Inventory Management for Perishable Goods using Simulation Methods

by

Nicola Tan
S.B. Mechanical Engineering, Massachusetts Institute of Technology, 2007

Submitted to the MIT Sloan School of Management and the Mechanical Engineering Department in Partial Fulfillment of the Requirements for the Degrees of

Master of Business Administration
and
Master of Science in Mechanical Engineering

In conjunction with the Leaders for Global Operations Program at the Massachusetts Institute of Technology

June 2014

Signature of Author Signature redacted
MIT Sloan School of Management, MIT Department of Mechanical Engineering
May 9, 2014

Signature redacted
Certified by
Itai Ashlagi, Thesis Supervisor
Assistant Professor of Operations Management, MIT Sloan School of Management

Certified by Signature redacted
Daniel Whitney, Thesis Supervisor
Senior Research Scientist, Emeritus, Engineering Systems Division

Signature redacted
Accepted by
David E. Hardt, Chair
Mechanical Engineering Committee on Graduate Students

Accepted by Signature redacted
Maura Herson, Director of MIT Sloan MBA Program
MIT Sloan School of Management

Inventory Management for Perishable Goods using Simulation Methods

by

Nicola Tan

Submitted to the MIT Sloan School of Management and the MIT Department of Mechanical Engineering on May 9, 2014 in Partial Fulfillment of the Requirements for the Degrees of Master of Business Administration and Master of Science in Mechanical Engineering

Abstract

Amazon.com is the world's largest online retailer, and continues to grow its business by expanding into new markets and new product lines that have not traditionally been sold online. These product categories create new challenges to inventory and operations management. One example of this new type of products sold online includes the category of perishable goods. Perishable goods provide a unique inventory challenge due to the fact that products may expire at unknown times while in stock, making them unavailable for the customer to purchase.

This thesis discusses a method for managing perishable goods inventory by characterizing the key variables into empirical probability distributions and developing a computational model for determining the key inventory attribute: the reorder point. This model captures both the demand and loss due to shrinkage based on the age of the product in inventory.

The resulting model results in a 25% improvement in simulated inventory levels with more accurate results than current methods. This improvement is shown to come from accounting for the known variability in lead time, as well as survival rate of the product.

Thesis Supervisor: Itai Ashlagi
Title: Assistant Professor of Operations Management, MIT Sloan School of Management

Thesis Supervisor: Daniel Whitney
Title: Senior Research Scientist, Emeritus, Engineering Systems Division

Acknowledgments

I'd like to thank Amazon for hosting this internship and for their continued support of the LGO program. I owe thanks to many individuals, beginning with Miriam Park, my manager who continually drove and challenged me throughout this project. I worked with an amazing team, which included David Mabe, Brent Hill, Mike Huffaker, Jason Taam, and many others, all of whom taught me along the way. Jackie Underberg was an inspiring mentor, and Russell Murnen helped show me the ropes of Datanet and also the best food trucks in South Lake Union.

I also owe a debt of gratitude to my LGO classmates who helped me navigate Amazon—Gold Truong, Chuck Cummings, and especially Mike Chun, with whom I shared too many cups of coffee to count. Thanks also to the Sloanie crew from the summer of 2013 – Olga, Patricia, ZZ, Harvey, Oren, and more – and the support and friendships forged over lunch and Seattle coffee.

Thank you to my academic advisors, Itai Ashlagi and Dan Whitney, for their guidance and support. Finally, thank you to my family for all the continued support along this whole journey.

Table of Contents

List of Figures

List of Tables

1 Introduction

1.1 Amazon.com Introduction

Amazon.com is the world's largest online retailer, founded in 1994 by Jeff Bezos, and headquartered in Seattle, WA. The company initially sold books online, but has expanded into a variety of product categories, from electronics and toys to clothing and industrial supplies. Amazon focuses on selection, price, and convenience of these items, touting itself as "Earth's most customer centric company." Additionally, they have expanded into international markets, including Japan, China, United Kingdom and Germany. This strategy has enabled strong, consistent sales revenue growth, with a ten-year average revenue growth of 31%.

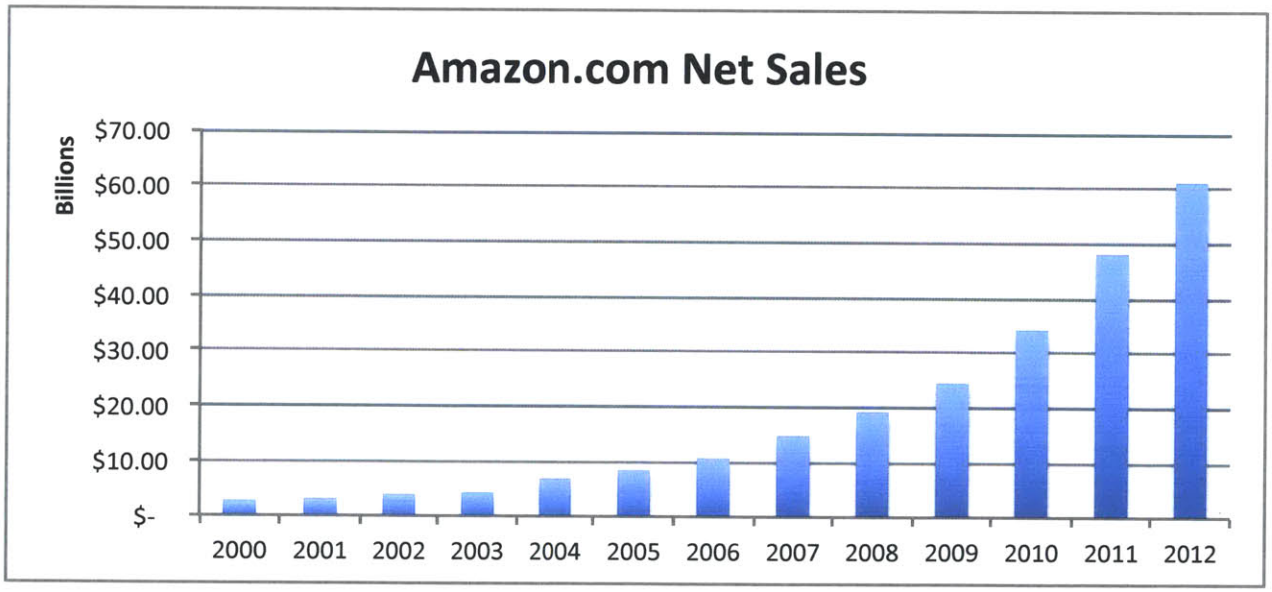

Figure 1: Amazon sales growth (from Amazon.com Annual Reports)

However, this rapid growth comes with some risk. As outlined in their 2013 Annual Report: "Our expansion places a significant strain on our management, operational, financial,

and other resources". In particular, the report mentions, "Fulfillment Center operation is key to profitability. If we do not successfully optimize and operate our fulfillment centers, our business could be harmed. If we do not adequately predict customer demand or otherwise optimize and operate our fulfillment centers successfully, it could result in excess or insufficient inventory or fulfillment capacity, result in increased costs, impairment charges, or both, or harm our business in other ways."

"If we do not stock or restock popular products in sufficient amounts such that we fail to meet customer demand, it could significantly affect our revenue and our future growth. If we overstock products, we may be required to take significant inventory markdowns or write-offs, which could reduce profitability," the report continues. These excerpts reflect the fact that effectively managing inventory is a key factor to Amazon's continued success.

1.2 Problem Statement

As Amazon expands into new product categories, new challenges arise from the specific constraints these categories bring. One example is in the category perishable goods. The term perishable goods refers to products that have a limited shelf life. Nahmias [1] draws an important distinction between inventory subject to obsolescence and inventory subject to perishability. He describes obsolescence occurring when a product has been superseded by a better version. In this case, the inventory does not change form, but the environment around it changes. Sometimes the remaining value is zero, but more often, goods retain some, albeit reduced, value. Common examples include electronics or fashion items.

Perishability refers to products that maintain a constant value until a certain expiration date, which may be deterministic (known) or stochastic (unknown). After the expiration date, the value of the inventory goes to zero. Initial studies in perishable inventory focused on blood bank

management, where blood has a legal lifetime of 21 days. However, other examples might include plants, pharmaceuticals, film, or packaged foods.

Improper management of perishable goods can lead to a number of costly issues:

1. Too much inventory leads to high holding costs, as well as high shrink costs due to slower inventory turns
2. Too little inventory results in stock-outs, leading to lost sales

Shrinkage, or shrink, refers to the difference between actual and recorded inventory. In retail environments, this can be caused by theft, damage, or lost products. In systems with perishable inventory, the perishability of an item, where an item expires and is no long sellable, is often a significant factor contributing to shrink.

Managing this category of products is difficult. In particular, a number of parameters are variable, requiring inventory managers to make judgments in how much product to order and when. In particular, when an item goes out of stock, the typical reaction is to order higher quantities of the product. However, this increases the length of time the product sits in inventory, resulting in higher shrink rates as the product outdates. When large shrink events occur, the product goes out of stock. This has resulted in a destructive cycle of high inventories and low customer service levels, as seen in Figure 2.

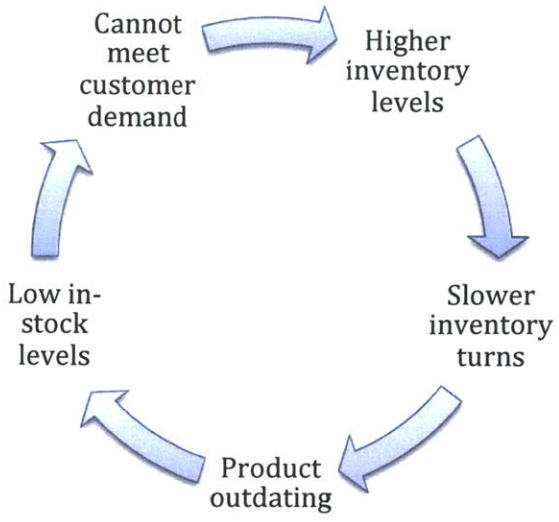

Figure 2: Destructive inventory and service cycle from improper management of perishable goods

1.3 Project Goals

This thesis will develop a methodology for measuring the key parameters for maintaining perishable goods, and will utilize these models to develop a simulation to determine the optimum inventory policy for this product category. This will result in minimizing costs at a given service level. This method will be applicable to a number of products that may exhibit varying demand, lead-time, and perishability behaviors.

1.4 Thesis Overview

This thesis will provide a method of characterizing existing data, as well as an inventory model that provides an optimal reorder point for a perishable inventory system.

Chapter 1 of this thesis provides an overview of the context and project goals, while Chapter 2 details the fulfillment process at Amazon to provide situational context. Chapter 2 also includes an overview of the computing resources developed by Amazon that allows new

approaches to be taken in inventory management. Chapter 3 details common methods in inventory control, while also providing a review of past investigations in the field of perishable inventory. Chapters 4-6 go into detail the methods for characterizing the key variable inputs: demand, lead time, and perishability. Chapter 7 provides a method for determine they key parameter in this system, the reorder point, while Chapter 8 concludes with a discussion and recommendations for future research.

In this thesis, several terms are used interchangeably, particularly with regards to the perishability of the product. These terms that refer to an item being inspected and deemed no longer sellable and include: expiration, discard, outdating, and shrink.

2 Amazon Fulfillment and IT Overview

In order to better understand the specific challenges of managing perishable inventory, we must understand the context in which it is managed. This chapter will provide insight into the basics of Amazon fulfillment processes, which are used for nearly all products that go through its warehouses. This fulfillment process will be used to determine the management processes and principles which will be used to determine proper inventory management.

Additionally, this chapter will discuss the information technology (IT) environment at Amazon at the time of this thesis. Amazon as a company has grown into a worldwide leader in IT management, and this strength has enabled new uses of computation that have not been previously widely used.

2.1 Fulfillment Center Operations

Amazon ships products to customers out of a network of fulfillment centers, where products are received from suppliers and packed and shipped to customers. Although a variety of fulfillment centers types exist for different types of product, the general process flow is typically very similar. This process is outlined in Figure 3.

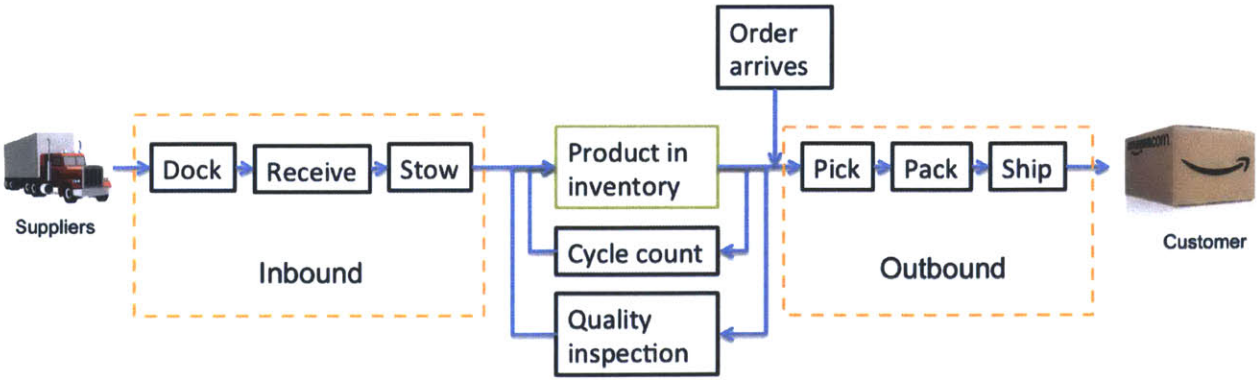

Figure 3: Amazon Fulfillment Center process flow

Amazon operations are separated into inbound and outbound operations, known simply as "inbound" and "outbound". Inbound refers to steps taken before a customer order has been placed. First, trucks arrive and product is unloaded in the dock operation. In receive, the product is identified and labeled, as well as scanned into database systems that track inventory. Then in stow, the product is placed in inventory, with its location constantly tracked virtually.

