NORTHWESTERN UNIVERSITY

Empirical Research in Service Operations, Sustainability and Supply Chain Management

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Operations Management

By

Kejia Hu

EVANSTON, ILLINOIS

June 2017

ABSTRACT

Empirical Research in Service Operations, Sustainability and Supply Chain Management

Kejia Hu

Data is a myth and a treasure. Empirical analysis is the key to unlock the myth and discover the valuable information in the treasure. My research during PhD study centered around empirical analysis and extracted insights to improve the operations in services, the sustainability regulation of government and the demand prediction in supply chain management.

In Chapter 1, it is studied how retrial behavior is related with service speed and service quality. A key dilemmas faced by all service providers is how to trade off between high quality of services and timely responses. In reality it's too expensive to offer both. When either features is lacking in the service systems, customers' retrial occurs – a calling back behavior for resolving the same request. According to the reason for retrial, we classify retrial into the congestion retrial where customers abandoned in the previous call due to a slow system and the fitness retrial where customers received unsatisfactory services in the previous call due to poor quality. In this paper we want to understand retrial by connecting customers behavior with their preferences for service aspects: the speed in service

access and the quality in service delivery. We use a random-coefficient dynamic structural model on a call-by-call dataset from a hybrid service system, where three service groups of different quality are sequentially brought in to serve customers. The unique feature of such hybrid service system allows us to quantify different preferences for service speed and service quality across different customer segments while confirming that in general high service quality and speedy delivery reduce retrial. Interestingly, business customers have a stronger preference for service speed compared to private customers while private customers are more sensitive to the service quality. Realizing the different preferences across customer segments, we suggest two economical viable strategies to reduce retrial by tailoring service to meet customers' distinct preferences. The first approach, without expanding the service team, can improve business customers' surplus by 37.9% and private customers' surplus by 18.2% by wisely allocating the current service groups along the timeline; The second approach, by expanding the service team with more cheap labor resources of call center agents, they can improve customers' surplus up to a certain level. However, they should be aware that the surplus will go down if there are too many call center agents because call center agents provide timely responses but not the best quality. The greatest extent of surplus increase is 2.35% for private customers and 26.3% for business customers.

In Chapter 2, it is studied how strictness of standards and intensity of competition drive carmakers' misconduct in emission. From 2000 to 2012, though the EU Commission tightened the emission standards three times, the actual emission per vehicle didn't reduce as expected and the fraction of cars emitting more NOx than the emission standards during on-road driving actually increased. In our research, we use theoretical models to

suggest that when facing both fierce competition and tight standards, carmakers are more likely to misconduct by failing to meet standards. Using empirical analysis on 13-year records of car-by-car on-road emission, we confirm the findings in practice. We find that a 1% increase in market-level competition intensity increases the probability of misconduct by 0.58%; a 1% tightening in standard limits increases the probability of misconduct by 1.72%; and the addition of one more vehicle model substitute increases misconduct by 0.48%. Our research suggests that regulators set the strictness of standards accounting for competition intensity and monitoring effectiveness. Once the competition intensity exceeds a threshold, regulators should ensure that tightening standards are accompanied by improved monitoring to avoid an increase in misconduct and ensure social welfare increase. Our counterfactual analysis finds that the EU decision to relax standards for the next few years while working to improve monitoring effectiveness is justified. The EU action is likely to decrease the probability of misconduct by between 9.56% to 11.04%.

In Chapter 3, it is developed a forecasting method to predict demand over the life cycle ahead of products' launch. We present an approach to fit product life cycle (PLC) curves from historical customer order data and use them to forecast customer orders of ready-to-launch new products that are similar to past products. We propose three families of curves to fit the PLC: the BASS diffusion curves, polynomial curves and piecewise-linear curves. Using a large data set (133 products) of customer orders for short lifecycle products, we compare goodness-of-fit and complexity for these families of curves. Our key empirical findings from PLC fitting are that simple, piecewise-linear curves are very effective at fitting the PLC in our data set, and the products in our data rarely have a "mature" or "sustain" phase often represented in traditional PLC curves. Using time-series clustering

techniques, we cluster the fitted PLC curves into several representative curves and use these curves to generate forecasts for the products in our data set. Our forecasts result in absolute errors approximately 9% lower than the company forecasts.

Acknowledgements

Thank you, God for looking after me and guiding me. For the glory are Yours, now and forever.

I also want to thank my parents. Thank you for the support and encouragement.

Thank my husband. It's nice to have a peer who walks along the same road.

I could never thank enough for my brilliant advisors: Jan, Sunil, Gad, Achal. Without you, I won't become who I am today.

Thanks to Jiming and Jianqiang. You led me into the world of research and looked after me during my exploration.

There are also many wonderful friends during my Ph.D. study. You bring sunshine to my life and make me happy.

Thanks to whoever is reading this thesis. Thank you for the interests and time.

Table of Contents

List of Tables

List of Figures

CHAPTER 1

Understanding Customers Retrial in Call Centers: Preferences for Service Quality and Service Speed (joint with Gad Allon and Achal Bassamboo)

1.1. Introduction

One of the key dilemmas faced by all service providers is how to trade off between high quality of services and timely responses. Although customers always want a fast and accurate response when contacting a call center, it is expensive for service providers to offer both. When either features is lacking in the service systems, customers' retrial occurs – a calling back behavior for resolving the same request. Customers may retry after abandoning their previous call due to long waiting in a slow service system. Customers may also retry after receiving unsatisfactory services in the previous call due to poor quality. In service industries, the metrics closely related with retrial is called First-Contact Resolution (FCR). According to the International Customer Management Institute, FCR is the percentage of initial calls that do not require any further contacts to address the customers' requests. In other words, higher FCR means less retrial rate. Service providers actively seek solutions to increase FCR because 1% improvement in FCR can reduce 1% operation cost, improve 1% customer satisfaction and 1% $\sim$ 5% employee satisfaction, increases selling opportunities and retain customers for longer term.[1] Though FCR is one

[1] http://sqmgroup.com/call-center-first-contact-resolution-benchmarking-study

of the most important operation metrics in services, the industry average is merely 70% (Babel (2014)) because it is not easy to cost-effectively provide both good quality and service speed. To reduce retrial in an economic viable manner, service providers need first to clarify customers' preferences regarding good quality and timely services, and then to balance the service offerings between these two aspects. Ultimately, this paper aims 1) to capture the drivers behind customers' retrial behavior, (2) to understand customers' preferences between service speed and quality, and (3) to enlighten the research of service priority and staffing choices related with the service speed and quality.

According to the reason for retrial, either due to long waiting or due to poor service quality, we classify retrial into two types. We will refer to the first type as the congestion retrial.

Congestion retrial occurs when customers call back for the same request after abandoning their previous call. A service system equipped with insufficient number of agents leads to congestion, causing long waiting time for the customers. Eventually, after the customers' patience is exhausted, customers choose to abandon the call without getting their request resolved, and have to call back later.

In our research, we use a call-by-call dataset from a medium-sized Israeli bank which adopts a hybrid service model.[2] This hybrid model incorporates three service groups which differ in the quality of services they offer and their speed in reaching customers. Illustrated in Figure 1.1, the hybrid service model includes three service groups: the target agent group, the branch backup group and the general call center group. The target agent group is composed of customers' private bankers. These private bankers know most about

[2]We thank the Service Enterprise Engineering (SEE) lab at the Technion for generously providing us with the call center records.

the customers' information and are expected to deliver the best quality of services. The branch backup group is composed of personal bankers of other customers from the same physical branch. The last service group is the large general call center group which is composed of the regular call center agents. When a customer calls in, he is first placed in the queue leading to the target agent group. If his call doesn't get answered within one minute, he will also take a spot in the queue to the branch backup group. After another minute, if both the target agent and the branch backup team cannot answer the phone, then the customer will take another spot in the queue to the call center group. At this point of time, the customer takes one spot in each of the three queues and the first available group will provide the service. By first allocating the customer into the queue of target agents and later bringing in the help from other service groups, the purpose of the hybrid model is to combine high-quality personal banking services with a quick response. However, this brings up a new concern: a customer who could have been served by the target agent are now served by a call center agent. If the call center agent is unable to provide good quality, the customer will retry later. In fact, an industrial survey about contact centers (Dimension Data, 2009) states that poor agent capabilities and lack of access to customer information are the top two reasons for retrial.

Here we define the second type of retrial related with service quality. **Fitness retrial** is customers' behavior of calling back after receiving poor service quality in the previous call. Compared with the congestion retrial, in which customers call back because they did not speak with a service agent in the previous call, the fitness retrial occurs when customers speak with service agents but are not happy with the service quality.

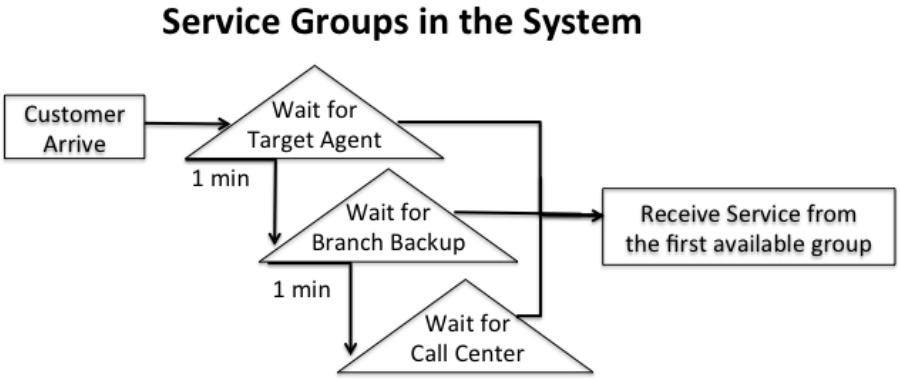


Figure 1.1. Call Center Service Groups.

In this paper we want to understand congestion retrial and fitness retrial by connecting customers' behavior with their preferences for service aspects and further to provide suggestions to improve services.

To begin with, we use a Probit regression to explore how the probability of retrial is connected with with aspects of services delivered in the previous call. Our results suggest that the decisions of retrial are significantly impacted by the outcome of the previous call (whether the caller abandoned or was served), the service provider and the length of the provided service.

Then we turn to a random-coefficient dynamic structural model to capture customers' behavior responding to the service speed and quality. Fitting the structural model with the call-by-call dataset, we found that in general customers value good service quality and timely responses. The services offered by the target agents are perceived as the best quality while the call center agents offers the most ordinary quality. Across two

customer segments, we find that in terms of retrial, business customers are less sensitive to the service groups but have a stronger preference for timely responses compared to the private customers.

Lastly, we use counterfactual analysis, to suggest economic viable approaches to improve services. We know that in order to reduce retrial, the ideal service system immediately serves every customer with their target agent upon arrival. However, this strategy is very costly because such high-quality and timely services require a large number of target agents who are expensive to train and hire. Hence we suggest two strategies that tailoring services to meet customers' distinct preferences. In our first approach, we suggest improving customers' surplus by efficiently allocating the service teams along the timeline based on customers' preferences. Instead of the generic one-minute time lag between adding new service groups in the original system, we suggest a 5-second time lag for business customers and a 25-second time lag for private customers. The time lag plays a role in trading-off between timely responses and good quality. A large time lag increases the likelihood of accessing good quality but causes longer waiting while a short time lag offers the reverse. Without expanding the service teams, we improve business customers' surplus by 37.9% and private customers' surplus by 18.2%. In our second approach, we suggest to improve services by hiring more cheap resources, the call center agents. The call center agents reduces' customers waiting cost but also lowers the chances to get good quality. After adding a certain number of call center agents, the surplus will start decreasing when the marginal loss in service quality outweighs the marginal gain from shortening online waiting. Hence the service provider need to understand customers' preferences in speed and quality before deciding the number of call center agents to hire. In the current

service teams, business customers' surplus will be increased by 26.3% with 5 more call center agents and the private customers' surplus will be increased by 2.35% with 3 more call center agents.

Our paper offers several contributions. First, to our knowledge, our paper is one of the first to understand how customers' retrial behavior is impacted by service quality and speed. In particular, utilizing the rich data from the hybrid service model, we are able to disentangle the effects on retrial from two service aspects: timely responses and good quality. Based on our classification of retrial, the fitness retrial and the congestion retrial, we empirically structures the mechanism of retrials with customers' preferences for speed and quality of services. Secondly, this is the first empirical model we are aware of in the service domain that comprehensively captures customers' behavior when they are online (waiting in line or talking with an agent) and offline (waiting outside of the service system). Most empirical service models only study customers behavior when the customers are observed online. However, in our model, we acknowledge that customers also make decisions in the offline stage of whether to retry and return to the service system. Service providers should realize their online service offering affects such offline decisions and the customers offline decisions will also impact the online service operations. Thirdly, we demonstrate the importance of accounting customers' preferences when making operation decisions in a multi-skill setting. Specifically, we show adding agents, an action usually conceived beneficial to the customers, may in fact hurt customers' overall utility beyond a certain point. Given the fact that adding a general call center agent means trading off good quality for a quick response, eventually customers will be hurt when their loss in quality cannot be compensated by the quick service speed. Lastly, our study develops a

methodology framework to analyze customers' preferences in speed and quality and the mechanism between customer behavior and offered services. This methodology framework can be applied to a wide range of digital and virtual service industrial practices beyond call center management.

The results of our analysis indicate that customers across different segments vary in their preferences between service quality and service speed. To be specific, this research highlights that business customers have a stronger preference for timely responses but care less about the service quality compared with the private customers. Hence service providers need to first understand how their customers' retrial impacted by the quality and speed in the service system, and then align the service features with customers' interests. We suggest several economic viable plans to improve the service system: (1) without expanding the service team, we can efficiently allocate the current service groups along the timeline for different customers segments; (2)we can increase the customers' surplus with cheaper resource by hiring call center agents, as long as the gain from shorter waiting time outweighs the loss in quality.

In the rest of the paper, we discuss literature review in Section 1.2, data and retrial in the call center in Section 1.3, the dynamic structural model, its estimation and results in Section 1.4, 1.5 and 1.6 respectively and the counterfactual analysis in Section 1.7. In the end, the conclusion is summarized in Section 1.8.

1.2. Literature Review

Studies in retrial date back into a long history: there are many theoretical models that incorporate retrial behavior. Initially, studies assumed that retrial occurs mainly

after abandonment due to unavailability of servers and customers' impatience in waiting. (Sze (1984), Kulkarni (1983), Falin (1995), Artalejo and Lopez-Herrero (2000), Aguir et al. (2004), Aguir et al. (2008), Reed and Yechiali (2013), Mandelbaum et al. (1999),Mandelbaum et al. (2002),Shin and Choo (2009)). This type of retrial is what we call the congestion retrial. Gradually, researchers realized that beyond the unavailability of servers, the quality of servers may also impact retrial. Aissani (1994) and Kulkarni and Choi (1990) model retrial due to unreliable of the server such as breakdowns. De Véricourt and Zhou (2005) considers retrial happens when customers' requests are not completely resolved by the service representatives. This type of retrial is what we call as fitness retrial. Ding et al. (2015) also distinguish fitness retrial from the congestion retrial and call the first one reconnect and the second one redial. For furthur retrial modeling work, we refer interested readers to the well-summarized literature surveys (Yang and Templeton (1987), Falin (1990), Falin and Templeton (1997), Artalejo (1999),Gans et al. (2003), Aksin et al. (2007), Artalejo (2010)).

One group of theoretical work related to our study is to model congestion retrial as a strategic behavior of customers. Customers will choose an optimal retrial rate considering both the system condition and the others' choices in the system (Elcan (1994)). For most of the cases, the self-interested consumers' optimal retrial rate is not equal to the social optimal rate (Cui et al. (2014)). Some studies suggest regulating retrial by changing the service system design, such as imposing tolls on retrials (Hassin and Haviv (1996)) or requesting advance payment (Armony et al. (2009)). While the listed research only studies the congestion retrial and focuses on customers' preference for quick responses, we

study the fitness retrial and the congestion retrial when customers make decisions based on service speed as well as service quality.

Another group of theoretical work related to our study focuses on the speed-quality trade-off faced by the service providers. Anand et al. (2011) considers spending longer time with customers increases the service quality yet slows the service delivery and leads to long waiting time for customers. Zhan and Ward (2013) considers the trade-off between agents' speed in handling calls and capability at resolving customers' inquiries. They then develop a threshold routing rule to allocate customers' calls to agents of various service speed and quality. For our research, we use an empirical approach to investigate the trade-off between service speed measured by the average waiting time in line and the quality of service agents. Then we provide suggestions for improving the system based on the quantified customers' preferences for service speed and quality.

Despite the abundance of theoretical work, there is very few empirical studies about retrial. Shen (2010) said "little is known about the actual retrial behavior of customers". The two empirical retrials studies we are aware of are Hoffman and Harris (1986), which estimates the call volume considering the presence of congestion retrial and Ding et al. (2013), which estimates the call volume considering the presence of both congestion retrial and fitness retrial. Instead of estimating the volume of calls or percentage of retrial calls, we want to understand what drives retrial behavior from the service aspects.

There are several empirical papers to understand customer behavior in the call center. Aksin et al. (2013) studies the abandonment behavior of customers in the call center. Yu et al. (2016) studies the impact of delay announcement on abandonment behavior. Both of the papers use structural models to capture the customer behavior in the call center.

Structural models origin from economics (Rust (1987)) and now are adopted in operations management. Our model is innovative because it comprehensively captures customers' behavior when they are online (waiting in line or talking with agent) and offline (waiting outside of the service system). The previous literature mainly addresses customers' online behavior such as abandoning or waiting in the line. However, our model acknowledges that customers also make decisions in the offline stage of whether to retry and return to the service system. Furthurmore, we connect the retrial behavior with the speed-quality trade-off to instruct service providers to improve services.

1.3. Data and Retrial in Call Center

In this section, we first describe the dataset used in our research, quantitaively define retrial and then present a descriptive analysis of customer retrial behavior in the call center. The dataset is composed of call-by-call records from a medium-sized Israeli bank.

The data provides us with information about the identity of customers and each customer's online session. Customers are uniquely identified with the customer id and are classified into two segments, business customers and private customers. The customer's session starts from his initial contact with the call center regarding a particular issue. Upon his arrival, he enters into the **"online waiting"** period where he waits for an agent. In this period, we observe his waiting time and abandonment decisions. If the customer chooses not to abandon and waits until an agent becomes available, he will be served. For served customers, we observe the service time and the service group (note that a customer can be served by the target agent group, the branch backup group, or the call center group). Up to this point, we observe the customer's entire online system flow as

illustrated in Figure 1.2. The online system starts upon the customer's arrival, continues throughout his online waiting period, and ends either when he chooses to abandon the call or when the service has been received. Upon the customer leaving the online system, he enters into the **"offline waiting"** period. During this period, the customer considers whether to retry the online system. We then observe his offline waiting time and the retrial decisions. If a customer chooses to retry, he returns to the online system. However, if a customer chooses not to retry and has no action for a long time in the offline system, we consider for this current issue he will no longer contact the call center again and mark it as the termination of his session. The offline system starts from the moment customers leave the online system and ends when customers retry or terminate their session. The entire customer session can iterates between the online system and the offline system several times if retrial occurs.

An overview of the call center data is presented here. Altogether the dataset recorded 269,035 calls. 78.91% of the callers were private customers and the others were business customers. In terms of the outcome of calls, 18.91% of private callers chose to abandon and 30% of them were served by the call center group, 22% by the branch backup group and 28.91% by the target agent group. 18% of business callers chose to abandon and 26% of them were served by the call center group, 30% of them were served by the branch backup and 26% were served by the target agent group. In the online system, for private customers, the average online waiting time is 51 seconds with the interquartile range from 7 seconds to 75 seconds and the average service time is 129.1 seconds with the interquartile range from 26 seconds to 167 seconds. For business customers, the average online waiting time is 49 seconds with the interquartile range from 6 seconds to 67 seconds and the

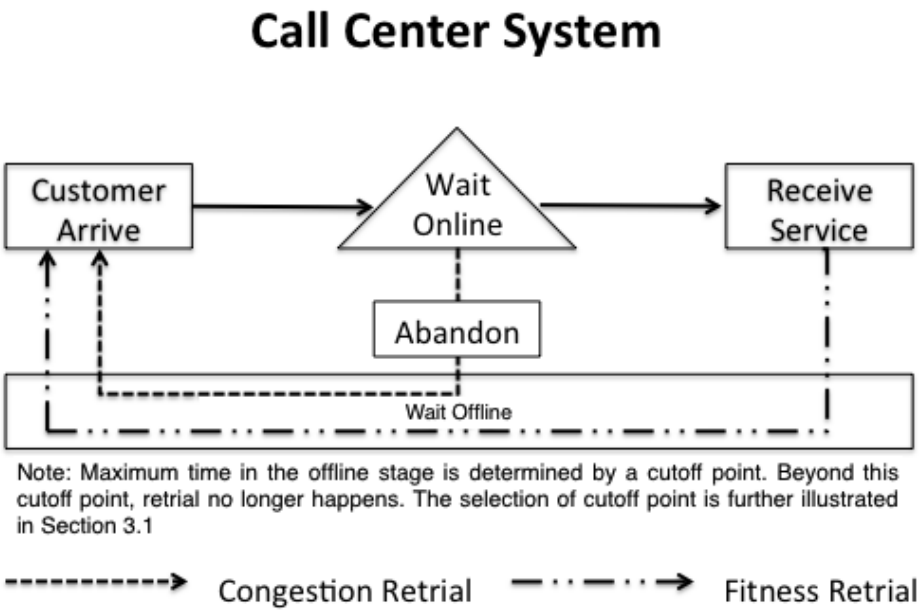


Figure 1.2. Call Center System.

average service time is 92.2 seconds with the interquartile range from 18 seconds to 124 seconds.

1.3.1. Quantitative Definition of Retrial

In this section, we quantitatively define retrial. In Section 1.1, we introduce retrial as customers' behavior of calling back for the same request. If we do not observe a customer calling back to the system, then he has no retrial. Based on this criterion, 65.1% of calls are not observed with follow-up calls. The remaining 34.9% of calls are considered to be the candidates of retrial behavior. To qualify as retrial, the callbacks must be for the same request as the one in the previous call. Though we do not observe direct information

about the content of calls, we detect retrial based on the following justification and two statistical tests. Customers contact the call center either due to new requests or retrial. Suppose the new requests occur in a constant manner for each customer, then, without retrial, a customer contacts the call center with a constant arrival rate. However, when retrial occurs, the arrival rate will be higher because now calls are from both new requests and retrial. The following two statistical tests are build on the above rationale to detect retrial.

We first use the Kolmogorov-Smirnov test to detect the length of retrial window for each customer. Suppose, for customer i, his regular arrival rate is λ_i. When retrial occurs, the arrival rate is λ_i^r. We identify the retrial threshold (i.e., retrial window length) T_i^r based on whether the arrival rate during the retrial window is significant higher than the regular one. Specifically, we use Kolmogorov-Smirnov test to examine whether the CDF (cumulative distribution function) of calling gaps in the retrial window is higher than the CDF of gaps out of the window. The testing hypothesis is that $H_0 : \lambda_{(0,T_i^r)} = \lambda_{(T_i^r,\infty)}; H_a : \lambda_{(0,T_i^r)} > \lambda_{(T_i^r,\infty)}$ where $\lambda_{(0,T_i^r)}$ is the arrival rate from the moment the customer ended the previous call to T_i^r later.

We then use the change point detection test to detects the start of the retrial window. As discussed above, retrial means abnormal high calling frequency. Hence we identify the start of retrial with detected surges in calling frequency using a change point detection method. Calls corresponding to surges in calling frequency are considered as retrial. If there are more than one consecutive retrial calls, the second one will not be detected given the calling frequency is already in the high-zone and there is no detectable surge. Hence, the detected calls mark the beginning of retrial calls, and the retrial window determines

whether the follow-up calls are retrial. If the follow-up calls happen within the retrial window, they are also considered as retrial.

Recognizing customers have different calling frequency, our two methods, the change point detection method and the Kolmogorov-Smirnov test, are designed to detect individual behavior. The change point detection method is customized to individuals' calling behavior. It detects surges in calling frequency for a particular customer compared to his own calling behavior. Then, the Kolmogorov-Smirnov test is carried out for each individual to detect a customer's own unique retrial window.

Now we formally define retrial.

DEFINITION 1 Retrial: For customer i, a call at time t is retrial if either the call at time t is identified as a surge or the call following a previous surge-identifieded call ended at time j where $j > t - T_i^r$.

DEFINITION 2 Retrial can be classified into two types given the outcome of the previous call.

Congestion Retrial: For customer i, a call at time t is congestion retrial if the call is identified as retrial and it's previous call is abandoned.

Fitness Retrial: For customer i, a call at time t is fitness retrial if the call is identified as retrial and it's previous call is served.

Based on the quantitative definition of retrial, the retrial behavior observed in the data is summarized in Figure 1.3. Among 269,035 calls observed in the record, 34.9% of them had a follow-up call from the same customer. 23.1% of calls are identified as retrial. Given the outcome of their last call, 4.7% of calls were congestion retrial and 18.4% were fitness retrial.

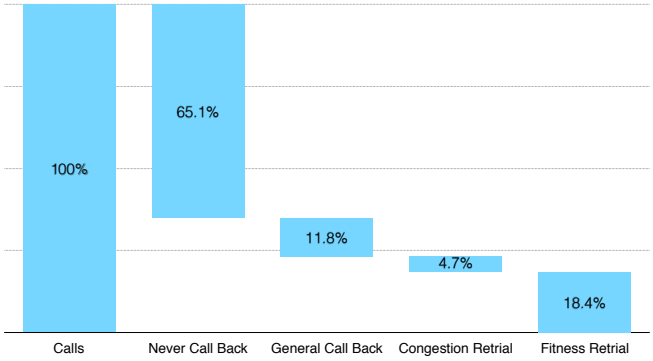


Figure 1.3. Summary of Retrial

1.3.2. Data Explorative Analysis on Retrial

In this subsection, we explore what factors relate to retrial. This exploration helps to construct the structural model in Section 1.4. First, we investigate how retrial is affected by whether customers received service or not in the previous call. If they received service, we then study how service experiences lead customers to conduct retrial.

First we discover that the probability of retrial is greater when the customer is not served in his previous call. In (1.1), a Probit regression describes the relationship by linking the dummy variable of retrial with the dummy variable of outcome in the previous call. $\Phi(\cdot)$ is the CDF of the standard normal distribution. The exact form of regression is

$$\begin{aligned} &Pr(\text{Retrial} = 1|\text{Outcome}) = \\ &\Phi(\alpha + \beta \times \mathbb{I}(\text{Outcome} = \text{No Service Received in the Previous Call})). \end{aligned} \tag{1.1}$$

In Table 1.1, the estimated parameters are α, the base impact on retrial regardless the service outcome, and β, the marginal impact if the previous call is abandoned. We find that if a customer abandons his last call and receives no service, the z-score for retrial is

Parameter	Estimate	Std. Error	z value	P-value
α	-0.710	0.003	-274.42	$< 2 \times 10^{-16}$
β	0.220	0.006	39.63	$< 2 \times 10^{-16}$

Table 1.1. Estimation of Probit Regression in Equation (1.1)

increased by 0.22. In other words, the estimated probability of retrial is 0.31 for those not served customers and 0.24 for those served customers. Hence, the first factor that influences retrial is whether the customer receives a service or not in his previous call.

Secondly, we find that retrial is also impacted by the service characteristics: the service group and the length of service time. We denote the service time as S_{time} and the service group as S_{group}. Recall the service groups l can be the target agent group, the branch backup group or the call center group. The Probit regression to characterize such relationship is

$$Pr(\text{Retrial} = 1 | S_{group}, S_{time}) = \Phi(\beta_{\text{Time}} \times S_{time} + \sum_{l} \beta_l \times \mathbb{I}\{S_{group} = l\}). \tag{1.2}$$

In Table 1.2, the estimation results suggest that service groups and service time are significantly related to the probability of retrial. Among service groups, the target agent group reduces retrial the most effectively, given that the z-score is lower by 0.86. Considering the target agents are the personal banker of the customers, it is reasonable to conclude that they provide the best service quality. Regarding service time, one more minute of service reduces the probability of retrial, given that the z-score is lower by 0.005. Hence, two additional factors that influence retrial are service time and service groups.

