐ 4 ☐ 5 (very likely)

Bibliography

Aboolian, R., Berman, O., and Krass, D. (2007a). Competitive facility location and design problem. *European Journal of Operations Research*, 182:40–62.

Aboolian, R., Berman, O., and Krass, D. (2007b). Competitive facility location model with concave demand. *European Journal of Operational Research*, 181(2):598–619.

Abrate, G., Capriello, A., and Fraquelli, G. (2011). When quality signals talk: Evidence from the Turin hotel industry. *Tourism Management*, 32(4):912–921.

Acemoglu, D., Makhdoumi, A., Malekian, A., and Ozdaglar, A. (2017). Fast and slow learning from reviews. Technical report, National Bureau of Economic Research.

Achabal, D. D., Gorr, W. L., and Mahajan, V. (1982). MULTILOC-A MULTIPLE store location decision-model. *Journal of Retailing*, 58(2):5–25.

Ackerberg, D. (2003). Advertising, learning and consumer choice in experience goods markets: An empirical examination. *International Economic Review*, 44(3):1007–1040.

Ackerberg, D., Benkard, C. L., Berry, S., and Pakes, A. (2007). Econometric tools for analyzing market outcomes. *Handbook of Econometrics*, 6:4171–4276.

Admati, A. R. and Pfleiderer, P. (2004). Broadcasting opinions with an overconfident sender. *International Economic Review*, 45(2):467–498.

Aguirregabiria, V. and Mira, P. (2007). Sequential estimation of dynamic discrete games. *Econometrica*, 75(1):1–53.

Anderson, C. K. (2012). The impact of social media on lodging performance. *Cornell Hospitality Report*.

Anderson, E. W. (1998). Customer satisfaction and word of mouth. *Journal of Service Research*, 1(1):5–17.

Anderson, E. W. and Sullivan, M. W. (1993). The antecedents and consequences of customer satisfaction for firms. *Marketing Science*, 12(2):125–143.

Anderson, M. and Magruder, J. (2012). Learning from the crowd: Regression discontinuity estimates of the effects of an online review database. *The Economic Journal*, 122(563):957–989.

Andersson, S.-E. (1998). Passenger choice analysis for seat capacity control: A pilot project in scandinavian airlines. *International Transactions in Operational Research*, 5(6):471–486.

Askalidis, G., Kim, S. Y., and Malthouse, E. C. (2017). Understanding and overcoming biases in online review systems. *Decision Support Systems*, 97:23–30.

Bajari, P., Benkard, C. L., and Levin, J. (2007). Estimating dynamic models of imperfect competition. *Econometrica*, 75(5):1331–1370.

Banerjee, A. and Fudenberg, D. (2004). Word-of-mouth learning. *Games and Economic Behavior*, 46(1):1–22.

Ben-Akiva, M. and Lerman, S. (1985). *Discrete-Choice Analysis: Theory and Application to Travel Demand*. MIT Press, Cambridge, MA.

Ben-Akiva, M. and Watanatada, T. (1981). Application of a continuous spatial choice logit model. In Manski, C. F. and McFadden, D., editors, *Structural analysis of discrete data with econometric applications*, pages 320–343. MIT Press Cambridge, MA.

Benati, S. (1999). The maximum capture problem with heterogeneous customers. *Computers & Operations Research*, 26(14):1351–1367.

Berman, O., Drezner, T., Drezner, Z., and Krass, D. (2009). Modeling competitive facility location problems: New approaches and results. In *TutORials in Operations Research. INFORMS Annual Meeting: San Diego CA*, pages 156–181.

Berman, O. and Krass, D. (2002). Locating multiple competitive facilities: Spatial interaction models with variable expenditures. *Annals of Operations Research*, 111:197–225.

Berry, S. (1994). Estimating discrete-choice models of product differentiation. *The RAND Journal of Economics*, pages 242–262.

