UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA INDUSTRIAL

DATA DRIVEN ANALYTICS IN CAPACITY AND SERVICE QUALITY MANAGEMENT

TESIS PARA OPTAR AL GRADO DE DOCTOR EN SISTEMAS DE INGENIERÍA

DANIEL IVÁN YUNG MEYOHAS

PROFESOR GUÍA:
MARCELO OLIVARES ACUÑA

PROFESOR CO-GUÍA:
ANDRÉS MUSALEM SAID

MIEMBROS DE LA COMISIÓN:
VÍCTOR BUCAREY LÓPEZ
FELIPE CARO VALDÉS
CRISTIÁN GUEVARA CUE

SANTIAGO DE CHILE
2019

RESUMEN DE LA MEMORIA PARA OPTAR
AL TÍTULO DE DOCTOR EN SISTEMAS DE INGENIERÍA
POR: DANIEL IVÁN YUNG MEYOHAS
FECHA: 2019
PROF. GUÍA: MARCELO OLIVARES ACUÑA

DATA DRIVEN ANALYTICS IN CAPACITY AND SERVICE QUALITY MANAGEMENT

La administración de recursos humanos es de suma importancia, ya sea en una tienda física o una plataforma en línea, ya que contribuyen a determinar la relación entre la firma y sus clientes. En muchas industrias de servicio, la fuerza de trabajo representa uno de los principales costos, mientras que proporcionar una rápida atención puede representar una importante ventaja competitiva, sobre todo si los clientes son sensibles al tiempo de espera o requieren asistencia especializada para efectuar una compra. Cuando la demanda es variable y la dotación no puede ser ajustada rápidamente, las decisiones de capacidad requieren balancear la capacidad de respuesta de la firma con los costos operacionales a través de la utilización. Este trabajo extiende recientes estudios empíricos que estiman el impacto de la dotación y calidad del despacho en las ventas, combinando datos transaccionales y de demanda con datos de dotación, desarrollando un enfoque econométrico para modelar el comportamiento de los consumidores en el contexto de un call-center, un retailer online de ropa y una cadena de tiendas tradicionales. La estimación de los modelos es desafiante debido a problemas de endogeneidad. Este trabajo desarrolla herramientas de apoyo a la toma de decisiones para administrar la dotación de personal y la definición de turnos de trabajo, considerando la relación causal entre personal y ventas.

En el contexto del call-center estudiado en el capítulo 1, se encuentra un fuerte impacto del tiempo de espera en el comportamiento de los consumidores: ventas un 50% más altas si los clientes son contactados dentro de los primeros 5 minutos de haber visitado la plataforma en línea en lugar de una hora después. La heterogeneidad en la productividad de los ejecutivos del call-center es relevante. Un alza de $7,7\%$ en la utilización genera un alza de entre $0,9\%$ y $3,6\%$ en la conversión debido a una mayor retención de los ejecutivos más calificados. El capítulo 2 presenta una investigación aplicada que desarrolla una novedosa herramienta de apoyo en la toma de decisiones de personal en el sector del retail, combinando ideas de diferentes ramas de la gestión de operaciones. Los resultados empíricos revelan un efecto no lineal del tráfico y los niveles de dotación en las ventas. Este efecto fue desagregado en conversión y tamaño de la canasta, encontrando que la mayor parte del efecto es a través de la conversión, cuya magnitud es comparable a la de estudios previos y del orden del $2\% - 5\%$ al aumentar la fuerza de trabajo en tiendas sub-dotadas. En el capítulo 3 se investiga la relación entre la calidad del despacho y el comportamiento futuro de los consumidores. Se encuentra que experiencias negativas con el servicio de despacho impactan adversamente el desempeño del retailer al desalentar futuras compras. No se encontró evidencia de un efecto directo en el tamaño de la canasta, aunque sí a través de otros mecanismos, particularmente la frecuencia de compra. Mejorar la calidad del servicio de despacho puede aumentar las utilidades en hasta un $5,63\%$, a través de un aumento en la frecuencia de compra $(4,91\%)$ y del monto gastado $(0,69\%)$.

RESUMEN DE LA MEMORIA PARA OPTAR
AL TÍTULO DE DOCTOR EN SISTEMAS DE INGENIERÍA
POR: DANIEL IVÁN YUNG MEYOHAS
FECHA: 2019
PROF. GUÍA: MARCELO OLIVARES ACUÑA

DATA DRIVEN ANALYTICS IN CAPACITY AND SERVICE QUALITY MANAGEMENT

Personnel decisions are of the utmost importance and, whether a brick and mortar store or an online platform, they contribute to determine the relationship between the firm and its customers. In many service industries, labor costs are one of the largest expenses and providing prompt response to customers can be an important competitive advantage, specially when customers are time-sensitive or require specialized assistance to make a purchase. When demand for service is variable and staffing requirements cannot be adjusted quickly, capacity decisions require making a trade-off between responsiveness to customers versus controlling operating costs through worker utilization. This work extends recent empirical studies that estimate the impact of staffing levels and delivery performance on sales, combining detailed sales transaction and demand data with employee staffing data, and developing an econometric approach to model customer behavior in the context of an out-bound call center, an online apparel retailer and brick and mortar stores. Model parameters estimation is challenging due to endogeneity issues. This work develops decision support tools to manage staffing levels and schedule working shifts, balancing the sales contribution of employees with salary costs, accounting for the causal relation between personnel and sales.

For the outbound call center under study in chapter 1, a strong impact of waiting time on customer behavior was found: conversion rates are 50% greater if a customer is contacted within the first 5 minutes of visiting the platform's website rather than after an hour. Considering the role of agents, heterogeneity in worker productivity is relevant, since a 7.7% increase in utilization translates into a 0.9% to 3.6% increase in conversion rate due to higher retention of more skilled agents under medium traffic conditions. Chapter 2 presents a practice-based research project that develops a novel solution to build a decision support tool for workforce planning in the retail sector. The developed solution combines ideas from different research streams in Operations Management. The empirical results reveal a non-linear effect of traffic and staffing levels on sales. This effect was decomposed into conversion and basket value and found that most of the effect is through the former. The magnitude of the effect is comparable to previous results in the literature: increasing labor in under-staffed stores can increase sales in the order of $2\% - 5\%$. In chapter 3 the relation between order fulfillment and future customer behavior is investigated. Negative delivery experiences produce an adverse effect on the retailer performance by discouraging customers to purchase again. While evidence of a direct impact of delivery quality on basket size wasn't found, delivery performance can impact amount spent through other mechanisms, particularly purchase incidence. Improving delivery speed can yield a 5.63% increase in revenues if all customers are satisfied with their order's fulfillment, stemming from a 4.91% increase in purchase probability and a 0.69% increase in basket size.

A mis padres, que me han dado todo...

A mi hermano, que siempre me apoya...

A mi amada, que siempre me acompaña...

A mis amigos, que estuvieron conmigo...

... y a todos ustedes!

Los quiero mucho!!

Acknowledgements

Agradezco a mis padres, por su constante apoyo y amor incondicional. Todo lo que soy se lo debo a ustedes, muchas gracias por estar siempre conmigo, los quiero mucho!!

Agradezo a mi hermano, que siempre está conmigo... en las buenas y en las malas. Su cariño, aunque a la distancia, es invaluable.

Agradezco a Ale por acompañarme siempre, por quererme y tenerme paciencia. Desde que te conocí ya no estoy solo...

Agradezco a mis amigos, por los buenos momentos y las memorias que hemos creado. Siempre han estado cuando los he necesitado, espero estar a la altura y estar para ustedes cuando me necesiten.

Agradezco a Marcelo y Andrés por su constante apoyo durante esta etapa, no solo en el plano académico. Ha sido un privilegio aprender de ustedes.

I'm very grateful to Nitish and Karan. Thank you for trusting me and giving me the opportunity to work with you.

Agradezco al programa de formación de capital humano avanzado de CONICYT-PCHA. La beca de Doctorado Nacional año 2015 - folio 21150899 financió mis estudios de doctorado y financió parcialmente mi trabajo de investigación.

Table of Contents

List of Tables

List of Figures

Introduction

Personnel decisions are of the utmost importance and, whether a brick and mortar store or an online platform, they contribute to determine the relationship between the firm and its customers. In many service industries, labor costs are one of the largest expenses and providing prompt response to customers can be an important competitive advantage, specially when customers are time-sensitive or require specialized assistance to make a purchase. When demand for service is variable and staffing requirements cannot be adjusted quickly, capacity decisions require making a trade-off between responsiveness to customers versus controlling operating costs through worker utilization. This work extends recent empirical studies that estimate the impact of staffing levels and delivery performance on sales, combining detailed sales transaction and demand data with employee staffing data, and developing an econometric approach to model customer behavior in the context of an out-bound call center, an online apparel retailer and brick and mortar stores. Model parameters estimation is challenging due to endogeneity issues. This work develops decision support tools to manage staffing levels and schedule working shifts, balancing the sales contribution of employees with salary costs, accounting for the causal relation between personnel and sales.

Chapter 1 focuses on providing insights on how to manage the capacity of flexible agents in the context of a service platform where customers are sensitive to waiting time. Analyzing this problem requires to study two empirical questions: (i) what is the impact of waiting time and agent productivity on revenues? (ii) what is the impact of workload on the attrition of agents? The answers to these empirical questions can be used to evaluate how changes in the agent workload ultimately affect revenues and thereby choose an optimal level of capacity utilization for the service platform. To study this research question and demonstrate its applicability to a real problem, we partnered with a company that operates an online market platform to facilitate the search and purchasing process of insurance products. The platform operates an outbound call center to provide customer support to their purchase decision, which has a significant impact on the conversion of quotes.

Chapter 2 combines empirical analysis with optimization methods to build a decision support tool that can be used by managers to plan labor allocation and schedule working shifts to maximize store profitability, balancing gross margins with labor costs. Operations Research literature has a long history in developing models to optimize shift schedules in service delivery systems using tools from mathematical programming. However, these models typically take the labor requirements as a fixed input, finding a feasible schedule that achieves these requirements at minimum costs. In our work, the desired labor requirements is one of the key decisions, which accounts for the trade-off between costs and sales. The methodology

is separated into three steps. First, an econometric model is developed to estimate the effect of labor on sales in an hourly basis. This model provides an input to an optimization problem that seeks to optimize the hour-by-hour labor requirement based on traffic projections. While useful, this second step provides an ideal labor requirement that cannot be implemented in practice as it is unlikely to comply with labor regulatory laws, company hiring policies and employees' preferences. The third step consists in translating this ideal labor plan into a feasible working schedule, achieved through a mathematical program. These three steps have been developed into a prototype that is currently been integrated into a decision support tool that complements the current services provided by our industry collaborator to its clients.

Chapter 3 studies the relation between order fulfillment, specifically delivery speed, and future customer behavior using one year of comprehensive transactional data from a large online apparel retailer. Literature relies on the theoretical assumption that higher service quality increases demand and consumer satisfaction, though existing evidence is largely anecdotal rather than quantitative. We quantify the impact of delivery quality over time on purchase incidence and amount of future purchases.

Chapter 1

Online Service Platform

1.1 Introduction

In many service industries, including the financial sector, restaurants, transportation, retail and call centers, two major drivers of service quality are the support provided by experienced employees to customers and the speed of the service. In terms of the latter, many of these service systems exhibit variable demand and relatively inflexible capacity — under these circumstances, making capacity decisions involves balancing operating costs with responsiveness to customers. To provide prompt service to unpredictable requests of time-sensitive customers, the service system has to operate at low levels of capacity utilization, keeping some idleness in the service capacity in order to accommodate unexpected peaks in the demand. In services that are labor intensive with fixed wages, low utilization implies higher labor costs, generating a trade-off between waiting times and operational costs.

The 'gig' economy has enabled an emerging business model — the on-demand service platform (Taylor [2018], Allon et al. [2018]) — that seeks to break this trade-off. Flexibility is achieved by accessing a large pool of employees (service agents) that are willing to be compensated with piece-rate wages (per customer served). Examples of this business model include taxi-style transportation (e.g. Uber, Lyft), food delivery (GrubHub, Seamless, InstaCart), customer contact centers (Liveops) and home services (Handy), among other industries.[1] The key feature of these service platforms is to have access to a large crowdsourced pool of agents to meet spikes in demand, while managing labor operating costs using variable pay instead of fixed wages. Past research has studied the role of financial incentives in self-scheduling labor platforms to better align demand and supply (Cachon et al. [2017], Allon et al. [2018]). In this work, we examine the long term implications of increasing or reducing the workload of agents in terms of the selection and retention of the flexible workforce.

Online service platforms can operate with low levels of agent utilization in order to maximize the responsiveness to customer requests. However, agent idleness can potentially hurt performance through agent retention. For services that require higher levels of employee

[1] The aforementioned examples are business operating in the United States, but similar business models can be found in other countries.

training, exhibit learning-by-doing or have significant heterogeneity in agent productivity, retention of highly-skilled/experienced employees is a critical factor affecting operational performance. In order to reduce agent attrition, the service platform should also consider agent satisfaction, which can be negatively affected by idleness during working hours, specially so when wages are based on piece-rates. Therefore, in choosing service capacity, the service platform has to balance customer waiting times with agent utilization.

The main goal of this research is to provide insights on how to manage the capacity of flexible agents in the context of a service platform where customers are sensitive to waiting time. Analyzing this problem requires us to study two empirical questions: (i) what is the impact of waiting time and agent productivity on revenues? (ii) what is the impact of workload on the attrition of agents? The answers to these empirical questions can be used to evaluate how changes in the agent workload ultimately affect revenues and thereby choose an optimal level of capacity utilization for the service platform. To study this research question and demonstrate its applicability to a real problem, we partnered with a company that operates an online market platform to facilitate the search and purchasing process of insurance products (auto, travel, home, etc). The platform operates an outbound call center to provide customer support to their purchase decision, which has a significant impact on the conversion of quotes. This call center is operated with freelance agents who work remotely and receive variable pay based on the number of converted quotes. Customers visiting the platform's website and not making a purchase join a queue to be contacted by call center agents.

Overall, this research makes three main contributions. First, we empirically analyze an outbound call center, focusing on how service speed and workload influence customers and agents, respectively. Although the empirical analysis of inbound contact-centers is well established in the literature, we show that outbound call centers have important differences that require a different modeling approach. Specifically, outbound call centers need to decide when to call customers (since customers are not waiting in line to be served). We develop an empirical framework to incorporate this important feature. In our setting, we find a strong impact of waiting time on customer behavior: conversion rates are 50% greater if a customer is contacted within the first 5 minutes of visiting the platform's website rather than after an hour. Second, our empirical results show that there is a trade-off between contacting customers more rapidly and retaining high-performance agents. This trade-off arises due a positive effect of agent utilization on retention: a 10% increase in workload translates into a 25% percentage decrease in weekly agent attrition rate which yields a 33% increase in agent tenure. Finally, we show that combining our empirical results with a queuing model and an agent selection policy can be used to compare the profitability of different choices of capacity utilization. In the context of our application, we show that customer sensitivity to waiting time has a predominant effect, and in the margin the outbound call center would benefit by reducing capacity utilization with more agents, further exploiting the flexibility provided by online labor platforms – increasing revenues from 2.7% to 4.6% under medium traffic and from 0.9% to 3.2% under low traffic conditions. This "what-if" analysis can be used as a decision support tool to manage capacity in service platforms.

The rest of this chapter is organized as follows. Section 1.2 reviews the relevant literature and develops the hypotheses that will be tested. Section 1.3 describes the empirical context

and the data used for this study. Section 1.4 presents the econometric formulation of customer behavior. Section 1.5 presents the results of the econometric model of customer purchases. Section 1.6 formulates a model of agent attrition and presents the corresponding empirical results. Section 1.7 uses the results from the empirical analysis to evaluate the profitability of changes to agent utilization levels. Finally Section 1.8 concludes the chapter with a summary of the main results and a discussion of avenues for future research.

1.2 Literature Review and Hypothesis

The main research goal of this study is to characterize the impact of agent utilization on the performance of a service system where both customer waiting time and worker retention affect profitability. Figure 1.1 describes the different mechanisms through which agent utilization can affect the performance of the service platform under study, where the main measure of profitability is sales conversion. We elaborate on each of these mechanisms next, relating them to the existing literature, the context of the application and using them to motivate our main hypotheses.

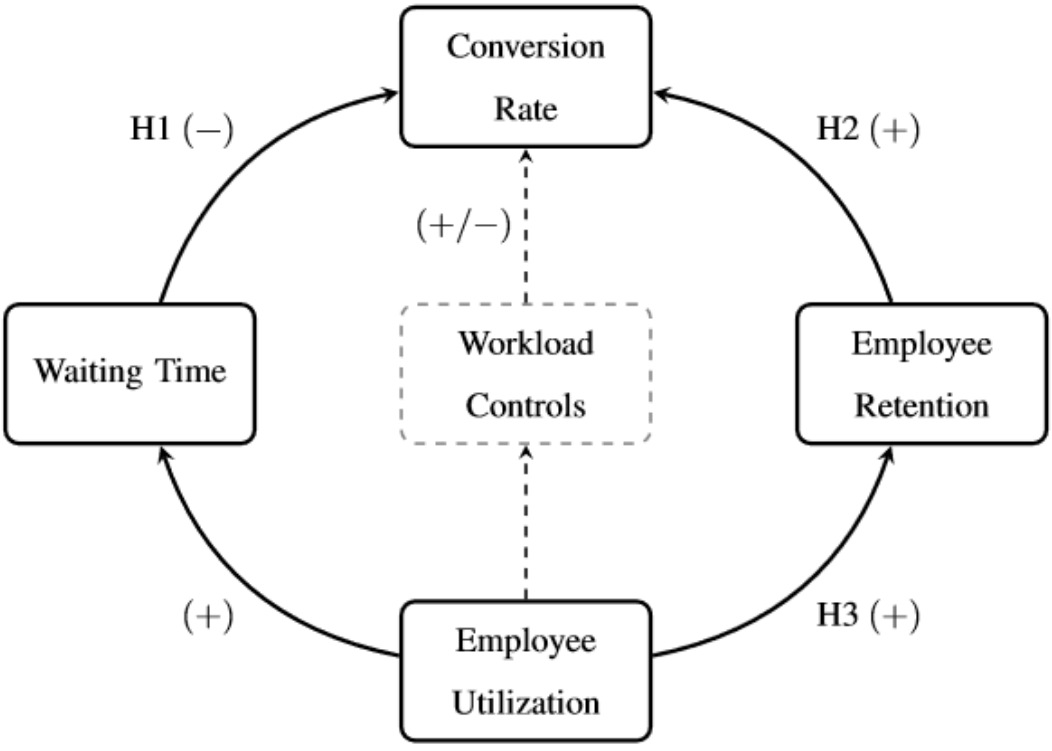

Figure 1.1 Hypothesis schema. Representation of the mechanisms through which employee utilization can affect the conversion rate. H1, H2 and H3 indicate the main hypothesis of this study and plus (+) and minus (-) signs indicate the direction of the anticipated effects.

There is a large body of literature in service operations that studies the relationship between capacity utilization and waiting time, with many applications to call centers (see Gans et al. [2003] and Aksin et al. [2007] for recent surveys). The variability of incoming calls together with the inflexibility of adjusting capacity generates congestion effects: at high

levels of agent utilization, waiting time increases dramatically even for small increases in workload (Cachon and Terwiesch [2009]). Queuing models and discrete event simulation capture this effect and can be used to measure the effect of capacity utilization on waiting time (Gross [2008]; Carson et al. [2005]).

Whereas the effect of utilization on waiting time can be studied using analytical models, the impact of waiting time on sales requires an empirical analysis of customer behavior, which has been studied in the fields of operations management, marketing and economics. Along this line, Lu et al. [2013] studies the effect of waiting time on customer purchases, in the context of a deli counter in a supermarket. Batt and Terwiesch [2015] studies the abandonment of incoming patients in an emergency room. Allon et al. [2011] study the effect of waiting time on demand and prices in the fast-food industry. Png and Reitman [1994] study how gasoline stations compete to attract demand by reducing waiting times. In all of these examples, the queue is visible to the customer, whereas in many call-centers (including the one in this study) the queue is not directly observable. Aksin et al. [2013] develops a dynamic structural model to capture customer abandonment based on an optimal stopping problem, where customers form expectations about waiting time without directly observing the queue. This work was extended by Akşin et al. [2016] and Yu et al. [2016] to incorporate delay announcements, which can directly affect customer expectations about waiting time. Although most empirical studies focus on inbound call centers, Hathaway et al. [2018] studies a setting where customers are offered a call-back option, which resembles an outbound call-center.

The aforementioned studies find that customers dislike waiting time and therefore longer waits lead to lower system performance. But there is also theory that predicts the opposite. Veeraraghavan and Debo [2009] presents an analytical model where uninformed customers can learn about the quality of the service through the length of the queue, inducing a herding effect that implies a positive association between waiting time and demand. Another example is in the case of purchases that require time for the customer to evaluate options and decide, in which case it may be preferable to wait before calling. Furthermore, in an out-bound call center it may not be desirable to call immediately after a failed attempt as the customer may be unavailable to answer the phone (Samuelson [1999]).

Hence, the effect of waiting time on purchasing behavior is, in theory, ambiguous. However, given the context of our application – sales of auto insurance – we hypothesize a *negative effect* of waiting time on conversion (H1 in Figure 1.1) because: (i) auto insurance is a standard product and customers are well informed about its characteristics, and therefore herding effects are unlikely to occur; (ii) insurance markets are competitive with low price dispersion, therefore time to respond to customers is important to generate conversion (De Treville and Van Ackere [2006]).

The middle part of Figure 1.1 illustrates a direct effect of workload on performance, which has been the focus of several recent studies in the operations management literature. This literature encompasses different mechanisms with an ambiguous prediction of the effect of workload on productivity and performance. Tan and Netessine [2014] empirically study restaurant waiters and show that they "fill-in" time when workload is low, decreasing their performance. KC and Terwiesch [2009] study the effect of employee workload in hospitals

and its impact on service quality, accounting for the effect of fatigue (a negative effect) and pressure to work faster (a positive effect). KC [2013] studies hospital emergency departments, looking at how workload affects multitasking of the doctors and measure how this ultimately affects service quality. Freeman et al. [2017] studies maternity units and finds that these units treat patients differently depending on their workload. Although these are all important factors to consider, they have been well documented in the literature and are not central to the main trade-off analyzed in this study. Nevertheless, the existing literature is useful to specify important control variables that need to be included in the econometric models described in Section 1.4.2.

Next, we discuss the hypothesis related to employee retention. We identify two mechanisms through which retention can affect sales conversion. First, employee experience can directly impact productivity. There is substantial work in manufacturing and service industries showing that learning curves are significant (Lapré and Van Wassenhove [1998]; Lapré et al. [2000]; Lapré [2011]; Lapré and Tsikriktsis [2006]; Ramdas et al. [2018]). Focusing on the sales function, Misra et al. [2004] empirically measures learning curves for salespeople in an office supply company. Other papers have documented the importance of considering different features of worker experience, such as volume and scope (Staats and Gino [2012]; Kc and Staats [2012]; Huckman et al. [2009]; Huckman and Pisano [2006]), and the worker interactions within a team (Huckman and Staats [2011]). In our setting, the task performed by agents is relatively standard and performed individually by each agent, so task variety and team structure are less important; moreover, work is performed remotely without any interaction among the agents.

Employee retention can affect sales through other mechanisms too, even when learning/ experience effects are not present. Service employees typically exhibit heterogeneity in their skills and productivity. Gans et al. [2010] studies productivity of agents in a call center, measured by call duration. This work shows significant heterogeneity across agents, both in the level of productivity as well as their learning curve (rate of improvement as they gain experience). Arlotto et al. [2013] develop an analytical model to optimize employee selection policies when agents are heterogeneous in their productivity. The selection policy needs to balance "exploration" – testing agents to infer their productivity – with "exploitation", achieved by allocating work to the more productive agents. In section 1.6 we show a simpler version of this model which shows that as the retention rate of agents increases, the selection policy become more effective at screening agents with higher conversion rates. Hence, in the presence of agent heterogeneity, increasing employee retention increases sales. Based on this existing literature, our second hypothesis (H2 in Figure 1.1) predicts a positive effect of worker experience on sales productivity.

We now turn to our third hypothesis (H3 in Figure 1.1) relating agent utilization with employee retention. In our application, agent's salary is proportional to converted calls, and therefore workload has a direct effect on their compensation. Sager et al. [1989] present a causal model of the turnover process for salespeople, which suggests that satisfaction with pay and tenure are negatively associated with salesperson turnover. Also related to our work is Emadi and Staats [2018], which develops a structural dynamic model to analyze factors that affect voluntary attrition in a business process management company. Their findings reveal that employee attrition is relatively insensitive to salary and depends instead on the

characteristics of their supervisors. In that study employees receive fixed salary with annual adjustments, whereas in our setting pay is variable and can change from week to week depending on worker utilization, and so the effect could be different. Griffeth et al. [2000] conducts a meta-analysis of related work analyzing employee attrition and finds that satisfaction with pay has a relative small effect size relative to other factors. However, the salary effect appears to be larger for performance-based compensation schemes, suggesting that variable payment schemes have a stronger effect on retention (relative to fixed payment).

Our research is also related to recent studies about flexible working schedules and how they impact labor supply in service settings. Mas and Pallais [2017] study a call-center to analyze how agents value different schedule arrangements, which is relevant to calculate labor supply. Chen et al. [2017] use data from a ridesharing company to analyze how agents (i.e. drivers) choose their working hours. In that setting, agents have full flexibility to decide when to work on the platform making a trade-off between the expected income and their reservation wage, which can vary across the day (i.e. some agents dislike driving in late hours whereas other agents with day jobs prefer driving at night). Although in our work the main focus is to study the impact of workload on agent attrition (which is not studied in the aforementioned studies), the trade-off is similar: workers choose to participate in the service platform based on the expected pay, which is directly related to the assigned workload. Hence, the work by Mas and Pallais [2017] and Chen et al. [2017] provide some support for the hypothesis that the level of workload should increase agent's participation and thereby reduce attrition, in settings where workers have flexibility in their working hours and receive variable payment. In summary, based on the literature and the context of application studied, hypothesis 3 (H3) predicts that increasing labor utilization leads to higher retention rates of agents.

1.3 Empirical Setting

The context of our application is an online market platform selling auto insurance plans to car owners. The service platform works with several insurance companies that provide insurance quotes. The platform acts as a broker that facilitates the comparison of alternative auto insurance plans that vary in terms of deductibles, premiums and other characteristics. Transparency is a key aspect of the business model of the platform, providing unbiased information to customers to assist them during the purchase process. To achieve this, revenues are acquired through a flat fee charged by the platform to the insurance company for each plan sold. This fee is constant across all insurance plans and, hence, it is independent of the premium of each plan. To further assist customers in their search process, the platform combines online information with personalized assistance over the phone, using an outbound call center with freelance agents that work remotely from home with flexible hours. This flexibility allows the company to adjust service capacity when facing variability in customer arrivals. The freelance agents are paid on commissions for each converted sale, which is also flat across plans and premiums.

Figure 1.2 illustrates the operation of the service platform. The process is initiated when the customer generates a quote online, entering vehicle characteristics (make, model, year) and demographics (age, address, phone number). This information is directed in real-time to insurance companies that provide one or more insurance plans with different characteristics (e.g. deductible, coverage and premiums), which are displayed online to the customer in a

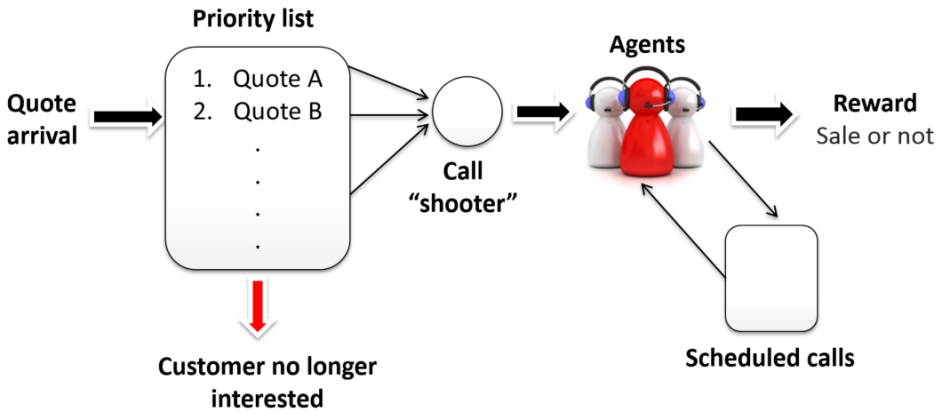

Figure 1.2 Operation of the service platform.

table format that facilitates comparison. Customers may decide to immediately purchase online choosing from one of the products available. When customers do not purchase within 3 minutes, the quote is added to a priority list of potential outbound calls.

Each quote in the list has a priority score which depends on a set of quote characteristics and the elapsed time since the quote was generated. The priority list is sorted based on this priority function and adjusted every minute. An automated dispatcher continuously monitors the activity of agents and decides how many customers in the top of the priority list to call. This decision can be based on the number of available agents and the priority scores of the quotes in the list. When the dispatcher calls and the customer answers, it gets automatically routed to one of the available agents. When the customer doesn't answer the call, the quote is returned to the priority list and its priority score is reduced.

After being assigned a call, the agent observes the list of quotes presented to the customer and provides assistance in the purchase process. The remaining sales process is entirely delegated to the agent, who may require additional outbound calls which are coordinated by the agent based on the customer's availability. These scheduled calls are considered as busy time of the agent where he cannot be assigned new outgoing calls.

