# Reducing Total Fulfillment Costs through Distribution Network Design Optimization

by

David Guasch Rodríguez

B.S. Mechanical Engineering, Universidad Pontificia Comillas, 2006
B.S. Industrial Engineering, Ecole Centrale Paris, 2006

Submitted to the MIT Sloan School of Management and the Department of Mechanical Engineering 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 2013

Signature of Author........................................................................................................
Department of Mechanical Engineering, MIT Sloan School of Management
May 10, 2013

Certified by ........................................................................................................
Stanley Gershwin, Thesis Supervisor
Senior Research Scientist, Department of Mechanical Engineering

Certified by ........................................................................................................
Stephen Graves, Thesis Supervisor
Professor of Management Science, Department of Mechanical Engineering and Engineering Systems

Accepted by........................................................................................................
David Hardt
Associate Department Head, Mechanical Engineering

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

*This page intentionally left blank.*

# Reducing Total Fulfillment Costs through Distribution Network Design Optimization

by

David Guasch Rodríguez

Submitted to the MIT Sloan School of Management and the Department of Mechanical Engineering on May 10, 2013 in Partial Fulfillment of the Requirements for the Degrees of Master of Business Administration and Master of Science in Mechanical Engineering

## Abstract

Compared to legacy retailers, online retailers have the potential to better accommodate buyer needs by offering more service time and inventory options. One fundamental operational challenge faced by most online businesses is designing a cost effective distribution network. Based on a fixed number of locations with finite resources, companies strive for finding the cost minimizing formula for fulfilling each customer order while meeting rigorous time constraints. In practice this involves allocating specific geographies to each warehouse and defining the logistic routes serving each customer. In an attempt to address this question, a Mixed Integer Linear Programming model has been developed as a decision-making tool for determining the optimal carrier-destination combination at each facility. The resulting algorithm is capable of analyzing thousands of potential shipping lanes and selecting those that minimize overall shipping cost. Based on historical data from customer orders, the model consistently finds an optimal network configuration yielding operational savings on the order of 1.5%. Furthermore, the algorithm can be used to identify near-optimal solutions requiring minor tweaks on the current configuration that produce significant economic gains. This simulation tool can be used on a regular basis to adapt the outbound network to demand fluctuations. However, this phenomenon evinces the existence of a fine trade-off between economic gains and operational feasibility. For that reason, a heuristic for selecting the most robust solution is also proposed.

Thesis Supervisor: Stanley Gershwin
Title: Senior Research Scientist
MIT Department of Mechanical Engineering

Thesis Supervisor: Stephen Graves
Title: Professor of Management Science
MIT Department of Mechanical Engineering and Engineering Systems Division

*This page intentionally left blank.*

## Acknowledgments

I would like to thank the Leaders for Global Operations (LGO) program at MIT and Amazon for giving me the opportunity to carry out this project. Within Amazon, special mention goes to my supervisors Adam Baker and Mingang Fu for their guidance and unconditional support throughout the duration of the internship. In addition, I would also like to acknowledge Russell Allgor, John McDonald and Akshay Katta for their time and dedication, without their expertise and sharp insights this work would not have been possible.

From MIT, I would like to acknowledge my academic advisors, Stephen Graves and Stanley Gershwin, for their contribution has been of great importance for the development of this thesis. Specifically, I most obliged to them for offering their invaluable advice and providing me with the analytical fundamentals used in the development of this project. By constantly challenging my assumptions and asking the right questions I was able to overcome every major obstacle.

Finally and most importantly, I am most grateful to my beloved wife Almudena and my daughter Elisa. Their sacrifice and unconditional encouragement in difficult times have motivated me to do my best. Looking back into these two past years at LGO, it would only be fair to say that Almudena deserves the merit of my academic and professional successes.

*This page intentionally left blank.*

# Table of Contents

## List of Figures

# 1 Introduction

The purpose of this thesis is to develop an analytical approach for designing a cost effective distribution network. The resulting decision-making tool evaluates different carrier options, shipping routes and regional markets to produce an optimal configuration that meets customer expectations and geographic constraints. This research has been conducted as part of an internship in collaboration with Amazon.com and the Leaders for Global Operations program from the Massachusetts Institute of Technology. The following is a brief overview of the company, problem statement and project goal, which will provide the reader with the imperative background for understanding the problem at hand.

## 1.1 Project Overview

In 2011, the global online retail industry had revenues of $530.2 billion[1]. By 2016, e-commerce is expected to be worth $1,096 billion. Just in the US, online sales are expected to increase form $142 billion in 2010 to $279 billion by 2015, representing a 10% compound annual rate[2]. Similarly, in the EU the online market will grow from €83 billion to €134 billion during the same time period[2]. For China and India, growth is expected to be much faster, with China overtaking the US as the biggest internet retailing market by 2015 with $314 billion in sales[3].

In the e-commerce market, Amazon stands out as the world's leader in online retailing. In 2011, its revenue amounted to $48.1 billion[4]. For all those sales, the company incurred $4.6 billion in fulfillment costs[4]. *Fulfillment* is defined as the process by which an order is picked, packed and shipped to the customer. Given that fulfillment costs represent around 9.5% of total revenues while profit margins for large online retailers are in the order of 2% to 5%[5], any reduction on fulfillment costs no matter how little will have an impact in profitability. This study looks at a portion of Amazon's supply chain, namely its

[1] Global Online Retail 2011 report by MarketResearch.com.
[2] Forrester Research, 2012.
[3] Bloomberg News, 2011.
[4] Amazon 10k, 2012.
[5] Seeking Alpha, 2012.

distribution network, and proposes analytical methodology for cutting down outbound transportation costs. Before moving into the problem per se, it is necessary to grasp the basic context surrounding logistics and operations of this business segment.

The operational principles of online retail businesses differ substantially from those of conventional retailers. Principle among these is their supply chain structure. In a traditional retail business supply chain, vendors supply distribution centers, which in turn supply physical selling points, namely stores. Customers travel to these selling points to purchase inventory on the shelf. In this business model, as long as there is enough inventory available, buyers are served immediately when checking out at the cash register.

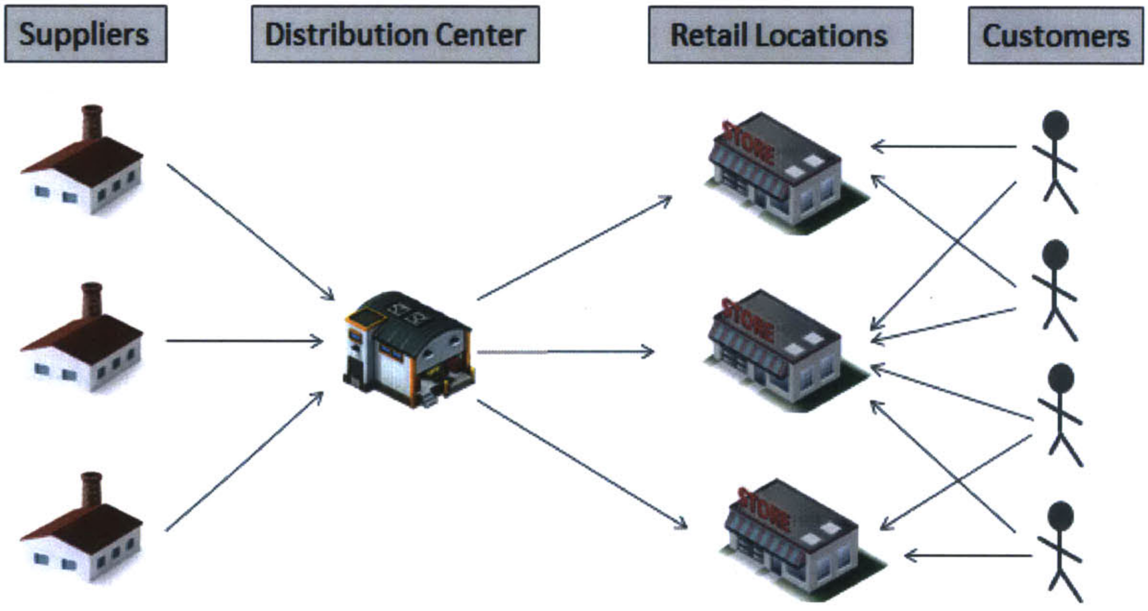


**Figure 1: Traditional retailer supply chain structure.**

Albeit the upstream network structure is very similar, it is the downstream portion that makes an e-commerce so different. Instead of having a central warehouse or distribution center, online retailers have fulfillment centers which hold different items or Stock Keeping Units (SKU's) awaiting to be processed in the form of a customer order. These fulfillment centers are generally located in strategic sites that maximize geographic coverage. In addition, inventory is usually not grouped by product family, but rather by size, weight and popularity.

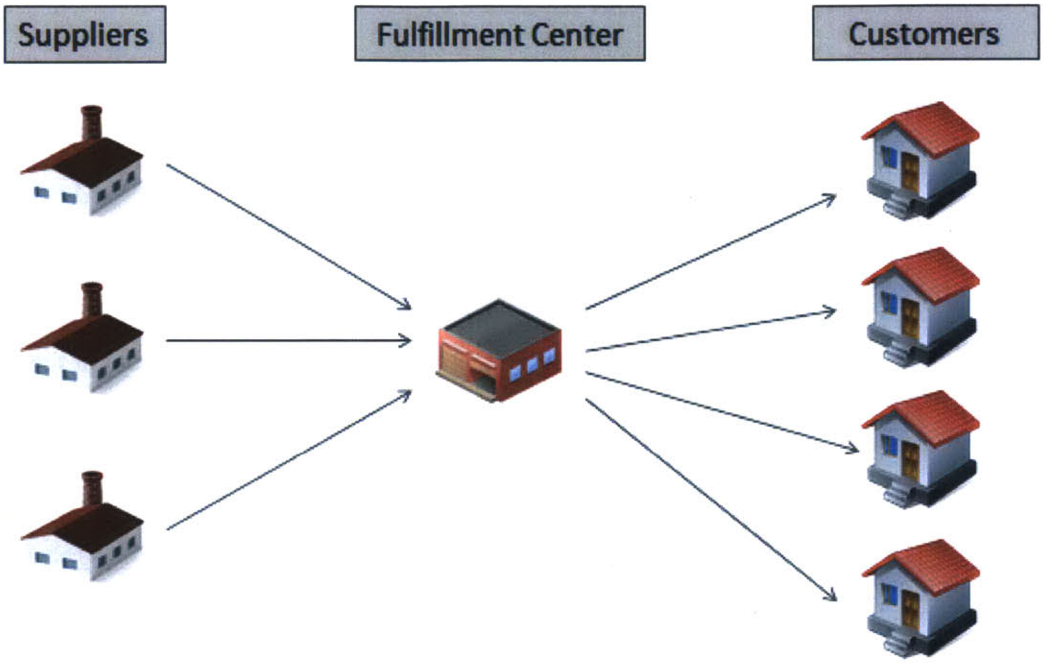


**Figure 2: Online retailer supply chain structure.**

In addition, online retail supply chains differ not only in structure but also in the way they function. First, in the traditional retail business model, customers determine where to buy from, which often involves visiting different locations. In contrast, online retailers can choose from where to serve an order based on transportation costs and service times. Secondly, in online retail supply chains there is a time delay between an order placement and its fulfillment, which again allows the retailer to decide on the optimal fulfillment solution. Finally, customers are usually given the option to select a service class or delivery option going from same day delivery to slower ship methods. From an operations management point of view, all of these characteristics pose some unique challenges. Yet, at the same time, they offer some unparalleled improvement opportunities.

Broadly speaking, there are two levers that can be used to reduce fulfillment costs. The first one controls inventory allocation throughout the network by selecting how much inventory and what items are placed at each location. This non-trivial problem is outside the scope of this project. The second option concerns how orders are served to customers, by which means and what infrastructure is required to do so. The answers to these questions occupy the remainder of this document.

## 1.2 Problem Statement

Fulfillment centers use a finite number of shipping lanes to ship orders to customers. *Shipping lane* is defined as an origin, destination and carrier triplet. For instance, a fulfillment center in Phoenix could serve Los Angeles area market-segment with USPS. At an operational level, a shipping lane can be thought of as a route carrying discrete customer orders to a particular market region. The route can be as simple as a direct path between the fulfillment center and the regional carrier hub from which orders leave for the final mile delivery, or it can be more complex with intermediate hubs where packages are sorted and sent to the next stage of the process. At the same time, two ship modes can be used: air shipment and ground shipment.