While product sits in inventory, cycle counts and quality checks are performed regularly to maintain an accurate account of inventory quantities, as items may become damaged, misplaced, or stolen over time. For perishable goods, this process also includes regular inspection to make sure that the product meets conditions for sale. If the product fails inspection, it is removed from inventory and is not available for sale.

When an order is placed by the customer and received by Amazon, it goes into outbound operations. First, the order is received by an Amazon associate, where it is picked; that is, it is located and taken off the shelf. Then it is consolidated with other items in that order, if necessary, and goes to pack, where the order is boxed and labeled. Finally, the order is sent to shipping, where it is placed in the appropriate truck for delivery to the customer.

2.2 Amazon Web Services

Amazon has also expanded beyond product fulfillment into cloud computing services. Amazon Web Services is now one the world's largest computing platforms, and provides these services in a flexible environment to the public. Although revenue figures are not published, it is estimated to generate $2 billion in revenue for Amazon [2], while it holds five times the market share than the next 14 competitors combined [3]. These computing services were initially developed for internal Amazon infrastructure, allowing Amazon to quickly scale computing

power for its quickly growing operations [4]. This history and current AWS capabilities are indicative of the IT infrastructure available not only at Amazon, but now to the general public.

2.3 Conclusion

The context in which inventory is managed is integral in determining what kind of system to utilize. The fulfillment center operations will determine key parameters such as how inventory is inspected and how often, which will be key inputs for the eventual model. Additionally, the computational environment available at Amazon, as well as at other organizations through distributed computing environments, opens the doors to utilizing more computationally intensive approaches than have been previously used in industry.

3 Literature Review

This chapter will provide a review of the academic literature surrounding inventory management in general, and specifically, in the field of perishable inventory. This theory will lay the groundwork for the inventory management in this specific instance. The first section, classifications for order policies, will provide the outline of the key variables that need to be considered when determining an inventory policy. Section 3.2 will provide a primer on the basic inventory management theory for non-perishable items. Section 3.3 provides a review of key studies in perishable inventory that are built upon in this thesis.

3.1 Classifications for Order Policies

Significant work has been done in determining various models for perishable inventories. Silver [5] and Raafat [6] provide an initial classification, which is further expanded on here. Key considerations in determining an inventory management model include the following components: order policy, demand pattern, lead time, inventory type, deterioration rate, sortages, and other factors.

3.1.1 Order Policy

Order policy refers to a number of possible review and order quantity policies. This can be either a fixed input, based on existing systems, or can be an adjustable variable based on the product type and other specifics of the order. Determining factors are how often the inventory status is determined, when a replenishment order should be placed, how large the replenishment order should be. Silver [7] outlines four basic types of control systems:

1. **Order-Point, Order-Quantity (s,Q) system** – inventory is reviewed continuously, and when the inventory position reaches below a level s, an order quantity Q is placed.

2. **Order-Point, Order-Up-to-Level (s,S) system** – inventory is reviewed continuously, as in (s,Q) and when inventory position reaches level s, an order is placed to replenish inventory up to level S.
3. **Periodic-Review, Order-Up-To-Level (R,S) system** – inventory is reviewed only periodically, every R units of time. At each review, enough units are ordered to raise inventory position to level S.
4. **(R,s,S) system** – this is a combination of the (s,S) and (R,S) systems. Every R units of time, inventory position is checked. If inventory position is below reorder point s, enough inventory is ordered to raise the position to S. If inventory position is above s, no action is taken.

3.1.2 Demand Pattern

Demand can be either deterministic, as in the case of factory orders that are placed well in advance; or stochastic, where demand is probabilistic and variable. In stochastic cases, the demand fluctuations can be modeled by a known probability distribution (normal, binomial, etc.) or can follow a more arbitrary distribution Additionally, demand can be relatively static in time, or it can vary over time, either increasing or decreasing, or seasonally.

3.1.3 Lead Time

Lead time is the time that elapses between the time an order is placed and the time it is received. It can be treated differently in different models. In the simplest case, lead time is zero, where an order is placed and instantaneously received into inventory. In most cases however, lead time is a positive value. This can be known and fixed, such as two weeks from delivery to order; it can also be modeled probabilistically, following a known or arbitrary probability distribution.

3.1.4 Inventory Type

Inventory can take on specific shelf life characteristics. Goyal [8] divides inventoried goods into three meta-categories. The first type is inventory assumed to have infinitely long shelf life. This includes products whose shelf life is significantly longer than the time spent in inventory, and thus whose shelf life is irrelevant and can be ignored. Inventory can also experience obsolescence. Nahmias describes obsolescence occurring when an item is superseded by a better version, such as in electronics and fashion, in which the introduction of new products can reduce or eliminate the utility of a product. In the case of obsolescence, the product itself does not change, but the environment surrounding it changes. Because the overall environment is constantly changing, obsolescence cannot be typically modeled. This is in contrast to perishable inventory. Perishable inventory loses utility over time due to a change in the product itself. The rate of this deterioration can vary, as described in the next section.

3.1.5 Deterioration Rate

Of items that are perishable, they can exhibit multiple types of behavior in regards to deterioration. First is a constant fractional rate of decay, in which decay is proportional to overall volume and time. This is can be seen in volatile liquids subject to evaporation, such as alcohol, or in radioactive materials. The constant rate of decay can be modeled using an exponential time model. Next is a fixed lifetime model. This is the case in which a product holds it utility for a fixed and known period of time, then the product loses its value. A classic example of this case is in blood bank management, where the legal lifetime of blood is 21 days, after which the product is no longer usable. Other cases include prescription drugs or packaged goods which include a known expiration date. The last case is the random lifetime case, in which the lifetime of a product is unknown. Several studies have assumed various probability models for the lifetime of

a product, including Covert and Philip [9] who applied a Weibull distribution to the deterioration of a product, and Tadikamalla [10] who instead used a gamma distribution. Tadikamalla [10] also compared similar Weibull and gamma distributions and found that similar Weibull and gamma distributions can have significantly different instantaneous deterioration rate functions (hazard functions).

3.1.6 Shortages

The treatment of shortages is also an important variable. If demand can be bounded, then a constraint that no shortages are allowed can be applied. Otherwise, shortages can either result in backorders, where demand is filled when a inventory is replenished; or they can simply result in lost sales.

3.1.7 Other Factors

Many others have studied various cases including:

- Price changes (discount, known price increase)
- Two warehouses
- Quantity discount
- Delivery rate
- Multiple items
- Shortages allowed
- Single period (newsvendor) vs. multi period

3.2 Review of Economic Order Quantity

The basic Economic Order Quantity (EOQ) theory assumes [7]:

1. Demand is constant and deterministic

2. Order quantity does not have any restrictions on size
3. Unit cost does not depend on replenishment quantity (no quantity discounts)
4. Cost factors (e.g., inflation) do no vary in time
5. The item can be ordered independently (there are no effects of joint review or replenishment)
6. Replenishment lead time is of zero duration, or it is a known nonzero value
7. No shortages are allowed
8. The entire order quantity is ordered at the same time
9. Parameters will hold the same values indefinitely.

From these assumptions, the economic order quantity is determined, where the holding cost is balanced with the order costs to find the lowest overall costs. The EOQ, or the most cost efficient order quantity, is found to be

$$EOQ = \sqrt{\frac{2AD}{vr}} \quad (1)$$

where

A = the fixed cost per order

D = demand rate per item (in units/time)

v = unit variable cost per item

r = holding cost (in $ cost/$ inventory/unit time)

Because demand is constant, the EOQ can also determine how often an order should be placed. This time to replenishment, or T_{EOQ} is

$$T_{EOQ} = \frac{EOQ}{D} \quad (2)$$

If lead time is a known non-zero value, a reorder point can be placed such that a new order arrives just as the current inventory will run out. This results in inventory levels as seen in Figure 4.

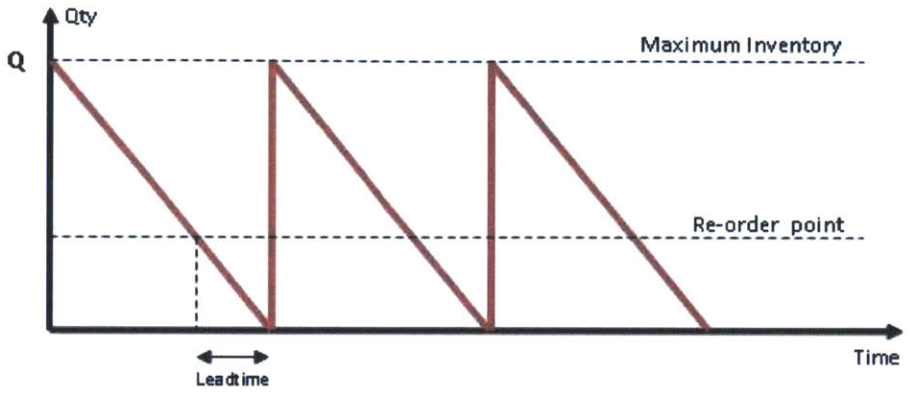

Figure 4: Example of basic EOQ inventory policy

This model is then extended to items where demand is variable. In order to account for variance in demand, a safety stock is established, which covers the difference in expected (mean) demand and a specific service level of demand, as seen in Figure 5.

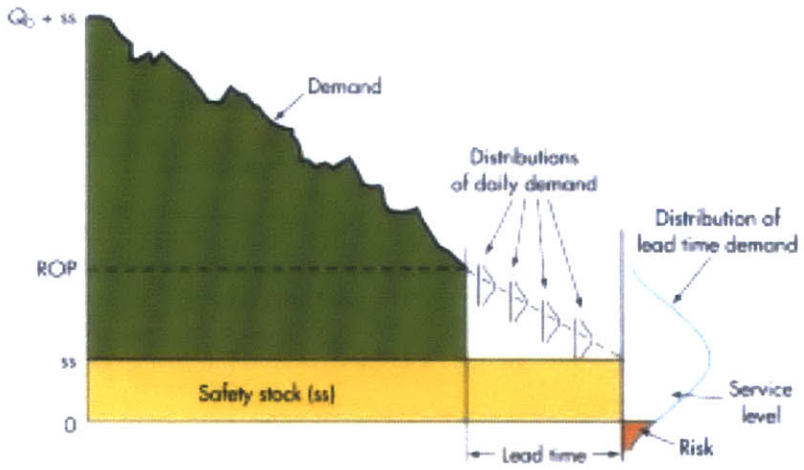

Figure 5: Example of inventory with random demand

The service level can be optimized given the purchase costs, out of stock costs, and salvage costs. This is done using a newsvendor model analysis, which optimizes between overstock and understock costs; or it can be a given input based on organizational goals.

3.3 Perishable Inventory Models

Perishable inventory models build upon these basic models to provide optimal solutions in specific cases. One of the early analyses of perishable inventory was done by Ghare and Schrader [11], who assumed a constant rate of decay and a constant demand. This constant rate of decay results in an exponential probability density function, which is shown Figure 6.

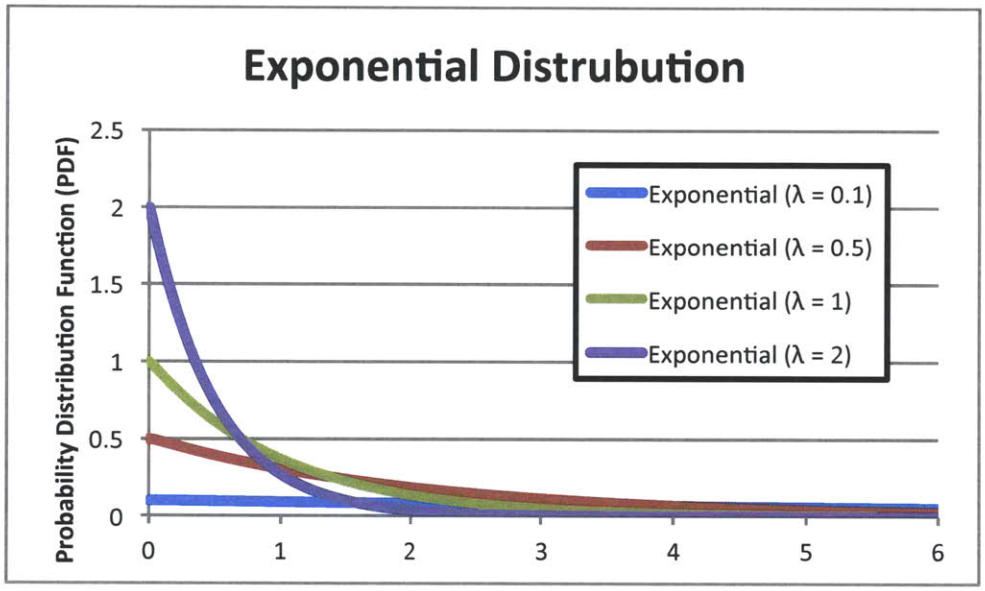

Figure 6: Examples of the exponential distribution

This resulted in a change in inventory level, I(t):

$$I(t)' = -\theta \mathrm{I}(\mathrm{t}) - \mathrm{d} \qquad (3)$$

where:

I(t) is the inventory level at time t

θ is the decay rate, which is assumed to be constant

d is the demand, which is assumed to be constant

Graphically, this results in an inventory curve such as that in Figure 7.