We collected eight months of comprehensive data regarding each auto insurance quote including: the date and time when the quote was generated, the online acquisition channel where the customer reached the website (email promotion, organic/sponsored online search, etc), customer demographics (age, gender and residence area), car characteristics and the complete list of offers, including the insurance company, deductible, coverage, premiums and other characteristics (replacement car option, personal injury insurance, etc.). For each quote, we observe the number of calls, which may include failed call attempts (where the customer couldn't be reached), the first answered call and subsequent scheduled calls. Upon the first contact, the quote is assigned to an agent who handles subsequent calls. Agent information includes age, gender and the job starting date. For each answered call, we obtain the corresponding timestamp, duration and outcome of the call (purchase, no-purchase or whether a new call was scheduled).

A key descriptive statistic in this study is customer waiting time as measured by the time

Call attempt	Percentiles of elapsed time [minutes] 25%	50%	70%	90%	Contact rate
1st	3.7	5.7	125.3	2,636.0	56.3%
2nd	5.5	61.9	182.1	1,593.0	36.4%
3rd	2.2	68.2	1,003.1	2,853.7	30.5%

Table 1.1 Descriptive statistics. Elapsed times [minutes] since the generation of the quote until the first call attempt and between calls for the second and third attempt, showing the percentiles of the distribution of elapsed time and the contact rate.

elapsed between the instant the quote is generated, as a result of a consumer search at the platform's website, until the customer is successfully contacted by an agent. Three minutes after a quote was generated, the quote enters the priority list to schedule the first call attempt (a few calls were attempted before three minutes but were excluded from the sample). Table 1.1 shows descriptive statistics of the first three call attempts, showing the percentiles of the distribution of the elapsed time and the contact rate. The elapsed time of the first call is highly skewed: 50% of the customers contacted in less than 5.7 mins, but the 70% percentile reaches more than two hours. The contact rate for the first call (i.e., the fraction of first attempts that are answered by a customer) is 56%. If the first attempt fails, a second call is made, with a median of 62 mins from the time the first call was made. This time is also skewed and a histogram shows the distribution to be bi-modal, where some customers are called soon after the first call and others about an hour later. The contact rate of the second call is lower, around 36%. The distribution of the third call follows a similar pattern as the second, but the upper tail is thicker because a higher fraction of these calls are made the following day.

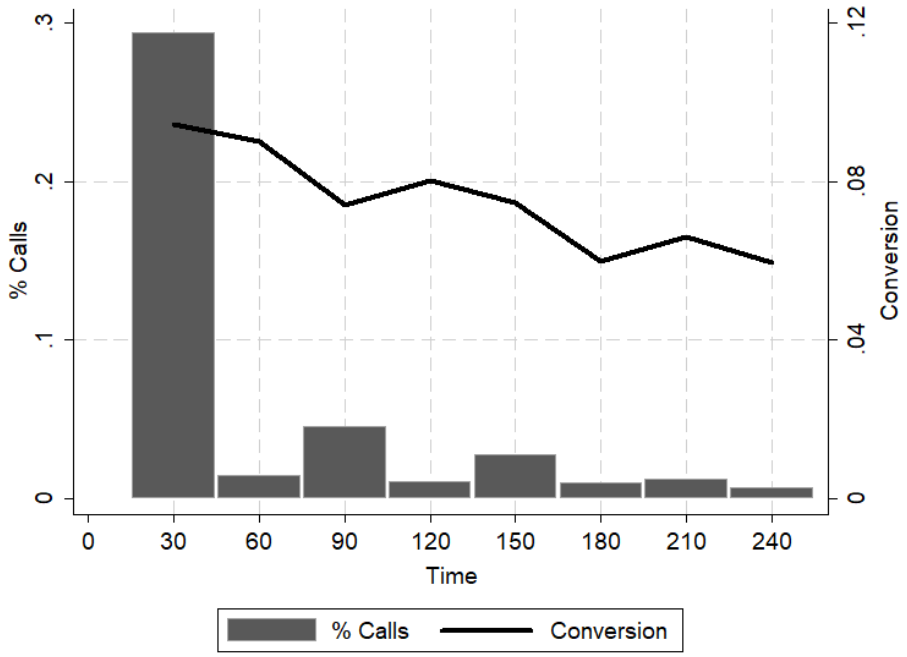

Figure 1.3 Histogram of customer waiting time. Showing time between first contact and quote and conversion rate.

One of the main empirical questions in this study relates to the causal effect of contact times on conversion. To visualize this relationship, Figure 1.3 shows a histogram of the elapsed time to contact the customer (bars) and the average conversion rate (solid line) as a function of this elapsed time. The data shows that conversion peaks within the first hour and then declines rapidly as waiting time increases. This could be interpreted as a negative effect of contact time on conversion, but there are several confounding factors that preclude us from drawing this conclusion. First, the dispatcher prioritizes calls in order to contact first those customers that are more likely to purchase. Second, a contact may be delayed because the customer doesn't answer the calls. Table 1.1 shows that the contact rate goes down as the number of failed called attempts increases, suggesting a possible selection effect where uninterested customers remained unreached after several calls. Hence, the negative association between contact time and conversion may be driven by prioritization of calls or a skimming process where interested customers are more likely to answer a call attempt. Accounting for these confounding factors –which are applicable to most out-bound call centers– is the main empirical challenge to test hypothesis 1; the econometric model developed in the next section provides an empirical strategy to identify the causal effect of contact time on conversion.

1.4 Econometric Modeling of Customer Purchases

Recall that for the outbound call center to succeed at selling an insurance policy to a customer, the call center needs to be able to reach the customer and then the interaction between the customer and the agent has to result in an insurance policy purchase. Hence, the revenue generated from a quote in the service platform depends on two aspects of customer behavior: (i) the contact rate, defined as a customer's probability of answering a call attempt; (ii) the conversion rate, defined as the probability of purchasing conditional on answering the call. We model each of these probabilities separately. Consider w_{it} as the elapsed time since quote i was generated until the time of call attempt t ($t = 1$ is the first call attempt, $t = 2$ is the second, and so on). Define A_{it} as an indicator of whether call attempt t is answered. Whenever $A_{it} = 1$, the binary r.v. Y_{it} is an indicator of whether the customer purchased after being contacted. Normalizing the service fee minus the bonus paid to the agent to one, the expected revenue associated to call attempt t for quote i can be expressed as:

$$R(w_{it}) = \Pr(A_{it} = 1|w_{it}) \times \Pr(Y_{it} = 1|w_{it}, A_{it} = 1) \tag{1.1}$$

Two separate econometric models are developed to estimate the effect of the elapsed time w_{it} on the probability functions of A_{it} and Y_{it}. For both models, an important challenge arises because the elapsed time w_{it} is endogenous. Below we describe in detail the specification of these two models and how to handle endogeneity to identify the effect of elapsed time.

1.4.1 Customer Contact Probability

Consider the first call attempt for a new quote, with $t = 1$. We seek first to estimate how the probability of answering the call is affected by the elapsed time since the quote was generated, w_{it}, and other factors. The probability of making contact is parametrized through a random

utility model, where a customer's utility U_{it}, a construct that captures the customer's interest in the quote and availability to respond to the call, is represented by:

$$U_{it} = f(w_{i1}; \alpha_t) + \beta_t X_{it} + \varepsilon_{it} \tag{1.2}$$

where $f(w_{it}, \gamma)$ is a parametric function that captures the effect of the elapsed time between the quote was generated and the time the call attempt is made. X_{it} is a vector of observed quote characteristics and ε_{it} is a set of unobserved characteristics. The probability of making contact is specified by $\Pr(A_i = 1|X_{it}, w_{i1}) = \Pr(U_i > 0)$, which becomes a parametric function for specific choices of the distribution of the unobservable error ε_{it}. Hereon, we use a binary logit model to estimate this random utility model (with ε_{it} following a double exponential distribution).

Since the dispatcher uses a priority rule to schedule calls, customers with a higher probability of answering will have a shorter elapsed time for the first call attempt. Hence, it is necessary to include controls in X_{it} that capture the relevant variables considered by the dispatcher in prioritizing calls, to avoid an omitted variable bias. Detailed information was collected about the priority rules used by the platform during the study period. The priority rule used considers a score that is based on the acquisition channel of the customer (via email, organic search, direct URL or others), where email has the lowest score and organic search/direct URL has the highest. To control for this, we focused only on high priority quotes, acquired through organic search, all of which have the same priority for the first call attempt. If the first call is not answered by the customer, the priority score is reduced. This is an important factor that is controlled in the model by constructing separate models for each call attempt (so that we have customers with the same priority within the same model) and including all the relevant factors considered in those rules as control variables in the econometric specifications.

In some cases, a reduction to the score could be applied to customers that have visited the website before and has not answered call attempts in the past. The company did not record these data, and although this rule was apparently used infrequently, it can become an omitted variable that could potentially generate an endogeneity bias. Therefore, in addition to the set of controls used to account for priority, an additional identification strategy is used based on instrumental variables (IVs). Following the approach by Kim et al. [2014] and Dong et al. [2018], the proposed IVs are based on the level of congestion of the system. We calculated several metrics that capture the system utilization (e.g. the ratio of the number of new quotes divided by the total available agent hours), and found that higher utilization is indeed positively correlated with elapsed time w.

The estimation with IVs is implemented through a logistic regression with a control function. To construct the congestion instrument, the time period comprised by w_{it} is divided into one-hour length intervals. For each interval, a measure of congestion was calculated by taking the ratio of the number of labour hours staffed divided by the number of new quotes[2]. An average is taken across the time intervals to construct the instrumental variable Z_{it}. To implement the IV via a control function, the first step is a linear regression of w_{it} on X_{it} and

[2]For robustness, an alternative IV was calculated based on the number of call attempts that were made for other quotes besides the focal, and the results were similar.

Z_{it}. A function of the residual of this regression is added as an additional covariate to the logistic model (equation 1.2) to account for a potential correlation between ε_{it} and w_{it}[3].

For the IV to be exogenous, it must be unrelated to unobserved factors that affect a customer's interest in answering the call. It is therefore important to control for demand shocks that can increase traffic of customers interested in purchasing and thereby induce congestion. Part of these demand shocks are captured by the seasonal controls included in the model (dummies for month, day of the week and hour of the day). In addition, we obtained the number of Google searches related to auto insurance in the relevant markets where the service platform is operating. Recall that the sample focuses on quotes that were acquired through online search, and as expected, the number of searches explains part of the residual variation in quote volume (after controlling for seasonality).

Other controls in X_{it} include quote characteristics such as the car model and year, the age and gender of the customer and seasonality factors (days of the week, hour of the day). The 22 most popular vehicle models after model-year 2007 accounted for 75% of the quotes, and the sample was focused on this set of quotes.

Insurance premiums are also an important factor affecting purchases. However, price is also endogenous because insurance companies set premiums based on customer risk profiles calculated from proprietary data not available to the platform. Riskier customers are offered higher premiums, but their outside options are also likely to be different. Estimating this price elasticity requires an identification strategy that accounts for the endogeneity of premiums. One option is to control for customer risk, which in part is captured by the customer demographics included in the model. However, insurance companies have much more detailed data on customers, including their past claim settlements, which are not available to the service platform. Phillips et al. [2015] face a similar problem analyzing customer loans in banking, and resolve it using Instrumental Variables for price. They use a Hausman-type instrument, using the prices offered to customers in other regions. In our context, this type of instruments are weak and may actually exacerbate the endogeneity bias (Hahn et al. [2004]). For this reason, we opted to exclude price from the main model, but provide some robustness tests (reported in section 1.5) suggesting that omitting price from the model does not introduce bias in the estimation of the coefficients of interest (waiting time and agent experience).

To model subsequent call attempts ($t = 2, \ldots$), a similar approach was used but including additional covariates to better capture the effect of elapsed time. It is plausible that, in addition to the elapsed time w_{it} from the generation of the quote, the time from the previous failed attempt also impacts the probability of answering a call. Calling immediately after a failed called attempt may result in a low probability of answering, because the customer may be unavailable to answer the phone. Hence, for call attempts after the first one, we include an additional set of covariates capturing the time elapsed from the previous call attempt.

[3]The main results use a quadratic polynomial of the residual to specify the control function. Other specifications yield similar results.

1.4.2 Conversion After Contact

The second term of the revenue equation (1.1), $\Pr(Y_\text{i} = 1|w_{\text{i}t}, A_{\text{i}t} = 1)$, relates to the conversion of a contacted customer. For modeling purposes, a quote is considered as converted if the agent finishes the sale of the assigned quote anytime after the first contact is made (which may be in the first contact call or subsequent calls scheduled by the agent).

In contrast to the model of contact probability, which only considers characteristics of the quote and elapsed time, conversion is also affected by factors related to the agent assigned to the quote, which include: (i) the idiosyncratic characteristics of the agent; (ii) the agent's experience; (iii) the workload of the agent. The first two are directly related to H2 (see Figure 1.1), whereas the third defines important controls that need to be included in the model.

The empirical model of the purchase probability $\Pr(Y_\text{i} = 1|w_{\text{i}t}, A_{\text{i}t} = 1)$ is a binary logit choice model with random utility:

$$V_\text{i} = f_W(w_\text{i}^c; \gamma_W) + f_K(K_\text{i}; \gamma_K) + f_L(L_\text{i}; \gamma_L) + \beta X_\text{i} + \delta_{j(\text{i})} + \mu_\text{i} \tag{1.3}$$

where f_W, f_K and f_L are parametric functions of the elapsed time between the quote generation and the first contact (w_i^c), the experience of the agent assigned to the quote (K_i) and the workload handled by the agent at the time of first contact with the customer (L_i). Both K_i and L_i, described in detail later in this section, can vary across quotes assigned to the same agent. To control for time-invariant characteristics, an agent fixed effect $\delta_{j(\text{i})}$ is also included in the model (where j(i) indexes the agent assigned to quote i). The set of control variables X_i are similar to those included in model (1.2), including characteristics of the customer and vehicle, the relevant variables that are used by the service platform to prioritize calls, and the controls for demand shocks. The term μ_i is an error term representing all other unobserved factors that affect the purchase probability. As with the customer contact model, we opted to exclude insurance premiums from the covariates; Section 1.5 presents a robustness analysis to show that this omitted variable does not bias our estimates.

Similar to the contact model (1.2), the contact time w_i^c in model (1.3) is endogenous because it is affected by the priority rules used to schedule calls. This endogeneity problem can be addressed using the instrumental variables previously discussed. The identification comes from comparing identical customers (with the same X_i) that may have different contact time depending on the level of congestion of the service at the time their quotes were generated. If the level of congestion is unrelated to μ_i, then congestion is a valid instrument for the contact time w_i^c.

Agent experience K_i is measured as the number of weeks the agent has been working in the service platform at the time the quote is assigned. An alternative measure would be the number of quotes handled by the agent, but because our dataset comprises only 6 months, we cannot observe the volume of quotes for agents with longer tenure. The specification includes a linear and quadratic term of K_i that can incorporate non-linear and non-monotone effects. The workload of the agent, L_i, is measured by the number of new quotes assigned to the agent previously during the same day of the focal quote. Note that the quote is assigned at random to any of the idle agents, so that both K_i and L_i should be unrelated to the unobservables μ_i, thereby providing an ideal setting for the identification of these effects (Aksin et al. [2015]).

We end this section providing further details of the specification of the customer contact model (1.2). Specifically, the specification of $f(w_{it}, \gamma)$ was implemented with a set of dummy variables capturing the level of w_{it}, in order to allow for a flexible estimation that allows for non-linearities. The thresholds for the levels was based on the actual distribution of the elapsed time w_{it}, which varied depending on the call attempt. For the first call ($t = 1$), w_{i1} starts 3 minutes after the call is received, with about 32% in the 3-5 minutes range, and 8% in the 5-10 minutes interval. More than half of the quotes receive their first call within the first two hours after the quote is generated. Hence, the thresholds were placed at minutes 4, 5, 6, 10, 20, 40, 60, 80 and 120. This specification puts special focus on measuring time-sensitivity in the first minutes after the quote, which is where most of the first call attempts are concentrated. Other specifications were also considered, yielding similar results.

If the first call attempt fails to make contact, about 30% of the second call attempts are made within 15 minutes after the first call. About 15%, are done within 60 and 80 minutes after the first call. The time between the 2^{nd} and 3^{rd} call follows a similar pattern. This variability together with the variability of the first call attempt generates more dispersion in the elapsed time of the 2^{nd} call attempt. Similar thresholds as with the first call were used to set the levels of the elapsed time of the second call attempt. An additional set of dummy variables were used to capture the elapsed time between the 1^{st} and 2^{nd} call, with threshold levels set every 30 minute intervals.

1.5 Estimation Results

The complete set of results of each of the contact models (1.2) can be found in Tables 1.5 and 1.6 (Appendix A). All of the models were estimated with IVs for the elapsed contact time, implemented via a control function. Figure 1.4 shows the estimated effect of the elapsed time for the first call. The estimates suggest a strong time-sensitivity within the first 20 minutes after the quote was generated: the probability of answering the call drops from 62% to about 52%, i.e. a 17% reduction in the contact rate. The estimates have high precision in the first 10 minute interval (where most of the calls are concentrated), and become less precise after 20 minutes. The estimates also suggest a stable pattern after the first 20 minutes.

For the second call attempt, the estimates suggest that both the elapsed time from quote generation and the previous call affect the contact probability. To visualize the overall effect, consider the case where the first failed call attempt was made within the 10 minutes of the quote generation. Figure 1.5 shows the contact probability of the second call as a function of the elapsed time from the quote generation (which starts at 10 minutes, after the first call attempt). The estimates suggest that the probability of answering a call following shortly after the first one has a low probability of success: the probability of answering increases sharply 40 minutes following the first call and then becomes relatively stable, at around 45%. Hence, it seems reasonable to wait at least 30 minutes before making the second call. The results for the third call attempt (not reported) are similar to those obtained in the second call.

The results for the conversion model (1.3) are described next. To capture non-linearities, the impact of contact time was modeled through a non-linear function specified in levels. Several models were estimated varying the thresholds that determine the levels, and using

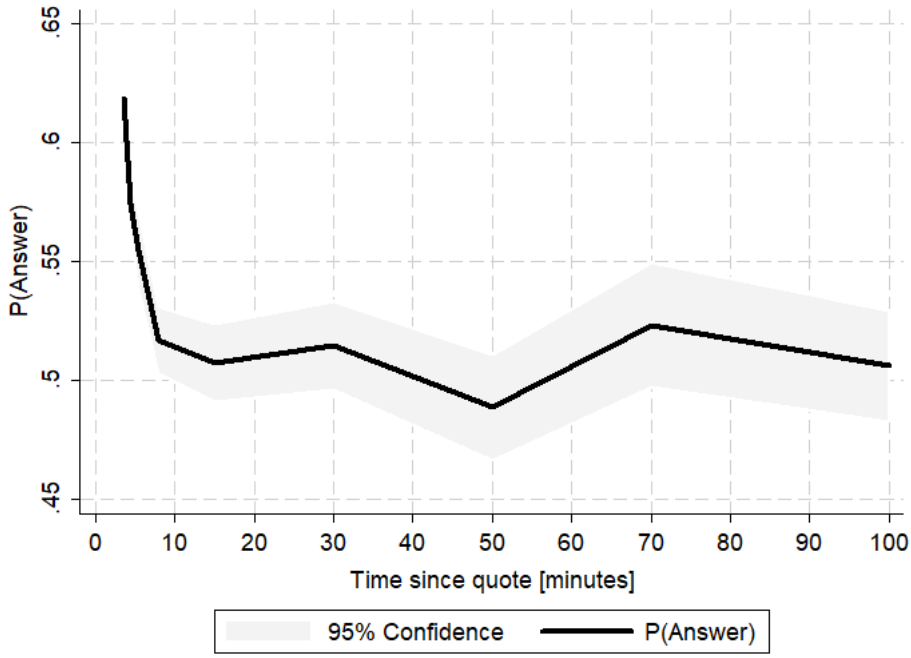

Figure 1.4 Effect of elapsed time on the probability of answering the first call attempt. The gray area shows the 95% confidence interval.

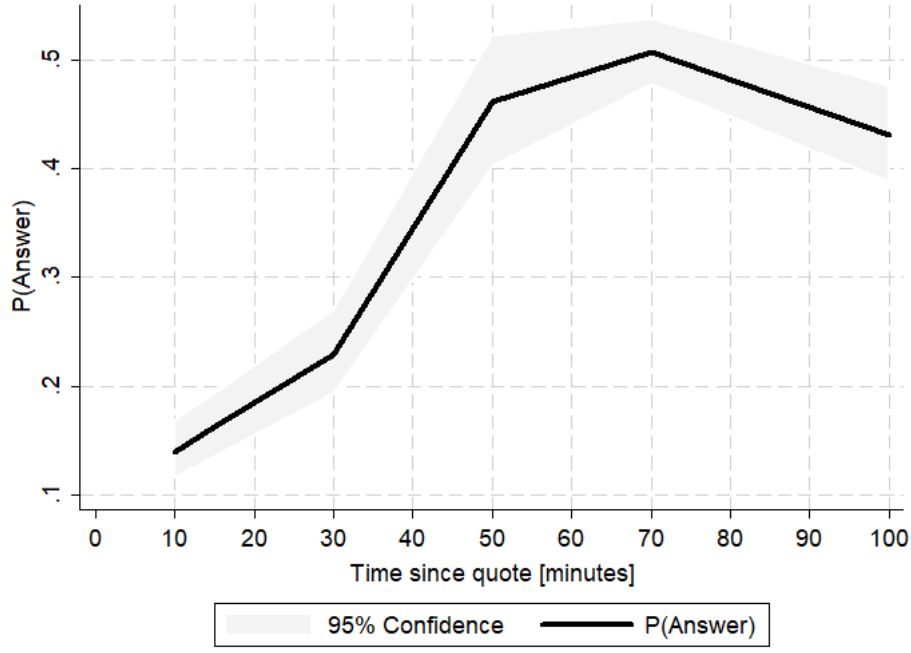

Figure 1.5 Effect of elapsed time on contact probability for the second call attempt. The first call was made withing the first 10 minutes since the quote was generated.

more granularity in the first 30 minutes where most of the first calls are concentrated.

Tables 1.7 and 1.8 in the Appendix A describe the details of the estimation, and Figures 1.6 and 1.7 illustrate the impact of contact time. The estimates of the contact time illustrated

in Figure 1.6 suggest that conversion is highly time sensitive, dropping from 6% to 4% during the first 90 minutes. The effect of contact time appears to be weaker after 70 minutes. The analysis with alternative thresholds to specify the levels of contact time yields similar results.

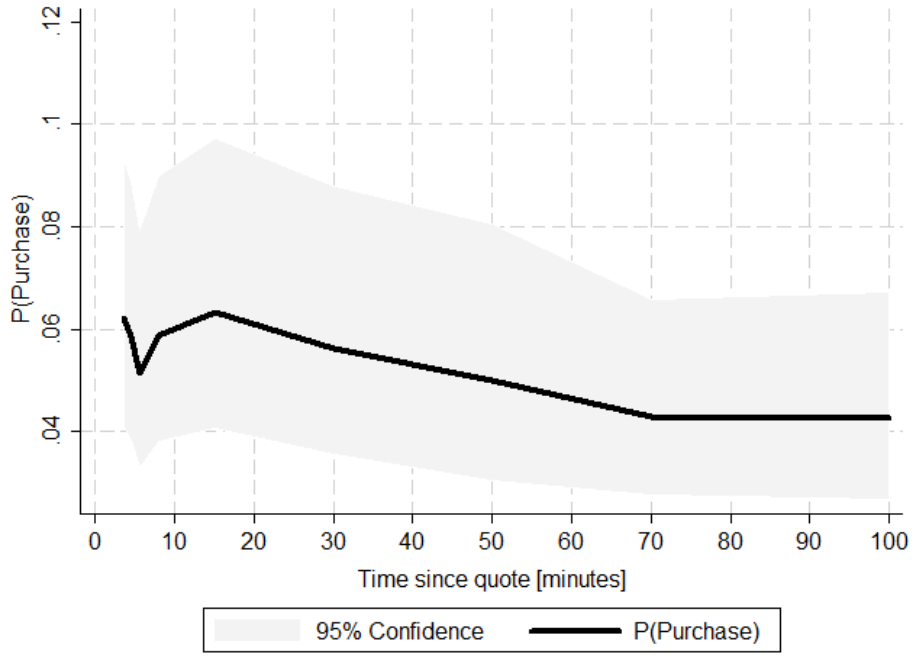

Figure 1.6 Effect of contact time on conversion.

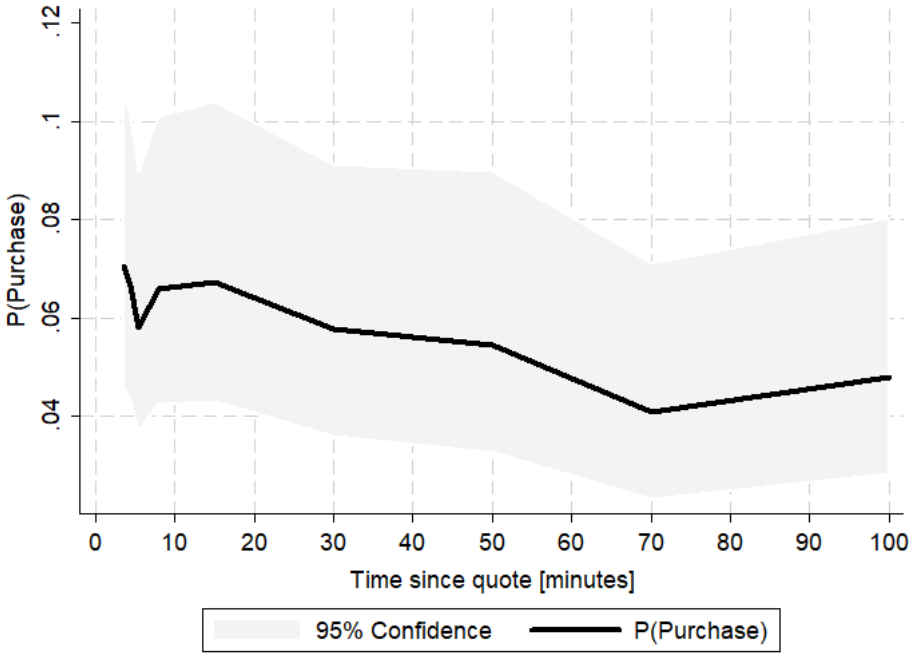

Figure 1.7 Effect of contact time on conversion
Clients contacted in the first call.

The estimates of Figure 1.6 include customers that were called either one, two or three times until the first contact. These three groups of customers may have different preferences,

which could lead to an endogenous sample selection: customers that did not answer the first call constitute a higher proportion of the sample of customers with higher contact time. The lower conversion rate observed at higher levels of contact time may be due to this sample selection mechanism and not through a causal effect of waiting time on conversion.

Our instrumental variable identification strategy should address this concern, because identification comes from variation in contact time due to congestion, thereby removing the sample selection mechanism (under the assumption that congestion levels are not related to customer's propensity to purchase). Nevertheless, to further validate our results, a separate model was estimated using only the sample of customers that answered the first call attempt (using similar instrumental variables to control for endogeneity). The estimates of the effect of contact time (which in this case coincides with the time of the first call attempt) are reported in Figure 1.7. The estimates look similar to those obtained with the complete customer population. The standard errors are higher for the estimated effect beyond 60 minutes after the quote is generated, which is expected because most of the first call attempts are done within the first hour. Overall, the main results regarding customer sensitivity to waiting time appear to be robust to potential sample selection bias.

Finally, the estimates of the contact model and conversion models were combined to measure how waiting time affects the expected revenue generated by a call (see equation (1.1)). We focus on the first call attempt because the empirical results suggest that: (i) customer sensitivity to answer the call is high within the first 50 minutes of the call (Figure 1.4); (ii) the probability of answering a second call after a failed first attempt is low within the first hour (Figure 1.5); (iii) conversion rates tend to be stable when customers are contacted after the first hour (Figure 1.6). Figure 1.8 illustrates the expected revenue function obtained by multiplying the contact model of Figure 1.4 with the conversion model of Figure 1.7 (which is estimated only for customers that answer the first call). These results suggest high customer sensitivity to waiting time, specially within the first 10 minutes of the call. After 70 minutes, the expected revenue is insensitive to waiting time.

We now consider Hypothesis 2, which requires us to evaluate the effect of agent related factors on conversion rate. The results reveal that agent experience does not have a statistically significant effect on conversion. That is, as an agent increases the length of time working for the platform, productivity as measured by the conversion of quotes remains unchanged. In contrast, we do find heterogeneity across employees, as reflected by the variation in the agent fixed effects estimated in the model. There is a moderate degree of heterogeneity in conversion rates across agents (captured by the estimated fixed effects), as shown in Table 1.2. The average agent achieves a conversion rate of 8.6%, and the corresponding standard variation yields a coefficient of variation of the conversion of rate of 0.1.

2.5%	Median	97.5%	Mean	Sd	CV
0.080	0.085	0.11	0.086	0.009	0.10

Table 1.2 Conversion rate distribution across agents. Computed from the estimated fixed effects in the conversion model. Values were computed under average conditions of the model covariates.