In this section, the results of the Probit regressions suggest retrial is impacted by three factors: whether a customer received service or not in his previous call, the service group

Parameter	Estimate	Std. Error	z value	p-value
$\beta_{Call\ Center}$	-0.529	0.005	-116.19	$< 2 \times 10^{-16}$
$\beta_{Branch\ Backup}$	-0.816	0.006	-140.40	$< 2 \times 10^{-16}$
$\beta_{Target\ Agent}$	-0.859	0.007	-130.38	$< 2 \times 10^{-16}$
β_{Time}	-0.005	0.001	-4.75	$< 2 \times 10^{-6}$

Table 1.2. Estimation of Probit Regression in Equation (1.2)

that provided services and the service time. In Section 1.4, we explore this further using a structural model that includes the three factors. Firstly, the service reward is positive if a customer receives a service or zero otherwise. Secondly, the service reward depends on the service group. Lastly, the service reward is larger when the service time is longer.

The essential reason for us to adopt the structural model is to use operations and economic theory so that we can clarify how institutional and economic conditions affect retrial and to simulate counterfactuals.

First, a structural model of retrial can estimate unobserved behavioral parameters that could not otherwise be inferred from the original non-experimental data. For example, we cannot infer the customers' unit cost of waiting online and offline from the above explorative analysis. We cannot measure the relative service reward offered by different service groups. Those parameters, though essential to guide the call center service system design, are not identifiable in the descriptive analysis. The structural model, on the other hand, provides the institutional and economic conditions regarding the retrial behavior and allows us to quantify the waiting cost and service rewards. The estimation results are presented in Section 1.6.

Secondly, the structural model provides us opportunities to simulate counterfactuals. One goal of our paper is to offer strategies to the service provider to reduce retrial. Therefore, we should examine the performance of the call center services under different system designs. The design simulation, however, is not achievable through descriptive analysis. With descriptive analysis, it can only examine the design which generates the observed data. With the structural models, we can test how alternative service designs perform in terms of the service speed, service quality, and customer surplus. The counterfactual analysis is presented in Section 1.7.

1.4. The Structural Model

To understand how customers abandon and retry in the call center system, we present a dynamic structural model. The dynamic structural model captures customers' decisions regarding abandonment and retrial as they try to maximize their total utility from resolving an issue. The dynamic decision process starts from the first time a customer contacts the call center system, i.e. arrives online, for that particular issue. While the customer waiting in the online stage for a service, he decides whether to abandon or to keep waiting online by trading off the online waiting cost against the potential service reward. Note that the reward depends on which service group provides the service: the target agent, the branch backup agent or the call center agent. A customer will move to an offline stage in the system after he abandons the queue or finishes a service. In the offline stage, the customer decides whether to retry by trading off potential gains from further services or some cost for leaving the issue partially resolved. If a customer retries, he will move back to an online stage. If a customer stays offline for a sufficiently long

period without retrying, we consider the customer will not pursue the issue again. And this marks the end of the customer's dynamic decision process for resolving the issue for which he initially contacts the call center.

We call this entire dynamic decision process described above as an **episode**. For each episode, it starts with an online stage, then alternates between the online stage and the offline stage, and ends when the caller waits a sufficient long period (length of retrial window) in the offline stage without retrying.

In order to set up the dynamic model, we next discuss the decision process during the online and offline stages. For each stage, we outline the decisions the callers can make and the utility function associated with the callers' decisions.

1.4.1. Structural Model for the Online Waiting Stage

For each caller, the online waiting stage covers the entire period from the moment he starts to wait online till the moment when he chooses to abandon or when he finishes a service with an agent. We assume a caller does not choose to abandon while talking to an agent. Hence, the actual decision-making process is from the moment he starts to wait online and ends either when he chooses to abandon or when he starts a conversation with an agent. We divide the total length of the online decision making process into small time periods of equal length. Similar to Yu et al. (2016) and Aksin et al. (2013), we use 5 seconds for the length of each time period. Recall each episode may have multiple online stages. If retrial occurs in the dynamic decision process, we use index k to denote the

kth online waiting stage. By denoting the total number of periods during the kth online waiting stage as N_k^w, we index each period by t where $t = 1, 2, .., N_k^w$. [3]

The value-to-go for caller i at time period t during his kth stage at the online waiting stage is given by

$$\Psi_{ikt}(\epsilon_{ikt}(a_{ikt}), a_{ikt}) = \psi_{ikt}(\theta^w, a_{ikt}) + \epsilon_{ikt}(a_{ikt}). \tag{1.3}$$

where a_{ikt} is caller i's decision about whether to abandon at time period t or wait until time period $t+1$ during his kth online stage. The function $\psi_{ikt}(\theta^w, a_{ikt})$ is the nominal value-to-go of caller i. It is important to note that $\psi_{ikt}(\theta^w, a_{ikt})$ includes not only the nominal utility at time period t but also the utility beyond period t. The last term $\epsilon_{ikt}(a_{ikt})$ is the idiosyncratic shock which is observed only by caller i but not the researchers. The idiosyncratic shock captures callers' lack of adherence to pure rational decision making or callers' private information. Note that all these factors may vary with callers' decisions on whether to abandon the system. Thus, we let the idiosyncratic shock be a function of caller i's action a_{ikt}. We assume the idiosyncratic shock with mean set to 0 follows a Gumbel distribution as $\epsilon_{ikt}(a_{ikt}) \sim Gumbel(\gamma\theta^w, \theta^w)$ where γ is the Euler-Mascheroni constant. We assume that $\epsilon_{ikt}(a_{ikt})$ is independent across time periods, the online waiting stages, callers' decisions, and the nominal value-to-go. It is important to note that $\psi_{ikt}(\theta^w, a_{ikt})$ is the expected utility of time period t and beyond, over the idiosyncratic shock. Hence it is a function of only the idiosyncratic parameter θ^w rather than any realization of the shocks.

[3]Recall an online waiting stage may have follow-up online waiting stages if the caller chooses to retry during the offline stages, the index of time periods starts from 1 for each online waiting stage.

At period t, caller i makes a decision whether to abandon or to wait until period $t+1$ to maximize his utility. In particular, the optimal action $a^*_{ikt} \in \{\text{wait online}, \text{abandon}\}$ is given by

$$a^*_{ikt} = \arg\max \ \Psi_{ikt}(\epsilon_{ikt}(a_{ikt}), a_{ikt}).$$

If caller i decides to abandon, he will do so immediately at the beginning of time period t and move to the offline waiting stage in the next period. In this case, the nominal value-to-go at time period t is given by χ_{ik0}, the expected total utility of caller i from the kth offline stage. Since the episode starts from an online waiting stage and then alternates between the online and offline stages, a caller will enter into the kth offline stage after leaving the kth online stage and χ_{ik0} is his aggregated offline utility function measured at the beginning of the kth offline stage. The functional form of χ_{ik0} is specified in details in the following subsection when we introduce the structural model for the offline waiting stage.

If caller i chooses to wait online at period t, he incurs one unit waiting cost c^w_i and stays in the queue until the next time period. Since each caller may have different preferences for online waiting, we incorporate the randomness of unit online waiting cost with a random variable z^w_i being i.i.d $N(0,1)$. To ensure the online waiting cost is positive, we write it as $c^w_i = |\mu^w + \sigma^w z^w_i|$. As he waits at period t, he may get a service from group $l \in \{\text{call center, branch backup, target agent}\}$ with probability p_{lt}. The probability of being served by group l at period t conditional on the fact that the caller is still in the queue at time period t is defined as

$$p_{lt} = \frac{F_l(t+1) - F_l(t)}{1 - F_l(t)} \tag{1.4}$$

where $F_l(t)$ is the cumulative distribution function (CDF) of the waiting time of callers who receive a service from group l. By computing p_{lt} from the observed dataset, the expected payoff from being served is $\sum_l (p_{lt} R_{ilk})$ where R_{ilk} is the service reward from the group l. The service reward R_{ilk} depends on the service group l, and the length of the anticipated service time T_{ik} for caller i's issue in the kth online stage. The longer the anticipated service time, the more complicated and important the issue is to the caller. This correlation is also supported by the exploration analysis in Section 1.3.2. Hence we consider the service reward is positively correlated with the service time. Denoting the service payoff per unit of anticipated service time from group l as r_l, we write the anticipated service payoff as $R_{ilk} = r_l T_{ik} + \chi_{ik0}$. The χ_{ik0} captures the nominal value-to-go for caller i after he finishes the current service and enters into the offline stage where he has the option to retry. This term will be defined mathematically in the next subsection about the offline structural model. On the other hand, there is a probability $(1 - \sum_l p_{lt})$ that the caller may not get served during this online waiting period. Then he will decide again whether to abandon or continue to wait as he enters the next time period $t+1$. In this case, the caller get $(1 - \sum_l p_{lt}) V_{ikt}$ where V_{ikt} denotes the expected total future utility of caller i beyond time period t conditional on the fact that he decides to wait until time period $t+1$. We refer to it as the aggregated future online utility function. It is given by

$$V_{ikt} = \mathbb{E}[\max_a \Psi_{ikt+1}(\epsilon_{ikt+1}(a), a)].$$

To summarize what we've discussed above, the nominal value-to-go function associated with customer i's action at period t of the kth online waiting stage is

$$\psi_{ikt}(a_{ikt}) = \begin{cases} -c_i^w + \sum_l (p_{lt} R_{ilk}) + (1 - \sum_l p_{lt}) V_{ikt} & : a_{ikt} = \text{wait online} \\ \chi_{ik0} & : a_{ikt} = \text{abandon} \end{cases} . \tag{1.5}$$

1.4.2. Structural Model for the Offline Waiting Stage

Once a caller finishes a conversation with an agent or chooses to abandon in the kth online waiting stage, he immediately enters into the kth offline waiting stage. In the offline waiting stage, he can choose to retry or to keep waiting offline. By retrying, the caller will return to the online stage. Hence, the offline waiting stage ends either when the caller decides to retry or when the caller's offline waiting time at the current offline stage reaches the retrial threshold. The statistical process for selecting the retrial threshold is explained in Section 1.3.1. Recall a caller's episode starts from the first time he joins the online waiting stage and then alternates between the online and offline stages, and the episode ends whenever during one offline waiting stage the waiting time reaches the retrial window length. It marks the end of his pursue to solve the issue he called for in this episode. The total length of the offline waiting stage is divided into small time periods of equal length. We use 1 hour for the length and index the offline time period by τ. Recall each episode may have multiple offline stages. If retrial occurs in the dynamic decision process, we use index h to denote the hth offline waiting stage. By denoting the total time periods of the hth offline waiting stage as N_h^o, the period index τ is equal to $1, 2, ..., N_h^o$ where N_h^o is determined by the retrial threshold and 1-hour interval. Note that

one episode may contain multiple offline waiting stages if a caller decides to retry in the offline stages. The period index τ starts from 1 for each offline waiting stage.

The value-to-go of caller i at time period τ during the hth offline stage is given by

$$\Phi_{ih\tau}(\eta_{ih\tau}(d_{ih\tau}), d_{ih\tau}) = \phi_{ih\tau}(d_{ih\tau}) + \eta_{ih\tau}(d_{ih\tau}). \tag{1.6}$$

where $d_{ih\tau}$ is caller i's decision about whether to retry at time period τ or wait until time period $\tau + 1$. The function $\phi_{ih\tau}(.)$ is the nominal value-to-go of caller i which includes not only the nominal utility at time period τ but also the utility beyond period τ. The last term $\eta_{ih\tau}$ is the idiosyncratic shock which is observed only by caller i but not by the researchers. For the same reasons stated in the structural model for the online waiting stage, this offline idiosyncratic shock is a function of callers' actions. Furthermore, we assume the idiosyncratic shock with its mean set to 0 follows the Gumbel distribution, $Gumbel(\gamma\theta^o, \theta^o)$ where γ is the Euler-Mascheroni constant. The shocks $\eta_{ih\tau}$ are independent across different time periods, different offline stages, different decisions and also independent from the nominal value-to-go $\phi_{ih\tau}(.)$. The nominal value-to-go captures the expected utility of time period τ. Hence it is a function of only the idiosyncartic parameter θ^o rather than any realization of the shocks.

In the beginning of each period τ, caller i makes a decision whether to retry or to wait until the next time period $\tau + 1$ to maximize his utility. In particular, the optimal action $d^*_{ih\tau}$ is given by

$$d^*_{ih\tau} = \arg\max \ \Phi_{ih\tau}(\eta_{ih\tau}(d_{ih\tau}), d_{ih\tau}).$$

Recall the aggregated future offline utility χ_{ik0} caller i gets when he leaves the kth online waiting stage, its detailed formula is $\mathbb{E}[\max_d \Phi_{ik1}(\eta_{ik1}(d_{ik1}), d_{ik1})]$, in words, the value-to-go at the beginning of the kth offline waiting stage.

Next we clearly specify the nominal value-to-go for caller i associated with his action $d_{ih\tau}$. If a caller chooses to retry, he will do so immediately at the beginning of time period τ and will move to the $(h+1)$th online stage in the next period. Since the episode starts from an online waiting stage and then alternates between the online and offline stages, the caller will enter into the $(h+1)$th online stage after leaving the hth offline stage. Hence, the nominal value-to-go he receives at this period is $\mathbb{E}(V_{i,h+1,0})$, the aggregated future online utility. It measures the value-to-go at the beginning of the online stage, i.e.,

$$V_{i,h+1,0} = \mathbb{E}[\max_a \Psi_{i,h+1,1}(\epsilon_{i,h+1,1}(a), a)].$$

If a caller chooses to wait offline, he faces the offline waiting cost of the current period and then he decides again whether to retry or to continue waiting as he enters the next time period. The offline waiting cost captures the disutility a caller i incurs as he leaves the calling issue unsolved or partially solved as time passes. The waiting cost first comes from the regular waiting cost c_i^o that shows up in each period of the offline waiting stage. Since each caller may have a different offline waiting cost, we incorporate randomness in the unit offline waiting cost with a random variable z_i^o being i.i.d $N(0,1)$ across all customers. To ensure the offline waiting cost is positive, we write it as $c_i^o = |\mu^o + \sigma^o z_i^o|$. Moreover, the waiting cost jumps periodically with a daily pattern. In bottom panel of Figure 1.4, we found a clear daily jump in calling back probability. We consider there is a time-dependent offline cost ν. One explanation of the time-dependent cost is that a

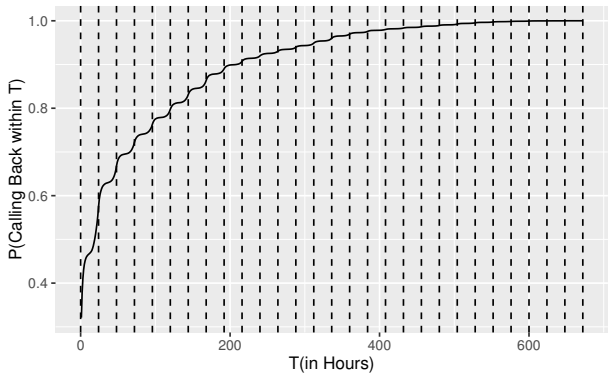


Figure 1.4. Calling Back Behavior (dashed lines to indicate daily cycles)

caller may choose to contact the call center only at some specific time of the day due to the convenience. For example, a private caller may prefer to contact the call center only during his lunch break. Hence we are more likely to observe his retrial behavior during lunch time for the next following days. Another explanation of the time-dependent cost is that a caller may face a penalty for leaving the issue unsolved in a daily cycle. In order to avoid such penalty, the caller is more likely to retrial just before the he incurs the daily penalty. We set this periodically-occurred waiting cost by $\nu\mathbb{I}(\tau \in \{1 \text{ day}, 2 \text{ day}, \dots\})$. If ν is equal to 0, that means the caller doesn't have any time of the day preferences for contacting the call center. If ν is very large, that means the caller i is restricted to a limited window of time to contact the call center everyday.

To summarize what we've discussed above, the nominal utility function at period τ of the offline waiting stage is

$$\Phi_{ih\tau}(d_{ih\tau}) = \begin{cases} V_{i,h+1,0} & : d_{ih\tau} = \text{retry} \\ -c_i^o - \nu\mathbb{I}(\tau \in \{1 \text{ day}, 2 \text{ day}, \dots\}) + \chi_{ih\tau} & : d_{ih\tau} = \text{wait offline} \end{cases}. \tag{1.7}$$

where $\chi_{ih\tau}$ denotes the expected total future utility of caller i beyond time period τ conditional on the fact that he decides to wait until time period $\tau + 1$. We refer to it as the aggregated future offline utility function. It is given by.

$$\chi_{ih\tau} = \mathbb{E}[\max_{d} \Phi_{ih\tau+1}(\eta_{ih\tau+1}(d_{ih\tau+1}), d_{ih\tau+1})].$$

Terminal Condition marks the end of an episode. Recall an episode is a dynamic decision process involving a caller who contacts the call center to solve an issue, and the episode ends when the caller stops pursuing the issue. Specifically, the episode ends when the caller does not contact the call center for a period of time exceeding the retrial window. The terminal condition is to set the future utility beyond the retrial threshold in a single offline stage as 0. Mathematically speaking, the terminal condition of the dynamic program is

$$\chi_{i,h,T_i^r} = 0, \forall i, \forall h \tag{1.8}$$

Note T_i^r is the length of retrial window for customer i as introduced in Section 1.3.1.

We next explain how to estimate the above structural model of customers' behavior in the service system.

1.5. Estimation

In this section, we discuss the estimation strategy for the dynamic structural model constructed above. Specifically, we use the maximum likelihood estimation to obtain the parameters in our dynamic structural model. The main idea is to first compute the

likelihood of observing a sequence of decisions, (i.e. waiting online/ abandoning during online stages and waiting offline/retrying during offline stages) for each caller, and then to compute the overall likelihood of all callers for a given parameter set. We estimate the model parameters by maximizing the overall likelihood function, which best explains the observed behavior.

To construct the likelihood function, we first characterize the probabilities of callers' decisions whether to wait or abandon during the online waiting stage and whether to wait or retry in the offline waiting stage. We define $P^w_{ikt}(a_{ikt}; c^w_i(Z^w), \theta^w, r_l, \forall l)$ as the probability that caller i chooses the action a_{ikt} by time period t during his kth online waiting stage and $P^o_{ih\tau}(d_{ih\tau}; c^o_i(Z^o), \theta^o, \nu)$ as the probability that caller i chooses the action $d_{ih\tau}$ by time period τ during his hth offline waiting stage. Using the corresponding choice probability at each time period, we obtain the probability of observing the sequence of choices over time for each caller. The following proposition characterizes the probability of the callers' decisions.

PROPOSITION 1.1 Let the idiosyncratic shock $\epsilon_{ikt}(a_{ikt})$ be i.i.d. $Gumbel(\gamma\theta^w, \theta^w)$ distributed, and $\eta_{ih\tau}(d_{ih\tau})$ be i.i.d. $Gumbel(\gamma\theta^o, \theta^o)$ distributed. With the nominal value-to-go specified in Equation (1.5) for the online waiting stage and in Equation (1.7) for the offline waiting stage, we have the closed forms for the online and offline aggregated future utility function

$$V_{ikt} = \theta^w log(1 + exp(\frac{\psi_{ikt+1}(a_{ikt+1} = \text{“wait online”})}{\theta^w}))$$

$$\chi_{ih\tau} = \begin{cases} \theta^o log(1 + exp(\frac{\phi_{ih\tau+1}(d_{ih\tau+1} = \text{“wait offline”})}{\theta^o})) & : \tau < T^r_i \\ 0 & : \tau \geq T^r_i \end{cases} . \tag{1.9}$$

Note T_i^r is the length of retrial window for customer i.

The choice probabilities at each time period of each online and offline stage are

$$
\begin{aligned}
P_{ikt}^w(a_{ikt}; c_i^w(Z^w), \theta^w, r_l, \forall l) &= \frac{exp(\psi_{ikt}(a_{ikt})/\theta^w)}{1 + exp(\psi_{ikt}(a_{ikt} = \text{“wait online”})/\theta^w)}, \\
P_{ih\tau}^o(d_{ih\tau}; c_i^o(Z^o), \theta^o, \nu) &= \frac{exp(\phi_{ih\tau}(d_{ih\tau})/\theta^o)}{1 + exp(\phi_{ih\tau}(d_{ih\tau} = \text{“wait offline”})/\theta^o)}.
\end{aligned} \tag{1.10}
$$

The total likelihood is computed by tracking the sequence of decisions for all callers over a given parameter set. For caller i, his episode may contain multiple online stages and offline stages as described in Section 1.4,. We use K_i to denote his total number of online waiting stages and H_i to denote his total number of offline waiting stages during his episode. Moreover, each online waiting stage and each offline waiting stage are divided into small time periods. For caller i, the total number of periods for the kth online waiting stage is denoted as N_{ki}^w and the total number of periods for the hth offline waiting stage is denoted as N_{hi}^o. The total log likelihood of observing the data, denoted by L, is equal to the sum of the log likelihood of observing the sequence of caller i's actions a_{ikt} during his kth online waiting stage for $k = 1, ..., K_i$ and $d_{ih\tau}$ during his hth offline waiting stage for $h = 1, ..., H_i$, over $i = 1, 2, ..., N$ where N is the total number of callers. Thus, the total log likelihood is given by

$$
L = \sum_{i=1}^{N} ln\mathbb{E}[\prod_{k=1}^{K_i}\prod_{t=0}^{N_{ki}^w} P_{ikt}^w(a_{ikt}; c_i^w(Z^w), \theta^w, r_l, \forall l) \prod_{h=1}^{H_i}\prod_{\tau=0}^{N_{hi}^o} P_{ih\tau}^o(d_{ih\tau}; c_i^o(Z^o), \theta^o, \nu)]
$$

where $c_i^w(Z^w) = |\mu^w + \sigma^w z_i^w|$ and $c_i^o(Z^o) = |\mu^o + \sigma^o z_i^o|$ with z_i^o and z_i^w being i.i.d $N(0,1)$ distributed. To estimate the parameters $\omega = (\mu^o, \sigma^o, \theta^o, \mu^w, \sigma^w, \theta^w, r_l, \nu) \in \Omega = \{(\mu^o, \sigma^o, \theta^o, \mu^w, \sigma^w, \theta^w, r_l, \nu) : \mu^o \in \mathbb{R}, \sigma^o \in \mathbb{R}^+, \theta^o \in \mathbb{R}^+, \mu^w \in \mathbb{R}, \sigma^w \in \mathbb{R}^+, \theta^w \in \mathbb{R}^+, r_l \in$

$\mathbb{R}^+, \nu \in \mathbb{R}^+\}$ where Ω is the feasible set of the parameters, is equivalent to solving the optimization problem as follows:

$$\begin{aligned}
\max_{\omega\in\Omega} L(\omega) &= \max_{\omega\in\Omega} \sum_{i=1}^{N} L_i(\omega) \\
L_i(\omega) &= \ln \int\int \prod_{k=1}^{K_i} \prod_{t=0}^{N_{ki}^w} P_{ikt}^w(a_{ikt}; c_i^w(Z^w), \theta^w, r_l, \forall l) \\
&\quad \prod_{h=1}^{H_i} \prod_{\tau=0}^{N_{hi}^o} P_{ih\tau}^o(d_{ih\tau}; c_i^o(Z^o), \theta^o, \nu)\phi(Z^w)\phi(Z^o)dZ^w dZ^o
\end{aligned} \tag{1.11}$$

where $\phi(.)$ is the probability density function of the standard normal distribution.

Identification strategy addresses how to identify the parameter set ω in determining customers behavior based on our dataset. We argue that the identification arguments used in Aksin et al. (2013) can be applied to our structural model. First, it is important to point out that, combining the nominal utility, the aggregated future utility and the terminal condition, as described in (1.5), (1.7), (1.10) and (1.8), one can show that the customers' choice probability given by (1.10) can also be expressed by a well defined function of the parameters $\left(\left\{\frac{-c_i^w+\sum_l(p_{lt}r_lT_{ik}))}{\theta^w}\right\}_{t=1,\ldots,max(N_{ki}^w)}, \left\{\frac{-c_i^o-\nu\mathbb{I}(\tau\in\{1 \text{ day, } 2 \text{ day, } \ldots\})}{\theta^o}\right\}_{\tau=1,\ldots,T_i^r}\right)$. Specifically, we have

$$\begin{aligned}
&P_{ikt}^w(a_{ikt}; c_i^w(Z^w), \theta^w, r_l, \forall l) = \\
&\zeta\left(\left\{\frac{-c_i^w + \sum_l(p_{lt}r_lT_i)}{\theta^w}\right\}_{\forall t}, \left\{\frac{-c_i^o - \nu\mathbb{I}(\tau \in \{1 \text{ day, } 2 \text{ day, } \ldots\})}{\theta^o}\right\}_{\forall \tau}\right) \\
&P_{ih\tau}^o(d_{ih\tau}; c_i^o(Z^o), \theta^o, \nu) = \\
&\xi\left(\left\{\frac{-c_i^w + \sum_l(p_{lt}r_lT_i))}{\theta^w}\right\}_{\forall t}, \left\{\frac{-c_i^o - \nu\mathbb{I}(\tau \in \{1 \text{ day, } 2 \text{ day, } \ldots\})}{\theta^o}\right\}_{\forall \tau}\right)
\end{aligned} \tag{1.12}$$

Shape Parameter	Target Agent	Branch Backup	Call Center
Private Customer	1.02	0.41	0.19
Business Customer	0.88	0.55	0.28

Table 1.3. Shape Parameter of the Fitted Gamma Distribution of the Waiting Time for Different Customer segments and Service Groups

where ζ and ξ are the corresponding well defined functions for the online choices and the offline choices.

For the online stage, as shown by Aksin et al. (2013), one can identify $\frac{c_i^w}{\theta^w}$ and $\frac{r_l}{\theta^w}$ with $l \in \{$target agent, branch backup, call center$\}$ separately iff the service probability p_{lt} varies across different time periods for each l. One can see that the service probability p_{lt}, given in (1.4), is a discrete approximation of the hazard rate of the waiting time associated with group l. Thus, to verify that our data shows significant intertemporal variation in service proabability, we turn to the distribution of the waiting time associated with different service groups. Specifically, Table 1.3 reports the shape parameter of the fitted Gamma distribution of the waiting time for different customer segments and service groups. We observe that none of the shape parameters of the Gamma distribution equals to 1, which implies that the hazard rate of the waiting time for each service group and each customer segment varies over time. Thus, following the arguments in Aksin et al. (2013), we claim that one can identify $\frac{c_i^w}{\theta^w}$ and $\frac{r_l}{\theta^w}$ with $l \in \{$target agent, branch backup, call center$\}$ seperately.

For the offline stage, one can identify $\frac{c_i^o}{\theta^o}$ and $\frac{\nu}{\theta^o}$ iff the offline waiting time exceeds more than one day for some callers. When the offline waiting time is beyond 1 day, the waiting cost is $c_i^o + \nu$. When the offline waiting time is less than 1 day, the waiting cost is c_i^o. In Figure 1.4, the empirical density plot of the offline waiting time verifies that

some callers' offline waiting time exceeds 1 day and some callers' offline waiting time is less than one day. Hence we can identify $\frac{c_i^o}{\theta^o}$ and $\frac{\nu}{\theta^o}$ separately following the arguments in Aksin et al. (2013).

1.6. Results

In previous sections, we explained the structural model and the estimation strategy. In this section, the estimation results are presented. We first report the results for the structural model and then use cross-validation to show that our structural model has the ability to predict retrial behavior.

1.6.1. Estimation Results

In the structural model, the customers make decisions regarding abandonment and retrial based on their evaluation of online waiting cost, offline waiting cost and the expected service rewards. In this subsection, we first discuss the estimation results for the online stage: the online waiting cost c_i^w, the service reward associated with the three service groups $\{r_l : l \in \{\text{call center, branch backup, target agent}\}\}$ and the dispersion of the online idiosyncratic shock θ^w. Then we will do the same for the offline stage estimators: c_i^o, ν, and θ^o.

