Measuring the Effect of Queues on Customer Purchases[*]

Yina Lu · Marcelo Olivares Andrés Musalem Ariel Schilkrut

Columbia Business School *Duke University* *Scopix Solutions*

May 24, 2011

Abstract

Capacity decisions in service operations often involve a trade-off between operating cost and the level of service offered to customers. Although the cost of attaining a pre-specified level of service has been well-studied, there isn't much research studying how customer service levels affect revenue and profit. This paper conducts an empirical study to analyze how waiting in a queue in the context of a retail store affects customer purchasing behavior. Our methodology uses a novel technology based on digital imaging to record periodic information about the queuing system. Our econometric methodology integrates these data with point-of-sales information to estimate the effect of queues on purchases. We find that waiting in queue has a non-linear effect on purchase incidence and that customers appear to primarily focus on the length of the queue rather than the actual expected wait when making their purchase decisions. We also find that customer sensitivity to waiting is heterogeneous and negatively correlated with price sensitivity. We discuss implications of these results in the context of service design and category pricing.

Keywords: queuing, service operations, retail, choice modeling, empirical research, operations/marketing interface.

[*] Acknowledgments: We thank seminar participants at the Consortium for Operational Excellence in Retail (COER) for their useful feedback. We also thank Scopix Solutions for providing the data used in this study.

1 Introduction

Capacity management is an important aspect in the design of service operations. These decisions involve a trade-off between the costs of sustaining a service standard and the value that customers attach to this level of service. Most of the work in the operations management literature has focused on the first issue, developing models that are useful to quantify the costs of attaining a given level of service. Because these operating costs are more salient, it is frequent in practice to observe service operations rules designed to attain a given quantifiable service stantard. For example, a common rule in retail stores is to open additional check-outs when the length of the queue surpass a given threshold. However, there isn't much research focusing on how to choose an appropriate target service level. This requires measuring the value that customers assign to objective service levels measures and how this translates into revenue. The focus of this paper is to measure the effect of service levels– in particular, customers waiting in queue– on actual customer purchases, which can be used to attach an economic value to customer service.

Lack of objective data is an important limitation to empirically study the effect of waiting on customer behavior. A notable exception is call centers, where recent studies have focused on measuring customer impatience while waiting on the phone line (Gans et al. (2003)). Our focus is to study *physical* queues in services, where customers are physically present at the service facility during the wait. This type of queues is common in retail stores, banks, amusement parks and health care delivery. Because objective data on customer service is typically not available in this type of service facilities, most of the previous research relies on surveys to study how customers' *perceptions* of waiting affect their *intended* behavior. However, previous work has shown that customer perceptions of service do not necessarily match with the actual service level received, and their purchase intentions do not always translate into actual revenue (e.g. Chandon et al. (2005)). In contrast, our work uses objective measures of actual service collected through a novel technology – digital imaging with image recognition – that tracks operational metrics such as the number of customers waiting in line. We develop an econometric framework that uses these data together with sales transactions to estimate the impact of customer service levels on purchase incidence and choices among products. We test our methodology using field data collected in a pilot study conducted at the deli section of a big-box supermarket.

Our model provides several metrics that are useful for the management of service facilities. First, it provides estimates on how service levels affect the effective arrivals to a queuing system when customers may balk. This can be useful to set service and staffing levels optimally balacing operating costs against lost revenue. Second, it can be used to identify the relevant visible factors in a physical queuing system that drive customer behavior, which can be useful for the design of a service facility. Third, our models can be used to estimate how the performance of a queuing system may affect how customers substitute among alternative products or services accounting for heterogeneous customer preferences. Finally, our methodology can be used to attach a dollar value to the cost of waiting experienced by customers and to segment customers based on their sensitivity to waiting factors.

There are two important challenges in our estimation. A first issue is that congestion is highly dependent on store traffic and therefore periods of high sales are typically concurrent with long waiting lines. Consequently, we face a reverse causality problem: while we are interested in measuring the causal effect of waiting on sales, there is also a reverse effect by which spikes in sales generate congestion and therefore longer waits. The correlation between waiting times and aggregate sales is a combination of these two simultaneous effects and therefore cannot be used directly to estimate the causal effect of waiting on sales. To address this issue, we collected a detailed panel data with the history of individual customer transactions, which we use to control for this congestion effect.

Using customer transaction data produces a second estimation challenge. The imaging technology captures snapshots that describe the state of the queuing system at specific time epochs and does not provide an exact measure of the actual service experienced by an individual customer (technological and legal limitations preclude us from tracking the identity of customers in the queue). Therefore, the actual state of the queue observed by each customer is missing data which needs to be handled appropriately in the estimation. Our approach relies on using queuing theory to describe the stochastic process driving the transient behavior of the queue, and use this to infer the actual state of the queue observed by each customer based on the periodic snapshot data. We believe this is a valuable contribution that will facilitate the use of periodic store operational data in other studies involving customer transactions obtained from point-of-sales (POS) information.

Our empirical results suggest that purchase incidence is mainly affected by the number of customers in line, and does not seem to be affected by the level of staffing of the queue. This is consistent with customers using the number of people waiting in line as the only visible cue to assess the expected waiting time, not fully accounting for the actual speed at which the line moves. This empirical finding could have important implications in the design of the service facility. For example, we show that pooling multiple queues into a single queue with multiple servers may lead to more customers walking away. We also find significant heterogeneity in customer sensitivity to waiting, and that the degree of waiting sensitivity is negatively correlated with customer's sensitivity to price. We show the implications of these results in pricing decisions of product categories under the presence of congestion effects. Finally, our results suggest that the effect of queues is economically significant and appears to have a strong non-linear effect. Moderate increases in the number of customers in queue can generate sales reduction equivalent to a 5% price increase.

2 Related Work

In this section, we provide a brief review of the literature studying the effect of waiting on customer behavior and its implications for the management of queues. The literature is extensive, including empirical research using experimental and observational data in the fields of operations management, marketing and economics. We focus this review on a selection of the relevant literature, which helps us to identify hypotheses that are useful in developing our econometric model (described in section 3). At the same time, we also reference survey articles that provide a more exhaustive review of different literature streams.

Recent studies in the service engineering literature have analyzed customer transaction data in the context of call centers. See Gans et al. (2003) for a survey on this stream of work. Customers arriving to a call-center are modeled as a Poisson process where each arriving customer has a "patience threshold". Customers join the queue waiting to be served, but abandon the queue if they wait more than their patience threshold. This is typically referred to as the Erlang-A model or the M/M/c+G, where G denotes the generic distribution of the customer patience threshold. Brown et al. (2005) estimate the distribution of the patience threshold based on call-center transactional data and use it to measure the effect of waiting time on the number of lost (abandoned) customers.

Customers arriving to a call center typically do not directly observe the number of customers ahead of the line. In contrasts, for physical customer queues at a retail store, the length of the line is observed and may become a visible cue to assess the expected waiting time. Hence, the length of the line is an important factor in the customer's decision to join the queue, which is not captured in the Erlang-A model. In these settings, given a fixed amount of serving capacity, arrivals to the system can be modeled as a Poisson process where a fraction of the arriving customers may *balk* – that is, not join the queue – depending on the number of people already in queue (see Gross et al. (2008), chapter 2.10). Our work focuses on estimating how visible aspects of physical queues, such as queue length and capacity, affect choices of arriving customers, which provides an important input to this class of models.

Png and Reitman (1994) empirically study the effect of waiting time on the demand for gas stations. They identify service time as an important differentiating factor in this retail industry. Their estimation is based on aggregate data on gas station sales and uses measures of a station's capacity as a proxy for waiting time. Allon et al. (2010) study how service time affects demand across outlets in the fast food industry, using a structural estimation approach that captures price competition across outlets. Both studies use aggregate data from a cross-section of outlets in local markets. The data for our study is more detailed as it uses individual customer panel data and periodic information on the queue state, but it is limited to a single service facility.

Several empirical studies suggest that customer responses to waiting time are not necessarily linear. Larson (1987) provides anecdotal evidence of non-linear customer disutility under different service scenarios. Laboratory and field experiments have shown that customer's perceptions of waiting are important drivers of dissatisfaction and that these perceptions may be different from the actual (objective) waiting time, sometimes in a non-linear pattern (e.g. Davis and Vollmann (1993); Berry et al. (2002)). Mandelbaum and Zeltyn (2004) use analytical queuing models with customer impatience that can explain non-linear relationships between waiting time and customer abandonments. Indeed, in the context of call-center outsourcing, the common use of service level agreements based on delay thresholds at the upper-tail of the distribution (e.g. 95% of the customers wait less than 2 minutes) is consistent with non-linear effects of waiting on customer behavior (Hasija et al. (2008)).

Larson (1987) provides several examples of factors that affect customers' perceptions of waiting, such as: (1) whether the waiting is perceived as socially fair; (2) whether the wait occurs before or during the actual service begins; and (3) feedback provided to the customer on waiting estimates and the root causes generating the wait, among other examples. Berry et al. (2002) provides a survey of empirical work testing some of these effects. Most of this research relies on surveys which measure objective and subjective waiting times, linking these to customer satisfaction and intentions of behavior. Although surveys are useful to uncover the mechanism by which waiting affects customer behavior and the factors that mediate this effect, it also suffers from some disadvantages. In particular, there is a sample selection of non-respondents which tend to have higher opportunity cost for their time. In addition, several papers report that customer's intentions of purchase do not always match well with actual purchasing behavior (e.g. Chandon et al. (2005)). Our work uses measures on actual customer purchases and operational drivers of waiting time which have the benefit of being objective metrics, albeit at the expense of being somewhat limited to study some of the underlying behavioral mechanisms driving the effect of waiting time.

Several other studies use primary and secondary observational data to study the effect of service time on customer behavior. Forbes (2008) analyzes the impact of airline delays on customer complaints, showing that customer expectations play an important role in mediating this effect. Campbell and Frei (2010) study multiple branches of a bank, providing empirical evidence that teller waiting times affect customer satisfaction and retention. Their empirical study reveals significant heterogeneity in customer's sensitivity to waiting time, some of which can be explained through demographics and the intensity of competition faced by the branch. Our study also looks at customer heterogeneity in waiting sensitivity but in addition we relate this sensitivity to customers' price sensitivity. This has important implications for pricing, as we show later in section 5.