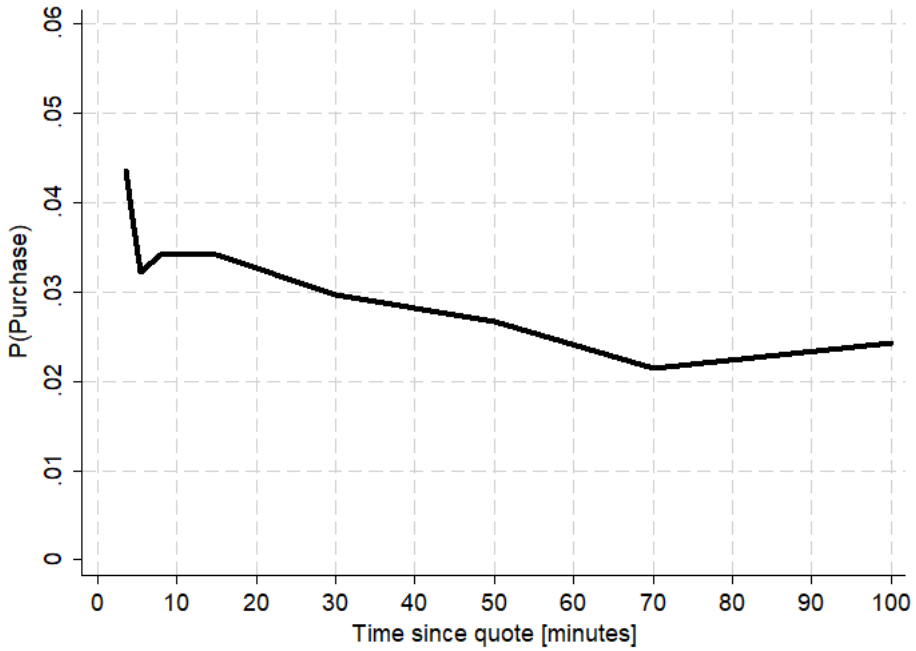

Figure 1.8 Effect of contact time on expected revenue. Given by Pr(Contact) × Pr(Conversion|Contact).

1.5.1 Discussion and Robustness of The Results of Customer Behavior Models

In summary, the main conclusions of the customer behavior analysis are: (i) customers are highly sensitive to waiting time within the first hour after a quote is generated; (ii) this sensitivity is both due to a a decline in their probability of answering the phone and a lower conversion rate after making contact; (iii) experience –measured by agent tenure– doesn't affect the conversion rate of the agent; (iv) there is moderate heterogeneity in conversion rates across agents. We now discuss alternative specifications to assess the robustness of these conclusions, particularly with respect to the inclusion of prices in these models.

Price is an important factor that can potentially affect both the probability of answering a call and the conversion after making contact. If price promotions increase a customer's interest in an insurance policy offered through the platform, it is possible that these price promotions may increase the probability of answering a call, which translates into a positive correlation between prices and the error term ε_{it} in model (1.2). But because the platform doesn't prioritize calls based on prices, there is no association between prices and the elapsed time until the first call and subsequent attempts (w_{it}). Hence, omitting prices from the model should not bias the coefficients of interest in the contact model (1.2). Furthermore, because insurance companies use pricing strategies that account for customer risk, it is likely that prices are positively correlated with purchase intentions (because risky customers have fewer outside options). This would induce a positive association between price premiums and μ_{i} (the error term in equation (1.3)). Moreover, if price also affects the probability of answering calls —for example, if price promotions increase a customer's interest in the products offered by the platform— it will also be correlated with the contact time w^c. Under

these circumstances, omitting price from the model can lead to bias in the estimated effect of contact time w^c. However, the IV strategy helps to mitigate this potential bias. The level of congestion of the system is unrelated to price, because both customer traffic and staffing are not affected by prices in this setting. Hence, using congestion as an IV helps to correct for all of the potential sources of endogeneity bias that can affect the estimation of coefficient γ_1 associated with the contact time.

To further assess whether the exclusion of price from the model could affect the results, we ran additional specifications including prices as a covariate. More specifically, we calculated a metric of the attractiveness of the prices offered to a customer as follows. We first ran a hedonic regression of insurance policy prices as a function of policy and customer characteristics. We then used the residual as a proxy for the (un)attractiveness of the offer (the residual indicates the deviation of the price from the expected price given the product characteristics). Accordingly, the average and the minimum value of the residuals of the offered prices are included as a measure of price in models (1.2) and (1.3). In both cases, the results and conclusions obtained from the empirical analysis were similar to those discussed earlier in this section.

1.6 Modeling Agent Retention

This section describes an econometric model to test Hypothesis 3 (see Figure 1.1): the effect of agent utilization on retention. To model turnover, we use a duration model to estimate how agents' churn probability is affected by their workload, among other factors that are used as controls. A discrete duration model is specified, using weekly observations following each agent during our study period. Each agent k is observed for $t = 1 \dots T_k$, where $t = 1$ is the week the agents starts receiving calls in the platform and T_k is the week in which the agent leaves the platform. The binary dependent variable y_{kt} is equal to one when the agent quits and zero for all previous periods. The probability that the agent leaves on tenure week t is modeled as $\Pr[Y_{kt} = 1] = p(X_{kt}, \theta)$ (conditional on not leaving prior to t), which represents the hazard rate of the discrete time duration model. A parametric specification of the hazard rate based on a logistic regression is used. The covariates of this regression, X_{kt}, are discussed next.

To provide flexibility in the estimation of the hazard rate, the logistic regression model includes a linear and quadratic term of the agent experience –measured as the number of weeks working for the platform– in order to capture a time-varying baseline hazard rate of attrition. An additional set of covariates were included to capture the characteristics of the agent: (i) dummy variables capturing different levels of agent age in years ([18-35];[36,50] and 51 or older); (ii) a dummy for the agent's gender.

The main objective of this model is to measure the impact of workload on the attrition probability $p(X_{kt}, \theta)$. One possible approach to capture this effect is to use the number of converted calls, which directly affects the agent's income from working on the platform. There are two reasons why we do not include the number of converted calls as measure of workload. First, the platform cannot directly control the number of converted calls – this can only be affected indirectly by allocating more calls to an agent, which is the main decision variable that this study focuses on. Second, the conversion rate of an agent is endogenous: more

skilled and motivated agents determine both their retention and productivity (as measured by conversion). An agent's intrinsic skill and motivation is an omitted variable that confounds the causal effect of converted calls on agent retention.

For these reasons, instead of using converted calls, we opted to use the assigned workload to an agent to measure its effect on attrition. The workload assigned to an agent in a given week depends on: (i) the number of hours that the agent was available to attend calls; and (ii) the number of quotes that arrived during the working hours of the agent. Unfortunately, the data don't provide a precise measure of how many hours an agent was available to attend calls. However, using the time-stamps of the calls that were attended, we can infer the morning and afternoon shifts that an agent worked during the week (because agents always receive at least one call during a working shift). Using this information, a measure of weekly workload was computed using the number of calls attended per shift worked by the agent, *Calls per Shift*.

	Logit		Logit with RE	
	(1)	(2)	(3)	(4)
	1wk lag	2wk lag	1wk lag	2wk lag
Calls per Shift	-3.160^{***}	-2.081^{***}	-3.160^{***}	-2.081^{***}
	(0.552)	(0.552)	(0.555)	(0.556)
Observations	1,515	1,515	1,515	1,515

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 1.3 Estimation results of the of agent retention model. Coefficients represent the estimated elasticity of the attrition probability to changes in the assigned workload, measured by the Calls per Shift.

Table 1.3 shows the estimation results of four different specifications of the agent attrition model. All models include controls for age, gender and the agent's experience (not shown in the table for brevity). Columns (1) and (2) show the estimates of a logistic regression model with different lags to measure past workload. Column (1) uses the calls per shift on the most recent past week (weeks were the agent did not work were dropped from the sample). Column (2) uses the average calls per shift during the last 2 weeks (excluding weeks were the agent didn't work in the calculation). Columns (3) and (4) estimate similar specifications but including an agent random effect to capture heterogeneity (a fixed effect cannot be used because a single spell is observed for each agent).

To facilitate the interpretation of the results, Table 1.3 shows the estimated elasticity of Calls per Shift on the retention probability, for an average female agent of 40 years old and 10 months of experience. The elasticity – measured as the percent change in attrition probability – is negative and statistically significant in all specifications, and economically meaningful, in the range of -2 to -3 depending on the specification. The models with a one week lag tend to have a larger elasticity: a 10% increase in the calls assigned per shift lowers the probability of quitting by 30%, decreasing it from 2% to 1.4% in every period. The effect

is smaller when considering the last 2 weeks of assigned workload, with an elasticity of -2. Comparing the fit of the models (1) and (2), the Akaike and Bayesian Information criterion tend to favor the one-week lag model (AIC: 291.3 vs 313.8; BIC: 339.2 vs. 361.7). The estimates of the model including random effects (columns (3) and (4)) are nearly identical.

To evaluate the magnitude of the effect of workload on attrition and experience, we conducted the following counterfactual. Each agent i is retained for T_{i} weeks, which follows a geometric distribution with attrition probability p and expected lifetime $E(T_{\text{i}}) = 1/p$. For an average agent, $p = 0.02$ and $E(T_{\text{i}}) = 50$ weeks. Based on the estimates of Table 1.3, using the average elasticity from columns (1) and (2) of approximately -2.5, a 10% increase in the number of calls per shift drops the hazard rate by 25%. This in turn increases the expected lifetime of a randomly picked agent by 33.3%, which is economically significant. Note, however, that an arriving customer is more likely to encounter an agent with a longer lifetime: the expected lifetime of an agent as seen by an arriving customer is $E(T_{\text{i}}) + Var(T_{\text{i}})/E(T_{\text{i}})$, equal to $2/p - 1$ when T_{i} is geometric[4]. Consequently, an arriving customer is attended by an agent with expected lifetime of 100 weeks, which is twice the average lifetime of an agent. Repeating the above calculations reveals that the relative change in this expected lifetime due to a 10% increase in workload is 33.7%, similar to what is obtained for the expected lifetime of a randomly picked agent.

1.7 Balancing Customer Waiting Time and Agent Retention

The empirical results show that it is important to contact the platform visitors quickly to improve the chances that they will buy one of the insurance policies. This speed could be improved by increasing the number of call center agents. This would decrease their utilization and hence decrease the average time to contact customers. At the same time, a lighter workload, as shown in the previous section, may lead to greater employee attrition, which in turn may affect the platform's effectiveness at converting visitors into buyers. We study this trade-off by considering scenarios where we modify the call center capacity and for each of them determine expected revenues.

To implement this analysis, note that platform revenues depend on the time elapsed since a quote is generated until an attempt is made to contact a customer (w_{it}). Since this is a random variable which depends on the capacity of the system (i.e., the number of active agents), it is necessary to integrate its distribution when determining expected revenues.

We derive the distribution of contact time (w_{it}) by formulating an M/M/c queuing model. The arrival rate and capacity are estimated from the data. Focusing on peak hours (i.e., weekdays between 10 and 11 am), the platform makes 5.2 calls per minute and has an average of 45 agents. The service rate of the M/M/c model can be estimated by matching the observed elapsed times to the ones predicted by the model. Figure 1.9 shows that this model provides a reasonable approximation of the empirical elapsed times.

A change in capacity also affects the probability of retaining agents. Agents abandoning

[4]This result is known as the *inspection paradox* in renewal theory. See Kulkarni [2016], chapter 7.

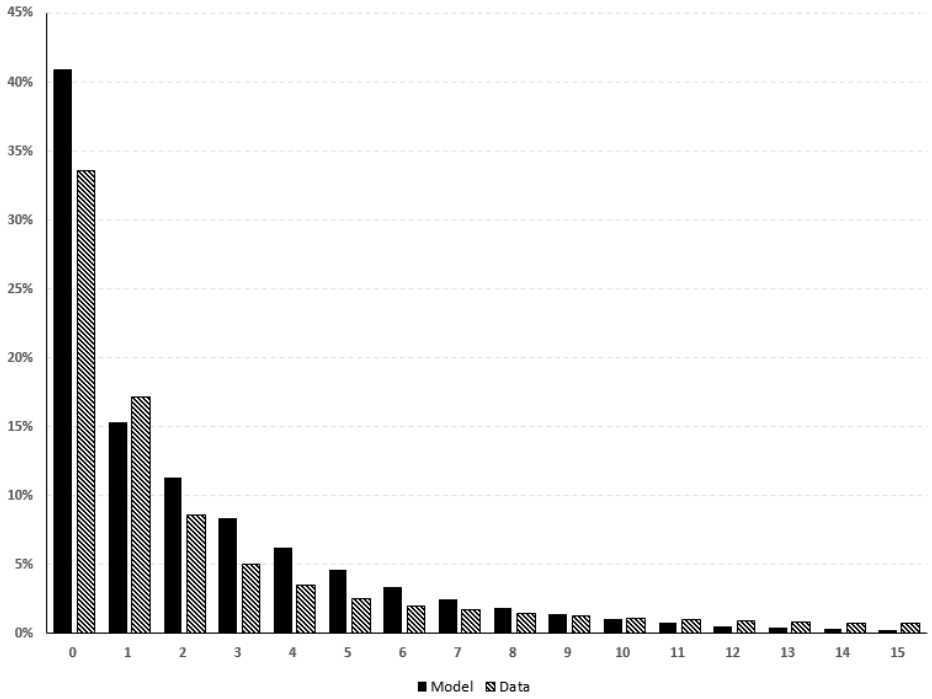

Figure 1.9 Waiting time distribution: M/M/c model vs data.

the platform will be replaced and to compute expected revenues, it is necessary to specify an agent recruiting and retention process. Assume that the service platform can hire agents from a pool of candidates. These candidates are heterogeneous in the average revenue $s \in \mathbb{R}$ they generate per contacted customer. This heterogeneity is characterized by the distribution $F(x) \equiv Prob(s < x)$, which can be estimated from the empirical results in Section 1.5. In our simulations we will use three alternative specifications of $F(x)$, all of them centered around the empirical mean, but with a variation coefficient of either 0.1, 0.3 or 0.5, where 0.1 is consistent with the degree of agent heterogeneity estimated from our data. This will be helpful to illustrate the impact of agent heterogeneity on our results.

Define a tenure process such that an agent hired by the service platform will be evaluated at the end of R periods. After those R periods, the call center learns the agent's average revenue s and if it falls below a threshold θ, then the agent's contract is not renewed and the call center hires a new agent with an average revenue per customer drawn from the distribution $F(\cdot)$, which will be learnt after R periods. Based on our empirical results, after $R = 8$ weeks the platform should be able to estimate an agent's conversion rate with a reasonable precision (i.e., with a standard error equal to 5% of the population mean of s).

Agents with average revenue below (above) θ will be denoted as low (high) quality agents. Therefore, the fraction of low quality agents in the population is given by $q \equiv F(\theta)$. In addition to this evaluation process, at the end of every period, an agent hired by the service platform may voluntarily quit with probability p. To characterize the revenues of the platform, we determine the fraction of customers that will be served by a high quality agent. Appendix B shows the derivation of this fraction and of the optimal retention threshold (θ) using results from renewal theory.

Given this assumptions, we begin by analyzing a medium traffic scenario with $\lambda = 2.5$ customers arriving per minute, a service rate of $\mu = 0.2$ calls per minute and $c = 14$ servers. Under these conditions, server utilization reaches 89% and on average customers are contacted 2 minutes after they visited the platform's website (see Table 1.4).

Consider next a change in capacity such that one of the servers is removed from the system: $c = 13$. As expected, agent utilization increases to 96%. This higher level of utilization has two consequences. On the one hand, it leads to longer elapsed times (8.4 minutes on average), which in turn decrease conversion by 10.2%. On the other hand, a higher utilization level leads to greater agent workload (+7.7%), which lowers employee attrition, increasing the number of periods that a high quality agent can be retained by the platform. Assuming a low level of heterogeneity among potential agents (i.e., $CV(\text{agent}) = 0.1$), this translates into a 0.9% revenue improvement. Higher levels of heterogeneity yield stronger improvements of 3.6%, partially compensating the 10.2% revenue decrease from longer contact times.

Similarly, consider a capacity increase such that one server is added to the system: $c = 15$. As a consequence, agent utilization decreases to 83%. As before, there are two implications of this lower utilization level: a 5.3% increase in conversion due to shorter contact times (0.8 minutes) and a 0.6%-2.5% decrease in conversion due to lower agent experience depending on the degree of heterogeneity among agents. Overall, under these medium traffic conditions, increasing capacity leads to a more favorable forecast than a capacity reduction.

It is also interesting to consider other scenarios. As shown in Table 1.4, under low traffic conditions, with $c = 11$ servers, a negative or positive change in capacity is associated with a smaller impact on waiting times, and hence a smaller impact on revenues (-7% and $+4\%$). In addition, the impact on agent workload is greater than under medium traffic conditions, yielding stronger impacts on revenue associated with agent retention.

	Medium Traffic		**Low Traffic**	
Rate and capacity	$\lambda = 2.5, \mu = 0.2, c = 14$		$\lambda = 1.8, \mu = 0.2, c = 11$	
Utilization & $\bar{W}$	89%, $2.0'$		82%, $1.1'$	
	↑ Workload	↓ Workload	↑ Workload	↓ Workload
Change in Capacity	$(c-1)$	$(c+1)$	$(c-1)$	$(c+1)$
Δ Workload	+7.7%	−6.7%	+10%	−8.3%
Effect of workload through Waiting Time				
Utilization	96%	83%	90%	75%
Waiting time $\bar{W}$ (avg)	$8.4'$	$0.8'$	$3.3'$	$0.4'$
Δ Exp Rev	−10.2%	+5.3%	−7%	+4%
Effect of workload through Retention				
Δ Exp Rev [CV(agent)=0.1]	+0.93%	−0.63%	+1.27%	−0.77%
Δ Exp Rev [CV(agent)=0.3]	+2.45%	−1.67%	+3.34%	−2.03%
Δ Exp Rev [CV(agent)=0.5]	+3.63%	−2.47%	+4.96%	−3.02%

Table 1.4 Balancing waiting time with employee retention.

1.8 Conclusions

In this study we consider the problem of managing utilization of workers for a service platform. We focus on settings with two key features: i) customer response is time sensitive and ii) the heterogeneity in skill of the agents providing this service is relevant. Under these conditions, the platform faces an important trade-off between increasing speed of service and providing a sufficiently attractive workload to its workers.

This research shows the importance of empirically measuring this trade-off when managing a labor platform. For the outbound call center under study, we find that customers are indeed sensitive to the time elapsed between their visit to the platform's website and when they are reached by a platform agent. In particular, the probability that a customer answers a phone call from the platform drops from 62% to 52%, as this elapsed time increases from 3 to 20 minutes. In addition, conditional on successfully reaching a customer, the probability that a customer will purchase one of the insurance policies drops from 6% to 4% as the elapsed time increases from 5 to 90 minutes.

Considering the role of agents, we find that heterogeneity in worker productivity is relevant, since a 7.7% increase in utilization translates into a 0.9% to 3.6% increase in conversion rate due to higher retention of more skilled agents under medium traffic conditions. Our agent attrition model results show that agent attrition depends on their assigned workload. For this platform, we find that a 10% increase in workload translates into a 25% percentage decrease in weekly agent attrition rate which yields a 33% increase in agent tenure. In conclusion, these results highlight the importance of considering and modeling the behavior of both customers and workers when making capacity and utilization decisions for service platforms.

Finally, there are interesting avenues for future research focusing on improving both sides of the labor platform, i.e. increasing agent retention and their conversion rate. In particular, future work could consider call routing policies that take into account agent retention and conversion. For example, regarding agent retention, a larger volume of calls could be assigned to agents that are at a higher risk of abandoning the platform. In terms of conversion rates, the platform could identify customers that are less sensitive to contact time (e.g., those that after one hour have still not been contacted) and assign them a lower priority. This is an important avenue for future research, since most of the relevant literature focuses on inbound call centers (e.g., Allon et al. [2011]). Overall, the complexity of labor platforms offers several interesting research opportunities and we hope this work may stimulate multidisciplinary research aimed at improving their performance and deepening our understanding of the gig economy.

Appendices

Appendix A: Detailed Estimation Results

	1^{st} Call IV	2^{nd} Call IV
Time elapsed since quote		
$\leq 4[m]$	0.37232***	−0.48448**
	(0.02440)	(0.17772)
$5[m]$	0.18904***	−1.87195***
	(0.02877)	(0.18788)
$6[m]$	0.10953**	−1.41776***
	(0.03779)	(0.18884)
$10[m]$	−0.04424	−0.80365***
	(0.03459)	(0.13309)
$20[m]$	−0.08290*	0.41247***
	(0.03844)	(0.10023)
$40[m]$	−0.05313	0.85035***
	(0.04222)	(0.09838)
$60[m]$	−0.15673**	0.22593
	(0.04938)	(0.12286)
$80[m]$	−0.01807	−0.03864
	(0.05649)	(0.04101)
$120[m]$	−0.09035	−0.12569*
	(0.05054)	(0.05741)
Observations	113,808	46,521

Standard errors in parentheses * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 1.5 Estimation results of the contact model, elapsed time since quote covariates. Table shows estimated parameters of the logit model defined by equation (1.2).

	1^{st} Call IV	2^{nd} Call IV
Time elapsed since previous call		
$\leq 30[m]$		$-2.05303***$
		(0.04524)
$30[m] - 60[m]$		$-0.37060***$
		(0.09096)
$60[m] - 90[m]$		$0.06625*$
		(0.03088)
$90[m] - 120[m]$		-0.06692
		(0.07026)
Price Index	$-0.09938***$	-0.07855
	(0.02460)	(0.04186)
Google Searches	0.00028	0.00063
	(0.00041)	(0.00072)
Res	$0.00001***$	$0.00001**$
	(1.89e−6)	(2.51e−6)
Res^2	−2.94e−11	−2.22e−11
	(1.52e−11)	(1.81e−11)
Constant	−1.15314	0.18806
	(1.16632)	(0.32958)
Observations	113,808	46,521

Standard errors in parentheses * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 1.6 Estimation results of the contact model, other covariates. Table shows estimated parameters of the logit model defined by equation (1.2).

	Customers contacted in 1^{st} call	Customers contacted in 1, 2 or 3 calls
$\leq 4[m]$	$0.394 * **$	$0.483 * **$
	(0.0513)	(0.0396)
$5[m]$	$0.330 * **$	$0.422 * **$
	(0.0623)	(0.0533)
$6[m]$	$0.186*$	$0.286 * **$
	(0.0865)	(0.0798)
$10[m]$	$0.323 * **$	$0.427 * **$
	(0.0781)	(0.0706)
$20[m]$	$0.345 * **$	$0.505 * **$
	(0.0944)	(0.0846)
$40[m]$	0.179	$0.380 * **$
	(0.120)	(0.103)
$60[m]$	0.121	0.252
	(0.160)	(0.137)
$80[m]$	-0.182	0.0908
	(0.198)	(0.0656)
$120[m]$	-0.0146	0.0895
	(0.170)	(0.105)
Observations	51,680	73,218

Standard errors in parentheses * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 1.7 Estimation results of the conversion model, elapsed time covariates. Table shows estimated parameters of the logit model defined by equation (1.3).

	Customers contacted in 1^{st} call	Customers contacted in 1, 2 or 3 calls
#quotes$\times \mathbb{1}_{<4month}$	-0.000209	$-0.000264*$
	(0.00013)	(0.000112)
#quotes$\times \mathbb{1}_{\geq 4month}$	-0.000118	-0.000097
	(0.00011)	(0.0000941)
Agent Experience	-0.000495	-0.000125
	(0.00741)	(0.00638)
Agent Experience Sq.	0.00000121	-0.000015
	(0.0000243)	(0.0000213)
Price Index	$-0.860 * **$	$-0.910 * **$
	(0.0675)	(0.0585)
Google Searches	0.00361*	0.00258*
	(0.00141)	(0.00122)
Res	$-0.0000478***$	$-0.0000343***$
	(0.00000854)	(0.00000528)
Res Sq.	3.27e$-$10***	2.50e$-$10***
	(6.27e $-$ 11)	(3.57e $-$ 11)
Constant	$-4.016 * **$	$-4.951 * **$
	(0.960)	(0.942)
Observations	51,680	73,218

Standard errors in parentheses * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 1.8 Estimation results of the conversion model, other covariates. Table shows estimated parameters of the logit model defined by equation (1.3).

Appendix B: Optimal Tenure Process

In this appendix we derive the optimal agent retention policy. We begin by characterizing the revenues of the platform, which depends on the fraction of customers that will be served by a high quality agent. Using results from renewal theory, define a cycle as a period of time that begins when a high quality agent is hired and ends when the next high quality agent is hired. Note that this cycle may include cases where:

- a high quality agent replaces another high quality agent.
- a high quality agent is consecutively replaced by several low quality agents until the next high quality agent is hired.

Denote the number of low quality agents in a cycle by N_L. Using the properties of a geometric distribution, its expected value corresponds to $\frac{q}{1-q}$. Denote by T the number of periods an agent would voluntarily wish to work, which follows a geometric distribution with parameter p. Using Theorem 3.4.4 in Ross et al. [1996], the long run probability of observing a high quality agent within a cycle is given by:

$$\begin{aligned}
\pi_H \equiv \lim_{t\to\infty} \text{Prob[H in time t]} &= \frac{E(T)}{E(T) + \frac{q}{1-q}E(\min\{T,R\})} \\
&= \frac{E(T)(1-q)}{(1-q)E(T) + qE(\min\{T,R\})} \\
&= \frac{E(T)(1-q)}{E(T) - q(E(T) - E(\min\{T,R\})} \\
&= \frac{E(T)(1-q)}{E(T) - qE(\max\{T-R,0\})}
\end{aligned}$$

Using once again the properties of the geometric distribution, $E(\max\{T-R,0\}) = (1-p)^R E(T)$. Hence:

$$\pi_H = \frac{1-q}{1-q(1-p)^R} = \frac{1-F(\theta)}{1-F(\theta)(1-p)^R}.$$

where the last equality follows from the definition of q, i.e. $q = F(\theta)$. Accordingly, the long run probability of observing a low quality agent within a cycle is given by:

$$\pi_L = \frac{F(\theta)(1-(1-p)^R)}{1-F(\theta)(1-p)^R}.$$

Using π_L and π_H, long run average revenues per contacted customer Y are given by:

$$
\begin{aligned}
Y &= E(s|L)\pi_L + E(s|H)\pi_H \\
&= \frac{1}{1-F(\theta)(1-p)^R}\left[F(\theta)(1-(1-p)^R)\frac{\int_{-\infty}^{\theta} s\mathrm{d}F(s)}{F(\theta)} + (1-F(\theta))\frac{\int_{\theta}^{\infty} s\mathrm{d}F(s)}{1-F(\theta)}\right] \\
&= \frac{1}{1-F(\theta)(1-p)^R}\left[(1-(1-p)^R)\int_{-\infty}^{\theta} s\mathrm{d}F(s) + \int_{\theta}^{\infty} s\mathrm{d}F(s)\right] \\
&= \frac{1}{1-F(\theta)(1-p)^R}\left[E[s] - (1-p)^R\int_{-\infty}^{\theta} s\mathrm{d}F(s)\right] \\
&= \frac{1}{1-F(\theta)(1-p)^R}\left[E[s] - (1-p)^R F(\theta)E[s|s<\theta]\right]
\end{aligned}
$$

Under regularity conditions for $F(\cdot)$, an interior solution for θ maximizing long run expected revenues solves:

$$
\begin{aligned}
\frac{\mathrm{d}Y}{\mathrm{d}\theta} &= \frac{f(\theta)(1-p)^R}{(1-F(\theta)(1-p)^R)^2}\left[E[s] - (1-p)^R\int_{-\infty}^{\theta} s\mathrm{d}F(s)\right] + \frac{1}{1-F(\theta)(1-p)^R}\left[-(1-p)^R\theta f(\theta)\right] = 0 \\
&= \frac{f(\theta)(1-p)^R}{1-F(\theta)(1-p)^R}\left[\frac{E[s] - (1-p)^R\int_{-\infty}^{\theta} s\mathrm{d}F(s)}{1-F(\theta)(1-p)^R} - \theta\right] = 0 \\
&= \frac{f(\theta)(1-p)^R}{(1-F(\theta)(1-p)^R)^2}[E[s] - (1-p)^R F(\theta)E[s|s<\theta] - \theta + \theta F(\theta)(1-p)^R] = 0
\end{aligned}
$$

Therefore, θ^* solves the following fixed point equation:

$$\theta^* = E[s] + F(\theta^*)(1-p)^R(\theta^* - E[s|s<\theta^*]).$$

Note from this last equation that an interior solution for the optimal threshold θ^* is greater than the population mean $E[s]$.

Chapter 2

Brick and Mortar Stores

2.1 Introduction

Labor operating costs are among the largest expenses in the retail sector, and one of the main sources of employment in developing countries. Whereas retail operations management has made significant progress in improving inventory management (the largest operating expense in retail), labor management has now become critical to improve operational efficiency, in part enhanced by new technologies that provide real-time data on customer traffic and employees. These data has propelled empirical research in retail labor management, starting with the seminal work by Fisher et al. [2006] that studies how changes in the staffed labor affects the sales of retail stores. That study and others that follow show that labor is an important driver of revenues. This work builds over this important stream of work to combine empirical analysis with optimization methods to build a decision support tool that can be used by managers to plan labor allocation and schedule working shifts to maximize store profitability, balancing gross margins with labor costs.

Following Fisher et al. [2006], Perdikaki et al. [2012] studies the relationship between labor and sales, but account for customer traffic as an additional factor. Chuang et al. [2016] expands this work to develop a decision model to plan labor hours at a weekly level by store, accounting also for heterogeneity across stores in a chain. Although this is similar to our work, our goal is to provide a more detailed plan of labor staffing in an hour-by-hour, day-to-day basis, including working shifts of full-time and part-time employees. Our work is related to Mani et al. [2015], that seeks to measure over and under staffing of stores. Their approach uses structural estimation to uncover the gross margin and costs of labor, assuming that the retail store staffs optimally at the weekly aggregate level. We use a different methodology that does not assume that the store sets staffing levels optimally; moreover, the objective of our work is to provide a more detailed description of the different time slots at which a store can be over and understaffed. We also consider labor regulatory restrictions and worker's scheduling preferences to jointly optimize the staffing hours and shift scheduling.

The Operations Research literature has a long history in developing models to optimize shift schedules in service delivery systems using tools from mathematical programming (e.g. Thompson [1995]). However, these models typically take the labor requirements as a fixed

input, finding a feasible schedule that achieves these requirements at minimum costs. In our work, the desired labor requirements is one of the key decisions, which accounts for the trade-off between costs (employee salaries) and margins (sales).