The estimation results for the online stage across the two customer segments are summarized in Table 1.4. In the online stage, customers decide whether to abandon or keep waiting for the service based on their unit online waiting cost and the expected service reward at each time period t. By normalizing the reward rate of a call center agent

Customer	$E(c_i^w) = \mu^w$	$SD(c_i^w) = \sigma^w$	$r_{\{\text{branch backup}\}}$	$r_{\{\text{target agents}\}}$	θ^w
Private	0.07 (0.11)	0.16 (0.11)	1.12 (1.19)	24.49 (4.00)	29.96 (1.74)
Business	0.28 (0.79)	0.13 (0.07)	2.45 (2.41)	23.11 (6.48)	17.87 (3.36)

Table 1.4. Estimates (and Std) of Online-Stage Parameters

Customer	$E(c_i^o) = \mu^o$	$SD(c_i^o) = \sigma^o$	θ^o
Private	0.60 (0.11)	0.61 (0.12)	0.79 (0.10)
Business	0.45 (0.14)	0.45 (0.16)	0.66 (0.15)

Table 1.5. Estimates (and Std) of Offline-Stage Parameters

$r_{\{\text{call center}\}}$ to be 1[4], the other estimates are then ratios compared to this base metric. Note that the random online cost $c_i^w = |\mu^w + \sigma^w z_i|$ is characterized by the location parameter μ^w and the dispersion parameter σ^w.

The estimates suggest that:

(1) A quick access to services matters, especially to the business customers. From the cost-reward ratio estimates, we see that the online waiting costs for both customer segments are distinct above 0 and the business customers face higher costs from online waiting. A private customer is willing to wait almost 3 times longer than a business customer for equivalent service from a call center agent. (A business customer is willing to wait 3.57(= 1/0.28) minutes for 1-minute talk with a call center agent while a private customer is willing to wait 14.29(= 1/0.07) minutes.)

(2) The service quality from target agents is the best and the private customers are more sensitive to the service providers compared with the business customers. From the service rewards ratio estimates, we see that for the private customers,

[4]Because we can only identify the ratio $\frac{c_i^w}{\theta^w}$ and $\frac{r_l}{\theta^w}$ with $l \in \{\text{target agent, branch backup, call center}\}$, we need to normalize one parameter to obtain a unique set of estimates.

the service reward from the target agent is 24.49(= 24.49/1) times that of the call center and 21.87(= 24.49/1.12) times that of the branch backup. However, for business customers, the reward rates across service groups do not vary significantly: the service reward from the target agent is 23.11(= 23.11/1) times that of the call center and 9.43(= 23.11/2.45) times that of the branch backup.

(3) The private customers exhibit more diversified behavior during online waiting stage than the business customers. We see that both the variance of online waiting cost and the dispersion of the online idiosyncratic shock are larger for private customers.

Our key finding for the online stage is that customers value both service quality and quick access to services, but their preferences differ across customer segments. Ultimately, business customers, when compared to private customers, place more value on speed and care less about quality.

The estimation results for the offline stage across two customer segments are summarized in Table 1.5. At this stage, customers decide whether they should retry or stay offline by evaluating their offline waiting cost and the potential gains if retry. The offline waiting cost is composed of a regular unit offline waiting cost c_o and a daily-occurred waiting cost ν. By normalizing the daily-occurred waiting cost ν to be 1[5], the other estimates are then ratios compared to this base metric. Note that the random offline cost $c_i^o = |\mu^o + \sigma^o z_i|$ is characterized by the location parameter μ^o and the dispersion parameter σ^o.

The estimates suggest that:

[5]Because we can only identify the ratio $\frac{c_i^o}{\theta^o}$ and $\frac{\nu}{\theta^o}$, we need to normalize one parameter to obtain a unique set of estimates.

(1) Business customers have lower regular offline waiting costs. For business customers, they incur less loss if not consulting the call center again to solve their issue. This might be explained by that the business customers have other professional alternatives to solve the issues instead of contacting the call center. However, for private customers, there are much fewer side options.

(2) The time-dependent cost ν is more prominent for business customers compared to the private customers. The unit offline waiting cost at the daily cycle for business customers is $3.22(= (0.45 + 1)/0.45)$ times more than the regular offline waiting cost while, for private customers, the incremental is just $2.67(= (0.60 + 1)/0.60)$ times. This is consistent with our knowledge that business customers have a more regular working routines to follow which impacts the time they contact the call center.

(3) The private customers exhibit more variation during offline stages than the business customers. We see that both the variance of regular offline waiting cost and the dispersion of the offline idiosyncratic shock are larger for private customers.

Utilizing the obtained estimates in the structural model, we can estimate the probability of abandonment and retrial.

The estimated probabilities of abandonment for the business customers and private customers are plotted with the corresponding 95% confidence interval in Figure 1.5. Compared with the observed probabilities of abandonment, our model accurately captures this behavior in the call center.

In Figure 1.6, the estimated probabilities of retrial segmented by the retrial types are illustrated. Recall we define two types of retrial in Section 1.3: the congestion retrial

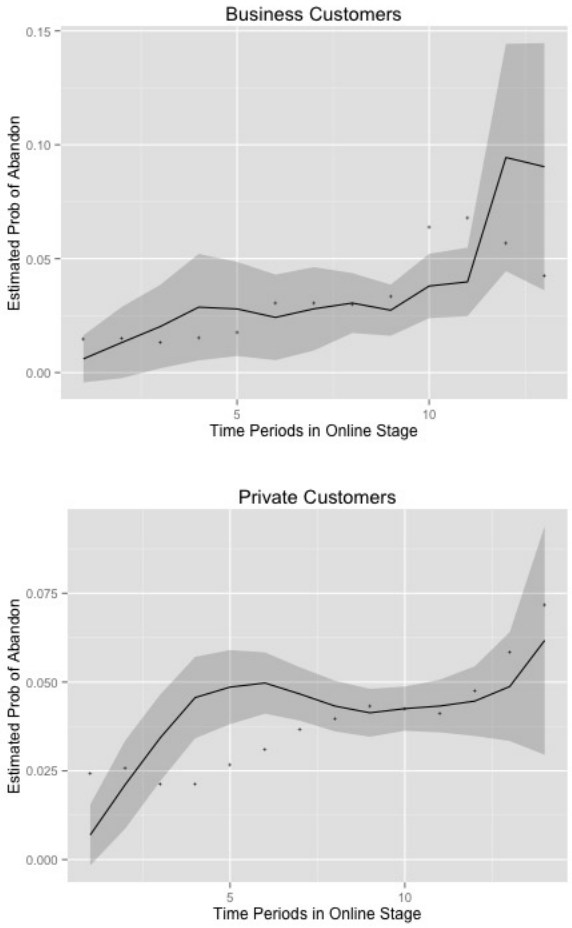


Figure 1.5. 10-fold Cross-Validation: Estimated Prob of Abandonment (line) v.s. Observed Prob of Retrial (dots)

which customers retry after abandoning the previous call; and the fitness retrial which customers retry after an unsatisfactory service in the previous call. The probability of

retrial for each type is calculated by:

(1.13)

P(congestion retrial at offline period τ)

$= \sum_t P$(abandonment in previous call at online period t, retrial at offline period τ)

P(fitness retrial after a service from group l at offline period τ)

$= \sum_t P$(receive a service from group l in previous call at online period t,

retrial at offline period τ)

where $l \in$ {call center, branch backup, target agent}

In Figure 1.6, we see that in general retrial is more likely to happen in the early periods of the offline stage. Moreover, the daily-occurred offline cost pushes people to retry before the daily cycle ends. Compared across the retrial types, congestion retrial tends to happen early in the offline stage while the fitness retrial happens across the entire stage. This aligns with our conjecture that customers who didn't receive a service in the previous call tend to call back sooner to obtain services.

1.6.2. Cross-Validation

In this subsection, we use cross validation to assess our model's performance in predicting abandonment and retrial behavior. To use the ten-fold cross-validation, we first randomly partition the original sample into ten equal-sized subsamples. Of the ten subsamples, a single subsample is retained as the validation data for testing the structural model, and the remaining subsamples are used as training data to estimate the parameters of retrial

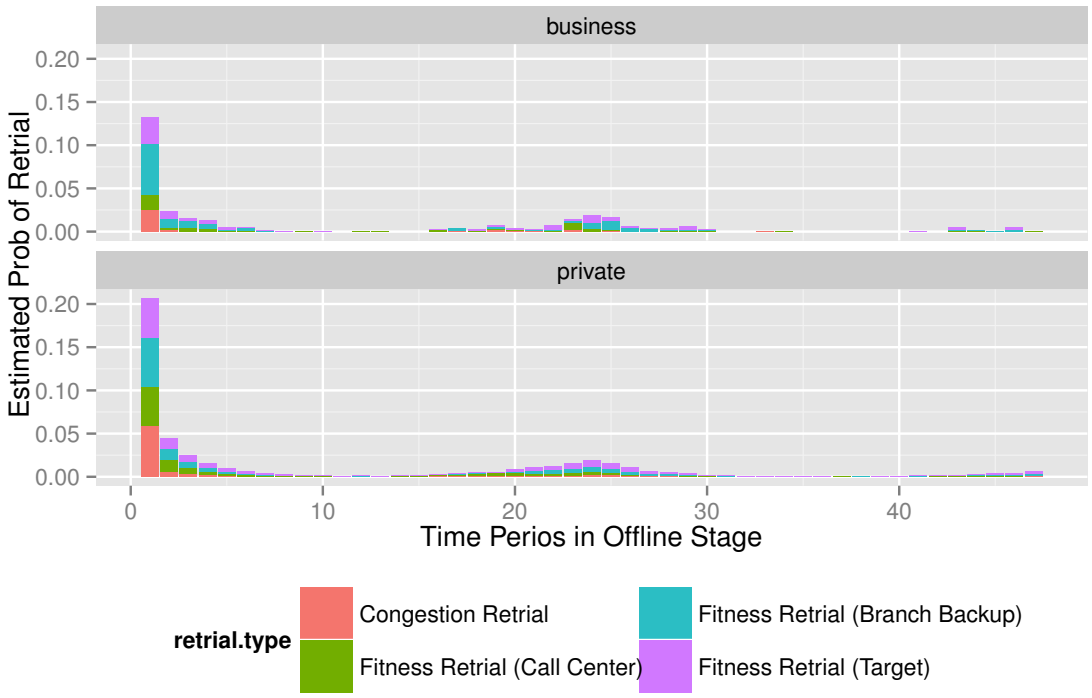


Figure 1.6. 10-fold Cross-Validation Estimated Prob of Retrial

behavior. The cross-validation process is then repeated ten times, with each of the ten subsamples used exactly once as the validation data. For each validation process, we calculate the MSE (mean-squared errors) based on the predicted probabilities of abandonment and retrial along with the observed ones in the testing subsample. The average of the ten MSEs calculated from the ten-fold cross-validation measures the predictability of our structural model. The resulted average MSE from the ten-fold cross validation is merely 3.26×10^{-4} for the business customers dataset and 1.01×10^{-4} for private customers dataset. This suggests our model predicts fairly well for the customers behaviors from both segments.

1.7. Counterfactual Analysis

In this section, we conduct a counterfactual analysis to demonstrate how our methodology can help call centers design better service systems. In previous sections, the structural model and estimation results convey the idea that both service speed and service quality are desired for customers. Moreover, for different customer segments, their preferences in terms of speed and quality are different. In order to improve the services, we want to test two strategies, one without expansion in the service teams and the other one expands the service teams with the cheap labor resources, the call center agents. The first strategy examines the length of preferred time lag for each customer segment instead of the generic one-minute time lag for all customers in the original system. Secondly, we examine whether the service provider can improve services by adding more call center agents. We use customer surplus, the expectation of service reward minus the total waiting cost from online and offline stages, to evaluate whether a change should be implemented into the current system.

Before diving into the counterfactual analysis, let us briefly describe the essence of our call center model: the equilibrium between the perceived service probabilities and the experienced service probabilities. Given the customers' historical interaction with the call center, they form beliefs about the service probabilities from the three service groups. Based on these beliefs, customers make abandonment and retrial decisions, which in turn impact the service probabilities experienced by other customers (since the service probabilities are jointly determined by the staffing size of service groups and the number of customers online). For a given service system, we are interested in characterizing the equilibrium where the customers' beliefs about the service probabilities (namely the

perceived service probabilities) equal to the actual ones experienced by the customers (namely the experienced service probabilities).

To compute this equilibrium, we use the following iterative procedure: for any service system, we start with an initial guess of customers' perceived service probabilities from the three service groups. We then simulate the service system and the decisions of customers under such anticipation. Based on the simulation results, we compute the actual service probabilities experienced by the customers. If the experienced service probabilities are different from the perceived ones, we will use the experienced service probabilities as the new anticipation of customers and rerun the simulation. We stop once this simulation converges in the sense that the difference between the experienced service probabilities and the perceived ones is less than 1% for each period. Once we reach the equilibrium state, we measure the merits of the service system by using the average of customers' surplus. For one customer, his surplus is equal to the total service rewards minus the total waiting costs. [6]

Using the iterative procedure described above, we simulate the current service system and determine the staffing size that leads to the closest service probabilities as the observed ones. The final workforce we determined are 2 target agents, 1 branch backup agent and 4 call center agents.

First we examine how the time lag between adding new service groups should be altered to align with customers' preferences. The current system's generic one-minute time lag for all customers sets the benchmark. We will determine the preferred time lag based on the improvement in customer surplus. Recall the time lag determines the delay when

[6]Remark: Given the estimates of reward and waiting cost is obtained when normalizing the reward from call center agents to be 1, the consumer surplus is also measured based on this normalized value.

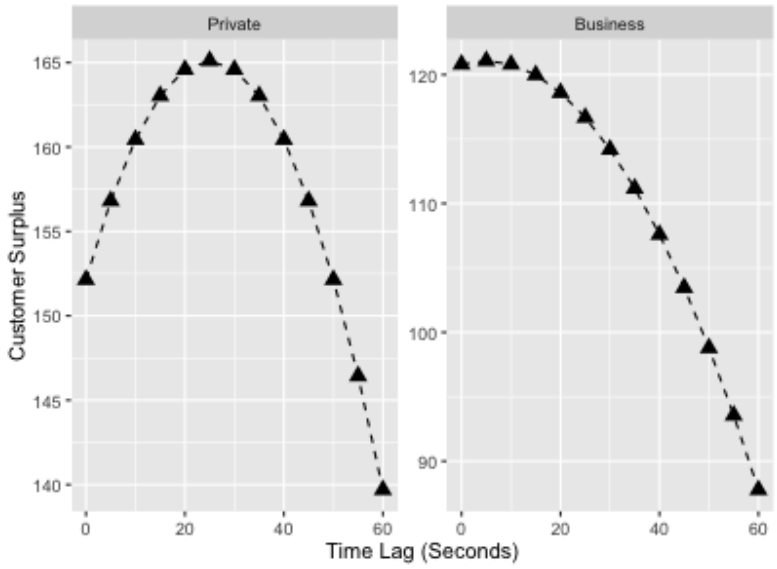


Figure 1.7. Optimal Lag between Adding Service Groups

sequentially introducing the three service groups with diminishing service quality. If the gap is very large, customers have higher chances to get served by high-quality agents but endure longer waiting time. Conversely, when the gap is close to zero, customers face the least likelihood to get good quality but also experience the shortest waiting time. Hence, there should exist a sweet spot of such gap so that customers' preferences in quality and speed are aligned with the service offering. We anticipate different customers segments should prefer different time lags considering their distinct preferences in timely services and good quality. In our counterfactual analysis, we test the time lags from 0 seconds to 1 minute with a five-second increment. Shown in Figure 1.7, we find that the system should use a 5-second time lag for business customers and a 25-second time lag for private customers. Updated the hybrid service model with the preferred time lags, we find that the business customers obtain 37.9% more surplus and private customers obtain 18.2% more surplus.

From this counterfactual analysis, we highlight the different preferences between speed and quality for each segment and also quantify the preferred time lag for each customers segment. In general, the one-minute time lag is too long for both segments. The current system needs to shorten the time lag in service delivery to prevent customers from waiting too long in the online stage. Moreover, we see the business customers prefer a much shorter time lag compared with private customers. The business customers prefer almost no delay in access to the general call center agents which greatly shortens their waiting time online. The crucial improvement of the business customer surplus is from the shortened online waiting time. Private customers, however, would prefer certain amount of delay to receive better quality of services. With the new shorter time lag, they are able to reduce their waiting time online without loosing too much in good quality. This result once again highlights our crucial message that business customers have a stronger preference for timely services but care less about the service quality compared with the private customers.

Secondly we examine whether an additional call center agent is always beneficial to the customers. The reason to have both target agents and call center agents is because call center agents are cheaper to hire but provide ordinary-quality services while target agents are expensive to hire but deliver good quality. In order to improve the service system cost-effectively, we want to see whether we can improve customer surplus by hiring more call center agents. In the baseline system, we have two target agents, one branch backup agent and four call center agents. In Figure 1.8, we examine how customer surplus changes when expanding the general call center service group from 4 agents to 30 agents. In the original system, private customers' surplus is 139.7 and business customers' surplus is

87.8. Initially, we see the surplus for both segments increases. The peak of the surplus is 143.7 (a 2.35% increase) for private customers and 110.9 (a 26.3% increase) for business customers. The increase in surplus is mainly due to the shortened waiting time and less congestion retrial resulting from the abandoned calls. The percentage of increase is more significant for business customers because they care more about online waiting compared to the private customers. However, the surplus starts to decrease after adding a certain amount of call center agents. For business customers, the surplus starts to reduce after adding the 5th call center agents. For private customers, the surplus starts to reduce after adding the 3rd call center agents. At the turning points in surplus, customers suffer too much loss from losing good-quality and such loss outweighs their gain in shortened online waiting time. The results suggest that adding more general call center agents is not always beneficial. Since business customers care more about online waiting and are less sensitive to service quality, they prefer to add more call center agents compared to the private customers.

In this counterfactual, we learn that for service providers, whether to add one more call center agent depends on the service regime. If the marginal loss in service quality outweighs the marginal benefit in shortening online waiting, the service provider should not add a call center agent. The maximum increase in customer surplus and also the maximum number of call center agents that benefits customers' surplus are both related with customers' preferences. This once again highlights that service providers need to first understand customers' preferences in service speed and quality, and then align the service offering with customers' preferences to provide the highest customers surplus.

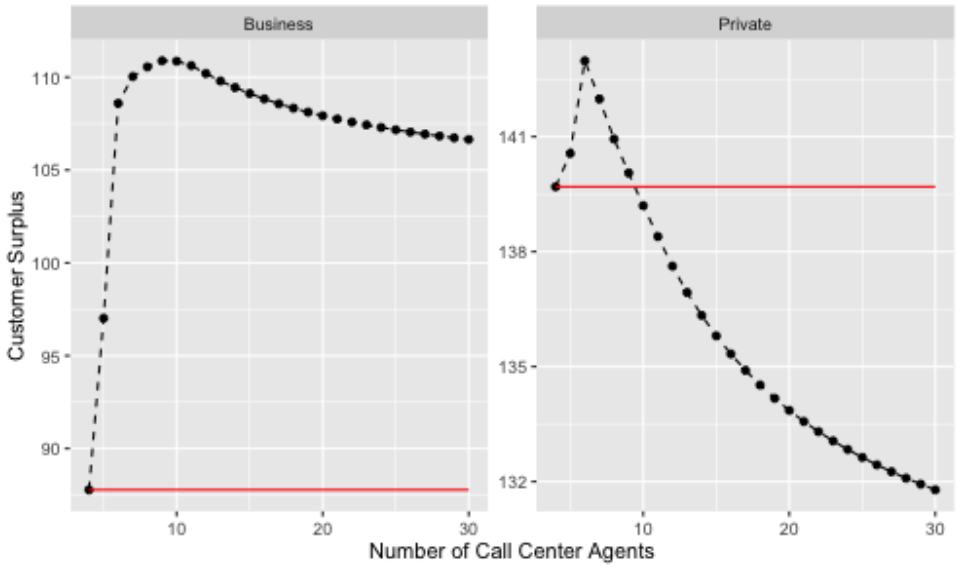


Figure 1.8. Customers are not always happier with more general call center agents.

In this section, we conduct two types of counterfactual analysis to examine how we can improve the current service system. The first analysis suggest customers surplus will be improved significantly when the time lag between adding new service groups is optimized for each customer segment. Business customers' surplus is increased by 37.9% with a 5-seconds time lag and private customers' surplus is increased by 18.2% with a 25-seconds time lag. By choosing different time lags, it offers customers a service system balancing between good quality and timely services to align with their preferences. The second analysis suggests the service providers can hire the cheap labor resource, the general call center agent, to improve customers surplus when the marginal gain from shortening online waiting outweighs the marginal loss in service quality. For service providers, these two analysis offer two strategies to improve services: 1. without expanding the service team, they can improve customers' surplus by wisely allocating the current service groups along

the timeline; 2. By expanding the service team with more call center agents, they can improve customers' surplus up to a certain level. However, they should be aware that the surplus will go down when there are too many call center agents on duties. The number of call center agents to add is related with customers' preferences.

1.8. Conclusion

This empirical study focuses on customers' retrial behavior and studies how service providers can improve services via a well designed hybrid system balancing between service speed and service quality. The two important features in services are service quality, i.e. how effectively the services resolve customers' requests, and service speed, i.e. how long customers need to wait before reaching services. The empirical results suggest that customers across segments weights differently on service speed and service quality. It proposes an important notion for service management that: service providers need to first understand customers' preferences in service speed and quality, and then to align their service offering with customers' preferences.

We classify retrial behavior into two types: the congestion retrial resulted from abandonment after long waiting and the fitness retrial resulted from unresolved requests after low-quality services. The ideal service system offers both the best quality and the quickest delivery. However, this is not the most cost-effective way to improve customers welfare. In order to design an economic viable service system, the first step is to understand customers' preferences in speed and quality.

We use a call-by-call dataset from a hybrid call center that enables us to disentangle customers' preferences in speed and quality. In this call center, there are three groups of

service providers that join the customers' service pool sequentially: the target agent group first, then the branch backup group and lastly the call center group. In this paper, we study congestion retrial and fitness retrial by connecting customers' behavior with their preferences for service aspects and further to provide suggestions to improve services.

We first use Probit regressions to connect the probability of retrial with the outcomes of the previous call. The results suggest retrial is impacted by three factors: whether a customer received services or not in his previous call, the service group that provided services and the service time.

We then use a random-coefficient dynamic structural model to study the fundamental mechanism between retrial and service features suggested by the Probit regressions. We model customers' behavior in the two stages of the service system: the online stage where customers wait for services and decide whether to abandon or not; and the offline stage where customers decides whether to retry and go back to the online stage or not. The online waiting cost suggests customers prefer speedy delivery and services from taget agents as they provide the best quality. Moreover, our estimation of customers' preferences suggests that business customers value more speedy delivery and are less sensitive to service quality compared to private customers.

Lastly, we suggest two approaches to improve services by tailoring the services to meet customers' distinct preferences. In our first approach, we suggest improving customers' surplus by efficiently allocating the service teams along the timeline based on customers' preferences. Instead of the generic one-minute time lag between adding new service groups in the original system, we suggest a 5-second time lag for business customers and a 25-second time lag for private customers. The time lag plays a role in trading-off between

timely responses and good quality. Without expanding the service teams, we improve business customers' surplus by 37.9% and private customers' surplus by 18.2%. In our second approach, we suggest to improve services by hiring more cheap resources, the call center agents. The service provider should be aware that the call center agents reduces' customers waiting cost but also lowers the chances to get good quality. After adding a certain number of call center agents, the surplus will start decreasing when the marginal loss in service quality outweighs the marginal gain from shortening online waiting. In the current service teams, business customers' surplus will be increased by 26.3% with 5 more call center agents and the private customers' surplus will be increased by 2.35% with 3 more call center agents.

This paper contributes to the existing service management literature in the following manners. First, our study promotes an understanding of how retrial behavior is impacted by the service offering, in particular, speed and quality. We categorize retrial into two types: the congestion retrial resulted from untimely services and the fitness retrial resulted from poor quality. Secondly, our empirical model is innovative in capturing customers behavior in both the online stage and the offline stage. We acknowledge that customers also make retrial decisions in the offline stage that are heavily impacted by the service aspects. Thirdly, we demonstrate the importance of accounting customers' preferences when making operation decisions in a multi-skill setting. Our counterfactual analysis suggests two economic viable options to improve a hybrid service system. Lastly, our study develops a methodology framework to analyze customers' preferences in speed and quality and the mechanism between customer behavior and offered services. This methodology

framework can be applied to a wide range of digital and virtual service industrial practices beyond call center management.

Future extensions of our research are worth exploring. In our structural model, we consider service time is exogenous, which is not determined by the customers. One extension could model customers' decisions in the online service stage, particularly whether customers choose to hang up or keep talking with an agent. By modeling customers' decisions in the service stage, the service time is endogenous and determined by the customers' experience of the on-going service.

CHAPTER 2

Macro-environmental Forces that Drive Carmakers to Misconduct: Intense Competition and Stringent Standards (joint with Sunil Chopra, Yuche Chen)

2.1. Introduction

In 2008, Volkswagen's announced its new "clean" diesel cars and claimed to offer superior fuel efficiency while complying with the strictest vehicle emission standards at that time. Soon Volkswagen swept the US diesel car market with strong sales, environmental awards and tax breaks for its "innovative" products. This success, however, was built on a very shaky foundation. On Sept. 3rd 2015, Volkswagen officially admitted that cheating devices had been used on their vehicles since 2009 to manipulate the Nitrogen Oxides (NOx) emission. The NOx output for the modified vehicles' could meet standards in lab tests and standard testing environments but emitted up to 40 times more than the limits when the vehicles were driven on-road. The shocking news didn't stop there. Shortly after the Volkswagen scandal, the U.S. EPA found the same cheating software on Audi diesel models and a Porsche model[1]. In EU, an official investigation revealed that none of the 37 top-selling vehicle models which claimed to satisfy Euro 5 actually met

[1]http://www.nytimes.com/interactive/2015/business/international/vw-diesel-emissions-scandal-explained.html

the standards when driven on-road[2]. In this paper we refer to the automakers' failure to meet standards on-road as misconduct because the firms were aware that there was no testing regimen in place for on-road performance and sold the cars even though they did not meet standards when driven on-road. In our classification of misconduct, we do not distinguish between the case where carmakers made an effort to meet standards but were unable to do so and the case where they set out to cheat. In our paper, any failure to meet standards is classified as misconduct.

Responding to the discovery of misconduct in on-road NOx emission, EU decided to strengthen its monitoring by adopting the Real Driving Emissions test in place of the decades-old lab test. Simultaneously, EU introduced a conformity factor, which relaxes the emission standard for the carmakers. From September 2017, vehicle models are allowed to emit up to twice as much NOx as the current limits for on-road testing. After 2019, they will still be permitted a 50% overshoot.[3]

It has been known for a long time that NOx has both direct and indirect harmful effects on human health. Directly it can cause respiratory disorders, headaches and even cardiovascular diseases. Indirectly, it affects humans by damaging the ecosystem through acid rain and smog. It is estimated that in Britain alone, 23,500 people are killed by NOx emission every year (Colbeck and Lazaridis (2010)). While all fossil-fuel burning processes result in NOx emission, road transport is the largest contributor to NOx emission, accounting for 39% of NOx emissions in 2008 and 46% in 2013 in EU.[4] With a goal

[2]https://www.theguardian.com/business/2016/apr/21/all-top-selling-cars-break-emissions-limits-in-real-world-tests
[3]https://www.theguardian.com/environment/2016/feb/03/eu-parliament-gives-green-light-for-loopholes-in-car-emissions-tests
[4]http://ec.europa.eu/environment/air/transport/road.htm

Standard	Euro 3	Euro 4	Euro 5	Euro 6
Imposed Year	2000	2005	2009	2014
NOx Limits (g/km)	0.50	0.25	0.18	0.08

Table 2.1. EU Emission Standards for Diesel Passenger Vehicles

of reducing NOx emissions, the EU Commission has tightened emission standards four times from 2000 to 2014 (Table 2.1).