Our study uses discrete choice models based on random utility maximization to measure substitution effects driven by waiting. The same approach was used by Allon et al. (2010), who incorporate waiting time factors into the customer's utility modeled through a multinomial logit (MNL). We instead use a random coefficient MNL, which incorporates heterogeneity and allows for more flexible substitution patterns (Train (2003)). The random coefficient MNL model has also been used in the transportation literature to incorporate the value of time in consumer choice (e.g. Hess et al. (2005)).

Finally, all of the studies mentioned so far focus on settings where waiting time and congestion generate disutility to customers. However, there is theory suggesting that longer queues could create value to a customer. If a customer's utility for a good depends on the number of customers that consume it (as with positive network externalities), then longer queues could attract more customers. Another example is given by herding effects, which may arise when customers have asymmetric information about the quality of a product. In such a setting, longer queues provide a signal of higher value to uninformed customers, making them more likely to join the queue (see Debo and Veeraraghavan (2009) for several examples).

3 Estimation

This section describes the data and models used in our estimation. The literature review of section 2 provides several findings and hypotheses that are useful for specifying our econometric model. These can be summarized into the following testable hypotheses: (1) the effect of waiting time on customer purchasing behavior may be non-linear, such that a customer's sensitivity to a marginal increase in waiting time may vary at different levels of waiting time; (2) the effect may not be monotone– for example, although more anticipated waiting is likely to negatively affect a customer's purchase intentions, herding effects could make longer queues attractive to customers; (3) customer purchasing behavior is affected by perceptions of waiting time which may not necessarily match actual waiting time; (4) customer's sensitivity to waiting time is heterogeneous and possibly related to demographic factors, such as income or price sensitivity.

The first subsection describes the data used in our empirical study, which motivates the econometric framework developed in the rest of the section. Section 3.2 describes an econometric model to measure the effect of queues on purchase incidence. It uses a flexible functional form to measure the effect of the queue on purchasing behavior that permits non-linear and potentially non-monotone effects (hypotheses (1) and (2) above). Different specifications are estimated to test for factors that may affect customer's perceptions of waiting (hypothesis (3) above). Section 3.3 describes how to incorporate the periodic queue information contained in the snapshot data into the estimation of this model. The last subsection develops a discrete choice model that captures additional factors not incorporated into the purchase incidence model, including substitution among products, prices, promotions, and state-dependent variables that affect purchases (e.g., household inventory). This choice model is also used to measure heterogeneity in customer sensitivity to waiting (hypothesis (4) above).

3.1 Data

We conducted a pilot study at the deli section of a super-center located in Santiago, Chile. The store belongs to a leading supermarket chain in this country and is located in a working-class neighborhood. The deli section sells about 8 product categories, most of which are fresh cold-cuts sold by-the-pound.

During a pilot study running from October 2008 to May 2009 (approximately 7 months), we used digital snapshots analyzed with image recognition technology to periodically track the number of people waiting at the deli and the number of sales associates serving it. Snapshots were taken periodically every 30 minutes during the open hours of the deli, from 9am to 9pm on a daily basis. Figure 1 shows a sample snapshot that counts the number of customers waiting (top figure) and the number of employees attending customers behind the deli counter (bottom figure).[1] Throughout the paper, we denote the length of the deli queue at snapshot t by Q_t and the number of employees serving the deli by E_t.

During peak hours, the deli uses numbered tickets to implement a first-come-first-served priority in the queue. The counter displays a visible panel intended to show the ticket number of the last customer attended by a sales associate. This information would be relevant for the purpose of our study to complement the data collected through the snapshots; for example, Campbell and Frei (2010) use ticket-queue data to estimate customer waiting time. However, the ticket information was not stored in the POS database of the retailer and we learned from other supermarkets that this information is rarely recorded. Nevertheless, the methods proposed in this paper could also be used with periodic data collected via a ticket-queue, human inspection or other data collection procedures.

In addition to the queue and staffing deli information, we also collected point-of-sales (POS) data for all transactions involving grocery purchases from Jan 1st, 2008 until the end of the study period. In the market

[1] In counting the number of employees, the image recognition technology counts only those employees whose role is to serve customers, which wear a different uniform. Employees focused in other tasks (slicing, cleaning, etc.), which also appear in Figure 1, are not counted.

area of our study, grocery purchases typically include bread and about 78% of the transactions that include deli products also include bread. For this reason, we selected basket transactions that included bread to obtain a sample of grocery-related shopping visits. Each transaction contains check-out data on the basket of purchases, including a time-stamp of the check-out and the stock-keeping units (SKUs) bought along with unit quantities and prices (after promotions). We use the POS data prior to the pilot study period– from January to September of 2008 – to calculate metrics employed in the estimation of some our models (we refer to this subset of the data as the *calibration* data).

Using detailed information on the list of products offered at this supermarket, each cold-cut SKU was assigned to a product category (e.g. ham, turkey, bologna, salami, etc.). Some of these cold-cut SKUs include prepackaged products which are not sold by the pound and therefore are located in a different section of the store. For each SKU, we defined an attribute indicating whether it was sold in the deli or pre-packaged section. About 29.5% of the transactions in our sample include deli products, suggesting that deli products are quite popular in this supermarket.

We then examined how the number of transactions, queue length and number of employees varied throughout the course of the day. In weekdays, peak traffic hours are observed around mid-day, between 11am and 2pm, and in the evenings, between 6 and 8pm. Although there is some adjustment in the number of employees attending, this adjustment is insufficient and therefore queue lengths exhibit an hour-of-day pattern similar to the one for traffic. A similar effect is observed for weekends, although the peak hours are different. Congestion generates a positive correlation between sales and queue lengths, making it difficult to study the causal effect of queues on traffic using aggregate POS data. For this reason, our empirical study uses instead detailed *customer transaction* data. More specifically, the supermarket chain in our study operates a very popular loyalty program where more than 60% of the transactions are matched with a loyalty card identification number. Using this information we constructed a panel of individual customer purchases. To better control for customer heterogeneity, we focus on grocery purchases of loyalty card customers who visit the store one or more times per month on average. This accounts for a total of 284,709 transactions from 13,103 customers. Table 1 provides some summary statistics describing the queue snapshots, the POS and the loyalty card data.

3.2 Purchase Incidence Model

Recall that the POS and loyalty card data are used to construct a panel of observations for each individual customer. Each customer is indexed by i and each store visit by v. Let $y_{iv} = 1$ if the customer purchased a deli product in that visit, and zero otherwise. The objective is to model how the probability of purchase at the deli is affected by the state of the queue during a customer's visit. Define $\tilde{Q}_{iv}$ and $\tilde{E}_{iv}$ as the number of people in queue and the number of employees, respectively, that were observed by the customer during visit v. Note that we (the researchers) do not observe $\tilde{Q}_{iv}$ and $\tilde{E}_{iv}$ directly in our data. For now, we assume that these are observable but in the next section we show how to handle the unobserved data on $\tilde{Q}_{iv}$ and $\tilde{E}_{iv}$ in the estimation. As described in section 2, the effect of waiting time on customer behavior may be non-linear. Accordingly, let $f(\tilde{Q}_{iv}, \tilde{E}_{iv})$ be a vector-valued function that captures the functional form by which the state of the queue experienced by the customer affects its purchase incidence. The probability of a deli purchase is then modeled through the following generalized linear model:

$$h\left(\Pr[y_{iv} = 1]\right) = \beta_q \cdot f(\tilde{Q}_{iv}, \tilde{E}_{iv}) + \beta_x X_{iv}, \tag{1}$$

where $h(\cdot)$ is a link function, X_{iv} is a set of covariates that capture other factors that affect purchase incidence and (β_q, β_x) are parameters to be estimated, with dimensions that match their respective covariates. Changing the link function $h(\cdot)$ leads to different statistical models. For example, an identity link function leads to a linear probability model that can be estimated via Ordinary Least Squares. A logit link function, $h(x) = \ln[x/(1-x)]$, gives a logistic regression model which can be estimated via maximum likelihood methods (ML). We tested alternative link functions and found the results to be similar. For brevity, we only report the results from the logistic regression specification later in section 4.

Based on our discussion of section 2, we consider several specifications for $f(\tilde{Q}_{iv}, \tilde{E}_{iv})$ to test for multiple factors that affect a customer's perception of waiting. Upon arrival to the queue, a customer may estimate the waiting time and make a decision based upon that information. The number of customers in queue divided by the number of employees serving the deli, $W_{iv} = \tilde{Q}_{iv}/\tilde{E}_{iv}$, is a reasonable proxy for the expected waiting time of an arriving customer.

As shown in some of the experimental results reported in Carmon (1991), customers may use the length of the line, $\tilde{Q}_{iv}$, as a visible cue to assess the expected waiting time. The only information provided to customers is the number of customers ahead of them (through the display showing the ticket number of the last customer being served); no other metric of expected waiting time is provided. Hence, the length of the queue is highly visible in the environment we study, whereas the number of employees attending is not. Therefore, we test some specifications where the effect of the state of the queue is only a function of the queue length, $f(\tilde{Q}_{iv})$.

We consider alternative formulations for the function $f(\cdot)$: (1) a simple linear specification, to measure the average effect of expected time in queue on purchase incidence; (2) a piece-wise linear specification, to capture non-linearity and potentially non-monotone effects; and (3) a quadratic polynomial, which allows for non-linear and monotone effects in a more parsimonious way. Different criteria for model selection (e.g. Chow test and Akaike Information Criteria) are used to identify which models are best supported by the data.

There are two important challenges to estimate the model in equation (1). The first is that we are seeking to estimate a causal effect– the impact of $(\tilde{Q}_{iv}, \tilde{E}_{iv})$ on purchase incidence – using observational data rather than a controlled experiment. In an ideal experiment a customer would be exposed to multiple $(\tilde{Q}_{iv}, \tilde{E}_{iv})$ conditions holding all other factors (e.g., prices, time of the day, seasonality) constant. For each of these conditions, her purchasing behavior would then be recorded. In the context of our pilot study, however, there is only one $(\tilde{Q}_{iv}, \tilde{E}_{iv})$ observation for each customer visit. This could be problematic if customers with a high purchase intention visit the store around the same time. These visits would then exhibit long queues and high purchase incidence, generating a bias in the estimation of the causal effect. One example of this is when customers are heterogeneous in their purchase incidence and deli-customers visit the store at specific weekend hours. The data suggests such an effect: the average purchase probability is 34.2% on weekends at 8pm when the average queue length is 10.3, and it drops to 28.3% on weekdays at 4pm when the average queue length is only 2.2. Another example of this potential bias is when the deli runs promotions: price discounts attract more customers which increases purchase incidence and also generates higher congestion levels.