Our methodology is separated into three steps. First, an econometric model is developed to estimate the effect of labor on sales in an hourly basis. This model provides an input to an optimization problem that seeks to optimize the hour-by-hour labor requirement based on traffic projections. This second step is important, because although some previous work has provided valid approaches to plan labor requirements at a weekly level (e.g. Fisher et al. [2017]; Chuang et al. [2016]), these solutions don't provide details to a store manager on how to allocate his labor budget across days of the week and hourly shifts. While useful, this second step provides an ideal labor requirement that cannot be implemented in practice as it is unlikely to comply with labor regulatory laws, company hiring policies and employees' preferences. The third step consists in translating this ideal labor plan into a feasible working schedule, achieved through a mathematical program that seeks to find the best feasible schedule (with several regulatory and practical constraints) to maximize store profitability. These three steps have been developed into a prototype that is currently been integrated into a decision support tool that complements the current services provided by our industry collaborator to its clients.

Evaluating the staffing requirements at an hourly level requires to analyze granular sales data. Although previous work by Mani et al. [2015]; Fisher et al. [2017]; Chuang et al. [2016]; Perdikaki et al. [2012] have also studied the impact of staffing levels on sales, their analysis is at a more aggregate level – daily or weekly. At an hourly level, it is quite common to observe zero sales, which can complicate the specification of the econometric model. To account for this, we decompose the effect of labor on sales into two parts: (i) conversion, which is modeled through a Poisson regression or other similar models for count data that can account for zero transactions during the time interval; and (ii) ticket value, which is always positive and can be modeled through a log-linear regression.

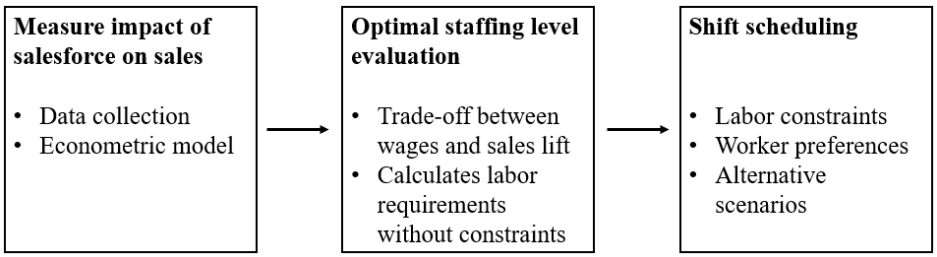

Figure 2.1 General scheme of the decision support tool. The first step uses an econometric model to estimate the impact of employees on sales. The second step uses optimization tools to set an ideal staffing requirement for each store, per hour. The third step translates this ideal labor requirement into a feasible schedule that meets labor regulatory restrictions and practical constraints to ensure compliance with company labor policies and employee's preferences.

Closest to our work is Kabak et al. [2008], who develop a decision model to optimize workforce planning and shift scheduling considering the impact of staffing levels on sales. They use a model similar to Lam et al. [1998] to estimate the sales response to changes in labor. We make several contributions to that work. First, our math programming formulation is

easier to adapt to incorporate alternative constraints that limits the set of possible feasible schedules. This is particularly important in the markets that we study, where the retail sector has been facing a growing number of legal restrictions that make the workforce planning process challenging. In addition to the regulation, employee preferences for working schedules have been shown to be important in determining productivity. Mas and Pallais [2017] study the relationship between schedule preferences and productivity in the context of call-centers. In retail, the recent field study by Williams et al. [2018] shows that providing stable schedules to store employees leads to meaningful lift in sales and productivity, mostly due to improvements in employee retention, which has been shown to impact performance (Ton and Huckman [2008]). The mathematical formulation developed in this research is flexible to incorporate alternative sets of constraints, which we evaluate in section 2.6.

A second difference with Kabak et al. [2008] is on the econometric model used to measure the impact of staffing on sales. Our model is tailored specifically to handle granular hourly sales data with frequent zeros. Moreover, we develop a novel specification to capture the interaction of traffic and staffing levels, and we show that this formulation improves the goodness-of-fit compared to the models of Lam et al. [1998] and Chuang et al. [2016]. Finally, our econometric model handles some potential endogeneity problems that were not considered by Kabak et al. [2008] and Lam et al. [1998]. On this regard, the recent work by Fisher et al. [2017] shows that identifying a causal effect of labor on sales is challenging because staffing levels are endogenous: a positive correlation between staffing and sales can be generated by omitted variables in the regression model which are observed and accounted for by the store managers in the labor planning process. To identify the causal effect of labor on sales, we follow the same strategy as Fisher et al. [2017], using deviations between the planned and actual staffing level of the store as exogenous variation. However, our deviations are measured at a more granular level using hourly planned and actual schedules, which were collected from the stores archival data (Fisher et al. [2017] uses weekly labor hours).

The empirical results reveal a non-linear effect of traffic and staffing levels on sales, similar to the findings of Chuang et al. [2016] and Perdikaki et al. [2012]. We decompose this effect into conversion and basket value and find that most of the effect is through the former. The magnitude of the effect is comparable to the results of Fisher et al. [2017]: increasing labor in under-staffed stores can increase sales in the order of $2\% - 5\%$ (depending on the level of traffic of the store).

In terms of the optimization of hourly staffing requirements, the ideal staffing plan suggested by the optimization is compared to the historical staffing levels used by the retailer to evaluate potential inefficiencies. In general, we find that most stores are both over and understaffed at different time periods and days of the week. This patterns of over and understaffing by hour and day of the week is a contribution to the literature, which has usually analyzed this problem at a more aggregated level, which misses some of the inefficiencies that can be identified with the more detailed approach we propose. The results suggest that the stores analyzed exhibited severe understaffing during Fridays and weekends from 3pm to 6pm and moderate overstaffing during the weekday lunch period (12-3pm). We provide some insights of potential causes of these inefficiencies based on traffic patterns and the observed deviations between planned and actual labor.

In summary, our work integrates several features of previous work, combining hourly customer traffic, planned and actual staffing levels and optimization methods to develop a decision support tool that can be used by store managers to optimize the labor plan together with a detailed shift schedule that complies with labor regulations and other real-world constraints. We effectively test this methodology in a children apparel retail chain in South America. We also use the model to quantify the economic impact of labor regulatory restrictions on store profitability, which are relevant in the context of our application due to ongoing changes in the regulatory environment.

2.2 Industry Collaboration and Data

This research is in collaboration with a company focused on providing technology services to retailers, focused in solutions to analyze customer traffic and store operations with objective metrics. One of the main products is a system to monitor customer traffic in retail stores of different format, including specialty stores and big-box retailers, providing descriptive data reporting traffic patterns and forecasts. The company realized that traffic counters and other related technologies are rapidly becoming "commodities" in a competitive market of providers that offer similar solutions for retailers. Consequently, the company is motivated to expand its services with its current clients to provide additional solutions that go beyond descriptive reports, developing decision support systems to prescribe improvements to the operational execution of retail chains. Big-box retailers and large international chains use enterprise system solutions for workforce management (e.g. Kronos), but these systems are to complex and costly for smaller retail chains. Moreover, retail labor tends to be more heavily regulated in Latin America relative to the U.S., and adapting the existing solutions to the idiosyncrasies of each country is not straightforward. Hence, we started a research collaboration with the objective of developing a decision support tool aimed for this market – a workforce management solution for medium-sized specialty retail stores in Latin America.

The first stage of the research project involved discussing with retail managers how they planned working schedules across the chain. In general, labor budgets for each store were calculated using historical sales, projecting labor allocations based on future sales forecasts and a fixed sales to labor ratio. However, because sales are a function of labor, this approach misses the opportunity of realizing the full demand potential in periods where the stores have been systematically understaffed. An important improvement of the new tool is to allocate labor in relation to traffic, which is a better measure of the demand potential on the store and is not directly affected by staffing levels. It was also observed that the decisions of translating labor requirements to working schedules was delegated to store managers, and each manager used a different approach to implement the workforce scheduling. Consequently, another improvement of the new tool is to provide a systematic approach to translate labor requirements into a feasible working schedule, through an interactive system that be used by store managers to evaluate alternative scenarios.

Based on these observations, we concluded that the proposed solution had to be built with three complementary modules, represented in Figure 2.1, addressing three interrelated challenges in retail workforce management. The first module uses as input traffic data and staffing levels to generate a prediction of sales, which is used to build a sales curve as a function of staffing levels. The second module uses this sales response curve and traffic forecasts

to conduct a cost-benefit analysis that trades-off revenues (times margins) with employee salaries, to determine labor requirements. This information is useful for retail executives to plan labor budgets across the chain. The third module is focused on implementing a feasible workforce schedule in order to attain the labor requirements, providing a systematic approach to support the decisions currently made by store managers.

To develop a prototype of this decision support system, we also partnered with a medium sized retail chain of children apparel, with more than 30 stores located in Chile, open from 9am to 9pm seven days a week. Eleven stores were chosen for the purpose of the research study, located in Santiago (Chile's capital), which had customer traffic data for more than one year of history. The following data were collected for the period of February to December of 2016:

- Transactional data: This information was collected through the detailed transactions of each sales ticket, which were aggregated to calculate the number of tickets sold and the monetary value of each ticket.
- Personnel data: The actual number of employees working in each store were calculated using the attendance control system, which captures the times when employees check in and out of their shift, as well as lunch breaks. In addition, we also collected data on the planned schedule for employees, which are used to handle some of the endogeneity issues discussed in section 2.3.
- Traffic data: This information is collected via video cameras and image recognition algorithms, tracking inbound and outbound traffic in the store (including multiple entrances in some stores).

A summary of the data for the stores in our study is presented in Table 2.1. The table reveals some heterogeneity across stores in terms of sales, traffic, conversion rates and staffing levels. Later in the analysis, store 9 was chosen as a representative store to illustrate some of the results.

Store	**Traffic**	**Sales**	**Conversion**	**Staffing**
1	313.84	0.96	0.28	2.73
2	328.91	0.57	0.15	2.51
3	340.11	0.64	0.15	2.50
4	351.35	0.58	0.13	2.06
5	331.66	0.61	0.14	2.43
6	252.66	0.56	0.17	2.37
7	397.96	0.68	0.14	2.74
8	590.68	1.00	0.14	3.01
9	485.04	0.94	0.17	3.11
10	376.17	0.70	0.14	3.01
11	196.40	0.54	0.20	2.22

Table 2.1 Summary statistics. Daily averages for the eleven stores in the study, covering February to December of 2016. Sales column was rescaled by store 8 sales volume.

The next section describes how the data is used to estimate the impact of staffing levels on sales, which corresponds to the first module of Figure 2.1.

2.3 Empirical Model of Sales Response to Labor

Choosing appropriate staffing levels in a retail store requires to evaluate a trade-off between the cost of employee wages versus the additional profits accrued from increasing labor. Hence, the first step to develop a model that can predict sales as a function of the staffed labor in the store. This section describes an econometric model that measures the causal effect of increasing labor on sales, which is moderated by the level of customer traffic to the store.

Our empirical model of sales is based on the original work by Lam et al. [1998] and the extensions developed by Chuang et al. [2016]. The model is formulated over a panel of stores, where i indexes the store and t a time period. The exact specification of the period depends on the application, and will be discussed in further detail later in this section.

Let y_{it} denote the total sales (in \$) of store i in a specific period. Let E_{it} denote the average number of employees that worked in the store during that period, and denote by T_{it} the flow of customers that entered the store. Lam et al. [1998], Chuang et al. [2016] and Mani et al. [2015] use an econometric model of the form:

$$y_{it} = \alpha_i T_{it}^{1-\rho} \exp(f(T_{it}, E_{it}; \beta) + \text{Controls} + \varepsilon_{it}) \tag{2.1}$$

which essentially captures the effect of traffic and labor on sales. Equation 2.1 captures a saturation effect through traffic, as measured by the coefficient ρ. When $0 < \rho < 1$, there is a marginally decreasing effect of traffic on sales, which is in part determined by the limited resources of the store. When customer traffic increases, the store offers limited capacity to provide customer service, which reduces the sales per customer. For example, the cashiers may be operating at capacity, leading to long lines and customers that opt to walk out of the store without purchasing to avoid the waiting (Lu et al. [2013]). In apparel stores such as the one studied in this application, fitting rooms become congested thereby precluding customers to try on clothes, which leads to lower sales. In all the aforementioned studies that use this class of model, the coefficient ρ is positive taking values in the range of $0.7 - 1.0$, consistent with this saturation effect. In other studies, this saturation effect is so severe that sales may actually decrease at high levels of traffic, generating a inverted U-shape effect of traffic on sales (Perdikaki et al. [2012]).

Equation (2.1) also captures the interaction between customer traffic and store labor through the parametric function $f(T_{it}, E_{it}; \beta)$. Employees are one of the key resources in the store to provide valuable service-related activities, including assisting customers, attending cashiers, organizing the inventory, among other functions. As customer traffic increases, the utilization of employees goes up and reaching at some point a critical level where service starts to degrade. This lower service-level can reduce sales per customer, contributing to the saturation effect mentioned earlier. This effect can be moderated by increasing staffing levels in the store, so that the marginal effect of traffic on sales is in part determined by the number of employees in the store, as captured by $f(T_{it}, E_{it}; \beta)$ in equation (2.1). As we describe later, this function is specified in order to capture the level of worker capacity utilization, so that $f(T, E)$ increases with demand (traffic) and decreases with the number of employees (E). For this reason, it is also referred to as the employee utilization effect.

Equation (2.1) provides the basis to develop an empirical model that captures sales response to customer traffic and staffing levels. However, there are specific issues that require adapting the model to be used in the context of labor staffing and scheduling. In particular, the objective of this work is to have a sales response model that can be used to plan staffing levels at a granular level, in order to provide flexibility to plan working shift schedules. Many of previous work has defined the unit of analysis at a daily (Perdikaki et al. [2012]) or weekly (Chuang et al. [2016]; Fisher et al. [2017]) level, which works well for assigning labor budgets but is too aggregated for the purpose of scheduling working shifts. In the context of the application studied, the daily schedule was divided into 4 time blocks separated at noon, 3pm and 6pm (stores open at 9am and close at 9pm), which provides enough flexibility to define working schedules of full-time and part-time employees.

The usual approach to estimate equation (2.1) is to take logs and use ordinary least squares or some variant of linear regression methods. A problem arises when using a model in more granular periods as in our study, because zero sales within a block of 1-4 hours can happen frequently and the log-linear regression model cannot be applied directly. Dropping periods with zero sales can induce a sample selection bias in the estimation. To address this issue, the method proposed in this work is to decompose the sales equation (2.1) into two separate outcomes: (i) conversion and (ii) average sales per ticket, and using a different parametric model to model each outcome. This methodology is described next.

2.3.1 Model Conversion and Basket Sales

Sales can be decomposed as the product of the number of tickets, N_{it}, times the average sales per ticket, B_{it} (also referred as basket sales). If the time unit of analysis t is short, then the number of tickets can take the value zero frequently. In the context of our application, zero sales occur quite frequently: grouping transactions in three hour blocks, the histogram of the number of tickets per block shown in figure 2.2 reveals more than 6% of observations with no transactions.

A Poisson regression model is appropriate for count data of these characteristics. The number of transactions N_{it} is modeled as a Poisson process with rate :

$$E(N_{it}) = T_{it} \exp(\delta_i - \rho \log(T_{it}) + f(T_{it}, E_{it}; \beta) + \text{Controls}) \tag{2.2}$$

where the exponential term captures the conversion of traffic into tickets. This conversion includes a traffic saturation effect ρ and a moderation effect of customer traffic with the number of employees, plus additional controls. Provided a specification for $f(T_{it}, E_{it}; \beta)$, the model can be estimated via Maximum Likelihood using standard statistical packages.

To model sales value per ticket, using the monetary value b_{in} of each ticket $n = 1 \dots N_{it}$ in period t, compute the average basket (ticket) value as $B_{it} = \frac{1}{N_{it}} \sum_{n=1}^{N_{it}} b_{in}$ for all periods with positive sales ($N_{it} > 0$) and specify the linear regression:

$$\log(B_{it}) = \delta_i + \alpha \log(T_{it}) + f(T_{it}, E_{it}; \beta) + \text{Controls} + u_{it} \tag{2.3}$$

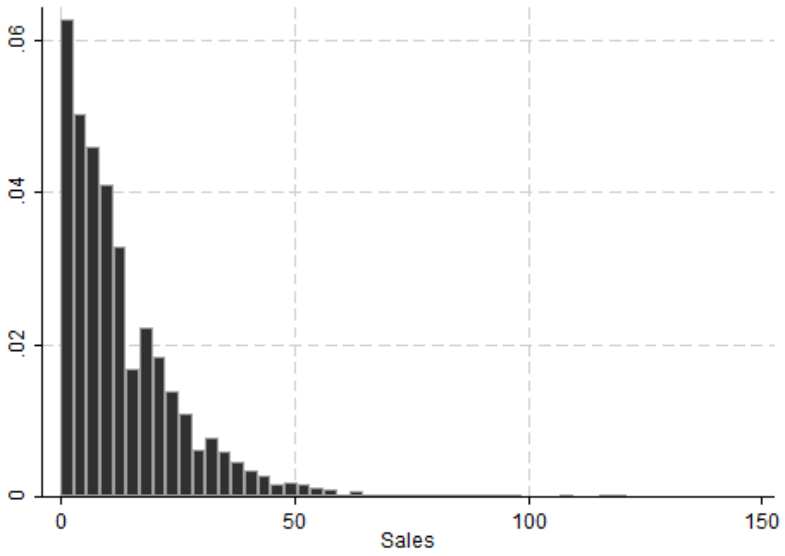

Figure 2.2 Zero sales. Histogram showing the frequency of zero sales hour-block periods in the data.

In this model, in contrast to the conversion model (2.2), average basket value is not necessarily increasing in sales. Therefore, a traffic-saturation effect in this model would be captured by a negative α coefficient. Models (2.2) and (2.3) are estimated independently, with restrictions on the parameters across the models. Note that the sample size of model (2.3) drops observations with $N_{it} = 0$ and therefore has a smaller sample size relative to the estimation of equation (2.2). Next, we discuss important endogeneity issues that need to be accounted for proper identification of the model parameters.

The main purpose of this econometric approach is to predict how sales respond to changes in the level of staffing in the store, E_{it}. Hence, it is important that the model estimation properly identifies the causal effect of E_{it} on sales (through N_{it}, B_{it} or both), in order to predict the outcome of alternative staffing scenarios for the purpose of optimizing labor schedules. Estimating the causal effect of employee staffing on conversion and basket value through an econometric analysis using historical transactions data is challenging because the staffing levels are endogenous: they were decided by store managers based on sales forecasts. Many retail stores, including the retail chain in this application, plan their staffing levels based on projected sales, which induces a positive correlation between store labor and sales which cannot be interpreted as a causal effect of labor on demand. It is therefore important that the controls included in the econometric model account for factors that retail managers may use to predict demand and allocate labor resources, in order to avoid a potential omitted variable bias.

The following covariates were included as controls in the estimation. First, a detailed set of special dates, including all the main holidays (Christmas, New years, Independence day, Religious holidays and others) and special promotion dates (including a "children's day" and winter and summer vacations). Month and week dummies are included in some specifications, as well as day of the week and hour of the day indicators. Weather patterns, which in some cases may be used to forecast sales, were included using a binary indicator variable of rainy

days.

It is still possible that retail managers used additional factors to predict demand and choose staffing levels. To account for this, we collected data on the planned staffing level at each store and hour-block, to compute the average planned labor hours P_{it} for each store and time block. A comparison of the predicted (P_{it}) and actual (E_{it}) labor across time blocks is illustrated in Figure 2.3 for one store in the sample. The figures reveal that, on average, the planned and actual staffing are aligned in weekday up to 3pm, with some dis-adjustment later in the day. In weekends, however, actual staffing is on average above planned labor. Our forthcoming analysis shows that currently the stores tend to be understaffed on weekends, and it appears that store managers are adjusting upwards the planned staffing levels during the weekend to mitigate this understaffing.

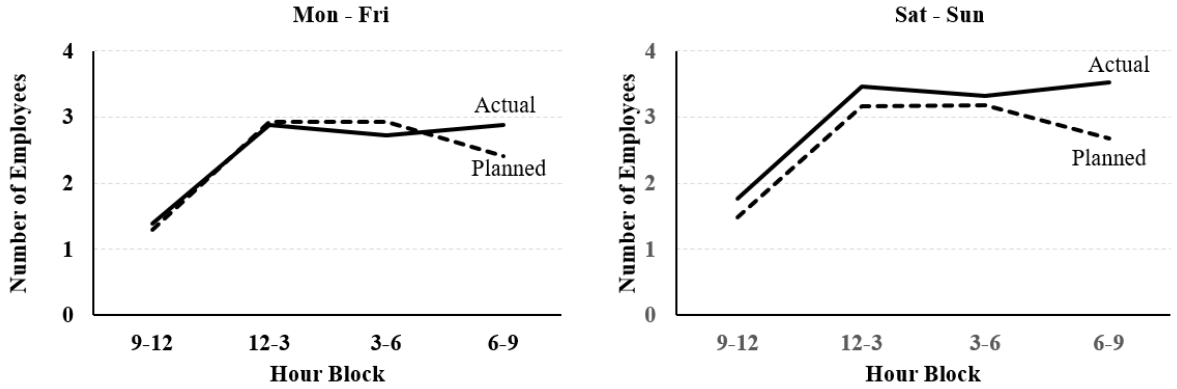

Figure 2.3 Actual versus planned labor. Comparison of average planned labor (P_{it}) and actual labor (E_{it}) in a typical store of the retail chain, for weekday and weekend days.

Figure 2.4 shows a histogram of the deviations from the planned labor hours, as measured by the relative difference $(P - E)/P$, for the 12-3pm and 6-9pm hours blocks. The figure reveals that although on average the actual and planned labor are relatively well aligned, there is significant deviations (the same pattern was observed for the hour blocks not included in the figure). The 12-3pm hours blocks reveals a large proportion of less-than-planned labor around the lunch hour, which could be due to longer lunch periods taken by the staff or idle times generated during changes of shift for part-time workers. The 6-9pm hour block reveals the opposite, a higher fraction of observations suggesting more-than-planned labor (this pattern is consistent with the differences in averages shown in Figure 2.3 for the last hour block).

It can be concluded that, although planned and actual labor tend to be aligned on average, there are substantial deviations across time, as observed by Fisher et al. [2017]. That study uses the deviation from planned labor as an exogenous source of variation to estimate the effect of staffing on sales, and we follow a similar approach. The underlying assumption is that the deviations from planned labor due to absenteeism, late arrivals, employees leaving early, among other factors, are exogenous. Fisher et al. [2017] actually validate this with a field experiment, showing that deviations from planned labor seem to be an appropriate source of variation to estimate the effect of labor on sales. However, Fisher et al. [2017] use a week as the period analysis, whereas models 2.2 and 2.3 use more granular time periods of 3 hour blocks. Moreover, Fisher et al. [2017] does not use customer traffic, therefore it

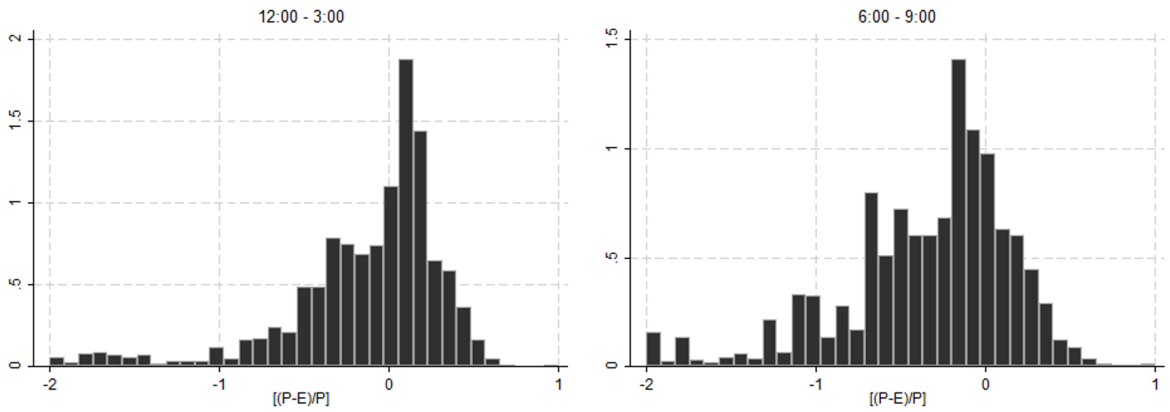

Figure 2.4 Deviation from planned labor. Histogram of the relative deviation from the planned labor hours, $(P - E)/P$.

is not clear how to specify the interaction between traffic and the deviations from planned labor. Hence, we adapted the specification of Fisher et al. [2017] to use it in our econometric specification in the following way. Rather than computing the difference between planned and actual number of employees, planned labor P_{it} was added as an additional control variable in the model. The main objective here is to use planned labor as a proxy for omitted variables that affect demand that are also taken into account by store manager. Note that the estimated coefficient of P_{it} cannot be interpreted as a causal effect: it is the actual labor in the store that affects sales, not the planned labor. High levels of P_{it} should be interpreted as a signal that the retail management planned for a higher service capacity, possible driven by high expected demand. Planned labor P_{it} was included as a set of dummy variables capturing different levels of planned labor, a total set of eight dummies for discrete value of planned labor hours.Planned labor was averaged over a time block and rounded up, creating a categorical variable that was implemented with a set of dummies. It remains to specify the functional form of $f(T, E; \beta)$ which captures the effect of the number of employees and its interaction with customer traffic, which is described next.

2.3.2 Specification of the Interaction Effects Between Labor and Customer Traffic

As discussed by Chuang et al. [2016], there are alternative specifications for the traffic/employee moderation effect $f(T_{it}, E_{it}; \beta)$, which aims to capture the effect of employee capacity utilization. Previous literature has used alternative specifications; for example, Lam et al. [1998] uses the inverse of the number of employees, $f(T, E; \beta) = \beta/E$ (Mani et al. [2015] use a similar specification). Chuang et al. [2016] argue that the level of service provided to customers is affected by the utilization of the employees, which is better captured by the demand to capacity ratio T/E, so that the elasticity to labor depends on the level of customer traffic. We tested these alternative specifications and also included a quadratic term $(T/E)^2$ to account for more flexible non-linear effects.

From our experience in working with various specialty-retail chains, practitioners' view employee utilization as a function of the ratio between the number of customers shopping in the store (as opposed to the customer flow T) and the number of employees staffed, similar

to what is used in hospitals through the patient to nurse ratio. By Little's Law, the average inventory of customers in the store is equal to the average time spent in the store times the traffic. Hence, if the average time spent in the store is constant and similar across stores and time blocks, then using the ratio T/E is equivalent to use the customer inventory to employee ratio: the covariate is simply rescaled by a constant. However, it is possible that the type of customers vary across stores, days of the week and time blocks, in which case it may be useful to account for this heterogeneity.

This idea motivates an alternative specification to capture the effect of staffing levels on sales. Defining ω_{it} as the average time that customers spend in the store, the ratio $(\omega_{it}T_{it})/E_{it}$ is a reasonable measure of capacity utilization if demand for service is better captured by the inventory of customers in the store, rather than the flow (which relates to the T_{it}/E_{it} measure used by Chuang et al. [2016]). Kabak et al. [2008] also discusses the issue of customer sojourn time in the store, using a weighted average of concurrent traffic and lagged traffic as an approximate measure of employee utilization.

One approach to estimate the average time spent in the store, ω_{it}, is to use a sample of customers, measure their store visit time and use it to calculate averages by store and time-block. This approach requires substantial data collection effort, because the people tracking technology used in the stores cannot recognize the identity of each customer to record their individual entry and exit time. Instead, we develop an alternative approach that uses aggregate customer inflow and outflow data to estimate the distribution of the time spent in the store, which is based on the method developed in Lu et al. [2013].

Consider a single store (store index i is suppressed) during a specific time-block (e.g. 12-3pm). Time is discretized in 2 minute intervals, and let I_t and O_t be the inflow and outflow of customers in the interval t. Let θ_s denote the probability that a customer spends s time intervals in the store, and $\mathcal{I}_t = (I_t, I_{t-1}, \ldots)$ the history of inflows up to time interval t. The expected outflow of customers in period t is given by:

$$E[O_t|\mathcal{I}_t] = \sum_{s=0}^{S} \theta_s I_{t-s} \tag{2.4}$$

with a number of lags S such that the total probability mass $\sum_{s=0}^{S} \theta_s$ is close to one. A linear regression of O_t on I_t and its lags, without an intercept, provides consistent estimates of $\{\theta_s\}_{s=1}^{S}$. In the context of our application, we found that $S = 12$ was enough to capture the distribution of the time spent in the store. The regression was run separately for each of the 11 stores, during the weekend and weekday, providing a total of 22 sets of estimated coefficients $\{\theta_s\}_{s=1}^{S}$. Figure 2.5 shows the estimates for one of the stores for weekends; for weekdays the estimates were similar. The figure reveals that there is an important fraction of customers that spend less than 2 minutes in the store (about 30%). The total probability accounted for the coefficients is more than 90%. Based on the estimated coefficients for all the stores, an average customer visit time is calculated for every store on weekends and weekdays. In general, the overall average visit time is around 6.4 minutes and is relatively stable across stores and week/weekends, with a coefficient of variation of 6. Nevertheless, there is some heterogeneity across stores and weekdays/weekends that is relevant to specify

the ratio of customer inventory to employees, $\omega_{it}T_{it}/E_{it}$, as an alternative measure of employee utilization.