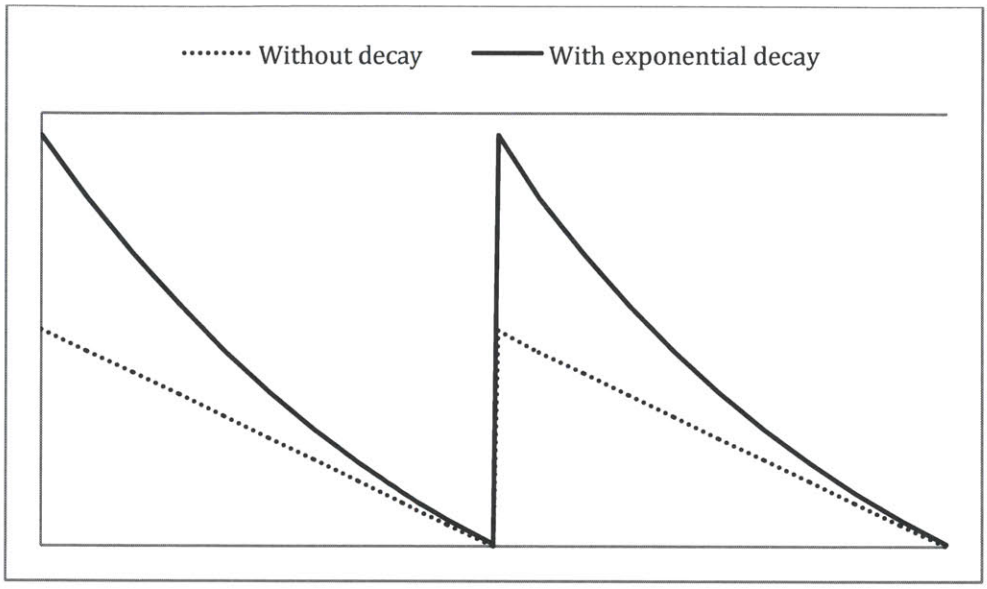

Figure 7: Inventory with decay

They then derived an equation for optimum order quantity Q*, as a function of inventory cycle time, T:

$$Q^* = \frac{d}{\theta}\exp(\theta T) - 1 \qquad (4)$$

Utilizing a Taylor series expansion, inventory cycle time T can be determined with the following equation:

$$\frac{C_1 d}{2} + \frac{C_1 dT\theta}{2} + \frac{C_4 \theta d}{2} + \frac{C_3}{T} = 0 \qquad (5)$$

where C_1, C_3, C_4, are inventory carrying cost, ordering cost, and deteriorating cost. The optimal inventory cycle time can be determined by solving Equation 5 for T, giving T*, the optimum time between orders.

Covert and Philip [9] expanded on this work to determine an economic order quantity for a variable deteriorating rate, assuming a two parameter Weibull distribution. The Weibull hazard function, which provides the instantaneous deterioration rate, g(t), is given as

$$g(t) = \alpha\beta t^{(\beta-1)} \qquad (6)$$

where α and β are the scale and shape parameters of the distribution. Examples of Weibull probability density functions is seen in Figure 8:

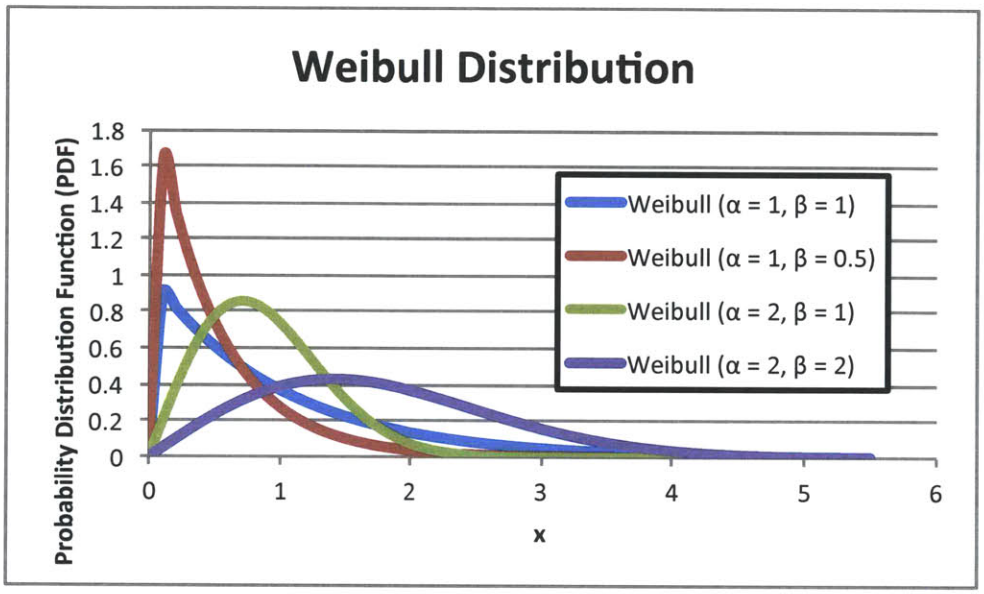

Figure 8: Examples of the Weibull distribution

This results in the following differential equation describing inventory levels

$$I(t) = -g(t)I(t) - d \tag{7}$$

The lot size is then given by

$$Q = \int_0^T [d\exp(\alpha t^{\beta})]dt$$

$$Q = \frac{d\sum_{n=0}^{\infty}\alpha^n T^{(n\beta+1)}}{n!(n\beta+1)} \tag{8}$$

If average inventory is assumed to be Q/2 the following equation can be obtained:

$$\frac{C_1 d\exp(\alpha T^{\beta})}{2} + (C_4 d)\left[\sum_{n=0}^{\infty}\frac{\alpha^n n\beta T^{(n\beta-1)}}{(n\beta+1)n!}\right] + \frac{C_3}{T} = 0 \tag{9}$$

Solving for T results in the optimum order period in this scenario.

Tadikmalla [10] expanded on Covert and Philip, utilizing a gamma distribution for product deterioration. Examples of gamma probability density distributions are shown in Figure 9.

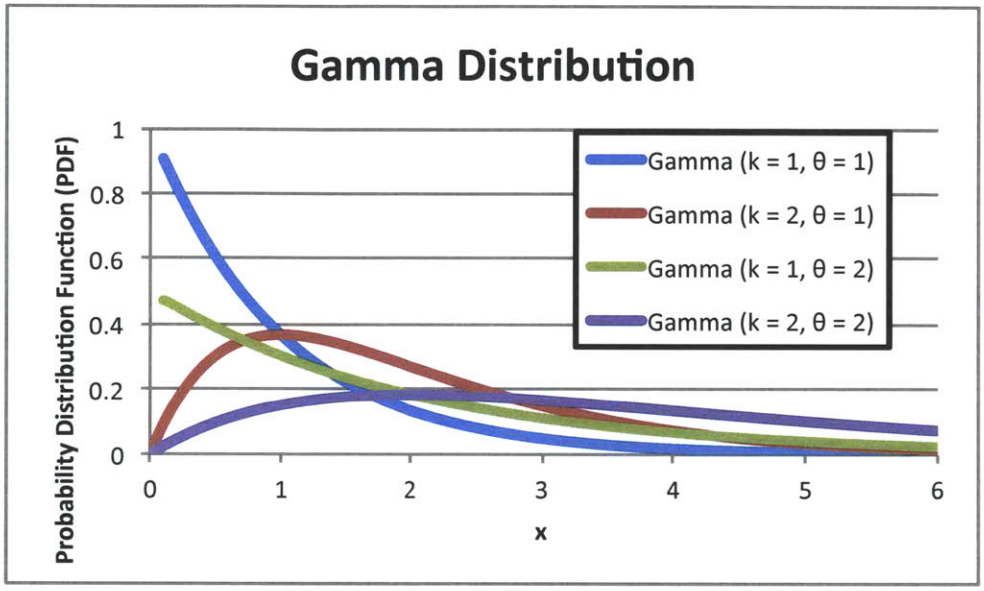

Figure 9: Examples of gamma distribution

Tadikmalla compared his findings to that of Covert and Philip and found that even for Gamma functions that are have a similar Weibull counterpart (see Figure 10), instantaneous decay functions can vary dramatically between the two models.

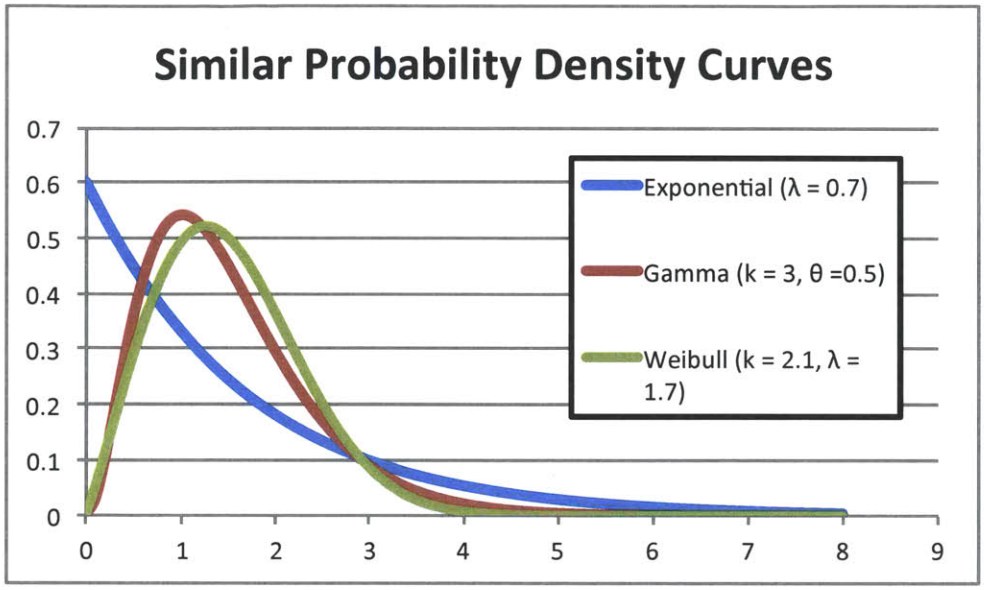

Figure 10: Comparison of PDF for various models

A number of other studies have been done to characterize perishable inventory systems in various contexts—permutations of the classifications outlined in section 3.1. Nahamias [1], Raafat [6], and Goyal et. al [8] provide extensive reviews of the various models that have currently been developed.

3.4 Computational Considerations for Perishable Inventory Models

Silver notes the value of using heuristic approaches, such as Wagner-Whitin [12] or Silver-Meal [13] algorithms, is that it captures the complexity without requiring lengthy computations. However, these heuristics are only applicable in specific cases, and are difficult to extend into more specific and complex situations. Raafat [6] notes that when the deterioration rate of a perishable good is not constant, closed form solutions are generally not possible, and calculations become extremely difficult. Additionally, the classification system proposed by Raafat [6] and expanded upon in this paper illustrates the complexity of this category of problems, with each variation of the problem having a different solution.

3.5 Conclusion

Chapter 3 summarizes key research relevant to developing new perishable inventory approaches. First, we outline the various factors that go into developing an order policy, which is relevant in both perishable and non-perishable inventory systems. Then, we build upon basic inventory theory to expand into preliminary models for perishable goods. This provides a basis for building upon in the remainder of this thesis.

4 Demand Forecasting

After a review of Amazon processes and key literature, Chapters 4-6 provide implementation details of an improved inventory management system. The first key part to understand is demand, which is the focus of Chapter 4. Many basic models assume static demand; however, the actual realized demand in this context is constantly changing. Thus it is important to forecast demand accurately. This forecast will then feed into the overall inventory model. This chapter will describe the key criteria used to develop various forecasting methods, a description of methods tested, and methodology towards determining not only the most robust model, but the most robust parameters as well.

4.1 Key Factors

The first step in determining how much product to hold in inventory is to understand and forecast demand. Demand can vary based on a number of different factors, while the relative strength of these factors can vary between products as well as over time. Possible factors that affect the product demand in the particular setting studied were identified by the author as:

Number of customers

Some products are only available to select markets in the early stages of a new product introduction. As the product expands into new markets, it is expected that the demand will grow.

