Forecasting Linehaul Transit Times & On Time Delivery Probability Using Quantile Regression Forests

ARCHIVES

by
Gold Truong

B.S. Applied Mathematics, Columbia University, 2008

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

Certified by Signature redacted
Yanchong Karen Zheng, Thesis Supervisor
Assistant Professor of Management Science, MIT Sloan School of Management

Certified by Signature redacted
Christopher Caplice, Thesis Supervisor
Executive Director CTL, Center for Transportation and Logistics

Certified by Signature redacted
Warren Seering, Mechanical Engineering Reader
Weber Shaughness Professor of Mechanical Engineering

Accepted by Signature redacted
David E. Hardt, Chairman of the Committee of Graduate Students
Department of Mechanical Engineering

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

This page intentionally left blank.

Forecasting Linehaul Transit Times & On Time Delivery Probability Using Quantile Regression Forests

by

Gold Truong

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

The complexity of linehaul scheduling is due to the numerous root causes associated with delays on the road and variabilities introduced by the major participants in the process, ie: distribution centers, drivers, etc. These sources of variability also make it difficult to measure the impact changes in transit time have on on-time performance. This paper focuses on trying to identify indicators of variability and incorporates them into quantile regression forest, a black box forecasting model, that will provide estimated scheduled transit times for a given probability of on-time arrival at the destination.

With the use of Amazon's Q1 & Q2 2013 linehaul data, an analysis on performance trends based on length of haul were categorized to develop an understanding linehauls in North America. The outbound transportation team at Amazon faces the complex trade off between providing a sufficient amount of scheduled transit time to ensure on-time delivery to destination and the utilization rate of a truck. The ability to quantify how changes in scheduled transit time impact the performance of a particular linehaul allows transportation managers to assess this trade off.

The paper explores a machine learning regression technique called quantile regression forests. The model was developed in R using the *quantregforest* package. It incorporates numerous factors about linehaul including: origin, destination, historical reporting on sources of late to arrivals, time to depart from origin and time of departure. The strengths of this black box model are in its ability to handle a large amount of data and continuously update its predicting structure to provide more accurate recommendations. Quantile regression forests also enable the user to specify the on-time performance percentage, p, that he/she wants the model to predict based on historical data. The final model at p = 95% provided a weight mean absolute percent error of 4.57% and a root mean square error of 2.22%. A four-week pilot was conducted to validate these predictions and the results are discussed.

Thesis Supervisor: Yanchong Karen Zheng
Title: Assistant Professor of Management Science, MIT Sloan School of Management

Thesis Supervisor: Christopher Caplice
Title: Executive Director of CTL, Center for Transportation and Logistics

Thesis Reader: Warren Seering
Title: Weber-Shaugness Professor of Mechanical Engineering

This page intentionally left blank.

Acknowledgments

This thesis would not have been possible without the generous help and support of many individuals. I would like to thank my manager Mark Michener, who without his support this project would not have been possible. A special thanks goes to Jack Cox, whose expertise in the trucking industry provided valuable insight that helped develop the many iterations of the model. This project also would not have been possible without the help of Dr. John McDonald and Daniel Toone. Thank you for the hours you spent working through the algorithm and code with me. I'd also like to thank the Amazon outbound transportation team for all their help from feedback to scheduling the pilot to test the validity of my model. I am grateful for the opportunity to have worked with Amazon.com on this project and for their continued support of the Leaders for Global Operations (LGO) program.

I would also like to thank my advisers, Dr. Karen Zheng and Dr. Chris Caplice, who provided valuable guidance to help drive this thesis. In addition to my advisers, my Course 2 reader, Dr. Warren Seering, who provided invaluable advice on how to structure the writing process. I am grateful to have had the opportunity to work with such talented individuals.

My time at MIT would not have been possible without the loving support of my family, whose faith in me helped brighten many frustrating moments over the last two years. To my LGO classmates, you all have made this experience a joyful and truly memorable one. I would also like to acknowledge the Leaders for Global Operations Program for its support of this work. A particularly special thank you goes out to Molly Boatright, Wonder Haas, Kristin Lien, Moira Lein, and Bijal Mehta who made the hours of data analysis go by much faster.