To partially overcome this challenge, we include covariates in X that control for customer heterogeneity. A flexible way to control for this heterogeneity would be to include customer fixed-effects, which controls for the average purchase incidence of customers. Purchase incidence could also exhibit seasonality– for example, consumption of fresh deli products could be higher during a Sunday morning in preparation for a family gathering during Sunday lunch. To control for this, the model includes a set of time of the day dummies interacted with weekend-day indicators. Finally, we also include a set of dummies for each day in the sample which controls for seasonality, trends and promotional activities (because promotions typically last at least a full day).

Although customer fixed effects account for purchase incidence heterogeneity across customers, they don't control for heterogeneity in purchase incidence across visits of the same customer. Furthermore, some of this heterogeneity across visits may be customer specific, so that they are not fully controlled by the seasonal dummies in the model. State-dependent factors, which are frequently used in the marketing literature (Neslin and van Heerde (2008)) could help to partially control for this heterogeneity. In addition, we note that the purchase incidence model (1) cannot be used to characterize substitution effects with products sold in the pre-packaged section, which could be important to measure the overall effect of queue-related factors on total store revenue and profit. The choice model described in subsection 3.4 addresses these and other limitations of the purchase incidence model (1). Nevertheless, these additions require focusing on a single product category, whereas the purchase incidence model considers all product categories sold in the deli. For this reason and due to its relative simplicity, the estimation of the purchase incidence model (1) provides valuable insights about how consumers react to different levels of service.

A second challenge in the estimation of (1) is that $(\tilde{Q}_{iv}, \tilde{E}_{iv})$ are not directly observable in our data set. The next subsection provides a methodology to infer $(\tilde{Q}_{iv}, \tilde{E}_{iv})$ based on the periodic data captured by the snapshots (Q_t, E_t) and describes how to incorporate these inferences into the estimation procedure.

3.3 Inferring Queues From Periodic Data

We start by defining some notation regarding the times associated to the relevant events in the data set, which are summarized in Figure 2. Time ts denotes the observed checkout time-stamp of the customer transaction. Time $\tau < ts$ is the time at which the customer observed the deli queue and made his decision on whether to join the line. The snapshot data of the queue were collected periodically, generating time intervals $[t-1, t)$, $[t, t+1)$, etc. For example, if the checkout time ts falls in the interval $[t, t+1)$, τ could fall in the intervals $[t-1, t)$,$[t, t+1)$, or in any other interval before ts (but not after). Let $B(\tau)$ and $A(\tau)$ denote the snapshots just before and after time τ. In some applications, such as the one we analyze, there is no record of the time at which the customer visited the queue. Therefore τ is not observed and we model it as a random variable for the purpose of estimation, and denote $F(\tau|ts)$ its conditional distribution given the checkout time ts.[2]

In addition, the state of the queue is only observed at pre-specified time epochs, so even if the deli visit time τ was known, the state of the queue may not be known exactly. It is then necessary to estimate (Q_τ, E_τ) for any given τ based on the snapshot data (Q_t, E_t). The snapshot data reveals that the number of employees in the system, E_t, doesn't fluctuate much: for about 60% of the snapshots, consecutive observations of E_t are identical. When they change, it is typically by one unit (81% of the samples). When $E_{t-1} = E_t = c$, it seems reasonable to assume that the number of employees remained constant in the interval $[t-1, t)$ and hence $E_\tau = c$. When changes between two consecutive snapshots E_{t-1}, E_t are observed, we assume (for simplicity) that the number of employees is equal to E_{t-1} throughout the interval $[t-1, t)$.

Assumption 1. *In any interval $[t-1, t)$, the number of servers in the queuing system is equal to E_{t-1}.*

[2] Note that in applications where the time of joining the queue is observed– for example, as provided by a ticket time stamp in a ticket-queue – it may still be unobserved for customers that decided not too join the queue. In those cases, τ may also be modeled as a random variable for customers that did not join the queue.

A natural approach to estimate Q_τ would be to take a weighted average of the snapshots around time τ: for example, two snapshots prior to τ ($Q_{B(\tau)-2}$ and $Q_{B(\tau)-1}$) and one snapshot after τ ($Q_{A(\tau)}$). If, for example, τ falls in the $[t-1,t)$ interval but lies closer to t than $t-1$, it would be reasonable to give more weight to the snapshot at t. However, it is not clear which is the exact weight that each observation should carry when calculating this average, or how many observations before and after τ should be considered. In what follows, we show a formal approach to use the snapshot data in the vicinity of τ to get a point-estimate of $\tilde{Q}_\tau$. Our methodology requires the following additional assumption about the evolution of the queuing system:

Assumption 2. *In any snapshot interval $[t-1,t)$, arrivals follow a Poisson process with an arrival rate $\lambda_{t-1}(Q,E)$ that may depend on the number of customers in queue and the number of servers. The service times follow an exponential distribution with a constant rate identical for all servers within each snapshot interval.*

Assumptions (1) and (2) together imply that in any interval between two snapshots the queuing system behaves like an Erlang queue model (also known as M/M/c) with balking rate that depends on the state of queue. The Markovian property implies that the conditional distribution of $\tilde{Q}_\tau$ given the snapshot data only depends on $Q_{B(\tau)}$and $Q_{A(\tau)}$, which simplifies the estimation. We now provide some empirical evidence to validate these assumptions.

Given that the snapshot intervals are relatively short (30 minutes), stationary Poisson arrivals within each time interval seem a reasonable assumption. To corroborate this, we did some analysis on the number of cashier transactions on every half-hour interval by comparing the fit of a Poisson regression model with a Negative Binomial (NB) regression. The NB model is a mixture model that nests the Poisson model but is more flexible, allowing for over-dispersion – that is, a variance larger than the mean. This analysis suggests that there is a small over-dispersion in the arrivals counts, so that the Poisson model provides a reasonable fit to the data.[3] As we show shortly, the arrival rate during each time period $\lambda_t(Q,E)$ will be useful in the estimation. This state dependent arrival rate is modeled as $\lambda_t(Q,E) = \bar{\lambda}_t \cdot d(Q,E)$, where $\bar{\lambda}_t$ is an average arrival rate that captures seasonality and variations across times of the day, and $d(Q,E) \in [0,1]$ is a discount factor that captures customer balking and is assumed to be time-independent. To estimate $\bar{\lambda}_t$, we grouped the time intervals into different days and hours-of-the-day and calculated the average number of transactions within each group. For example, we calculate the average number of customer arrivals across

[3] The NB model assumes Poisson arrivals arrivals with a rate λ that is drawn from a Gamma distribution. The variance of λ is a parameter estimated from the data; when this variance is close to zero, the NB model is equivalent to a Poisson process. The estimates of the NB model imply a coefficient of variation for λ equal to 17%, which is relatively low.

all time periods corresponding to "Mondays between 8-10am" and used this as an estimate of $\bar{\lambda}_t$ for those periods. Note that the balking effect $d(Q,E)$ is also unknown; in fact, it is exactly what the purchase incidence model (1) seeks to estimate. To make the estimation feasible, we get a first rough estimate of $d(Q,E)$ by estimating model (1) replacing $\tilde{E}_\tau$ by $E_{B(\tau)}$ and $\tilde{Q}_\tau$ by a simple average of three snapshots around time τ: $Q_{B(\tau)-2}, Q_{B(\tau)-1}$ and $Q_{A(\tau)}$. We later show how this rough estimate can be refined.

Since we do not observe the service times, we cannot estimate its distribution directly from the data. Therefore, to further validate assumption 2, we compared the distribution of the observed samples of $\{Q_t\}$ in the snapshot data with the distribution predicted by the Erlang model. To do this, we first group the time intervals into *buckets* $\{C_k\}_{k=1}^{K}$, such that intervals in the same bucket k have the same number of servers E_k and a similar average arrival rate $\bar{\lambda}_k$. For example, one of these buckets corresponds to "Mondays between 8-10am, with 2 servers". Provided an estimate of $\lambda_t(Q,E)$ (the previous paragraph showed how to obtain one), the only unknown primitive of the Erlang model is the service rate μ_t, or alternatively, the utilization level $\rho_t = \frac{\bar{\lambda}_t}{E_t \cdot \mu_t}$. The idea is then to estimate a utilization level ρ_k for each bucket so that the predicted stationary distribution implied by the Erlang model best matches the empirical distribution observed in the periods within each bucket. In our analysis, we estimated ρ_k by minimizing the L_2 distance between the empirical distribution and the predicted Erlang distribution.

Overall, the Erlang model provides a good fit for most of the buckets: a chi-square goodness of fit test rejects the Erlang distribution only in 4 out of 61 buckets (at a 5% confidence level). By adjusting the utilization parameter ρ, the Erlang model is able to capture shifts and changes in the shape of the empirical distribution across different buckets. The implied estimates of the service rate μ vary between 3 and 6 minutes, which also seems reasonable. We found that this service rate has a negative correlation (-0.46) with the average queue length, suggesting that servers speed up when the queue is longer (Kc and Terwiesch (2009) found a similar effect in the context of a healthcare delivery service).

The Markovian property (given by assumptions 1 and 2) implies that the conditional distribution of $\tilde{Q}_\tau$ depends only on the snapshots just before and after time τ, $B(\tau)$ and $A(\tau)$. Given the primitives of the Erlang model, we can use the transient behavior of the queue to estimate the distribution of $\tilde{Q}_\tau$, as we describe next. The length of the queue can be modeled as a birth-death process in continuous-time, with transition rates that depend on the primitives E_t, $\lambda_t(Q,E)$ and ρ_t. Note that we already showed how to estimate these primitives. The transition rate matrix during time interval $[t, t+1)$, denoted $\mathbf{R_t}$, is given by: $[\mathbf{R_t}]_{i,i+1} = \lambda_t(i, E_t)$, $[\mathbf{R_t}]_{i,i-1} = \min\{i, E_t\} \cdot \mu_t$, $[\mathbf{R_t}]_{i,i} = -\Sigma_{j \neq i}[\mathbf{R_t}]_{i,j}$ and zero for the rest of the entries.