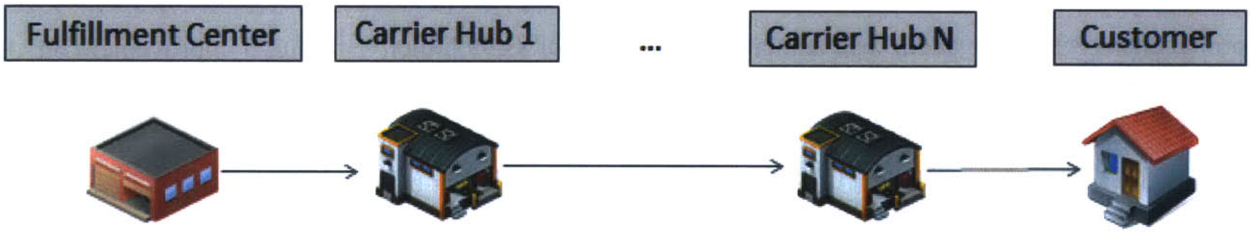


**Figure 3: Schematic shipping path.**

The number of shipping lanes at a particular facility is limited by the physical space as well as capacity constraints, including labor and maximum throughput that can flow through the system. Given the large number of destinations and the limited number of shipping lanes that can be used at a particular point in time, deciding on which ones to operate at each fulfillment center becomes a challenging question. One alternative consists in maximizing geographic coverage of each building by connecting it to as many market regions as possible. Doing so would increase the likelihood of Amazon being able to serve a larger portion of its customers from every facility. However, this could result in an overly expensive network as carriers with the broader coverage tend to be more costly. Conversely, another strategy is network specialization, by which each fulfillment center focuses on certain key geographies. Alas, this option can result in marginalization of some customers who would then turn to other online retailers. Consequently, the optimal shipping lane allocation consists of a combination of both approaches.

To further aggravate the issue, there is a second component to be taken into account. Companies in the online retailing market face a fierce competition. It is not enough being able to serve every customer, it is vital to do so as fast as possible. For that reason, the challenge becomes designing an outbound distribution network reaching the maximum number of customers in a cheap and timely manner.

Up until recently, the shipping lane allocation process consisted in simulating customer demand at every fulfillment center to determine the cost-minimizing configuration. Such approach ignores the effect each node has over the others, which is characteristic of networks and complex systems. While at the individual shipment level the cost difference between using the current network configuration and an optimal one is relatively small, with annual shipping costs of $3.99 billion[6], any small improvement on outbound shipping cost will result in substantial bottom-line savings. Given the difficulty of setting up a distribution network by evaluation of every node individually, there is a sizeable opportunity in reducing shipping costs by evaluating the system as a whole.

## 1.3 Project Goals

This project aims to devise a new method for designing a more effective outbound network from an economic and service time perspective. In other words, the objective is to develop a decision-making tool capable of analyzing thousands of potential shipping lanes and selecting those that minimize overall outbound shipping cost while meeting a demand forecast and lead time constraints, thus determining which shipping lanes to operate at each building. This study includes a close examination of current tools

[6] Amazon 10k, 2012.

and schemes used for evaluating shipping performance as well as identifying new lanes. Ultimately, success of the project will be measured by the model's ability to capture operating principles and intricacies of Amazon's distribution network in order to generate a list of recommendations to improve network performance.

## 1.4 Approach

The problem at hand is approached through a mixed-integer network optimization program, consisting of a series of origin, intermediate and destination nodes. A comprehensive sensitivity analysis is conducted in order to provide evidence that the algorithm works well. In addition, a pilot program focusing on a smaller portion of the network is run, making possible to identify real improvement opportunities and demonstrate the benefits of the formulation proposed. In addition, it also reveals the limitations of this approach. For that reason, alternative heuristics and future opportunities for research are presented.

## 1.5 Summary

Despite its renown for developing and running a state-of-the-art supply chain, Amazon can further enhance its operations management to deliver better results. This project pursues one particular improvement opportunity, namely optimizing its outbound distribution network, to demonstrate the advancement of the overall fulfillment scheme.

# 2 Current Process Overview

This chapter provides a brief description of the structure and operations behind Amazon's supply chain, starting with a general overview of its fulfillment center network, to then concluding with an analysis of its shipping lane allocation process.

## 2.1 Fulfillment Center Network

Central to Amazon's strategy of enhancing customer experience are fulfillment centers, also known as FC's. A *fulfillment center* is a specially designed warehouse capable of holding inventory and shipping customer orders at the same time. This broad definition hides the fact that facilities are specialized by the type of inventory they carry. A piano and a book require very different handling and shipping methods. In this way, fulfillment centers are classified by physical features of the products they hold.

Another distinctive element of every fulfillment center is its geographic location. The closer a building is to a market region, the better and cheaper service it can offer. For that reason, the outbound distribution network structure has recently shifted from connecting every FC to as many markets as possible to a grouping model by which delivery coverage of every building is limited to a particular area. As a result, the US market has been broken down into different regions.

The rationale behind this division is that fulfillment centers located in a given region should only serve that particular area, thus limiting the number of shipments across different regions. This allows Amazon to reduce overall distances between origin and destination pairs, resulting in reduced shipping costs and lead times. However, this approach is not without its limitations. For one, the network is unbalanced, meaning that customers of some territories consume more than the combined fulfillment capacity of all fulfillment centers within the region. At the same time, other regions have idle capacity and are able to handle more volume than required by local customers. Subsequently, a purely regional strategy is not feasible. Some regions act as net exporters, while others have to import a considerable fraction of their orders from other areas.

## 2.2 Outbound Distribution Network

As more FC's populate the national grid, overall distances to households are reduced. This constitutes an important competitive advantage. Yet, a larger distribution network gives rise to redundancies and inefficiencies, which hinder fulfillment performance. With a larger selection of buildings at its disposal, Amazon can choose how to serve every market region in a more effective way. Rationalizing its network structure will not only improve the bottom line, but it will also enhance the customer experience by reducing transit times. In an effort to take advantage of the expanding FC network, the outbound distribution network has moved to the regional plan described in the previous section. However, this approach is thus far incomplete for the shipping lane allocation process has not evolved accordingly.

The current lane selecting process is governed by inventory availability. Based on the inventory mix at a particular FC along with the geographic distribution of customer orders, shipping lanes are designated for linking supply to demand. There have been a number of projects focused on determining the optimal inventory allocation for each FC based on demand patterns of certain market regions. While these studies have yielded impressive operating savings, the solution is incomplete from an overall supply chain perspective. Indeed, these models tend to neglect the intrinsic complexity of distribution networks, overly simplifying shipping costs or assuming lanes as an immovable given. As a result, the distribution network is relegated to a second place in the order of operations excellence.

The optimal solution can only be found if the totality of the supply chain is analyzed at the same time, weighing the fine trade-off between inventory holding costs and transportation costs while allowing the model to move inventory around and create new shipping arcs. Unfortunately, stock allocation is out of the scope of the present project. However, careful attention to this issue has been placed when formulating this model. For that reason, customer demand is pre-assigned to a fulfillment center, which in turn makes possible to run several scenarios with different product mixes to evaluate the impact of different inventory allocation policies.

As far as the process of selecting which shipping lanes to operate is concerned, it is currently done using a simulation by which individual buildings are modeled one at a time. The program takes a demand forecast and generates a good shipping lane allocation based on economic criteria. As mentioned earlier, this method fails to account for the influence each node exerts on the others. As an example, two FC's could be connected to the same carrier hub which has a limited capacity. By looking at every building separately, potential effects of saturating the hub are ignored. Similarly, considering every fulfillment center as an independent operation disregards the possibility of transshipments across facilities. Furthermore, optimizing one building at a time benefits the first FC's to be evaluated, yet it forces unnecessary constraints into subsequent ones. Therefore, a new approach evaluating all of the buildings at once is required. Fortunately, the regional distribution network reduces the size of the problem since there are fewer nodes to look at.

Another ongoing effort for reducing shipping costs is cutting down the number of air shipments in favor of ground shipment. In general, it is more expensive to ship an item by air than by ground. With the regional distribution plan, it is easier to fulfill most of the orders by truck as fulfillment centers are closer to customers. For that reason, this project only looks into ground shipment solutions. Similarly, it is more expensive to ship a multi-item order in multiple packages than to ship a single package from a single location. This constitutes another reason for using the pre-assigned demand approach mentioned earlier, as per definition it considers packages and customer orders both alike. From now on, the terms package and customer order will be used synonymously, as both refer to the smallest unit that can be shipped from an origin to an end destination.

Finally, there is the operational aspect of enabling a new shipping lane to be considered. Doing so requires of a non-negligible effort from different actors within the organization. From negotiating with the carrier to putting into place the appropriate configuration changes in the IT systems, along with adapting the physical space at the fulfillment center and educating the hourly associates, every recommendation poses some implementation challenges. For that reason, every change should be robust enough to endure for a reasonable amount of time. Otherwise it risks causing resentment and push back from the rest of the establishment. A shipping configuration that is constantly changing can cause problems that outweigh the sought economic benefits. This does not mean that the outbound network should not adapt to changes in demand; on the contrary, the best configuration solution is flexible enough to absorb those fluctuations while keeping the number of physical modifications to a minimum. This fine trade-off will be referred to as *network robustness*.

## 2.3 Summary

Up until now, shipping lanes were allocated with a myopic view, considering nodes as independent entities. Such course of action not only ignores dynamics of complex systems in which relationship among separate elements cannot be understood without evaluating the network as a whole, but also results in a sub-optimal ranking system by which some fulfillment centers enjoy a privileged position to the detriment of the others. Regardless, influence of one node over the others should not be ignored. The crux of the matter is developing an analytical approach capable of evaluating the entire outbound network at once, which in turn will reveal where the improvement opportunities are concealed.

# 3 Literature Review

Extensive literature has been written on the topic of distribution network design and supply chain optimization. This brief chapter is not intended as an exhaustive review of those fields, but rather identifies the main principles underlying the problem at hand. In addition, various approaches developed by different authors are discussed and their applicability to the present project is evaluated.

## 3.1 Mixed Integer Linear Programming Approach

Mixed integer linear programming (MILP) constitutes a general framework for modeling problems involving integer and discrete variables. In general, MILP problems belong to the NP-hard computational class of decision problems. A special case of MILP problems is binary integer linear programming in which decision variables can take 0 or 1 values. Generally speaking, linear optimization problems with binary decision variables are better understood. There are several commercial applications capable of solving this problem family in a timely manner.

Modeling a distribution network as a mixed integer linear program is not a new concept. Manzini et al. (2006) propose the development of a decision support system platform as a response to the so called Production Distribution Logistic System Design (PDSD) problem[7]. Such tool is in fact a general MILP applied to a generic supply chain consisting of production plants, distribution centers and customers. This work illustrates the benefits of using an optimization approach while at the same time reveals one of its main limitations. As mentioned earlier, MILP are NP-hard problems that require special computing capabilities. Alas, in real life applications the size of the problem can rapidly exceed computational limits of conventional numerical solvers. Alternatively, heuristics and local optimization algorithms can be used as a compromise solution.

[7] This problem involves dealing simultaneously with the design, management and control of logistic supply chains.

Tsiakis and Papageorgiou (2006) developed a mixed integer linear programming model applied to a large organization with a complex operational structure. Given a number of fixed production plants and customer zones, the model evaluates a number of possible distribution centers and selects those that minimize overall shipping costs while meeting various constraint types, including production caps and quality restrictions. The resulting model constitutes a strategic decision making tool that addresses financial and operational challenges. Despite being very similar in nature, the present project is not concerned with which intermediate nodes to choose, but rather with which arcs to enable.

In a similar work, Melachrinoudis and Min (2007) approach the warehouse redesign problem through a MILP formulation. A transit time constraint is introduced, which effectively identifies transportation performance and reliability as an additional challenge to be modeled. Besides determining which facilities to operate, the formulation develops a regional operation plan. This is, deciding what customers should be served by which warehouses. Moreover, it also measures the sensitivity of the optimal solutions to small changes of the network constraints. This methodology, while relevant to problem at hand, ignores the multi-commodity problem in which customer orders fall into different lead time categories.