Table of Contents

Abstract 3
Acknowledgments 5
1 Introduction 11
1.1 Company Overview/Project Motivation 11
1.2 Thesis Overview 11
1.2.1 Outline of Transit Time Problem 11
1.2.2 Components of Linehaul Process 12
1.2.3 Variation 13
1.2.4 Cost Savings Through Accurately Scheduled Transit Times 14
1.3 Outbound Transportation Overview 14
1.3.1 Description of Amazon and Carrier relationships 14
1.3.2 Key Terms 15
1.3.3 Current state of calculating transit time 15
1.3.4 Process for Changing Scheduled Transit Time 15
2 Literature Review 17
2.1 Industry Standards - Comparison to Airline Scheduling Problem 17
2.2 Ordinary Least Squares Linear Regression 18
2.3 Quantile Linear Regressions 20
2.4 Random Forests 22
2.5 Quantile Regression Forests 24
3 Understanding Truck Departure Process from Amazon Facilities 26
3.1 Overview of Truck Departure Process & the Last Truck Out Time 27
3.2 Kaizen Teams & Findings 28
3.2.1 Dock Team 28
3.2.2 Ship Clerk Team 30
3.2.3 Yard Team 31
3.3 Kaizen Results & Implications on Modeling Transit Time 32
4 Developing An Analytical Understanding of Transit Time 34
4.1 Current Methodology: Scheduling Transit Time With Respect To Distance 34
4.2 Estimating Components of Transit Time 36
4.2.1 Validating Nominal Drive Time Estimates Using Transportation Software 36

4.2.2 Estimating Time to Depart a Truck From Amazon Facilities 37
4.2.3 Understanding Performance Variations 39
4.3 Data Cleaning 40
4.4 Variables in Model 40
5 Quantile Regression Forests 43
5.1 Selecting Lanes for the Model 43
5.2 Training & Test Data Set 43
5.3 Initial Model 44
5.4 Stability of the Model 45
5.4.1 Number of Trees 45
5.4.2 Nodesize 51
5.4.3 Final Model & Maintenance of Model 52
5.5 Testing for Variable Significance 53
5.5.1 Methodology: Permutation Test for Variable Significance 53
5.5.2 Results of the Permutation Test 54
5.5.3 Trends Between Variables 55
5.5.4 Trend across percentiles 58
5.6 General comments on predictions 59
6 Implementation of Model & Pilot 61
6.1 Selection of Lanes 61
6.2 Selection of Service Level to Test & Expected Error Rates 62
6.3 Generating Predictions 64
6.4 Results from Pilot 65
6.5 Evaluating Accuracy of Model & Performance 66
6.6 Pilot Implications on Supplier Engagement 67
6.6.1 Carrier & Truck Driver Response to Pilot 67
6.6.2 Traffic – Potentially Overused Carrier Reported Issue 68
6.7 Implementation of Model as a Tool 68
6.8 Updating and Maintaining the Model 69
6.9 Predicting Transit Times for New Linehauls 70
7 Recommendations and Conclusion 72
7.1 A Perfect Schedule? 72
7.2 Recommendations 72

7.2.1 Operational Recommendations 73
7.2.2 Data Recommendations 74
7.3 Conclusion 76
8 Appendix 77
8.1 Appendix A: Performance factors 77
8.2 Appendix B: Speed Limits For Trucks by State 78
8.3 Appendix C : Pearson Correlation Matrix 79
8.4 Appendix D: Sample Output of Tool 80
8.5 Appendix E: Note on Linear Models Attempted 81
8.6 Appendix F 84
9 References 85

Table of Figures

Figure 1-1: Graphical Representation of Nominal Transit TIme 13

Figure 2-1: Studentized Residuals of Linear Model 19

Figure 2-2: Normal Probability Plot of Linear Model 19

Figure 3-1 Results from Time Study of Yard Depart Times 26

Figure 3-2 Truck Departure Process at Amazon 27

Figure 3-3 Associate at the Dock During LTOT 29

Figure 3-3-4 Example of Yard Congestion 32

Figure 4-1 Composition of Transit Time 34

Figure 5-1 Forest Size Stability Test: Weight Mean Absolute Percent Error from 46