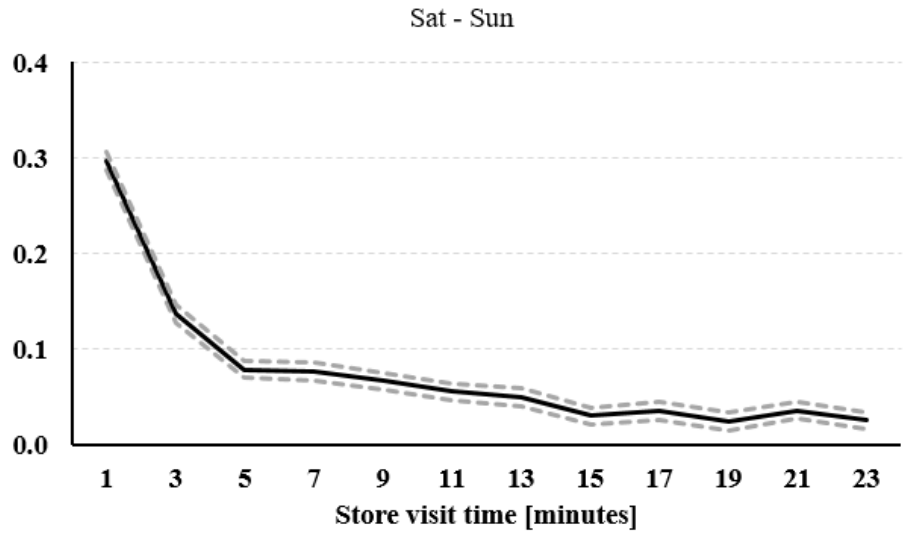

Figure 2.5 Time spent in store. Distribution of the time spent in the store, estimated via a linear regression model of equation 2.4. Dotted lines show the 95% confidence interval of the estimates.

2.3.3 Estimation Results

This section reports the estimation results of alternative specifications of equations 2.2 and 2.3. Table 2.2 reports the results of the conversion model defined by equation 2.2. Recall that the model is estimated with a Poisson regression to account for the count data associated with the number of transactions, as well as the frequent zeros observed in time blocks of low rotation. The first column of Table 2.2 shows the specification with employee capacity utilization defined as $f(T_{it}, E_{it}; \beta) = \beta/E_{it}$, the second column with $\beta T_{it}/E_{it}$ and the third column with $\beta\omega_{it}T_{it}/E_{it}$. The coefficients associated to these covariates are reported in the row labeled "EmpUtil". Columns (4) to (6) replicate the specifications adding a quadratic term, with the corresponding covariate labeled as "EmpUtil Sq". All the specification include store fixed effects and dummies for day of the week, time block interacted with weekend dummy, holidays and special promotion dates, a week dummy and rainy day dummy.

In all specifications, the coefficient associated with log traffic $(1 - \rho)$ is between 0.93 and 0.99; the saturation effect of traffic appears to be small in this application in comparison to the results of previous studies. However, an alternative explanation is that the company partner has made special effort to improve the precision of their traffic counters. If indeed the traffic metric used in this study is more precise, then it would mitigate the attenuation error that arises in econometric models with measurement error in the covariates, leading to a larger and less biased coefficient.

In columns (1)-(3), the coefficient for the linear term "EmpUtil" is negative and significant, suggesting a positive effect of the number of employees on sales (E_{it} is in the denominator of "EmpUtil"). Among the models with the quadratic term for "EmpUtil", columns (5) and (6) report a positive and significant coefficient on the quadratic term, suggesting a

marginally decreasing effect of employee utilization (the effect of the number of employees on sales is reported later in Figure 2.6). The bottom of the table reports statistics to evaluate the goodness of fit for the purpose of model selection. The AIC and BIC criteria was used to compare across models (which are not nested among each other), suggesting that the quadratic model with $Ratio = \omega_{it} T_{it}/E_{it}$ has the best fit (column (6)). This suggest that modeling heterogeneity in customer visit time is relevant for the purpose of measuring the effect of labor, which corroborates the intuition of the store managers that we have interviewed suggesting that the number of customers in the store is a more appropriate metric to capture the demand for service.

Recall that the model has covariates that capture the number of planned employees for the store, with the purpose of controlling for potential omitted variables that affect sales and are considered in the staffing decisions of retail managers. The estimates suggest that, in this application, most of these additional controls are not statistically significant, which appears to be at odds with the results of Fisher et al. [2017] who show that controlling for planned labor is important to mitigate endogeneity bias. However, an important difference between this study and Fisher et al. [2017] is that the present study includes customer traffic as a covariate in the model. Hence, it is possible that factors that are accounted for by store managers in planning their labor are being captured by store traffic, and therefore traffic alone is enough to address the endogeneity problem. This arguments were presented by Chuang et al. [2016] and Perdikaki et al. [2012] to justify their empirical strategy (which doesn't use planned labor as a control), and our empirical analysis seems to corroborate that justification.

The full set of coefficient estimates are reported in Table 2.3 for the specification corresponding to column (6) of Table 2.2, the model with the best fit. The coefficients for day of week dummies (base level is Sunday) suggest that Monday has the lowest sales, which gradually progress increasing during the week to reach its peak on Saturday. The morning time block (the base dummy) has the lowest sales on average, whereas the other blocks have similar coefficient estimates. Weekends tend to have lower sales after noon, as suggested by the interaction terms of time block and weekend.

Table 2.4 shows the estimation results of the basket value model defined by equation (2.3), with the same structure as the previous table. In this model, the effect of "EmpUtil" is not significant in all the models. The effect of staffing levels appears to operate mainly by increasing the conversion of customer traffic into tickets, but does not affect the value of these purchases.

To visualize the magnitude of the effect of staffing levels on sales, the graph in Figure 2.6 uses the estimates of Table 2.2 column (6) to compute the sales response to the number of employees for different levels of store traffic in a representative store. This model specification is chosen because: (i) AIC and BIC suggest this has the best fit among all the conversion models evaluated; (ii) the empirical results suggest that basket value is not affected by staffing levels, therefore all the impact in sales is captured through the conversion model. The left panel shows the sales response curve for moderate traffic scenarios, the 3-6pm block during two weekdays. The y-axis has been rescaled in order to conceal sensitive information of the retail company that collaborated in the study. These curves show an increasing effect

	Different specifications of EmpUtil, $f(T,E)$					
	$1/E$	T/E	$\omega T/E$	$1/E$	T/E	$\omega T/E$
ln(Traffic)	0.94***	0.97***	0.97***	0.94***	0.99***	0.99***
	(0.008)	(0.010)	(0.010)	(0.008)	(0.011)	(0.011)
EmpUtil	-0.100***	-0.002***	-0.020***	-0.098*	-0.005***	-0.048***
	(0.022)	(0.0005)	(0.005)	(0.040)	(0.001)	(0.008)
EmpUtil Sq				-0.0008	0.00003***	0.003***
				(0.023)	(0.00001)	(0.001)
N	13,481	13,481	12,440	13,481	13,481	12,440
pseudo R-sq	0.639	0.639	0.644	0.639	0.640	0.644
AIC	70,549	70,555	65,423	70,551	70,536	65,403
BIC	71,443	71,448	66,299	71,452	71,437	66,287

Standard errors in parentheses * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 2.2 Conversion model, estimation results. Model in equation 2.2.

	Coef.	**Sd. Err**		**Coef.**	**Sd. Err**
ln(Traffic)	0.995***	(0.011)	Mon	-0.172***	(0.025)
EmpUtil	-0.048***	(0.008)	Tue	-0.136***	(0.024)
EmpUtil Sq	0.003***	(0.001)	Wed	-0.101***	(0.024)
			Thu	-0.052*	(0.024)
Store2	-0.554***	(0.014)	Fri	0.025	(0.024)
Store3	-0.564***	(0.013)	Sat	0.158***	(0.011)
Store4	-0.746***	(0.013)			
Store5	-0.665***	(0.014)	12-3pm	0.407***	(0.018)
Store6	-0.484***	(0.014)	3-6pm	0.376***	(0.019)
Store7	-0.620***	(0.013)	6-9pm	0.397***	(0.018)
Store8	-0.558***	(0.015)	Wknd x 12-3pm	-0.119***	(0.024)
Store9	-0.598***	(0.013)	Wknd x 3-6pm	-0.144***	(0.024)
Store10	-0.666***	(0.014)	Wknd x 6-9pm	-0.206***	(0.024)
Store11	-0.305***	(0.015)			
Rain	0.002	(0.0115)	Constant	-1.382***	(0.083)

Standard errors in parentheses * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 2.3 Conversion model, full estimation results. Model in equation 2.2, corresponding to the specification with EmpUtil $= \omega T/E$, column (6) in Table 2.2.

of employees on sales that is also marginally decreasing. The magnitude of the effect is important: increasing from 1 to 4 employees increases sales in the order of 10%. The right panel shows scenarios of high traffic on Saturday, which reveals an even larger effect of increasing staffing from 1 to 4 employees, in the order of 15%.

	Different specifications of EmpUtil, $f(T,E)$					
	$1/E$	T/E	$\omega T/E$	$1/E$	T/E	$\omega T/E$
ln(Traffic)	0.004**	0.005***	0.005***	0.004**	0.006***	0.005**
	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)	(0.002)
EmpUtil	-0.011**	-0.00004	-0.0005	-0.013*	-0.0002	-0.002
	(0.004)	(0.0001)	(0.001)	(0.006)	(0.0001)	(0.001)
EmpUtil Sq				0.0008	0.000002	0.0001
				(0.002)	(0.000001)	(0.0001)
N	12,941	12,941	11,944	12,941	12,941	11,944
pseudo R-sq	0.000	0.000	0.000	0.000	0.000	0.000
AIC	53,917	53,917	49,779	53,919	53,919	49,781
BIC	54,806	54,806	50,651	54,815	54,816	50,660

Standard errors in parentheses * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 2.4 Basket value model, estimation results. Model in equation 2.3.

The highest traffic scenario, Saturday 3-6pm, shows that at low staffing levels the sales response curve is not concave. When the store is too congested, increasing from one to two employees has a negligible effect, where the positive impact of increasing staffing levels requires a minimum level of staffing, above two employees, to become significant. This is in contrast to Chuang et al. [2016] who prove (analytically) that the sales response function is concave under their specification. The main difference between their model and ours is that we include a quadratic term of the ratio traffic/employees, which in our application is highly statistically significant, providing the flexibility to accommodate non-concave sales response curves.

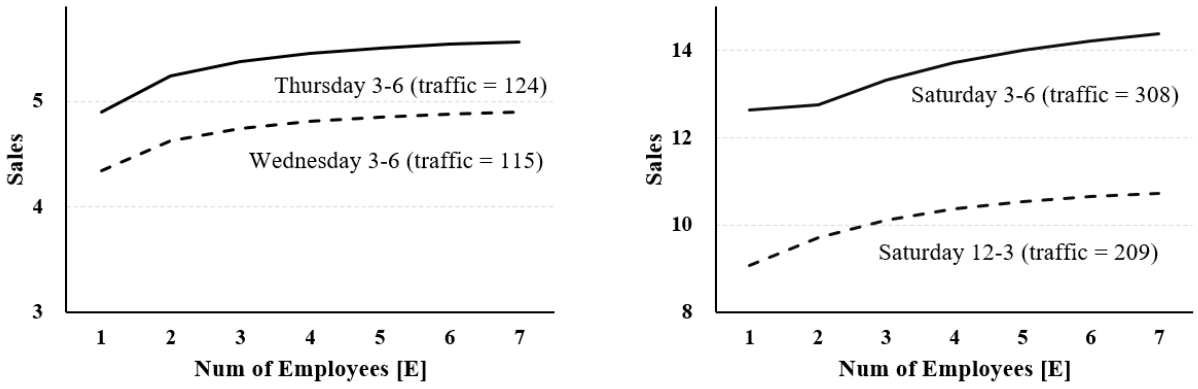

Figure 2.6 Sales response to staffing level. Representative store for different scenarios of customer traffic. The graph is computed using the estimates of Table 2.2, column (6).

Overall, the results suggest that employee utilization, as captured by the ratio of customer inventory to staffing level, has a significant impact on sales. The effect appears to operate

mainly by increasing conversion rates; the effect on basket size is not significant, suggesting that employees do not play an active role in cross-selling efforts in this retail chain. The next section shows how to use these empirical results to evaluate over and understaffing across store and time block and how to allocate labor budgets across stores.

2.4 Evaluate Under and Over-Staffing

The estimates of the sales response to staffing levels and customer traffic can be used to evaluate: (i) the optimal staffing level at each store and hour block; (ii) assess whether the current number of employees at each store is over or understaffed.

An ideal staffing plan for a store would be to evaluate the number of employees for each hour block, adding labor until the hourly cost exceeds its marginal profit contribution. If labor schedules are fully flexible, in the extreme where workers can be hired for a single hour block, then the problem is decoupled and each store/hour-block is solved separately. This solution is unlikely to be feasible in practice, because it would not comply with labor regulation which requires some stability of working schedules. Moreover, unbalanced working shifts are unappealing for workers and may induce high turnover, which in the long run translates into lower productivity (Williams et al. [2018]). Nevertheless, it serves as a baseline that provides a first order evaluation of the labor budgets that should be allocated across stores. It also provides an upper bound on the profit that could be achieved in a more realistic situation, that is also useful to asses the economic cost of regulatory constraints and labor preferences on the scheduling costs of a retail chain. A similar analysis is conducted by Lam et al. [1998] to allocate labor.

Consider a store/time slot combination, with a traffic forecast $\hat{T}_{it}$ and projected basket value $\hat{B}_{it}$ (which is independent of labor and traffic, as suggested by the empirical results of section 2.3.3). Let $N_{it}(\hat{T}_{it}, E_{it})$ denote the expected number of transactions as a function of the staffing level and the projected traffic. Define "Margin" as the contribution margin (as a fraction of sales) and w_t the salary of employees for hour block t. The optimal number of employees in store/time block (i, t) solves:

$$\max_{E_{it}} \quad \text{Margin} \times \hat{B}_{it} \times N_{it}(\hat{T}_{it}, E_{it}) - w_t \times E_{it} \tag{2.5}$$

Chuang et al. [2016] shows that the objective function is concave when the employee capacity utilization function $f(E, T) = \beta T/E$ with positive β. In this case, the optimal solution is to set staffing levels until the marginal contribution Margin $\times \hat{B}_{it} \times \mathrm{d}N_{it}/\mathrm{d}E$ equals the hourly salary w_t. For other specifications, including the one used to construct the sales response function used in our application, this is not necessarily the case. Recall that Figure 2.6 shows that for hour-blocks with high customer traffic, the revenue function is not concave. Nevertheless, because the problem is decoupled by hour block, it is possible to do a fast search to find the optimal staffing level. This can be done for any store and hour-block, as long as a projection for traffic, basket value and the specification of $N_{it}(\cdot)$ are provided. Estimates of $N_{it}(\cdot)$ and $\hat{B}$ are obtained from the empirical models described in section 2.3. Next, we discuss how to generate traffic projections.

2.4.1 Forecasting Customer Traffic

Lam et al. [1998] develops a time series econometric model to make traffic projections based on historical traffic data. We replicated this model at the hour-block level, using a logarithmic transformation on traffic $\ln(T_{it})$ and differencing with a one week lag (28 hour-block periods) to account for the weekly seasonality in the data. Additional autoregressive and moving average components were included to adjust for autocorrelation in the residuals. The resulting model is defined by $z_{it} = ln(T_{it})$, the one-week lag difference $\Delta_{it} = z_{it} - z_{it-28}$ and the time-series model specified by:

$$\Delta_{it} = \mu + \alpha Holiday + \sum_{l}^{L} \theta_l \Delta_{it-l} + \varepsilon_{it} \tag{2.6}$$

$$\varepsilon_{it} = \sum_{k}^{K} \phi_k \varepsilon_{it-k} \tag{2.7}$$

where μ is the mean of the differences variable capturing possible trends and $Holiday$ is an indicator for holidays. We estimated different specifications for the model separately for each store using data from February 2016 to August 2016; the estimated coefficients of the model for a representative store are reported in Table 2.5. The specification with 2 autoregressive and one moving average term shows the best fit to the data according to the AIC and BIC measures. This model was used to conduct an out-of-sample forecast for October 2016, shown in Figure 2.7.

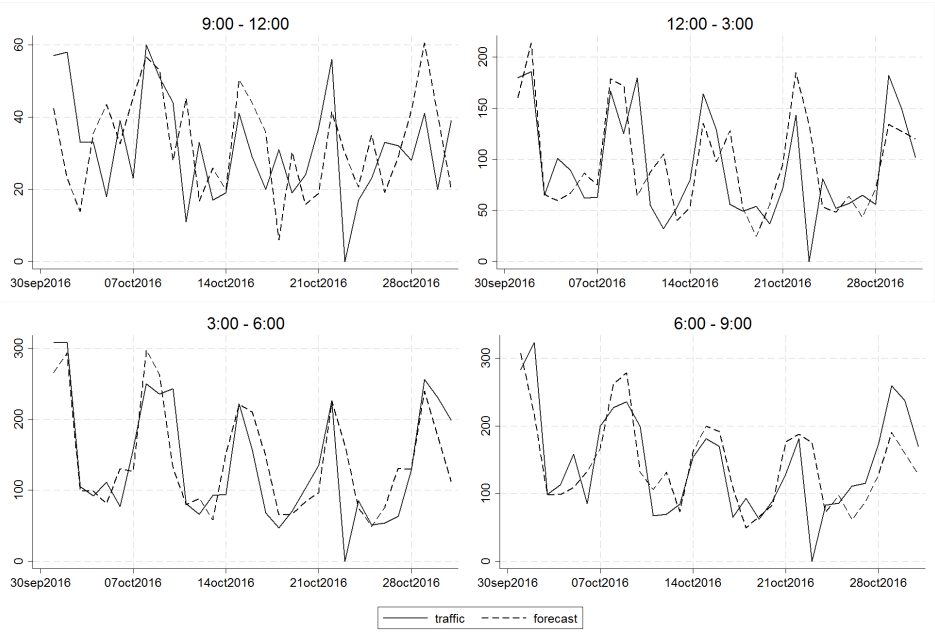

Figure 2.7 Traffic forecast accuracy. Out-of-sample traffic forecast vs. actual traffic for October 2016.

Different specifications of the ARIMA model			
	AR(3) MA(2)	AR(3) MA(1)	AR(2) MA(1)
Holiday	0.107	0.108	0.106
	(0.0697)	(0.0695)	(0.0684)
μ	0.0100	0.00998	0.0102
	(0.0522)	(0.0522)	(0.0519)
ARMA			
θ_1	0.548	1.093***	1.080***
	(0.581)	(0.0653)	(0.0593)
θ_2	0.494	-0.111*	-0.131**
	(0.591)	(0.0537)	(0.0458)
θ_3	-0.115	-0.0293	
	(0.0643)	(0.0435)	
ϕ_1	-0.273	-0.818***	-0.802***
	(0.585)	(0.0603)	(0.0481)
ϕ_2	-0.444		
	(0.449)		
$\sigma(\varepsilon)$	0.394***	0.394***	0.394***
	(0.00751)	(0.00752)	(0.00754)
N	808	808	800
AIC	803.3	802.3	800.8
BIC	840.8	835.1	828.9

Standard errors in parentheses * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 2.5 Estimates of the traffic forecast regression model.

2.5 Evaluating Ideal Staffing Levels

To provide some insights, we illustrate how the methodology can be used to compute the ideal staffing level for the representative store, and compare this to the actual staffing to evaluate whether the store is systematically over or understaffed for different hour-blocks of the week. The model is also used to compute the profit loss due to sub-optimal staffing levels.

The top panel of Table 2.6 shows the average traffic of the representative store during the period of June-August of 2016, for different hour-blocks. Margins are set at 55%, average basket value at $16,600CLP (25 USD) and hourly salary at $1,250 CLP (2 USD) (based at 10% above minimum wage in Chile). Using these parameters, we computed the optimal staffing level for each hour-block, and compare it to the average staffing level used in the store. The bottom panel of Table 2.6 shows this comparison, where the number on each cell

is the average staffing level that was observed in the data and the number in parenthesis is the difference with respect to the optimal staffing. A minus sign (-) represent over-staffed hour blocks and a plus sign (+) the under-staffed.

According to the projected ideal staffing levels suggested by the model, Table 2.6 reveals that the store is severely understaffed on Fridays and weekends after 3pm. On Saturday, which exhibits high traffic in the 12-3pm hour block, understaffing is extended to that hour block. Section 2.3 (Figure 2.3) revealed that actual staffing tends to be higher than planned during weekend afternoon/evening, suggesting that store managers are working around the schedule to compensate this understaffing. This accommodation is not enough, as suggested by the comparison of the actual staffing and the ideal staffing for these hour blocks.

Table 2.6 also suggests that, during weekdays, the 12-3pm period is overstaffed. This could explain why the actual labor during weekday 12-3pm hour block is lower than the planned labor. Overstaffing implies that the workers are idle and therefore they can take a longer lunch break or arrive late for the afternoon part-time shift.

What drives the mismatch between the actual staffing and the ideal staffing? Our conversation with retail managers suggest three underlying mechanisms. First, the economic value of increasing staffing levels is not straightforward to quantify and is usually evaluated in qualitative terms. Planned staffing levels are partially adjusted in response to predictable patterns of customer traffic, but our analysis – which is based on a quantitative measurement of the impact of staffing on revenues – reveals that this adjustment is insufficient. Second, managers tend to set staffing levels based on projections of sales, not traffic. If the stores are understaffed on weekends, there is sales potential that the managers over-look and they continue to systematically understaff the stores week after week. Instead, if the staffing requirements were based on traffic projections, planning can realize the unrealized sales potential of understaffed periods and correctly adjust labor requirements to capture this potential. This is the main reason why our company collaborator seeks to complement their traffic counter services with a workforce planning system.

Third, the overstaffing observed during the weekdays is in part caused by labor regulatory restrictions. The retail labor sector has experienced numerous changes in recent years which have restricted the use of flexible labor practices. Moreover, these changes have also increased the complexity of the workforce scheduling task. Labor regulations in Chile require a minimum number of hours per shift, a minimum of resting days during a week, a maximum of Sundays that an employee can work during a six month period, among other restrictions. With these constraints, the staffing problem cannot be decoupled to analyze each hour-block independently. This may explain why managers are not able to properly staff stores during low traffic periods. Motivated by this, the next section develops a mathematical program to solve for the optimal staffing level, balancing labor operating costs with contribution margin and complying with labor regulations.

Traffic	**Blocks**			
	9-12	**12-3**	**3-6**	**6-9**
Monday	25	76	120	123
Tuesday	37	78	108	115
Wednesday	33	80	115	126
Thursday	32	82	124	141
Friday	36	98	162	196
Saturday	59	209	308	287
Sunday	32	140	238	227

Staffing Levels	**Blocks**			
	9-12	**12-3**	**3-6**	**6-9**
Monday	1.6 [–0.6]	3.2 [–1.2]	3.2 [–0.2]	3.2 [–0.2]
Tuesday	1.8 [–0.8]	3.6 [–1.6]	3.3 [–0.3]	3.2 [–0.2]
Wednesday	1.6 [–0.6]	3.2 [–1.2]	3.4 [–0.4]	3.3 [–0.3]
Thursday	1.6 [–0.6]	3.2 [–1.2]	3.2 [–0.2]	3.1 [+0.9]
Friday	1.7 [–0.7]	3.2 [–0.2]	3.4 [+0.6]	3.5 [+2.5]
Saturday	2.5 [–1.5]	4.5 [+1.5]	4.7 [+2.3]	4.8 [+2.2]
Sunday	1.9 [–0.9]	3.4 [+0.6]	3.6 [+2.4]	3.8 [+2.2]

Table 2.6 Comparison of staffing levels. Actual staffing level versus optimal staffing level for a representative store during June-August 2016. Store traffic is shown in the top panel. The bottom panel shows the average staffing observed in the data, with the difference with the optimal staffing in parenthesis. A minus sign (-) indicates overstaffing, a plus sign (+) indicates understaffing.

2.6 Workforce Scheduling Accounting for Labor Regulation and Employee Preferences

The approach to optimize staffing developed in the previous section is unfeasible to implement in practice. To generate a staffing plan that is feasible, we develop an integer linear program to maximize store profit subject to constraints that capture regulatory restrictions and worker preferences for working shifts. Our method is based on previous work in workforce scheduling; Ernst et al. [2004] provides a comprehensive review of applications of workforce scheduling in different application domains. The traditional modeling approach in workforce scheduling is to solve for feasible schedules that comply with a pre-specified labor requirement; see Musliu et al. [2002] for a review. For example, in the context of a call center, a first step is to determine the minimum number of agents that are required to fulfill a certain service level agreement (e.g. average waiting time); in a second step, the scheduling algorithm is used to find a feasible shift schedule at minimum cost that fulfills the requirement.

However, in the retail context it is not clear how to specify the pre-specified labor requirement. In this context, Kabak et al. [2008] uses a variation of the model of Lam et al. [1998] to first determine a labor requirement based on a cost-benefit analysis that trade-offs labor costs with sales, similar to what is done in section 2.4. This labor requirement is then used as an input for a scheduling problem that seeks to meet the labor requirement subject to constraints. Although we could replicate the approach of Kabak et al. [2008], we found to

be more efficient to skip the first step and directly solve a math programming formulation that maximizes profit as the objective function, including revenues, margins and labor costs. We show that the proposed formulation can be solved quickly, in the order of minutes, and can therefore be used by a store manager to evaluate alternative workforce schedule plans interacting with the decision support tool.

The first step in workforce scheduling is to define working shifts that are feasible to implement. For this purpose, we define the set of possible shift types S, where each shift type is indexed by s. In the context of this application, the retail industry in Chile uses three types of shifts:

1. Full time (FT): Shift of five full days per week.
2. Part time, half-day shift (PT30): Typically implemented with a working shift combining 6 half days per week. The typical configuration for these type of shift are: (i) Three full days, (ii) Two full days and 2 afternoons, (iii) One full day and four afternoons and (iv) six afternoons.
3. Part time, full-day shift (PT10): Typically consist on working one full day per week.

These shift types have a minimum of half day, so it is reasonable to aggregate time blocks into the "morning block" (9am-3pm) and the "evening" block (3-9pm). Each shift type has a set of possible shift configurations, $C(s)$, where each shift in the set is indexed by i and specifies the time blocks assigned to the shift. Figure 2.9 provides some examples of shifts for the different types. The set of feasible shifts depends on labor restrictions, employee preferences and business policies.

To formulate the math programming formulation, it is useful to define the following parameters:

- $U_{\mathrm{d}tm}$: Expected sales in period t of day d with $E = m$ workers attending the store. This parameter is calculated directly from the estimates of the empirical model described in section 2.3.
- $W_{\mathrm{i}s}$: Wages associated to shift i of type s.
- $A_{\mathrm{i}sdt} : \begin{cases} 1 & \text{Shift i of type } s \text{ includes period } t \text{ in day d} \\ 0 & \sim \end{cases}$

The decision variables are defined as:

- $X_{\mathrm{i}s}$: Number of shifts i of type s
- $Z_{\mathrm{d}t}$: Number of employees on period t and day d
- $Y_{\mathrm{d}tm} : \begin{cases} 1 & Z_{\mathrm{d}t} = m \\ 0 & \sim \end{cases}$

The objective function is given by:

$$\max \text{Margin} \times \sum_{m=1}^{M} \sum_{t \in T} \sum_{\mathrm{d} \in D} Y_{\mathrm{d}tm} U_{\mathrm{d}tm} - \sum_{s \in S} \sum_{\mathrm{i} \in C(s)} X_{\mathrm{i}s} W_{\mathrm{i}s} \tag{2.8}$$

where M is the maximum number of employees that can be present in the store at any time, T is the set of periods (morning / evening), D is the set of days (Mon-Sun), and Margin is the contribution margin as a fraction of sales.

The following constraints link the decision variables:

$$\sum_{s \in S} \sum_{\mathrm{i} \in I(s)} X_{\mathrm{i}s} A_{\mathrm{i}sdt} = Z_{\mathrm{d}t} \qquad \forall \mathrm{d} \in D, t \in T \tag{2.9}$$

$$\sum_{n=1}^{N} n \times Y_{\mathrm{d}tn} = Z_{\mathrm{d}t} \qquad \forall \mathrm{d} \in D, t \in T \tag{2.10}$$

$$\sum_{n=1}^{N} Y_{\mathrm{d}tn} = 1 \qquad \forall \mathrm{d} \in D, t \in T \tag{2.11}$$

Retailers often want to limit the number of part-time employees and have a minimum of full-time employees; define the parameters Min_s and Max_s as the minimum and maximum quantity of type s shifts to hire. Two constraints are added to allow limits on the shift types:

$$\sum_{\mathrm{i} \in I(s)} X_{\mathrm{i}s} \geq Min_s \qquad \forall s \in S \tag{2.12}$$

$$\sum_{\mathrm{i} \in I(s)} X_{\mathrm{i}s} \leq Max_s \qquad \forall s \in S \tag{2.13}$$

Full-time employees typically prefer consecutive resting days in their schedule. Defining the parameter DC_{i} as indicator whether shift i has consecutive resting days, this requirement can be incorporated in the model with the following linear constraint:

$$(1 - DC_{\mathrm{i}}) X_{\mathrm{i}s} = 0 \qquad \forall \mathrm{i} \in C(s) \wedge s \in S_{FT} \tag{2.14}$$

Another important regulation in the Chilean legislation for retail labor is a restriction in the number of consecutive working Sundays for full-time employees. This constraint requires a rotation of full-time employees on each store. Define SUN_{i} as an indicator whether shift i includes a Sunday. The rotation of full-time employees can be achieved when the number of assigned shifts that work on Sunday is balanced with the same number of full-time shifts that exclude Sunday. This requirement is equivalent to the following linear constraint:

$$\sum_{s \in S_{FT}} \sum_{\mathrm{i} \in C(s)} X_{\mathrm{i}s} SUN_{\mathrm{i}} \leq \sum_{s \in S_{FT}} \sum_{\mathrm{i} \in I(s)} X_{\mathrm{i}s} (1 - SUN_{\mathrm{i}}) \tag{2.15}$$

The model was used to evaluate alternative scenarios with different sets of constraints, to analyze how reducing flexibility in the schedules affects profits. The base scenario is the ideal staffing level calculated in Section 2.4, where each hour block is optimized separately with no constraints. This base case is an upper bound that can be compared with each of the following scenarios which are solved by optimizing the objective (2.8) subject to different sets of constraints:

- Scenario 1: three shift types FT, PT30 and PT30, with the constraints 2.9, 2.10, 2.11.
- Scenario 2: similar to scenario 1 with the additional constraint of no consecutive Sundays (equation 2.15) and consecutive resting days (equation 2.14).
- Scenario 3: similar to scenario 2 but without PT10 shifts.
- Scenario 4: similar to scenario 2 but without PT30 shifts.
- Scenario 5: similar to scenario 2 but only with full-time shifts.