Berry, S., Levinsohn, J., and Pakes, A. (1995). Automobile prices in market equilibrium. *Econometrica: Journal of the Econometric Society*, pages 841–890.

Besanko, D., Gupta, S., and Jain, D. (1998). Logit demand estimation under competitive pricing behavior: An equilibrium framework. *Management Science*, 44(11-part-1):1533–1547.

Besbes, O. and Scarsini, M. (2018). On information distortions in online ratings. *Operations Research*.

Bickart, B. and Schindler, R. M. (2001). Internet forums as influential sources of consumer information. *Journal of Interactive Marketing*, 15(3):31–40.

Birkin, M., Clarke, G., and Clarke, M. P. (2002). *Retail geography and intelligent network planning*. John Wiley & Sons.

Boulding, W., Kalra, A., Staelin, R., and Zeithaml, V. A. (1993). A dynamic process model of service quality: From expectations to behavioral intentions. *Journal of Marketing Research*, 30(1):7.

Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press.

Cai, H., Chen, Y., and Fang, H. (2009). Observational learning: Evidence from a randomized natural field experiment. *The American Economic Review*, 99(3):864–882.

Caro, F. and Gallien, J. (2012). Clearance pricing optimization for a fast-fashion retailer. *Operations Research*, 60(6):1404–1422.

Chen, P., Wu, S., and Yoon, J. (2004). The impact of online recommendations and consumer feedback on sales. In *ICIS'04*, pages 711–724.

Chevalier, J. and Mayzlin, D. (2006). The effect of word of mouth on sales: Online book reviews. *Journal of Marketing Research*, 43(3):345–354.

Chintagunta, P., Goettler, R., and Kim, M. (2012). New drug diffusion when forward looking physicians learn from patient feedback and detailing. *Journal of Marketing Research*, 49(6):807–821.

Clemons, E. K., Gao, G. G., and Hitt, L. M. (2006). When online reviews meet hyperdifferentiation: A study of the craft beer industry. *Journal of Management Information Systems*, 23(2):149–171.

Coleman, J. S., Katz, E., and Menzel, H. (1996). *Medical innovation: a diffusion study*. Bobbs-Merrill, Indianapolis, IN.

Collard-Wexler, A. (2013). Demand fluctuations in the ready-mix concrete industry. *Econometrica*, 81(3):1003–1037.

Cronin, J., Joseph, J., and Taylor, S. A. (1992). Measuring service quality: A reexamination and extension. *The Journal of Marketing*, 56(3):55–68.

Cui, R., Gallino, S., Moreno, A., and Zhang, D. (2017). The operational value of social media information. *Production and Operations Management*.

De Langhe, B., Fernbach, P. M., and Lichtenstein, D. R. (2015). Navigating by the stars: Investigating the actual and perceived validity of online user ratings. *Journal of Consumer Research*, 42(6):817–833.

De Palma, A., Ginsburgh, V., Labbe, H. M., and Thisse, J. F. (1989). Competitive location with random utilities. *Transportation Science*, 23:244–252.

Dellarocas, C. (2003). The digitization of word of mouth: Promise and challenges of online feedback mechanisms. *Management Science*, 49(10):1407–1424.

Dellarocas, C. (2006). Strategic manipulation of internet opinion forums: Implications for consumers and firms. *Management Science*, 52(10):1577–1593.

Dellarocas, C., Awad, N. F., and Zhang, M. (2007). Exploring the value of online product reviews in forecasting sales: The case of motion pictures. *Journal of Interactive Marketing*, 21(4):23–45.

Deutsch, C. (1990). Managing; 007 It's not. But intelligence is in.

Dong, L., Kouvelis, P., and Tian, Z. (2009). Dynamic pricing and inventory control of substitute products. *Manufacturing & Service Operations Management*, 11(2):317–339.

Drezner, T. (1994). Locating a new facility among existing, unequally attractive facilities. *Journal of Regional Science*, 34:237–252.

Drezner, T. (2014). A review of competitive facility location in the plane. *Logistics Research*, 7(1):1–12.