Time from market or product launch

When a product is available to a market, it may take a certain amount of time for customers to become aware of new offerings and to purchase these items through Amazon. Thus, even within a market, we expect uptake rates to grow over time.

Seasonality

Many products face seasonal changes in demand, where demand for the product cycles in regular intervals of time. Amazon acknowledges this in their annual report: "We expect a disproportionate amount of our net sales to occur during our fourth quarter." Demand for products can not only cycle annually, but quarterly, monthly, or weekly as well, depending on the product and purchasing drivers.

Noise

Finally, there is random variability in customer demand. This random noise must be decoupled from driving factors; however this can be difficult, making it difficult to accurately forecast items.

4.2 Methodology

With these factors in mind, a number of possible forecast methods were tested against historical data for accuracy. Initial screening tests were run against nine products, with three randomly chosen from each of slow (bottom 40% in terms of product velocity), medium (middle 30%), and fast (top 30%) moving products. The product velocity was determined by the relative number of items per time unit sold in the test period, which utilized three months data. Because the relative strength of each factor was still unknown, a number of models were evaluated and tested. These models ranged from simple to complex, incorporating the factors mentioned in the previous section to varied degrees. The following definitions are used in regards to forecast periods:

Period: a single forecast period, regardless of the season

Season: the specific season, which repeats regularly

Cycle: The entire set of forecast periods until a season is repeated

Cycle I					Cycle II					Cycle III				
Season A	Season B	Season C	Season D	Season E	Season A	Season B	Season C	Season D	Season E	Season A	Season B	Season C	Season D	Season E
Period 1	Period 2	Period 3	Period 4	Period 5	Period 6	Period 7	Period 8	Period 9	Period 10	Period 11	Period 12	Period 13	Period 14	Period 15

Figure 11: Cycle, season, and period distinctions

4.3 Models Evaluated

4.3.1 Model A: Average of Past *n* Forecast Periods

This model takes a simple moving average of the past n periods, factoring a lead time:

$$f(t_0) = \frac{1}{n}\sum_{i=1}^{n} d(t_0 - l - i) \qquad (10)$$

where

$f(t_0)$ = forecasted demand for time t_0

$d(t_i)$ = past demand for time t_i

n = number of forecast periods to measure

l = lead time

For example, a forecast cycle can be defined as a calendar year, with a season or period defined as one calendar month. Consider n = 3, the period being forecasted is July, and lead time is one month, so the forecast is being generated in June. The forecast would be the average sales of the past three months: March, April, and May.

The benefits of this approach are that it is intuitive and simple to measure. Additionally, it uses all past data points, so the number of data points used for a given period of time is relatively large. If n is chosen to be relatively high, the forecast will remain more stable to random demand fluctuations that occur month to month. A smaller value of n will be more reactive to changing demand patterns.

4.3.2 Model B: Average of the Past *n* Forecast Seasons

This method assumes that the seasonality of the product demand is a strong indicator of future demand, and that demand is independent of other seasons' demands. The forecast is determined using

$$f(t_0) = \frac{1}{n}\sum_{i=1}^{n} d(t_0 - c*i) \qquad (11)$$

where c is the number of seasons in a full cycle.

Using the above example which defines one season as a calendar month, if n = 2, the forecast for June 2014 will be the average of June 2013 and June 2012. Similar to method A, the larger the value n, the more stable to random fluctuations; the smaller the value n, the more responsive to demand changes the forecast is. One drawback of this method is that, compared to method A, less data is used for the same time period. For example, a value of n=3 requires three years of data in the above example, while only using three data points. In method A, three years of data would include 36 data points.

4.3.3 Model C: Linear Extrapolation Based on Past n Forecast Seasons

Model C accounts for the fact that in a growing market, Model B will systematically under-predict demand because it ignores any growth rate. Here, we assume a linear growth rate, and determine this rate using a least-squares regression of the past n seasons. This forecast is given as

$$f(t_0) = a(t_0) + b(t_0) * t_0 \qquad (12)$$

where a, b are the linear slope and intercept of $[t_0 - c * i, d(t_0 - c * i)]$ for $i = n$ points

4.3.4 Model D: Exponential Smoothing by Factor α of Past n Forecast Seasons (Weighted Average)

One challenge in Model C is that the line drawn using the least squares regression can be very sensitive to random noise elements, particularly if only a few points are used. Exponential smoothing is used as a way to weigh more recent data points more heavily than older data points. Thus, an outlier data point is mitigated by past data, providing a more consistent, accurate forecast. However, this method of weighted averaging is much more responsive to changing demand compared to Model B, with this responsiveness controlled using the parameter α, with larger values of α weighing more recent data more heavily.

$$f(t_0) = \alpha * d(t_0 - c) + (1-\alpha) * f(t_0 - c) \quad (13)$$

Because each forecast is the weighted average of past demands, Equation 13 can be expanded to show this relationship:

$$f(t_0) = \alpha * d(t_0 - c) + \alpha(1-\alpha) * d(t_0 - 2c) + \alpha(1-\alpha)^2 * d(t_0 - 3c) + \alpha(1-\alpha)^3 * d(t_0 - 4c)... \quad (14)$$

Thus we can see that all past data is considered, but each period contributes a decreasing amount to the overall forecast.

4.3.5 Model E: Winters Exponential Smoothing Procedure for a Seasonal Model

This method is the most complex tested. This method deconstructs baseline demand from seasonal weighting, and applies exponential smoothing to both baseline demand and the seasonal affects. The process for determining the forecast using the Winters Exponential Smoothing Procedure is outlined in Silver [7].

4.3.6 Model F: Average of Past n Forecast Periods, Weighted by Long Term Seasonal Trend

Because the Winters model requires multiple equations and is relatively complex to implement, this method takes a simplified approach to the general concept. Demand over n complete cycles is averaged by period. Because entire cycles are averaged, any seasonal effects are lost. The seasonal effects are then added in by taking the historical weighting of each season over m cycles. Thus, if baseline demand is changing, but the seasonal effect is strong but constant, this method can use a small value of n with a large value of m. This can provide the seasonal sensitivity that Model B provides while utilizing all data points in a time period, as Model A allows. The results in the following formula for the forecast:

$$f(t_0) = \frac{F(t_0)}{cn} \sum_{i=1}^{cn} d(t_0 - l - i) \qquad (15)$$

where

$$F(t_0) = \frac{\frac{1}{m} \sum_{i=1}^{m} d(t_0 - ci)}{\frac{1}{cm} \sum_{k=1}^{cm} d(t_0 - l - k)} \qquad (16)$$

4.3.7 Models G-L: Above Models with Demand Normalized by Aggregate Weekly Forecasts

The above forecasts capture only a single dimension of demand growth. In order to account for the hypothesis that demand within a market is growing and the number of active customers is also growing, demand used in the above forecast is normalized by the number of possible customers available.

4.4 Model Downselection

For each model and each SKU, parameters were chosen to minimize mean squared error (MSE) over a training set of data. This process was relatively trivial for one and two parameter models, where sensitivity tables were used to find the optimum parameters. For the Winters Exponential Smoothing model, three parameters were required. Excel Solver was used to determine these parameters, using the LP Evolutionary algorithm.

These parameters were then used to forecast demand over a test set of data. The MSE of the forecast to the actual demand was then calculated and averaged over the time frame of test data.

Table 1: Average mean square error (MSE) for each model and product tested

Average Mean Square Error						
	Model A	Model B	Model C	Model D	Model E	Model F
Product A	4.566	4.662	5.245	4.817	5.163	4.555
Product B	1.286	1.438	1.638	1.433	1.494	1.424
Product C	1.004	1.114	1.211	1.092	1.097	1.073
Product D	1.423	1.494	1.834	1.429	1.813	1.412
Product E	1.402	1.396	1.609	1.431	1.542	1.396
Product F	8.299	8.620	9.987	9.057	8.120	8.400
Product G	2.091	2.050	2.433	2.090	2.219	2.050
Product H	2.119	2.279	2.631	2.412	2.602	2.207
Product I	1.775	1.900	2.265	1.940	1.915	1.900

This value was used to determine the rank of models for each particular SKU. The average rank was used to determine the top models to continue evaluation.

Table 2: Model rank for each product tested

	Rank					
	Model A	**Model B**	**Model C**	**Model D**	**Model E**	**Model F**
Product A	2	3	6	4	5	1
Product B	1	4	6	3	5	2
Product C	1	5	6	3	4	2
Product D	2	4	6	3	5	1
Product E	3	2	6	4	5	1
Product F	2	4	6	5	1	3
Product G	4	2	6	3	5	1
Product H	1	3	6	4	5	2
Product I	1	3	6	5	4	2
Average Rank	**1.889**	**3.333**	**6.000**	**3.778**	**4.333**	**1.667**

For this particular system, a moving average (Model A), a seasonal moving average (Model B), and a seasonally weighted average (Model F) most consistently ranked highest amongst models tested. These models were then used for parameter optimization.

4.5 Parameter Optimization

With the number of base models downselected, universal parameters were then chosen. Each model was run with a range of parameters, and average adjusted MAPE was calculated. MAPE, or mean average percentage error, is typically calculated using:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{A_t - F_t}{A_t}\right| \qquad (17)$$

where A_t is the actual demand at time t and F_t is the forecasted demand at time t.

MAPE was chosen over MSE for this particular analysis because values are then averaged across a number of SKUs, or stock keeping units. Because items with higher demand tend to also have larger absolute forecast errors, we divide by the demand to normalize for this. However, traditional MAPE calculation incurs two key problems. First, in cases in which there was no demand ($A_t = 0$), a divide by zero error occurs. Additionally, when the forecast difference

and possible demand variation is relatively equal, MAPE can return highly variable values depending on whether actual demand is higher or lower than forecasted. For example, if the forecast is for 3 units, and actual demand is 2, the forecast difference is 1, and the percentage error is 50%. However, if the actual demand is 4, the absolute difference is still 1, but the percentage error is now 25%. In order to minimize this effect, adjusted MAPE divides by the average actual demand ($\overline{A_t}$) of the past cycle.

$$\text{Adjusted MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{A_t - F_t}{\overline{A_t}}\right| \qquad (18)$$

Adjusted MAPE is then averaged across SKUs. The parameters that resulted in the lowest average adjusted MAPE values were then chosen. Because of small differences in overall results, two sets of parameters for each of Model A (corresponding to the average of one and two full cycles, respectively) and Model B were chosen, as well as one set of parameters from Model F.

4.6 Test and Compare Forecasts

With forecasts and parameters selected, all SKUs are forecasted against this smaller number of models and adjusted MAPEs are calculated.

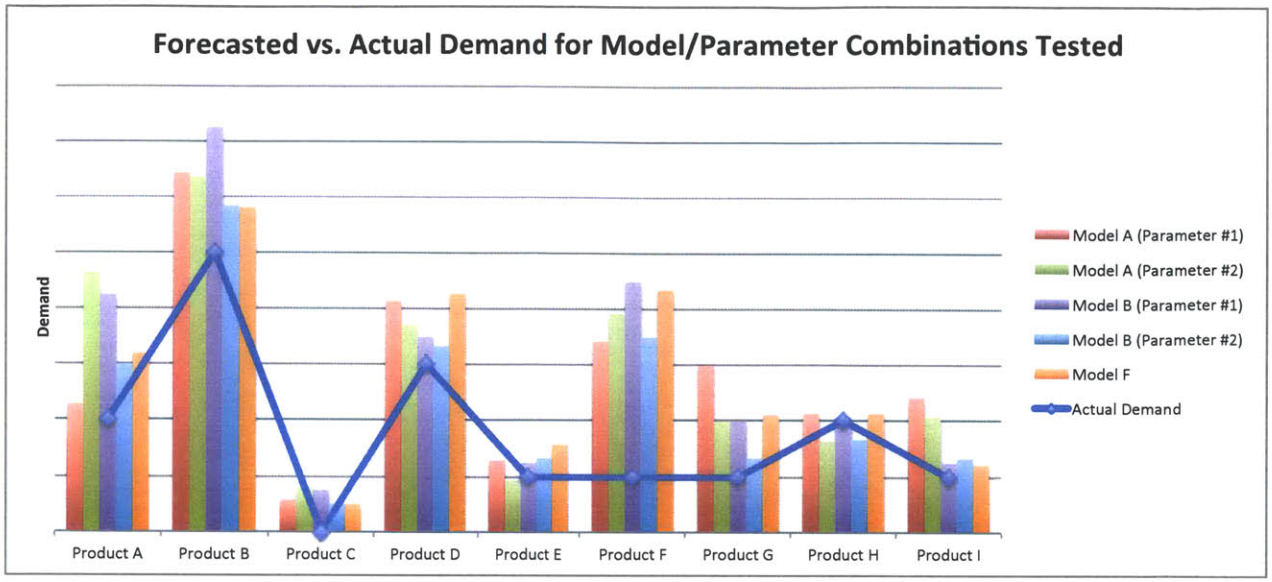

Figure 12: Actual vs. forecasted demand for model and parameter combinations tested

For each forecast method, the adjusted MAPE values are then compared to the baseline forecast using a t-test with unequal variances. Each method can then be evaluated on both forecast accuracy (absolute value of average adjusted MAPE) as well as consistency (p-value of t-test). A forecast method may have significant improvement in one set of SKUs but also worsen the forecast slightly for another set of SKUs. Thus the average error may go down, but the effect may be skewed by a few large numbers. In fact, the majority of SKUs may actually have error go up, so it is important to consider both factors.