The outcome of the tightening standards, however, was not quite as anticipated with regards to on-road emissions. While on-road emissions of NOx initially decreased, they essentially flattened out after 2006 (see Figure 2.1). As standards tightened, the fraction of automakers failing to meet on-road NOx standards increased. Tightening NOx standards seem to have been accompanied by greater misconduct on the part of automakers. The International Council on Clean Transportation (Franco et al. (2014)), one of the first to discover Volkswagen's misconduct, tested 15 vehicles from 6 carmakers and discovered that few vehicle models sold in EU actually met the standard limits for on-road driving even though all of them passed the lab emission test.

The extent of misconduct observed in our data[5] as standards tightened is shown in Figure 2.1. Figure 2.1 shows the boxplots of on-road NOx emission from 2000 to 2012 recorded in our dataset. The solid line shows the NOx standard limits in place for each year. Figure 2.1 shows that even though standards were tightened after 2006, the actual on-road emissions essentially flattened out and the fraction of automakers failing to meet on-road standards increased.

[5]Our data comes from Chen and Borken-Kleefeld (2014) who obtained an extensive dataset covering on-road vehicle emissions in the EU auto market between 2000 and 2012.

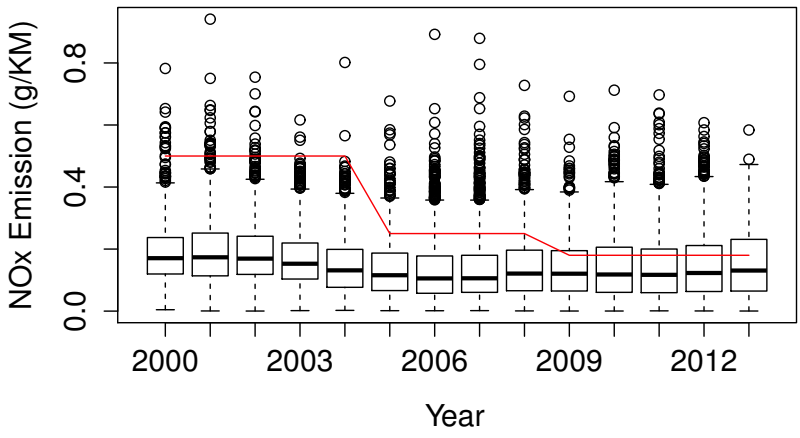


Figure 2.1. Boxplots of on-road NOx emission compared to the EU Standards Limits (Solid Line)

The increasing misconduct observed in the data indicates the importance of understanding factors that influence misconduct. It is the goal of our research to use simple theoretical models and empirical analysis of on-road emission data in the EU market, to identify factors that drive misconduct in the auto market. Our goal is to better inform regulators as they set future emission standards.

In order to understand misconduct, it is important to consider the automakers (and customers) perspective. Contrary to regulators, carmakers are likely to view a reduction in NOx emission as hurting their product performance. When questioned about their NOx emission misconduct, Volkswagen Chairman Hans-Dieter Potesch said that they "couldn't find a technical solution within the company's time frame and budget to build diesel engines that would meet U.S. emissions standards." Upon learning the stringent NOx emission standard released in 2004 for diesel cars in the U.S., Mazda, Honda, Nissan and

Hyundai decided not to introduce diesel vehicles and said, "the main challenge was that it was too difficult to meet the new standards while maintaining the engine performance and staying on budget."[6] From the carmakers' perspective it is clear that reducing NOx emission requires either a sacrifice in engine performance or higher investment. This perspective is validated in the scientific literature which discusses two approaches that carmakers can use to reduce NOx emissions. One is to install an expensive catalyst (Gandhi et al. (2003)), which increases the retail price of vehicles. The other is to sacrifice engine performance by downsizing the engine or reducing fuel efficiency (Yeh (2007)). Customers tend to care more about good engine performance and low vehicle prices than the reduction in NOx emission (Deloitte (2014)). Moreover, NOx emission (especially on-road) is harder for customers to detect compared to the selling price and fuel consumption. Hence, even for customers who favor low NOx emission, purchase decisions are unlikely to be based on the vehicles' on-road NOx emission. This fact coupled with the absence of on-road testing may increase misconduct as standards tighten if carmakers are willing to risk high NOx emission in a bid to survive competition in the auto market by attracting customers who value low price and good engine performance.

Our theoretical and empirical analysis suggests that both market competition and regulation influence misconduct. Our theoretical models suggest that when competition strengthens and standards tighten, carmakers are more likely to commit misconduct and emit in excess of standards. Our empirical analysis using on-road data between 2000 and 2012 confirms this implication and quantifies the impact of both competition and regulation on misconduct. Our results suggest that a 1% increase in market level competition

[6]http://www.newsweek.com/2015/12/25/why-volkswagen-cheated-404891.html

increases the probability of misconduct by 0.58%; a 1% tightening of standard limits increases the probability of misconduct by 1.72%; and each additional model substitute available in the market increases the probability of misconduct by 0.48%. We also find a link between the probability of misconduct and vehicle features that are important to customers. We find that the probability of misconduct is lower by 8.57×10^{-6}% for a 1% increase in price (auto makers meet standards with a higher probability for high price cars); the probability of misconduct is lower by 1.73×10^{-3}% for a 1% decrease in vehicle weight (auto makers meet standards with a higher probability for lighter cars); and the probability of misconduct is lower by 4.26×10^{-2}% for a 1% decrease in engine power (auto makers meet standards with a higher probability for cars with lower power).

Our research also sheds light on EU's decision to loosen NOx limits in the short term while introducing sophisticated on-road testing in the longer term. We use both theoretical comparative statics and empirical counterfactual analysis to study the issue. Our theoretical model suggests that in the absence of monitoring effectiveness and beyond a certain threshold of competition intensity, the emission standards should be set looser as the competition intensity increases. Our empirical counterfactual analysis suggests that doubling the NOx standard limits (which is what EU did) decreases the probability of misconduct by 9.56% to 11.04%, depending on the model-level competition in the market. This finding supports EU's decision to loosen the NOx limits. Moreover, our theoretical model suggests that the introduction of more sophisticated monitoring reduces the influence of competition on misconduct.

Our paper contributes to the literature in the following ways. First, we empirically identify the role of strict standards (along with competition) in driving misconduct.

Though previous modeling work (Branco and Villas-Boas (2015), Chen (2001)) has implied the link between strict standards and misconduct, to the best of our knowledge, there is no prior empirical study that confirms this relationship. Secondly, our research provides both theoretical and empirical guidance to regulators for setting up NOx emission standards. We suggest that regulators should decide the strictness of standards considering the competition intensity and the monitoring effectiveness. Once competition intensity exceeds a certain level, regulators should consider improving the effectiveness of the monitoring system each time standards are tightened to prevent misconduct on the part of automakers.

The remainder of the paper is organized as follows. A review of the related literature is provided in Section 2.2. In Section 2.3 we discuss a simple theoretical model that links competition and standards to the probability of misconduct. The results of this model provide the basic hypotheses for our empirical analysis. In Section 2.4 we describe the data and define variables used in our analysis. The empirical models are presented in Section 2.5 and results discussed in Section 2.6. In Section 2.7, we discuss how the regulators should choose the strictness of standards considering the competition intensity and their monitoring ability. With counterfactual analysis, we examine how misconduct is likely to change under the upcoming loosened EU on-road emission standards. In the end, we summarize the results in Section 2.8.

2.2. Literature Review

Prior research related to misconduct has identified competition as a key factor influencing the extent of misconduct. For example, several authors (Kulik et al. (2008),

Pierce and Snyder (2008), Cai and Liu (2009), Bagnoli and Watts (2010), Markarian et al. (2014), Du and Lai (2015)) have identified that competition drives the failure of firms to truthfully report their financial accounts. Others have identified the role of competition with regards to fraud in credit ratings and certification (Kinney et al. (2004), Becker and Milbourn (2011), Jiang et al. (2012), Short et al. (2013), Nistor and Tucker (2015)). Similarly Mayzlin et al. (2014) have identified the link between competition and the over-claiming of quality. Competition not only increases misconduct (Schwieren and Weichselbaumer (2010)) but also promotes the spread of misconduct across firms because even ethical firms are forced to mimic the questionable practices of their less ethical counterparts so that they won't be forced out of business (Reynolds (1940), Staw and Szwajkowski (1975), Cummins and Nyman (2005), Shleifer (2004), Kilduff et al. (2015)). In particular, competition drives misconduct when the interest of the market is not well aligned with the interest of social welfare (Hart (1983)). For example, Snyder (2010) finds that liver transplant centers with greater competitive pressures overstate the health problems for their patients to gain priority on the liver waiting list. Bennett et al. (2013) show that vehicle emission testing facilities facing fiercer competition passed a higher fraction of unqualified vehicles in order to please customers. Utgård et al. (2015) find that retail stores facing tougher competition tend to sell more alcohol to underage teens.

While there is no empirical work linking the strictness of standards to misconduct, there are some theoretical papers that have studied how regulation affects firms' effort to improve quality and the total social welfare under competition. Several papers find that the reward and punishment environment influences individual decisions and the extent of compliance or misconduct (Becker (1968), Hegarty and Sims (1978), Coleman

(1987)). Related to our work, the following papers find that stricter regulation does not necessarily improve social welfare (Melumad and Ziv (2004)) or result in greater effort to meet standards (Chen (2001), Branco and Villas-Boas (2015)). There are other papers, however that find situations where firms go beyond what regulations require (and sometimes benefit as a result). Dowell et al. (2000) find that multi-national firms that adopt a more stringent global environmental standard (compared to some host country standards) have higher market values, as measured by Tobin's q. Kraft et al. (2013) find that in the presence of competition, the uncertainty of upcoming regulation may lead to firms implementing ahead of potential regulation.

In the context of automakers and NOx emissions, however, we find that both increased competition and tighter standards lead to increased misconduct. Our research uses both theoretical models and empirical analysis to identify the strictness of standards as a key factor that along with competition affects the level of misconduct in the auto industry. As suggested in prior research that stricter standards are not always beneficial (Melumad and Ziv (2004), Chen (2001), Branco and Villas-Boas (2015)), our research quantifies the impact of stricter standards on misconduct in the automotive market.

2.3. A Simple Theoretical Model for Misconduct

In this section, we describe a simple principal-agent model to motivate how competition and regulation influence misconduct. The model builds on Branco et.al (2015) to capture how competing carmakers invest in costly efforts to reduce NOx emission under different levels of competition and strictness of standards.

We assume that there are N homogenous carmakers competing in the auto industry. Each carmaker i chooses to produce quantity q_i, resulting in a total supply of $Q = \sum_i q_i$. The price is determined by the the inverse demand function assumed to be $P(Q)$ with $P'(Q) < 0$ and $2P'(Q) + QP''(Q) < 0$.[7]

We assume that the strictness of standards for NOx emission set by regulators is denoted by $\lambda \in [0, 1]$. $\lambda = 0$ indicates the absence of any standards whereas $\lambda = 1$ indicates the strictest standards.

Each carmaker i selects an effort level $\gamma_i \in [0, 1]$ when designing and producing its products. The strictness of standards λ and the effort level γ_i determine the unit production cost $c(\gamma_i, \lambda)$ for carmaker i. We assume that the unit cost of production is increasing in both the effort and the strictness of standards, i.e., $\frac{dc}{d\gamma_i} > 0$ and $\frac{dc}{d\lambda} > 0$. We also assume that the unit cost of production is convex in both the effort and the strictness of standards, i.e., $\frac{d^2c}{d\gamma_i^2} > 0$, $\frac{d^2c}{d\lambda^2} > 0$ and $\frac{d^2c}{d\lambda d\gamma_i} > 0$.

We denote the effectiveness of the regulatory monitoring system by $m \geq 0$ with higher values of m indicating more effective monitoring. Lab testing has been the monitoring system in place for NOx emission over the last 15 years. The extensive amount of misconduct for on-road emissions has indicated the low effectiveness of lab testing. As a result, the EU Commission decided to improve the effectiveness of monitoring by choosing a more advanced and sophisticated on-road testing procedure.

If an automaker fails to meet standards, we denote the probability of not being caught by $p(m, \gamma_i)$, a function of both the effort γ_i and the monitoring effectiveness m. We assume that increasing the level of effort increases the probability of not being caught, i.e., $\frac{dp}{d\gamma} > 0$.

[7]The first inequality ensures that price is decreasing in quantity. The second inequality ensures that profit is concave in quantity.

We also assume that increasing the monitoring effectiveness decreases the probability of not being caught, i.e., $\frac{dp}{dm} < 0$. We also assume that the marginal effect of each of greater effort and monitoring effectiveness on the probability of not being caught is decreasing, i.e. $\frac{d^2p}{dm^2} > 0$ and $\frac{d^2p}{d\gamma^2} < 0$.

Consistent with Volkswagen's 18.28 billion settlement for its NOx emission misconduct, carmakers pay a penalty if they are detected violating standards. We assume that the penalty equals the profit on all vehicles that violated standards. In other words, an automaker makes a profit on the cars sold only if no misconduct is detected (either because standards were not violated or the violation was not detected). If misconduct is detected, the carmaker makes a profit of 0. Under this assumption the profit function of carmaker i is given by $\pi_i(q_i, Q, \gamma_i) = p(m, \gamma_i)q_i[P(Q) - c(\gamma_i, \lambda)]$. We use this simplified form of the penalty but the main messages of this model do not change if the penalty is proportional to the profit or depends on the degree of misconduct.[8]

Each carmaker must select its effort level γ_i and its production quantity q_i. A higher effort level decreases the probability of being caught for misconduct but increases the unit production cost. Carmakers choose their production quantity and effort level to maximize expected profits given the current strictness of standards and level of competition. The first order conditions for carmaker i with respect to quantity q_i and effort γ_i imply that

$$P(Q) - c(\gamma_i, \lambda) + q_i P'(Q) = 0 \tag{2.1}$$

[8]If the penalty depends on the extent to which a firm violates the market standards, however it is defined, then one may obtain in some cases that competition does not affect the degree of investment in satisfying the market standards. However, if there is some uncertainty as to how the monitoring authority evaluates the extent to which a firm violates the market standards then, again, if there is limited liability and sufficient uncertainty, the main messages presented above would still be present.

$$\frac{dp}{d\gamma_i}[P(Q) - c(\gamma_i, \lambda)] - p(m, \gamma_i)\frac{dc}{d\gamma_i} = 0 \tag{2.2}$$

Under the assumption that all carmakers are homogenous, each of them will select the same quantity q and effort level γ. Thus, we can drop the subscript of quantity and effort for each carmaker. Totally differentiating the equilibrium q and γ, we obtain

$$\frac{d\gamma}{dN} = \frac{qP'(Q)^2\frac{dp}{d\gamma}}{D} < 0$$

where $D = [P(Q)-c(\gamma,\lambda)]\{-[(N+1)P'(Q)+QP''(Q)]\frac{d^2p}{d\gamma^2}\}+[(N+2)P'(Q)+2QP''(Q)]\frac{dc}{d\gamma}\frac{dp}{d\gamma}+[(N+1)P'(Q)+QP''(Q)]p(m,\gamma)\frac{d^2c}{d\gamma^2}$

This result suggests Proposition 2.1 linking effort level to competition intensity.

PROPOSITION 2.1 As the intensity of competition increases (the number of carmakers increases), each carmaker decreases the optimal effort to meet emission standards.

Consistent with previous research (Bennett et al. (2013), Utgård et al. (2015)), our model indicates that increasing competition increases misconduct. The rationale behind this is "a bid to survive". As fiercer competition leads to lower prices, carmakers lower their effort in order to reduce their unit production cost even though this increases the probability of misconduct being detected. Our empirical analysis in Sections 2.5 and 2.6 validates the implication of Proposition 2.1 in the context of NOx emissions.

For our next result, we need a stronger assumption on the inverse demand function. We assume that $P'(Q) + QP''(Q) < 0$, i.e., the second derivative of the inverse demand

function is dominated by the first derivative. This condition is called the strongly dominated first-order effect and is satisfied by most common inverse demand functions.[9]

Using equation (2.1) and equation (2.2) we obtain

$$\frac{d\gamma}{d\lambda} = -\frac{[(P'(Q) + QP''(Q)]\frac{dc}{d\lambda}\frac{dp}{d\gamma} + [(N+1)P'(Q) + QP''(Q)]p(m,\gamma)\frac{d^2c}{d\gamma d\lambda}}{D} < 0$$

if $P'(Q) + QP''(Q) \leq 0$.

This result suggests Proposition 2.2 linking effort level to the strictness of standards.

PROPOSITION 2.2 If the inverse demand function satisfies the strongly dominated first-order effect ($P'(Q) + QP''(Q) \leq 0$), the optimal effort exerted by each carmaker decreases as emission standards become stricter.

The rationale behind Proposition 2.2 is that as standards get tighter, it becomes more costly for carmakers to comply with the standards. As carmakers now have less to lose from being caught for misconduct, the penalty for being caught violating standards becomes less threatening to them. As a result, carmakers exert a lower effort with tighter standards even though it increases the risk of being caught violating standards.

In summary, the findings from our theoretical model suggest that misconduct is more likely to occur when competition becomes fiercer and when standards become tighter. Next, we examine these findings with empirical analysis to see whether they hold with regards to NOx emissions.

[9]Common inverse demand functions (Huang et al. (2013)) include linear models: $P(Q) = a - bQ$ where $a, b > 0$; logarithmic models: $P(Q) = log(\frac{Q}{a})/b$ where $a > 0, b < 0$ and exponential models $P(Q) = (a - e^Q)/b$ where $a, b > 0$.

2.4. Data and Variable Definition

We construct our dataset by linking four distinct data sources between the years 2000 and 2012: (i) car-by-car on-road emission data from a European country road; (ii) EU vehicle sales catalog covering all vehicle models offered; (iii) EU Vehicle Registration dataset; and (iv) EU Vehicle Emission Standards. Below, we discuss some characteristics of these data sources in more detail, along with definitions of variables in our empirical analysis.

2.4.1. Dataset

The on-road emission dataset is the core of our study and enables us to analyze actual NOx emission from cars traveling on European roads. The dataset is collected by sensors installed on European country roads and captures each passing vehicle's emission information and features. For each vehicle, the license plate tells the vehicles' basic information about the carmaker, the model, and its first registration date; the speed sensors record the speed and acceleration at the moment of measurement; and the emission sensors measure the NOx emission on-road. Chen and Borken-Kleefeld (2014) provide more details about the techniques for measurement and collection of the data. The original data sample is collected in the first day of June, July and August[10] for each year from 2000 to 2014. Altogether we have 288350 records of pass-by vehicles. In our research, we focus on diesel passenger cars sold in the EU market that are under the enforcement of the three NOx-related emission standards, Euro 3 to Euro 5 (from 2000 to 2012). Moreover,

[10]The reason to choose these three months is because the temperature in the summer months is similar to the lab environment. Vehicles operated in colder temperature emit more NOx. In order to make the on-road measurement comparable to the lab-testing environment, the sensors only measure vehicles during summer months.

we restrict our attention to vehicles that are less than two-year old[11] with a driving speed and acceleration[12] when the emission data is recorded that are comparable to lab testing conditions. The empirical analysis is conducted based on a 13-year on-road NOx emission dataset of 41883 passenger vehicles from 57 carmakers[13]. The emission data allows us to identify whether cars were emitting NOx within standards or not.

To obtain key features of each car we use the publicly available EU vehicle sales catalog. Based on the information about each vehicle's carmaker and model in the on-road emission dataset, we use the sales catalog to obtain each vehicle's MSRP (manufactures suggested retail price), horsepower and vehicle gross weight. These vehicle features and price have a significant influence on customer preferences as well as the main trade-offs made by carmakers when deciding on the level of effort to put towards reducing NOx emission. We also use this data to cluster and identify substitute models in the market.

The publicly available EU vehicle registration dataset from 2000 to 2012 records the vehicle registration number of each car sold during that period. This data is used to obtain the annual market share of each carmaker and thus implies the market-level competition intensity.

The EU emission standards from the EU Commission website provide the limits of NOx emission for each vehicles to be granted market entry. Here we focus on the limits

[11]Vehicles' emission performance gets worse as they age. The emission tests are enforced on new cars. Hence we restrict our attention to new cars to accurately detect misconduct.

[12]The lab emission test measures vehicles' emission performance under controlled driving conditions because speed and acceleration affect emission. Hence, we restrict vehicles on the road to comparable levels of speed (between 20 km/h and 58 km/h) and acceleration (between 0 m/s^2 and 2.8 m/s^2).

[13]Given the variety of ownership changes over the sample, we construct a stable definition of carmakers based on the logo of vehicles. For example, we keep Audi and Volkswagen separate throughout the sample even though they both belong to the Volkswagen Group.

that regulate diesel-fueled passenger cars between 2000 and 2012. Across the 13 years, the market limits were tightened three times as shown in Table 2.1.

2.4.2. Variable Definition

We define the misconduct, M, for each pass-by vehicle as a dummy variable with $M = 1$ if the on-road NOx emission exceeds the standards limits and $M = 0$ otherwise.

We measure the competition intensity at two-levels: the annual market-level competition intensity, C^{market} which impacts all vehicles sold in that year, and a model-level competition intensity, C^{model} which measures the product substitutability of each vehicle model. According to prior research (Raith et al. (2003), Sutton (2007), Vives (2008)), market-level competition alone may not capture all competition faced by firms. Using product substitutability can complement the market concentration in measuring the competition intensity.

We define the market-level competition, C^{market}, for each year using the Herfindahl Index. Calculated from carmakers' market share, the Herfindahl index measures market concentration. A market is highly concentrated if the Herfindahl Index is above 0.25 and is highly competitive if the index is below 0.01. Define $S_{j,y}$ to be the market share of carmaker j in year y and N_y to be the number of carmakers competing in year y. We construct the variable C^{market} as $C_y^{market} = \sum_{j=1}^{N_y} S_{j,y}^2$.

We define the model-level competition, C^{model}, faced by each vehicle model as the number of substitute models available in the market. We determine the substitutes by clustering analysis on vehicle features (power, and weight) and price. Vehicle models with similar vehicle features and at comparable price levels offered one-year before or one-year

after a model are also considered as potential close substitutes. We use clustering analysis using vehicle features and selling price on the potential substitute data to group them into several clusters. For each model, its model-level competition is measured as C^{model}, the number of distinct vehicle models in its cluster.

We define the tightness of the standards, R, for each year by normalizing the limit value relative to the limit in 2000. For example, the NOx limit in 2000 was 0.50 and the limit in 2012 was 0.18. This implies that in 2000 the tightness of standards was equal to 1, R_{2000}=1 and in 2012, the tightness of standards was equal to $R_{2012} = 0.18/0.50 = 0.36$. The smaller the number R, the tighter the standards.

We use the variable $Maker_i$ to identify the carmaker for vehicle i and variables X_i to identify vehicle specifications of price, horsepower, gross weight and vehicle age that are used in the empirical analysis.

We summarize the definition and source of variables used in the empirical analysis in Table 2.2.

2.5. Hypotheses and Empirical Models

In this section, we empirically examine the insights from the theoretical model in Section 2.3. Using the dataset described in Section 2.4, we analyze how competition and strict standards link to misconduct in the auto industry. We propose six econometric models to properly check the robustness of our empirical analysis.

Our theoretical model in Section 2.3 indicates that both the level of competition (Proposition 2.1) and the strictness of standards (Proposition 2.2) are linked to the level of misconduct observed. Our first goal is to confirm this fact empirically. Our second

Variable	Definition	Source
M_i	Dummy variable. 1 if vehicle i's on-road NOx emission surpasses the enforced limit. 0 otherwise.	On-road Emission Dataset and EU Vehicle Emission Standards
C_{year}^{market}	Continuous variable. The smaller the value, the fiercer the competition in the market.	Annual Vehicle Registration in EU
C_i^{model}	Continuous variable. The larger the value, the fiercer the model-level competition faced by vehicle i.	EU Vehicle Sales Catalog
R_{year}	Continuous variable. The larger the value, the looser the standard in specified year.	EU Vehicle Emission Standards
X_i	Continuous variables. Vehicle i's features including vehicle prices, horsepower, vehicle weight, and vehicle age.	EU Vehicle Sales Catalog and On-road Emission Dataset
$Maker_i$	Categorical variable. Vehicle i's carmaker.	On-road Emission Dataset

Table 2.2. Lists of the name, definition and sources of the variables

goal is to check whether the strictness of standards directly impacts misconduct or does so through the level of competition. These two goals lead to our first set of hypotheses below:

HYPOTHESIS 1 Increasing market-level competition leads to more misconduct.

HYPOTHESIS 2 a. Stricter standards directly lead to more misconduct.

b. Stricter standards indirectly lead to more misconduct via the level of market competition.

To test whether competition and standards impact misconduct, we start with Probit models that link the probability of misconduct with the linear form of the strictness of standards and market-level competition intensity.

$$\text{Model 2.0: } Pr(M_i = 1) = \Phi(\beta_0 + \beta_2 R_{year} + e)$$

$$\text{Model 2.1: } Pr(M_i = 1) = \Phi(\beta_0 + \beta_1 C_{year}^{market} + \beta_2 R_{year} + e)$$

In Model 2.0 and Model 2.1, $Pr(M_i = 1)$ denotes probability of misconduct, and Φ is the cumulative distribution Function (CDF) of the standard normal distribution.

Using Model 2.0 and Model 2.1, we can test whether competition and the strictness of standards jointly affect misconduct. If the coefficients of competition intensity (β_1) and strictness of standards (β_2) are both significantly different from zero in Model 2.1, then both factors influence misconduct. If, however, the coefficient of strictness of standards (β_2) in Model 2.0 (where competition is absent) is significantly different from zero, but becomes insignificant in Model 2.1 (where market level competition is present), then we would conclude that standards have an indirect effect on misconduct via market level competition. Moreover, based on the signs of coefficients β_1 and β_2 from Model 2.1, we can obtain the direction of the effect of market level competition and the strictness of standards on misconduct.

Our next objective is to examine the monotonicity and linearity of the impact of the strictness of standards and the level of market competition on misconduct. We do so because some previous research suggests that the effects of competition can be nonlinear (Bennett et al. (2013), Bresnahan and Reiss (1991)) and some research suggests that such relationship can also be non-monotone (Olivares and Cachon (2009)). The following hypotheses assume both effects to be linear:

HYPOTHESIS 3 The effect of increased market competition on misconduct is linear.

HYPOTHESIS 4 The effect of stricter standards on misconduct is linear.

We use Model 2.2 to test the two hypotheses. Model 2.2 incorporates quadratic terms for both the level of market competition and the strictness of standards.

$$\text{Model 2.2} : Pr(M_i = 1) = \Phi(\beta_0 + \beta_{1,1} C_{year}^{market} + \beta_{1,2} C_{year}^{market^2} + \beta_{2,1} R_{year} + \beta_{2,2} R_{year}^2 + e)$$

In Model 2.2, if the coefficients of the second-order terms ($\beta_{1,2}$ and $\beta_{2,2}$) are significantly different than zero, we will conclude that the relationships are nonlinear. Moreover, based on the coefficients in front of the first and second-order terms combined with the value of competition and strictness of standards ($\beta_{1,1} + 2\beta_{1,2} C_{year}^{market}$) and ($\beta_{2,1} + 2\beta_{2,2} R_{year}$), we can examine whether the relationships are monotone.

To examine the robustness of the results from Models 2.0 to 2.2, we examine Hypotheses 1-4 again with Models 2.3 to 2.5 that include additional variables in the Probit regression.