Drezner, T. and Drezner, Z. (1998). Facility location in anticipation of future competition. *Location Science*, 6:155–173.

Duan, W., Gu, B., and Whinston, A. B. (2008). Do online reviews matter? An empirical investigation of panel data. *Decision Support Systems*, 45(4):1007–1016.

Dubois, P. and Nauges, C. (2010). Identifying the effect of unobserved quality and expert reviews in the pricing of experience goods: Empirical application on bordeaux wine. *International Journal of Industrial Organization*, 28(3):205–212.

Dunne, T., Klimek, S. D., Roberts, M. J., and Xu, D. Y. (2013). Entry, exit, and the determinants of market structure. *The RAND Journal of Economics*, 44(3):462–487.

Eiselt, H. A., Laporte, G., and Thisse, J. F. (1993). Competitive location models: A framework and bibliography. *Transportation Science*, 27(1):44–54.

Eliashberg, J. and Shugan, M. (1997). Film critics: Influencers or predictors? *Journal of Marketing*, 61(2):68–78.

Erdem, T. and Keane, M. (1996). Decision-making under uncertainty: Capturing dynamic brand choice processes in turbulent consumer goods markets. *Marketing Science*, 15(1):1–20.

Erdem, T., Keane, M., and Sun, B. (2008). A dynamic model of brand choice when price and advertising signal product quality. *Marketing Science*, 27(6):1111–1125.

Factual (2016). Factual US Hotels Api. `http://www.factual.com/products/global#hotels`. Accessed on 2016-11-30.

Farahani, R. Z., Rezapour, S., Drezner, T., and Fallah, S. (2014). Competitive supply chain network design: An overview of classifications, models, solution techniques and applications. *Omega*, 45:92–118.

Ferreira, K. J., Lee, B., and Simchi-Levi, D. (2015). Analytics for an online retailer: Demand forecasting and price optimization. *Manufacturing and Service Operations Management*, 18(1):69–88.

Fisher, M., Gallino, S., and Li, J. (2017). Competition-based dynamic pricing in online retailing: A methodology validated with field experiments. *Management Science*, 64(6):2496–2514.

Fisher, M. and Vaidyanathan, R. (2014). A demand estimation procedure for retail assortment optimization with results from implementations. *Management Science*, 60(10):2401–2415.

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

Ghosh, A., McLafferty, S., and Craig, C. S. (1995). Multifacility retail networks. In Drezner, Z., editor, *Facility location: A survey of applications and methods*, pages 301–330. Springer-Verlag: New York.

Gilad, B. and Fuld, L. M. (2016). Only half of companies actually use the competitive intelligence they collect. `https://hbr.org/2016/01/only-half-of-companies-actually-use-the-competitive-intelligence-they-collect`.

Glaeser, C. K., Fisher, M., and Su, X. (2017). Optimal retail location: Empirical methodology and application to practice. Working Paper.

Godes, D. and Mayzlin, D. (2004). Using online conversations to study word-of-mouth communication. *Marketing Science*, 23(4):545–560.

Grove, A. S. (2015). *High Output Management*. Vintage.

Haase, K. and Müller, S. (2014). A comparison of linear reformulations for multinomial logit choice probabilities in facility location models. *European Journal of Operations Research*, 232:689–691.

Hakimi, S. (1983). On locating new facilities in a competitive environment. *European Journal of Operations Research*, 12:29–35.

Hanson, W. and Martin, K. (1996). Optimizing multinomial logit profit functions. *Management Science*, 42(7):992–1003.

Hayashi, F. (2000). Econometrics. *New Jersey, USA: Princeton University*.

Hernandez, T. and Bennison, D. (2000). The art and science of retail location decisions. *International Journal of Retail & Distribution Management*, 28(8):357–367.

Ho, Y. C., Tan, Y., and Wu, J. (2013). Effect of disconfirmation on online rating behavior: A dynamic analysis. Working Paper.