Figure 5-2 Forest Size Stability Test: Root Mean Square Error 49

Figure 5-3 Nodesize Test: Weighted Mean Absolute Percent Error by Nodesize 52

Figure 5-4 Gate Departure Estimates Using EDI Data (Q1 2013) 59

Figure 8-1 Late Depatures vs. Number of Trucks Departing Simultaneously 82

List of Tables

Table 1 Lane Types by Average Speed 36

Table 2 EDI Compliance Rates 39

Table 3 Distribution of Load Types Within Test & Training Datasets 43

Table 4 Values for the Initial Model Developed 44

Table 5 Forest Size Test: Weighted Mean Absolute Percent Error of Untouched Dataset 48

Table 6 Forest Size Stability Test: Root Mean Square Error for Untouched Data Set 51

Table 7 Baseline WMAPE Rates by Probabiliity of On Time Delivery 54

Table 8 Results from Permutation Test: Percent Error from Baseline Error Rate 54

Table 9 Error Rates of Model at Specified On Time Probability 62

Table 10 Historical Performance Prior to Pilot 63

Table 11 Scenarios By Percentage in Network & Subset of Lanes in Model 64

Table 12 Summary of Results from Pilot 65

Table 13 Reasons for Lateness Due to Reduction in Transit Time 65

Table 14 Predicted vs. Demonstrated Error Values for Model 67

1 Introduction

1.1 Company Overview/Project Motivation

Jeff Bezos founded Amazon.com as an online textbook retailer in 1995. Today, Amazon is a global retailer that aims to be the "Earth's most customer centric company." As part of this mission, Amazon's Transportation department strives to deliver their goods as quickly and cost efficiently as possible. As Amazon continues to offer more competitive shipping options, such as same day, next day and second day, balancing on-time delivery and transportation costs becomes a bigger challenge (Thomas). The purpose of this project is to better forecast scheduled transit times between Amazon distribution centers and cross dock facilities.

The estimation of transit time is both a logistical and financial issue. There is an inherent tradeoff between allotted time to fulfill an order and the allotted time to deliver the order. The shorter the transit time, the greater the number of orders that may be fulfilled before the truck must depart the facility. However, a shorter scheduled transit time increases the risk of a late delivery, and can negatively impact customer experience. Conversely, a long transit time decreases the risk of a late arrival, with the tradeoff of a lower utilization rate of the transportation assets. This project attempts to quantify the trade off between on time delivery with regards to the calculation of scheduled transit time.

1.2 Thesis Overview

1.2.1 Outline of Transit Time Problem

Amazon's Transportation team works with transportation carrier companies to transport packages from Amazon warehouses to regional transporters, who then complete the final mile delivery to customer. These carriers are also contracted to perform intra-Amazon network inventory transfers. Both types of routes are referred to as outbound linehauls. The project is focused on outbound linehauls within North America. More specifically, the dataset considered in this paper are linehauls with single destination, which make up 99% of the routes at Amazon in North America.

Rudimentary calculations of transit time can be made using length of haul. This type of calculation does not examine the factors that can influence on-time performance, such as: time to depart from origin or variations in performance based on day of the week, time of day of

departure, destination type (external customer vs. internal customer), and seasonality in weather trends. By ignoring sources of variability in the transit time calculation, this type of calculation cannot anticipate how these factors might affect the on-time delivery performance. Additionally, the ability quantify the change in transit time to a change in percentage of on-time delivery is also needed to understand how incremental changes impact performance. This gap in the ability to evaluate changes in transit time has led to one off adjustments, based on short-term performance, to improve on time delivery. Without a more sophisticated calculation methodology, overly conservative scheduled transit time estimates may be unnecessarily increasing the cost per package for a retailer. My thesis investigates forecasting techniques that attempt to incorporate these factors of variability.

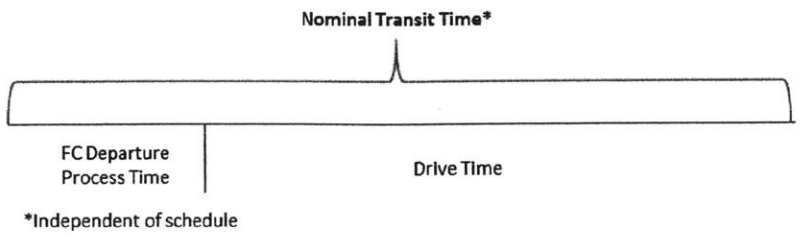
1.2.1.1 A Perfect Schedule

Scheduled transit time is defined as the time needed for a truck to be pulled off of the dock until the time it arrives at its destination. This thesis attempts to answer the question: how would a perfect schedule be designed? In order to do so, characteristics of specific loads will be examined to determine which characteristics are most influential to transit time estimation. To qualify as a perfect schedule, every truckload should be scheduled such that it is maximizes the number of packages on a truck and minimizes the amount of scheduled transit to minimize transportation cost per package. The difficulty in creating a perfect schedule not only lies in predicting the necessary transit time for each route but also negotiating arrival times with carriers and departure times with the distribution centers. The scope of this project will only address the calculation of scheduled transit time and process improvements to the truck departure process.