In Model 2.3, we add variables corresponding to vehicle features (price, weight, and power) that influence customer buying behavior as well as the trade-offs considered by carmakers when selecting their effort towards NOx emission reduction.We also include the age of the vehicle because an older vehicle is likely to have higher emissions. In the Probit regression, the corresponding variables used are X_i^{price} (vehicle price), X_i^{hp}(horsepower), X_i^{wt}(vehicle weight), and X_i^{age}(vehicle age).

$$\text{Model 2.3} : Pr(M_i = 1) = \Phi(\beta_0 + \beta_{1,1} C_{year}^{market} + \beta_{1,2} C_{year}^{market^2} + \beta_{2,1} R_{year} + \beta_{2,2} R_{year}^2 + \beta_3 X_i^{price} + \beta_4 X_i^{hp} + \beta_5 X_i^{wt} + \beta_6 X_i^{age} + e)$$

In Model 2.4, we also include carmakers' fixed effects ($Maker_i^m$). Many empirical models in the literature have included manufacturer fixed effects because different carmakers may

adopt different business strategies (Olley and Pakes (1992), Berry et al. (1995), Goldberg (1998), Knittel (2011)) and embrace different attitudes towards misconduct (Bennett et al. (2013), Utgård et al. (2015)). Hence in Model 2.4 we examine whether standards and competition still affect misconduct while controlling for the carmakers' effects.

$$\text{Model 2.4} : Pr(M_i = 1) = \Phi(\beta_0 + \beta_{1,1} C_{year}^{market} + \beta_{1,2} C_{year}^{market^2} + \beta_{2,1} R_{year} + \beta_{2,2} R_{year}^2 + \beta_3 X_i^{price} + \beta_4 X_i^{hp} + \beta_5 X_i^{wt} + \beta_6 X_i^{age} + \sum_{m \in Maker} \beta_m Maker_i^m + e)$$

In Model 2.5, we include model-level competition (C_i^{model}) as an additional measure of competition intensity. C_i^{model} is evaluated as the number of substitutes for each carmaker's model in year i. In Models 2.1-2.2, we measure competition only at the market level with the Herfindahl Index to indicate market concentration level. Some prior research (Sutton (2007), Raith et al. (2003), Vives (2008)) suggests that product substitutability can also indicate competition intensity and should be examined along with market-level competition. Hence, we incorporate both first-order and second-order terms for model-level competition into Model 2.5. With Model 2.5, we examine whether the previous findings still hold and whether model level competition has any impact on misconduct.

$$\text{Model 2.5} : Pr(M = 1) = \Phi(\beta_0 + \beta_{1,1} C_{year}^{market} + \beta_{1,2} C_{year}^{market^2} + \beta_{2,1} R_{year} + \beta_{2,2} R_{year}^2 + \beta_3 X_i^{price} + \beta_4 X_i^{hp} + \beta_5 X_i^{wt} + \beta_6 X_i^{age} + \sum_{m \in Maker} \beta_m Maker_i^m + \beta_7, 1 C_i^{model} + \beta_7, 2 C_i^{model^2} + e)$$

This leads to our final two hypotheses related to the effect of model level competition.

HYPOTHESIS 5 Increasing model-level competition leads to more misconduct.

HYPOTHESIS 6 The effect of increasing model-level competition on misconduct is linear.

2.6. Empirical Results

In this section, we analyze the estimation results of the empirical models in Section 2.5 and examine how strictness of standards and competition impacts misconduct. The estimation results for all models are shown in Table 2.3. Table 2.3 shows the estimated coefficient for each variable in Models 2.0 to 2.5 along with the level of significance with which the coefficient is different from 0.

Observe that Hypotheses 1 and 2a are empirically supported by the results of Model 2.0 and Model 2.1. The negative coefficient of R (strictness of standards) in Model 2.0 confirms that tightening standards leads to more misconduct. The negative coefficients of both R and C^{market} (level of market competition) in Model 2.1 confirm that both stricter standards and fiercer competition together drive misconduct. In other words, the empirical evidence suggests that stricter standards independently influence misconduct and not via the level of market competition.

The results of Models 2.2-2.4 provide a robust confirmation of Hypotheses 1 and 2a. From Table 2.3 observe that the coefficients of both R and C^{market} remain significantly negative across Models 2.2-2.4. In other words the effect of strictness of standards and market competition on misconduct is observed even when we include vehicle characteristics and carmaker fixed effects. The results of Models 2.2-2.4 also support Hypotheses 3 and 4 that the effect of stricter standards and fierce competition on misconduct is linear.

From Table 2.3 observe that while the coefficients of R and C^{market} are significant in Models 2.2-2.4, the coefficients of the corresponding quadratic terms are not significantly different from 0. The results of Models 2.3 and 2.4 also show that vehicle characteristics such as horsepower and weight have a significant influence on misconduct. In other words, carmakers seem to account for vehicle characteristics that are important to customers when selecting their effort level towards NOx emission reduction.

Hypothesis 5 is empirically validated by the results of Model 2.5. The results show that increasing model level competition increases the level of misconduct though the relationship is not linear. The estimation results from Model 2.5 indicate that as the number of substitutes increases, the marginal increase in likelihood of misconduct also increases. In other words, vehicle models with more substitutes are more sensitive to the pressure of model-level competition in terms of misconduct.

We summarize our hypothesis testing results in Table 2.4.

All our empirical results confirm the main message of our paper that increasing competition intensity (whether market level or model level) and stricter standards lead to a higher level of misconduct.

To understand the impact of competition and regulation, we look at the estimation results from the most complete model, Model 2.5. The marginal effect of variable X on the probability of misconduct is measured by $\Phi(X\beta)\beta_x$ and the results for all variables are shown in Table 2.5. Our estimation results show that a 1% increase in market level competition increases the probability of misconduct by 0.58%; a 1% tightening of standard limits increases the probability of misconduct by 1.72%; and each additional model substitute available in the market increases the probability of misconduct by 0.48%. Moreover, the

Variable	Model 2.0	Model 2.1	Model 2.2	Model 2.3	Model 2.4	Model 2.5
R	-0.27***	-0.31***	-12.61***	-24.25***	-23.96***	-22.95***
	(0.05)	(0.11)	(2.22)	(2.67)	(2.74)	(2.74)
R^2			-2.35	1.84	0.016	0.69
			(2.75)	(2.86)	(2.52)	(2.94)
C^{market}		-14.44***	-9.43***	-8.16***	-8.66***	-7.70***
		(3.32)	(2.41)	(2.46)	(2.52)	(2.54)
${C^{market}}^2$			2.5	1.05	2.20	2.31
			(2.63)	(2.69)	(2.73)	(2.72)
Prices				-2.82×10^{-6}	-5.83×10^{-7}	-1.14×10^{-6}
				(1.52×10^{-6})	(1.96×10^{-6})	(1.96×10^{-6})
Horsepower				0.01***	0.01***	5.67×10^{-3}***
				(6.47×10^{-4})	(7.26×10^{-4})	(7.29×10^{-4})
Weight				2.53×10^{-4}**	2.48×10^{-4}**	2.30×10^{-4}**
				(6.91×10^{-5})	(8.51×10^{-5})	(8.51×10^{-5})
Vehicle Age				1.01×10^{-3}*	1.13×10^{-3}*	1.15×10^{-3}*
				(5.43×10^{-4})	(5.52×10^{-4})	(5.52×10^{-4})
Maker's Effect					Yes	Yes
C^{model}						-1.07
						(5.00)
${C^{model}}^2$						6.43*
						(2.72)

Note: "***" means significance at 0.1 percent level, "**" at 1 percent level, "*" at 5 percent level.

Table 2.3. Model Estimation Results

results in Table 2.5 also show the trade-offs between vehicle features and NOx reduction. The probability of misconduct will be higher by 8.57×10^{-6}% with a 1% decrease in price, higher by 1.73×10^{-3}% with a 1% increase in vehicle weight, and higher by 4.26×10^{-2}% with a 1% increase in engine power. As vehicles age, they emit more NOx emission. The probability of emitting NOx on-road beyond limits increases by 8.65×10^{-5}% when the vehicle becomes 1 year older.

Hypotheses	Testing Results
Hypothesis 1. Increasing market level competition leads to more misconduct.	True
Hypothesis 2a. Stricter standards directly lead to more misconduct.	True
Hypothesis 2b. Stricter standards indirectly lead to more misconduct via increased market competition.	False (2a is correct)
Hypothesis 3. The effect of increased competition on misconduct is linear.	True
Hypothesis 4. The effect of stricter standards on misconduct is linear.	True
Hypothesis 5. Increasing model-level competition leads to more misconduct.	True
Hypothesis 6 The effect of increasing model-level competition on misconduct is linear.	False (Monotone but convex)

Table 2.4. Hypotheses Testing Results

Variable	Coefficient Estimates	Marginal Effect of 1% Increase in Variable on Probability of Misconduct
R	-22.95***	-1.72%
R^2	0.69	
C^{market}	-7.70***	0.58%
C^2_{market}	2.31	
Prices	-1.14×10^{-6}	-8.57×10^{-6}%
Horsepower	5.67×10^{-3}***	4.26×10^{-4}%
Weight	2.30×10^{-4}**	1.73×10^{-5}%
Vehicle Age	1.15×10^{-3}*	8.65×10^{-5}%
Maker's Fixed Effect	Yes	
C_{model}	-1.68	
C^2_{market}	6.43*	0.48%

Table 2.5. Marginal Effects of Variables on Probability of Misconduct

2.7. Counterfactual Analysis

The findings in our paper suggest that misconduct is more likely when standards are tightened or competition intensity increases. In this section, we will first use theoretical models to evaluate how optimal standards should be set to improve social welfare and

examine whether the EU decision to relax emission standards is reasonable. Then we will use counterfactual analysis to examine the extent to which misconduct is likely to change in response to the new EU NOx emission standards.

This section is motivated by the actions taken by the EU Commission after misconduct on the part of automakers was detected. The commission took two main actions. Their first action was to replace the old lab testing (which was unable to monitor on-road emissions) with the Real Driving Emissions test procedure starting from September 1, 2017. The second action was to relax the NOx emission standards. In the short term until September 1, 2019, carmakers will be allowed to emit up to twice the current NOx standards. After that, the on-road vehicles will still be allowed to emit up to 50% more than the current standards. The changes can be characterized as improving the monitoring effectiveness but loosening the emission standards.

While the objective of improved monitoring is relatively clear, our goal is to better understand and motivate the rationale that may have the led the EU Commission to temporarily relax its NOx emission standards. Both moves would make sense if they lead to a lower level of misconduct. We provide a simple model and empirical counter factual analysis that supports the case that improved monitoring effectiveness and somewhat looser standards are likely to reduce the extent of misconduct related to NOx emission.

2.7.1. Theoretical Models

Continuing with the theoretical models built in Section 2.3, we examine how regulators can set the strictness of standards to improve social welfare and use the analytical results to empirically review the two actions taken by the EU Commission.

Regulators' objective is to maximize the social welfare composed of two parts: the total surplus received by customers and carmakers through trading and the social cost to the environment related to emissions. The cost related to emissions is affected by both the strictness of the standards and the extent of carmakers' misconduct . Responding to the number of competing carmakers in the market N, the strictness of standards λ and the effectiveness of monitoring m, in equilibrium carmakers choose the production quantity q and effort level γ (to improve emissions). From Section 2.3, recall the inverse demand function $P(.)$ and carmakers' cost function $c(\gamma, \lambda)$ related with their effort γ to comply with the standards λ. The total surplus through market trading is $\int_0^Q P(x)dx - c(\gamma, \lambda)Q$ where $Q = Nq$ is the total quantity of production. Regarding the social cost to the environment, we look at the impact of the total supply, Q units, in two parts: $(1-\gamma)Q$ units emit NOx beyond standard limits due to misconduct whereas γQ units comply with the standards. We assume the social cost for each unit emitting beyond standards is k and the social cost for each complying unit is $G(\lambda)$ (this cost depends on the strictness of standards). We assume that $G(\lambda)$ becomes smaller as the standards become stricter. When standards are the strictest, i.e., $\lambda = 1$, the social cost for each complying unit is 0. When there is no standard, i.e., $\lambda = 0$, the social cost for each complying unit is equal to k. Regulators choose the strictness of standards λ and the effectiveness of monitoring m to maximize the social welfare. The social welfare function is written as

(2.3)

$$S(Q(m), \gamma(m), \lambda) = \int_0^{Q(m)} P(x)dx - c(\gamma(m), \lambda)Q(m) - k(1-\gamma(m))Q(m) - G(\lambda)\gamma(m)Q(m).$$

The social welfare function is concave in the strictness of standards λ. The optimal strictness of standards λ^* corresponding to the highest social welfare satisfies $\frac{dS(Q,\gamma,\lambda^*)}{d\lambda} = 0$.

Though the following results hold in general functional forms[14], here we use a linear example to illustrate how the optimal standards should be set corresponding to the competition intensity and monitoring effectiveness. Suppose the inverse demand function is $P(Q) = 1 - Q$. In this case, the cost function is $c(\gamma, \lambda) = c_0 + \alpha\gamma + \beta\lambda$ (where $\alpha, \beta > 0$) and the probability for regulators to detect misconduct is $1 - p(m, \gamma)$ equal to $1 - \gamma^m$. In equilibrium, carmakers choices for effort and quantity are $\gamma = m\frac{1-c_o-\beta\lambda}{\alpha(1+m+N)}$ and $q = \frac{1-c_0-\beta\lambda}{1+m+N}$. After inserting the relevant parts into the social welfare function, we obtain that the optimal strictness of the standards λ^* should satisfy

$$G(\lambda^*)m\frac{2}{\alpha} - G'(\lambda^*)m\frac{1-c_0-\beta\lambda^*}{\alpha\beta} + k\frac{1+m+N}{1-c_0-\beta\lambda^*} - m\frac{2k}{\alpha} - N - 2 = 0.$$

Moreover, the optimal standards λ^* reacts to the competition intensity in the following way:

$$\frac{d\lambda}{dN} = \frac{[k-(1-c_0-\beta\lambda)](1-c_0-\beta\lambda)}{G'(\lambda)m\frac{3}{\alpha}(1-c_0-\beta\lambda)^2 - G''(\lambda)m\frac{(1-c_0-\beta\lambda)^3}{\alpha\beta} + k(1+m+N)\beta}.$$

Thus, we have

$$\frac{d\lambda}{dN} < 0 \tag{2.4}$$

if $N > \frac{-3G'(\lambda)m(1-c_0-\beta\lambda)^2+G''(\lambda)m\frac{(1-c_0-\beta\lambda)^3}{\beta}-k(1+m)\alpha\beta}{\alpha k\beta} = C^*$

[14]The results still hold for convex inverse demand function, convex cost function in effort and the strictness of standards.

We summarize the finding in Proposition 2.3 to explain how the optimal standards should be set according to the competition intensity.

PROPOSITION 2.3 For linear demand and cost functions, there exists a threshold of competition intensity C^*. When competition is fiercer than the threshold C^*, the optimal standards should be set looser as competition becomes fiercer. When competition is milder than the threshold C^*, the optimal standards should be set stricter as competition becomes fiercer.

The analytical results rationalize the action of the EU Commission to stretch the standard limits. Given the relatively high competition intensity in the EU auto market, the previous emission standards were set stricter than the optimal level. As standards were too strict, carmakers exerted less effort in complying with the standards, which led to more misconduct (Proposition 2.2), thus hurting social welfare. Thus, it is reasonable for the EU commission to relax standards in the short term.

Next, we look at the EU Commission's decision to improve monitoring effectiveness. As the EU Commission replaces the outdated lab testing with the more effective Real Driving Emissions testing, the effectiveness of monitoring, m, increases. In equation (2.4), it shows the threshold of competition intensity C^* increases as monitoring effectiveness m increases ($\frac{dC^*}{dm} > 0$). That is to say, as monitoring improves, regulators can tighten standards under conditions of higher competition intensity. Suppose the current market has N_0 carmakers, if the monitoring is strengthened to the level m_0 where $\frac{-3G'(\lambda)m_0(1-c_0-\beta\lambda)^2+G''(\lambda)m_0\frac{(1-c_0-\beta\lambda)^3}{\beta}-k(1+m_0)\alpha\beta}{\alpha k\beta} > N_0$, then the regulators can always tighten the standards to improve social welfare. Moreover, from equation (2.1) and (2.2), it shows that carmakers exert more effort in complying with the standards under a more

effective monitoring system ($\frac{d\gamma}{dm} > 0$). We summarize these two effects of improved monitoring in Proposition 2.4.

PROPOSITION 2.4 If regulators improve the effectiveness of monitoring, carmakers exert more effort in complying with the emission standards. Under linear demand and cost functions, the threshold of competition intensity C^* increases as monitoring improves. Hence regulators can tighten standards under conditions of higher competition intensity as monitoring effectiveness improves.

By adopting the new Real Driving Emission tests, the EU Commission incentivizes carmakers to exert more effort in complying with the standards. Better monitoring will also allow the EU to continue tightening standards in the future independent of competition intensity. This explains the feasibility of EU's plan to initially relax standards and then quickly tighten them in just two years.

2.7.2. Empirical Analysis

Based on the most sophisticated empirical model, Model 2.5 in Section 2.5, we conduct a counterfactual analysis to examine how the probability of misconduct will change as EU relaxes its emission standards to be twice as large as the one of Euro 5. In Figure 2.2, we visualize the impact calculated from the counterfactual analysis. Overall, loosening standards has an impact on misconduct reduction. Our counterfactual analysis predicts that by relaxing the standards limits to twice the existing level, the probability of misconduct will be reduced by 11.04% in a perfectly differentiated market (with the model-level competition to be the case "no substitutes") and will be reduced by 9.56% in a perfectly substituted market (with the model-level competition to be the case "all substitutes")

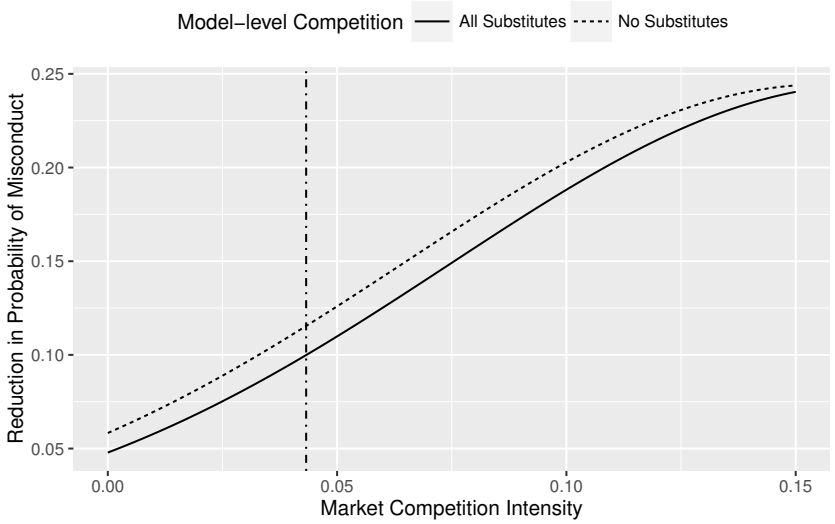


Figure 2.2. Probability of misconduct under various standards tightness and competition intensity

at the latest market-level competition intensity in our sample[15], marked by the vertical black line. Notice our Model 2.5 doesn't include the effectiveness of monitoring because EU has used the same monitoring system to test NOx emission from 2000 to 2016, which covers the timespan of our sample. Hence, our counterfactual analysis, which excludes the monitoring variable, calculates the minimum reduction in the probability of misconduct with the introduction of the new standards. The actual reduction should be larger than the estimates because carmakers exert more effort in NOx control under a strengthened monitoring system as shown in Section 2.7.1.

[15]The latest market-level competition intensity measured in our sample is 0.043. It shows the market is not yet highly competitive but certainly not concentrated.

2.8. Conclusion

In this research, we use our theoretical and empirical analysis (using data from 13-year records of car-by-car on-road emission on European roads) to show that carmakers' tendency to commit misconduct increases as competition becomes more fierce and standards become stricter. Our counterfactual analysis points out that the regulators should set the strictness of standards considering the competition intensity. In general, improved monitoring should always accompany a tightening of standards when competition intensity exceeds a threshold.

Our research mainly looks at misconduct based on the economic and regulatory environment. In future studies, it is worth studying how carmakers' ownership structure and their interaction impact the misconduct as the studies conducted in other industries (Bertrand and Lumineau (2015)). Moreover, prior research mentions the design of optimal monitoring system (Duflo et al. (2014), Pierce and Snyder (2008)), our research didn't study such aspects due to the lack of changes in the monitoring system of EU auto market from 2000 to 2012. However, after the new EU emission standards have been enforced for years, research can also empirically study the effect of improved monitoring on misconduct.

CHAPTER 3

Forecasting Product Life Cycle Curves: Practical Approach and Empirical Analysis (joint with Jason Acimovic, Francisco Erize, Doug Thomas, Jan A. Van Mieghem)

3.1. Introduction

Many companies seek to innovate and bring new products and services to market, and growth and product innovation remain top priorities for executives. CEOs have indicated their commitment to new product development growing over time (PWC 2016) and at least one survey reports new product development as their top investment priority (KPMG 2016). One common metric used to evaluate the success of a firm's innovation efforts is the percentage of revenue derived from new products. Based on a cross-industry survey, Cooper and Edgett (2012) report that an average of 27% of a firm's revenue comes from new products. (This percentage varied dramatically across respondents; 27% would be quite high for a food or consumer goods manufacturer and quite low for a technology firm.) The same survey also reports that the percentage of *profit* coming from new products lags the percentage of revenue coming from new products. This suggests that while new products are essential to growth, they are expensive to support.

One of the largest challenges in managing new product introductions is creating sales forecasts. Here it is important to distinguish how firms approach forecasting in general, and how the approach may differ for new product forecasting. Several studies indicate that statistical methods play a major role for sales forecasts of mature products. For example, based on a survey of 144 forecasting practitioners, Fildes and Goodwin (2007) report that 75% of all forecasts are generated or at least influenced by a statistical forecast. Contrast this with new product forecasting where market-research based methods and executive opinion dominate (Kahn 2002). While these approaches may be best, or even the only viable approach, for completely new market entries, most new products are not unlike anything we have ever seen before. Focusing just on new product forecasting, Kahn (2002) separates new products according to their "newness," ranging from incremental cost or product-attribute improvements to "new-to-the-world" market entries. Perhaps surprisingly, the survey results in Kahn (2002) indicate that the most popular three techniques used in practice—market research, executive opinion and sales force input—are the same across the range of product newness. That is, even when a firm has historical data on a similar product, market research, executive opinion and sales input are still the most commonly used approaches.

Statistical forecasting for new, but not earth-redefining-new products is the focus of this paper and where we seek to make a contribution. Our objective is to develop an approach that can be effectively applied to generate forecasts for new products that are similar to previous products. Our industrial partner is Dell, and the personal computer industry, characterized by very high reliance on new product revenue and short product lifecycles, is our motivating setting. We describe the business environment in greater

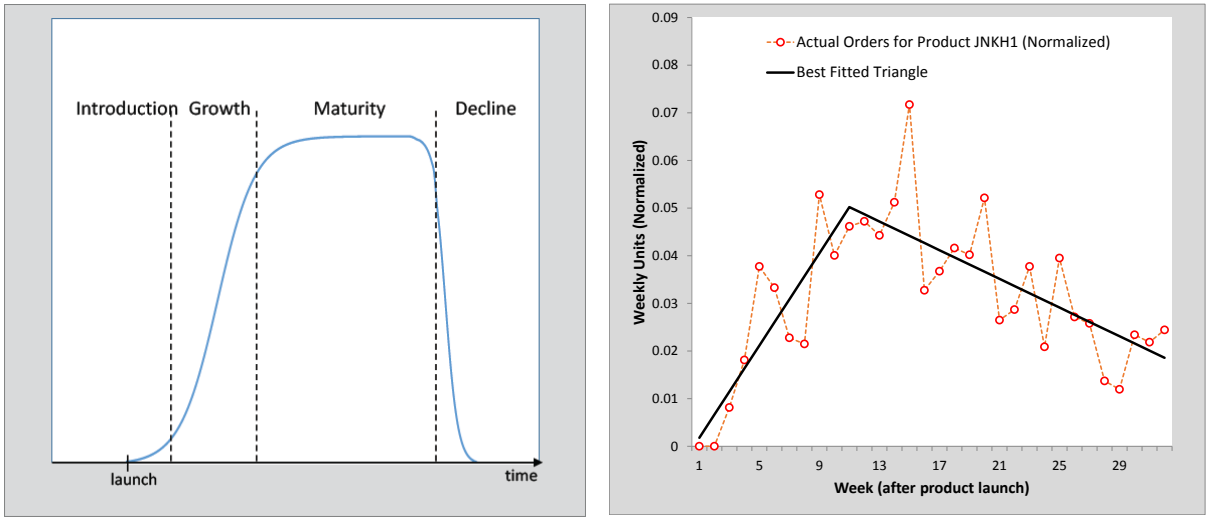


Figure 3.1. A typical PLC curve (left) has four phases. The actual orders of the majority of short-lifecycle technology products at our partner company are best described by a triangular PLC curve (right) with two phases.

detail below. The difference in product lifecycles for computers was observed quite early, with Goldman (1982) pointing out that computer sales were "...characterized by a short life on the market, a steep decline stage and the lack of a maturity stage." Figure 3.1 shows a typical product lifecycle curve with introduction, growth, maturity and decline phases next to actual customer orders for one of the products in our data set. The simple, triangular PLC curve shown in Figure 3.1 fits historical orders quite well. As we will see, the pattern in Figure 3.1 is representative as many products in our data set show no "mature" or "sustain" stage and are well fit by simple, piecewise-linear PLC curves.

While new product forecasting is of critical importance for personal computers, and this industry is an important one, our approach for product lifecycle forecasting is general and could be applied in other settings. The central idea behind our approach is to (1) use the historical product life cycle (PLC) customer order information of previous similar

products to fit a PLC curve and to (2) use the PLC curve to forecast the entire customer order evolution of ready-to-launch new products that are similar to past products. We use and compare several families of functional forms for fitting PLC curves that permit presence or absence of typical phases in the PLC such as the maturity phase.

Four elements in our approach are important. First, we use normalized product lifecycle data with clustering to operationalize "similarity" between products. This normalization allows us to look for similar patterns across items that may have different volumes and lifecycles. Clustering could be provided exogenously (e.g., by the company's product hierarchy) and/or could be automated or refined using a clustering algorithm as we propose and demonstrate. Second, we focus on static forecasting of the entire PLC just before the product launch. The key reason for and advantage of forecasting the entire PLC curve is that operations must plan capacity, sourcing, production, transportation, and inventory before product launch. Given the leadtimes involved (e.g., transportation leadtimes from China to North America by ocean is 8 weeks at the computer company we study), long range (8 week) forecasts are required even before the product is launched. Utilizing a PLC curve meets that challenge. Third, our approach considers both robustness and effectiveness of several families of PLC curves. Robustness takes into account both the goodness-of-fit and the complexity of the curves and is measured by the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). Forecast accuracy is our measure of effectiveness. It is well known that forecast accuracy directly drives safety inventory and capacity requirements, but it also significantly impacts sourcing and transportation decisions, which leads us to the final element in our approach. Fourth, we apply our approach on a large set of actual customer order data (133 products) for Dell

and find that piecewise-linear curves have significantly lower fitting errors than smooth curves (Bass (Bass 1969) and polynomial) with comparable number of parameters. Moreover, our approach reduces absolute errors by 9% relative to Dell's forecasts. According to an internal study conducted at Dell during the time period we study, an improvement in forecast accuracy of the magnitude we report would result in transportation and inventory expense savings of $2-$6 per unit on annual volumes in the millions.

3.2. Literature review

Half a century ago, Levitt (1965) wrote that "most alert and thoughtful senior marketing executives are by now familiar with the concept of the product life cycle." His critical review of strengths and weaknesses, including the importance of forecasting its shape to put the concept to work, remains surprisingly relevant. According to Rink and Swan (1979), the idea of the PLC was introduced in 1950 yet 30 years later there remained a paucity of empirical evidence. Golder and Tellis (2004) and Stark (2015) provide contemporary overviews of the vast field of PLC theory and management. The remainder of this section focuses on how our work contributes to selective relevant strands of literature on product life cycle forecasting.