Hotelling, H. (1929). Stability in competition. *Economic Journal*, 39:41–57.

Hu, N., Pavlou, P. A., and Zhang, J. (2006). Can online reviews reveal a product's true quality? Empirical findings and analytical modeling of online word-of-mouth communication. In *ACM'06*, pages 324–330.

Huff, D. L. (1964). Defining and estimating a trade area. *The Journal of Marketing*, 28(3):34–38.

Isaac, M. (2017). Uber's C.E.O. plays with fire. `https://nyti.ms/2p90N43`.

Isaac, M. and Lohr, S. (2017). Unroll.me service faces backlash over a widespread practice: Selling user data. `https://nyti.ms/2pYH0Eb`.

Jing, B. (2011). Social learning and dynamic pricing of durable goods. *Marketing Science*, 30(5):851–865.

Jurca, R. and Faltings, B. (2009). Mechanisms for making crowds truthful. *Journal of Artificial Intelligence Research*, 34(1):209.

Kamal, I. (2012). Metrics are easy; insight is hard. `https://hbr.org/2012/09/metrics-are-easy-insights-are-hard`.

Knittel, C. R. and Metaxoglou, K. (2014). Estimation of random-coefficient demand models: Two empiricists' perspective. *Review of Economics and Statistics*, 96(1):34–59.

Kök, A. G. and Fisher, M. L. (2007). Demand estimation and assortment optimization under substitution: Methodology and application. *Operations Research*, 55(6):1001–1021.

Koren, Y., Bell, R., and Volinsky, C. (2009). Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37.

Kuksov, D. and Xie, Y. (2010). Pricing, frills, and customer ratings. *Marketing Science*, 29(5):925–943.

Lambrecht, B. and Perraudin, W. (2003). Real options and preemption under incomplete information. *Journal of Economic Dynamics and Control*, 27(4):619–643.

Las Vegas City Hall (2014). Las Vegas restaurant seat and size information. `https://www.lasvegasnevada.gov/portal/faces/home`. Accessed on 2017-04-20.

Leeflang, P. S. H., Wittink, D. R., Wedel, M., and Naert, P. (2000). *Building Models for Marketing Decisions*. Kluwer, Dordrecht, Netherlands.

Lerman, S. R. (1985). Random utility models of spatial choice. In *Optimization and Discrete Choice in Urban Systems*, pages 200–217. Springer.

Li, H. and Huh, W. T. (2011). Pricing multiple products with the multinomial logit and nested logit models: Concavity and implications. *Manufacturing & Service Operations Management*, 13(4):549–563.

Li, H., Ye, Q., Law, R., and Wang, Z. (2014a). The influence of hotel price on perceived service quality and value in e-tourism: An empirical investigation based on online traveler reviews. *Journal of Hospitality and Tourism Research*, 38(1):23–39.

Li, J., Granados, N., and Netessine, S. (2014b). Are consumers strategic? structural estimation from the air-travel industry. *Management Science*, 60(9):2114–2137.

Li, X. and Hitt, L. M. (2008). Self-selection and information role of online product reviews. *Information Systems Research*, 19(4):456–474.

Li, X. and Hitt, L. M. (2010). Price effects in online product reviews: An analytical model and empirical analysis. *MIS Quarterly*, 34(4):809–831.

Liu, Y. (2006). Word of mouth for movies: Its dynamics and impact on box office revenue. *Journal of Marketing*, 70(3):74–89.

Liu, Y. (2010). Understanding price/value rating in online consumer review. In *ICEC'10*, pages 48–57.

Lounamaa, P. H. and March, J. G. (1987). Adaptive coordination of a learning team. *Management Science*, 33(1):107–123.

Luca, M. (2011). Reviews, reputation, and revenue: The case of yelp.com. Working Paper, Harvard Business School NOM Unit, Harvard University.