Results of the optimization are summarized in Figure 2.8, where the profits account for gross operating profits and labor costs, normalized to 100 for the base scenario. The profit loss of scenario 1 – the feasible schedule with three types of shifts – reduces profits by 1%. This is a relevant profit reduction for retail chains with a high fixed cost component in their cost structure. The additional regulatory restrictions included in scenarios 2 and 3 have a negligible effect on profits. Figure 2.9 illustrates the optimal solutions for scenarios 1 and 2. The optimal shift for scenario 1 assigns three employees to full-time shifts, working on Sundays with consecutive resting days. This solution is not feasible in scenario 2 because of the consecutive Sundays constraint (equation 2.15). To account for this constraint, the solution is adapted to have two rotating full-time shifts, one with free Sunday, and compensating with an additional PT30 and PT10 shift.

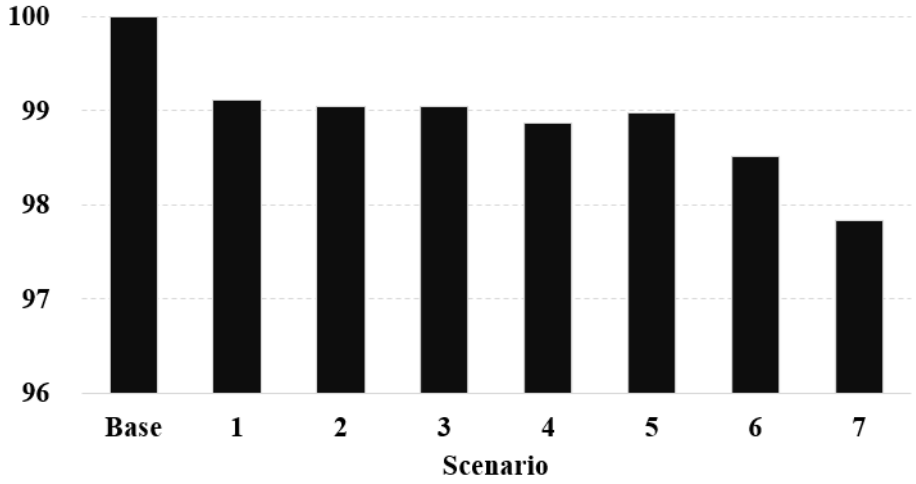

Figure 2.8 Workforce schedule optimization results across different scenarios. The base case corresponds to the ideal schedule calculated by optimizing each hour block separately, an upper bound for all scenarios. Profits are normalized to 100 for that scenario.

The last three bars in the chart of Figure 2.8 shows the impact of removing part-time shifts. Eliminating PT10 shifts has a small impact, because the optimal solution in scenario 2 uses three such shifts which can be accommodated by an an extra PT30 shift. But remov-

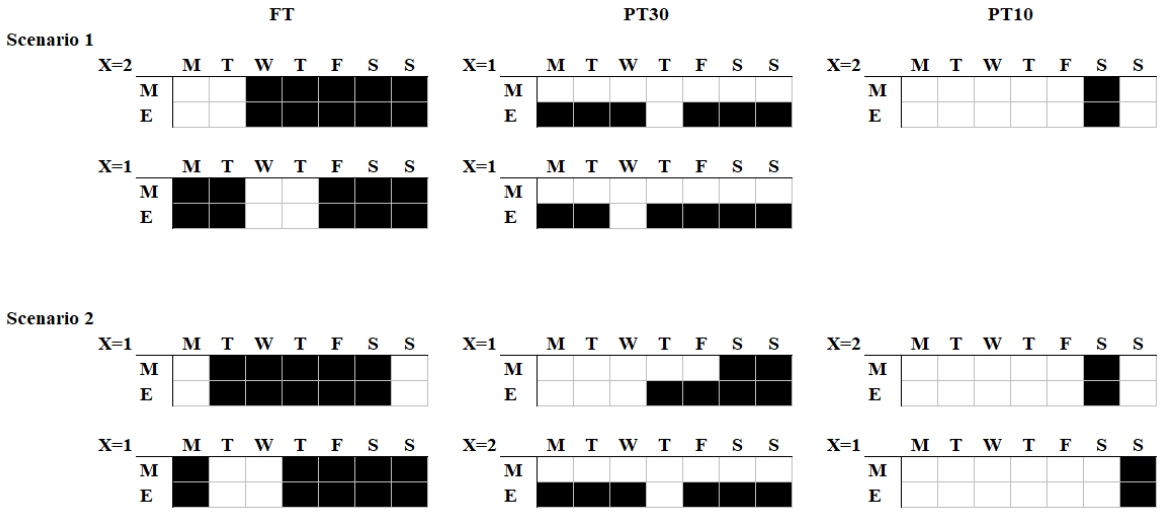

Figure 2.9 Optimal solution, model output. Feasible solution specifying the optimal working shifts for scenarios 1 and 2. Shifts are grouped by shift type, FT corresponding to full-time shift. PT30 is 30 hrs. part-time shift and PT10 is for 10 hrs. Black cells indicates the working hours in the corresponding shift, columns indicate day of the week and rows the Morning/Evening hours. The "X=" in the top left of each table indicates the number of workers assigned to that shift.

ing the PT30 shifts is more costly, dropping profits by an additional 0.5%: this shifts are valuable because they permit hiring part-time employees that work half-days, which is useful to accommodate the peaks in traffic during the evening hours (recall that PT10 shifts are required to work full days). The last scenario uses only full-time shifts, taking an extra drop in profit, with an accumulated loss of more than 2% relative to the base case.

The analysis of scenarios is specific to traffic patterns of a store and to the institutional regulation of the retail environment, and it is not always possible to generalize to different conditions. Nevertheless, the analysis above illustrates how the solution provided by the proposed decision support tool can be used to improve workforce planning and understand the implications of imposing restrictions on the workforce schedule. We have developed a prototype of this optimization tool, which can be operated interactively by a store manager to evaluate alternative scenarios. The optimization was implemented in Julia (Bezanson et al. [2017]) using Gurobi as an optimization package. We are currently working with our industry partner to deploy a full-scaled version of this prototype which is planned to be commercialized in 2019.

2.7 Conclusions

This practice-based research project develops a novel solution to build a decision support tool for workforce planning in the retail sector. The research is conducted in collaboration with a technology company that offers traffic counters and other technologies to monitor store operations, with the objective of broadening the set of services that the company can offer to its clients in Chile and Latin America. Workforce management is particularly challenging in Latin American countries, which have been recently adding more restrictions through labor regulations.

The developed solution combines ideas from different research streams in Operations Management. It uses recent developments in empirical OM research to evaluate the impact of staffing levels in sales. An important contribution in this domain is that we decompose the impact of staffing into conversion and basket value, which is useful in the context of workforce scheduling in order to measure this effect at a more granular level – in blocks of three hours – versus the aggregated models used previously in the literature which cannot be used directly to optimize working shifts. Other innovation is that we combine traffic information with deviations from the planned scheduled, showing how these are useful to mitigate endogeneity bias associated to managers staffing decisions.

We are currently working on the following improvements to the model and the decision support tool:

- The models developed so far to set staffing levels are based on a deterministic optimization problem. In reality, the uncertainty in the traffic forecast make the objective function stochastic. We plan to include this uncertainty into the model and use simulation-based optimization to generate staffing levels. The evaluation of the prescribed staffing levels is then calculated using simulation, to account for the impact of uncertainty in customer traffic.
- In our analysis, we evaluated the impact of some regulatory constraints that have been recently introduced in the Chilean labor legislation. We plan to evaluate the impact of additional labor restrictions in Chile, as well as other regulatory constraints in other countries. Many of the retail chains that operate in Chile also have presence in other countries in the region and therefore it is of interest to show the applicability of this labor planning tool with alternative labor regulations.
- In addition to labor laws, it is also important to account for employee preferences for schedules. We are planning to conduct a survey about employee preferences, to identify schedules that, from the employees perspective, are not desirable and should therefore be removed from the set of feasible shifts.

Chapter 3

Online Retailer

3.1 Introduction

For a customer of an online retailer, the shopping experience relates both to the physical products offered by the retailer and the services that enable and facilitate the purchase. Delivery is one of the most important service components for online retailers (Heim and Sinha [2001]) and delivery speed and reliability can have a significant impact on the retailer's performance as shown by Fisher et al. [2016] and Cui et al. [2018] respectively. Thus, logistics play a critical role for online vendors and can be a key competitive differentiator, so companies are enhancing purchase delivery options Mottl [2018].

While attention has been given to the optimization of the order fulfillment process, like the works by Xu et al. [2009] and Acimovic and Graves [2014] who develop decision support models for real-time order fulfillment decisions, there is not much empirical evidence in the literature on whether and the extent to which delivery quality matters to customers and shapes their shopping decisions.

Theoretical models rely on the assumption that higher quality increases demand and consumer satisfaction Kumar et al. [1997], though existing evidence is largely anecdotal rather than quantitative. Otim and Grover [2006] studies consumer loyalty on web-based services using survey data and concludes that post-purchase services positively influence consumer loyalty. Recent work by Fisher et al. [2016] quantifies the impact of speed, a delivery quality dimension, on demand while Cui et al. [2018] estimates the impact of both reliability and speed.

Our work studies the relation between order fulfillment, specifically delivery speed, and future customer behavior. It is similar in that sense to Rao et al. [2011]. Using one year of comprehensive transactional data from a large online apparel retailer, we quantify the impact of delivery quality over time on incidence and amount of future purchases, contributing to the growing literature in the field (Rust and Huang [2014] and Caro et al. [2019]). We found that retailer revenues can increase by 5.63% by improving delivery service quality. This increase originates from improvements in purchase incidence (4.91%) and basket size (0.69%).

The rest of this chapter is organized as follows. Section 3.2 describes the empirical context and the data used for this study. Section 3.3 presents the econometric formulation of customer behavior. Section 3.4 presents the results of the econometric models of purchase incidence and basket size. Section 3.5 uses the results from the empirical analysis to quantify the effect of delivery quality profits. Finally Section 3.6 concludes the study with a summary of the main results.

3.2 Empirical Setting

We obtained a full year of transactional data from a major online apparel retailer from India. This information was collected through the detailed transactions of each sales ticket, allowing us to compute detailed features of each transaction, such as: delivery time, discount, price paid, basket composition, destination city for each order, date and time the order was placed, processed and delivered, and a unique identifier for the customer. Our data also includes information on whether the goods where sold by the retailer or by a third party vendor. While in both cases the order is fulfilled by the online retailer, third party items are usually just in time inventory, which could potentially extend the delivery time of orders containing those items.

Our data comprises more than 7 million purchases by more than 2.4 million customers that were delivered to almost ten thousand locations in India, although it is important to note that 80% of all purchases where delivered to about a thousand different locations and 93% of orders to approximately $2,000$ locations, as can be seen in Figure 3.1. This suggests that we have enough data to compute delivery time distributions for each location for a subset of cities, which however cover the vast majority of purchases.

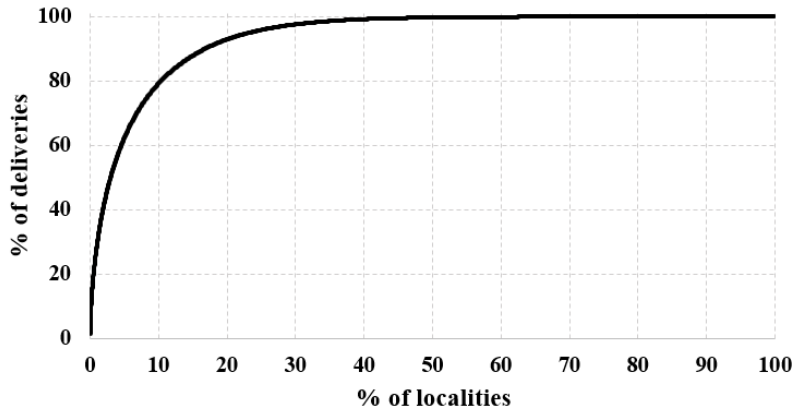

Figure 3.1 Distribution of deliveries per location. It's possible to see that almost all purchases are delivered to a relatively small subgroup of localities.

A detailed summary for the number of purchases made per client is presented in Table 3.1 and Figure 3.2. Most than half the customers made 1 or 2 purchases during the period covered by the data, which reduces our sample by more than half since we are studying future purchase behavior based on past experiences. On the other hand, some customers

made numerous purchases, 5% of customers made more than 15 purchases during this 1 year period, behavior we believe is more consistent with small retailers rather than household clients. In addition, Figure 3.3 shows that most than 80% customers had their purchases delivered to a single location.

Summary: Purchases per customer

Percentiles				Summary	
1%	1	**99%**	34	Max	1,301
5%	1	**95%**	15	Avg	4.5
10%	1	**90%**	10	Mode	2
25%	1	**75%**	5	Min	1

Table 3.1 Summary statistics: Purchases per customer.

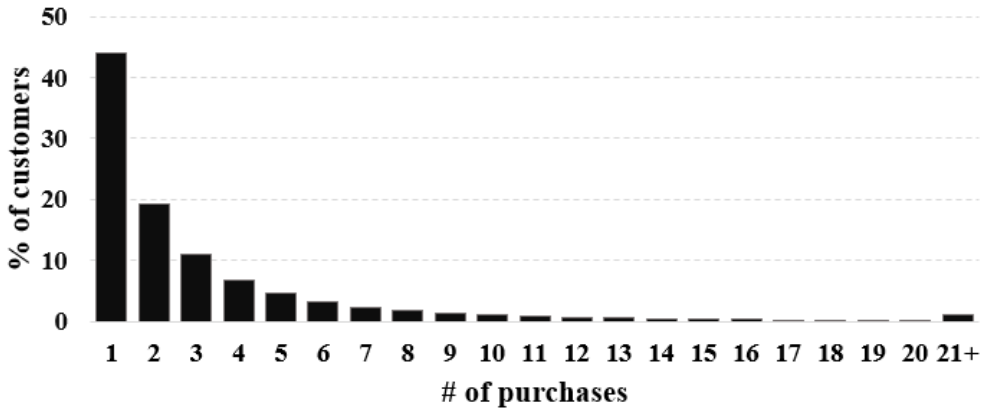

Figure 3.2 Purchases per customer. While almost 75% of active customers made 3 purchases or less, a small percentage ($\sim 1\%$) purchased in more than 20 occasions.

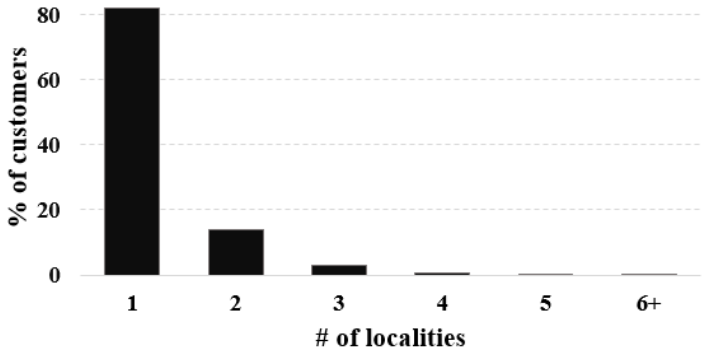

Figure 3.3 Locations per customer. More than 80% of customers ship to only one location.

3.2.1 Delivery Time Distribution

Figure 3.4 shows the empirical distribution of delivery time, computed as the number of days elapsed since the date the order is placed until the day it is delivered to the customer. Delivery usually takes no longer than a week and it's worth noting that only standard shipping was available to customers, since the retailer didn't offer any kind of expedited or premium shipping option.

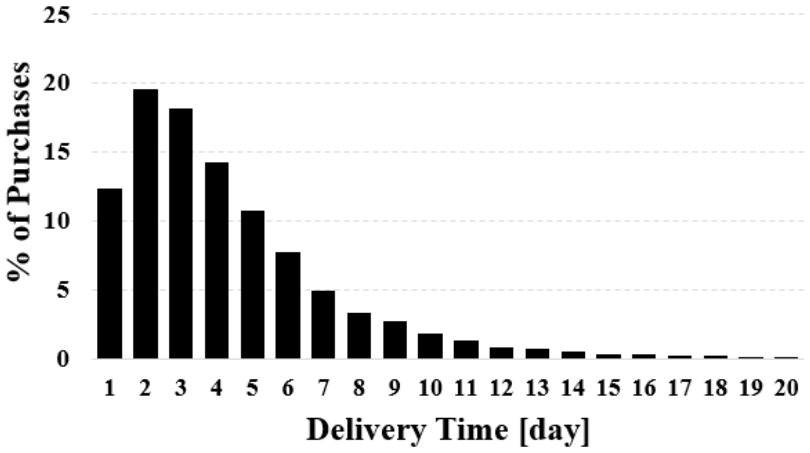

Figure 3.4 Delivery time. Distribution of delivery time, all locations and items.

Some items are procured just in time, typically those sold by third party vendors, while others are kept as on hand inventory. This could impact delivery speed, so we obtained the distribution of delivery times by supply type. Less than 5% of baskets had mix supply types, so we computed delivery time distributions using orders containing only just in time items and orders containing only on hand goods. It's not surprising to see that orders containing on hand items were delivered faster, around 2 days earlier, than orders for just in time procured products, as can be seen in Figure 3.5.

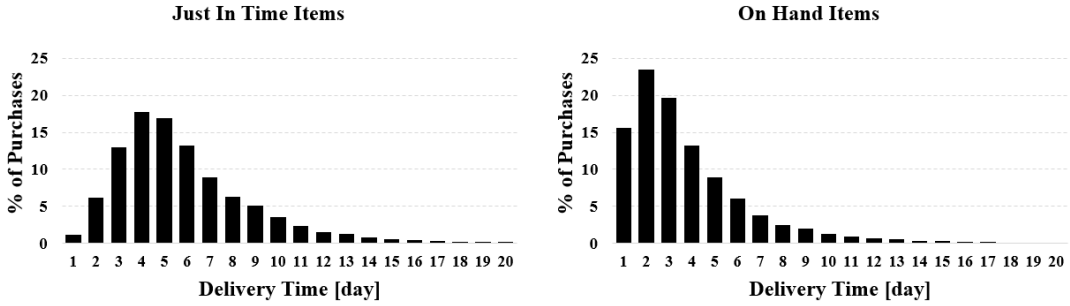

Figure 3.5 Delivery time by supply type. Distribution of delivery time, all locations by item supply type.

Shipping times could also vary by location and city characteristics. Figure 3.6 presents delivery time distributions by city tier[1]. It's clear from the figure that delivery time is shorter

[1]Tier 1 cities correspond to larger more populated areas than tier 2 cities, with tier 4 cities being smaller

for larger cities than it is for smaller towns.

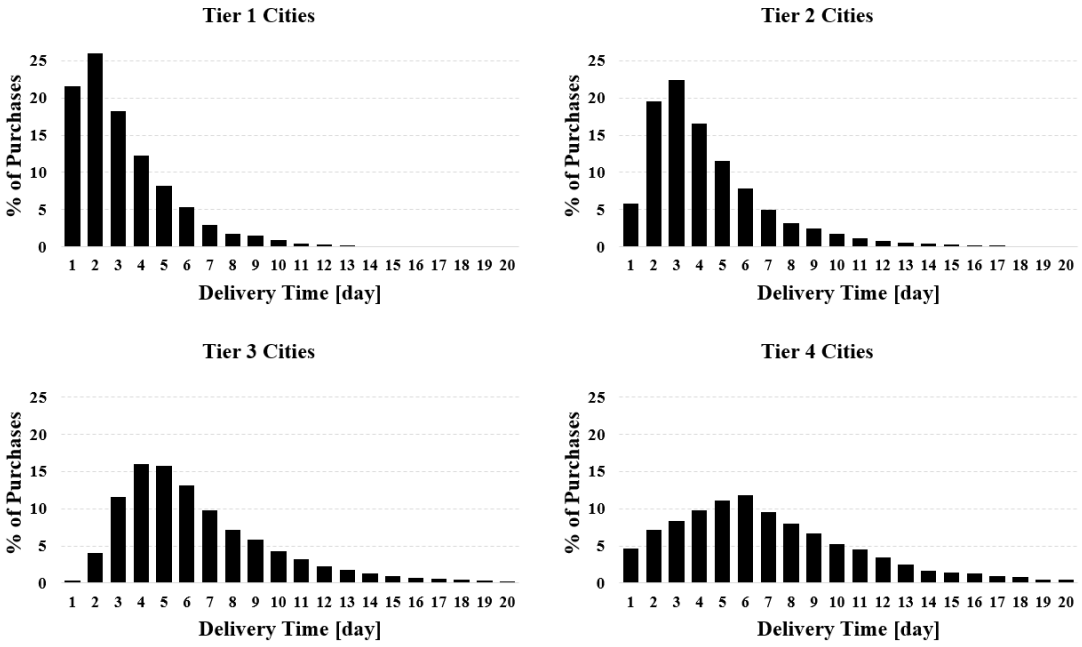

Figure 3.6 Delivery time by city tier. Distribution of delivery time, all items by city tier.

3.2.2 Delay Distribution

Another metric of delivery performance relates to promised delivery dates. At the time of making a purchase, the retailer informs the customer of an estimated delivery date which we'll refer to as "promised date". Delivery delay can be computed as the difference between delivery date and promised date – equation (3.1) – so d*elay* takes negative values for orders delivered earlier than anticipated.

$$\text{d}elay = \text{d}elivery\ \text{d}ate - promised\ \text{d}ate \tag{3.1}$$

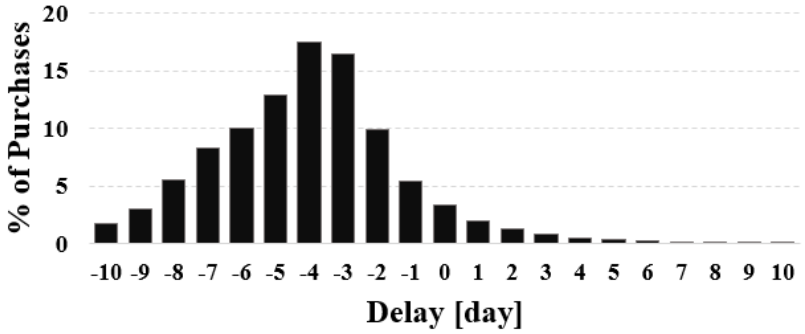

Figure 3.7 Delay. Distribution of delay in delivery, all locations and items.

and less populated zones.

Figures 3.7, 3.8 and 3.9, replicate the analysis from the previous section for the de*lay* metric. We can see that more than 90% of orders arrived before the promised date, while delays are observed for just under 6% of transactions. Customers experience more delays with just in time (8%) than with on-hand items (5%), they also face more uncertainty in delivery speed. A similar pattern is observed among city tiers, with bigger cities experiencing less variability in shipping times and suffering delays less frequently than smaller towns.

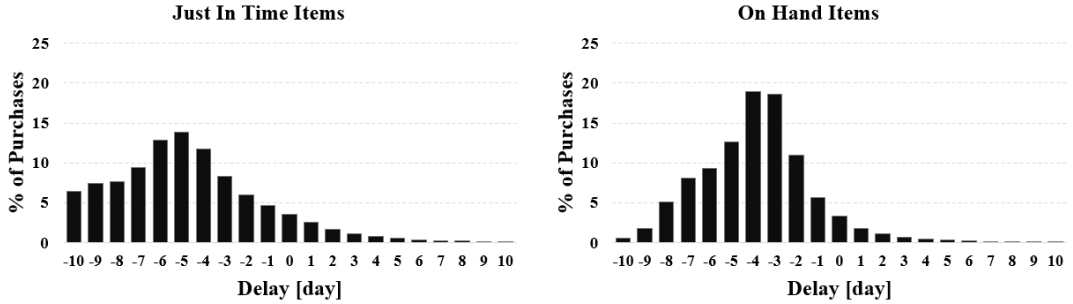

Figure 3.8 Delay by supply type. Distribution of delay in delivery, all locations by item supply type.

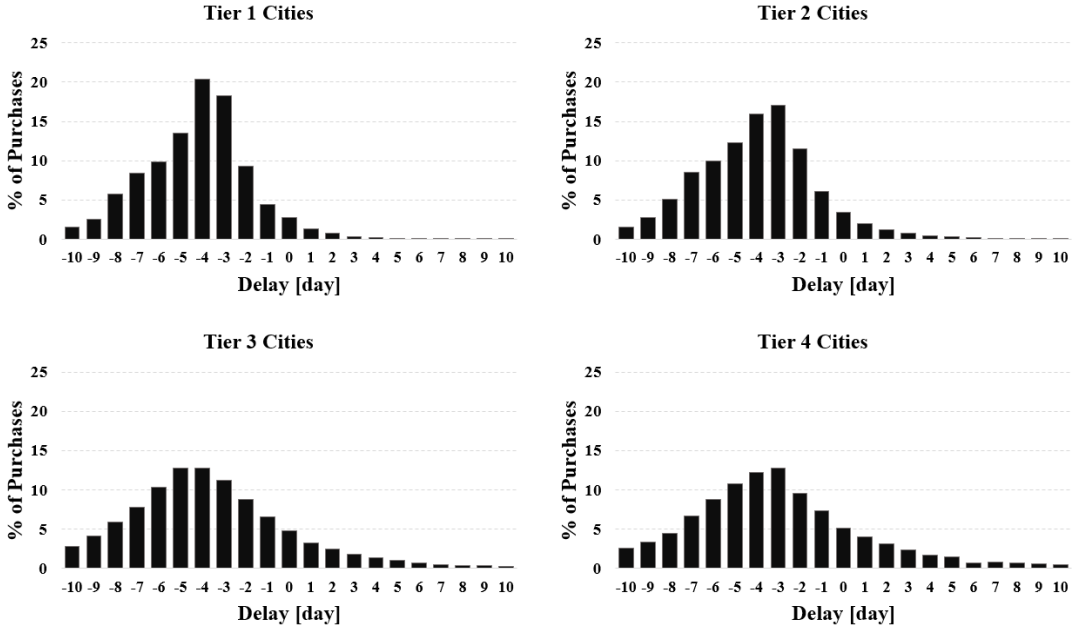

Figure 3.9 Delay by city tier. Distribution of delay in delivery , all items by city tier.

3.3 Model

Choosing which shipping and delivery options to make available to its customers requires us to understand consumer preferences, not only what they value, but how much they value it. Shipping and delivery is a post-purchase service that previous research has shown as having an influence on customer loyalty (Otim and Grover [2006]), sales (Fisher et al. [2016]; Cui et al. [2018]), company performance and competitiveness (Weise [2018]). This research

aims to quantify the effects of past delivery experiences on future shopping behavior in a large online apparel retailer. Our work is similar to Rao et al. [2011] since we both study order incidence and order size. Rao et al. [2011] uses a difference in difference approach in a setting where a subset of customers experienced a one time a failure to deliver an order within a previously promised time limit, while we model each customer's propensity to make a purchase during a series of discrete consecutive time periods. In that sense our model is dynamic because it considers the outcome of every delivery experienced by consumers in the past.

We follow an approach similar to Schweidel and Knox [2013] as we also model purchase incidence separately from order size introducing a correlation between both models. Customers' purchase history is introduced as a state dependence (Rosen [1981]) in both the purchase incidence and order size model, which are specified in detail in sections 3.3.1 and 3.3.2 respectively.

3.3.1 Purchase Incidence Model

We model customers' purchase decisions as a discrete-time Bernoulli process over two week periods with probability p_{it}.[2] We allow the purchase probability for customer i in period t to vary over time based on seasonality, promotion depth and customers' previous delivery experiences:

$$p_{\mathrm{i}t} = \Phi\left(\beta_{loc(\mathrm{i})} + X_t\gamma + \sum_{j=1}^{t-1}\left[\beta_{\mathrm{i}}^{exp} hasD_{\mathrm{i}j} + \beta_{\mathrm{i}}^{\mathrm{d}} D_{\mathrm{i}j} + \gamma_{bse} BSE_{\mathrm{i}j}\right]\mathrm{e}^{-\gamma_m j}\right) \tag{3.2}$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution, $\beta_{loc(\mathrm{i})}$ is a location-specific random effect and X_t is a vector of seasonal dummies and promotional depth variables. Covariates $hasD_{it}$ and D_{it} represent whether customer i received a delivery during period t and the quality of that delivery experience respectively (please refer to section 3.4.1 for a detailed description of each variable). State dependence, which can capture behavior changes due to past purchase behavior (Dubé et al. [2010]; Dubé et al. [2008]), is introduced in the summation term through $hasD_{it}$, which is also a proxy for a purchase being made, since most orders are received in the same period they are placed. The summation term accounts for customers' delivery experiences prior to period t. BSE_{it} accounts for the existence of a big sales event at the time items delivered in period t were purchased. β_{i}^{exp} and $\beta_{\mathrm{i}}^{\mathrm{d}}$ are customer level random effects, while $\mathrm{e}^{-\gamma_m}$ is a memory term allowing the impact of delivery on next period shopping behavior to potentially decay over time. Since our model is dynamic, care must be taken to ensure that the zero order heterogeneity is fully accounted for in order to distinguish it from time dynamics. Heckman [1981] suggests that ignoring heterogeneity could lead to a strong spurious state dependence, even when the actual choices are not correlated over time. Similarly, a model that accounts for heterogeneity, but ignores state dependence may overestimate the degree of heterogeneity (Keane [1997]).