3.2.1. PLC curves: theory

The diffusion model introduced by Bass (1969) remains a cornerstone in PLC theory. Its ensuing differential equation can be solved explicitly (Lemma 1 in Kumar and Swaminathan (2003) and reproduced later); we will refer to its particular bell-shape as the Bass curve. It should be noted that the classic Bass model has been extended or modified

in several dimensions. In terms or market and company structure, (Tigert and Farivar 1981) account for whether or not there is a monopoly and whether or not the company is public. From an operational perspective, for example, Ho et al. (2002) and Kumar and Swaminathan (2003) include supply availability and constraints. Niu (2006) proposes a stochastic Bass model and allows its parameters to vary over time periods.

3.2.2. PLC curves: empirical studies

In addition to theory-inspired Bass curves, authors have fitted other curves to empirical data. One popular family of curves are polynomials, which have been validated with demand data for recreation programs (Crompton 1979), municipal library services (Crompton and Bonk 1978) and grocery products (Headen 1966). Another popular family are piecewise-linear curves, as suggested by demand data for ethical drugs (Cox 1967), food (Buzzell and Nourse 1967) and chemicals (Frederixon 1969). Our empirical analysis will include Bass curves, as well as polynomial and piecewise-linear curves.

3.2.3. New Product and PLC Forecasting

Goldman (1982) states that many high-tech companies frequently face a situation of a long lead time and a short PLC. Such a situation is most demanding managerially. It is noteworthy to point out that classical time series models require the knowledge of some realized demands or sales to generate their forecasts, which is problematic in this situation. To help that challenge, our paper analyzes how to forecast entire PLC curves using historical data of similar previous products. Earlier, Fisher and Raman (1996) demonstrated the importance of initial forecast (not necessarily of the entire PLC curve)

quality and of forecast updating and responsive fulfillment. Gallien et al. (2015) analyze flexibility in replenishment for new product launches where initial forecasts are updated (for Zara). Another approach to address the challenge is a product portfolio approach where some product demands may serve as leading indicators for a group of products Wu et al. (2006).

3.2.4. Integrated PLC Forecasting and Operational Planning and Execution

Hayes and Wheelwright (1979) advocate how to link the manufacturing process and PLC, and a large literature has coupled forecasting with operational planning and execution. For example, Fisher and Raman (1996), Kurawarwala and Matsuo (1996), Zhu and Thonemann (2004) and Gallien et al. (2015) analyze joint forecasting and inventory decisions. As do we, Kurawarwala and Matsuo (1996) consider computers and forecast PLC curves, but they focus on, and in-sample test, only four make-to-order products using only Bass curves. As we describe further below, our focus is on make-to-stock products at Dell, where higher forecast accuracy would reduce expedited air transportation in favor of ocean transportation. For this situation, integrated forecasting and dual sourcing models are desired; Boute and Van Mieghem (2014) may provide some initial ideas by coupling of exponential smoothing with dual sourcing. Forecasting over the PLC implies non-stationary demand, and Graves (1999) addresses (single-sourcing) inventory management for non-stationary demand. Our paper focuses on PLC forecasting; future work will integrate with operations planning execution.

3.3. Context and Business Environment at Dell

Our industrial partner, Dell, is the third largest producer of personal computers globally. Historically, Dell has fulfilled personal computer demand using a configure-to-order (CTO) approach. With such an approach, forecasts and inventory decisions must be made at the component level. Kapuscinski et al. (2004) provide an overview of Dell's CTO operations and discuss the challenges of forecasting and managing component inventory to support CTO fulfillment. In recent years, Dell made the strategic decision to shift to fulfilling significant volume with a make-to-stock (MTS) model. Products selected to be managed MTS span multiple product categories (e.g., laptop, fixed workstation) and multiple target markets (e.g., business and consumer). The intent is to select products to be managed MTS where customers may value a simplified ordering process and fast delivery over the ability to customize their product. Some MTS products are available on Dell's website under a program called Smart Selection with the stated aim to provide "a simplified ordering process for our best value, prebuilt systems custom-designed based on customer feedback."[1]

Our data set, described in further detail below, is for North America only although Dell uses this MTS approach globally. For North America, the most cost effective product flow for Dell is to have their contract manufacturing partner in China produce and ship products via ocean—with an 8 week lead time—into fulfillment centers in the United States. For laptops, air freight from China can be used for faster delivery, but at substantially higher expense. Desktops can be delivered to U.S. fulfillment centers more quickly by having them produced in Mexico rather than China, but this results in substantially

[1] `http://www.dell.com/learn/us/en/04/campaigns/smart-select-consumer?c=us&l=en&s=dhs` accessed on Sept. 10, 2016.

higher manufacturing cost. In both these cases, the additional transportation or manufacturing cost associated with faster delivery may make adopting the MTS approach financially unattractive; thus, generating accurate forecasts is critical.

The forecasting process for a new MTS product starts with a product team providing a lifetime quantity forecast and a projected lifecycle length. Using these inputs, a demand planner creates a weekly forecast. The planner may include adjustments for known seasonal effects, planned promotions or sales initiatives and the potential impact of the introduction of other new products that may cannibalize demand. Planners may also examine orders for similar products from the past in creating the weekly forecast. We expect promotional effects to be somewhat limited in our data set for two reasons. First, the majority of the products in our data set are aimed at business customers where promotions are limited. Second, for products targeted at consumers, promotions can occur, but these plans may not be known at the time of product launch in which case they would not affect the *initial* forecasts made, and it is these initial forecast we use for comparison. Since we do not have information regarding promotions and cannibalization, we do not incorporate these effects into our PLC fitting and forecasting process.

3.4. Data

Our data set, which will be available online, includes weekly North American customer orders and forecast data for 52 complete product lifecycles from November 2013 until June 2014. For these 52 products, we do not have product category information. Our data set also includes 81 complete product lifecycles from April 2014 until January 2016. For these 81 products, we do not have forecasts but we do have product category information. We

note that the older data set is over a smaller period of time (8 months) than the newer data set (21 months). As such, the products in the older data set for which we have complete lifecycles will be biased towards short lifecycle products and product launches between November and January. We first provide a summary of this data, discuss its limitations, and then describe how we clean and prepare it for analysis. Cleaning and preparation of the data is necessary so that the PLC curves are normalized: this allows PLC curves to be compared to each other regardless of total volume or length of the PLC.

3.4.1. Overview

These 133 products belong to one of four product categories—laptops, desktops, mobile workstations, and fixed workstations—and are all managed with a MTS model. Dell holds these items in its fulfillment centers to serve individual consumers, institutions, and retailers. There is one exception: this dataset also includes some bulk custom orders requested by large organizations. Often, these very large orders are not filled from stock but rather added to the production schedule at a long lead time to the customer in a build-to-plan workstream. We try to filter these large orders out (see discussion below in Section 3.4.3) as our intent is to forecast MTS customer orders satisfied from Dell's fulfillment centers. We share summaries of the data's volume and launch distribution in Tables 3.1 and 3.2 respectively.

For 81 of the 133 products, we also know the company-defined category. In the forecasting process we will ignore any product category information; instead, we will cluster products by the shapes of their PLCs only as opposed to company defined attributes. We will, however, report on each cluster's breakdown of product category types.

Metric	25th percentile	Median	75th percentile
Weekly orders of slowest 20% products	8	29	52
Weekly orders of product with median volume	51	98	148
Weekly orders of fastest 20% products	312	575	883
Number of weeks of PLC	18	30	44

Table 3.1. Summary statistics of the data. (For 'weekly' net customer orders, the 25th, 50th, and 75th percentiles are over the observations for that product across the periods in its own lifecycle.)

Month	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Number launched	27	21	3	5	1	20	0	2	20	8	20	6

Table 3.2. Distribution of number of new products' launches across different calendar months.

For 52 of the 133 products we also have Dell's complete PLC forecasts, updated weekly. That is, at week 0 (time of launch) we have the point customer order forecasts for week $1, 2, \ldots$ until the last week. Additionally, for any week t we have the forecasts for weeks $t+1, t+2, \ldots$ until the predicted end of the lifecycle.

Above, we mentioned that we assume that demand planners know the lifetime quantity, the launch date, and the product lifecycle length. Our approach relies on these assumptions because it essentially provides merely the shape of the PLC with cumulative orders normalized to 1 and a normalized lifecycle length of 1. Scaling this curve to the true length of the PLC and to the true volume are tasks we do not attempt to tackle based on curve shape or category alone: as outlined in Section 3.3, we propose that to implement our approach, one would work with a demand planning team to estimate these two values. The launch date and length of PLC are also important because of seasonality. Once we can identify a normalized PLC shape for a product, we can easily create a

forecast for actual weekly customer orders by scaling this shape for lifetime volume and lifetime length and adjusting for seasonality.

How accurate are these two assumptions? We validate the knowledge of total volume and length of PLC in Figure 3.2. These are derived from the weekly forecasts—all made in week zero—over the lifetime for the 31 products for which we had forecast data and which were not removed in the data preparation stage outlined in section 3.4.3. In general, the company has very good estimates of total lifetime volume (bottom subfigure), either because it only sells all units produced (which is the forecast) or because it has very good lifetime forecasts. The company does not do as well predicting lifecycle length based on the data we have access to; in general it predicts lifecycle length to be longer than the actual length. While not ideal, there seems to be some consistency in the overestimate suggesting that there might be room to improve the forecasts. Nevertheless, when we compare our forecasts to the company's forecasts in Section 3.7, we use the company's imperfect estimates of lifetime quantity and lifecycle length as inputs into our forecasts in order to enable a fair comparison.

3.4.2. Limitations

The data we have is the same data available to Dell's demand planners. We note here some limitations.

(1) Net customer orders only: For each week, we observe only the net customer orders. First of all, this is the sum of total orders placed minus returns and cancellations in that week. We do not know the true total customer orders in a given week, nor do we know the breakdown (whether it was one big order or lots of

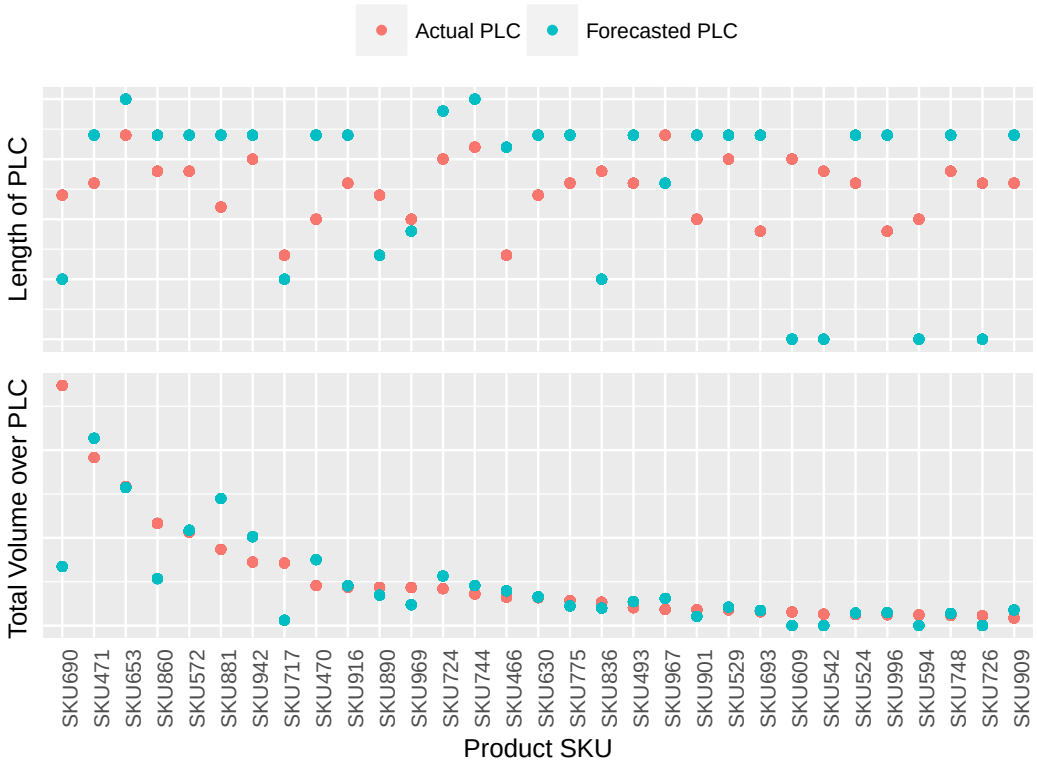


Figure 3.2. Actual versus company forecasted values of lifecycle length (top) and total volume (bottom) for 31 products. Company estimates of lifecycle length are often longer than the true length, while the volume forecasts appear to be very close to actuals. The vertical axis is disguised. There are four products for which we had no forecasts, and thus the forecasted volumes and PLC lengths are zero. There are only 31 products because we have company forecasts for only 52 products, and of these 52 we eliminate 21 in the data preparation stage described in Section 3.4.3.

small orders). Additionally, if a large order is placed in one week and returned the next week, it may lead to negative net customer orders observed in the following week. We discuss how we treat this below in Section 3.4.3.

(2) Censored demand: We observe only customer orders, and we do not have access to the inventory information. From the customer point of view, if an item is not in a fulfillment center she will not see the item as explicitly out of stock. Rather

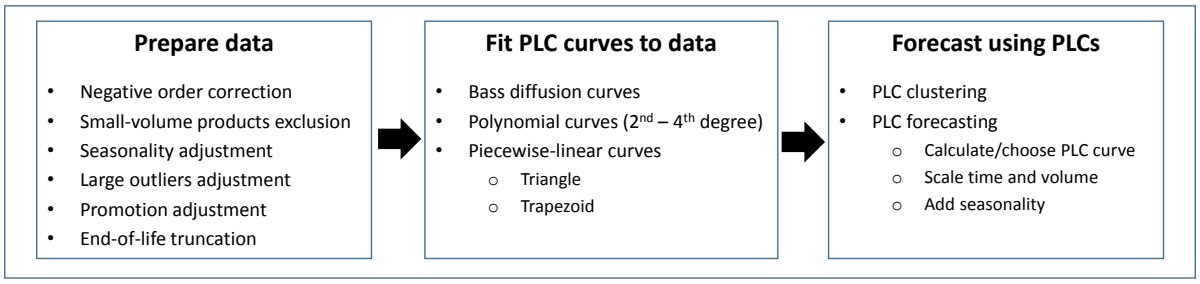


Figure 3.3. Forecasting using PLC curves requires several steps, including data preparation, curve fitting, and forecasting itself.

the lead time will be longer for items which are not in the fulfillment center. The customer may or may not decide to continue placing her order. Due to data limitations, we have no choice but to ignore potential censored demand at this time.

3.4.3. Preparation

For each product, we have access to only the total net customer orders for each week of its lifecycle. Thus, before forecasting the customer orders, it is necessary to prepare the historical raw data to address the phenomena of negative orders, very small volume items, customer order seasonality, and managed end-of-life behavior. Figure 3.3 summarizes our overall approach including treatment of the raw data. As noted above, since we do not have promotional information, we do not include that step in data preparation for this Dell data set. We retain that step in the figure as it may be an important step in application of our approach to another data set. We met with a demand planner at the partner company who informed us as to the root cause of the phenomena and helped guide us as to the proper treatment of these phenomena.

Detecting large cancellations. In the raw data, we observe several large negative net customer orders which occur in the middle of the product life cycle. Let D_t^i be the raw observed net customer order value in week t for product i. T_i is the length of the lifecycle (in weeks) of product i, and $t \in \{1, \ldots, T_i\}$ is the product specific week relative to that product's launch week. The total observed customer orders for a product in week t is the sum of all the actual customer orders minus all the order cancellations and returns for that week. Returns are relatively rare, and the large negative data points suggest that a large order which was placed in an earlier week is being cancelled. Based on conversations with a demand planner at Dell, we define a net customer order value of D_t^i at week t to be a *large negative order* if 1) the customer order value is negative and the week t is between the first and last positive order values for product i; and 2) the difference between D_t^i and D_{t-1}^i is larger than the average differences of neighboring customer order values. Based on the definition above, we identify 4 negative net customer order values from 4 products among 133 products as shown in Figure 3.4. We 'correct' negative customer order values by making two adjustments: we reduce the customer order value in the week in which the large order occurred and we increase the order value in the week in which the negative order occurred. We achieve this by averaging the two values: $D_t^{C[1],i} = D_{t-1}^{C[1],i} = \frac{D_t^i + D_{t-1}^i}{2}$, where $C[1]$ denotes cleaning step number 1.

Excluding short lifecycle and low volume products. Some products in our dataset have fairly short product life cycles on the order of the slower lead time of 8 weeks. Because we consider a lead time of 8 weeks from China to the United States, we exclude products whose product life cycle is no more than one and a half times this lead time. Additionally, we focus on medium- to high-volume products for three reasons: 1) this set of

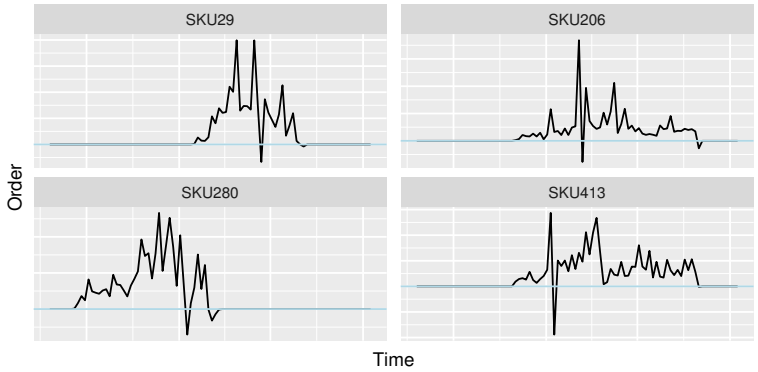


Figure 3.4. The 4 products with negative orders we correct. Note that for products SKU029, SKU206, and SKU413, these negative orders can clearly (visually) be matched to abnormally high order weeks just before.

products as a whole represents higher-volume products the company presumably chooses to manage make-to-stock and low-volume products are likely products that in hindsight should not have been selected for MTS; 2) low-volume products may be better represented by forecasting approaches developed for discrete distributions when integrality matters; and 3), the lower-tail of low-volume products does not contribute much to the overall revenue. Specifically, we remove products that fit one or both of the following criteria: 1) product life cycle is less than 12 weeks long; 2) average weekly volume is fewer than 20 units. Of the 133 products total, we retain 97 products, 31 in the older dataset with forecasts and 66 in the newer dataset without forecasts but with category information. Thus, we retain 73% of the products overall, and 95% percent of the total customer order volume of the original 133 products. Although for this processing step no customer order values are adjusted within a product, we define the new customer order observations as $D_t^{C[2],i} = D_t^{C[1],i}$ for all i such that i is not excluded by the criteria mentioned.

Adjusting for seasonality and normalization.

Products in different categories for different markets have different seasonality patterns. For instance, different industrial customers and governmental organizations have different fiscal years; thus the phenomenon of 'end-of-fiscal-year-buying' will occur in different (but perhaps fixed) months throughout the year. Individual customer orders may be driven by holiday gift-giving or 'back-to-school' shopping. Industrial and governmental organizations may purchase more workstations and fewer entertainment-focused laptops as compared to individual customers. We do not have access to these effects, and in our analysis, we do not attempt to adjust for seasonality for two reasons. First, we have no individual product lifecycle that covers two complete years, so we cannot estimate any seasonal effects at the product level. Second, since we do not know which products are aimed at which market, we do not know which subsets of products to group together to try to estimate seasonal effects.

If one had enough data to group products together by market (namely, by seasonal buying patterns of primary customers of each product), we suggest one approach here to adjust for seasonality:

(1) Normalize data so that cumulative volume of each product equals 1. In this way, the seasonal effect will not be disproportionally affected by high volume products.

(2) For each group, apply an additive or multiplicative seasonal effect model to the normalized data. This model would estimate customer orders based on the following independent variables:

   (a) Month effect (the seasonality to be estimated)

   (b) A generalized group-wide PLC that the model would estimate

   (c) Fixed effects for year and other group attributes

As a robustness check, we applied the above multiplicative seasonal effects model. However, due to the different seasonal patterns of different products, the forecast quality of the seasonal model we implemented was significantly worse than the same approach ignoring seasonality.

Hence, the seasonality-adjusted data is then $D_t^{C[4],i} = f_{Season}^{-1}\left(D_t^{C[3],i}\right)$, where $f_{Season}^{-1}(\cdot)$ denotes the deseasonalization function derived from the seasonal effects model. For the reasons mentioned above, we set $f_{Season}^{-1}(\cdot)$ to be the identity function; that is, $f_{Season}^{-1}(x) = x$ and $f_{Season}(x) = x$.

Excluding build-to-order customer orders. We focus on forecasting PLCs for MTS products whose orders are filled from on-hand inventory when there is not a stockout. However, if an individual customer order is large enough, it will be moved from the MTS workstream to the build-to-order workstream, thereby incurring a longer lead time which includes production and transportation. Because these orders do not draw on on-hand inventory, we want to remove them. We do not know exactly which portion of a week's customer orders were due to these very large orders. As a proxy, we identify weeks with very large customer order totals using outlier detection, assuming that this outlier is actually mostly made up of a very large order with a different workstream. We replace these outliers with 'reasonable' values (defined below) because once the build-to-order workstream units are removed, there are still likely underlying MTS customer orders. We identify outliers by the time series outlier detection method described in Chen and Liu (1993). In essence, the method outlined in that paper fits a time series model to the data and then identifies outliers significantly deviating from this time series model. Only nine orders are identified as outliers (Figure 3.5) across all the weeks of data of the 133

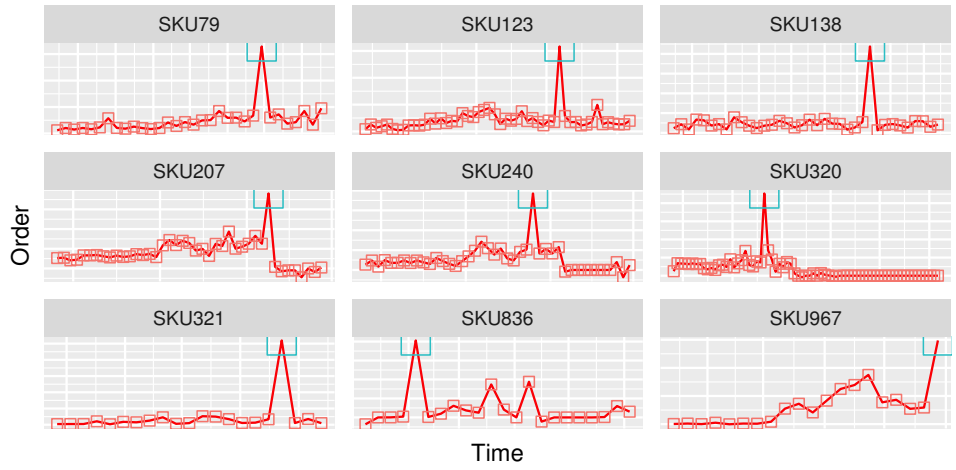


Figure 3.5. The nine products with large values (presumably directed to another workstream). Large values are denoted by blue squares.

products in the dataset. We replace the detected outliers for product i at week t with their weighted moving averages, as outlined in Roberts (2000). Recalling that T_i denotes PLC length of product i, we set:

$$D_t^{C[5],i} = \frac{D_1^{C[4],i} + 2D_2^{C[4],i} + \cdots + (t-1)D_{t-1}^{C[4],i} + (T_i - t)D_{t+1}^{C[4],i} + \cdots + D_{T_i}^{C[4],i}}{(1 + \cdots + t - 1) + (1 + \cdots + (T_i - t))}. \tag{3.1}$$

End-of-life truncation. Customer orders near the end of the lifecycle can be strongly influenced by managerial decisions such as promotions or timing of the introduction of a new product intended to replace an old one. Since we seek to focus on forecasting the "naturally occuring" product lifecycle, rather than orders that occur to an actively managed end-of-life, we exclude customer orders near the end of the lifecycle. In addition to managerial decisions affecting end-of-life orders, the end-of-life is 'far away' at time 0 (launch) and thus there is less value to forecasting it well versus forecasting the nearer term

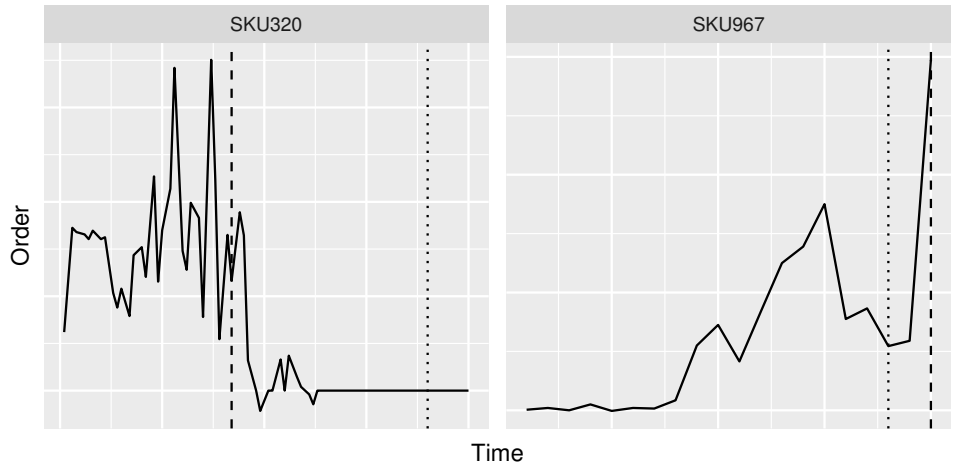
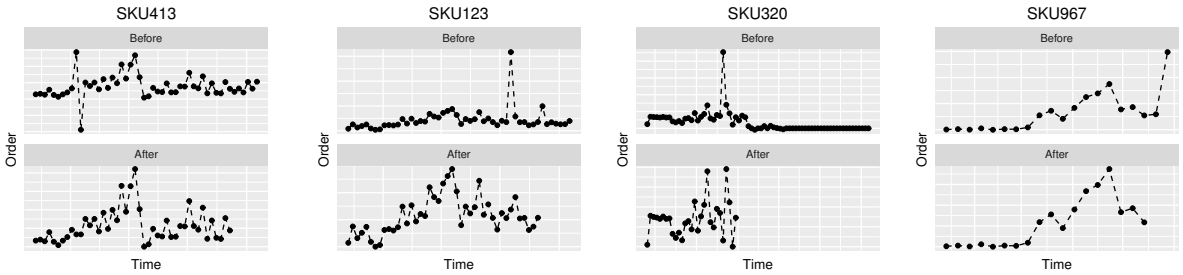


Figure 3.6. Illustration of our end-of-life truncation method. Dotted line: cut-off based on PLC length; Dashed line: cut-off based on volume. Our method would cut off all data points that occurred after the *earliest* cut-off point from either method.

majority of the PLC. In our data set, the length of the product lifecycle can be artificially extended past when a product's life has essentially ended, since a single customer order may occur or a return or cancellation may be made weeks later. Figure 3.6 shows the third behavior (artificially extended PLC) on the left and the first behavior (managed end-of-life promotion) on the right.

We exclude the last $1 - \theta^{p_t}$ of the weeks of the products' PLCs and we exclude the last $1 - \theta^{p_v}$ of the total volumes of the products. We initially set $1 - \theta^{p_t} = 1 - \theta^{p_t} = 0.9$. The processed data is now defined as $D_t^{C[6],i} = D_t^{C[5],i}$ for t satisfying the above criteria. We note that even though we have a short lifecycle cutoff of 12 weeks, by truncating the end of life of products some resulting lifecycles in our dataset may be less than 12 weeks. T_i is redefined to be a smaller value as appropriate to account for the cut off data points.



(a) Large cancellation correction

(b) Build-to-order order removed

(c) End-of-life correction

(d) End of life sales spike

Figure 3.7. Example of four products' actual customer orders over time (top) and normalized orders after data preparation (bottom). Data preparation can help the data more accurately reflect the reality of the problem we are trying to solve.