Marianov, V., M., R., and M.J., I. (2008). Facility location for market capture when users rank facilities by shorter travel and waiting times. *European Journal of Operations Research*, 191:32–44.

Martínez-de Albéniz, V. and Talluri, K. (2011). Dynamic price competition with fixed capacities. *Management Science*, 57(6):1078–1093.

Mayzlin, D. (2006). Promotional chat on the internet. *Marketing Science*, 25(2):155–163.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In Zarembka, P., editor, *Frontiers in Econometrics*, pages 105–142. Academic Press, New York, NY.

McLean, S. and Samavi, M. (2015). Data for the taking: Using website terms and conditions to combat web scraping. `https://www.sociallyawareblog.com/2015/03/12/data-for-the-taking-using-website-terms-and-conditions-to-combat-web-scraping`.

Mitra, D. and Golder, P. N. (2006). How does objective quality affect perceived quality? Short-term effects, long-term effects, and asymmetries. *Marketing Science*, 25(3):230–247.

Moutinho, L., Curry, B., and Davies, F. (1993). Comparative computer approaches to multi-outlet retail site location decisions. *Service Industries Journal*, 13(4):201–220.

Murphy, C. (2007). *Competitive Intelligence: Gathering, Analyzing and Putting it to Work*. Routledge, London.

Nakanishi, M. and Cooper, L. G. (1974). Parameter estimates for multiplicative competitive interaction models-least square approach. *Journal of Marketing Research*, 11:303–311.

Narayanan, S., Manchanda, P., and P.K., C. (2005). Temporal differences in the role of marketing communication in new product categories. *Journal of Marketing Research*, 42(3):278–290.

NASA (2016). NASA Socioeconomic data and applications center (Sedac). `http://sedac.ciesin.columbia.edu/data/collection/gpw-v4`. Accessed on 2016-11-30.

Neelamegham, R. and Chintagunta, P. (1999). A bayesian model to forecast new product performance in domestic and international markets. *Marketing Science*, 18(2):115–136.

Nevo, A. (2001). Measuring market power in the ready-to-eat cereal industry. *Econometrica*, 69(2):307–342.

Nevskaya, Y. (2012). Consumer information asymmetry in online product reviews. Working Paper, New York: William E. Simon Graduate School of Business, University of Rochester.

Newman, J. P., Ferguson, M. E., Garrow, L. A., and Jacobs, T. L. (2014). Estimation of choice-based models using sales data from a single firm. *Manufacturing & Service Operations Management*, 16(2):184–197.

Oliver, R. L. (1980). A cognitive model of the antecedents and consequences of satisfaction decisions. *Journal of Marketing Research*, 17(4):460–469.

Olley, G. S. and Pakes, A. (1996). The dynamics of productivity in the telecommunications equipment industry. *Econometrica*, 64:1263–1298.

OpenTable (2016). Opentable: Restaurant reservation software. `http://www.opentable.com`. Accessed on 2016-11-30.

Owen, S. H. and Daskin, M. S. (1998). Strategic facility location: A review. *European Journal of Operations Research*, 111(3):423–447.

Pakes, A., Ostrovsky, M., and Berry, S. (2007). Simple estimators for the parameters of discrete dynamic games (with entry/exit examples). *The RAND Journal of Economics*, 38(2):373–399.

Pancras, J., Sriram, S., and Kumar, V. (2012). Empirical investigation of retail expansion and cannibalization in a dynamic environment. *Management Science*, 58(11):2001–2018.

Papanastasiou, Y., Bakshi, N., and Savva, N. (2015). Social learning from early buyer reviews: Implications for new product launch. Technical report, Working Paper, London Business School, London, UK.

Parasuraman, A., Zeithaml, V. A., and Berry, L. L. (1985). A conceptual model of service quality and its implications for future research. *The Journal of Marketing*, 49(4):41–50.

Parsa, H. G., Self, J. T., Njite, D., and King, T. (2005). Why restaurants fail. *Cornell Hotel and Restaurant Administration Quarterly*, 46(3):304–322.