Additionally, SKUs were divided into categories that were hypothesized to have different demand behaviors, which may influence which forecast is most accurate. Divisions included low vs. high velocity items and new vs. established markets.

4.7 Results

While overall forecasts improved using the best overall forecast method, the greatest improvement was through categorizing items by velocity as well as market type, and applying different models to each category. In a mature market, low velocity items were best forecasted

using the simplest model, Model A, with a low value of n. For these items, seasonal effects were minimal, but potential for baseline demand changes was high, so being responsive to these changes was advantageous. For high velocity items, Model G, a simplified version of the Winters Exponential Smoothing Model, was most effective. The original Winters model perhaps contained too many parameters and thus overfit the data, while the simplified version required less variables. A short look-back period is used for determining baseline demand; this requirement is utilized mostly for new markets where velocity may be relatively high, but is still changing. However, the look-back period for seasonal effects is quite long. This indicates the strength and relative noisiness of the seasonality effect.

In an emerging market, a seasonal buying pattern has not yet emerged, and a simple moving average of one cycle provided a consistently improved forecast.

Table 3: MAPE improvement and p-value comparisons for top two models in subdivided product groups

	Model A (Parameter #1)		**Model F**	
	% imp	**P-value**	**% imp**	**P-value**
Mature	0.90%	0.40	1.20%	0.37
Low vol	4.10%	0.23	-0.20%	0.51
High vol	-4.20%	1.00	3.30%	0.02
Emerging	7.50%	0.03	4.80%	0.11
Low vol	10.80%	0.04	5.60%	0.18
High vol	2.50%	0.09	3.50%	0.03

It is also important to note that using historical demand through past sales will likely underpredict actual demand, as unfilled demand when a product is out of stock is not captured.

4.8 Residuals

Because the forecast varies for each period, variance for each forecast is measure through the resulting residuals of each forecast. These residuals, or the difference between the predicted and actual demand, are recorded and used to simulate the variability in demand in Chapter 7.

4.9 Conclusion

Multiple models were developed and tested, and a methodology was designed to optimize the models and parameters used in building a demand forecast. A combination approach to demand forecasting, based on both market maturity and product velocity were determined to most accurately forecast demand. These results will then be used to predict inventory needed in future periods in the reorder point model developed later in this thesis.

5 Lead Time

A second attribute that was found to be variable was in lead time, or the time between when an order is placed and when it is received. Variability in lead time was known to be an issue in this situation, but was not previously quantified and applied to inventory management. This section will provide an overview of how to convert fill rate data into a lead time distribution which will be later feed into the final inventory model.

5.1 Key Considerations

Uncertainty in fill rates is another cause of uncontrolled inventory levels in the system analyzed in this paper. Fill rate refers to the percentage of items ordered in a purchase order (PO) that is actually received into inventory. Assume that items not received are cancelled (as opposed to backordered). Because inventory managers cannot be confident that a new supply of inventory will arrive when ordered, they have been observed to "hoard" supplies, contributing to increased inventory levels.

A root cause analysis discovered two separate causes for items to not be delivered when ordered:

1. Short term supplier managed inventory issues
2. Systematic supply chain issues

The first issue refers to individual suppliers temporarily going out of stock of an item. This is most often caused by mis-forecasted demand or late deliveries from second-tier suppliers. This issue is typically resolved quickly, either by ordering the item from another supplier, or by placing a reorder of the item for the next scheduled delivery.

The second issue refers to issues that affect the entire supply chain. This is often a supply disruption to the upstream supply. Because there are a small number of upstream suppliers, an

issue affecting a particular supplier will often then disrupt multiple first tier suppliers. In this case, resolving inventory will take longer that in the previous case as supply alternatives are limited.

5.2 Model Development

These considerations came into play when deciding how to account for low fill rates—all missed deliveries are not the same. Occasional, isolated indents where the missed shipment is easily replaced by another supplier have a different effect than consecutive missed deliveries across suppliers. Data was captured as fulfilled or unfulfilled for a particular order. Fill rates are thus addressed by converting into lead time, by tracking how many times a SKU is ordered until it is received. Because multiple suppliers may supply the same SKU, this is tracked across suppliers to capture market wide supply disruptions. An example of the conversion of filled and unfilled orders to lead time is shown in Table 4. A filled order for a specific product is given a 1, a missed order is given a -1. If that product is not ordered, it is given a 0. In order to convert to lead time, the number of consecutive times a product is ordered until it is fulfilled is counted for each fulfilled order.

Table 4: Fill rates of individual products and conversion to lead time

Fill Rates

Week #	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**
Product A	1	0	0	1	0	1	0	0	1	1	0	1	0
Product B	0	1	-1	1	1	-1	-1	1	-1	1	1	1	1
Product C	0	0	-1	1	-1	1	0	-1	-1	1	0	1	1

Lead Time

Product A	1			1		1			1	1		1	
Product B		1		2	1			3		2	1	1	1
Product C				2		2				3		1	1

This lead time is then converted to an empirical probability distribution based on the number of times a lead time appears.

Table 5: Probability of lead time by product

Lead Time (Weeks)	1	2	3
Product A	1	0	0
Product B	0.625	0.25	0.125
Product C	0.4	0.4	0.2

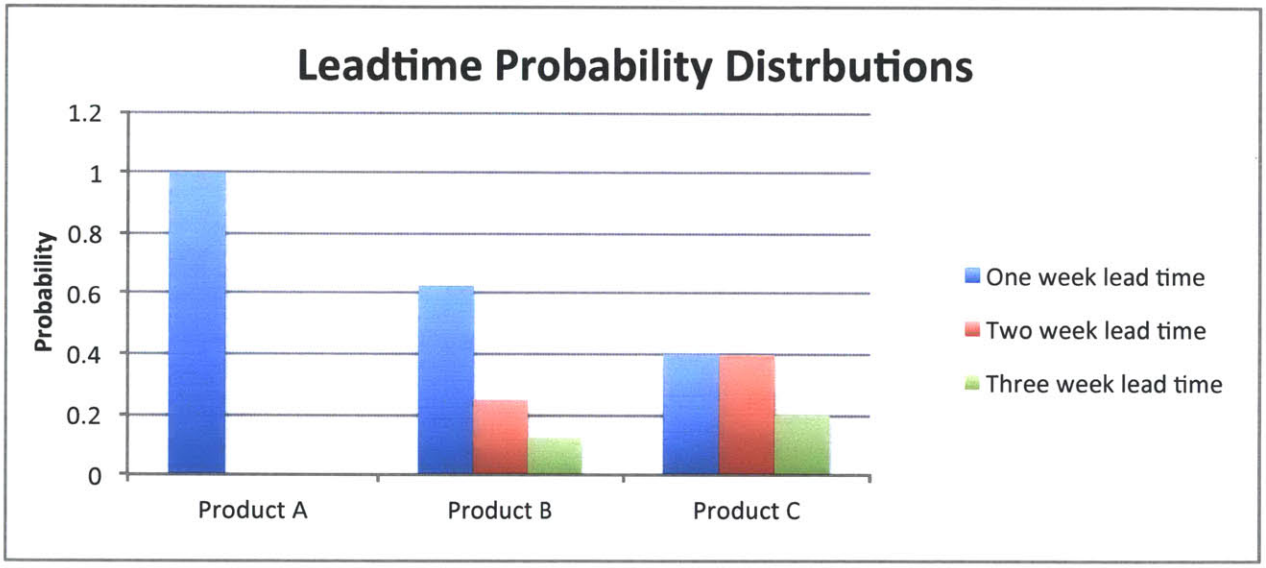

Figure 13: Lead time probability distribution by product

5.3 Results and Discussion

This method results in a distribution of historical lead times for each product received. The results from the sample data in Section 5.4 is shown in Figure 13. In this example Product A has a very reliable supply chain, in which it arrives every time it is ordered. Product B usually arrives when it is ordered, but has some probability of not arriving for two or three weeks. Product C has a very unreliable supply chain, where it will often take two or more orders until it is fulfilled.

Like the sample results above, some items are found to have reliable supply chains, with few supply disruptions. Others may have occasional short term disruptions, caused by individual

suppliers going temporarily out of stock. However, a third category of products exists, where long term stockouts across suppliers were more common, indicating fragility in the upstream supply chain. According to interviews, the anticipation of long term stockouts was a leading driver of inventory managers to keep additional inventory. However, once quantified, the number of items in this third category was lower than expected by inventory managers. This resulted in unnecessary over-ordering of supplies in some categories.

5.4 Conclusion

The fragility of the supply chain was known to lead to supply disruptions, but this data was not effectively used to manage inventory. Converting fill rates to a lead time distribution results in a more familiar inventory management problem. This distribution is then used to feed into the inventory model that is developed in Chapter 7.

6 Perishability

Although other factors were founds to be key variables necessary to characterize in this situation, the perishability of products was still the leading concern of inventory managers. Because products outdate at unknown intervals, inventory managers held additional inventory to protect against unforeseen shrink events. By using existing data, a probability model can be built that can be used to forecast shrink and drive proactive ordering when items are likely to soon outdate.

6.1 Key Considerations

The key challenge of this problem is in the perishability of the products. As noted in Chapter 3.1.5, there are multiple ways a product can deteriorate. Understanding the basic method of deterioration is key to determining the method with which to characterize deterioration rate. For this particular type of inventory, individual units expire stochastically—there is no known fixed date for perishability, and it can vary within a lot of items. Additionally, it has been observed that the rate is variable over time. Rather than a relatively stable proportion of items expiring each period, much more common is the situation where products deteriorate slowly initially, but this deterioration accelerates over time. This observation would eliminate an exponential decay model, and favor a Weibull or gamma distribution. However, Tadikamalla [10] notes that similar looking distributions generate very different instantaneous rates of change. This implies that incorrect choice of model can lead to significant inaccuracies in predicting decay rates. Additionally, with a large number of products to categorize, it is possible that different products may have different underlying models (including models not previously studied in this context). Thus, a non-parametric approach was found to be most applicable in this context.

6.2 Model Development

One challenge in developing a decay model is that only a portion of data is explicitly captured: we only know the full shelf life of items that expire while in inventory. For many products, this only captures a fraction of the total items. Additionally, utilizing only this data may lead to misleading assumptions. An extreme, though not entirely unrealistic, example may be of a product that sells at a relatively high velocity. The length of time the item is in inventory is shorter than its shelf life, established by the fact that few items expire while in inventory. However, due to the stochastic nature of that product's perishability, a few items outdate while in inventory. Let's assume that 10 days of inventory are typically held for this product, with few items discarded due to deterioration. With an average time in inventory being 10 days and a steady number of units sold per day, that would imply that an individual unit may be in inventory for up to 20 days, with the minimum amount of time being 0 days. Average demand is five units a day, so given a basic EOQ model, an order of 100 units is placed every 20 days. However, there are three units that, upon inspection, were discarded due to deterioration. Those items expired while in inventory at 5, 12, and 18 days.

Age	Units Sold	Units Discarded
1	5	0
2	5	0
3	5	0
4	5	0
5	5	1
6	5	0
7	5	0
8	5	0
9	5	0
10	5	0
11	5	0
12	5	1
13	5	0
14	5	0
15	5	0
16	5	0
17	5	0
18	5	1
19	5	0
20	5	0

Table 6: Sample sales and discard data

Using only the data points where the full shelf life is known, only three data points are used. This would imply that this particular product has a 33% probability of expiration at 5 days, 67% probability of expiration by 12 days, and 100% probability of deterioration at 18 days. A look at reality would show why this would not be true. Out of 100 units, 40 units are sold after being in inventory more than 12 days, yet the probability model we just constructed says that 67% of our inventory should expire by the time it has aged 12 days.

A better approach is to look at the total inventory in stock at a given age. For example, at 12 days, there are 39 units still in stock (12 days demand, with 5 units sold per day, plus one unit that outdated at 5 days). However, on day 12, one additional unit is discarded, so 38 out of 39 units are eligible for sale on day 13. Thus, for a product that has made it to day 12, the

probability of making it to day 13 is 38/39 or 97.4%. This can be calculated for each day, utilizing the available number of items in inventory at that day. To arrive at the condition that a product has survived to day i, you can multiply the probabilities of each preceding day. Thus the probability of surviving to day 4 is the probability that the product has survived day 1, and the probability the product has survived day 2, and the probability the product has survived day 3.

These dynamics are captured in the Kaplan Meier estimator [14] which builds a survival curve using the equation

$$S(t) = \prod_{t_i \leq t} \frac{n_i - d_i}{n_i} \qquad (19)$$

S(t) = the probability of survival through period t

n_i = the number of subjects at the beginning of period i

d_i = the number of "deaths" during period i

Kaplan Meier is useful because it accounts for items that are sold before they expire. Although we do not have expiry data on these items, they do account for items that are at least age n. This is known as censored data--specifically right side censorship, where the final end date is unknown (left side censorship refers to when the start date is unknown). For the above example, the survival curve, which plots the probability of survival up to a given age, is shown in Figure 14, compared to the naïve approach, utilizing only the three data points of deterioration.