Normalization of data After data preparation, we re-normalize the data so that for each product, the lifetime cumulative sum of customer orders is equal to 1. We did this in the seasonality adjustment, but the data needs to be renormalized due to other processing. Thus, we obtain the customer order series $\tilde{D}$, with $\tilde{D}_t^i = D_t^{C[6],i} / D^{C[6],i}$. The lack of a t subscript denotes summation: $D^i \equiv \sum_{t=1}^{T_i} D_t^i$. Thus, $\sum_t \tilde{D}_t^i = 1 \quad \forall i$. All the following analyses are carried out based on this normalized data series $\tilde{D}$. In this way, PLCs for products with dramatically different volumes or launch seasons can be compared with each other. Figure 3.7 shows two examples of pre- and post-cleaned data.

3.5. PLC Curves Fitting

Having cleaned and normalized customer order data, we can proceed with fitting lifecycle curves to these data for each product. Previous literature has suggested the following three families of curves: the n^{th}-order polynomial curve (*polyn*), the BASS diffusion curves (*BASS*), and the piecewise-linear 'curve' (*triangle* and *trapezoid*). We

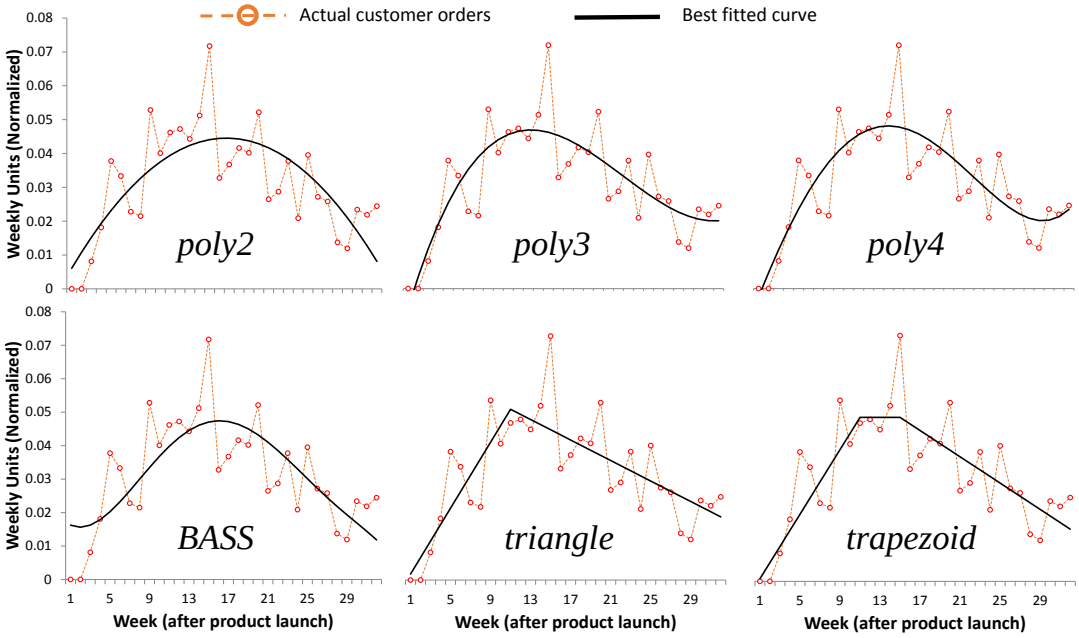


Figure 3.8. Six PLC curves fit to one product (JNKH1). A second order polynomial and the BASS curve overestimate demand in first few weeks. The fourth order polynomial might be overfitting the last few weeks of the lifecycle. While the trapezoid curve allows for a sustain phase, it is very short and visually it is difficult to identify that a clear sustain phase even exists (from the firm's point of view). Visually, the third order polynomial and triangle seem to provide 'good' fits in this example.

test and compare models from all three families, comparing both fitting accuracy and model complexity. Figure 3.8 shows one example product (JNKH1, the same product in Figure 3.1) fit by each of these curves, with polynomials of order 2, 3, and 4.

3.5.1. Overview of PLC category families

The BASS diffusion model uses three parameters (p, q, m). The parameter m represents lifetime volume, which we force to be 1. The resulting simplified BASS model for instantaneous rate of customer order total $\eta(t)$ is $\eta(t) = p + (q - p)N_t - q(N_t)^2$, where

$N_t = \sum_{s=0}^{t} \eta(s)$ is the cumulative sum of customer orders up to time t, and p and q are shape parameters (Kumar and Swaminathan 2003). Thus,

$$\breve{D}_t^{BASS} = \frac{p(p+q)^2 \exp(-(p+q)t)}{(p+q\exp(-(p+q)t))^2} \tag{3.2}$$

The $\breve{\cdot}$ notation denotes a customer order estimate provided by a particular PLC shape whose cumulative volume is 1 and whose lifecycle length is T_i.

The family of piecewise-linear curves fit the PLC with connected straight line segments. We explore two types of 'curves' in this family: the triangle (using two connected line segments) and the trapezoid (using three connected line segments with the middle segment forced to be flat). The triangle is suggested by Goldman (1982) and the trapezoid allows identification of maturity or "sustain" phase. The below piecewise-linear functions are valid for any lifetime sum of customer orders. When we force lifetime sum of customer orders to be 1, in terms of model complexity, the triangle and trapezoid functions will have 1 less free parameter each.

The triangle PLC requires four parameters (a, b, c, τ), and is defined as such:

$$\breve{D}_t^{triangle} = \begin{cases} at + b & 0 < t < \tau \\ c(t-\tau) + (a\tau + b) & \tau \leq t \leq T_i \end{cases} \tag{3.3}$$

where $\breve{D}_t^{triangle}$ is the order rate at time t, τ marks the period of transition and (a, b, c) are shape parameters.

The trapezoid is defined by five parameters $(a, b, c, \tau_1, \tau_2)$ which characterize the PLC as such:

$$\breve{D}_t^{trapezoid} = \begin{cases} at + b & 0 < t < \tau_1 \\ a\tau_1 + b & \tau_1 \leq t < \tau_2 \\ c(t - \tau_2) + (a\tau_1 + b) & \tau_2 \leq t \leq T_i \end{cases} \tag{3.4}$$

where $\breve{D}_t^{trapezoid}$ is the customer order rate at time t, $\tau_1 and \tau_2$ mark the two transition periods and (a, b, c) are shape parameters: a and c are two slopes and $\breve{D}_t^{trapezoid}$ will equal $a\tau + b$ at transition time τ.

The family of polynomial curves capture the PLC with smooth curvature. According to the literature (Crompton and Bonk 1978, Crompton 1979, Headen 1966), polynomial functions up to fourth orders are sufficient to capture a wide range of PLC curves. The n^{th} degree polynomial PLC curve is:

$$\breve{D}_t^{poly-n} = \sum_{i=0}^{n} a_i t^i \tag{3.5}$$

where $\breve{D}_t^{poly-n}$ is the orders at time t, and a_i for $i = 0, \ldots, n$ are the shape parameters. In terms of model complexity, when lifetime volume is forced to equal 1, an n^{th} order polynomial will have n free parameters.

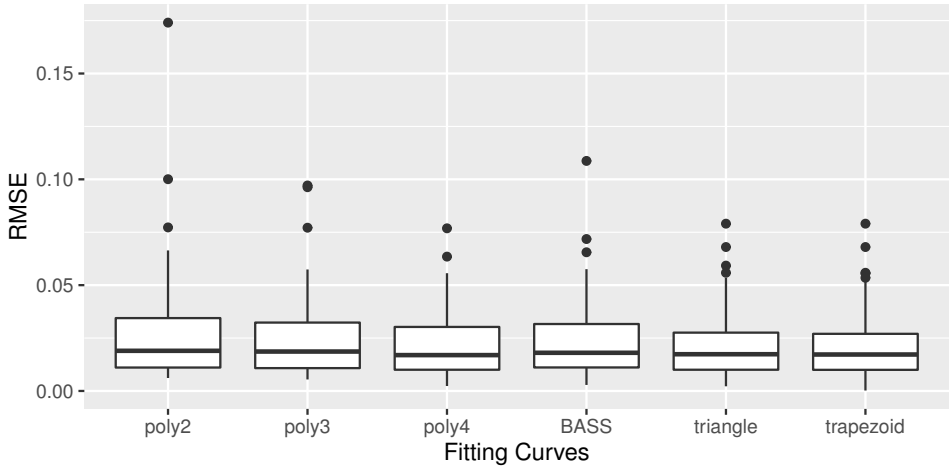
Variable	poly2	poly3	poly4	BASS	triangle	trapezoid
RMSE mean	0.0261	0.0238	0.0216	0.0236	0.0219	**0.0215**
RMSE stdev	0.0230	0.0180	**0.0150**	0.0171	0.0151	0.0150
RMSE 1st quantile	0.0111	0.0108	0.0100	0.0111	0.0100	**0.0100**
RMSE median	0.0190	0.0186	**0.0169**	0.0180	0.0174	0.0172
RMSE 3rd quantile	0.0344	0.0323	0.0302	0.0316	0.0276	**0.0270**
Loglikelihood	7,457	7,594	**7,731**	7,548	7,674	7,715
BIC	-13,914	-13,856	-13,796	**-14,096**	-14,014	-13,763
AIC	-14,331	-14,412	-14,493	-14,514	**-14,571**	-14,459
Number of parameters	3	4	5	3	4	5

Table 3.3. Summary statistics of PLCs' fits to the data. When adjusted for model complexity, the piecewise-linear curves fit the data the best (see AIC and BIC values). Values in **bold** denote they are the best in each row.

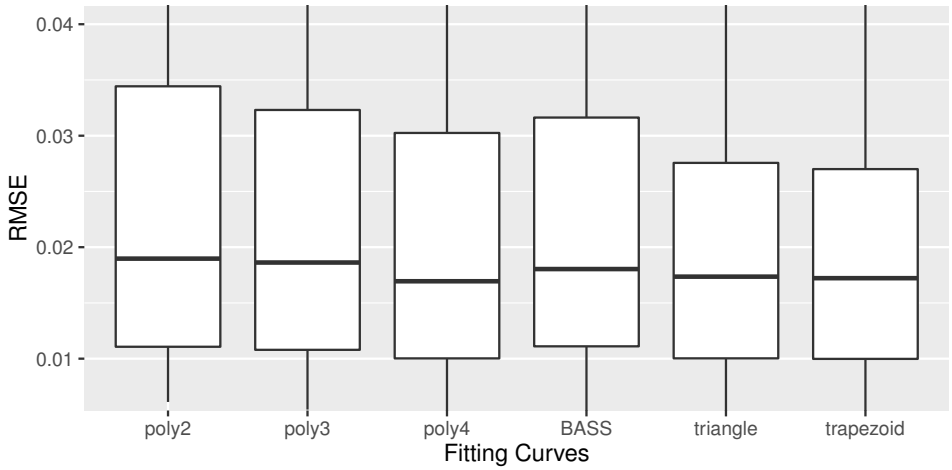
3.5.2. Quality of PLC fits

We fit all candidate PLC curves to each product's processed and normalized customer order data. The parameters for each curve are found through optimizing the root-mean-squared error (RMSE) of the candidate PLC applied to the customer order values across the weeks of the product's lifecycle. We utilized RMSE to not only fit the PLC curves, but also to evaluate the quality of the fit. In order to adjust the quality of fit for the model complexity (that is, the number of parameters) we report two measures intended to do just that: Akaike information criterion (AIC) and Bayesian information criterion (BIC). For AIC and BIC smaller values (in general, more negative) imply a better model. We summarize the performance in Table 3.3 and via boxplots in Figure 3.9.

We observe that smooth curves do not fit better than piecewise curves. In addition to triangle and trapezoid having the best AIC and BIC scores, of the 97 products, trapezoid is the best fit for 51, *poly*4 is the best fit for 38 and BASS is the best fit for 9. One explanation for a discontinuous derivative providing a better fit is that the business strategies and



(a) Full image with outliers.



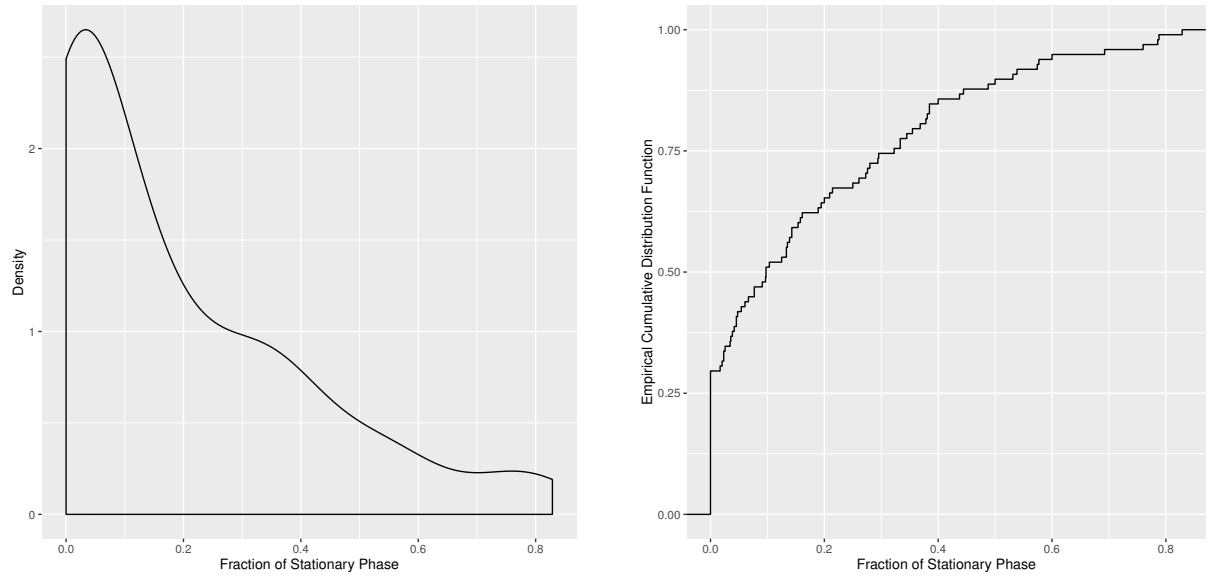
(b) Zoomed in (truncated vertical axis).

Figure 3.9. Distribution of RMSE of different curves' fits to the 97 products. *poly*4, *triangle*, and *trapezoid* appear to fit the data the best. *triangle*'s and *trapezoid*'s worst outliers are better (with respect to RMSE) than other families' worst outliers with the exception of *poly*4. The boxplot is drawn as such: the box shows the first, second (median), and third quantiles of the RMSE across the 97 products. The whiskers are 1.5 times the inter quartile range but will not extend beyond an actual observed value. Dots are outliers which extend beyond the whiskers.

marketing plans are changed at specific moments in time over the product life cycle. While a smoothed curve has to fit the customer order values from launch to near end-of-life, it lacks the flexibility to allow the curve to capture individual phases (and phase changes) separately. However, the piecewise-linear curves break the entire lifecycle into several distinct phases which can mimic the phase changes along the PLC. Regardless of the quality of fit of piecewise-linear curves, it may be advantageous to use them for practical reasons. Equally—if not more—important than fit, triangles and trapezoids are intuitive to explain and estimate. One needs only two slopes (growth and decline rates), and either one or two transition times.

Within the piecewise-linear curves, the trapezoid curve is slightly better than the triangle curves in terms of fitting performance but it requires one more parameter. The key difference between the two types of curves is that the trapezoid allows a flat sustaining phase in the middle. We now examine whether the sustain phase is significant within the PLC. In Figure 3.10, we observe that the fraction of the stationary phase relative to PLC length is small. Most products have some sustain phase but in general, half of products have sustain phases less than 10% of the entire lifecycle length and three quarters of products have sustain phases less than 30% of the entire product lifecycle length.

Thus, due to the high quality of fit of the triangle model, the good AIC score, and the fact that even with a trapezoid most products have very short sustain phases, we will model product lifecycles with triangles. The following analysis is based on using triangle curves.



(a) Empirical density

(b) Empirical cumulative distribution

Figure 3.10. Distribution of relative length of 'sustain' phase across 97 products. The relative length of a product's 'sustain' phase is calculated from its trapezoid PLC. It is the proportion of each product's PLC that is the flat middle line segment of the trapezoid. Note that three fourths of products have sustain phases significantly less than a third (on average) of their entire lifecycle lengths.

3.6. PLC Forecasting

Thus far, we have fitted to each product its own PLC curve. But this is useful only if we know the exact curve shape for each product prior to launch. As this may be unrealistic, we propose clustering similar PLCs together into several representative PLC shapes. Then we can generate PLC forecasts based on the cluster-level PLC curves and other available supplementary information. That is, if we can identify to which cluster a product belongs, we can use that cluster's representative PLC shape as a basis for the forecast of the product's customer orders. The true forecast is then based on this PLC shape, but adjusted for estimates of lifetime quantity, lifecycle length, and seasonality.

3.6.1. Clustering

Each PLC curve is in essence a time series and thus we can utilize time-series clustering as outlined in Chouakria and Nagabhushan (2007). This clustering method is based on proximity of both scale and behavior. The authors present a distance measure to address the proximity of values in two time series at the same point in time as well as the temporal correlation for behavior similarity. We outline the exact implementation of their ideas as well as the selection of number of clusters in the appendix. We note that their distance measure ($\delta_{CORT}(X_t, Y_t)$) is model-free: it allows us to cluster the fitted PLC curves based on their features in terms of temporal structure regardless of whether a polynomial, Bass, or piecewise-linear curve was used to model the PLC shape. We note that in order to cluster these PLC curves together, we must normalize the length of the PLCs to be the same (we choose 100 time periods). Thus a 30 week and 60 week PLC might be clustered together if they have the same shape, scaled by time.

Once a distance measure is established, we need to determine an appropriate number of clusters. We first plot 'sum of squared distances within clusters' versus number of clusters in Figure 3.11. Using our judgement, we choose 6 clusters because beyond 6, there is not much improvement in 'sum of squared distances.'

We summarize the overall attributes of each of the six clusters. First, we show the fitted triangle curves broken out by the six clusters in Figure 3.12. Tables 3.4, 3.5, and 3.6 show summary statistics and attributes of each cluster.

Figure 3.12 shows a few distinct patterns (and at least one outlier in the sixth cluster, and maybe a few anomalies in the fourth and fifth clusters worth further investigation by the firm). Cluster 1 tends to rise fast for more than half the lifecycle then trend down.

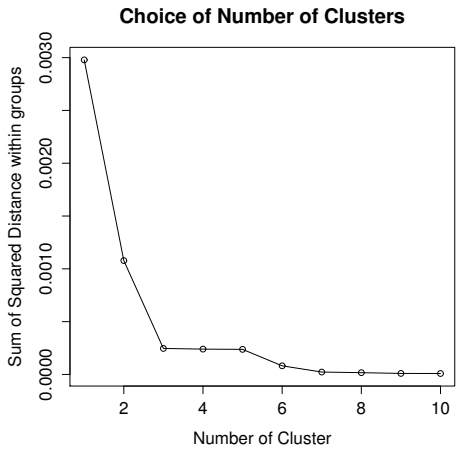


Figure 3.11. Sum of squared distances within clusters versus number of clusters. We chose six clusters because there is little reduction in sum of squared distances for values above 6.

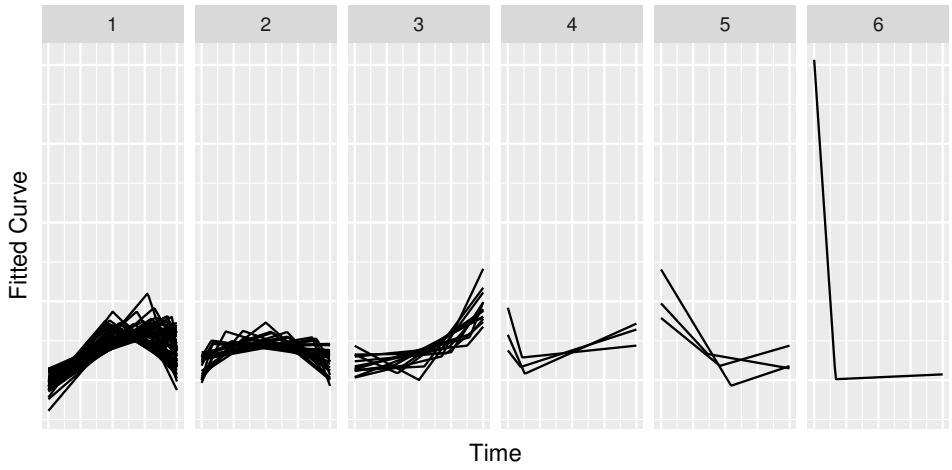


Figure 3.12. Each product's triangle curves broken out by cluster. Clustering clearly identifies products with similarly shaped curves, even selecting out an anomalous one by itself in cluster 6.

Cluster 2 tends to rise and fall at different times more gradually. Cluster 3 tends to be ramping up the entire time. Cluster 6 contains one product whose customer orders occurred almost all in the single week after Thanksgiving (the first week of its lifecycle).

Cluster group	Mean total volume (scaled)	Mean lifecycle length
1	1.00	32
2	0.34	24
3	0.29	25
4	0.13	20
5	0.57	16
6	0.36	7

Table 3.4. Breakdown of clusters by volume and lifecycle length. The mean scaled total volume for each cluster is proportional to the average volume per product within that cluster. We scale the raw volume means in order to disguise the data. Clusters 1 and 5 have the highest volumes while cluster 6 has the shortest lifecycles.

Cluster group	Product category					Fraction of total
	Fixed Workstation	Laptop	Mobile Workstation	Desktop	Unknown	
1	3	18	1	23	10	0.57
2	5	0	2	5	10	0.23
3	1	4	2	0	6	0.13
4	1	0	0	0	2	0.03
5	0	0	1	0	2	0.03
6	0	0	0	0	1	0.01

Table 3.5. Breakdown of clusters by product category (which were not actually used in the clustering process). Some category-cluster pairs that emerge are consumer laptops and desktops in cluster 1, workstation products in cluster 2, and laptops and unknown in cluster 3.

After that, the weekly volumes of customer orders were two orders of magnitude lower for the rest of the very short lifecycle.

Tables 3.4, 3.5, and 3.6 show that overall, products with differing attributes are spread across the clusters. This suggests that clustering may be able to identify hidden product attributes not represented in the raw data itself and possibly not known to demand planners. Thus, the mere act of clustering (without forecasting anything) may by itself be beneficial to firms in forecasting new (or almost new) products.

Cluster group	Jan-Mar	Apr-Jun	Jul-Sep	Oct-Dec
1	0.22	0.24	0.24	0.31
2	0.59	0.27	0.05	0.09
3	0.31	0.08	0.23	0.38
4	0.67	0.33	0.00	0.00
5	0.33	0.00	0.00	0.67
6	0.00	0.00	0.00	1.00

Table 3.6. Breakdown of clusters by launch month. Clusters 2 and 4 tend to have January to March launches while other clusters' launches are either spread out across the year or slightly concentrated in October to December.

3.6.2. Generating forecasts from clusters

In order to forecast the weekly orders of new products, we will use the information from the fitted curves of historical products as well as the company's knowledge of launch date as well as its estimates of lifetime volume ($\hat{D}^i$) and lifecycle length ($\hat{T}_i$), as mentioned in 3.4.1. Here, the $\hat{\cdot}$ notation denotes the company's estimates, not necessarily the true values. The exact steps we propose to move from PLC curve shape to actual forecast are as follows:

Step 1: **Obtain the fitted PLC curve for the ready-to-launch product.**

(a) If the company knows the fitted PLC curve, then the curve is given as

$$\bar{D}_t^{i,PLC,knownPLC}, \text{ for } t = 1, \cdots, T = 100 \tag{3.6}$$

where here, we utilize the trapezoid curve so that $PLC = trapezoid$. The $\bar{\cdot}$ notation denotes that we are working with time-normalized PLC curve. That is, $\breve{D}_t^{i,PLC}$ is defined over $t = 1, \ldots, T_i$, whereas $\bar{D}_t^{i,PLC}$ is defined over $t = 1, \ldots, 100$ for all products. For each, though, cumulative volumes are

normalized to 1 and other than the time-normalization $\breve{D}_{t_{act}}^{i,PLC} = \bar{D}_{t_{100}}^{i,PLC}$ for product i for the appropriately paired t_{act}'s and t_{100}'s.

(b) If the company knows that the product belongs to the kth cluster, then the fitted PLC curve is generated as

$$\bar{D}_t^{i,PLC,knownClust} = \frac{\sum_{i \in J_k} \bar{D}_t^{i,PLC,knownPLC}}{N_k}, \text{ for } t = 1, \cdots, T = 100$$

where N_k is the number of fitted curves in cluster k and J_k is the index of fitted curves in cluster k. Note that we also tested fitting a PLC to the clustered data which led to very similar results as taking the average of the PLCs.

(c) If the company knows only a prior probability p_k^i that the product i belongs to cluster k, then the fitted PLC curve is generated as

$$\bar{D}_t^{i,PLC,unknown} = \sum_k \frac{\sum_{i \in J_k} \bar{D}_t^{i,PLC,knownPLC}}{N_k} p_k^i, \text{ for } t = 1, \cdots, T = 100.$$

Unless otherwise stated, we assume that $p_k^i = p_k \ \ \forall i$ is the empirical probability/frequency of that cluster occurring.

Step 2: **Scale the time and add the seasonality effect $f_{Season}(\cdot)$ based on the information of launch week.** First, recall the PLC shape has 100 periods in the time-scaled version regardless of the true lifecycle length (this was necessary for clustering). We first scale this PLC shape to the actual lifecycle length. Thus, $\bar{D}_t^{i,PLC,\eta}$ for $t = 1, \ldots, 100$ is transformed into $\breve{D}_t^{i,PLC,\eta}$ for $t = 1, \ldots, \hat{T}_i$ by

scaling the lifecycle length by the *estimated* length of the PLC of $\hat{T}_i$ for product i. $\eta \in \{knownPLC, knownClust, unknown\}$ denotes the 'level of knowledge' we have about a particular product. Then, the fitted PLC curves are updated as:

$$\breve{D}_t^{i,PLC,\eta,Season} = f_{Season}\left(\breve{D}_t^{i,PLC,\eta}\right), \text{ for } t = 1, \cdots, \hat{T}. \tag{3.7}$$

Recall we do not adjust for seasonality in this analysis. Thus, we define $f_{Season}(\cdot)$ as the identity function (that is, $f_{Season}(x) = x$).

Step 3: Generate a forecast for product i for every period using the estimate of total volume $\hat{D}^i$. The forecast is then

$$\hat{D}_t^i = \hat{D}^i \frac{\breve{D}_t^{i,PLC,\eta,Season}}{\sum_t \breve{D}_t^{i,PLC,\eta,Season}}, \text{ for } t = 1, \cdots, \hat{T} \tag{3.8}$$

3.7. Forecast Evaluation

In this section, we compare the quality of the curve-based forecasts against how much information is known about the PLC curve of each product, and we also compare the curve-based forecasts with the company's own forecasts. The comparison metric is mean absolute scaled error (MASE), a measure of forecast accuracy proposed by Hyndman and Koehler (2006) and defined as

$$\text{MASE}(T_i') = \frac{1}{T_i'} \sum_{t=1}^{T_i'} \left(\frac{|\hat{D}_t^i - D_t^i|}{\frac{1}{T_i'-1} \sum_{t'=2}^{T_i'} |D_{t'}^i - D_{t'-1}^i|} \right) \tag{3.9}$$

MASE has the desirable properties that it is invariant to scale and it is not skewed when the data points are near 0. Mean absolute percentage error (MAPE), on the other hand, divides the period t error by period t demand and thus can be undefined (very large) when a period's demand is zero (near zero). Mean absolute error (MAE) is not scaled for volume or underlying variation, and so comparing MAE across products may have little meaning. We believe MASE to be a nice compromise between being scaled (to enable cross-product comparisons) and providing reasonable values even for products with periods of no demand. We write MASE as a function of T' because below we will calculate the MASE on differing fractions of each product's actual lifecycle length in weeks. We do this to understand how forecast quality changes with portion of lifecycle as a firm may place more importance on forecasting the first half of a product lifecycle than the entire lifecycle because the first half is more relevant to short term decisions.