Phelps, N. A. and Wood, A. M. (2018). The business of location: site selection consultants and the mobilisation of knowledge in the location decision. *Journal of Economic Geography*.

Phocuswright (2011). Phocuswright U.S. consumer travel report fourth edition. `http://www.phocuswright.com/products/4074`. Accessed on 2011.

Pioch, E. and Byrom, J. (2004). Small independent retail firms and locational decision-making: outdoor leisure retailing by the crags. *Journal of Small Business and Enterprise Development*, 11(2):222–232.

Plastria, F. and Carrizosa, E. (2004). Optimal location and design of a competitive facility. *Mathematical Programming*, 100(2):247–265.

Plastria, F. and Vanhaverbeke, L. (2008). Discrete models for competitive location with foresight. *Computers & Operations Research*, 35(3):683–700.

Poole, R., Clarke, G. P., and Clarke, D. B. (2006). Competition and saturation in west european grocery retailing. *Environment and Planning A*, 38(11):2129–2156.

Ramakrishnan, R. (2017). I have data. i need insights. where do i start? `https://towardsdatascience.com/i-have-data-i-need-insights-where-do-i-start-7ddc935ab365`.

Ratliff, R. M., Rao, B. V., Narayan, C. P., and Yellepeddi, K. (2008). A multi-flight recapture heuristic for estimating unconstrained demand from airline bookings. *Journal of Revenue and Pricing Management*, 7(2):153–171.

Resnick, P. and Zeckhauser, R. (2002). Trust among strangers in internet transactions: Empirical analysis of eBay's reputation system. In Bay, M. R., editor, *The Economics of the Internet and E-Commerce, Advances in Applied Microeconomics*, volume 11, pages 127–157. Elsevier Science, Amsterdam.

Roberts, J. and Urban, G. (1988). Modeling multiattribute utility, risk and belief dynamics for new consumer durable brand choice. *Management Science*, 34(2):167–185.

Rogers, D. S. (1984). *Store Location and Store Assessment Research*. John Wiley & Sons Inc.

Senecal, S. and Nantel, J. (2004). The influence of online product recommendations on consumers' online choices. *Journal of Retailing*, 80(2):159–169.

Sharma, K. K. and Krishna, H. (1994). Asymptotic sampling distribution of inverse coefficient-of-variation and its applications. *IEEE Transactions on Reliability*, 43(4):630–633.

Song, J. S. and Xue, Z. (2007). Demand management and inventory control for substitutable products. Working Paper, Fuqua School of Business, Duke University, Durham, NC.

Sreenivasan, S. (1998). Taking in the sites; corporate intelligence: A cloakhold on the web.

Statista (2016). Average occupancy rate of U.S. hotels. `http://www.statista.com/statistics/200161/us-annual-accomodation-and-lodging-occupancy-rate/`. Accessed on 2016-11-30.

Su, C. L. and Judd, K. L. (2012). Constrained optimization approaches to estimation of structural models. *Econometrica*, 80(5):2213–2230.

Sun, M. (2012). How does the variance of product ratings matter? *Management Science*, 58(4):696–707.

Talluri, K. and Van Ryzin, G. (2004). Revenue management under a general discrete choice model of consumer behavior. *Management Science*, 50(1):15–33.

Talluri, K. T. (2009). A finite-population revenue management model and a risk-ratio procedure for the joint estimation of population size and parameters. Technical report, Working Paper 1141, Department of Economics, Universitat Pompeu Fabra, Barcelona, Spain.

Tebaldi, C. and West, M. (1998). Bayesian inference on network traffic using link count data. *Journal of the American Statistical Association*, 93(442):557–573.

Tellis, G. J. and Johnson, J. (2007). The value of quality. *Marketing Science*, 26(6):758–773.

Toubia, O. and Stephen, A. T. (2013). Intrinsic versus image-related utility in social media: Why do people contribute content to twitter? *Marketing Science*, 32(3):368–392.