The transition rate matrix $\mathbf{R_t}$ can be used to calculate the transition probability matrix for any elapsed time s, denoted $\mathbf{P_t}(s)$.[4] Let $p_{ij}(s) = [\mathbf{P_t}(s)]_{ij}$ be the probability of transitioning from i to j customers in elapsed time s during time period $[t, t+1)$ (the t index is omitted for notational convenience). For any $\tau \in [t, t+1)$, the distribution of Q_τ conditional on the snapshot data can be calculated as:

$$\Pr(Q_\tau = k) = K \cdot p_{Q_{B(\tau)}k}(\tau - B(\tau)) \cdot p_{kQ_{A(\tau)}}(A(\tau) - \tau), \tag{2}$$

for all $k \geq 0$ (there is an implicit conditioning on $Q_{B(\tau)}$ and $Q_{A(\tau)}$ which is omitted). K is a constant that normalizes the probability distribution so that $\sum_{k=0}^{\infty} \Pr(Q_\tau = k) = 1$.

In applications where τ is not observed, such as ours, it is necessary to integrate over all possible values of τ to obtain the posterior distribution of $\tilde{Q}_{iv}$, so that $\Pr(\tilde{Q}_{iv} = k|ts) = \int_\tau \Pr(Q_\tau = k) dF(\tau|ts_{iv})$, where ts_{iv} is the observed checkout time of the customer transaction. Therefore, given a distribution for τ, $F(\tau|ts_{iv})$, we can compute the distribution of $\tilde{Q}_{iv}$, which can then be used in equation (1) for model estimation. In particular, the unobserved value $\tilde{Q}_{iv}$ can be replaced by the point estimate that minimizes the mean square prediction error, which corresponds to its expected value $E[\tilde{Q}_{iv}]$.

In our application, we discretize the support of τ so that each 30-minutes snapshot interval is divided into a grid of 30 one-minute increments. Accordingly, for every minute in the grid, we calculate the probability of each possible value of the queue length using equation (2). Because we do not have additional information about how customers distribute their shopping time in the store (e.g., which sections of the store they visit first), we assume that the distribution of the deli visit time τ follows a discrete uniform distribution over the 60 minutes prior to the check-out time ts_{iv}.

Figure 3 illustrates some estimates of the distribution of the observed queue length $\tilde{Q}_\tau$ for different values of τ (for display purposes, the figure shows a continuous distribution but in practice it is a discrete distribution). In this example, the snapshots were $Q_t = 2$ and $Q_{t+1} = 8$, the arrival rate is $\bar{\lambda}_t = 0.4$ arrivals/minute, and the utilization is $\rho = 80\%$. For $\tau = 5$ minutes after the first snapshot, the distribution is concentrated close to $Q_t = 2$, whereas for $\tau = 25$ (5 minutes before the second snapshot), the distribution is more concentrated around $Q_{t+1} = 8$. The proposed methodology provides a rigorous approach combining queuing theory and the periodic snapshot information to estimate the distribution of the unobserved data $\tilde{Q}_\tau$ at any point in time.

Finally, note that the calculation of (2) requires knowing $\lambda_t(Q, E) = \bar{\lambda}_t d(Q, E)$. We used a first rough estimate of the discount factor $d(Q, E)$ to estimate the transition rate matrix $\mathbf{R_t}$ which leads to the point estimates of $\tilde{Q}_{iv}$ to be used in the estimation of model (1). It is possible to run this process iteratively

[4] Using the Kolmogorov forward equations, one can show that $\mathbf{P_t}(s) = e^{\mathbf{R_t}s}$. See Kulkarni (1995) for further details on obtaining a transition matrix from a transition rate matrix.

by using the latest estimates of the purchase incidence model (1) to get new values of the discount factor $d(Q, E)$ and thereby update the transition rate matrix $\mathbf{R_t}$ and estimates of $\tilde{Q}_{iv}$. In our application, we found that the estimates converge quickly after 3 iterations.

3.4 Choice Model

There are three important limitations of using the purchase incidence model (1). The first limitation is that it doesn't account for changes in a customer's purchase probability over time, other than through seasonality variables. This could be troublesome if customers plan their purchases ahead of time, as we illustrate with the following example. A customer who does weekly shopping on Saturdays and is planning to buy ham by the pound at the deli section visits the store early in the morning when the deli is less crowded. This customer visits the store again on Sunday to make a few "fill-in" purchases at a busy time for the deli and does not buy any ham products at the deli because she purchased ham products the day before. In the purchase incidence model, controls are indeed included to capture the *average* purchase probability at the deli for this customer. However, these controls don't capture the *changes* to this purchase probability between the Saturday and Sunday visits. Therefore, the model would mistakenly attribute the lower purchase incidence on the Sunday visit to the higher congestion at the deli whereas in reality the customer would have not purchased regardless of the level of congestion at the deli on that visit.

A second limitation of the purchase incidence model (1) is that it cannot be used to attach an economic value to the disutility of waiting by customers. One possible approach would be to calculate an equivalent price reduction that would compensate the disutility generated by a marginal increase in waiting. Model (1) cannot be used for this purpose because it does not provide a measure of price elasticity. A third limitation is that model (1) does not capture substitution with products that do not require waiting (e.g., the pre-packaged section), which can be useful to quantify the overall impact of waiting on store revenues and profit.

To overcome these limitations, we use a random utility model (RUM) to explain customer choice. As it is common in these type of models, the utility of a customer i for product j during a visit v, denoted U_{ijv}, is described as a function of product attributes and parameters that we seek to estimate. Previous research in marketing and economics has effectively estimated RUM specifications using scanner data from a single product category (e.g., Guadagni and Little (1983) model choices of ground coffee products; Bucklin and Lattin (1991) model saltine crackers purchases; Fader and Hardie (1996) model fabric softener choices; Rossi et al. (1996) model choices among tuna products). Note, however, that deli purchases include multiple product categories. Hence, using a RUM to model customer choice requires us to select a single product category for which purchase decisions are independent from choices in other categories and where customers typically choose to purchase at most one SKU in the category. The ham category appears to meet these criteria. We calculated correlations between ham purchases and purchases on other cold-cut categories and they are relatively small (all less than 8% in magnitude). About 93% of the transactions with ham purchases included only one ham SKU. In addition, it is the most popular category among cold-cuts, accounting for more than 33% of the total sales. The ham category has 75 SKUs, 38 of them which are sold in the deli and the rest in the pre-packaged section. About 85% of ham sales are sold in the deli. In what remains of this subsection, we describe a RUM framework to model choices among products in the ham category. Table 2 shows statistics for a selection of products in this category.

One advantage of using a RUM to characterize choices among SKUs in a category is that it allows us to include product specific factors that affect substitution patterns. Although many of the product characteristics do not change over time and can be controlled by a SKU specific dummy, our data reveals that prices do fluctuate over time and should be an important driver of substitution patterns. Accordingly, we incorporate product-specific dummies, α_j, and product prices for each customer visit (PRICE$_{vj}$) as factors influencing customer's utility for a product j. Including prices in the model also allows us to estimate customer's price sensitivity, which we use to put a dollar tag on the cost of waiting.

As in the purchase incidence model (1), it is important to control for customer heterogeneity. Due to the size of the data set, it is computationally challenging to estimate a choice model including fixed effects for each customer. Instead, we control for customer's average propensity to buy by including a covariate measuring the average consumption rate of each customer, denoted CR$_i$. This consumption rate was estimated using calibration data as done by Bell and Lattin (1998). We also use the methods develop by these authors to estimate customer's inventory of ham products at the time of purchase, based on a customer's prior purchases and their consumption rate of ham products. This measure is constructed at the category level and is denoted by INV$_{iv}$.

We use the following notation to specify the RUM. Let J be the set of products in the product category of interest (i.e., ham). J_W is the set of products that are sold at the deli section and, therefore, potentially require the customer to wait. $J_{NW} = J \backslash J_W$ is the set of products sold in the pre-packaged section which require no waiting. Let T_v be a vector of covariates that capture seasonal sales patterns, such as holidays and time trends. Also let $\mathbf{1}[\cdot]$ denote the indicator function. Using these definitions, customer i's utility for purchasing product j during store visit v is specified as follows:

$$\begin{aligned} U_{ijv} &= \alpha_j + \mathbf{1}[j \in J_W] \cdot \beta_i^q \cdot f\left(\tilde{Q}_{iv}, \tilde{E}_{iv}\right) \\ &\quad + \beta_i^{price} \text{PRICE}_{jv} + \gamma^{cr} \text{CR}_i + \gamma^{inv} \text{INV}_{iv} + \gamma^T \cdot T_v + \varepsilon_{ijv}, \end{aligned} \tag{3}$$

where ε_{ijv} is an error term capturing idiosyncratic preferences of the customer and $f\left(\tilde{Q}_{iv}, \tilde{E}_{iv}\right)$ captures the effect of the state of the queue in the customer's preference. Note that the indicator function $\mathbf{1}[j \in W]$ adds the effect of the queue only to the utility of those products which are sold at the deli section (i.e., $j \in J_W$) and not to products that do not require waiting. As in the purchase incidence model (1), the queue state $(\tilde{Q}_{iv}, \tilde{E}_{iv})$ is not perfectly observed but the method developed in subsection 3.3 can be used to replace these by point-estimates.[5] An outside good, denoted by $j = 0$, accounts for the option of not purchasing ham, with utility normalized to $U_{i0v} = \varepsilon_{i0v}$. The inclusion of an outside good in the model enables us to estimate how changes in waiting time affect the total sales of products in this category (i.e., category sales).

Assuming a standard extreme value distribution for ε_{ijv}, the RUM described by equation (3) becomes a random-coefficient multinomial logit. The model includes random coefficients for PRICE (β_i^{price}) and for some of the coefficients associated with the effect of the queue (β_i^q). These coefficients follow a Multivariate Normal distribution with mean $(\theta^q, \theta^{price})'$ and covariance matrix Ω, which we seek to estimate from the data. Including random-coefficients for price is useful to accommodate more flexible substitution patterns based on this characteristic, overcoming some of the limitations imposed by the independence of irrelevant alternatives of standard multinomial logit models. Allowing for covariation between the price and the queue-state coefficients ($\beta_i^{price}$and β_i^q) provides useful information on how customer's sensitivity to the state of the queue relates to price sensitivity.