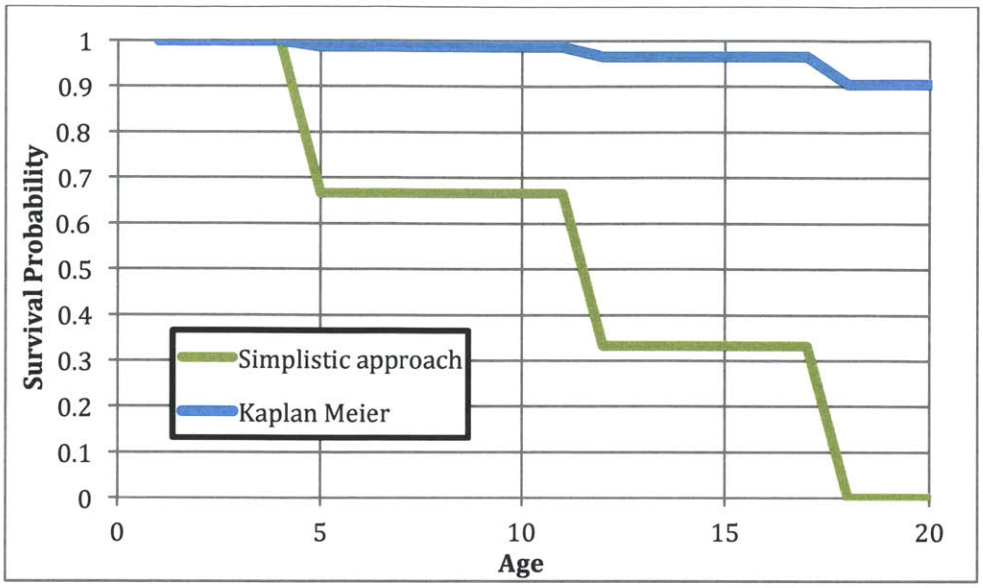

Figure 14: Comparison of survival probabilities using a simplistic approach and Kaplan Meier

This provides a empirical, non-parametric probability of survival over time. This probability represents only the data we have available, which is a sample of the overall population. In other words, we just constructed a survival distribution for a single lot of widgets, but it does not necessarily precisely represent all widgets in the world. However, we would like more insight into the overall population based on the sample information that we have.

6.3 Bootstrapping for Variance

Without knowing anything about the underlying distribution of the population, one technique that could be applied to the approximate the dataset is bootstrapping. As outlined in Diaconis and Effron [15], bootstrapping is a method enabled by computing technology to build a probability distribution of a sampled statistic. This is done by *resampling with replacement*. Suppose all given data points were placed in a box. One data point is taken from the box at a time, recorded, then returned, allowing the possibility for that point to be chosen again. This is repeated until the same number of points as the original sample has been taken, and the statistic

in question is calculated. This process is repeated a large number of times (say, 10,000). This gives a measure probability distribution of that statistic for the entire population.

A sample set is shown in Table 7. The original data set is shown in the left, containing 18 rows of data. On the right, the data is resampled four times, each time taking 18 rows of data from the original. Some data points may be repeated, while others may not appear in a resampled set. For each set, a new survival curve can be drawn, as shown in Figure 15.

Table 7: Bootstrapping data example

Row	Count	Age	Event
1	1	2	sale
2	30	3	shrink
3	1	4	sale
4	1	7	shrink
5	1	7	sale
6	1	8	sale
7	1	8	sale
8	1	10	shrink
9	4	11	shrink
10	4	12	shrink
11	1	12	sale
12	1	12	shrink
13	3	12	sale
14	2	13	sale
15	1	14	sale
16	2	15	shrink
17	1	15	sale
18	3	15	shrink

Original	Sample 1	Sample 2	Sample 3	Sample 4
1	16	10	5	16
2	15	5	17	11
3	5	9	16	14
4	2	1	4	3
5	13	17	16	7
6	9	18	7	13
7	12	8	14	13
8	5	4	16	10
9	6	15	17	1
10	15	15	17	11
11	4	14	15	4
12	7	5	5	3
13	11	14	12	17
14	14	17	5	11
15	8	4	10	1
16	8	4	8	15
17	8	1	11	13
18	18	8	12	4

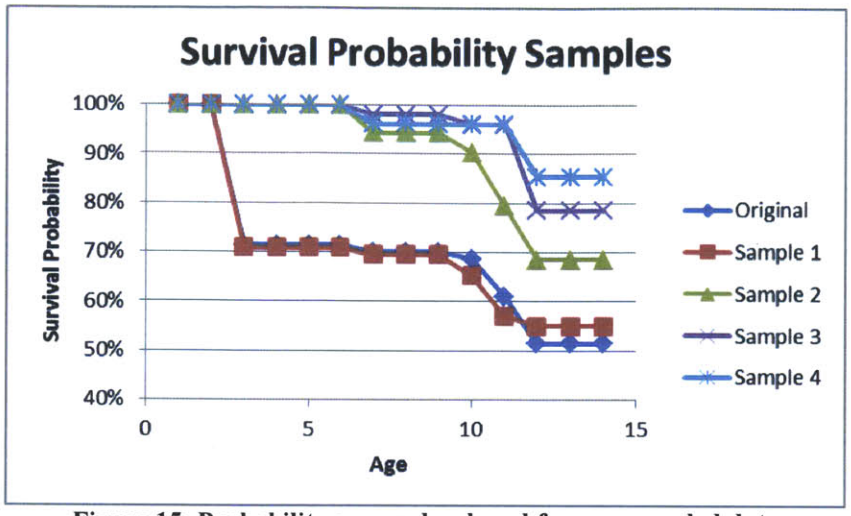

Figure 15: Probability curves developed from resampled data

The original data, shown in the blue diamonds, shows relatively low survival after age 3. However, this appears to be due to a single, possibly outlier, data point in row 2. When the data is resampled several times, this data point is removed and replaced by other data points, showing new possible survival curves. When repeated a larger number of times, these survival curves represent the range of possible actual survival probabilities given the data.

6.4 Conclusion

Utilizing the Kaplan Meier estimator for products allows us a method for capturing the survival probability of an item at a given age, utilizing both explicit data as well as censored data. Additionally, bootstrapping is used to develop a range of probabilities at a given age. This distribution, along with the distributions developed in Chapters 4 and 5, are then used in the inventory model that is outlined in Chapter 7.

7 Inventory Modeling

With they key variables characterized into their respective probability distributions, we can pull this information together to determine an optimum inventory strategy. Because lot sizes are fixed and there is assumed no per-order charges, they key variable is then the reorder point (ROP) of a product. The reorder point needs to account for the lead time variability, expected demand, and perishability of the item at its current age. This is done though Monte Carlo simulation, which samples from each distribution to calculate the reorder point needed to cover each scenario. The reorder point form the resulting distribution at the given service level is then used.

7.1 Key Considerations

With distributions now developed for demand, lead time, and survivability, a model can be created to determine the correct inventory strategy. First, several key properties of the supply chain and current process must be considered as we construct this model.

- With database managed systems, inventory is effectively continuously reviewed, however, orders can only be placed periodically
- There are no fixed order costs, or order costs are negligible
- Order quantities are fixed by case quantities
- Assume:
    - Shrinkage occurs before demand is fulfilled
    - New orders arrive at the beginning of the period
    - Inventory position is reviewed at time t, at the beginning of the period, after any new orders have arrived

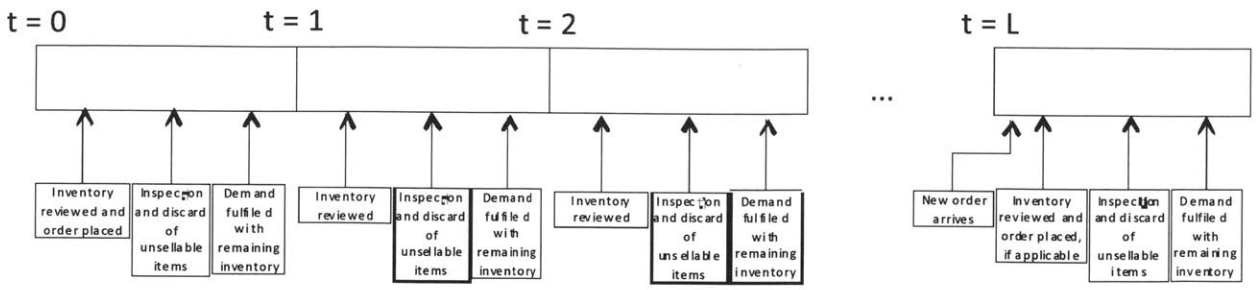

Figure 16: Relative timeline of inventory events

7.2 Initial Model Formulation

With a lack of fixed order cost and order quantities fixed, traditional EOQ approaches are not immediately relevant here. Typical EOQ calculations balances holding costs with the fixed order cost to find the minimum total cost. With no fixed order cost, holding cost should then be kept to a minimum—this is done by keeping minimum inventory and immediately ordering a replacement unit when one is sold, since there is no penalty for multiple orders.

What is minimum inventory? First, let's consider a system with no perishability, deterministic lead time, and deterministic demand. Additionally, demand during lead time is greater than the order quantity required by case size. In a continuous review system, an order is triggered when inventory position falls below the demand during lead time. This will mean that just as inventory reaches zero, the next shipment to replenish supply arrives. In a periodic review system, a new order is placed when inventory position falls below demand for lead time plus the review period.

In a stochastic system, we set a service level, and set the reorder point to the demand during lead time at that service level, as seen in Figure 5. Because both demand and lead time are stochastic, we need to combine these distributions into a single distribution of demand during lead time. When both demand and lead time are normally distributed, the reorder point takes the following form [7]:

$$R = \mu_D * \mu_L + z\sqrt{\mu_L \sigma_D^2 + \mu_D^2 \sigma_D^2} \quad (20)$$

Because our distributions are non-parametric, no closed form solution is possible. Instead, this must be found numerically using Monte Carlo methods.

Additionally, we need to consider the effect of perishability. At the beginning of the period, inventory is age a. At the beginning of the next period, it is age a + 1. Some percentage of items will expire between a and a + 1. In order to determine this percentage, we consider this in probabilistic terms: given that an item has survived to age a, what is the probability the item will age a + 1? In phrasing it in this way, we can apply Bayes theorem in determining this probability.

Bayes theorem states:

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)} \quad (21)$$

For this case, P(A) is the probability that the product survives to age a + 1, and P(B) is the probability the product survives to age a. P(B|A) is the probability that product survives to age a given the product survives to age a + 1, which is 1. Thus we get that the probability an item survives from age a to age a + 1 is the probability the product survives to age a + 1 divided by the probability the product survives to age a. Here we will refer to that value, the survival probability from age a to age a+1 as s(a), or the one period survival rate. From Equation 23 we can also see that the lifetime survival probability S(t) is the product of the past individual period survival rates.

$$s(a) = \frac{S(a+1)}{S(a)} \quad (22)$$

$$S(t) = \prod_{i=1}^{t} s(i) \quad (23)$$

To begin to formulate this problem, let us consider a sample product with time-varying demand, lead time, and perishability, all of which are deterministic. Lead time is three periods (LT=3), inventory review occurs every period (RP=1) and demand, age, and survival probabilities are shown in Table 8, with the overall survival curve shown in Figure 17.

Period	Demand	Age	Survival S(a)	One period survival rate s(a)
0		4	0.94	
1	12	5	0.89	0.95
2	20	6	0.79	0.89
3	10	7	0.63	0.80
4	16	8	0.44	0.70

Table 8: Demand and survival data for sample item

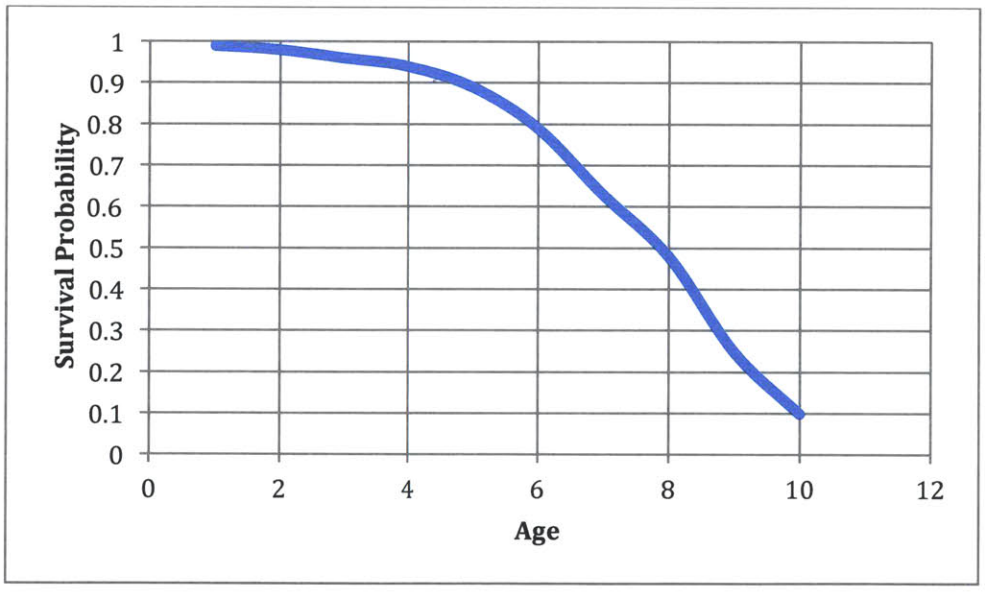

Figure 17: Sample product survival curve

As outlined in section 7.1, inventory is reviewed and perished items are discarded before items are shipped. Assuming we start with 100 units at the beginning of period 1, we can show the change in inventory at each step, as seen in Table 9.