In a different study, Gamus et al (2009) combine a MILP approach with a neuro-fuzzy demand forecast to optimize the design of a three-echelon supply chain. Said model successfully incorporates demand uncertainty into a network optimization problem to capture the realities of challenges faced by most companies. Once again, an increment on the number of elements to be modeled entails an exponential increase on the complexity of the problem, which requires non-conventional computing capabilities. For that reason, a simpler model accounting for all of the caveats mentioned earlier is proposed in this paper.

## 3.2 Predictive Modeling

Due to the size of the problem at hand, some simplifications are adopted in order to develop an accessible tool. Notably, some of the demand granularity is lost by aggregating customers at a meaningful cluster level. By this, the size of the problem is considerably reduced, yet it poses additional challenges. Because demand is no longer modeled at the customer level, some data precision is lost. Indeed, the current approach requires that shipping costs and transit times at the cluster level consist of an average of its individual values. Should the averaging be done on a population density basis, on an absolute number of packages or by geographic extension? Since there is no reasonable criterion to do so, a different approach based on statistical prediction is required.

Acimovic and Graves (2012) develop a shipping cost prediction model based on historical data. The model looks at a sample of fulfilled demand aggregated at the zip3[8] level and estimates a shipping cost function through linear regression. The resulting model consists of a series of step-wise functions that predict shipping costs based on origin-destination distance, ship option and weight. However, for purposes of the problem at hand, the aforementioned approach misses two important elements. First, it does not distinguish between carriers, resulting in an average cost across all shipping methods that does not capture the essence of the problem. Secondly, distance is measured as a straight line between FC and customer zone, ignoring the intermediate steps that a package follows in its delivery path.

In a different study, McDonald (2011) proposes an alternative to preconfigured transit times and advocates for a different perception of time-in-transit. *Transit time* is defined as the time spam encompassing the instant an order leaves an origin facility until it reaches its final destination. Given the uncertain nature of transportation problems, a probability measure is required. Thus far, most companies take a quantile function approach. They are concerned with the minimum number of days required for order to reach the customer for any given level of certainty. The new approach is based on cumulative

[8] Zip3 is an aggregate of postal codes which begin with the same prefix or the same three digits. A zip3 can contain as little as 1 postal code or as many as 100.

distribution functions (cdf), which looks at the probability that a package makes it to its final destination at any given number of days. Because of the discrete nature of transit times, this subtle difference has important implications. The use of quantiles entails some information loss, which makes it difficult to use. In contrast, cdf is lossless, enabling a finer trade-off between cost and surety. The same paper proposes the use of random forest regression to predict mean transit times. This method performs well in least square regression problems and has the advantage of running efficiently on large databases. However, it results in cumbersome prediction expressions which are not easy to incorporate. A simpler approach for predicting transit times is described in Section 4.3.

# 4 Model Formulation

This chapter describes the mathematical formulation behind the network optimization model proposed in this project. In addition, two prediction approaches for estimating shipping cost and transit time are advanced and integrated within the general decision-making tool.

## 4.1 Network Optimization Model

The problem at hand can be modeled through a mixed integer network optimization program consisting of a series of origin, intermediate and destination nodes. The origin node set represents the ensemble of fulfillment centers from which orders are shipped to the next stage. The intermediate node set consists of all of the carrier hubs, in which orders arrive, are processed and then shipped directly to the final customer. Finally, the destination node set encompasses the different customer clusters or demand aggregates. This grouping could be done by zip code, by state, by population density or by any other geographic rationale.

Likewise, shipments are categorized by service class and weight range. Four services classes are considered, each one with a different service time. They are: Next Day, Second Day, Standard and Super-saver. Next Day and Second Day deliveries belong to the premium category. Next Day packages have a 24 hour service time, while Second Day have 48 hours. The other two service classes belong to the standard group, which are characterized by longer delivery times. For Standard shipments it is between three and five days. For Super-savers it is between five and eight days. Finally, weight is broken down into four groups: light, medium-light, medium and heavy packages. The rationale behind this weight classification is explained on Section 4.2.

The mathematical formulation of the model relies on the following notation:

- $N$ : Set of network nodes. The network contains FC's, carrier hubs and customer zones.
- $A \subset N x N$: Set of arcs in the network. An arc between two nodes, FC and carrier hub, represents a shipping lane between those two nodes. Each arc has an associated per unit shipping cost.
- $K$: Set of service classes. As mentioned earlier, four service classes are considered: Next Day, Second Day, Standard and Super-saver shipment. Service class is designated by subscript *m.*
- W: Set of weight ranges. This set consists of four categories: Standards, SBPM, BPM and Parcels. Weight range is designated by subscript *w.*
- $U \subset N$: Subset of fulfillment center nodes in the network. FC's are designated by subscript *i.*
- $M \subset N$: Subset of carrier hubs. Carrier hubs are designated by subscript *j.*
- $Z \subset N$: Subset of customer zones in the network. Customer zones are designated by subscript *l.*
- $C_{ijlkw}$: Per unit cost for shipping product from FC *i* to customer zone *l* through carrier hub *j* in service class *k* and weight range *w.* For instance, if service class is Second Day delivery, then the cost parameter needs to account for the transportation cost to move product from the FC to the carrier, plus the carrier's cost to deliver to the customer zone and to do so within the two-day time window. This cost parameter will also account for transit time restrictions. That is, if a given lane cannot satisfy the time requirements of a particular service class, then its cost will be set to infinity or a very large number so the program automatically drops that particular shipping lane.

- $D_{ilkw}$: Daily demand at customer zone *l*, service class *k* and weight range *w* that is to be met from FC *i*.

- $V_i$ Capacity at FC *i* measured by the number of shipping lanes available.

- $H_j$: Capacity at carrier hub *j* measured by the maximum number of packages the carrier can handle on a single day at that particular facility.

- $Max_{ij}$ , $Min_{ij}$: Maximum and minimum volume allowed on link between FC *i* and hub *j*.

- $x_{ij}$: Binary decision variable to denote whether shipping lane serving carrier hub *j* is used at FC *i*. A value of 1 means the arc is active. A value of 0 the shipping lane is not used.

- $y_{ijlkw}$: Amount of product, measured in packages, shipped from FC *i* to customer zone *l* through carrier hub *j* satisfying service class *k* and weight range *w*.

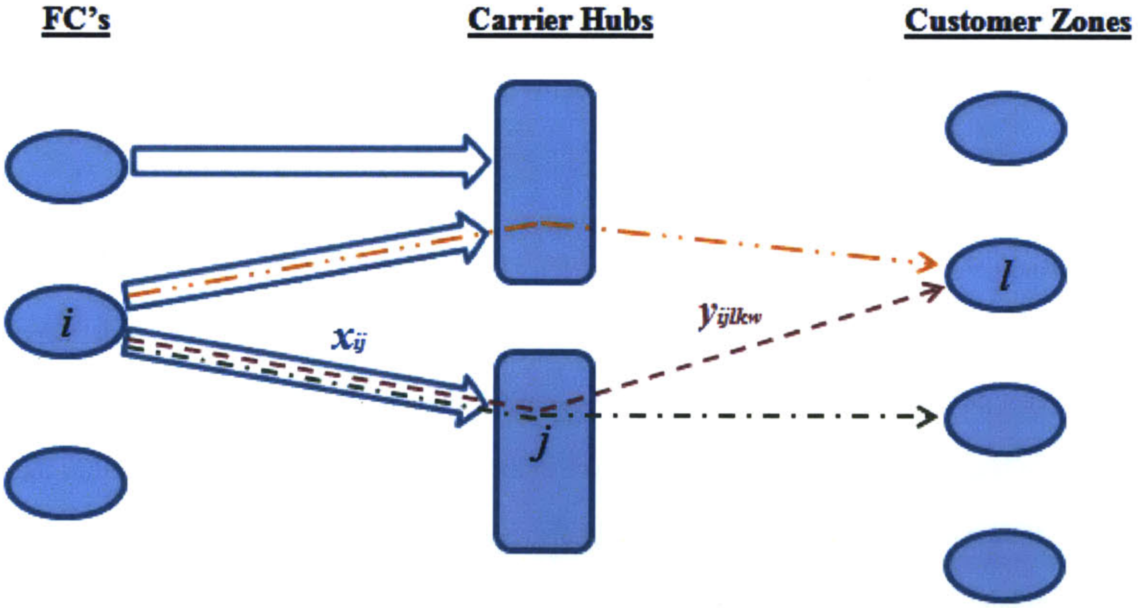


**Figure 4: Schematic representation of network optimization problem.**

The network optimization problem can be represented as depicted by Figure 4. Blue arrows symbolize a shipping lane going from FC $i$ to carrier hub $j$. Through that lane, multiple packages going to a particular destination with a given service class and weight range can flow. Each one of these flows is represented by a dotted line, which corresponds to decision variable $y_{ijlkw}$.

The mathematical formulation of the network optimization problem is:

Objective function minimizes outbound shipping cost:

$$\min \sum_{i \in U} \sum_{j \in M} \sum_{l \in Z} \sum_{k \in K} \sum_{w \in W} C_{ijlkw} y_{ijlkw} \qquad (1)$$

Subject to the following constraints and restrictions:

- Capacity constraints:

$$\sum_{j \in M} x_{ij} \leq V_i \qquad \forall i \in U \qquad (2)$$

$$\sum_{i \in U} \sum_{l \in Z} \sum_{k \in K} \sum_{w \in W} y_{ijlkw} \leq H_j \qquad \forall j \in M \qquad (3)$$

$$\sum_{l \in Z} \sum_{k \in K} \sum_{w \in W} y_{ijlkw} \geq Min_{ij}\, x_{ij} \qquad \forall i \in U, \forall j \in M \qquad (4)$$

$$\sum_{l \in Z} \sum_{k \in K} \sum_{w \in W} y_{ijlkw} \leq Max_{ij}\, x_{ij} \qquad \forall i \in U, \forall j \in M \qquad (5)$$

- Demand satisfaction:

$$y_{ijlkw} \leq x_{ij} D_{ilkw} \qquad \forall i \in U, \forall j \in M, \forall l \in Z, \forall k \in K, \forall w \in W \qquad (6)$$

$$\sum_{j \in M} y_{ijlkw} = D_{ilkw} \qquad \forall i \in U, \forall l \in Z, \forall k \in K, \forall w \in W \qquad (7)$$

- Non-negativity:

$$x_{ij} = 0, 1 \tag{8}$$

$$y_{ijlk} \geq 0 \tag{9}$$

The objective function to be minimized is the overall shipping cost of the network. Constraint (2) specifies the maximum number of shipping lanes available at every FC, while Constraint (3) limits the amount of packages that can be processed by every carrier hub on a given day. Constraint (4) can be used to specify the minimum volume required to maintain a shipping lane or open a new one. Expression (5) is the forcing constraint, reducing volume to zero when the shipping lane is disabled. Constraint (6) forces the algorithm to meet demand. Thus it is presumed that supply can always meet demand. Finally, Constraint (7) forces decision variable $x$ to take only two values (0 or 1) and Constraint (8) forces volume to be a positive number.

Three additional considerations are worth mentioning. First, fixed costs associated with enabling a shipping lane are negligible. However, opening a new lane is not an instantaneous process, meaning it takes a non-negligible amount of time to create a new one. Secondly, it is legitimate to aggregate product flows for purposes of the model, and it is not necessary to model products at a more detailed level. Finally, this model assumes that demand can be pre-assigned to each FC. That is, for each customer zone, service class and weight range triplet, it is possible to specify what fraction of demand will be satisfied by each FC. This reflects how inventory is allocated across the FC's, as well as the general operating plan for the network. Such definition of demand allows the assessment of inventory placement at a particular FC, which in turn serves to evaluate operational costs as a whole, including transportation and inventory holding costs. This definition overlooks how inventory is allocated across the network, which is a complex supply chain problem in itself. For that reason the model is not allowed to select which building will serve a particular order. Instead every order is allocated to a serving warehouse beforehand. This can be done because historical customer orders, for which the origin is known, are used as an input. An additional benefit of such an approach is that it further simplifies the problem because it omits the multi-item order splitting issue.

## 4.2 Cost Prediction Model

Per unit shipping cost in the system is defined in Expression (10). These cost coefficients are assigned to every arc and include not only shipping cost but also delivery time restrictions. The second term forces the cost coefficient to take a very large number when a particular shipping lane cannot meet a given service time. Incorporating a time dimension into the cost coefficients results in a fast performance heuristic, significantly reducing computing requirements.

$$C_{ijlkw} = h_{ijlkw} + \max(t_{ijlkw} - T_k, 0)M \tag{10}$$

Where:

- $h_{ijlkw}$: Per unit cost for shipping one unit of product from FC $i$ to customer zone $l$ through carrier hub $j$ in service class $k$ and weight option $w$.