The estimation of the model parameters is implemented using standard Bayesian methods (see Rossi and Allenby (2003)). The goal is to estimate: (i) the SKU dummies α_j; (ii) the effect of the consumption rate (γ^{cr}), inventory (γ^{inv}), and seasonality controls (γ^T) on consumer utility; and (iii) the distribution of the price and queue sensitivity parameters, which is governed by θ^q, θ^{price} and Ω. In order to implement this estimation, we define prior distributions on each of these parameters of interest: $\alpha_j \sim N(\bar{\alpha}, \sigma_\alpha)$, $\gamma \sim N(\bar{\gamma}, \sigma_\gamma)$, $\theta \sim N(\bar{\theta}, \sigma_\theta)$ and $\Omega \sim$ Inverse Wishart(df, Scale). For estimation, we specify the following parameter values for these prior distributions: $\bar{\alpha} = \bar{\gamma} = \bar{\theta} = 0$, $\sigma_\alpha = \sigma_\gamma = \sigma_\theta = 100$, df=3 and Scale equal to the identity matrix. These choices produce weak priors for parameter estimation. Finally, the estimation is carried out using Markov chain Monte Carlo (MCMC) methods. In particular, each parameter is sampled from its posterior distribution conditioning on the data and all other parameter values (Gibbs sampling). When there is no closed form expression for these full-conditional distributions, we employ Metropolis Hastings methods (see Rossi and Allenby (2003)). The outcome of this estimation process is a sample of values from the posterior distribution of each parameter. Using these values, a researcher can estimate any

[5] In our empirical analysis, we also performed a robustness check where instead of replacing the unobserved queue length $\tilde{Q}_{iv}$ by point estimates, we sample different queue lengths from estimated distribution of $\tilde{Q}_{iv}$. The results obtained with the two approaches are qualitatively similar.

moment of the posterior distribution, such as the posterior mean, variance and quantiles of each parameter.

4 Empirical Results

Using the methodology described in section 3.3, we obtained a point estimate of the state of the queue $(\tilde{Q}, \tilde{E})$ which is associated with each individual customer visit in the data. This section reports the estimates of the purchase incidence model (1) and the choice model (3) after replacing $(\tilde{Q}, \tilde{E})$ by these point estimates.

Purchase Incidence Model Results

Table 3 reports a summary of alternative specifications of the purchase incidence model (1). All the specifications include customer fixed effects and hour of the day dummies interacted with weekend/holiday dummies. The specifications differ in terms of: (1) the functional form for the queuing effect $f(\tilde{Q}, \tilde{E})$, including linear, piecewise linear and a quadratic polynomial; (2) the measure used to describe the state of the queue, either the expected waiting time, $\tilde{W} = \tilde{Q}/\tilde{E}$ or the queue length, $\tilde{Q}$ (we omit the tilde in the table). Accordingly, models I-III include linear, quadratic, and piecewise linear (with segments at $(0, 5, 10, 15)$) functions of $\tilde{W}$, respectively; while models IV-VI are the corresponding specifications based on $\tilde{Q}$ instead of $\tilde{W}$. We discuss the remaining models in Table 3 later in this section. The table also reports the number of parameters associated with the queuing effects (i.e., the dimension of β^q, dim(β^q)) and the log-likelihood achieved in the ML estimation. Because not all the models are nested, we provide two additional measures of goodness of fit, the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC), that are used for model selection.

Using AIC and BIC to rank the models, all the specifications with $\tilde{Q}$ as explanatory variables (models IV-VI) fit significantly better than those with $\tilde{W}$(models I-III), suggesting that changes in purchase incidence are better predicted by the queue length than the expected waiting time. Moreover, all of the β^q coefficients in models I-III are not statistically significant at the 5% level (not shown), whereas the coefficients in models IV-VI are statistically significant (discussed later in Table 4). In addition, we also estimated models that include both measures $\tilde{Q}$ and $\tilde{W}$ in a quadratic specification that nests models III and IV (not shown in the table). In this unrestricted model, none of the coefficients of $\tilde{W}$ are statistically significant (the average p-value is 35% and the smallest is 6%), whereas the coefficients on $\tilde{Q}$ are highly significant (a joint F-test has p-value less than 10^{-10}). This provides further support that it is the length of the line and not the expected waiting time what is driving customer purchase incidence.

Among the specifications that use $\tilde{Q}$, most of the nonlinear models outperform the linear one. Table 4 shows the estimated coefficients describing the non-linear effect of $\tilde{Q}$ in specifications V and VI. Note that for the quadratic model V, we de-meaned $\tilde{Q}$ to reduce multi-collinearity. The pattern obtained in these two models is similar: a slight positive effect in the purchase probability in the range $Q \in [0, 5]$, and then a marginally decreasing negative effect for queues longer that 5. To verify if the positive effect observed in the [0,5] range was driven by the limited flexibility of the function specification, we estimated additional specifications which allow for more flexibility in the range [0,6]: model VII is piecewise linear segmented at $(0, 3, 6, 10, 15)$; model VIII is piecewise linear segmented at $(0, 2, 4, 6, 10, 15)$; and model IX is a quadratic model similar to model V plus an indicator $I_{0 \leq \tilde{Q} \leq 1}$ to capture a "jump" near zero (labeled Quadratic+Jump). The AIC scores in Table 3 suggests that these more flexible models tend to provide a better fit than the less flexible models IV and VI. The BIC score, which puts a higher penalization for the additional parameters, tends to favor the more parsimonious quadratic models V and IX. A likelihood ratio test between the quadratic models V and IX suggests that the jump at zero is not statistically significant (p-value 0.64).

A detailed comparison of the estimates of these model is shown in figure 4. The purchase incidence pattern suggested by the estimates of the model is quite robust: customers balk when they experience long lines but they are less sensitive when the queue is short. The effect on purchase incidence can become quite large for queue lengths of 15 customers and more, reducing purchase incidence from 30% to 27%, which corresponds to a 10% drop in sales.

What is less intuitive is the slight increase in the purchase probability between 0 and 5. This result was puzzling at first to us, which motivated the additional estimation of models VII-IX. In all of these models there was a statistically significant increase in the purchase probability in the [0,5] range, corroborating that the effect was robust to alternative specifications.[6] We also discussed this pattern with managers from several supermarket chains to see if it made sense to them. In particular, when we presented the results to a senior executive of a leader first-tier supermarket chain in the US, he corroborated that some of their own data analysis (using a completely different methodology and data sources) showed a similar pattern.

Herding effects could explain this increasing pattern. Freshness is an important attribute of cold-cut products and there could be asymmetric information on this product characteristic across customers. Another potential explanation is that customers are not perfectly informed about the price promotions at the deli and observing people in line may lead customers to pay more attention to promotions. Finally, zero customers in the queue may be an indication that the deli is closed. To check this, we dropped observations during early and late hours and those for which the snapshots recorded no customers in line and zero staff. The estimated effects were similar in this sub-sample, suggesting this is not the main explanation of

[6] For model VII, the slope in the [3,6] range was not statistically significant, but it was positive and significant in the [0,3] range. For Model VIII, it was positive and significant in [0,2] but not significant in [2,4] and [4,6]. For Model IX, the Jump at zero was not statistically significant and the maximum of the quadratic polinomial is around $Q = 5$.

the observed effect. Although the increasing pattern is relatively small compared to the negative effects of long queues, it would be interesting to analyze in more detail the mechanism that drives this effect in future research.

Choice model results

In this subsection we present and discuss the results obtained for the choice model developed in section 3.4. The specification for the queuing effect $f(\tilde{Q}, \tilde{E})$ was based on the results of the purchase incidence model. In particular, we used a quadratic function of $\tilde{Q}$, which balanced goodness-of-fit and parsimony in the purchase incidence model. The utility specification includes product-specific intercepts, prices, consumption rate (CR), household inventory (INV) and controls for seasonality as explanatory variables. The model incorporates heterogeneity through random coefficients for price and the linear term of the length of the queue. We use 2,000 randomly selected customers in our estimation. After running 50,000 MCMC iterations and discarding the first 30,000 iterations, we obtained the results presented in Table 5 (the table omits the estimates of the product-specific intercept and seasonality). The left part of the table shows the estimates of the average effects, with the estimated standard error (s.e., measured by the standard deviation of the posterior distribution of each parameter). The right part of the table shows the estimates of the variance-covariance matrix (Ω) characterizing the heterogeneity of the random coefficients β_i^{price} and β_i^q.

Price, inventory and consumption rate all have the predicted signs and are estimated precisely. The average of the implied price elasticities of demand is -4.7. The average effects of the queue coefficients imply qualitatively similar effects as those obtained in the purchase incidence model: a small increase from $\tilde{Q} = 0$ to 6 and then a sharper decrease above $\tilde{Q} = 7$.

These results can be used to assign a monetary value to a customer's cost of waiting. For example, for an average customer in the sample, an increase from 5 to 10 customers in queue is equivalent to a 3.22% increase in price. Instead, an increase from 10 to 15 customers is equivalent to a 8.26% increase in price, illustrating the strong non-linear effect of waiting on customer purchasing behavior.

The estimates also suggest substantial heterogeneity on customers' price sensitivities (estimates on the right side of Table 5). The estimated standard deviation of the random price coefficients is 2.165, which implies a coefficient of variation of 26.4%. There is also significant heterogeneity in customers' sensitivity to waiting, as measured by the standard deviation of the linear queue effect, which is estimated to be 1.556. The results also show a significant negative correlation between the price and waiting sensitivity, estimated as -0.507.

To illustrate this negative relationship between price and queue sensitivity we consider the purchase probability of ham products in the deli section for three customer segments with different levels of price sensitivity: a price coefficient equal to the mean; one standard deviation below the mean, labeled high price sensitivity; and one standard deviation above the mean, labeled low price sensitivity. To compute these choice probabilities, we considered customer visits with average levels of prices, consumption rate and consumer inventory. Given the negative correlation between price and queue length sensitivity, customers with higher price sensitivity will in turn have lower sensitivity to the length of the queue. Figure 5 illustrates this pattern, showing a much stronger decline in the purchase probability in the customer segment with low price sensitivity. Interestingly, the low price sensitivity segment is also the most profitable, with average purchases that are 40% higher than those of the high price sensitivity segment. This has important implications for pricing product categories under congestion effects, as we discuss in the next section.