[2]Our year long data is discretized in 26 biweekly periods since it is an interval long enough for most goods to be delivered in the same period they are purchased, while providing sufficient observations per customer without producing too many periods with no purchase.

3.3.2 Amount Model

Like Schweidel and Knox [2013], we model basket size (amount spent) conditional on a purchase being made in period t as a lognormal random variable with parameters μ_{it}^A and σ^A as specified in equation (3.3). Similarly to the purchase incidence model – equation (3.2) – we allow basket size to vary over time due to seasonality, promotion depth and customers' previous experiences. Models in equations (3.4) and (3.5) also account for previous purchase amounts ($A_{\mathrm{i}t}$) and the number of periods since the last purchase ($T_{\mathrm{i}t}$). This last variable allows for correlation between purchase incidence and basket size. A positive value for α_T indicates a customer will purchase a larger than expected amount after a larger than expected span without making a purchase. The inclusion of time since last purchase as a covariate enables the model to account for this compensating behavior, that if ignored can lead to poorer predictive performance and customer overvaluation (Jen et al. [2009]).

$$ln(A_{\mathrm{i}t}) \sim N(\mu_{it}^A, \sigma^A) \tag{3.3}$$

$$\mu_{it}^A = \alpha_{\mathrm{i}} + X_t\alpha_x + \alpha_T T_{\mathrm{i}t} + \alpha_{\mathrm{i}}^a A_{\mathrm{i},\tau(\mathrm{i},t)} + \alpha_{\mathrm{i}}^{\mathrm{d}} D_{\mathrm{i},\tau(\mathrm{i},t)} + \alpha_{bse} BSE_{\mathrm{i},\tau(\mathrm{i},t)} \tag{3.4}$$

The amount model in equation (3.5) allows the effect of delivery experience to potentially decay over time. If this is the case this model should provide a better fit than the model in equation (3.4).

$$\mu_{it}^A = \alpha_{\mathrm{i}} + X_t\alpha_x + \alpha_T T_{\mathrm{i}t} + \alpha_{\mathrm{i}}^a A_{\mathrm{i},\tau(\mathrm{i},t)} + \alpha_{\mathrm{i}}^{\mathrm{d}} \frac{D_{\mathrm{i},\tau(\mathrm{i},t)}}{T_{\mathrm{i}t}} + \alpha_{bse} BSE_{\mathrm{i},\tau(\mathrm{i},t)} \tag{3.5}$$

where $\tau(\mathrm{i},t)$ is the last period before t in which customer i made a purchase/received a delivery. The state dependence term in equation (3.4) is different from that in equation (3.2) because the purchase incidence model considers all past purchases, while the amount model only considers the last purchase previous to the focal period. This also holds for the covariates that account for past delivery experiences ($D_{\mathrm{i}t}$ and $BSE_{\mathrm{i}t}$). This difference is addressed in detail in section 3.4.3.

3.4 Empirical Analysis

We fit the models in equations (3.2), (3.3) and (3.4) using a hierarchical Bayesian formulation and MCMC estimation (Kruschke and Vanpaemel [2015], Gamerman and Lopes [2006] and Smith and Roberts [1993]). Our goal is to quantify the impact of delivery quality – mainly delivery speed – on future consumer behavior. We calibrate models using different specifications for delivery quality, and also estimate a model in which we don't take into account the impact of delivery quality on consumer behavior as a benchmark model.

3.4.1 Covariates

The main covariate, which is the focus of our research, is delivery quality (D_{it}). To calculate this covariate, we first compute the empirical distribution of delivery time conditional on location and inventory system. Then, for a given percentile P we define a delivery to produce a negative experience if it took more time to reach the customer than the P^{th} percentile of delivery time (conditional on location and inventory system). D_{it} then corresponds to the fraction of deliveries that generated a negative experience among deliveries received by customer i during period t. Since the value of D_{it} depends on P, from now on we'll make explicit this dependence using the notation D_{it}^{P}. Other covariates included in our models are specified below.

X_t : Vector containing a promotional depth variable and dummies for each quarter. Promotional depth was computed as the average discount of every purchase made during period t, using a different sample of customers than the one used to estimate the model. It is re-scaled so it's mean is equal to 0 and variance equal to 1.

A_{it} : Summation of the amount of all purchases made by customer i in period t.

D_{it}^{P} : Proportion of orders delivered to customer i during period t which had a delivery time greater than a threshold. Thresholds are specific to location and supply type, i.e. they were computed using the data corresponding to the delivery times for each location and supply type combination. Note that supply type can be either just in time or on hand. Also note that this covariate takes the value of 0 if the customer received no orders in the period.

BSE_{it} : Proportion of items delivered to customer i during period t purchased during a big sales event. A big sales event is defined as date in which the average promotional discount is greater than a threshold. The threshold is computed as the 70% percentile of the empirical distribution of average promotional discount, using a different sample of customers. This variable takes a value of 0 if the customer received no orders.

$hasD_{it}$: Dummy variable equal to 1 if customer i received a delivery during period t, and is equal to 0 otherwise. This variable is used to capture state dependence in the purchase incidence model and when equal to 1, it often coincides with a purchase being made during that period.

T_{it} : Number of periods (from t) since the last purchase for customer i.

3.4.2 Sample Selection

Our data allows us to construct a panel in which for a given customer we consider all periods *after* her first purchase. This generates an unbalanced panel. Models were calibrated using a sample of 500 customers randomly chosen among those who met the following criteria:

- Having made a purchase in 6 or more periods, guaranteeing at least 5 observations per customer. This criteria keeps 27.2% of transactions.
- Receive deliveries in only one location, which in turn receives at least 365 deliveries in total. This allows us to estimate a location based random effect and to compute delivery time distributions, while keeping 49.5% of the remaining transactions.
- Having made no more than 15 purchases during the year that comprises our data. The goal of this criterion is to exclude small vendors from the sample. This excludes 16.7% of remaining transactions.

It's worth noting that 11.2% of transactions are left after excluding transactions of customers who didn't meet the aforementioned criteria. In addition, we only considered transactions that constitute a new purchase occasion, thus excluding the ones labeled as exchanges or returns.

3.4.3 Model Estimation Results

Models were fitted using MCMC estimation implemented in the programming language Stan Carpenter et al. [2017] utilizing the NUTS sampler.

Purchase Incidence Model

The model in equation (3.2) is estimated as a Bayesian hierarchical model (Kim et al. [2002]; Manchanda et al. [1999]). Table 3.2 shows the Widely Applicable Information Criterion (Gelman et al. [2014] and Watanabe [2010]) and Leave-One-Out cross validation for three delivery quality covariate D_{it}^P specifications and three distributional assumptions for the delivery speed coefficient: an homogeneous specification in the first and fourth columns, a normally distributed random effect in the second and fifth columns and a random effect that is constrained to be negative[3] in the third and sixth columns.

	WAIC			**LOO**		
		Random Effect			**Random Effect**	
P	**Homogeneous**	**Normal**	**Lognormal**	**Homogeneous**	**Normal**	**Lognormal**
50	8,609.6	8,618.4	8,609.7	8,610.8	8,620.7	8,611.7
70	8,609.6	8,616.8	8,611.7	8,610.8	8,618.5	8,613.5
90	8,623.9	8,628.5	8,618.6	8,625.6	8,630.4	8,619.9
	No Delivery Quality			**No Delivery Quality**		
	8,615.6			8,616.9		

Table 3.2 Purchase incidence model selection. Model selection criteria for different delivery quality variable specifications and distributional assumptions for its coefficient. Lower values are indicative of a model that better fits the data. Last row corresponds to a model that doesn't account for delivery quality.

[3]For this purpose we model β_i^P to be distributed as the negative of a lognormal random variable.

Table 3.2 shows that specifications with a high threshold for qualifying a delivery experience as negative ($P = 90$) and/or a normal distributional prior for the delivery random effect are outperformed by a model specification that doesn't account for delivery quality. On the other hand, lower thresholds produce models that better fit the data, suggesting longer than average delivery times –conditional on location and inventory type– are perceived as low quality service, and not just exceptionally long shipping times. Previous research suggests customers value delivery quality (Cui et al. [2018]) and delivery speed in particular (Fisher et al. [2016] and Otim and Grover [2006]), serving as theoretical support for constraining $\beta_{\mathrm{i}}^{\mathrm{d}}$ to be negative, as is the case with the models listed as "Lognormal" in Table 3.2.

	mean	se mean	sd	2.5%	5%	95%	97.5%	n_{eff}	$\hat{R}$
β^{d}	-1.060	0.010	0.538	-2.386	-2.093	-0.383	-0.291	3119.1	1.001
μ_{exp}	1.918	0.015	0.785	0.845	0.941	3.427	3.888	2594.5	1.000
σ_{exp}	1.616	0.020	0.902	0.476	0.582	3.287	3.884	1938.1	1.000
$\mu_{loc(\mathrm{i})}$	-0.900	0.001	0.062	-1.021	-1.001	-0.797	-0.779	3754.2	1.000
$\sigma_{loc(\mathrm{i})}$	0.035	0.000	0.025	0.001	0.003	0.083	0.095	2713.2	1.000
γ_m	1.678	0.008	0.386	1.020	1.085	2.349	2.468	2368.3	1.000
γ_{bse}	-0.089	0.004	0.381	-0.851	-0.689	0.527	0.715	7195.3	1.000
$\gamma_{discount}$	0.089	0.000	0.016	0.057	0.062	0.115	0.121	10326.1	1.000
γ_{autumn}	-0.126	0.001	0.067	-0.254	-0.237	-0.016	0.005	4085.6	1.000
γ_{winter}	-0.017	0.001	0.064	-0.145	-0.123	0.088	0.108	4185.7	1.000
γ_{spring}	-0.145	0.001	0.065	-0.272	-0.253	-0.039	-0.020	3882.7	1.000

Table 3.3 Estimation results, purchase incidence model "HE". Model in equation (3.2), homogeneous coefficient for delivery variable. Model was calibrated using MCMC, 1,500 post warm-up iterations, 4 chains, $P = 50$.

	mean	se mean	sd	2.5%	5%	95%	97.5%	n_{eff}	$\hat{R}$
μ_{d}	-0.774	0.020	0.866	-2.827	-2.336	0.385	0.572	1844.9	1.002
σ_{d}	1.142	0.034	0.774	0.051	0.105	2.577	2.954	514.2	1.005
μ_{exp}	1.849	0.019	0.788	0.820	0.903	3.419	3.822	1641.5	1.004
σ_{exp}	1.482	0.024	0.868	0.408	0.499	3.217	3.740	1320.7	1.003
$\mu_{loc(\mathrm{i})}$	-0.899	0.002	0.062	-1.021	-1.000	-0.796	-0.777	1601.1	1.000
$\sigma_{loc(\mathrm{i})}$	0.036	0.001	0.026	0.002	0.003	0.085	0.097	1458.0	1.002
γ_m	1.639	0.010	0.392	0.984	1.054	2.360	2.487	1621.9	1.003
γ_{bse}	-0.106	0.005	0.364	-0.841	-0.678	0.502	0.667	4393.1	1.000
$\gamma_{discount}$	0.088	0.000	0.017	0.055	0.060	0.116	0.121	6147.0	1.000
γ_{autumn}	-0.128	0.002	0.066	-0.257	-0.237	-0.019	-0.001	1739.5	1.000
γ_{winter}	-0.018	0.002	0.064	-0.143	-0.122	0.088	0.110	1720.9	1.001
γ_{spring}	-0.147	0.002	0.066	-0.274	-0.254	-0.038	-0.015	1695.9	1.000

Table 3.4 Estimation results, purchase incidence model "RE". Model in equation (3.2), random effect for delivery variable, constrained to be negative. Model was calibrated using MCMC, 1,500 post warm-up iterations, 4 chains, $P = 50$.

Tables 3.3 and 3.4 present estimation results for both the homogeneous (HE) and negative lognormal random effects model (RE) which are the models that better fit the data. Priors for all model parameters are detailed in Table 3.10 in Appendix A. We chose delivery quality variable thresholds as defined by $P = 50$, since it's the specification that best fits the data. $\hat{\beta}^{\mathrm{d}}$ is negative and significant above 95%, which is consistent with previous research and supports the negative prior used for the RE model. Since the prior is a negative lognormal, estimated hyper-parameters $\hat{\mu}_{\mathrm{d}}$ and $\hat{\sigma}_{\mathrm{d}}$ produce a distribution with mean -0.885 and standard deviation 1.450 for $\beta_{\mathrm{i}}^{\mathrm{d}}$, which is consistent with the magnitude and sign of $\hat{\beta}_{\mathrm{d}}$ in the HE model. Both model specifications produce similar values for the rest of the estimated parameters. As expected, $\gamma_{\mathrm{dis}count}$ and γ_m are positive so greater discounts increase purchase incidence and the impact of past experiences decrease over time. The lognormal RE model also suggest an economically significant degree of heterogeneity among consumers in their reactions to delivery speed (or lack thereof). Figure 3.11 in the Appendix D shows estimated distributions for relevant parameters. The plots presented in the figure suggests that parameter estimation has converged after 6,000 post warm-up iterations considering all 4 chains. Although not reported, model specifications with $P = 70$ produce similar results to those from specifications with $P = 50$.

We also explored different specifications for delivery quality such as the gap between actual delivery and expected delivery date[4], a binary indicator on whether the shipment is delayed and absolute delivery time. We found no meaningful results using those alternate specifications. This is consistent with customers relying more on their past experience than retailer information regarding service quality expectation, which is documented in the literature, Ofir and Simonson [2007] results show that pre-purchase expectations are indistinguishable from evaluations of the store's past performance.

Amount Model

We first attempted a similar approach to account for customers' past history with the retailer in modeling basket size –measured in amount spent– according to equations (3.3) and (3.6). Covariates are the same as in equation (3.4) and $\mathrm{e}^{-\alpha m}$ modulates the impact of customer history on next period's shopping behavior as $\mathrm{e}^{-\gamma_m}$ does in equation (3.2).

$$\mu_{\mathrm{i}t}^{A} = \alpha_{\mathrm{i}} + X_t\alpha_x + \alpha_T T_{\mathrm{i}t} + \sum_{j=1}^{t-1}\left[\alpha_{\mathrm{i}}^{a} A_{\mathrm{i}j} + \alpha_{\mathrm{i}}^{\mathrm{d}} D_{\mathrm{i}j} + \alpha_{bse} BSE_{\mathrm{i}j}\right]\mathrm{e}^{-\alpha_m j} \tag{3.6}$$

The model was estimated using MCMC, with random effect coefficients for α_a, α_{i} and α_{d}. Results for $P = 70$ are presented in Table 3.5, with fit criteria statistics WAIC $= 4,077.6$ and LOO $= 4,089.4$. The estimated value of $\hat{\alpha}_m = 16.3$ suggests that previous history might not have a strong effect on amount spent once the decision to make a purchase is made. Even though the mean and standard deviation for α_{d} are -5.679 and 34.256 respectively[5] those values are modulated by $\mathrm{e}^{-\alpha_m}$ rendering the effect of the summation term on basket size minimal. The fact that this model implies the impact of history declines so rapidly over time, led us to consider a model which only accounts for the last purchase/delivery experience,

[4]Expected delivery date is informed to the customer at the time of purchase.

[5]α_{d} was estimated using a negative lognormal prior.

	mean	se mean	sd	2.5%	5%	95%	97.5%	n_{eff}	$\hat{R}$
μ_{i}	7.088	0.002	0.067	6.959	6.978	7.199	7.216	1786.0	1.001
σ_{i}	0.343	0.001	0.029	0.287	0.295	0.391	0.399	1444.2	1.002
μ_a	4.349	0.183	9.314	-15.629	-12.266	18.911	21.647	2577.0	1.000
σ_a	20.400	2.660	138.409	0.274	0.504	52.545	97.712	2707.9	1.001
μ_{d}	-0.031	0.027	1.967	-3.915	-3.334	3.305	3.923	5234.9	1.000
σ_{d}	1.918	0.035	1.688	0.061	0.113	5.201	6.146	2284.6	1.001
α_m	16.279	0.119	4.348	11.408	11.807	25.114	28.090	1330.5	1.001
α_T	-0.011	0.000	0.006	-0.022	-0.020	-0.001	0.002	2832.4	1.000
α_{bse}	-0.003	0.138	9.811	-19.152	-16.344	16.141	19.419	5048.9	1.000
$\alpha_{\mathrm{discount}}$	0.044	0.000	0.020	0.006	0.012	0.077	0.083	3518.8	1.001
α_{autumn}	0.051	0.002	0.074	-0.091	-0.069	0.172	0.195	2068.0	1.001
α_{winter}	0.049	0.002	0.070	-0.094	-0.068	0.163	0.184	1978.1	1.001
α_{spring}	0.112	0.002	0.072	-0.034	-0.009	0.232	0.255	2146.0	1.002
σ^A	0.707	0.000	0.015	0.678	0.683	0.732	0.738	2439.3	1.000

Table 3.5 Estimation results, basket size model "RE+T". Model in equations (3.3) and (3.6), random effect for delivery variable, constrained to be negative. Model was calibrated using MCMC, 1,500 post warm-up iterations, 4 chains, $P = 70$.

which is the model presented in equation 3.4 in section 3.3.2. Since we are considering only one period of history for each observation, we will omit the modulating factor from the model, assuming the last experience will have a lasting effect in customers choice of basket size. If past delivery history indeed has a negligible effect on customers choice, we would expect an estimated distribution for α_{d} heavily skewed towards zero. Table 3.6 shows the WAIC for three delivery quality covariate D_{it}^P specifications and three model configurations for the delivery speed and time since last purchase covariates: an homogeneous specification for both coefficients in the first and fourth columns (HE+T), a random effect for delivery and an homogeneous effect for T in the second and fifth columns (RE+T) and a random effect for a delivery covariate defined as $D_{\mathrm{it}}/T_{\mathrm{it}}$ and an homogeneous effect for T in the third and sixth columns (RE/T), this is the model defined by equations (3.3) and (3.5). Both models use a negative lognormal prior for the delivery random effect, other priors are normally distributed (3.11 in Appendix B).

Best fitting models are RE/T, which corresponds to equation (3.5), with $P = 70$ and the specification that accounts for inter-purchase time but omits delivery speed (ND-T). This suggests that the effect of past delivery experiences does decay over time, although this specification fits the data much better than the one defined by equation (3.6). Results for model specifications ND-T and RE/T are presented in Tables 3.7 and 3.8. In both settings there is not much heterogeneity in the intercept and previous basket size coefficients. Discount coefficient is positive and significant at 95%, meaning that promotions increase basket size. α_T is negative and statistically significant, suggesting that longer than usual inter-purchase times also reduce spending in the retailer's website, which may indicate that customers may be substituting some goods and purchasing from another vendor. This values are robust and remain virtually unchanged among all model specifications tested. Estimated values

	WAIC			LOO		
		Random Effect			Random Effect	
P	**HE+T**	**RE+T**	**RE/T**	**HE+T**	**RE+T**	**RE/T**
50	4,020.9	4,021.3	4,026.0	4,024.2	4,026.9	4,028.0
70	4,021.4	4,020.9	4,019.1	4,024.2	4,023.5	4,022.8
90	4,022.2	4,021.2	4,019.3	4,024.2	4,025.5	4,024.5

No Delivery Quality	**No Delivery Quality**
4,021.6	4,023.4

No Delivery Quality + T	**No Delivery Quality + T**
4,019.4	4,021.2

Table 3.6 Basket size model selection. Model selection criteria for different delivery quality variable and model specifications, lower values are indicative of a model that better fits the data. Last rows correspond to models that doesn't account for delivery quality.

$\hat{\mu}_{\mathrm{d}}$ and $\hat{\sigma}_{\mathrm{d}}$ in Table 3.8 produces a negative lognormal distribution with mean -0.000 and standard deviation 0.001, strongly hinting at the ND-T specification as the preferred one. Figure 3.12 in the Appendix E shows estimated distributions for relevant parameters. The plotted graphs in the figure suggest that parameter estimation has converged after 20,000 post warm-up iterations considering all 4 chains. We also calibrated the RE/T model with $P = 50$, obtaining similar results for all coefficients, in particular the biggest difference, as expected, is in the distribution of the delivery random coefficients, with an estimated mean of -0.001 and standard deviation of 0.009.

	mean	se mean	sd	2.5%	5%	95%	97.5%	n_{eff}	$\hat{R}$
μ_{i}	7.011	0.001	0.071	6.871	6.893	7.128	7.152	3375.2	1.001
σ_{i}	0.176	0.001	0.041	0.089	0.106	0.238	0.251	1642.7	1.006
μ_a	0.00007	0.000	0.000	0.00003	0.00004	0.0001	0.0001	5703.5	1.002
σ_a	0.00002	0.000	0.000	0.00000	0.00000	0.00005	0.00006	1755.6	1.003
α_T	-0.011	0.000	0.006	-0.022	-0.020	-0.002	0.000	14626.1	1.000
α_{bse}	-0.075	0.000	0.039	-0.153	-0.140	-0.010	0.003	10584.8	1.000
$\alpha_{\mathrm{discount}}$	0.043	0.000	0.020	0.004	0.010	0.075	0.082	10276.2	1.001
α_{autumn}	0.069	0.001	0.076	-0.080	-0.056	0.194	0.217	3614.7	1.001
α_{winter}	0.047	0.001	0.071	-0.095	-0.071	0.163	0.185	3420.1	1.000
α_{spring}	0.119	0.001	0.073	-0.027	-0.004	0.240	0.261	3454.0	1.000
σ^A	0.749	0.000	0.015	0.721	0.726	0.774	0.779	4222.3	1.002

Table 3.7 Estimation results, basket size model "ND-T". Model in equations (3.3) and (3.4), delivery variable coefficient constrained to be zero. Model was calibrated using MCMC, 5,000 post warm-up iterations, 4 chains.

	mean	se mean	sd	2.5%	5%	95%	97.5%	n_{eff}	$\hat{R}$
μ_{i}	7.013	0.001	0.071	6.873	6.895	7.130	7.152	2842.5	1.002
σ_{i}	0.177	0.001	0.041	0.089	0.107	0.239	0.251	1517.3	1.004
μ_a	0.00008	0.000	0.000	0.00002	0.00003	0.00011	0.00011	5966.1	1.000
σ_a	0.00002	0.000	0.000	0.00000	0.00001	0.00006	0.00006	1429.8	1.004
μ_{d}	-12.148	0.133	6.046	-25.825	-23.422	-3.837	-3.191	2080.8	1.003
σ_{d}	2.315	0.035	1.851	0.094	0.183	5.945	6.872	2779.2	1.001
α_T	-0.011	0.000	0.006	-0.022	-0.020	-0.002	-0.000	12780.4	1.000
α_{bse}	-0.075	0.000	0.040	-0.153	-0.140	-0.008	0.004	10173.4	1.000
α_{discount}	0.043	0.000	0.019	0.005	0.011	0.075	0.081	8695.6	1.001
α_{autumn}	0.068	0.001	0.075	-0.079	-0.056	0.192	0.217	3056.0	1.002
α_{winter}	0.046	0.001	0.070	-0.092	-0.070	0.161	0.184	2970.2	1.002
α_{spring}	0.117	0.001	0.072	-0.023	-0.000	0.235	0.257	2981.1	1.002
σ^A	0.749	0.000	0.015	0.721	0.725	0.773	0.778	4316.4	1.001

Table 3.8 Estimation results, basket size model "RE/T". Model in equations (3.3) and (3.4), random effect for delivery variable, constrained to be negative. Model was calibrated using MCMC, 5,000 post warm-up iterations, 4 chains, $P = 70$.

3.5 Quantification of Delivery Effect

We have found that negative delivery experiences, particularly when a delivery is in the top 30 or 50 percentile of larger delivery times (adjusted by location and inventory type), produce an adverse effect on the retailer performance by discouraging customers to purchase again. While we didn't find strong evidence of a direct impact of delivery quality on basket size, delivery performance can impact amount spent through purchase incidence (section 3.4.3).

In order to quantify the effect of delivery on both purchase incidence and basket size, we simulated four scenarios: i) 0%, a scenario in which every delivery experience is positive ($D = 0$); ii) 30%, a scenario in which the delivery outcome is simulated with a 30% rate of negative experiences; iii) 50%, a scenario in which the delivery outcome is simulated with a 50% rate of negative experiences; iv) 100%, a scenario in which every delivery experience is negative ($D = 1$). Each scenario was analyzed based on the estimated parameters of the HE model for purchase incidence (Table 3.3) and RE/T model for basket size (Table 3.8) using $P = 70$ (the simulation algorithm can be found in Appendix C).

Results are presented in Table 3.9 and Figure 3.10. Note that the bottom panel of the Table 3.9 and Figure 3.10 compare the 0%, 30% and 100% scenarios to a base scenario (50%). From this analysis we conclude that turning all delivery experiences to positive ones can increase revenues by 5.63% relative to a scenario in which about half of delivery experiences are satisfactory. On the other hand, a 4.55% drop in revenues is predicted if delivery speed is completely neglected. The impact of delivery on purchase incidence can be up to 4.91% relative to the 50% base scenario, which is much larger than the effect on purchase amount (up to 0.69%).

	Negative Delivery Rate			
	0%	30%	50%	100%
Purchase probability	0.1923	0.1867	0.1833	0.1754
Purchase amount	1,732.28	1,725.36	1,720.42	1,716.02
Impact of delivery on performance (Base scenario with 50% negative delivery rate)				
Purchase probability	4.91%	1.85%		-4.31%
Purchase amount	0.69%	0.29%		-0.26%
Combined effect	5.63%	2.15%		-4.55%

Table 3.9 Quantification, simulation results. Simulated impact of delivery on the retailer performance. Simulated scenarios corresponds to: i) 0%, a scenario in which every delivery experience is positive ($D = 0$); ii) 30%, a scenario in which the delivery outcome is simulated with a 30% rate of negative experiences; iii) 50%, a scenario in which the delivery outcome is simulated with a 50% rate of negative experiences; iv) 100%, a scenario in which every delivery experience is negative ($D = 1$). Simulated values where computed using estimated parameters of the HE model for purchase incidence (Table 3.3) and RE/T model for basket size (Table 3.8).

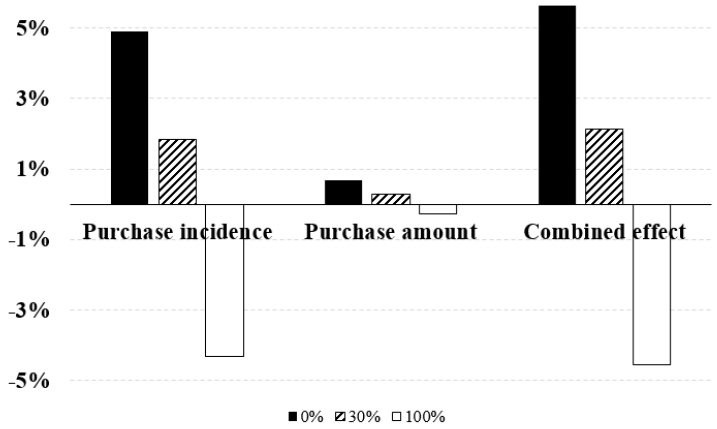

Figure 3.10 Quantification, impact of delivery quality. Effect of delivery quality on the retailer performance relative to a scenario in which half of the delivery instances produce a negative delivery experience.

3.6 Conclusions

We studied the relationship between service quality in the context of order fulfillment, specifically delivery speed, and future customer behavior. We focused on two aspects of future customer behavior: i) purchase incidence, and ii) purchase amount. We modeled purchase incidence in a given period as a hierarchical probit model with state dependence accounting for previous purchases, and found that negative delivery experiences produce an adverse effect on the retailer performance by discouraging customers to purchase again. We also found a small degree of heterogeneity in customers' future reactions to delivery quality service.

While we didn't find evidence of a direct impact of delivery performance on basket size, delivery performance can impact amount spent through other mechanisms, particularly purchase incidence. Our data suggests that longer than usual inter-purchase times can lead to smaller baskets, which is consistent with customers switching to other vendors to acquire products that otherwise would have been sourced from the retailer. Improving delivery speed can yield a 5.63% increase in revenues if all customers are satisfied with their order's fulfillment, steaming from a 4.91% increase in purchase incidence and a 0.69% increase in basket size.

Appendices

Appendix A: Purchase Incidence Model Priors

Parameters		Hyper-parameters	
β^{d}	$\sim N(0, 2^2)$		
$\beta_{\mathrm{i}}^{\mathrm{d}}$	$\sim$ -lognormal$(\mu_{\mathrm{d}}, \sigma_{\mathrm{d}}^2)$	μ_{d}	$\sim N(0, 2^2)$
		σ_{d}	$\sim$cauchy$(0, 2^2)$
β_{i}^{exp}	$\sim N(\mu_{exp}, \sigma_{exp}^2)$	μ_{exp}	$\sim N(0, 2^2)$
		σ_{exp}	$\sim$cauchy$(0, 2^2)$
$\beta_{loc(\mathrm{i})}$	$\sim N(\mu_{loc}, \sigma_{loc}^2)$	μ_{loc}	$\sim N(0, 2^2)$
		σ_{loc}	$\sim$cauchy$(0, 2^2)$
γ	$\sim N(0, 2^2)$		
$\gamma_{b\mathrm{se}}$	$\sim N(0, 2^2)$		
γ_m	$\sim N(0, 2^2)$		

Table 3.10 Priors: Purchase incidence model. β^{d} is the homogeneous parameter for the "HE" model estimated in Table 3.3, while $\beta_{\mathrm{i}}^{\mathrm{d}}$ is the negative lognormal random effect in the "RE" model estimated in Table 3.4. σ hyper-parameters use a cauchy prior truncated at zero, to ensure strictly positive values.