- $t_{ijlkw}$: Mean transit time for shipping one unit of product from FC $i$ to customer zone $l$ through carrier hub $j$ in service class $k$ and weight option $w$.

- $T_k$: Maximum transit time allowed for service class $k$. For instance, Second Day delivery is allowed a maximum of 48 hours. Maximum transit time allowed is a requirement per se, but instead of writing it as a constraint, it is included in the cost function (10)

- $M$: Large number.

## 4.3 Transit Time Prediction Model

Predicted transit times follow a different approach based on cumulative distribution functions (cdf). Given a particular origin, destination and ship method, the goal is it to determine the probability that a package makes it to its final destination at any given number of days. One simplistic approach is to look at historical data and evaluate the fraction of packages that made it on time. This method is synthesized in the following expression:

$$f_{ijlkw}(T) = \frac{x_{ijlkw(T)}}{X_{ijlkw}} \tag{11}$$

- $x_{ijlkw(T)}$: Number of packages shipped to customer region *l*, from FC *i* through carrier hub *j*, service class *k* and weight range *w* that made it within time T.
- $X_{ijlkw}$: Total number of packages shipped to customer region *l*, from FC *i* through carrier hub *j*, service class *k* and weight range *w*.

Therefore *f* represents the fraction of packages shipped to customer region *l*, from FC *i* through carrier hub *j*, service class *k* and weight range *w* that made it within promised service time T. For instance, in case of Next Day delivery, the aforementioned expression evaluates the fraction of packages that reached their final destination in 24 hours or less. Hence, the probability that an order is fulfilled in T days is equal to number of packages from the data sample that made it on time divided by the total number of packages flowing through the arc. This fraction can then be compared to on-time delivery (OTD) targets set by the organization for different service classes. Generally speaking, each service class has a targeted OTD set by the organization. If *f(T)* is greater or equal to the OTD, then the model assumes that the transit time for the arc is T. Otherwise, it evaluates the next service class. Seemingly, it can be expressed in mathematical notation:

$$\text{if } f_{ijlkw}(T) \geq \text{OTD(k)} \text{ then } t_{ijlkw} = T_k, \text{ else } t_{ijlkw} > T_k \tag{12}$$

As an example, if the fraction of packages delivered within 24 hours is smaller than the targeted Next Day OTD, then the fraction of packages delivered within 48 hours is evaluated. If this fraction is greater than the targeted Second Day OTD, then expected transit time of the customer region is assumed to be 48 hours. Conversely, if it is smaller than the targeted Second Day OTD, then the fraction of packages delivered within 72 hours is evaluated against the appropriate on-time delivery metric.

The model at hand presumes that historic transit time performance is representative of future events, which seems a legitimate assumption to make. However, more accurate predictions are obtained when using a larger sample of historic data. Estimated transit times are used in the cost prediction formulation discussed in the previous section to calculate cost coefficients $C_{ijlkw}$.

## 4.4 Summary

This chapter introduces a method for evaluating the entire distribution network at once, which is consistent with project expectations. In addition, some shipping cost and transit time prediction models are presented. Despite being somehow simplistic, they successfully capture the features and intricacies of the problem at hand.

# 5 Estimation of Coefficients

Chapter 4 presents the mathematical formulation behind the proposed optimization algorithm. However, there is one important piece missing: the shipping cost and transit time coefficients. This section describes a practical approach on how these parameters can be estimated. Note that in order to protect proprietary information, data presented in this chapter is fictitious. Furthermore, some simplifying assumptions have been adopted to preserve confidentiality. Nevertheless, the framework developed through this document mimics the decision-making tool built for the project. Thus findings and recommendations are still relevant for academic purposes.

## 5.1 Shipping Cost Coefficients

Section 4.2 introduces unit shipping cost parameter $C_{ijlkw}$ as a function of two coefficients. On the one hand *h* represents the actual shipping expenses. On the other hand, *t* adds a time dimension to the allocation problem. However, estimating both coefficients is as challenging as estimating *C*. One possible approach consists in running a least squares regression model using historical data on customer orders to develop a prediction function. This section looks at the cost prediction function while transit times are discussed in the next one.

If $\hat{h}$ is defined as an unbiased estimator of *h*, the objective is to minimize the residuals or error sum of squares:

$$SS_{error} = \sum_{i,j,l,k,w}(h_{ijlkw} - \hat{h}_{ijlkw})^2 \tag{13}$$

The regression uses five attributes: FC, carrier, origin-destination distance, service class and weight. Or alternatively:

$$\hat{h}_{ijlkw} = f(\text{FC, carrier, distance, service class, weight}) \tag{14}$$

Using historical data, it is possible to assess the influence of each one of these parameters. Distances between origin and destination pairs were estimated using the great circle formula with latitude and longitude of geometric center of demand clusters as inputs. Carrier is by far the largest cost driver followed by weight and origin-destination distance. Surprisingly, service class plays a minor role while the FC has no significant impact whatsoever. Figure 5 is a plot of actual cost versus predicted cost. The model accounts for 91% of variability.

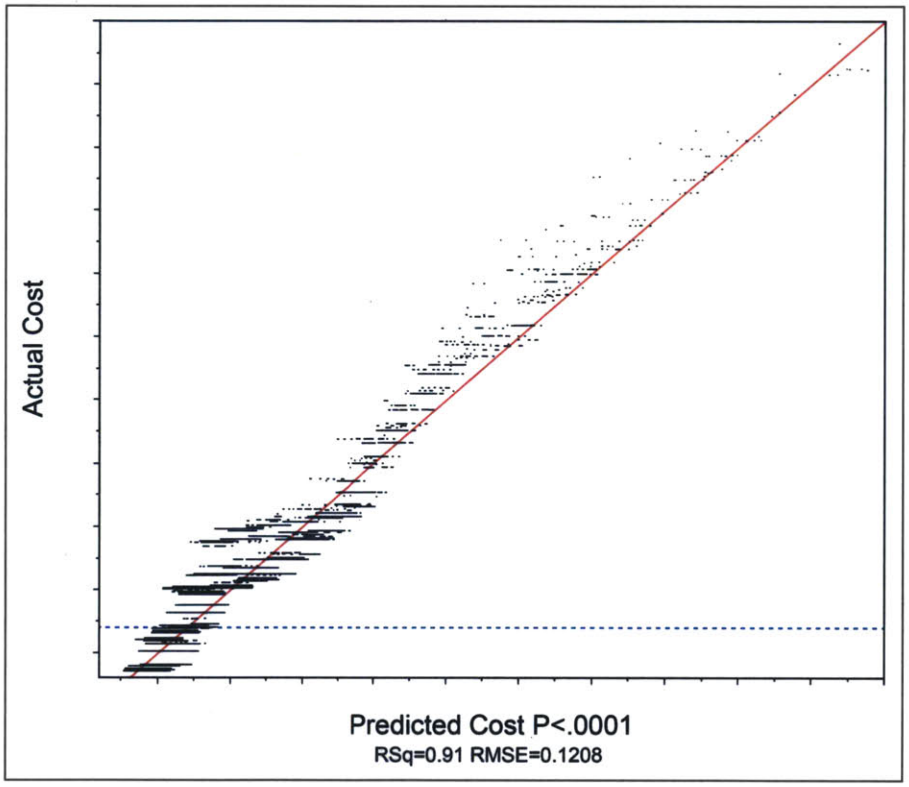
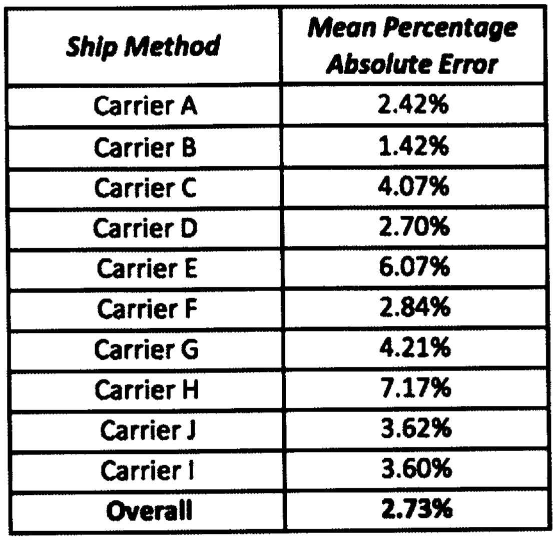


**Figure 5: Real shipping cost versus predicted shipping cost.**

At the same time, the accuracy of the shipping cost prediction model was assessed using a different data set from the one employed in the linear regression. Figure 6 shows expected errors for the ten carriers included in the model. The overall mean absolute percentage error is 2.73%.



| Ship Method | Mean Percentage Absolute Error |
|---|---|
| Carrier A | 2.42% |
| Carrier B | 1.42% |
| Carrier C | 4.07% |
| Carrier D | 2.70% |
| Carrier E | 6.07% |
| Carrier F | 2.84% |
| Carrier G | 4.21% |
| Carrier H | 7.17% |
| Carrier J | 3.62% |
| Carrier I | 3.60% |
| **Overall** | **2.73%** |

**Figure 6: Prediction cost model accuracy.**

Plotting predicted shipping cost as a function of package weight for various carriers reveals some meaningful insights. As weight increases, there are clear shifts on the cheapest ship option. The reason behind such cost structure is that some carriers do not want to handle bulkier packages, thus they inflate the price of heavier packages to remove themselves from competing at that particular market segment. The intersection between two or more of these curbs marks a transition in the overall shipping cost structure. Based on these results, four different weight categories were selected for the model:

- **Light Packages**: for little weight parcels.
- **Medium-light Packages**: for orders whose weight is between light and medium size parcels.
- **Medium Packages**: for medium size parcels.
- **Heavy Packages**: for bulkier parcels.

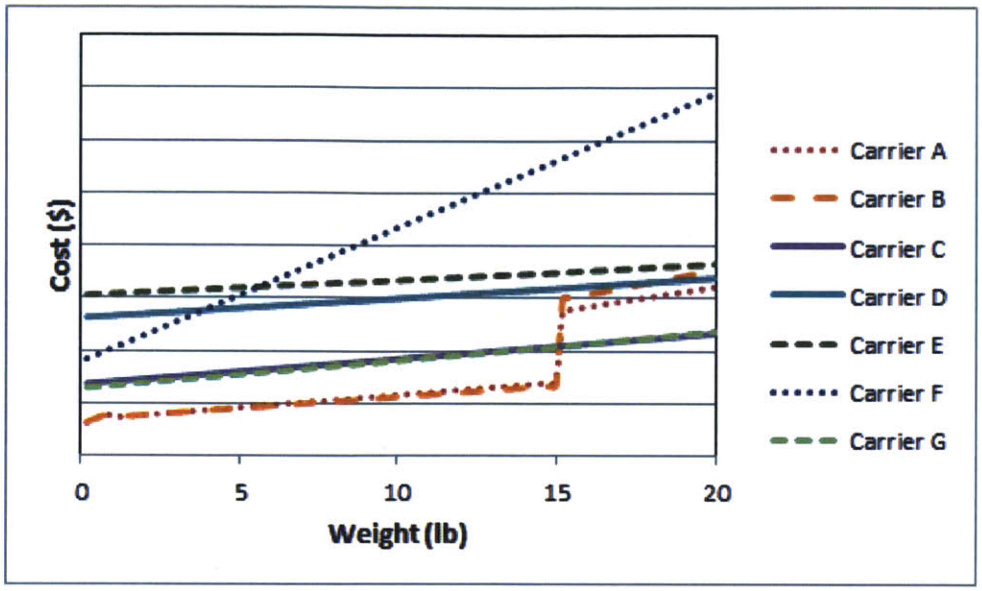


**Figure 7: Transportation cost as a function of weight for different carriers.**

There are further implications to this phenomenon, notably how a strategic assignment of heavier orders to certain carriers could result in significant savings. This will be discussed in Chapter 7.

## 5.2 Transit Time Coefficients

Early on, this model assumes that customer demand can be aggregated at the zip3 level without losing significant granularity. Given a region of the US, coverage footprint for every fulfillment center can be plotted. Figure 8 represents promised transit times for orders fulfilled by a FC using a particular carrier hub. Albeit these service levels are quoted by the carrier, historical data reveals that on average orders met or exceeded prevailing on-time delivery targets. Thus, they can be assumed to be representative. Regions with the same color coding have the same expected transit time. Moreover, the darker an area is, the longer its expected transit time.