5 Managerial Implications

The results of the previous section suggest that: (1) purchase incidence appears to be affected by the length of the line rather than the expected waiting time – which are not equivalent when the number of servers changes over time; and (2) there is heterogeneity in customers' sensitivity to the queue length, which is negatively correlated with their price sensitivity. We discuss two important managerial insights implied by these findings. The first one is that pooling multiple identical queues into a single multi-server queue may lead to an increase in lost sales. The second one discusses the implications of the externalities generated by congestion for pricing and promotions management in a product category.

5.1 Queuing Design

A relevent question for queuing design in services is whether to operate with multiple queues or to merge them into a single queue with pooled servers. It is well known that an $M/M/c$ pooled queuing system achieves lower waiting time than a system with separate $M/M/1$ queues operating under the same utilization. Therefore, if waiting time is the only measure of customer service, then pooling queues is beneficial. However, Rothkopf and Rech (1987) provide several reasons why pooling queues could be less desirable. For example, there could be gains from server specialization that can be achieved in the separate queue setting. The results in this paper provide another argument for why splitting queues may be beneficial: although the waiting time in the pooled system is shorter, the queue is longer. If customers base their decision of joining a queue based on the length of the queue, as our empirical results suggest, then a pooled system may lead to fewer customers joining the system and, therefore, increase lost sales. We illustrate this in more detail with the following example.

We study two alternative queuing systems. The first corresponds to a *pooled* system given by a $M/M/2$ queue with constant arrival rate λ. The second system corresponds to a *split join-the-shortest-queue (JSQ)* system with two parallel single-server queues with total arrivals given by a Poisson process with rate λ. In the split JSQ system, customers join the shortest queue upon arrival and no jockeying is allowed thereafter. If there is no balking– that is, all customers join the queue – it can be shown that the pooled system dominates the split JSQ system in terms of waiting time. However, the queues are longer in the pooled system, so if customers may walk away upon arrival and this balking rate increases with the queue length, then the pooled system may lead to fewer sales.

The following numerical example evaluates the differences in the average waiting time and revenue between the two systems. For the split JSQ system, the approximate model proposed by Rao and Posner (1987) is used to numerically evaluate the system's performance. The arrival rate when the queue has n customers is given by λd_n, where $d_n \leq 1$ is a discount factor. The discount factor is set to follow a similar pattern to that obtained in the estimates of the purchase incidence model (1): $d_i = 1$ for $i \in [0, 6]$ and then decreases down to 75% when the queue reaches 20 customers (we set the capacity of the queue to 20). Traffic intensity is defined as $\rho = d_0\lambda/\mu$ and revenue is defined as the number of customers that join the queue. Figure 6 shows the long-run steady-state average waiting time and average revenue of the two systems. As expected, the pooled $M/M/2$ system always achieves shorter waiting time but generates less revenue as it suffers more traffic loss due to long queues. The difference increases as the traffic intensity approaches one. In this example, the split JSQ system gains 2.6% more revenue while almost doubling the average waiting time at high levels of utilization compared to the pooled system.

5.2 Implications for Category Pricing

The empirical results also suggest that customers' sensitivity to waiting is negatively correlated with price sensitivity. This could have important implications for the pricing of products under congestion effects, as we show in the following example.

Consider two vertically differentiated products, H and L, of high and low quality respectively, with respective prices $p_H > p_L$. Customers arrive according to a Poisson process to join a M/M/1 queue to buy at most one of these two products. Following model (3), customer preferences are described by a multinomial logit model, where the utility for customer i of buying product $j \in \{L, H\}$ is given by $U_{ij} = \delta_j - \beta_i^p p_j - \beta_i^q \tilde{Q} + \theta_i + \epsilon_{ij}$. Customer may also choose not to join the queue and get a utility equal to $U_{i0} = \epsilon_{i0}$. In this RUM, δ_j denotes the quality of the product and $\tilde{Q}$ is a r.v. representing the length of queue observed by the customer upon arrival. Customers have heterogeneous price and waiting sensitivities, characterized by the parameters β_i^p and β_i^q, respectively. In particular, heterogeneity is modeled through two discrete segments, $s = \{1, 2\}$, with low and high price sensitivity, respectively, such that each segment accounts for 50% of the customer population. Let β_1^p and β_2^p be the price coefficients for these segments, with $0 < \beta_1^p < \beta_2^p$, and let θ_1 and θ_2 characterize the different utility intercepts of the two segments. In addition, the waiting sensitivity β_i^q is a random coefficient that can take two values: ω_h with probability r_s and ω_l with probability $1-r_s$, where s denotes the segment of customer i and $\omega_l < \omega_h$. This characterization allows for price and waiting sensitivity to be correlated: if $r_1 > r_2$ then a customer with low price sensitivity is more likely to be more wait-sensitive; but if $r_1 = r_2$, then there is no correlation.

Consider first a setting with no congestion so that Q is always zero (for example, if there is ample capacity to serve customers). For illustration, we fixed the parameters as follows: $\delta_H = 15$, $p_H = 5$, $\delta_L = 5$, $p_L = 1.5$, $\beta_1^p = 1$, $\theta_1 = 0$, $\beta_2^p = 10$ and $\theta_2 = 12$. In this example, the difference in quality and prices between the two products is sufficiently large so that most of the price sensitive customers ($s = 2$) buy the low quality product L. Moreover, define the cross elasticity E_{HL} as the percent increase in share of the H product from increasing the price of L by 1%, and vice-versa for E_{LH}. In this numerical example, we allow for significant heterogeneity with respect to price sensitivity so that, in the abscence of congestion, the cross elasticities between the two products are close to zero (to be exact, $E_{HL} = 0.002$ and $E_{LH} = 0.008$).

Now consider the case where customers do observe queues. In this setting, congestion effects generate an externality: increasing the demand of one product generates longer queues, which decreases the utility of some customers which may in turn decide not purchase. Hence, lowering the price of one product increases congestion and thereby has an indirect effect on the demand of the other product, which we refer to as the *indirect* cross elasticity effect.

We now show how customer heterogeneity and negative correlation between price and waiting sensitivity can increase the magnitude of the indirect cross elasticity between L and H. We parameterized the waiting sensitivity of each segment as $\omega_l = 1.25 - 0.5\Delta$ and $\omega_h = 1.25 + 0.5\Delta$, where Δ is a measure of heterogeneity in waiting sensitivity. We also considered different values of the conditional probabilities r_1 and r_2 to vary the correlation between waiting and price sensitivity while keeping the marginal distribution of waiting sensitivity constant (50% ω_l and 50% ω_h). Fixing all the parameters of the model (including prices p_H and p_L), it is possible to calculate the stationary probabilities of the queue length $\tilde{Q}$. Using the RUM together with this stationary distribution it is then possible to calculate the share of each product (defined as the fraction of arriving customers that buy it). Using finite differences with respect to prices, one can then calculate cross elasticities that account for the indirect effect through congestion.

Using this approach, we evaluated the cross elasticity of the demand for the H product when changing the price of the L product (E_{HL}) for different degrees of heterogeneity in customer sensitivity to wait (Δ) and several correlation patterns. The results of this numerical example are presented in Table 6. Note how in the absence of heterogeneity– that is, $\Delta = 0$ – the cross-price elasticity is low: the two products H and L appeal to different price-sensitivity segments and there is little substitution between them. However, adding heterogeneity and correlation can lead to a different effect. In the presence of heterogeneity, a *negative* correlation between price and waiting sensitivity increases E_{HL}, showing that the *indirect* cross-elasticity increases when the waiting sensitive customers are also the least sensitive to price. The changes in cross-elasticity due to correlation can become quite large for higher degrees of customer heterogeneity. In the example, when $\Delta = 2$, the cross elasticity changes from 0.011 to 0.735 when moving from positive to negative correlation patterns.

We now discuss the intuition behind the patterns observed in the example of Table 6. When there is heterogeneity in price sensitivity, lowering the price of the L product attracts customers who were not purchasing before the price reduction (as opposed to cannibalizing the sales of the H product). Due to this increase in traffic, congestion in the queue increases, generating longer waiting times for all customers. But when price and waiting sensitivity are *negatively* correlated, the disutility generated by the congestion will be higher for the less price sensitive customers. Hence, the less price sensitive customers are more likely to walk away after the price reduction in L. Since a larger portion of the demand for the H product comes from the less price sensitive buyers, the indirect cross-price elasticity will increase as the correlation between price and waiting sensitivity becomes more negative.

In summary, the relationship between price and waiting sensitivity is an important factor affecting the prices in a product category when congestion effects are present. Congestion can induce price-demand interactions among products which in the absence of congestion would have a low direct cross price-elasticity of demand. We illustrate how heterogeneity and negative correlation between price and waiting sensitivity can exacerbate these interactions through stronger indirect cross-elasticity effects. This can have important implications on how to set prices in the presence of congestion.

6 Conclusions

In this paper, we make use of a novel data set that links the purchase history of supermarket customers with objective measures of their service experience. Using this information we are able to study how an important component of the service experience – waiting in queue – affects customer purchasing behavior.

An important contribution of this paper is methodological. An existing barrier to study the impact of service levels on customer buying behavior in retail environments comes from the lack of objective data on waiting time and other customer service metrics. This work uses a novel data collection technique to gather high frequency store operational metrics related to the actual level of service delivered to customers. Due to the periodic nature of these data, an important challenge arises in linking the store operational data with actual customer transactions. We develop a novel econometric approach that relies on queuing theory to infer the level of service associated to each customer transaction. This allows us to estimate the effect of service on customer purchase incidence and choice. In our view, this methodology could be extended to other contexts were periodic service level metrics and customer transaction data are available.

This methodology allows us to estimate a comprehensive descriptive model of how waiting in queue affects customer purchase decisions. Based on several aspects of this descriptive model we provide useful prescriptions for the management of queues and other important aspects of service management in retail. In this regard, a first contribution of our work it to measure the overall impact of waiting on customer purchasing incidence, thereby attaching an economic value to the level of service provided. This value of service together with an estimate of the relevant operating costs can be used to choose an optimal target service level, a useful input for capacity and staffing decisions.