Period	Stating inventory	Discarded units	Inventory after discard	Demand	Inventory after demand is fulfilled
1	100	5	95	12	83
2	83	9	73	20	53
3	53	11	43	10	33
4	33	10	23	16	7

Table 9: Inventory level changes for sample inventory

With the starting inventory equal to the ending inventory of the previous period, we can represent the ending inventory for each period as:

$$EndInv_t = EndInv_{t-1} * s(a_t) - D_t \quad (24)$$

Similar to a typical periodic review system, the reorder point is set to cover demand plus discarded units for lead time plus review period. In this example, that would be for four periods. In this example, we can see whether we need to place an order at period 1—because there is still positive inventory remaining until the end of period 4, we know we can wait at least one additional period to place the order. Doing so will minimize holding costs.

The key question to answer is, at what level of inventory should an order be triggered; that is, what value of R (reorder point) should we set? This would be when the ending inventory at the end of review and lead time period is less than zero. At this point, waiting until the next review period to place an order will result in a stock out, so an order should be placed during the current period. This requires finding what value R at $StartInv_1$, results in $EndInv_{RP+LT}=0$, where RP+LT is the review period plus lead time. For a four-day review period plus lead time, Equation 24 expands to:

$$\Big(\Big(\big(R^* s(a_1) - D_1\big)s(a_2) - D_2\Big)s(a_3) - D_3\Big)s(a_4) - D_4 = 0 \quad (25)$$

Equation 25 can be rearranged to:

$$R = \frac{D_1}{s(a_1)} + \frac{D_2}{s(a_1)s(a_2)} + \frac{D_3}{s(a_1)s(a_2)s(a_3)} + \frac{D_4}{s(a_1)s(a_2)s(a_3)s(a_4)} \quad (26)$$

This can be generalized to:

$$R = \sum_{i=1}^{RP+LT} \left\lfloor \frac{D_i}{\prod_{k=1}^{i} s(a_k)} \right\rfloor \quad (27)$$

For the example in Table 8, this value is 86 units. Because the demand is time varying and the age of the products is constantly changing, the value for R is also constantly changing, depending on the current conditions. R must be recalculated compared to inventory position at each review period.

7.3 Monte Carlo Simulations

Equation 27 provides a formulation for a reorder point in this context. However, as outlined in this paper, lead time, demand, and survival rate are not known deterministically. Thus, in order to determine the correct reorder point, we must simulate the possible permutations of reorder point needed given the established probabilities for lead time, demand, and survival. This can be done using Monte Carlo simulation methods. In a Monte Carlo simulation the following steps are performed:

1. Define a domain of possible inputs.
2. Generate inputs randomly from a probability distribution over the domain.
3. Perform a deterministic computation on the inputs.
4. Aggregate the results.

Sections 4-6 in this paper reflect steps 1 and 2, where they key parameters (demand, lead time, and perishability) are identified and characterized into respective probability distributions. Step 3 is done using Equation 27. Step 4 will result in an overall probability distribution for the reorder point needed to cover possible demand and shrinkage scenarios over possible lead times.

From this probability distribution, a reorder point is chosen at the given service level to provide a high confidence probability that that reorder point will result in ordering patterns that allow only the acceptable percentage of stock-outs.

In the Monte Carlo simulation, the following computational steps are performed:

1. A lead time value (LT) is sampled from the lead time distribution developed in section 4.8.
2. The review period (RP) is added, resulting in the full period needed to be considered.
3. The forecasted demand is calculated for each period in the LT+RP.
4. Variation is added to each period's demand in step 3 from the residual distribution developed in section 4.8.
5. With age of inventory known at the time of review, the single period survival rate for each period in LT+RP is sampled from the bootstrap distribution in section 6.3.
6. Reorder point R is calculated using Equation 27.
7. Steps 1-6 are repeated a large number of times. The author recommends a minimum of 100 runs for consistent results, but 10,000 is recommended.
8. From the resulting probability distribution, chose the point R where the cumulative distribution function (CDF) is equal to the desired service level.
9. Compare current inventory position (current inventory plus orders already placed) to reorder point; if inventory level is below the reorder point, place an order.

A visual representation of this process is shown in Figure 18.

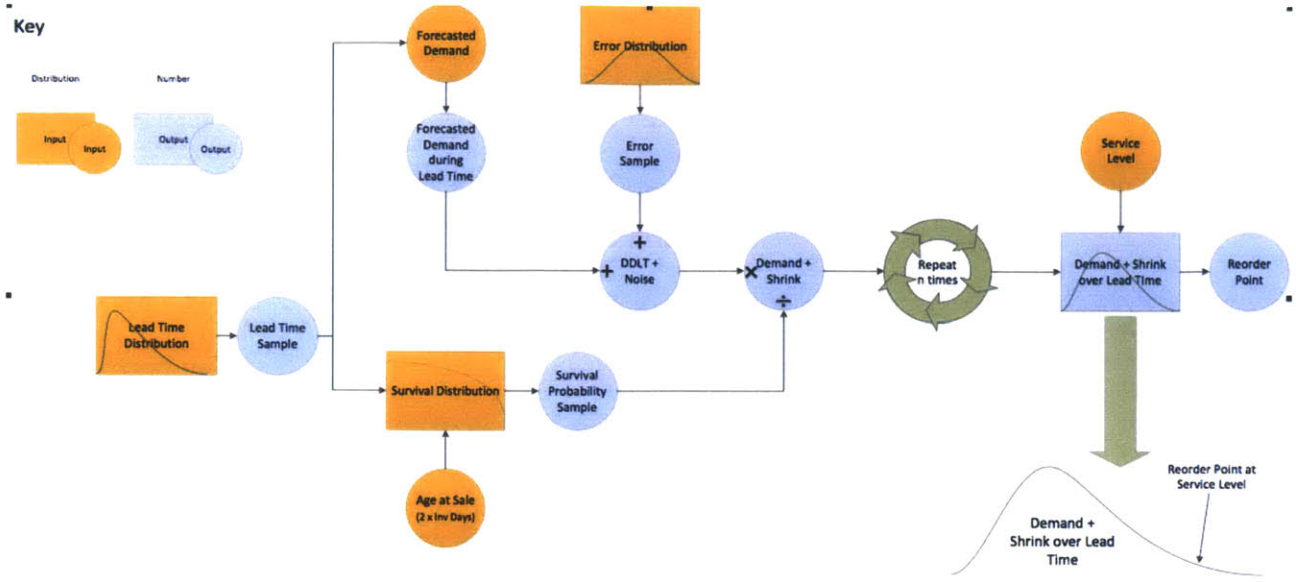

Figure 18: Process for simulating reorder point

7.4 Results Validation

In order to test the methods described above, reorder points were calculated from known data at a given point in time (simulation date). Data from before the simulation date was used to train the model, developing the probability distributions needed to run the Monte Carlo simulation, and data from after the simulation date was used to compare the data. However, a true comparison of results is impossible for the following reasons:

1. With reorder points changing, the age of items in inventory also changes; is it impossible to compare absolute shrink values during the evaluation period, because of the dependency of shrink to the age of the product
2. Additionally, inventory levels change. Because shrink levels are assumed to be proportional to inventory levels, shrink counts should reduce as inventory levels are reduced
3. Because orders may be triggered at different times, there is no actual order (and thus order lead time) to compare to each simulated order

Instead, data following the simulation date was used to create a new set of lead time distributions. Because of limited data with which to create new survival curves, absolute values for shrink per period were kept; however, data points of expired products older than the maximum expected age in the simulation were discarded. This provides a conservative account of shrink; assuming inventory levels are reduced, the number of units that are discarded also proportionally reduce (when controlled for age).

Inventory levels were calculated, and demand (actual) and shrink (sampled) were subtracted from inventory by period. If inventory was below the calculated reorder point, an order was triggered. A lead time was sampled from the new lead time distribution, and the appropriate number of days were taken until it arrived into inventory. Inventory levels by day, as well as days where the item was out of stock, were tracked.

7.5 Results Discussion

Using this method, service level is an input to the simulation to determine reorder point. Given the calculated reorder point, a simulated service level can be calculated based on actual demand following the simulation date. We can compare the inputted service level to the simulated output service level for both the original order method (which does not account for lead time variability and shrink) as well as the new order methods, utilizing both original and improved forecasts from section 3.

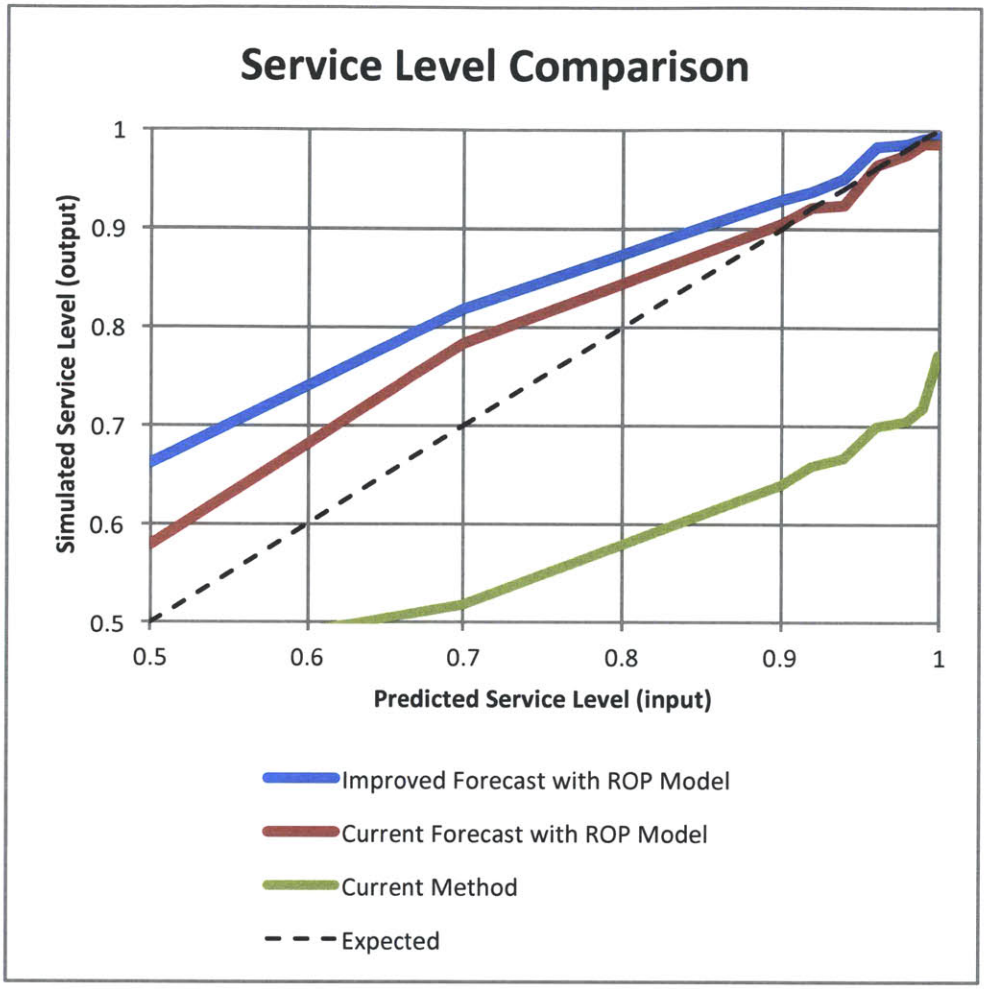

Figure 19: Predicted vs. simulated service levels

At low service levels, less than 90%, this method utilizing the new ROP calculation overperforms and provides actual service levels above the predicted service levels. From 90-99%, both forecast methods with the new ROP generally converge towards the expected values. This compare to the naïve approach, or current method, where perishability and lead time variability were not taken into account. In this case, the problem simplifies to a newsvendor-type model, utilizing only demand variation. Even when expected service level was set at high levels, increasing the reorder points and total amount of inventory, results fell far below expected values for service level. This result was observed in practice, as stock-outs occurred regularly even when service levels were intended to be high. This result indicates that not considering

perishability and variability in lead time in inventory policy will result in significant underperformance.

Another measure of performance in the model is to compare simulated service level with inventory levels. This provides a measure of efficiency and effectiveness of the model. In general, high inventory should result in a high service level. However, a well performing model will offer a higher service level for the same level of inventory, or lower inventory to meet a specific service level. Because the simulation was not able to achieve high service levels for the current method, we can focus on areas where results overlap and we can compare between models.