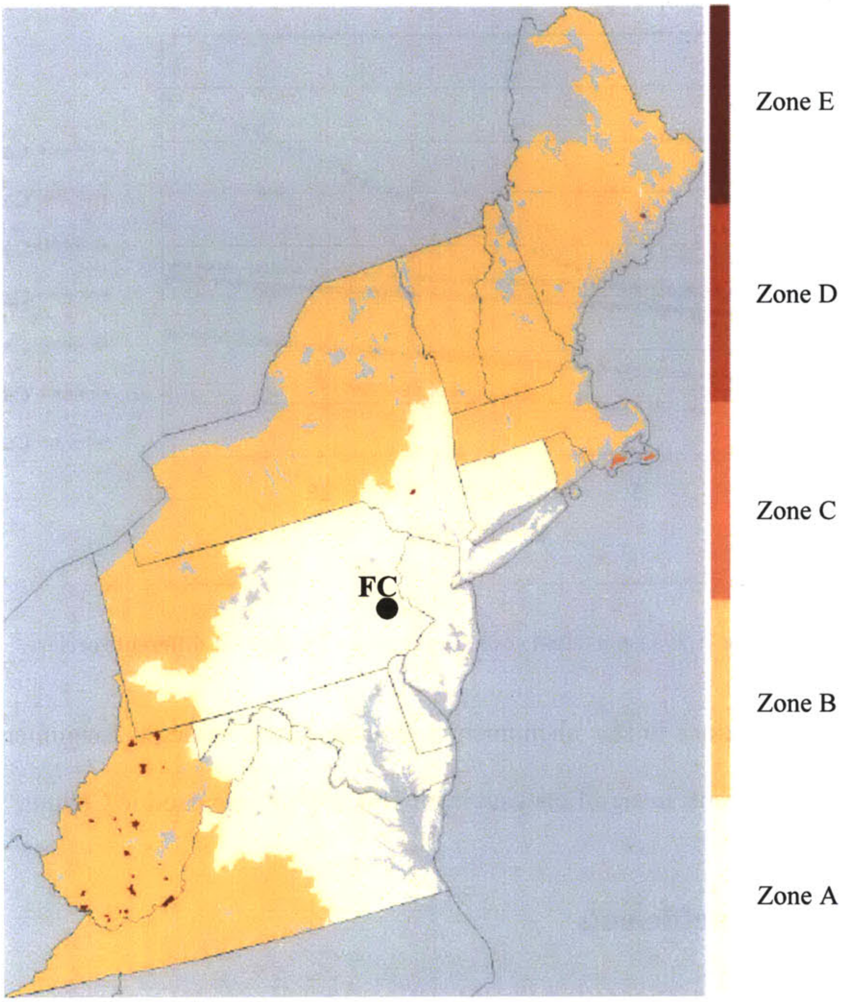


**Figure 8: Promised transit times at postal code level.**

In order to estimate actual transit times, historical data worth two months of customer orders was used. Expected transit times are calculated using Expression (12). Figure 9 represents transit times for the same ship method predicted with the model proposed on Section 4.3. In the first figure demand is defined at the postal code level while in the second one it is aggregated by zip3. Upon close examination, it can be concluded that both figures are very similar. The same analysis was carried out for every FC and ship method combination with analogous results. Therefore, aggregating customer orders by zip3 seems like a reasonable assumption.

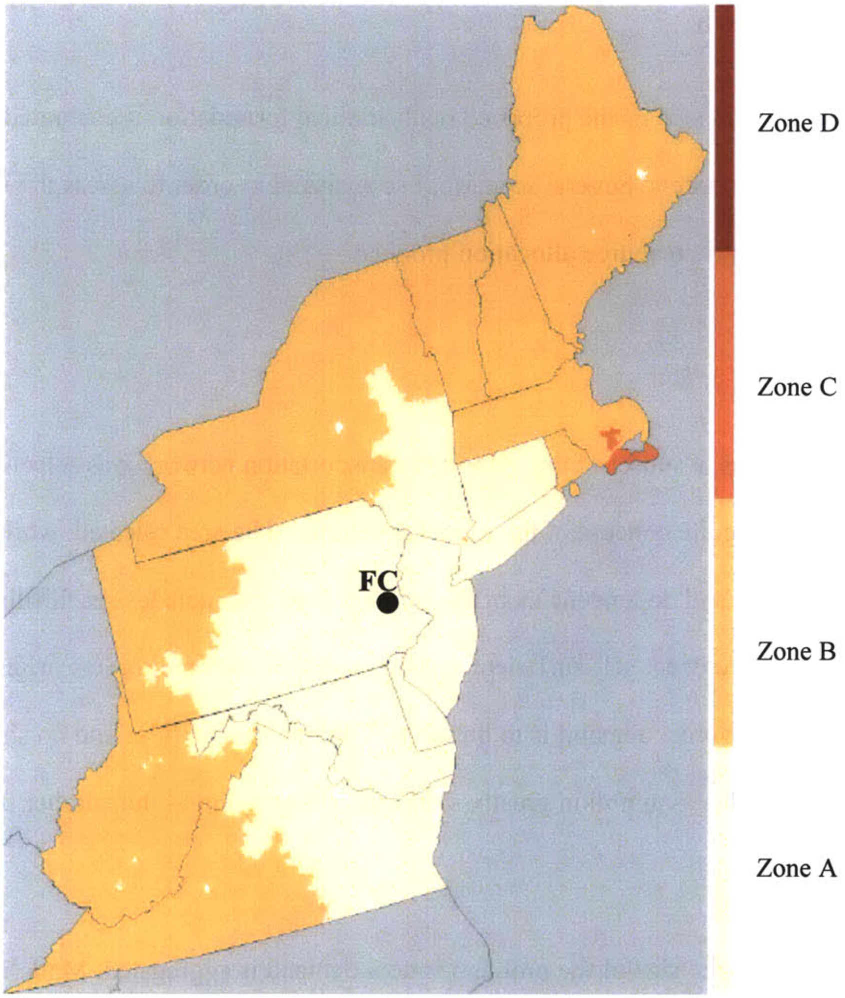


**Figure 9: Estimated transit times aggregated by zip3.**

Finally, the effective shipping cost *C* can be calculated with the estimated parameters *h* and *t*.

## 5.3 Summary

This section presents a possible approach for estimating shipping cost and transit time coefficients. The next chapter introduces the abovementioned parameters into the network optimization model in order to assess its validity.

# 6 Model Validation

In this chapter, the performance of the proposed mathematical formulation is evaluated using real data from Amazon's customer orders. Several scenarios are analyzed in order to assess the impact of each factor to the overall dynamic resource allocation problem.

## 6.1 Pilot Run

For computational reasons, a subset of the outbound transportation network was selected for running a pilot program and proving the concept of the model developed. The area selected, while still representative of features and dependencies of the original network, includes six fulfillment centers. This region is assumed to be a closed, self-sufficient system, in which customer orders originate and are fulfilled within. In other words, demand is to be served by those six facilities and no shipments outside the region are allowed. This assumption greatly simplifies the problem as the number of nodes and arcs is considerably reduced.

The largest contributor to the size of the problem is how demand is aggregated. Modeling each customer as a single node results in an intractable decision-making tool, hence the necessity to combine them into clusters. Given the geographic nature of the problem, demand can be aggregated by postal code or zip3. Choosing the first option implies dealing with approximately 7 million decision variables, whereas using zip3's instead requires around 150,000 decision variables. As granularity of demand is increased, the number of decision variables rapidly scales up. It is hard to determine if using a finer regional mesh justifies the increase in complexity. Given the geographical nature of the problem, it seems legitimate to assume that two neighboring regions could be served by the same shipping lane without incurring a significant cost difference. Therefore the use of zip3's seems justified.

The dataset used in this analysis consists of real orders, shipments and inventory details. All this information was compiled into a single database for which every customer order contains the fulfillment center serving the order, destination in form of zip3, service class, package weight and actual shipping cost. This data was then used to simulate customer demand for various days. Using different days not only accounts for demand variability within the period of time selected, but it also reflects the weekly customer order pattern that characterizes this particular online retail market.

The first step into validating accuracy of the model is comparing predicted overall shipping cost of the network with the real cost. For that, the model was constrained to use the actual network configuration, enabling those shipping lanes currently in use and forcing the rest to be suppressed. Seven data samples were used to introduce some variability. Each one consists of historical orders from a particular day. Once again, data has been modified in order to protect confidential information. In the second step, the algorithm is allowed to make at most six changes, which in practical terms provides the model with the ability to drop six shipping lanes from the current configuration and add up to six new ones in order to find the cheapest network set-up.

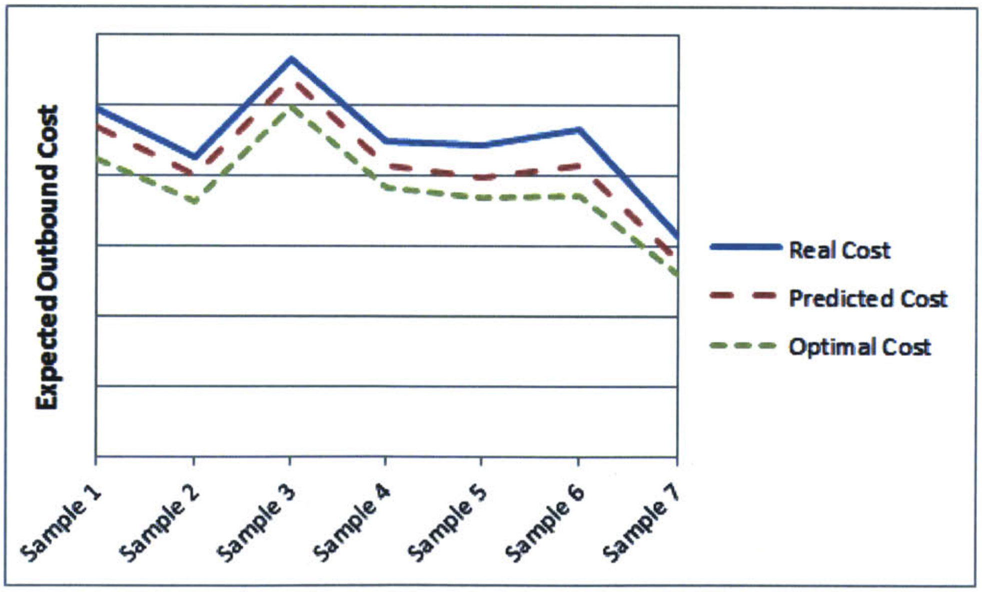


**Figure 10: Model performance for seven historic data samples.**

The figure above represents the real cost, cost predicted by the model using the current configuration and optimal solution for the different data samples. The network optimization model consistently undervalues the real shipping cost. Difference between real outbound cost and predicted cost has a mean value of 5.54%. Being a predictive model, some level of deviation is expected. The prediction error is consistent with the systematic error produced by the cost prediction model described in Section 4.2. On the other hand, analysis reveals that there are potential savings to be gained by implementing the optimal solution found by the decision-making tool. Let $C^{sim}$ be the cost of the current configuration simulated by the algorithm and $C^{op}$ the optimal cost or lower bound. Then, the overall outbound shipping cost improvement gap is defined as:

$$\frac{C^{sim} - C^{op}}{C^{sim}} \tag{15}$$

Total cost obtained through the simulation instead of the actual cost is used in order to have a common benchmark basis. Results suggest that there is a 1.39% improvement opportunity to be gained by just replacing four shipping lanes.

In terms of performance, the algorithm runs very fast using conventional computational resources. The numerical solver is able to find the optimal solution in a few seconds.

## 6.2 Model Robustness

The soundness of the model can be assessed by measuring the change in performance of the optimal solution when introducing a small perturbation in the system. A considerable delta would mean that the algorithm is very unstable and of limited applicability.

The optimal solution found earlier represents the best possible performance that can be achieved because every day has been optimized independently. Indeed, the algorithm finds the optimal network configuration for each day. Because of demand variability, the optimal solution for one day is not the optimal solution for the next one. In other words, the best shipping lane combination for one particular day is not necessarily the preferred choice for the rest of the days. Since new solutions in an industrial project lead to changes, and changes can causes disruption to some people, altering the network configuration every so often becomes intractable. Instead, it is more desirable to find a set of shipping lanes to be used at all times that outperforms the current network configuration, even if this does not yield the greatest possible savings.