First we show the forecasting accuracy within the sample. That is, we assume we know differing levels of information regarding the PLC curve of a given product, as outlined in Section 3.6.1: 1) the company knows the exact PLC curve; 2) the company knows the cluster to which the product belongs and we know the representative PLC curve of that overall cluster; 3) the company does not know the cluster to which the product belongs, but we know the prior probabilities of it belonging to each the different clusters.

For each level of knowledge about the PLC curve for a product, for each product, we measure the $MASE(T')$ (Equation (3.9)) for $T_i' \in \{0.5T_i, T_i\}$ where T_i is the actual length of the product lifecycle in weeks for product i. In Table 3.7 we present the summary values across the 97 SKUs in our sample.

We point to two insights derived from Table 3.7. The first insight mostly (but not entirely) confirms intuition: the PLC-based forecast performs better (MASE is smaller) when the exact PLC curve is known, but it is not clearly better (and sometimes worse) to know the cluster itself. We note that these percentiles are across the 97 products and MASE values are scaled by forecast variability. When we weight forecast errors by volume in Table 3.8 we see that cluster knowledge is actually very valuable. This suggests that the representative curve for a cluster is more representative for high volume products within that cluster, whereas for low volume products the company-wide curve may be a better choice.

The second insight is that the PLC-based forecast performs better for the 75th percentile of the 97 products for the entire PLC as opposed to the first half of the PLC. (Note there is the reverse behavior for the 25th percentile, but it is of smaller magnitude.) This suggests that for those products with the worst forecasts halfway through the PLC, the overall forecast will improve over the rest of the lifecycle under the assumption of known lifetime demand. Given that we assume we know the lifetime volume exactly for each product, this is not necessarily surprising. Oftentimes, a mismatch between estimated and true customer orders will be resolved near the end of the PLC due to this assumption of known-total-volume.

Secondly, we show the forecasting accuracy of our PLC-based model in a more realistic environment, comparing it with the company's own forecasting method. This can be performed only on 27 of the products for which we have non-zero forecasts (4 of the 31 products in the older dataset do not have forecasts). We aim to perform a fair comparison where each method has access to the same knowledge. Thus, we follow these steps:

Progress in PLC	Level of knowledge	25th Percentile	50th Percentile	75th Percentile
50%	Known PLC	0.67	0.83	1.03
50%	Known Cluster	0.75	0.88	1.16
50%	Unknown	0.74	0.95	1.20
100%	Known PLC	0.70	0.75	0.88
100%	Known Cluster	0.76	0.92	1.03
100%	Unknown	0.79	0.88	1.11

Table 3.7. Distribution of MASE of the 97 products broken out by progress in PLC (the fraction of the PLC's forecast quality being measured starting from day 1) and level of knowledge of the product's PLC shape. Knowing the exact PLC significantly improves the forecast quality.

(1) In week 0 (pre-launch), we have access to the company's entire lifetime forecast week by week (but made in week 0). We call these estimates $\hat{D}_t^{i,comp}$ for $t \in 1 \dots T^{i,comp}$. (*comp* is company and $T^{i,comp}$ is the company's estimate of the length of the product lifecycle.) Note these may not align with the realized volumes and lifecycle lengths.

(2) In week 0, we have generated a PLC curve for each product that is normalized by volume and time. This may be the PLC based on knowing the exact curve irself, knowing the cluster, or knowing only the cluster probabilities. We scale this normalized PLC curve by the *company's* estimates of lifetime volume and lifecycle length. Specifically, we scale the lifetime volume to be $\sum_t \hat{D}_t^{i,comp}$ and we scale the lifecycle length to be $T^{i,comp}$ weeks. Note that in this way, our method does not have extra knowledge about the product's total volume or lifetime length compared to the company, which is presumably making forecasts based on expert knowledge about previous products' customer order totals, observed lifecycles, etc.

Progress in PLC	Level of knowledge	25th Percentile	50th Percentile	75th Percentile
50%	Known PLC	0.68	0.96	1.21
50%	Known Cluster	0.73	1.05	1.29
50%	Unknown	0.78	1.04	1.45
50%	Company	0.82	1.10	1.78
100%	Known PLC	0.77	1.03	1.24
100%	Known Cluster	0.90	1.06	1.30
100%	Unknown	0.85	1.06	1.37
100%	Company	0.94	1.12	1.38

Table 3.8. Distribution of MASE for PLC forecasting versus the company's own week zero forecasts. MASE values of the 27 products are broken out by progress in PLC and level of knowledge of the product's PLC shape. The PLC forecasting method improves upon the company's forecasts, even when nothing is known ('unknown PLC') about each product's actual curve ('known PLC') or even peer products ('known cluster'). PLC and company forecasts both use imperfect day zero company forecasts of lifecycle length and lifetime volume.

(3) We compare the $MASE(T')$ for our PLC method (based on differening levels of knowledge about the PLC) and for the company's forecast. We do this for $T' \in \{0.5T, T\}$ where T is the true lifecycle length.

We present the results in Table 3.8. We see that the PLC-based forecast methods perform better than the company's own forecast, *even when we have no PLC-specific knowledge about the product* and use the (essentially) average product-wide PLC curve).

The metric MASE, while useful to compare forecast accuracy across products with differing volumes and levels of forecastability, does not include volume information. We also want to know the company-wide impact of a given forecasting method: if a method works better on a very high volume product we want to reward that method appropriately. Thus, we measure product-wide sum of absolute errors (SAE). This metric will measure all errors across all products, and to some extent reflects the weighted average of individual

Progress of PLC	Known PLC	Known Cluster	Unknown Cluster
50% of PLC	18.7%	12.7%	2.6%
100% of PLC	14.0%	9.2%	3.4%

Table 3.9. Summary of percent reduction in product-wide sum of absolute error (SAE) using PLC curves compared to using the company's forecasts. Percent reduction is measured relative to the company's SAE in the same 'progress within PLC' (50% versus 100%). Even when the cluster is not known, using the product-wide 'average' PLC ("Unknown cluster") improves the company's own forecast errors by 2%-3%. Knowing the PLC of similar products or the PLC itself leads to even more improvement. SAE is summed across all the non-normalized products: naturally products with higher volumes and higher forecast errors will contribute more to these values.

MAEs across products. We present results in Table 3.9 regarding percent reduction in SAE relative to the company's value for this metric. When we weight by volume, our method provides a significant improvement over the company's forecasts, on the order of 9% for known cluster. We also note – in contrast to Tables 3.7 and 3.8 – knowing the cluster significantly improves the product-wide forecast. Thus, we posit that forecasting by known cluster is more accurate for high volume products, which are exactly the products we want to forecast the best anyway.

3.8. Conclusion

In this paper, we address the problem of generating forecasts for new products that are similar to past products. To accomplish this, we fit several functional forms (Bass, piecewise linear and polynomial) to normalized product life cycle curves of historical data. Using complete product lifecycle order history for 133 products from Dell, we evaluate our data preparation and PLC fitting approach.

We find that simple, piecewise-linear curves are effective in fitting historical PLC curves. In particular, a simple triangle performed very well on our data. The triangle has the advantage that it is easy to explain and therefore easy to implement. In addition, we found that the products in our dataset (which had a median length of 30 weeks) had almost no "mature" phase of the PLC. While our approach is general and could be applied to other industries, the finding of a very short mature phase is of course specific to our dataset. Indeed, an opportunity for future research is to apply a PLC fitting approach such as ours to datasets from other industries.

We use the normalized PLC curves fit to historical data for forecasting by using time-series clustering techniques to cluster similar PLC curves and find representative curves for these clusters. A modification to our approach would be to use information provided by the company for clusters. This may work particularly well when a new product is the next version of a very similar past product. Since our approach uses normalized curves, we must scale the appropriate normalized curve for a new product by a lifetime quantity forecast. This means the performance of our method is dependent on the quality of these total lifetime forecasts.

For the subset of products where Dell forecasts are available, we quantify forecast accuracy improvements obtained by adopting our approach. To do this, we use lifetime quantity forecasts from Dell to scale our normalized curves, and we assume that we can use the representative curve from a product's cluster. (This approach is labeled *known cluster* throughout the paper.) We note that we did not have any seasonality, promotional or cannibalization information that may have been available to Dell. Our *known cluster* approach resulted in absolute errors 9% lower than Dell's historical forecasts. Given the

volumes involved, such an improvement in forecast accuracy would lead to very substantial savings.

Effective new product forecasting is critical for many companies, and many new products fall into the category of "similar" to past products; thus, our approach would be applicable. We hope our work, and the normalized data set that we make available, stimulates new research in this area.

References

Aguir, M Salah, O Zeynep Akşin, Fikri Karaesmen, Yves Dallery. 2008. On the interaction between retrials and sizing of call centers. *European Journal of Operational Research* **191**(2) 398–408.

Aguir, Salah, Fikri Karaesmen, O Zeynep Akşin, Fabrice Chauvet. 2004. The impact of retrials on call center performance. *OR Spectrum* **26**(3) 353–376.

Aissani, Amar. 1994. A retrial queue with redundancy and unreliable server. *Queueing systems* **17**(3-4) 431–449.

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

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

Anand, Krishnan S, M Fazil Pac, Senthil Veeraraghavan. 2011. Quality-speed conundrum: trade-offs in customer-intensive services. *Management Science* **57**(1) 40–56.

Armony, Mor, Erica Plambeck, Sridhar Seshadri. 2009. Sensitivity of optimal capacity to customer impatience in an unobservable m/m/s queue (why you shouldn't shout at the dmv). *Manufacturing and Service Operations Management* **11**(1) 19–32.

Artalejo, Jesus R. 1999. Accessible bibliography on retrial queues. *Mathematical and computer modelling* **30**(3) 1–6.

Artalejo, Jesus R. 2010. Accessible bibliography on retrial queues: Progress in 2000–2009. *Mathematical and computer modelling* **51**(9) 1071–1081.

Artalejo, Jesús R, MJ Lopez-Herrero. 2000. On the busy period of the m/g/1 retrial queue. *Naval Research Logistics (NRL)* **47**(2) 115–127.

Babel, Contact. 2014. The us contact center decision-makers' guide. Tech. rep., Contact Babel.

Bagnoli, Mark, Susan G Watts. 2010. Oligopoly, disclosure, and earnings management. *The Accounting Review* **85**(4) 1191–1214.

Bass, Frank M. 1969. A new product growth for model consumer durables. *Management Science* **15**(5) 215–227.

Becker, Bo, Todd Milbourn. 2011. How did increased competition affect credit ratings? *Journal of Financial Economics* **101**(3) 493–514.

Becker, Gary S. 1968. Crime and punishment: An economic approach. *The Journal of Political Economy* 169–217.

Bennett, Victor Manuel, Lamar Pierce, Jason A Snyder, Michael W Toffel. 2013. Customer-driven misconduct: How competition corrupts business practices. *Management Science* **59**(8) 1725–1742.

Berry, Steven, James Levinsohn, Ariel Pakes. 1995. Automobile prices in market equilibrium. *Econometrica: Journal of the Econometric Society* 841–890.

Bertrand, Olivier, Fabrice Lumineau. 2015. Partners in crime: The effects of diversity on the longevity of cartels. *Academy of Management Journal* .

Boute, Robert N, Jan A Van Mieghem. 2014. Global dual sourcing and order smoothing: The impact of capacity and lead times. *Management Science* **61**(9) 2080–2099.

Branco, Fernando, J Miguel Villas-Boas. 2015. Competitive vices. *Journal of Marketing Research* **52**(6) 801–816.

Bresnahan, Timothy F, Peter C Reiss. 1991. Entry and competition in concentrated markets. *Journal of Political Economy* 977–1009.

Buzzell, Robert Dow, Robert E Nourse. 1967. *Product innovation in food processing, 1954-1964*. Harvard University Press, New York.

Cai, Hongbin, Qiao Liu. 2009. Competition and corporate tax avoidance: Evidence from chinese industrial firms*. *The Economic Journal* **119**(537) 764–795.

Chen, Chialin. 2001. Design for the environment: a quality-based model for green product development. *Management Science* **47**(2) 250–263.

Chen, Chung, Lon-Mu Liu. 1993. Joint estimation of model parameters and outlier effects in time series. *Journal of the American Statistical Association* **88**(421) 284–297.

Chen, Yuche, Jens Borken-Kleefeld. 2014. Real-driving emissions from cars and light commercial vehicles–results from 13 years remote sensing at zurich/ch. *Atmospheric Environment* **88** 157–164.

Chouakria, Ahlame Douzal, Panduranga Naidu Nagabhushan. 2007. Adaptive dissimilarity index for measuring time series proximity. *Advances in Data Analysis and Classification* **1**(1) 5–21.

Colbeck, Ian, Mihalis Lazaridis. 2010. Aerosols and environmental pollution. *Naturwissenschaften* **97**(2) 117–131.

Coleman, James William. 1987. Toward an integrated theory of white-collar crime. *American Journal of Sociology* 406–439.

Cox, William E. 1967. Product life cycles as marketing models. *The Journal of Business* **40**(4) 375–384.

Crompton, John L. 1979. Recreation programs have life cycles, too. *Parks and Recreation* **Oct.** 52–57.

Crompton, John L, Sharon Bonk. 1978. An empirical investigation of the appropriateness of the product life cycle to municipal library services. *Journal of the Academy of Marketing Science* **6**(1-2) 77–90.

Cui, Shiliang, Xuanming Su, Senthil K Veeraraghavan. 2014. A model of rational retrials in queues. *Available at SSRN 2344510* .

Cummins, Jason G, Ingmar Nyman. 2005. The dark side of competitive pressure. *RAND Journal of Economics* 361–377.

De Véricourt, Francis, Yong-Pin Zhou. 2005. Managing response time in a call-routing problem with service failure. *Operations Research* **53**(6) 968–981.

Deloitte. 2014. Global automotive consumer study: the changing nature of mobility. Tech. rep.

Ding, Sihan, Ger Koole, Rob Van Der Mei. 2013. A method for estimation of redial and reconnect probabilities in call centers. *Proceedings of the 2013 winter simulation conference: Simulation: Making decisions in a complex world*. IEEE Press, 181–192.

Ding, Sihan, Maria Remerova, RD van der Mei, Bert Zwart. 2015. Fluid approximation of a call center model with redials and reconnects. *Performance Evaluation* **92** 24–39.

Dowell, Glen, Stuart Hart, Bernard Yeung. 2000. Do corporate global environmental standards create or destroy market value? *Management Science* **46**(8) 1059–1074.

Du, Xingqiang, Shaojuan Lai. 2015. Financial distress, investment opportunity, and the contagion effect of low audit quality: Evidence from china. *Journal of Business Ethics* 1–29.

Duflo, Esther, Michael Greenstone, Rohini Pande, Nicholas Ryan. 2014. The value of regulatory discretion: Estimates from environmental inspections in india. Tech. rep., National Bureau of Economic Research.

Elcan, Amie. 1994. Optimal customer return rate for an m/m/1 queueing system with retrials. *Probability in the Engineering and Informational Sciences* **8**(04) 521–539.

Falin, Gennadi. 1995. Estimation of retrial rate in a retrial queue. *Queueing systems* **19**(3) 231–246.

Falin, Gennadi, James GC Templeton. 1997. *Retrial queues*, vol. 75. CRC Press.

Falin, Gennadij. 1990. A survey of retrial queues. *Queueing systems* **7**(2) 127–167.

Fisher, Marshall, Ananth Raman. 1996. Reducing the cost of demand uncertainty through accurate response to early sales. *Operations Research* **44**(1) 87–99.

Franco, Vicente, Francisco Posada Sánchez, John German, Peter Mock. 2014. Real-world exhaust emissions from modern diesel cars. *communications* **49**(30) 847129–102.

Frederixon, Martin Shelton. 1969. An investigation of the product life cycle concept and its application to new product proposal evaluation within the chemical industry. Ph.D. thesis, Michigan State University.

Gallien, Jérémie, Adam J Mersereau, Andres Garro, Alberte Dapena Mora, Martín Nóvoa Vidal. 2015. Initial shipment decisions for new products at Zara. *Operations Research* **63**(2) 269–286.

Gandhi, HS, GWf Graham, R_W McCabe. 2003. Automotive exhaust catalysis. *Journal of Catalysis* **216**(1) 433–442.

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

Goldberg, Pinelopi Koujianou. 1998. The effects of the corporate average fuel efficiency standards in the us. *The Journal of Industrial Economics* **46**(1) 1–33.

Golder, Peter N, Gerard J Tellis. 2004. Growing, growing, gone: Cascades, diffusion, and turning points in the product life cycle. *Marketing Science* **23**(2) 207–218.

Goldman, Arieh. 1982. Short product life cycles: implications for the marketing activities of small high-technology companies. *R&D Management* **12**(2) 81–90.

Graves, Stephen C. 1999. A single-item inventory model for a nonstationary demand process. *Manufacturing & Service Operations Management* **1**(1) 50–61.

Hart, Oliver D. 1983. The market mechanism as an incentive scheme. *The Bell Journal of Economics* 366–382.

Hassin, Refael, Moshe Haviv. 1996. On optimal and equilibrium retrial rates in a queueing system. *Probability in the Engineering and Informational Sciences* **10** 223–228.

Hayes, Robert H, Steven C Wheelwright. 1979. Link manufacturing process and product life cycles. *Harvard Business Review* **57**(1) 133–140.

Headen, Robert Speir. 1966. *The Introductory Phases of the Life Cycle for New Grocery Products: Consumer Acceptance and Competitive Behavior*. Graduate School of Business Administration, George F. Baker Foundation, Harvard University.

Hegarty, W Harvey, Henry P Sims. 1978. Some determinants of unethical decision behavior: An experiment. *Journal of Applied Psychology* **63**(4) 451.

Ho, Teck-Hua, Sergei Savin, Christian Terwiesch. 2002. Managing demand and sales dynamics in new product diffusion under supply constraint. *Management Science* **48**(2) 187–206.

Hoffman, Karla L, Carl M Harris. 1986. Estimation of a caller retrial rate for a telephone information system. *European Journal of Operational Research* **27**(2) 207–214.

Huang, Jian, Mingming Leng, Mahmut Parlar. 2013. Demand functions in decision modeling: A comprehensive survey and research directions. *Decision Sciences* **44**(3) 557–609.

Hyndman, Rob J, Anne B Koehler. 2006. Another look at measures of forecast accuracy. *International Journal of Forecasting* **22**(4) 679–688.

Jiang, John Xuefeng, Mary Harris Stanford, Yuan Xie. 2012. Does it matter who pays for bond ratings? historical evidence. *Journal of Financial Economics* **105**(3) 607–621.

Kahn, Kenneth B. 2002. An exploratory investigation of new product forecasting practices. *Journal of Product Innovation Management* **19**(2) 133–143.

Kapuscinski, Roman, Rachel Q Zhang, Paul Carbonneau, Robert Moore, Bill Reeves. 2004. Inventory decisions in dell's supply chain. *Interfaces* **34**(3) 191–205.

Kilduff, Gavin, Adam Galinksy, Edoardo Gallo, James Reade. 2015. Whatever it takes to win: Rivalry increases unethical behavior. *Academy of Management Journal* amj–2014.

Kinney, William R, Zoe-Vonna Palmrose, Susan Scholz. 2004. Auditor independence, non-audit services, and restatements: Was the us government right?*. *Journal of Accounting Research* **42**(3) 561–588.

Knittel, Christopher R. 2011. Automobiles on steroids: Product attribute trade-offs and technological progress in the automobile sector. *American Economic Review* **2012**(101) 3368–3399.

Kraft, Tim, Feryal Erhun, Robert C Carlson, Dariush Rafinejad. 2013. Replacement decisions for potentially hazardous substances. *Production and Operations Management* **22**(4) 958–975.

Kulik, Brian W, Michael J O'Fallon, Manjula S Salimath. 2008. Do competitive environments lead to the rise and spread of unethical behavior? parallels from enron. *Journal of Business Ethics* **83**(4) 703–723.

Kulkarni, Vidyadhar G. 1983. On queueing systems with retrials. *Journal of Applied Probability* 380–389.

Kulkarni, Vidyadhar G., Bong Dae Choi. 1990. Retrial queues with server subject to breakdowns and repairs. *Queueing systems* **7**(2) 191–208.

Kumar, Sunil, Jayashankar M Swaminathan. 2003. Diffusion of innovations under supply constraints. *Operations Research* **51**(6) 866–879.

Kurawarwala, Abbas A, Hirofumi Matsuo. 1996. Forecasting and inventory management of short life-cycle products. *Operations Research* **44**(1) 131–150.

Levitt, Theodore. 1965. Exploit the product life cycle. *Harvard Business Review* **18** 81–94.

Mandelbaum, Avi, William A Massey, Martin I Reiman, Alexander Stolyar, Brian Rider. 2002. Queue lengths and waiting times for multiserver queues with abandonment and retrials. *Telecommunication Systems* **21**(2-4) 149–171.

Mandelbaum, Avishai, William A Massey, Martin I Reiman, Brian Rider. 1999. Time varying multiserver queues with abandonment and retrials. *Proceedings of the 16th International Teletraffic Conference*, vol. 4. 4–7.

Markarian, Garen, et al. 2014. Product market competition, information and earnings management. *Journal of Business Finance and Accounting* **41**(5-6) 572–599.

Mayzlin, Dina, Yaniv Dover, Judith Chevalier. 2014. Promotional reviews: An empirical investigation of online review manipulation. *The American Economic Review* **104**(8) 2421–2455.

Melumad, Nahum D, Amir Ziv. 2004. Reduced quality and an unlevel playing field could make consumers happier. *Management science* **50**(12) 1646–1659.

Nistor, Cristina, Catherine Tucker. 2015. Third party certification: The case of medical devices. *Available at SSRN 2554984* .

Niu, Shun-Chen. 2006. A piecewise-diffusion model of new-product demands. *Operations Research* **54**(4) 678–695.

Olivares, Marcelo, Gérard P Cachon. 2009. Competing retailers and inventory: An empirical investigation of general motors' dealerships in isolated us markets. *Management Science* **55**(9) 1586–1604.

Olley, G Steven, Ariel Pakes. 1992. The dynamics of productivity in the telecommunications equipment industry. Tech. rep., National Bureau of Economic Research.

Pierce, Lamar, Jason Snyder. 2008. Ethical spillovers in firms: Evidence from vehicle emissions testing. *Management Science* **54**(11) 1891–1903.

Raith, Michael, et al. 2003. Competition, risk, and managerial incentives. *American Economic Review* **93**(4) 1425–1436.

Reed, Josh, Uri Yechiali. 2013. Queues in tandem with customer deadlines and retrials. *Queueing Systems* **73**(1) 1–34.

Reynolds, Lloyd G. 1940. Cutthroat competition. *The American Economic Review* 736–747.

Rink, David R, John E Swan. 1979. Product life cycle research: A literature review. *Journal of Business Research* **7**(3) 219–242.

Roberts, SW. 2000. Control chart tests based on geometric moving averages. *Technometrics* **42**(1) 97–101.

Rust, John. 1987. Optimal replacement of gmc bus engines: An empirical model of harold zurcher. *Econometrica: Journal of the Econometric Society* 999–1033.

Schwieren, Christiane, Doris Weichselbaumer. 2010. Does competition enhance performance or cheating? a laboratory experiment. *Journal of Economic Psychology* **31**(3) 241–253.

Shen, Haipeng. 2010. Statistical analysis of call-center operational data: Forecasting call arrivals, and analyzing customer patience and agent service. *Wiley Encyclopedia of Operations Research and Management Science* .

Shin, Yang Woo, Taek Sik Choo. 2009. M/m/s queue with impatient customers and retrials. *Applied Mathematical Modelling* **33**(6) 2596–2606.

Shleifer, Andrei. 2004. Does competition destroy ethical behavior? *American Economic Review* **94**(2) 414–418.

Short, Jodi L, Michael W Toffel, Andrea R Hugill. 2013. What shapes the gatekeepers? evidence from global supply chain auditors. *Harvard Business School Technology and Operations Mgt. Unit Working Paper* 14–032.

Snyder, Jason. 2010. Gaming the liver transplant market. *Journal of Law, Economics, and Organization* **26**(3) 546–568.

Stark, John. 2015. Product lifecycle management. *Product Lifecycle Management*. Springer, 1–29.

Staw, Barry M, Eugene Szwajkowski. 1975. The scarcity-munificence component of organizational environments and the commission of illegal acts. *Administrative Science Quarterly* 345–354.

Sutton, John. 2007. Market share dynamics and the persistence of leadership debate. *The American Economic Review* **97**(1) 222–241.

Sze, David Y. 1984. Or practice—a queueing model for telephone operator staffing. *Operations Research* **32**(2) 229–249.

Tigert, Douglas, Behrooz Farivar. 1981. The Bass new product growth model: a sensitivity analysis for a high technology product. *The Journal of Marketing* **45** 81–90.

Utgård, Jakob, Arne Nygaard, Robert Dahlstrom. 2015. Franchising, local market characteristics and alcohol sales to minors. *Journal of Business Research* **68**(10) 2117–2124.

Vives, Xavier. 2008. Innovation and competitive pressure. *The Journal of Industrial Economics* **56**(3) 419–469.

Wu, S David, Berrin Aytac, Rosemary T Berger, Chris A Armbruster. 2006. Managing short life-cycle technology products for agere systems. *Interfaces* **36**(3) 234–247.

Yang, Tao, James G. C. Templeton. 1987. A survey on retrial queues. *Queueing systems* **2**(3) 201–233.

Yeh, Sonia. 2007. An empirical analysis on the adoption of alternative fuel vehicles: the case of natural gas vehicles. *Energy Policy* **35**(11) 5865–5875.

Yu, Qiuping, Gad Allon, Achal Bassamboo. 2016. How do delay announcements shape customer behavior? an empirical study. *Management Science* .

Zhan, Dongyuan, Amy R Ward. 2013. Threshold routing to trade off waiting and call resolution in call centers. *Manufacturing and Service Operations Management* **16**(2) 220–237.

Zhu, Kaijie, Ulrich W Thonemann. 2004. An adaptive forecasting algorithm and inventory policy for products with short life cycles. *Naval Research Logistics* **51**(5) 633–653.

APPENDIX A

Time Series Clustering

The proximity of behaviors between two series X_t and Y_t is evaluated by means of the first order temporal correlation coefficient, which is defined by

$$CORT(X_t, Y_t) = \frac{\sum_{t=1}^{T}(X_{t+1} - X_t)(Y_{t+1} - Y_t)}{\sqrt{\sum_{t=1}^{T}(X_{t+1} - X_t)^2}\sqrt{\sum_{t=1}^{T}(Y_{t+1} - Y_t)^2}} \tag{A.1}$$

$CORT(X_t, Y_t)$ falls into $[-1, 1]$ with 1 means that two series behave similarly, i.e. their increase or decrease at any instant of time are similar in direction and rate, -1 means that the two series have similar rate of change but oppositie in direction, and 0 means that the two series are stochastically linearly independent. The proximity on values are measured as the conventional Euclidean distance with $d(X_t, Y_t) = \sqrt{\sum_{t=1}^{T}(X_t - Y_t)^2}$. The dissimilarity index to measure the proximity between series X_t and Y_t is proposed as

$$d_{CORT}(X_t, Y_t) = \phi_k[CORT(X_t, Y_t)]d(X_t, Y_t) \tag{A.2}$$

where $\phi_m(\cdot)$ is an adaptive tuning function to adapt the distance metrics $d(X_t, Y_t)$ to the temporal correlation $CORT(X_t, Y_t)$. With m to be the tuning parameter, the function $\phi_k(u)$ is written as

$$\phi_m(u) = \frac{2}{1 + e^{\mu}}, m \geq 0.$$

In our case, we use k equal to 2, which is the default choice. When using other values for m, the results change very little. Note the dissimilarity measure $d_{CORT}(X_t, Y_t)$ is model-free, it allows us to cluster the fitted PLC curves based on their features in terms of temporal structure and scales.