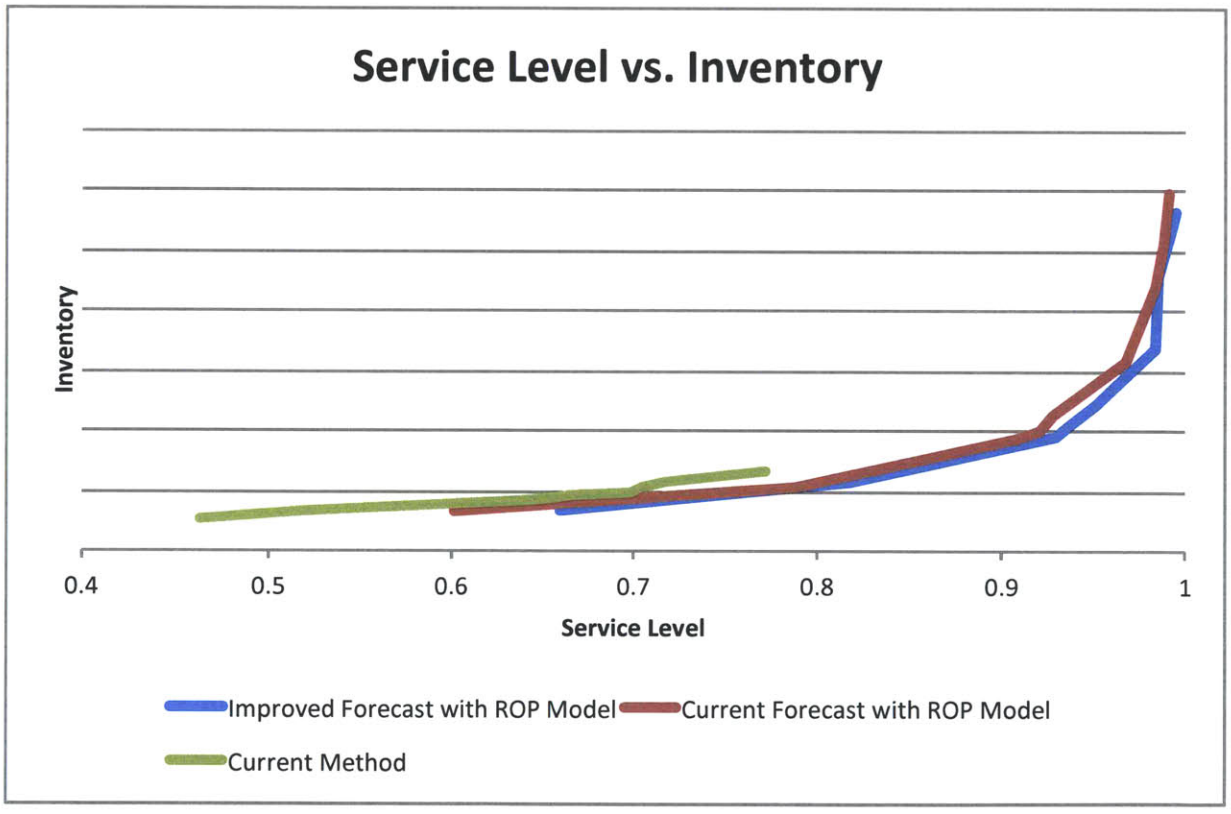

Figure 20: Service level vs. inventory

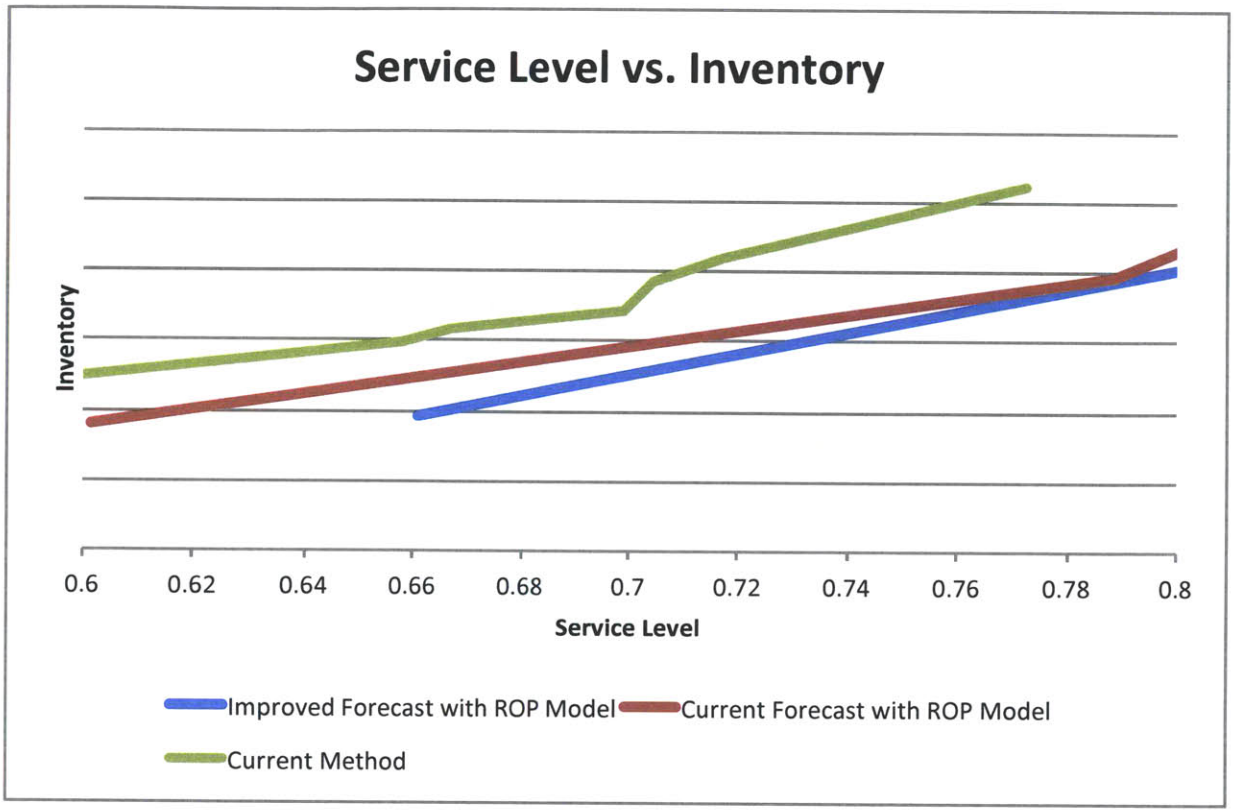

Figure 21: Service level vs. inventory (zoomed)

To achieve the same service level, the current, naïve approach requires more inventory than the new model. For example, at a 75% service level, the difference in inventory between the improved forecast and current forecast using the ROP model is 5%. However, the difference between the inventory levels for the current method and the ROP model, both utilizing the same demand forecast, is 25%. This indicates that the choice of forecast method is a relatively minor difference compared to the inclusion of lead time variability and perishability.

Inventory level differences can have important ramifications. First, because perishable inventory must be regularly inspected, the holding cost for inventory tends to be significantly higher for perishable goods than non-perishable goods. Additionally, higher inventory levels means a individual unit is held in inventory for a longer period of time. This can result in

increased shrinkage costs due to expiration. Thus, efficient inventory management is particularly important in dealing with perishable goods.

7.6 Conclusion

This section provides an overview of the method used to determine the reorder point for a perishable inventory system, utilizing the distributions derived in previous sections of this thesis. This model, which fully integrates lead time variability, perishability, as well as improved forecasting, provides more effective management of inventory, leading to higher service levels, and less inventory needed to meet those service levels.

8 Conclusions and Recommendations

This section will discuss overall findings and ramifications of the results of this thesis, as well as propose further areas of research.

8.1 Discussion

This thesis discusses a method for determining an inventory reorder point where multiple inputs are variable and inventory in perishable. Three key variables were identified as having significant impact on the inventory levels needed: customer demand, lead time variability, and product perishability. With each variable being stochastic, a probability distribution was developed using existing data. These distributions were then used in a Monte Carlo simulation to determine the optimal reorder point for this inventory system.

While much past research has been done on the field of perishable inventory in various situations, these methods are typically only applicable to very specific situations, often where the inputs are clearly parameterized. However, in reality, it is difficult, if not impossible, to identify the correct models and parameters, particularly in a rapidly changing and developing system. Thus, this method provides the opportunity to apply inventory theory to a noisy, less than ideal system and optimize inventory levels to a given service level. Results show that this method provides a reliable method of meeting intended service levels, while minimizing costs associated with holding the item in inventory.

In the field of perishable goods, optimizing inventory levels is especially beneficial. While a service-oriented company may typically keep additional inventory to meet customer need, this approach may not be appropriate for perishable goods. Because items may require more inspection than non-perishable goods in order to ensure its fitness for sale, holding costs can be particularly high. Additionally, high inventory levels typically indicate a longer length of time on

the shelf before it is sold. This results in a product aging more, increasing the probability of perishability. Thus, shrink rates may also increase. Keeping unnecessary additionally inventory may not only be costly, but ineffective as well.

8.2 Recommendations for Future Research

This research opens up several questions that require further research. First is in understanding the optimal level of service to provide. As mentioned in the discussion, high service levels are costly not only due to high holding costs, but are likely to result in high shrink costs as well. However, optimizing for the right service level requires understanding the true costs for an item being out of stock. Not only does this result in a lost sale (which in itself is difficult to track), but perhaps lost consumer confidence and so future sales may also be lost. This type of analysis may be impossible in a brick and mortar institution; however, this may be possible for an online retailer due to information on customer viewing and purchase histories.

Additionally, while the model is flexible in that it can accommodate any distribution of lead time and perishability, it can be further extended to additional situations. For example, order costs were assumed to be zero in this analysis; however this may not be an accurate assumption. Another common constraint is a minimum order quantity. In these cases, batching orders may be required. Perishability can play an important role—items with slower decay rates may have an advantage to being ordered in larger, fewer batches. This analysis also assumes that unit costs are constant. Additional studies should be performed to understand the effect of volume discounts. While unit costs go down, holding costs and shrink may increase, so it is unclear if this is an acceptable trade-off. How much should an organization pay to invest in flexibility in this case?

Another area of study is in analyzing the benefit or tradeoff of additional simulation runs to results. This includes how much history to include in building initial distributions, how often

should these distributions be updated, how many bootstrap runs should be used in building perishability distributions, and how many Monte Carlo simulations should be run for reorder point calculation. When the number of SKUs becomes large, the computation time can become unmanageably long and require costly computing bandwidth. At some point the marginal improvement of additional computation runs is diminishing; however it is currently unclear how much is "good enough." Depending on the consistency and fidelity of the data, this value will likely vary across organizations.

Bibliography

[1] S. Nahmias, *Perishable inventory systems*. 2011.

[2] D. Vellante, "Cloud Computing 2013: The Amazon Gorilla Invades The Enterprise - Wikibon." [Online]. Available: http://wikibon.org/wiki/v/Cloud_Computing_2013:_The_Amazon_Gorilla_Invades_the_Enterprise. [Accessed: 17-Feb-2014].

[3] L. Leong, "Toolkit: Comparison Matrix for Cloud Infrastructure as a Service Providers, 2013." [Online]. Available: https://www.gartner.com/doc/2575815. [Accessed: 17-Feb-2014].

[4] J. Clark, "How Amazon exposed its guts: The History of AWS's EC2 | ZDNet." [Online]. Available: http://www.zdnet.com/how-amazon-exposed-its-guts-the-history-of-awss-ec2-3040155310/. [Accessed: 17-Feb-2014].

[5] E. Silver, "Operations research in inventory management: a review and critique," *Oper. Res.*, 1981.

[6] F. Raafat, "Survey of Literature on Continuously Deteriorating Inventory Models," *J. Oper. Res. Soc.*, vol. 42, no. 1, pp. 27–37, 1991.

[7] E. A. Silver, D. F. Pyke, and R. Peterson, *Inventory Management and Production Planning and Scheduling*. Wiley, 1998, p. 784.

[8] S. K. Goyal and B. C. Giri, "Recent trends in modeling of deteriorating inventory," *Eur. J. Oper. Res.*, vol. 134, no. 1, pp. 1–16, 2001.

[9] R. Covert and G. Philip, "An EOQ model for items with Weibull distribution deterioration," *AIIE Trans.*, 1973.

[10] P. R. Tadikamalla, "An EOQ inventory model for items with gamma distributed deterioration," *A I I E Trans.*, vol. 10, no. 1, pp. 100–103, Mar. 1978.

[11] P. Ghare and G. Schrader, "A model for exponentially decaying inventory," *J. Ind. Eng.*, 1963.

[12] H. Wagner and T. Whitin, "Dynamic version of the economic lot size model," *Manage. Sci.*, 1958.

[13] E. Silver and H. Meal, "A heuristic for selecting lot size quantities for the case of a deterministic time-varying demand rate and discrete opportunities for replenishment," *Prod. Invent. ...*, 1973.

[14] M. K. Goel, P. Khanna, and J. Kishore, "Understanding survival analysis: Kaplan-Meier estimate.," *Int. J. Ayurveda Res.*, vol. 1, no. 4, pp. 274–8, Oct. 2010.

[15] P. Diaconis and B. Efron, "Computer-intensive methods in statistics," *Sci. Am.*, 1983.