Figure 11 compares expected outbound shipping cost of the current network configuration versus the optimal shipping lane combination from a given day. The day with the most customer orders placement is used as the baseline for which the optimal shipping lane allocation is estimated and then utilized on every single day. The simulation is run for two weeks. Note how the proposed configuration performs just as well if not better than the actual one. Not surprisingly, the greatest improvement corresponds the day with the greatest volume. At the same time, it is remarkable that the proposed network also improves results for the three subsequent days. As shown in the next chapter, this is due to the existing seasonality effect on customer orders throughout the week. The proposed configuration represents a 0.97% improvement over the actual scenario.

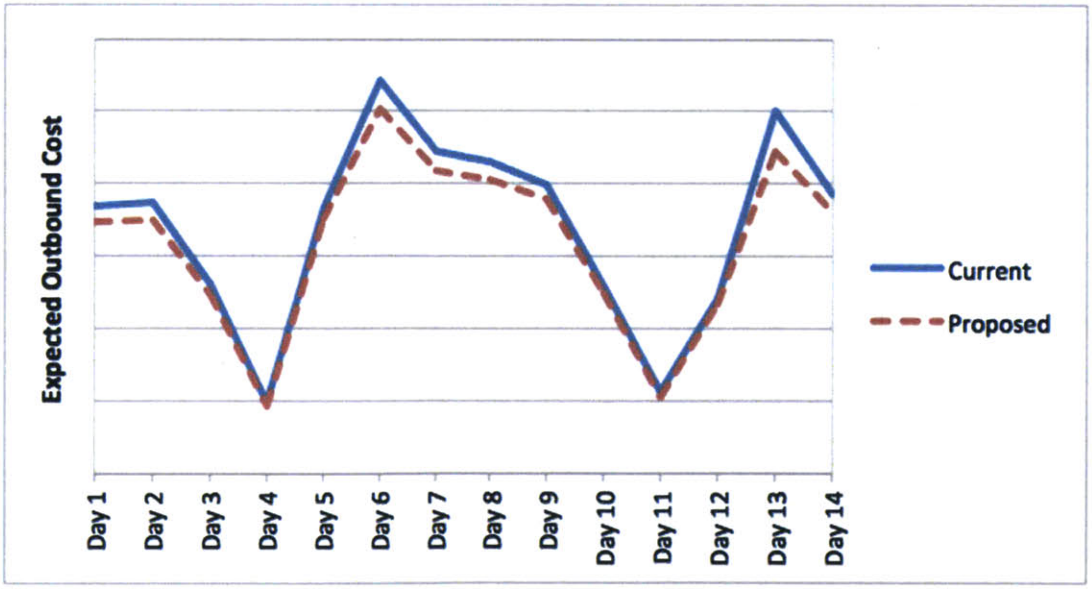


**Figure 11: Expected outbound cost from using an inflexible configuration.**

These results suggest that adding a small perturbation in the system in the form of demand variability results a small delta in performance. However, for the small difference in expected outbound costs there is a non-negligible disparity of shipping lanes selected by the algorithm. This fact highlights the sensitivity of the optimal solution to shifts in customer demand. As mentioned earlier, it would be difficult for Amazon to adapt its outbound network to such variation on a regular basis. Instead, it would be more desirable to focus on the underlying pattern of demand while overlooking the noise in the system.

A more robust model can be achieved by including a symbolic fixed cost associated with implementing changes in the current configuration. For instance, a new term could be incorporated into the objective function described in Expression 1, where F is the fixed cost associated with enabling a new shipping lane and $x_{ij}$ is the original binary decision variable. Summation of the second term includes the subset of new shipping lanes.

$$\min \left(\sum_{i,j,l,k,w} C_{ijlkw} \cdot y_{ijlkw} + \sum_{i^*,j^*} F \cdot x_{ij}\right) \tag{16}$$

Including this type of expression somehow limits the solution space to those shipping lanes that consistently yield savings outweighing the fixed cost associated with their initial setup. However, as discussed in the next chapter, demand will always have some variability that cannot be completely eliminated. For that reason, a classification system is proposed with the intent of guiding the decision of which shipping lanes to enable. The crux of the matter becomes determining the combination of shipping lanes that consistently yield the greater savings over a wide range of demand scenarios.

## 6.3 Sensitivity Analysis

No optimization program performance can be validated without a proper sensitivity analysis. Unfortunately, as noted by Guzelsoy and Ralphs (2010), duality for integer programs in not well understood yet. Computing shadow prices and reduced costs for a MILP is in itself a NP-hard problem. Current research efforts are focusing on generating dual functions to approximate the value function. Consequently, an approach based on evaluating different scenarios is required.

So far the impact of demand variability and customer aggregation have been evaluated. In addition, the associated LP relaxation problem can be solved to measure the delta between a discrete system and a continuous one. Again, seven data samples were used. By allowing the decision variables of the LP relaxation problem to take continues values, the algorithm is able to find better solutions. On the other hand, an MILP solution cannot be better than the associated LP relaxation solution because discrete numbers are a subset of real numbers. On average, the delta between both solutions is 2.43%. This implies that the mixed integer model is quiet robust.

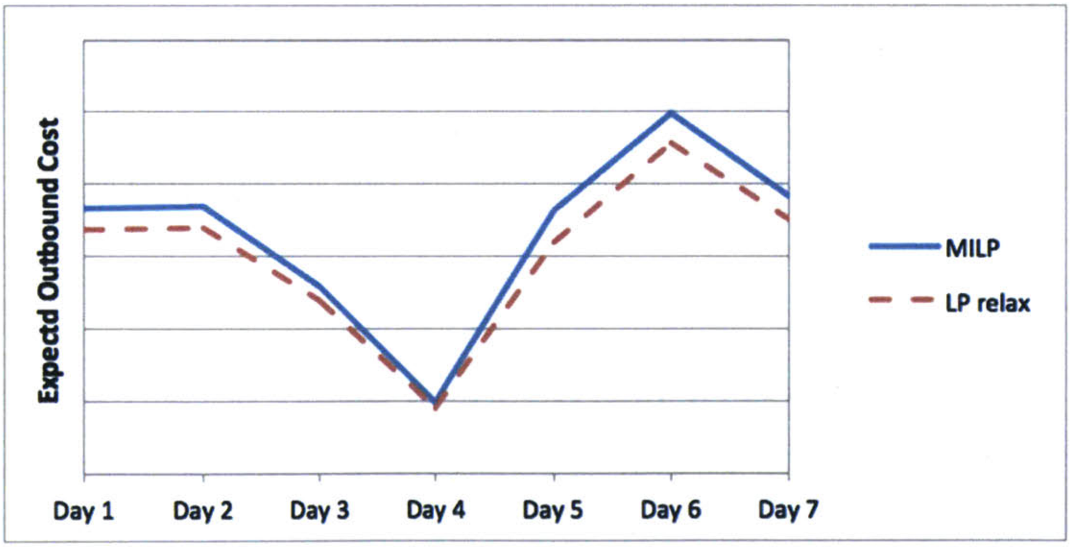


**Figure 12: Optimal solution found by MILP and LP relaxation.**

Another venue worth pursuing is the effect of capacity constraints. Figure 14 represents overall shipping cost for two fulfillment centers as a function of the number of shipping lanes available. As expected, with wider selection of shipping lanes available, the algorithm has more flexibility for finding a better solution. However, this phenomenon has a diminishing return behavior. Each facility has a threshold above which enabling additional lanes does not yield significant savings. This number should be used as the theoretical optimal solution above which building additional capacity does not report any extra benefits.

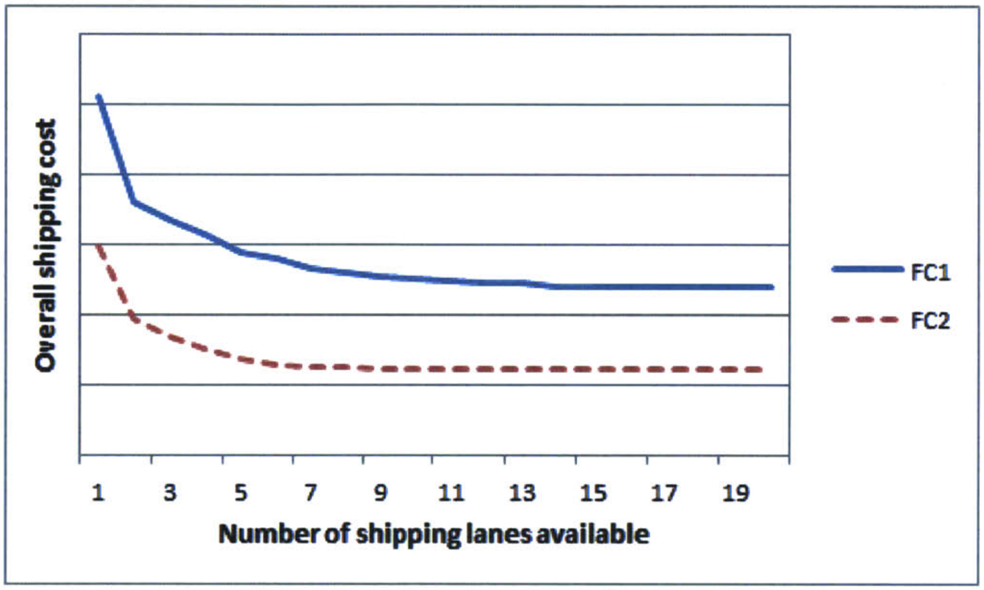


**Figure 13: Overall shipping cost as a function of shipping lanes available.**

Similar results are obtained when introduction different handling capacities at the carrier hubs.

## 6.4 Summary

Performance of the algorithm for a broad selection of scenarios suggests that the model is quiet robust. In the next chapter, a real setting is analyzed to produce a series of recommendations yielding significant bottom-line improvements.

# 7 Model Results

This section employs the refined model developed throughout the duration of the project to analyze performance of the current outbound configuration and benchmark improvement opportunities that directly translate into potential savings. The formulation proposed in this paper has been crystallized into a tool that looks at the network optimization problem in two successive steps. First, it identifies those new shipping lines with the highest saving potential. Then it simulates the outbound network by enabling those lanes and running a demand scenario. With this it is possible to obtain an estimate of the expected operational costs produced by any configuration.

## 7.1 Proposed Shipping Lane Allocation Methodology

Undoubtedly, customer demand is variable by nature. This has important implications on how much inventory and how it is allocated through the network, which in turn determines what facility can serve any particular order. Despite variability, customer orders seem to follow a general pattern by which demand peaks on Mondays and gradually decreases to a minimum on Sundays. For that reason, a robust solution cannot be found by just looking at single day, but rather considering at least an entire week.

On top of seasonality, placement of customer orders also varies within a day. Because of common trends in online buying habits, daily order placement follows a recurrent pattern by which certain time periods concentrate a significant fraction of customer orders while others have little activity. In practical terms this translates into most orders being placed between 8am and 7pm. This has important implications from a logistic standpoint. While early orders are more likely to be shipped out within the same day, late ones might not get picked until the next day. Moreover, truck departures are scheduled taking this phenomenon into consideration in order to capture as much demand as possible within the day. However, this also limits the operational range of certain ship options, as the later a truck departs the smaller service footprint it can offer. On the other hand, there are other factors to take into account, such as carrier working calendar. Some carriers do not work on the weekends, limiting the number of available shipping options to customers. Keeping this in mind, when checking out, Amazon customers are offered different ship options at different prices with a promised delivery date. All of these results in several time windows scattered throughout the week which try to accommodate customer's buying habits. From an operational stand point, this incentive system allows Amazon to group customer orders on a timely basis, facilitating order processing. For that reason, the data sample used hereafter consists of historical customer orders from a random week. The dataset has been broken down into separate days. Daily demand has then been run independently and results have been aggregated as weekly savings as shown in Figure 14. Once again, scale on the vertical axis is omitted to preserve proprietary information.