1.2.2 Components of Linehaul Process

The outbound linehaul process can be broken down into two components. The first component is the time it takes for a truck that has been loaded to depart the yard. The second component is the actual drive time to the destination. These two processes occur sequentially and independently of each other. Assuming that both events occur as efficiently as possible, the minimum amount of time required to do both will be defined as the Nominal Transit Time.

Figure 1-1: Graphical Representation of Nominal Transit Time



Actual transit time can then be decomposed into two components:

Actual Transit Time = Nominal Transit Time + factors that contribute to variability

Therefore, any variations in performance can be stated as the difference between Actual Transit Time and Nominal Transit time.

Variation = Actual Transit Time – Nominal Transit Time

The model incorporates characteristics that may be attributed to the variations at each part of the outbound linehaul process. It attempts to attribute these characteristics to quantifiable delays through the use of quantile regression forests.

1.2.3 Variation

Demonstrated performance can be affected by numerous factors; some of these factors are predictable, others are not. These factors, for the purposes in this paper, will be classified into three different categories: *Origin facility, carrier controllable* and *other*. A full list of these reason codes can be found in Appendix A. The *other* category contains stochastic variation reasons such as accidents, traffic, weather, DOT inspections, etc. which are reasonable and uncontrollable explanations for delays. *Origin facility* and *carrier controllable* can be considered areas for process and performance improvement.

1.2.3.1 Variation Related to the Departure Process at Origin Facilities

There are two types of delays that can occur at the origin facility: delays caused by warehouse operations and late truck arrival to the pick up site. Amazon may experience internal delays in loading the truck due to issues upstream in the warehousing process.

1.2.3.2 Variations Related to Drive Time

Drive time related variations include *carrier controllable* and *other* reason codes. In most cases the drive time makes up a majority of the scheduled transit time and is also where a significant amount of uncontrollable variability is introduced. This is where traffic, accidents and DOT inspections, which are classified in the *other* category, may cause truck drivers to be heavily delayed. Carriers can introduce variability through errors such as driver getting lost, error, mechanical breakdowns and dispatch errors. Effective carrier management practices are documented into a report and were available to provide a historical perspective on carrier performance.

1.2.4 Cost Savings Through Accurately Scheduled Transit Times

Overly scheduled transit times effectively increase shipping costs. Amazon prioritizes customer experience and will do everything possible to ensure the best customer experience, which includes on-time delivery (Green). With overly conservative transit times, many packages may be shipping with more expensive ship methods when they could be executed with lower shipping costs. Anecdotally, it is understood that today's scheduled transit times are heavily padded to ensure on-time delivery. By understanding how much transit time will ensure a certain service level of on-time performance, scheduling teams can be intelligently increase or reduce transit time where needed. Reduction of transit time will realize cost savings. Increasing transit time may increases costs but maintain Amazon's goal of ensuring on-time delivery to customer.

1.3 Outbound Transportation Overview

1.3.1 Description of Amazon and Carrier relationships

Transportation carriers within North America for trucking come in two forms: asset based carriers and brokerage carriers. Asset based carriers are trucking companies who own their own transportation assets, tractors and trailers, and employ drivers to execute driving assignments. Brokerage carriers may or may not own their own assets and generally auction the driving assignments to third party logistics companies who execute the assignment. Common performance metrics to evaluate a carrier on is on percentage on-time to destination. For late loads, it is common to work with carriers to identify root causes on a case-by-case basis. These

industry best practices have been adopted by Amazon and are used to help drive process improvement and evaluate carrier performance across the whole network.

1.3.2 Key Terms

This section will define transportation terms that will be used throughout this paper.