Trigeorgis, L. (1991). Anticipated competitive entry and early preemptive investment in deferrable projects. *Journal of Economics and Business*, 43(2):143–156.

U.S. Census Bureau (2014). U.S. Census Bureau. `https://www.census.gov/en.html`. Accessed on 2016-11-30.

Van den Bulte, C. and Lilien, G. L. (2001). Medical innovation revisited: Social contagion versus marketing effort. *American Journal of Sociology*, 106(5):1409–1435.

Vanderbei, R. J. and Iannone, J. (1994). An EM approach to O-D matrix estimation. Technical report, SOR 94-04, Princeton University.

Vardi, Y. (1996). Network tomography: Estimating source-destination traffic intensities from link data. *Journal of the American statistical association*, 91(433):365–377.

Vives, X. (2001). *Oligopoly Pricing: Old Ideas and New Tools*. MIT press.

Vulcano, G., Van Ryzin, G., and Chaar, W. (2010). Choice-based revenue management: An empirical study of estimation and optimization. *Manufacturing & Service Operations Management*, 12(3):371–392.

Vulcano, G., Van Ryzin, G., and Ratliff, R. (2012). Estimating primary demand for substitutable products from sales transaction data. *Operations Research*, 60(2):313–334.

Wang, M., Lu, Q., Chi, R., and Shi, W. (2015). How word-of-mouth moderates room price and hotel stars for online hotel booking. *Journal of Electronic Commerce Research*, 16(1):72–80.

Wood, S. and Browne, S. (2007). Convenience store location planning and forecasting–a practical research agenda. *International Journal of Retail & Distribution Management*, 35(4):233–255.

Wood, S. and Reynolds, J. (2012). Leveraging locational insights within retail store development? assessing the use of location planners' knowledge in retail marketing. *Geoforum*, 43(6):1076–1087.

Wood, Z. and Butler, S. (2018). Seven reasons why Marks & Spencer is in trouble. `https://www.theguardian.com/business/2018/may/23/seven-reasons-why-marks-spencer-is-in-trouble`.

Wrigley, N. (1996). Sunk cost and corporate restructuring: British food retailing and property crisis. In Wringley, N. and Lowe, M., editors, *The Oxford Handbook of Innovation*, pages 116–136. Longman, London.

Wu, C. F. J. (1983). On the convergence properties of the em algorithm. *The Annals of Statistics*, pages 95–103.

Ye, Q., Li, H., and Wang, Z. (2014). The influence of hotel price on perceived service quality and value in e-tourism: An empirical investigation based on online traveler reviews. *Journal of Hospitality and Tourism Research*, 38(1):23–39.

Yelp (2016). Yelp dataset challenge. `https://www.yelp.com/dataset_challenge`. Accessed on 2016-11-30.

Yoo, K. H. and Gretzel, U. (2008). What motivates consumers to write online travel reviews? *Information Technology & Tourism*, 10(4):283–295.

Yu, M., Debo, L., and Kapuscinski, R. (2015). Strategic waiting for consumer generated quality information: Dynamic pricing of new experience goods. *Management Science*, 62(2):410–435.

Zeithaml, V. A. (1988). Consumer perceptions of price, quality, and value: A means-end model and synthesis of evidence. *The Journal of Marketing*, 52(3):2–22.

Zhang, J. (2010). The sound of silence: Observational learning from the U.S. kidney market. *Marketing Science*, 29(2):315–335.

Zhang, M. and Luo, L. (2016). Can user generated content predict restaurant survival: Deep learning of yelp photos and reviews. University of Southern California.

Zhang, X. M. and Dellarocas, C. (2006). The lord of the ratings: How a movie's fate is influenced by reviews. In *ICIS'06*, page 117.

Zhao, Y., Yang, S., Narayan, V., and Zhao, Y. (2013). Modeling consumer learning from online product reviews. *Marketing Science*, 32(1):153–169.