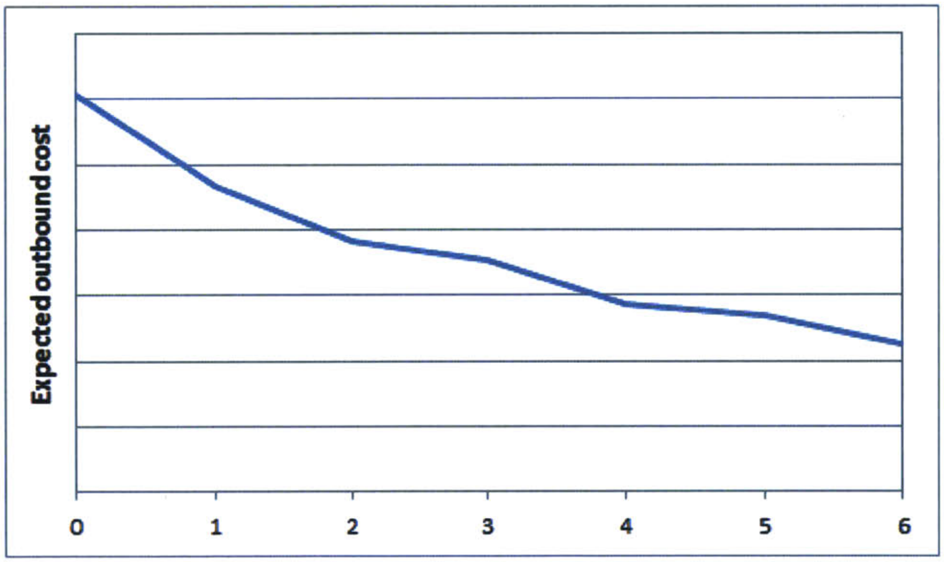
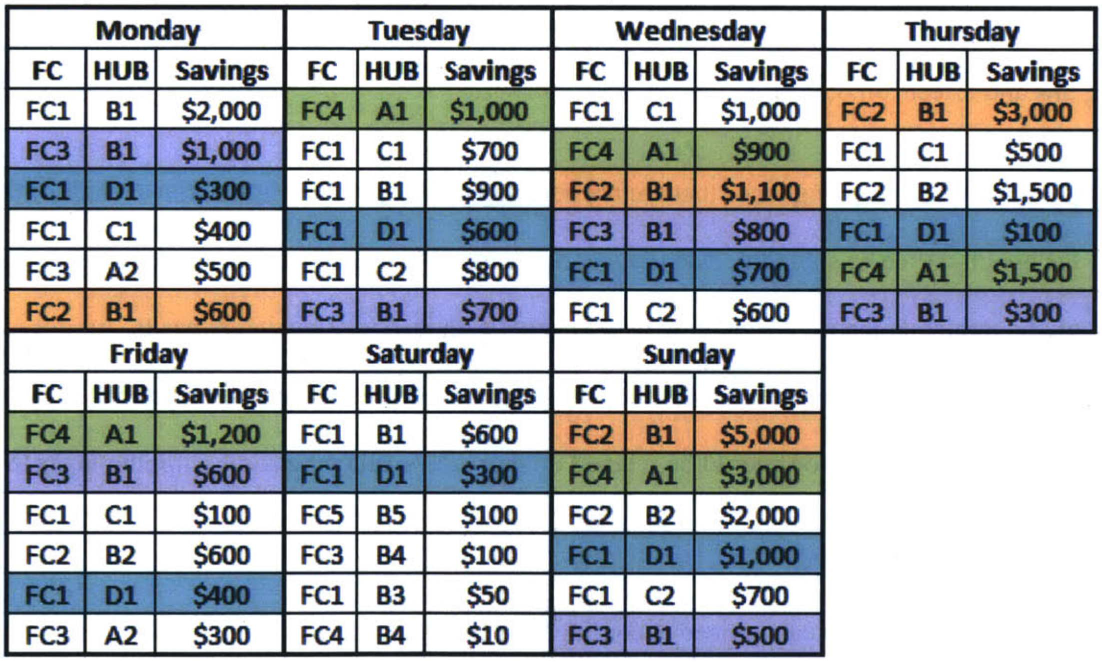


**Figure 14: Weekly fulfillment cost as a function of number of changes implemented.**

Instead of looking at the global optimal solution, the model looks at incremental changes of the outbound structure. Starting with the current network configuration, the model is allowed to make only one modification. On the next iteration, two changes are allowed. Then three changes and so on. With this approach it is possible to rank new shipping lanes on a potential savings basis and measure the delta between the current configuration and the desired scenario. Results seem to suggest that by implementing just 6 modifications in the network, expected shipping costs improve by some 1.64%. This represents the ideal upper bound for savings that could be achieved. However, because daily demand is run independently, recommendations are not consistent across the week. Indeed, some new shipping lanes are only used on a particular day, which as stated earlier poses some operational challenges.

To address this issue, a second step is included into the decision-making tool. In the first phase, the model identifies those new shipping lanes with the highest savings potential. A simulation is then run to measure the impact of implementing those recommendations. Given that not all proposals are used on every day of the week, the user can decide which ones to simulate. Figure 15 illustrates a sample output of the first step of the model. For every day of the week, top ranked six new shipping lanes are displayed with their expected daily savings.



| Monday | | | Tuesday | | | Wednesday | | | Thursday | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| FC | HUB | Savings | FC | HUB | Savings | FC | HUB | Savings | FC | HUB | Savings |
| FC1 | B1 | $2,000 | FC4 | A1 | $1,000 | FC1 | C1 | $1,000 | FC2 | B1 | $3,000 |
| FC3 | B1 | $1,000 | FC1 | C1 | $700 | FC4 | A1 | $900 | FC1 | C1 | $500 |
| FC1 | D1 | $300 | FC1 | B1 | $900 | FC2 | B1 | $1,100 | FC2 | B2 | $1,500 |
| FC1 | C1 | $400 | FC1 | D1 | $600 | FC3 | B1 | $800 | FC1 | D1 | $100 |
| FC3 | A2 | $500 | FC1 | C2 | $800 | FC1 | D1 | $700 | FC4 | A1 | $1,500 |
| FC2 | B1 | $600 | FC3 | B1 | $700 | FC1 | C2 | $600 | FC3 | B1 | $300 |

| Friday | | | Saturday | | | Sunday | | |
|---|---|---|---|---|---|---|---|---|
| FC | HUB | Savings | FC | HUB | Savings | FC | HUB | Savings |
| FC4 | A1 | $1,200 | FC1 | B1 | $600 | FC2 | B1 | $5,000 |
| FC3 | B1 | $600 | FC1 | D1 | $300 | FC4 | A1 | $3,000 |
| FC1 | C1 | $100 | FC5 | B5 | $100 | FC2 | B2 | $2,000 |
| FC2 | B2 | $600 | FC3 | B4 | $100 | FC1 | D1 | $1,000 |
| FC1 | D1 | $400 | FC1 | B3 | $50 | FC1 | C2 | $700 |
| FC3 | A2 | $300 | FC4 | B4 | $10 | FC3 | B1 | $500 |

**Figure 15: New shipping lanes identified in optimal solution.**

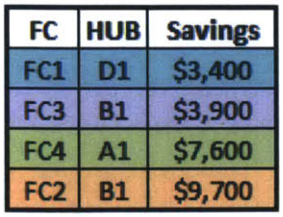
Based on their frequency or number of occurrences throughout the week, recommendations can be classified under three categories:

- **Regular**: this group contains new shipping lanes that are used every day of the week by the optimal solution. Because of its consistency, recommendations that fall under this category should be implemented whenever possible since they translate into automatic transportation savings. Lane FC1_D1, which is marked in blue, is a good example of this as it is selected on all seven days.

- **Average**: consists of shipping lanes that are used quite often, at least four days out of the week, yet not every day. Lanes FC2_B1 (orange), FC3_B1 (purple) and FC4_A1 (green) belong to this category. Implementation of this type of recommendation has to be weighed against its potential gains and effects on other existing lanes.

- **Sporadic**: those shipping lanes that are not selected in the optimal solution at least half of the time are classified as sporadic. Due to its variable nature, this genre of recommendation should only be implemented if its economic benefits outweigh its operational drawbacks. An exception to this rule could be applied on weekends. Since some carriers do not offer service on Saturday and Sunday, this criterion should be relaxed on those days, allowing some shipping lanes to be enabled only then.

Shipping proposals can also be categorized according to their potential gains. The same four shipping lanes identified above are ranked in increasing order of expected weekly savings in Figure 16. It can be seen that it is not the lane identified as regular that yields the greatest benefits. The three average lanes economically outperform the first one.



| FC | HUB | Savings |
|---|---|---|
| FC1 | D1 | $3,400 |
| FC3 | B1 | $3,900 |
| FC4 | A1 | $7,600 |
| FC2 | B1 | $9,700 |

**Figure 16: Expected weekly savings for recommended shipping lanes.**

These findings reveal an obvious trade-off between economic benefit and ease of implementation. A flexible network, which can easily adapt to demand shifts, would offer the greatest savings potential, yet its implementation poses some operational challenges. Notably, given the long lead-time for setting up a new shipping lane, putting in place recommendations from this model on a daily basis becomes unreasonable. Therefore, a more strategic approach has to be used, evaluating every proposal with different demand scenarios as well as pondering the cost benefits with its operational feasibility, which eventually can lead to a more robust network configuration.

Another element that has not been mentioned thus far is shipping lane replacement. Due to the problem formulation, there are shipping lanes that get automatically dropped because they represent more expensive ship methods than other available options. In the special case where the number of enabled shipping lanes hits the maximum capacity of the facility, then the smallest contributors to overall savings get dropped.

Going back to the decision-making tool, after identifying the desired changes to be implemented, the new network configuration can be simulated by forcing some lanes to be enabled and others to be disabled. Figure 17 shows expected benefits of implementing the same four recommendations mentioned earlier. Expected savings from adopting the four recommendations are in the order of 1.33% relative to the current outbound configuration.

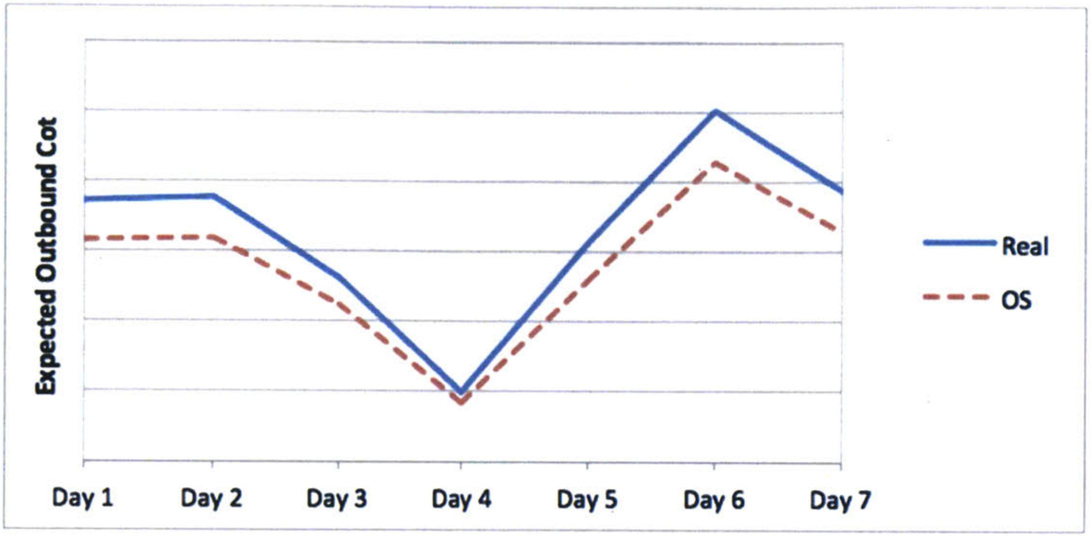


Figure 17: Expected outbound cost from implementing top 4 recommendations.

## 7.2 Carrier Hub Capacity

Besides determining the optimal network configuration, this model reveals some compelling insights, first of which is the effect of carrier hub capacity. In section 5.1, shipping cost functions for different carriers are plotted against weight of the package as shown in Figure 7. Above 15 pounds, two carriers, those identified as C and G, become the most attractive option from an economic standpoint. It just so happens that these two carriers have limited handling capacity. When running different demand scenarios using the model, their hubs get overwhelmed, soon reaching their maximum utilization as shown in Figure 18. As a result, a number of heavy orders that could benefit from the cheaper shipping rate offered by carriers C and G are instead diverted to more expensive carriers.