1. **Linehaul** – an origin-destination pair that has been assigned to a specific carrier to execute, also known as a lane or route
2. **Last Truck Out Time (LTOT)** – a calculated and scheduled time, the time a truck needs to be closed in order to ensure on-time delivery to destination
3. **Last Truck In Time (LTIT)** – a time provided by the destination in which loads must be at the destination in order to ensure on-time delivery to customer
4. **Scheduled Departure Time –** the scheduled time to close a trailer for each load
5. **Scheduled Arrival Time –** the scheduled arrival time communicated to a carrier to arrive for pick up
6. **Transit Time** – the time from Scheduled Pull Time to the Last Truck In Time at the destination
7. **Last Truck Out (LTO)** – the last truck at the specified Last Truck Out Time for the lane
8. **Sweeper Truck** – a truck that is scheduled but not the LTO
9. **Adhoc Truck** – a truck that is not scheduled and is created due to operational needs

1.3.3 Current state of calculating transit time

The current methodology for calculating transit time is a general formula that is applied across all routes in the Amazon North American network. Each route may have their transit time adjusted based on historical performance. The tendency is to extend transit time based on inconsistent carrier performance, rather than reduce transit time. About 70% of linehauls had scheduled transit times that were greater than the calculated value. The formula has two inputs: the distance between origin and destination in miles and the drive time estimated by transportation software based on 5 digit origin and destination zip codes. Transit times may also be modified based on operating hours of the origin and destination buildings

1.3.4 Process for Changing Scheduled Transit Time

Changes to the schedule are completed on an as-needed basis and are negotiated between the carrier managers, the carriers and the origin facilities. A typical request could begin as feedback

a carrier receives from a truck driver that a particular route has become more challenging to drive because of new road construction. This information is presented to the scheduling team to extend transit by a certain amount that was suggested by the carrier. The scheduling team may perform a back of the envelope calculation to determine the reasonability of the request. This calculation may consist of comparing recent changes requested by other lanes that drive through a similar area or comparing existing transit times of similar distances. There is no consistent methodology to determine if the request is valid.

After the scheduling team vets this request, it must be compared to the Last Truck In Times set by the destination. This is to ensure that the new extended transit time request would not require changing the existing Last Truck Out Time. If either LTIT or LTOT are compromised, negotiations with the warehouse begin. Changing the LTIT of a lane can influence staffing procedures and other processes at the warehouse, such as scheduled breaks.

These change requests can take anywhere from two weeks to one month to go into effect. Unfortunately the negotiation process is not documented. On time delivery is monitored after the change but there is no feedback system to determine if the change was sufficient or excessive. Amazon only knows if the change was deficient because on-time performance would remain low. The lack of documentation can cause overly inflated transit times. There is also no review process to reduce transit time once the issue that caused the increase in transit time is relieved. This is particularly true of transit times that are increased due to increased seasonal inclement weather.

A deliverable of this project will be to create a system that allows the transit time forecasting tool to take these changes into account.

2 Literature Review

2.1 Industry Standards - Comparison to Airline Scheduling Problem

The scheduling issue faced by Amazon is not dissimilar to the scheduling problem faced by airline companies. A thesis written by Gerasimos Skaltsas titled Analysis of Airline Schedule Padding on U.S. Domestic Routes (MIT, 2011) provides some insights into how to analyze the problem of scheduling transit times for trucks. The thesis focused on analyzing how airlines used block time to account for variabilities in flight operations. Block time is defined by Skaltsas as the "time interval between the gate departure and the gate arrival time for a given flight" (Skaltsas 41). Skaltsas examined correlations between buffer and flight time components to develop a linear regression model to analyze trends. Using his model, he was able to understand the impact of each component of variability on flight time and studied how various airlines accounted for them through their scheduling strategies.

One concept that Skaltsas discusses is Nominal Block Time. The Nominal Block Time is defined as the time required to complete the flight, including taxi time in and out of airports, in optimal conditions. The difference between Actual Transit and Nominal Block Time is defined as buffer. This concept was adopted in my analysis of transit time for Amazon, as defined earlier in section 1.2.2. The current method of calculating transit time utilizes a third party transportation software, to estimate a nominal drive time. My calculation of nominal transit time for transit time analysis is based on the same estimate from the third party transportation software.

An additional concept that was helpful in framing Amazon's transit time problem from Skaltsas' work was calculating gate delays. Gate delay is defined as the time difference between scheduled departure time and actual departure time. For airlines, gate delay can cause a late arrival to destination. These delays can be caused by any number of variables such as weather conditions, airline policies, mechanical problems, baggage handling etc. (Skaltsas 2011). Amazon has similar issues with the truck departure process from the dock. Since departing the yard is the first step in the delivery process, it is important to understand the variabilities that are introduced during this process. A more in-depth discussion of lessons from a kaizen will explore the causes of these variabilities in Chapter 3.