Second, our model describes the actual factors in a queuing system that influence customer behavior. The results suggest that customer seem to focus on the length of the line when deciding to join a queue, and seem to disregard information about the speed at which the queue is expected to move. This has implications for the design of a queuing system. For example, although there are several benefits of pooling multiple single-server queues into a single queue with multiple servers, the results in this paper suggest that some precautions should be taken. In moving towards a pooled system, it may be critical to provide information about the expected waiting time so that customers do not anchor their decision solely on the length of the line, which tends to increase when the system is pooled. In addition, our empirical analysis provides strong evidence that the effect of waiting on customer purchases is non-linear. Hence, measuring extremes in the waiting distribution – for example, the fraction of the time that 10 or more customers are waiting in queue – may be more appropriate than using average waiting time to evaluate the system's performance.

Third, our econometric model can be used to segment customers based on their waiting and price sensitivities. The results show that there is indeed a large degree of heterogeneity in how customers react to waiting and price. Moreover, there is a significant negative correlation between waiting and price sensitivity. This has important implications for the pricing of a product category where congestion effects are present. Pricing under congestion effects generates an externality across products. Heterogeneity and negative correlation in price and waiting sensitivity exacerbates this externality, and therefore should be accounted for in pricing decisions.

Finally, our study has several limitations that could be explored in future research. Our analysis focuses on studying the short term implications of queues by looking at how customer purchases are affected during a store visit. There could be long-term effects whereby a negative service experience also influences future customer purchases, for example, the frequency of visits and retention. Another possible extension would be to measure how observable customer characteristics – such as demographics – are related to their sensitivity to wait. This would be useful, for example, to prescribe target service levels for a new store based on the demographics of the market. Competition could also be an important aspect to consider; this would probably require data from multiple markets to study how market structure mediates the effect of queues on customer purchases.

On a final note, this study highlights the importance of integrating advanced methodologies from the fields of operations management and marketing. We hope that this work stimulates further research on the interface between these two academic disciplines.

References

Allon, Gad, Awi Federgruen, Margaret Pierson. 2010. 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. Working paper.

Bell, D.R., J.M. Lattin. 1998. Shopping behavior and consumer preference for store price format: Why "large basket" shoppers prefer EDLP. *Marketing Science* **17**(1) 66–88.

Berry, L.L., K. Seiders, D. Grewal. 2002. Understanding service convenience. *Journal of Marketing* **66**(3) 1–17.

Brown, L., N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, L. Zhao. 2005. Statistical analysis of a telephone call center: A queueing-science perspective. *Journal of the American Statistical Association* **100**(469) 36–50.

Bucklin, R.E, J.M. Lattin. 1991. A two-state model of purchase incidence and brand choice. *Marketing Science* **10** 24–39.

Campbell, D., F. Frei. 2010. Market Heterogeneity and Local Capacity Decisions in Services. *Manufacturing & Service Operations Management* .

Carmon, Ziv. 1991. Recent studies of time in consumer behavior. *Advances in Consumer Research* **18**(1) 703–705.

Chandon, P., V.G. Morwitz, W.J. Reinartz. 2005. Do intentions really predict behavior? Self-generated validity effects in survey research. *The Journal of Marketing* **69**(2) 1–14.

Davis, M.M., T.E. Vollmann. 1993. A framework for relating waiting time and customer satisfaction in a service operation. *Journal of Services Marketing* **4**(1) 61–69.

Debo, Laurens, Senthil Veeraraghavan. 2009. *Consumer-Driven Demand and Operations Management*, *International Series in Operations Research and Management Science*, vol. 131, chap. 4 , Models of Herding Behavior in Operations Management. Springer Science, 81–111.

Fader, P.S., B.G.S Hardie. 1996. Modeling Consumer Choice Among SKUs. *Journal of Marketing Reserach* **33.4** 442–452.

Forbes, S.J. 2008. The effect of air traffic delays on airline prices. *International Journal of Industrial Organization* **26**(5) 1218–1232.

Gans, N., G. Koole, A. Mandelbaum. 2003. Telephone call centers: Tutorial, review, and research prospects. *Manufacturing and Service Operations Management* **5**(2) 79–141.

Gross, D., J. Shortle, J. Thompson, C. Harris. 2008. *Queueing Theory*. 4th ed. Wiley.

Guadagni, P.M., J.D.C. Little. 1983. A logit model of brand choice calibrated on scanner data. *Marketing Science* **2** 203–238.

Hasija, S., E. Pinker, R. Shumsky. 2008. Call center outsourcing contracts under information assymetry. *Management Science* **54(4)** 793–807.

Hess, S., M. Bierlaire, J.W. Polak. 2005. Estimation of value of travel-time savings using mixed logit models. *Transportation Research Part A: Policy and Practice* **39**(2-3) 221–236.

Kc, D.S., C. Terwiesch. 2009. Impact of workload on service time and patient safety: An econometric analysis of hospital operations. *Management Science* **55**(9) 1486–1498.

Kulkarni, V.G. 1995. *Modeling and analysis of stochastic systems*. Chapman & Hall/CRC.

Larson, R.C. 1987. Perspective on queues: Social justice and the psychology of queueing. *Operations Research* **35**(6) 895–905.

Mandelbaum, A., S. Zeltyn. 2004. The impact of customers patience on delay and abandonment: Some empirically-driven experiments with the mmng queue. *OR Spectrum* **26**(3) 377–411.

Neslin, S.A., H.J. van Heerde. 2008. Promotion dynamics. *Foundations and Trends in Marketing* **3 (4)** 177–268.

Png, I.P.L., D. Reitman. 1994. Service time competition. *The Rand Journal of Economics* **25**(4) 619–634.

Rao, B.M., M.J.M. Posner. 1987. Algorithmic and approximation analyses of the shorter queue model. *Naval Research Logistics* **34** 381–398.

Rossi, P.E., G.M. Allenby. 2003. Bayesian statictics and marketing. *Marketing Science* **22 (3)** 304–328.

Rossi, P.E., R.E. McCulloch, G.M. Allenby. 1996. The value of purchase history data in target marketing. *Marketing Science* **15** 321–240.

Rothkopf, M.H., P. Rech. 1987. Perspectives on queues: Combining queues is not always beneficial. *Operations Research* **35**(6) 906–909.

Train, K. 2003. *Discrete choice methods with simulation*. Cambridge Univ Pr.

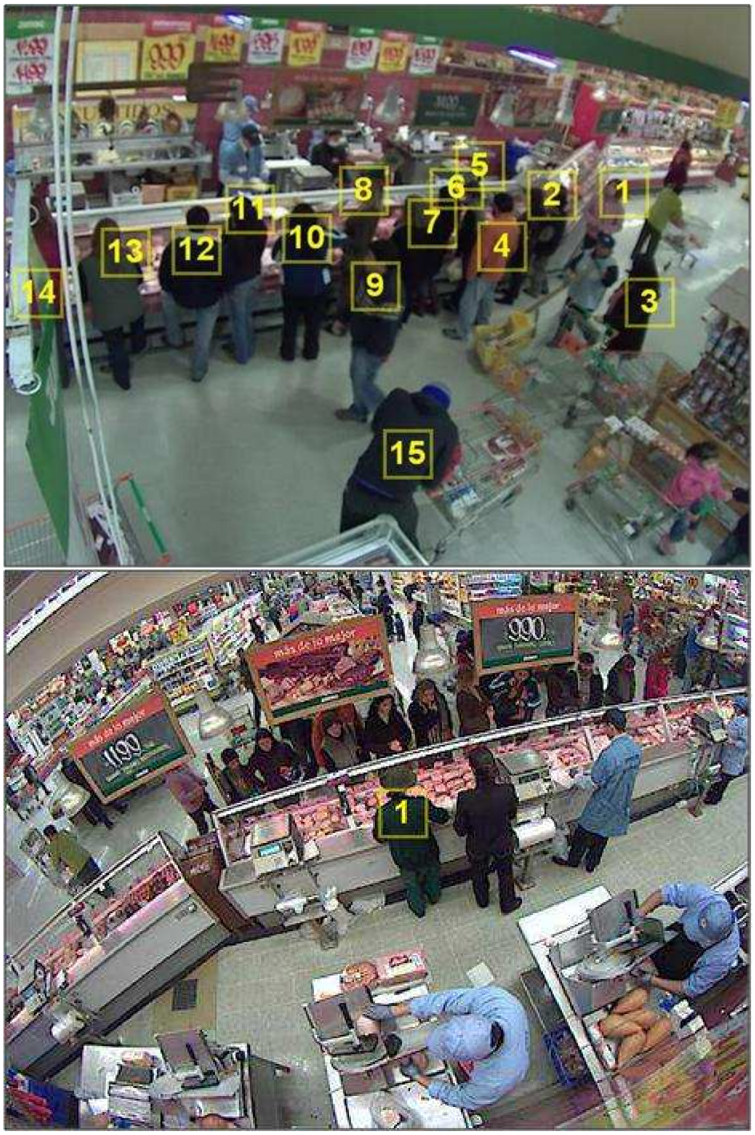

Figure 1: Example of deli snapshot showing the number of customers waiting (top) and the number of employees attending (bottom).

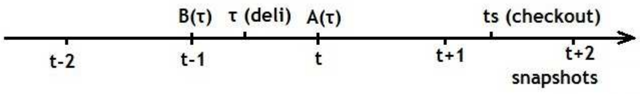

Figure 2: Sequence of events related to a customer purchase transaction.