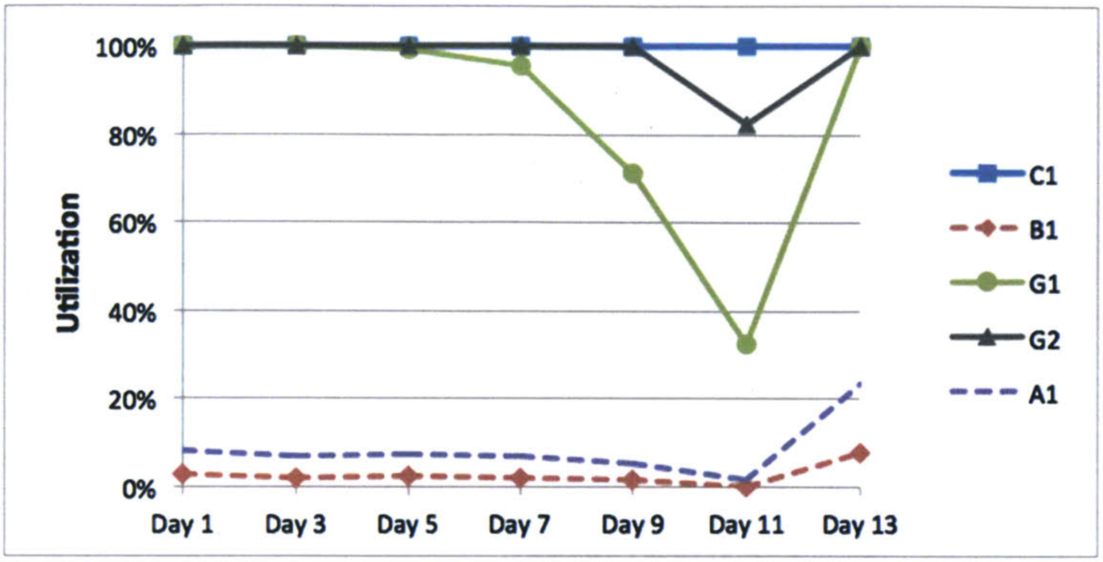


**Figure 18: Utilization rates for several carrier hubs.**

A sensitivity analysis can be used to understand the economic impact of this capacity limitation. By incrementally increasing the number of daily packages that can be processed by both carriers and running the model it is possible to quantify the potential gains from adding extra handling capacity. Figure 19 represents expected outbound cost of the network as a function of incremental handling capacity of carriers C and G. For comparison, current outbound cost and the optimal solution found on the previous section are also included. As the number of orders processed by both carriers increases, expected outbound cost is reduced. For instance, a 60% increase in handling capacity yields savings on the order of 2.56%, which results in a significant gain compared to the 1.33% found in the optimal solution. The 2.56% figure combines both, an optimal allocation of shipping lanes per FC along with increased handling capacity of carriers C and G. There is, however, a caveat to expanding capacity, as increasing the number of orders processed above 80% soon reaches a point of diminishing returns. In any case, expected savings from such an expansion have to be weighed against investment costs of increasing capacity.

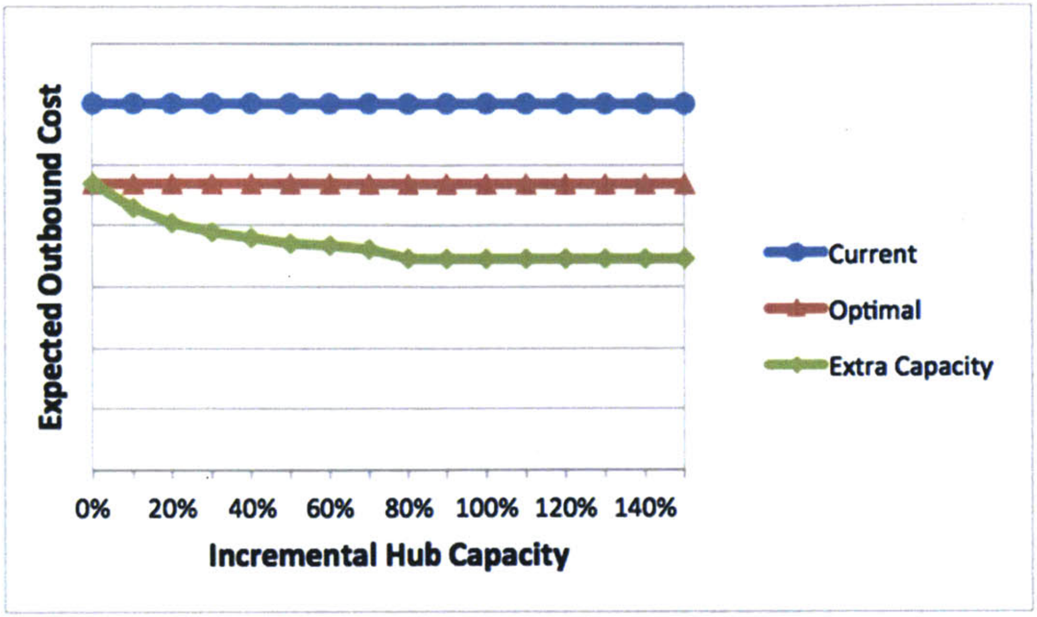


**Figure 19: Expected outbound cost as a function of carrier hub capacity.**

The other interesting finding concerns service class coverage. Most of the improvement opportunities identified by the model correspond to Standard and Super-saver order types from areas that had been historically served by the larger carriers, which have a better overall geographic coverage but are also more expensive. While smaller carriers cannot compete with them in the premium market except for specific small regions, they represent a cheaper alternative for non-urgent orders.

# 8 Conclusions and Recommendations

This section highlights the principal findings of this project and proposes future research venues to further improve performance of Amazon's supply chain.

## 8.1 Key Findings and Conclusions

While recognizing the complexity of Amazon's outbound network, the present work shows that it is possible to use an analytical approach for evaluating a distribution network as a whole and determining the most cost effective configuration. Specifically, a mixed integer network optimization model can be developed to evaluate a collection of potential shipping lanes and select those that minimize overall shipping costs. A pilot run with a wide variety of scenarios was used to prove robustness of the mathematical formulation, but also to evince the existence of a trade-off between economic gains and operational feasibility. For that reason, a criterion for selecting recommendations to be implemented is proposed, which can further be developed into a heuristic to easily evaluate and identify improvement opportunities throughout the network.

The main drawback of the model at hand is that it does not directly incorporate the inventory allocation dimension of the problem. Without evaluating the entire supply chain as a whole, solutions obtained by this type of approach do not correspond to the global optimal. The decision-making tool does not consider stock availability as a decision variable, but rather as a given. Different inventory allocation scenarios can be analyzed, but the output will be of limited use. A model assessing the trade-off between holding costs and shipping costs will reveal the greatest improvement opportunities.

The biggest challenge for extending this model to the entire distribution network is estimating the appropriate shipping cost and transit time cost functions for every region, which can be quite time consuming. However, doing so could reveal interesting system dynamics, such as the impact of transshipments across market regions.

Another caveat to keep in mind when enlarging the size of the network is the added complexity which will require additional computational resources. This could limit the granularity of demand clusters processed by the numerical solver. Nonetheless, the proposed mathematical formulation sets the foundation for developing similar optimization models, capable of handling more complex networks.

## 8.2 Opportunities for Future Research

Through the course of this project, additional related challenges were identified. Despite being outside of the scope of this work, future research could reveal interesting findings. One such opportunity is evaluating the impact of inventory allocation throughout the network. The present model presupposes how demand is assigned to the fulfillment centers. Indeed, early on customer demand was defined as allotted to a particular FC in advance. In reality, there is a complex algorithm in place that looks at every unit of inventory available at different facilities and chooses where to ship from based on economic considerations. Allowing the model to choose from which FC to fulfill every order would provide a baseline for evaluating total operational cost, including inventory holding cost as well as shipping cost. By running diverse demand scenarios, it could be possible to quantify shipment savings with different inventory mixes at a particular FC or group of facilities. Such numbers could justify a major change in stock allocation.

Another venue worth pursuing is evaluating when to expand existing capacity at the FCs. The present model can be used to assess the economic impact of adding extra lanes. In addition, it could also be used to guide decision on when to employ new carriers or invest in increasing their throughput capacity. All of these numbers would have to be weighed against the operational challenges such initiatives pose.

Finally, some tweaking would enable one to evaluate the impact of rearranging truck departures. Since some shipping lanes are more valuable than others, it would be possible to rank and estimate the savings of those that would yield greatest benefits. With this, coverage for different service classes could be enhanced to reach areas that fall outside the promised delivery date limit. Ultimately, it could be possible to develop a heuristic or basis to quickly identify those areas that would benefit from extending service footprint.

# References

[1] Acimovic J, Graves S. *Making better fulfillment decisions on the fly in an online retail environment* [working paper]. Cambridge (MA): Massachusetts Institute of Technology; 2011.March 9.

[2] Ahuja R, Magnati T, Orlin J. *Network flows: theory, algorithm, and applications*. 1st ed. Upper Saddle River (NJ): Prentice Hall; 1993.

[3] Anupindi R, Chopra S, Deshmukh S, Van Mieghem J, Zemel E. *Managing business process flows.* 2nd ed. Upper Saddle River (NJ): Prentice Hall; 2005.

[4] Bertsimas D, Freund R. *Data, models, and decisions: the fundamentals of management science*. 1st ed. Dynamic Ideas; 2004.

[5] Bloomberg News. *360buy says 2011 sales may double on China online retail demand* [internet]. 2011 February 14. Available from: http://www.bloomberg.com/news/2011-02-15/360buy-says-2011-sales-may-double-on-china-online-retail-demand.html

[6] Fourer R, Gay D, Kernigham B. *AMPL a modeling language for mathematical programming.* 2nd ed. Pacific Grove (CA): Duxbury Press; 2003.

[7] Gill M. *European online retail forecast, 2011 to 2016*. Forrester research [internet] 2012 February 27. Available from: http://www.forrester.com/European+Online+Retail+Forecast+2011+To+2016/fulltext/-/E-RES60745?docid=60745&src=RSS_2&cm_mmc=Forrester-_-RSS-_-Document-_-23

[8] Graves S, Willems S. *Optimizing the supply chain configuration for new products*. Management Science. 2005, 51(8): 1165-1180.

[9] Gumus A, Guneri A, Keles S. *Supply chain network design using an integrated neuro-fuzzy and MILP approach: a comparative approach study.* Expert Systems with Applications. 2009, 36: 12570-12577.

[10] Guzelsoy M, Ralphs T. *Integer Programming Duality*. Lehigh University. Computational Optimization Research at Lehigh [Internet]. 2010. Available from: http://coral.ie.lehigh.edu/~ted/files/papers/Duality-EOR10.pdf

[11] Manzini R, Gamberi M, Gebennini E, Regattieri A. *An integrated approach to the design the design and management of a supply chain system*. The international Journal of Advanced Manufacturing Technology. 2008, 37: 625-640.

[12] Market Research.com. *Global online retail* [internet]. 2012 July 23. Available from: http://www.marketresearch.com/MarketLine-v3883/Global-Online-Retail-7077580/

[13] McDonald J. *Options for transit time prediction* [company report]. Supply Chain Execution Team. Amazon.com. 2011 January 10.

[14] McDonald J. *A random forest kit* [company report]. Supply Chain Execution Team. Amazon.com. 2012 January 25.

[15] Melachrinoudis E, Min H. *Redesigning a warehouse network*. European Journal of Operational Research. 2007, 176(1): 210-229.

[16] Mulpuru S. *US online retail forecast, 2011 to 2016*. Forrester research [internet] 2012 February 27. Available from: http://www.forrester.com/US+Online+Retail+Forecast+2011+To+2016/fulltext/-/E-RES60672?docid=60672

[17] Simchi-Levi D, Kaminsky P, Simchi-Levi E. *Designing and managing the supply chain: concepts, strategies and case studies*. 3rd ed. New York: McGraw-Hill/Irwin; 2007.

[18] Sterman D. *If you own this tech giant, then it may be time to sell*. Seeking Alpha [internet] 2012 June 4. Available from: http://seekingalpha.com/article/635241-if-you-own-this-tech-giant-then-it-may-be-time-to-sell

[19] Tsiakis P, Papageorgiou L. *Optimal production allocation and distribution supply chain networks*. International Journal of Production Economics. 2008, 111: 468-483.

[20] US Securities and Exchange Commission. *Form 10-K from Amazon.com Inc* [internet]. Accessed 2012 December 5. Available from: http://www.sec.gov/Archives/edgar/data/1018724/000119312512032846/d269317d10k.htm