In general, Skaltsas found that there was a weak linearity between buffer and nominal airborne time. A practice that was common among all airlines examined was buffer increased with length of flight time (Skaltsas 76). The rationale for this was to account for the increased uncertainty

en-route, however buffer as a fraction of nominal block time decreases exponentially with nominal airborne time. This is believed to be due to the fact that absolute delays do not change with the nominal airborne time. Therefore on short haul flights, buffer is often a significant fraction of the scheduled block time. Skaltsas concluded that,

> *"An optimal padding strategy maximizes the number of flights that arrive on time and at the same time minimizes the total negative delay. The long right tail in the gate and block delay distribution reveals the existence of flights that require a large amount of buffer time to arrive on time. Because these flights are distributed over a wide time range, the gains in on-time performance would be very small compared to the cost of underutilizing the aircraft, the gates and the crew for every minute of early arrival" (Skaltasas 2011, page 49).*

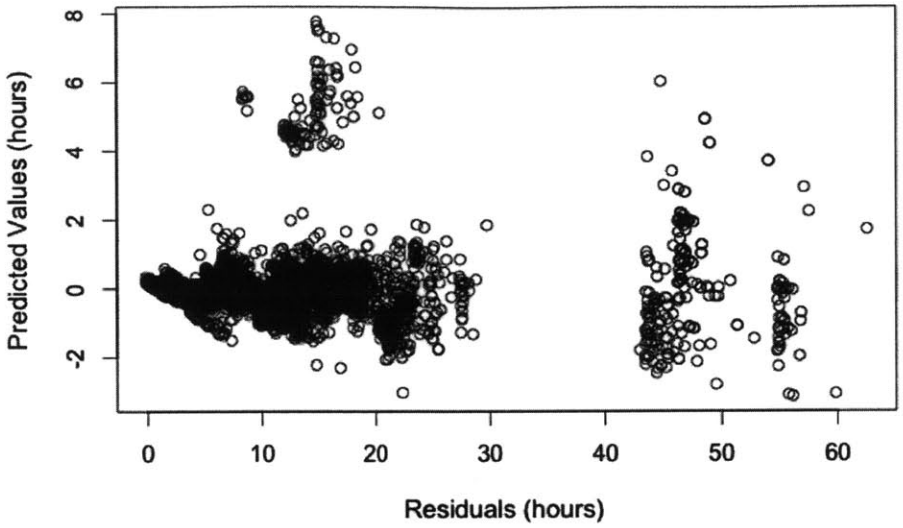
This conclusion may also be true for Amazon's scheduling issue. If this is true, the current goal of high percentage of on time delivery may result in highly buffered transit times and highly under utilized capacity on the trucks.

2.2 Ordinary Least Squares Linear Regression

The work conducted by Skaltsas was meaningful as a methodology to draw conclusions on the effects of various sources of variation associated with demontrated on-time performance. Unfortunately, his methodology of using ordinary least squares regression is not a good methodology for Amazon's data set. An ordinary least squares model makes several assumptions about the behavior of the data that are not true of Amazon's data set. Specifically, the assumption of linearity, homoscedasticity of the errors and normality of the error distribution are not true. By plotting the studentized residuals and normal probability plot of the standard residuals, it can easily be seen that the assumption of normal error distribution and homoscedasticity are not true.