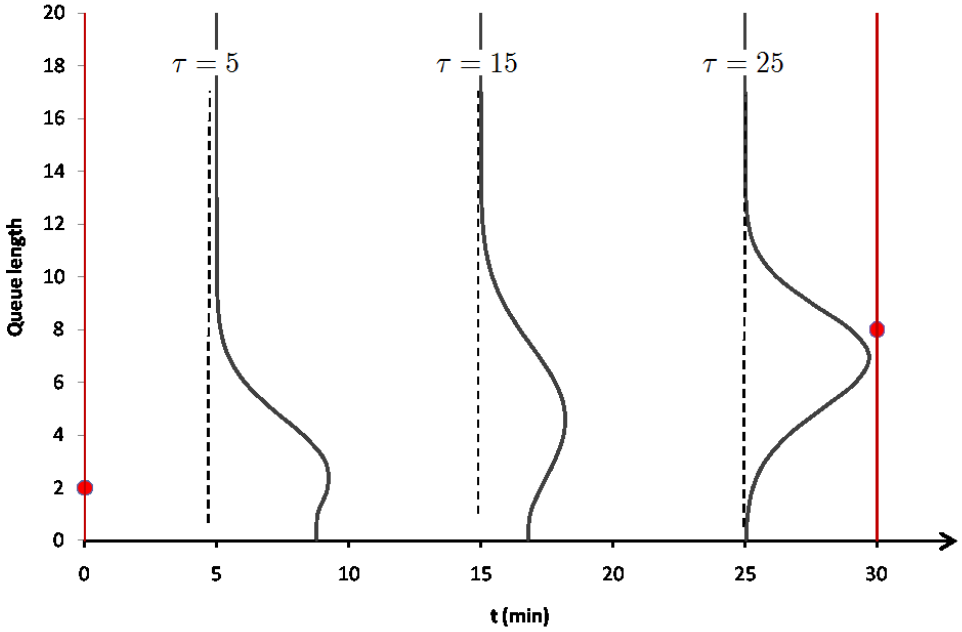

Figure 3: Estimates of the distribution of the queue length observed by a customer for different deli visit times (τ). In this example, the two circles in the vertical axes correspond to the snapshots before and after time τ, equal to 2 and 8, respectively.

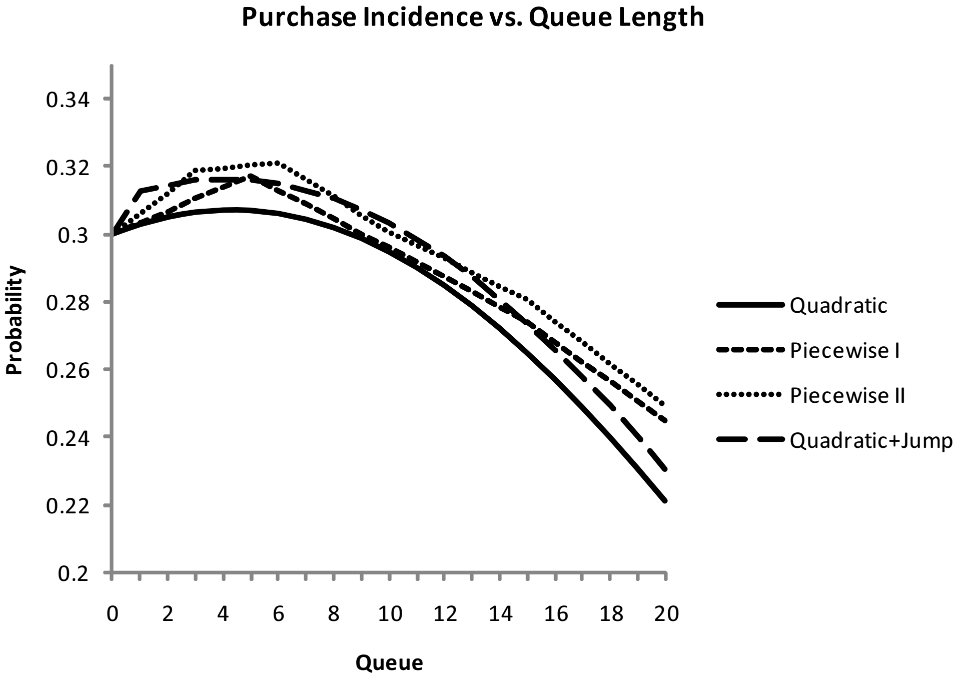

Figure 4: Results from the different specifications of the purchase incidence model.

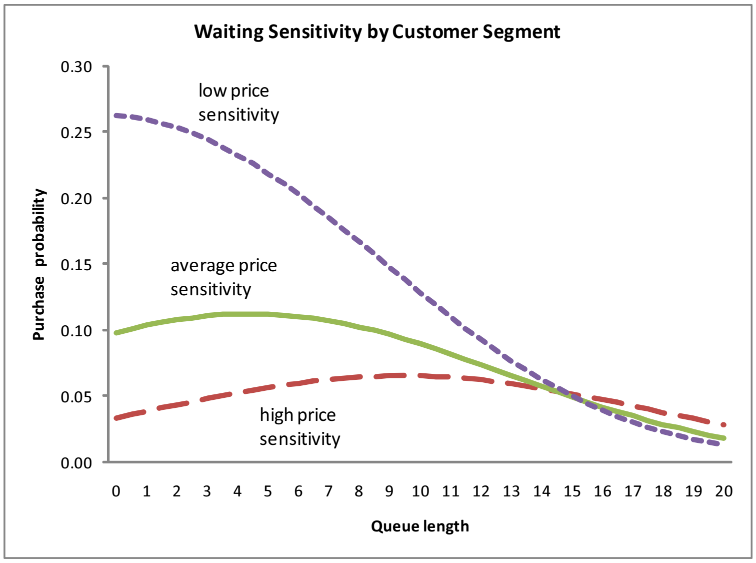

Figure 5: Purchased probability of ham products in the deli section versus queue length for three customers segments with different price sensitivity.

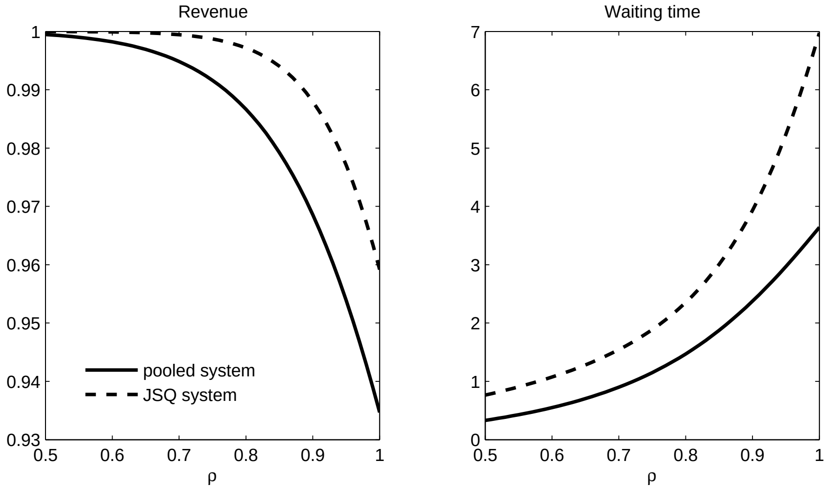

Figure 6: Comparison between the Split Join-Shortest-Queue (JSQ) and Pooled systems.

		# obs	mean	stdev	min	max
Periodic snapshot data						
Length of the queue (Q)	weekday	3671	3.76	3.81	0	26
	weekend	1465	6.42	4.90	0	27
Number of employees (E)	weekday	3671	2.11	1.26	0	7
	weekend	1465	2.84	1.46	0	9
Point-of-Sales data						
Purchase incidence of deli products		284,709	22.5%			
Loyalty card data						
number of visits per customer		13103	62.8	45.7	20	467

Table 1: Summary statistics of the snapshot data, point-of-sales data and loyalty card data.

Product	Avg Price	St.Dev. Price	Share
1	0.67	0.10	21.23%
2	0.40	0.04	9.37%
3	0.53	0.06	7.12%
4	0.59	0.06	6.13%
5	0.64	0.07	5.66%
6	0.24	0.01	5.49%
7	0.52	0.07	3.97%
8	0.54	0.07	3.10%
9	0.56	0.07	2.85%
10	0.54	0.08	2.20%

Table 2: Statistics for the ten most popular ham products, as measured by the percent of transactions in the category accounted by the product (Share). Prices are measured in ten thousands Chilean pesos per kilo (Ch$10,000 = US$20, approximately).

Model	Function form	Metric	dim(β^q)	logL	AIC	rank	BIC	rank
I	Linear	W	1	-121005.95	265503.90	9	388364.53	7
II	Quadratic	W	2	-121003.19	265500.38	8	388371.46	8
III	Piecewise I	W	4	-121000.35	265498.70	7	388390.70	9
IV	Linear	Q	1	-120991.61	265475.22	6	388335.84	3
V	Quadratic	Q	2	-120977.24	265448.48	3	388319.56	1
VI	Piecewise I	Q	4	-120975.48	265448.96	5	388340.96	4
VII	Piecewise II	Q	5	-120973.21	265446.42	1	388348.88	5
VIII	Piecewise III	Q	6	-120973.40	265448.80	4	388361.72	6
IX	Quadratic+Jump	Q	3	-120976.13	265448.26	2	388329.80	2

Table 3: Goodness of fit results on alternative specifications of the purchase incidence model (equation (1)).

	Variable	Coef.	Std. Err.	z
Model V	$\tilde{Q} - 5.7$	-.0050	.0028	-1.79
	$(\tilde{Q} - 5.7)^2$	-.0018	.0003	-5.35
Model VI	$\tilde{Q}_{0-5}$	.0161	.0062	2.58
	$\tilde{Q}_{5-10}$	-.0199	.0042	-4.79
	$\tilde{Q}_{10-15}$	-.0215	.0066	-3.25
	$\tilde{Q}_{15+}$	-.0306	.0230	-1.33

Table 4: Estimation result for selected specifications of the purchase incidence model (equation (1))

	Average Effect			Variance/Covariance (Ω)	
	estimate	s.e.		estimate	s.e.
Inv	-0.119	0.029			
CR	3.589	0.150			
Price	-8.202	0.221	Ω(Price)	4.689	0.292
$\tilde{Q}$	-0.221	0.073	$\Omega(\tilde{Q})$	2.419	0.188
$\tilde{Q}^2$	-0.792	0.097	Ω(Price,$\tilde{Q}$)	-1.708	0.176

Table 5: Estimation results for the choice model (equation 3). The estimate and standad error (s.e.) of each parameter correspond to the mean and standard deviation of its posterior distribution.

	Correlation between price and waiting sensitivity				
Heterogeneity	-0.9	-0.5	0	0.5	0.9
$\Delta = 0.0$	-	-	0.042	-	-
$\Delta = 1.0$	0.342	0.228	0.120	0.047	0.010
$\Delta = 2.0$	0.735	0.447	0.209	0.070	0.011

Table 6: Cross-price elasticities describing changes in the probability of purchase of the high priced product (H) from changes in the price of the low price product (L).