Appendix B: Basket Size Model Priors

Parameters		Hyper-parameters	
α_{i}^{c}	$\sim N(\mu_c, \sigma_c^2)$	μ_c	$\sim N(0, 10^2)$
		σ_c	$\sim$cauchy$(0, 10^2)$
α_{i}^{a}	$\sim N(\mu_a, \sigma_a^2)$	μ^a	$\sim N(0, 10^2)$
		σ_a	$\sim$cauchy$(0, 10^2)$
$\alpha_{\mathrm{i}}^{\mathrm{d}}$	$\sim$ -lognormal$(\mu_{\mathrm{d}}, \sigma_{\mathrm{d}}^2)$	μ_{d}	$\sim N(0, 2^2)$
		σ_{d}	$\sim$cauchy$(0, 2^2)$
α_m	$\sim N(0, 10^2)$		
α_T	$\sim N(0, 10^2)$		
$\alpha_b s\mathrm{e}$	$\sim N(0, 10^2)$		
α_x	$\sim N(0, 10^2)$		
σ^A	$\sim$cauchy$(0, 10^2)$		

Table 3.11 Priors: Basket size model. σ parameters/hyper-parameters use a cauchy prior truncated at zero, to ensure strictly positive values.

Appendix C: Simulation Algorithm

Algorithm 1: Simulation of shopping behaviour when all delivery experiences are positive

```
rD(negRate) := (negRate > Uniform())
for world = 1 : nWorlds do                    // loop through model parameter values
    pModel = Uniform(1, pIter × pChains)
    aModel = Uniform(1, aIter × aChains)
    for traj = 1 : nTrajectories do  // simulate random variable realizations
        row = 0
        for i = 1 : N do                                   // loop through customers
            Update D = rD(negRate), BSE, hasD = 1, A = A_0 and TSLP = 1
            for t = FirstPeriod(i):LastPeriod(i) do    // loop through periods
                P = Φ(f_p(D, BS, hasD, X, β(pModel)))
                if Uniform(0,1) < P then                          // purchase occurs
                    μ_A = f_A(D, BS, A, TSLP, X, β(aModel))
                    logA = μ_A + σ_A(aModel)×N(0,1)
                    A = exp(logA)
                    Update D = rD(negRate), BSE = avgDiscount_t >Threshold,
                     hasD = 1, TSLP = 1
                else                                          // no purchase occurs
                    A = 0
                    Update D = rD(negRate), BSE = 0, hasD = 0, TSLP += 1
                end if
                row += 1
            end for
        end for
    end for
end for
```

Appendix D: Estimated Distributions – Purchase Incidence

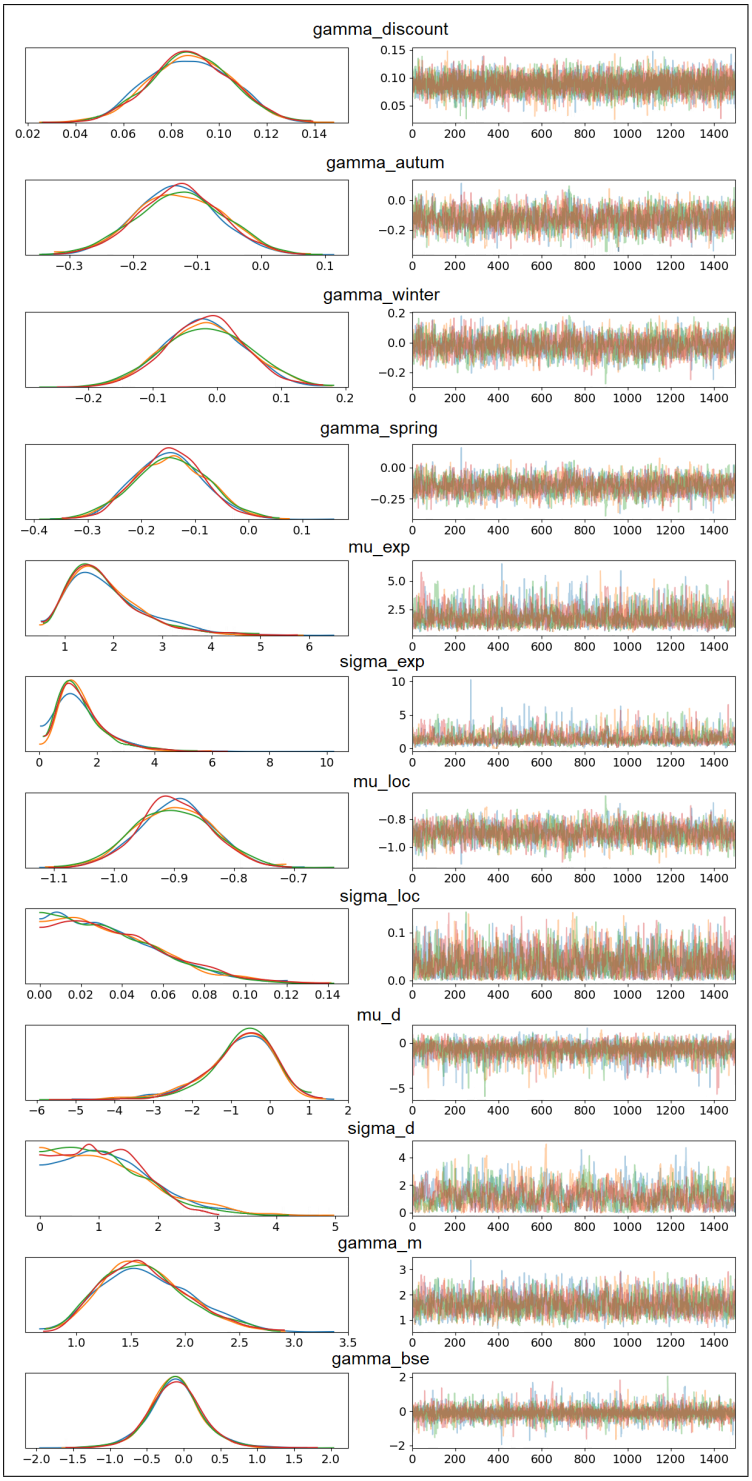

Figure 3.11 Estimated distributions, purchase incidence lognormal RE model.

Appendix E: Estimated Distributions – Basket Size

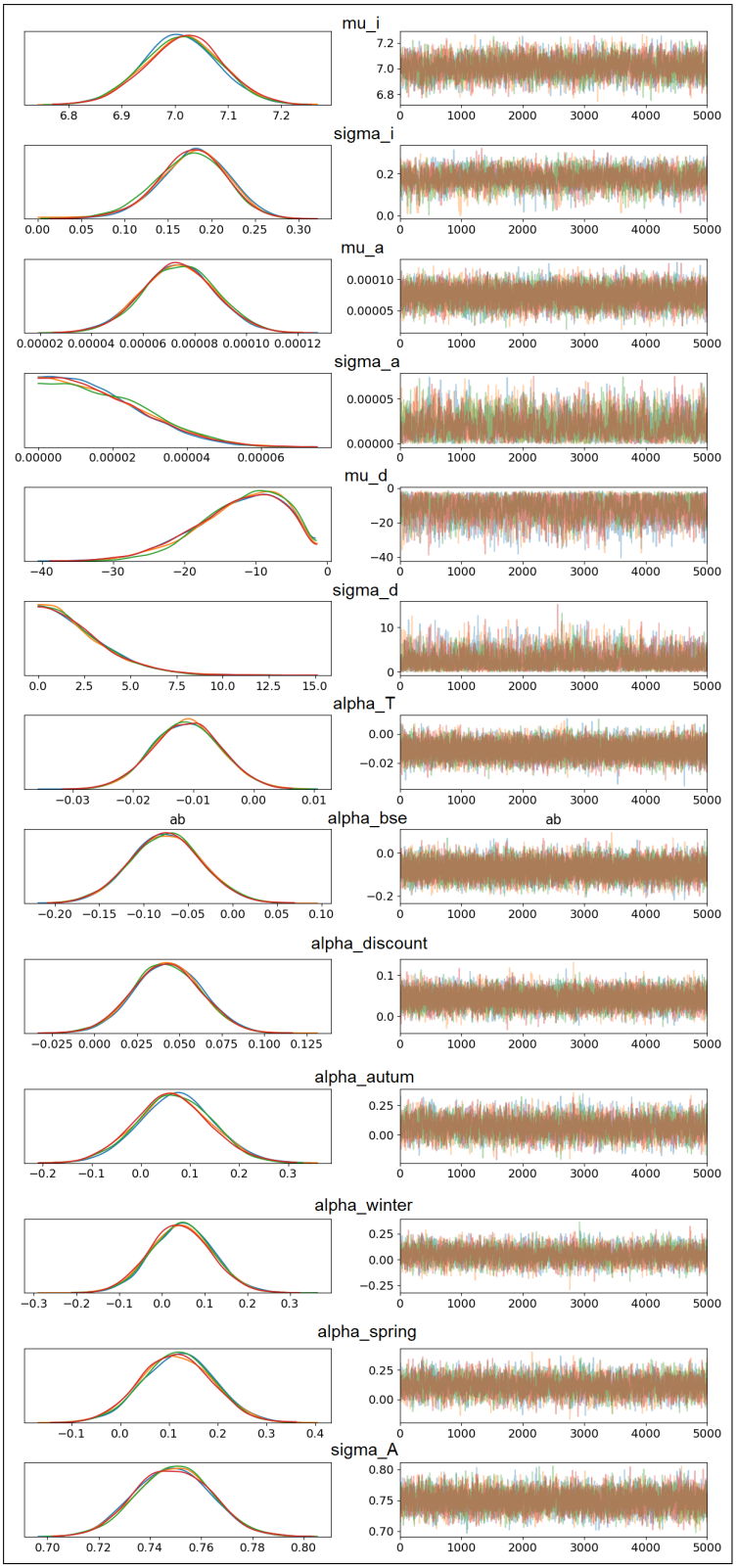

Figure 3.12 Estimated distributions, basket size RE/T model.

Conclusions

The three chapters that comprise this work address topics in management science deeply related to quality service. Chapters 1 and 3 address service quality through waiting time and quantify its effect on the online platform/retailer's revenue through a mechanism modulated by service quality. Chapters 1 and 2 tackle questions regarding personnel management accounting for the fundamental relation between human resources and customers. Studies in all three chapters found a statistical and economically significant effect of the quality of the service provided on business performance.

The work presented in chapter 1 shows the importance of empirically measuring the trade-off faced by the online platform when managing labor between increasing speed of service and providing a sufficiently attractive workload to its workers. For the outbound call center under study, we found that customers are indeed sensitive to the time elapsed between their visit to the platform's website and when they are reached by a platform agent. In particular, the probability that a customer answers a phone call from the platform drops from 62% to 52%, as this elapsed time increases from 3 to 20 minutes. In addition, conditional on successfully reaching a customer, the probability that a customer will purchase one of the insurance policies drops from 6% to 4% as the elapsed time increases from 5 to 90 minutes. Considering the role of agents, we also found that heterogeneity in worker productivity is relevant, since a 7.7% increase in utilization translates into a 0.9% to 3.6% increase in conversion rate due to higher retention of more skilled agents under medium traffic conditions.

Chapter 2 presents a practice-based research project that develops a novel solution to build a decision support tool for workforce planning in the retail sector. The developed solution combines ideas from different research streams in Operations Management. It uses recent developments in empirical OM research to evaluate the impact of staffing levels in sales. An important contribution in this domain is that decomposes the impact of staffing into conversion and basket value, which is useful in the context of workforce scheduling in order to measure this effect at a more granular level – in blocks of three hours – versus the aggregated models used previously in the literature which cannot be used directly to optimize working shifts. Other innovation is that we combine traffic information with deviations from the planned scheduled, showing how these are useful to mitigate endogeneity bias associated to managers staffing decisions. This kind of research would also be useful to economically evaluate the impact of labor policy on retail industry. The empirical results reveal a non-linear effect of traffic and staffing levels on sales. This effect was decomposed into conversion and basket value and found that most of the effect is through the former. The magnitude of the effect is comparable to previous results in the literature: increasing labor in under-staffed

stores can increase sales in the order of $2\% - 5\%$ (depending on the level of traffic of the store).

The study in chapter 3 investigates the relation between order fulfillment, specifically delivery speed, and future customer behavior. The work is focused on two aspects of future customer behaviour: i) purchase incidence, and ii) purchase amount. Purchase incidence was modeled as a hierarchical probit model with state dependence accounting for previous purchases, and found that negative delivery experiences produce an adverse effect on the retailer performance by discouraging customers to purchase again. We also found a small degree of heterogeneity in customers' future reactions to delivery quality service. While we didn't find evidence of a direct impact of delivery quality on basket size, delivery performance can impact amount spent through other mechanisms, particularly inter-purchase time. Our data suggests that longer than usual inter-purchase times can lead to smaller baskets, which is consistent with customers switching to other vendors to acquire products that otherwise would have been sourced from the retailer. Improving delivery speed can yield a 5.63% increase in revenues if all customers are satisfied with their order's fulfillment, stemming from a 4.91% increase in purchase incidence and a 0.69% increase in basket size.

Bibliography

[1] Jason Acimovic and Stephen C Graves. Making better fulfillment decisions on the fly in an online retail environment. *Manufacturing & Service Operations Management*, 17(1): 34–51, 2014.

[2] O Zeynep Aksin, Sarang Deo, Jónas Oddur Jónasson, and Kamalini Ramdas. Learning from many: Partner exposure and team familiarity in fluid teams. *Available at SSRN 2672133*, 2015.

[3] Zeynep Aksin, Mor Armony, and Vijay Mehrotra. The modern call center: A multi-disciplinary perspective on operations management research. *Production and operations management*, 16(6):665–688, 2007.

[4] Zeynep Aksin, Baris Ata, Seyed Morteza Emadi, and Che-Lin Su. Structural estimation of callers' delay sensitivity in call centers. *Management Science*, 59(12):2727–2746, 2013.

[5] Zeynep Akşin, Baris Ata, Seyed Morteza Emadi, and Che-Lin Su. Impact of delay announcements in call centers: An empirical approach. *Operations Research*, 65(1): 242–265, 2016.

[6] Gad Allon, Achal Bassamboo, and Itai Gurvich. "we will be right with you": Managing customer expectations with vague promises and cheap talk. *Operations Research*, 59 (6):1382–1394, 2011. doi: 10.1287/opre.1110.0976. URL `https://doi.org/10.1287/opre.1110.0976`.

[7] Gad Allon, Awi Federgruen, and Margaret Pierson. How much is a reduction of your customers' wait worth? an empirical study of the fast-food drive-thru industry based on structural estimation methods. *Manufacturing & Service Operations Management*, 13(4):489–507, 2011.

[8] Gad Allon, Maxime Cohen, and Wichinpong Sinchaisri. The impact of behavioral and economic drivers on gig economy workers. SSRN Working paper, 2018.

[9] Alessandro Arlotto, Stephen E Chick, and Noah Gans. Optimal hiring and retention policies for heterogeneous workers who learn. *Management Science*, 60(1):110–129, 2013.

[10] Robert J Batt and Christian Terwiesch. Waiting patiently: An empirical study of queue abandonment in an emergency department. *Management Science*, 61(1):39–59, 2015.

[11] Jeff Bezanson, Alan Edelman, Stefan Karpinski, and Viral B Shah. Julia: A fresh approach to numerical computing. *SIAM review*, 59(1):65–98, 2017.

[12] Gérard Cachon and Christian Terwiesch. *Matching supply with demand*, volume 2. McGraw-Hill Singapore, 2009.

[13] Gerard Cachon, Kaitlin M. Daniels, and Ruben Lobel. The role of surge pricing on a service platform with self-scheduling capacity. *Manufacturing & Service Operations Management*, 19(3):368–384, 2017.

[14] Felipe Caro, A Gürhan Kök, and Victor Martínez-de Albéniz. The future of retail operations. *Manufacturing & Service Operations Management*, 2019.

[15] Bob Carpenter, Andrew Gelman, Matthew D Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A probabilistic programming language. *Journal of statistical software*, 76(1), 2017.

[16] II Carson, David M Nicol, Barry L Nelson, Jerry Banks, et al. Discrete-event system simulation, 2005.

[17] M Keith Chen, Judith A Chevalier, Peter E Rossi, and Emily Oehlsen. The value of flexible work: Evidence from uber drivers. Technical report, National Bureau of Economic Research, 2017.

[18] Howard Hao-Chun Chuang, Rogelio Oliva, and Olga Perdikaki. Traffic-based labor planning in retail stores. *Production and Operations Management*, 25(1):96–113, 2016.

[19] Ruomeng Cui, Meng Li, and Qiang Li. Value of high-quality logistics: Evidence from a clash between sf express and alibaba. *Management Science, Forthcoming*, 2018.

[20] Suzanne De Treville and Ann Van Ackere. Equipping students to reduce lead times: The role of queuing-theory-based modeling. *Interfaces*, 36(2):165–173, 2006.

[21] Jing Dong, Pengyi Shi, Fanyin Zheng, and Xin Jin. Capacity management in inpatient wards with off-service placement and a network view. *Available at SSRN 3306853*, 2018.

[22] Jean-Pierre Dubé, Günter J Hitsch, Peter E Rossi, and Maria Ana Vitorino. Category pricing with state-dependent utility. *Marketing Science*, 27(3):417–429, 2008.

[23] Jean-Pierre Dubé, Günter J Hitsch, and Peter E Rossi. State dependence and alternative explanations for consumer inertia. *The RAND Journal of Economics*, 41(3):417–445, 2010.

[24] Seyed M. Emadi and Bradley R. Staats. A structural estimation approach to agent attrition. Working paper, 2018.

[25] Andreas T Ernst, Houyuan Jiang, Mohan Krishnamoorthy, and David Sier. Staff scheduling and rostering: A review of applications, methods and models. *European journal of operational research*, 153(1):3–27, 2004.

[26] Marshall Fisher, Santiago Gallino, and Joseph Xu. The value of rapid delivery in online retailing. *Available at SSRN 2573069*, 2016.

[27] Marshall Fisher, Santiago Gallino, and Serguei Netessine. Setting retail staffing levels: A methodology validated with implementation. *Working paper*, 2017.

[28] Marshall L Fisher, Jayanth Krishnan, and Serguei Netessine. Retail store execution: An empirical study. *University of Pennsylvania, the Wharton School and Research Center: Operations and Information Management Department. Retrieved December*, 1: 2006, 2006.

[29] Michael Freeman, Nicos Savva, and Stefan Scholtes. Gatekeepers at work: An empirical analysis of a maternity unit. *Management Science*, 63(10):3147–3167, 2017. doi: 10.1287/mnsc.2016.2512.

[30] Dani Gamerman and Hedibert F Lopes. *Markov chain Monte Carlo: stochastic simulation for Bayesian inference*. Chapman and Hall/CRC, 2006.

[31] Noah Gans, Ger Koole, and Avishai Mandelbaum. Telephone call centers: Tutorial, review, and research prospects. *Manufacturing & Service Operations Management*, 5 (2):79–141, 2003.

[32] Noah Gans, Nan Liu, Avishai Mandelbaum, Haipeng Shen, Han Ye, et al. Service times in call centers: Agent heterogeneity and learning with some operational consequences. In *Borrowing strength: theory powering applications–A Festschrift for Lawrence D. Brown*, pages 99–123. Institute of Mathematical Statistics, 2010.

[33] Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin. Bayesian data analysis (vol. 2). *Boca Raton, FL: Chapman*, 2014.

[34] Rodger W Griffeth, Peter W Hom, and Stefan Gaertner. A meta-analysis of antecedents and correlates of employee turnover: Update, moderator tests, and research implications for the next millennium. *Journal of management*, 26(3):463–488, 2000.

[35] Donald Gross. *Fundamentals of queueing theory*. John Wiley & Sons, 2008.

[36] Jinyong Hahn, Jerry Hausman, and Guido Kuersteiner. Estimation with weak instruments: Accuracy of higher-order bias and mse approximations. *The Econometrics Journal*, 7(1):272–306, 2004.

[37] Brett A. Hathaway, Seyed M. Emadi, and Vinayak Deshpande. Don't call us, we'll call you: An empirical study of callers' behavior under a callback option. Working paper, 2018.

[38] James J Heckman. Heterogeneity and state dependence. In *Studies in labor markets*, pages 91–140. University of Chicago Press, 1981.

[39] Gregory R Heim and Kingshuk K Sinha. Operational drivers of customer loyalty in electronic retailing: An empirical analysis of electronic food retailers. *Manufacturing &*

Service Operations Management, 3(3):264–271, 2001.

[40] Robert S Huckman and Gary P Pisano. The firm specificity of individual performance: Evidence from cardiac surgery. *Management Science*, 52(4):473–488, 2006.

[41] Robert S Huckman and Bradley R Staats. Fluid tasks and fluid teams: The impact of diversity in experience and team familiarity on team performance. *Manufacturing & Service Operations Management*, 13(3):310–328, 2011.

[42] Robert S Huckman, Bradley R Staats, and David M Upton. Team familiarity, role experience, and performance: Evidence from indian software services. *Management science*, 55(1):85–100, 2009.

[43] Lichung Jen, Chien-Heng Chou, and Greg M Allenby. The importance of modeling temporal dependence of timing and quantity in direct marketing. *Journal of marketing research*, 46(4):482–493, 2009.

[44] Özgür Kabak, Füsun Ülengin, Emel Aktaş, Şule Önsel, and Y Ilker Topcu. Efficient shift scheduling in the retail sector through two-stage optimization. *European Journal of Operational Research*, 184(1):76–90, 2008.

[45] Diwas S. KC. Does multitasking improve performance? evidence from the emergency department. *Manufacturing & Service Operations Management*, 16(2):168–183, 2013.

[46] Diwas S. KC and Christian Terwiesch. Impact of workload on service time and patient safety: An econometric analysis of hospital operations. *Management Science*, 55(9): 1486–1498, 2009.

[47] Diwas Singh Kc and Bradley R Staats. Accumulating a portfolio of experience: The effect of focal and related experience on surgeon performance. *Manufacturing & Service Operations Management*, 14(4):618–633, 2012.

[48] Michael P Keane. Modeling heterogeneity and state dependence in consumer choice behavior. *Journal of Business & Economic Statistics*, 15(3):310–327, 1997.

[49] Jaehwan Kim, Greg M Allenby, and Peter E Rossi. Modeling consumer demand for variety. *Marketing Science*, 21(3):229–250, 2002.

[50] Song-Hee Kim, Carri W Chan, Marcelo Olivares, and Gabriel Escobar. Icu admission control: An empirical study of capacity allocation and its implication for patient outcomes. *Management Science*, 61(1):19–38, 2014.

[51] John K Kruschke and Wolf Vanpaemel. Bayesian estimation in hierarchical models. *The Oxford handbook of computational and mathematical psychology*, pages 279–299, 2015.

[52] Vidyadhar G Kulkarni. *Modeling and analysis of stochastic systems*. Chapman and Hall/CRC, 2016.

[53] Piyush Kumar, Manohar U Kalwani, and Maqbool Dada. The impact of waiting time

guarantees on customers' waiting experiences. *Marketing science*, 16(4):295–314, 1997.

[54] Shunyin Lam, Mark Vandenbosch, and Michael Pearce. Retail sales force scheduling based on store traffic forecasting. *Journal of Retailing*, 74(1):61–88, 1998.

[55] Michael A Lapré. Reducing customer dissatisfaction: How important is learning to reduce service failure? *Production and Operations Management*, 20(4):491–507, 2011.

[56] Michael A Lapré and Nikos Tsikriktsis. Organizational learning curves for customer dissatisfaction: Heterogeneity across airlines. *Management science*, 52(3):352–366, 2006.

[57] Michael A Lapré and Luk N Van Wassenhove. Learning across lines. *Science*, 1998.

[58] Michael A Lapré, Amit Shankar Mukherjee, and Luk N Van Wassenhove. Behind the learning curve: Linking learning activities to waste reduction. *Management Science*, 46 (5):597–611, 2000.

[59] Yina Lu, Andrés Musalem, Marcelo Olivares, and Ariel Schilkrut. Measuring the effect of queues on customer purchases. *Management Science*, 59(8):1743–1763, 2013.

[60] Puneet Manchanda, Asim Ansari, and Sunil Gupta. The "shopping basket": A model for multicategory purchase incidence decisions. *Marketing science*, 18(2):95–114, 1999.

[61] Vidya Mani, Saravanan Kesavan, and Jayashankar M Swaminathan. Estimating the impact of understaffing on sales and profitability in retail stores. *Production and Operations Management*, 24(2):201–218, 2015.

[62] Alexandre Mas and Amanda Pallais. Valuing alternative work arrangements. *American Economic Review*, 107(12):3722–59, 2017.

[63] S. Misra, E.J. Pinker, and R.A. Shumsky. Salesforce design with experience-based learning. *IIE Transactions*, 36:941–952, 2004.

[64] Judy Mottl. Why delivery is playing a starring role in the retail customer experience. `https://www.retailcustomerexperience.com/blogs/why-delivery-is-playing-a-starring-role-in-the-retail-customer-experience`, 2018. Accessed: 2019-10-15.

[65] Nysret Musliu, Johannes Gärtner, and Wolfgang Slany. Efficient generation of rotating workforce schedules. *Discrete Applied Mathematics*, 118(1-2):85–98, 2002.

[66] Chezy Ofir and Itamar Simonson. The effect of stating expectations on customer satisfaction and shopping experience. *Journal of Marketing Research (JMR)*, 44(1), 2007.

[67] Samuel Otim and Varun Grover. An empirical study on web-based services and customer loyalty. *European Journal of Information Systems*, 15(6):527–541, 2006.

[68] Olga Perdikaki, Saravanan Kesavan, and Jayashankar M Swaminathan. Effect of traffic on sales and conversion rates of retail stores. *Manufacturing & Service Operations*

Management, 14(1):145–162, 2012.

[69] Robert Phillips, A Serdar Şimşek, and Garrett Van Ryzin. The effectiveness of field price discretion: Empirical evidence from auto lending. *Management Science*, 61(8): 1741–1759, 2015.

[70] I.P.L. Png and D. Reitman. Service time competition. *The Rand Journal of Economics*, 25(4):619–634, 1994. ISSN 0741-6261.

[71] K. Ramdas, K. Saleh, S. Stern, and H. Liu. Variety and experience: learning and forgetting in the use of surgical devices. *Management Science*, 64(6):2590–2608, 2018.

[72] Shashank Rao, Stanley E Griffis, and Thomas J Goldsby. Failure to deliver? linking online order fulfillment glitches with future purchase behavior. *Journal of Operations Management*, 29(7-8):692–703, 2011.

[73] Sherwin Rosen. *Studies in Labor Markets*. University of Chicago Press, 1981. URL `http://www.nber.org/books/rose81-1`.

[74] Sheldon M Ross et al. *Stochastic processes*, volume 2. John Wiley & Sons New York, 1996.

[75] Roland T Rust and Ming-Hui Huang. The service revolution and the transformation of marketing science. *Marketing Science*, 33(2):206–221, 2014.

[76] Jeffrey K Sager, Charles M Futrell, and Rajan Varadarajan. Exploring salesperson turnover: A causal model. *Journal of Business Research*, 18(4):303–326, 1989.

[77] Douglas A Samuelson. Predictive dialing for outbound telephone call centers. *Interfaces*, 29(5):66–81, 1999.

[78] David A Schweidel and George Knox. Incorporating direct marketing activity into latent attrition models. *Marketing Science*, 32(3):471–487, 2013.

[79] Adrian FM Smith and Gareth O Roberts. Bayesian computation via the gibbs sampler and related markov chain monte carlo methods. *Journal of the Royal Statistical Society: Series B (Methodological)*, 55(1):3–23, 1993.

[80] Bradley R Staats and Francesca Gino. Specialization and variety in repetitive tasks: Evidence from a japanese bank. *Management science*, 58(6):1141–1159, 2012.

[81] Tom Fangyun Tan and Serguei Netessine. When does the devil make work? An empirical study of the impact of workload on worker productivity. *Management Science*, 60(6): 1574–1593, 2014.

[82] Terry A. Taylor. On-demand service platforms. *Manufacturing & Service Operations Management*, 20(4), 2018.

[83] Gary M Thompson. Improved implicit optimal modeling of the labor shift scheduling

problem. *Management Science*, 41(4):595–607, 1995.

[84] Zeynep Ton and Robert S Huckman. Managing the impact of employee turnover on performance: The role of process conformance. *Organization Science*, 19(1):56–68, 2008.

[85] Senthil Veeraraghavan and Laurens Debo. Joining longer queues: Information externalities in queue choice. *Manufacturing & Service Operations Management*, 11(4):543–562, 2009.

[86] Sumio Watanabe. Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11(Dec):3571–3594, 2010.

[87] Karen Weise. Last-minute shoppers increasingly trust only amazon to deliver. `https://www.nytimes.com/2018/12/21/technology/amazon-last-minute-gifts-shopping.html`, 2018. Accessed: 2019-10-15.

[88] Joan C Williams, Susan J Lambert, Saravanan Kesavan, Peter J Fugiel, Lori Ann Ospina, Erin Devorah Rapoport, Meghan Jarpe, Dylan Bellisle, Pradeep Pendem, Lisa McCorkell, et al. Stable scheduling increases productivity and sales: The stable scheduling study. *University of California Hastings College of the Law, University of Chicago School of Social Service Administration, University of California Kenan-Flagler Business School*, 2018.

[89] Ping Josephine Xu, Russell Allgor, and Stephen C Graves. Benefits of reevaluating real-time order fulfillment decisions. *Manufacturing & Service Operations Management*, 11 (2):340–355, 2009.

[90] Qiuping Yu, Gad Allon, and Achal Bassamboo. How do delay announcements shape customer behavior? an empirical study. *Management Science*, 63(1):1–20, 2016.