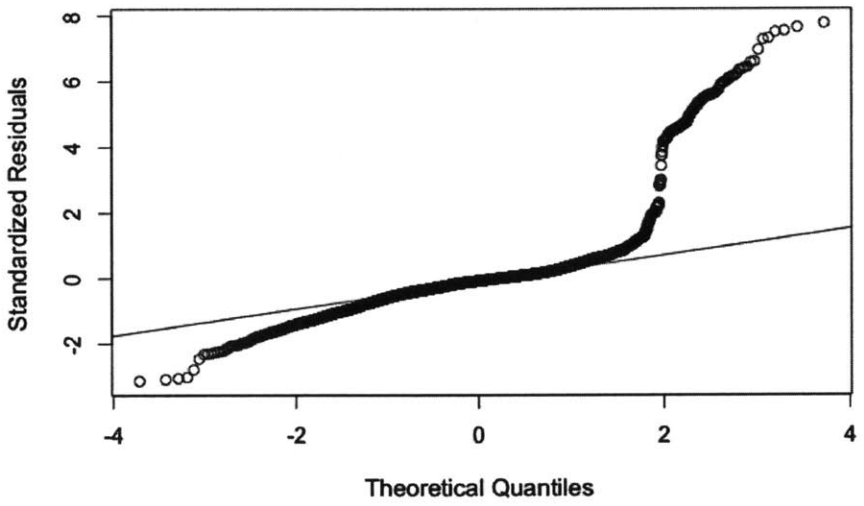
The first model that was attempted with the data set was a linear regression. It incorporated variables that measured Origin Arrival Delay, Dock Depart Delay, Gate Depart Delay, Destination Arrival Delay, Number of Scheduled Trucks Departing Simultaneously , Day of Week, Length of Haul, Time of Day and reported late reasons for historical loads. The initial OLS model was discarded however these variables were later considered in the quantile regression forests model.

Figure 2-1: Studentized Residuals of Linear Model



Note: If homoscedasticity were true, the plot of studentized residuals vs. predicted values would appear random about 0. The above plot clearly shows a pattern within the residuals that make us reject this assumption.

Figure 2-2: Normal Probability Plot of Linear Model



Note: If the errors associated with the linear regression were normal, the normal probability plot of studentized residuals would lie along a straight line. The normal probability plot above clearly points to reject the assumption of a normal distribution of errors.

Lastly, ordinary least squares regressions provide little ability to interpret service levels based on the predicted transit time. The ordinary least squares regression produces a value that is the conditional mean of the dependent variable. Unfortunately, that does not answer the question how much would be gained or lost through adding and reducing transit time. For this reason, the ordinary least squares regression model was abandoned.

2.3 Quantile Linear Regressions

i. Background & Definition

The need to specify transit times at specific service levels cannot be met by using linear regressions. With least square regressions, the following loss function is minimized:

$$\sum_{i=1}^{n} |Y - y_i|^2$$

$$where\ Y\ is\ the\ predicted\ value,$$

$$y_i\ is\ the\ i^{th}\ actual\ demonstrated\ value$$

The values produced from a least squares regression are an estimate of a conditional mean, *E(YDC | X=x),* based on independent variables x_i, where i={1,2,3...n} and n is the length of *x*. Unfortunately, a least squares regression is limited in its ability to recommend transit times for a given service level because it only estimates a conditional mean.

With quantile regressions, the following conditional loss function is minimized:

$$\sum_{i=1}^{n} p|Y - y_i|^+ + (1-p)|Y - y_i|^-$$

$$where\ Y\ is\ the\ predicted\ value\ ,$$

$$y_i is\ the\ i^{th}\ actual\ demonstrated\ value$$

$$and\ 0 \leq p\ \leq 1$$

By minimizing the above loss function, the regression produced provides greater information about *Y* than an OLS model. The quantile regression is a conditional distribution function, *F (y | X=x) that* is defined as:

$$F(y|X = x) = P(Y \leq y|X = x)\ \geq p$$

$$where\ \ 0 \leq p \leq 1$$

By interpreting *p* as the probability of a load delivering within a specified transit time *y*, the regression may be used to provide forecasts for a specified service level.

ii. Interpreting error rates at specified service levels

By interpreting *p* as the probability of a load delivering within a specified transit time *y*, the regression may be used to provide forecasts for a specified service level.

ii. Interpreting error rates at specified service levels

By minimizing the conditional loss function mathematically, we inherently favor transit times that overestimate the actual transit by definition. Therefore, rather than computing the mean absolute percent error (MAPE) of the prediction, we use the following formula to compute weighted mean absolute error rate, *WMAPE*:

$$WMAPE = \frac{1}{n}\left(\sum_{i=1}^{n}(p * |A_i - T_i|^{+} + (1-p) * |A_i - T_i|^{-}) * \frac{1}{A_i}\right) * 100$$

where p = % On-Time delivery, A_i = i^{th} Actual Transit Time, T_i = i^{th} Predicted Transit Time

Hence, an x% weighted mean absolute percent error rate for a given p implies that (1-x)% of time, the delivery will be on time with a probability of at least p, *where* $0 \leq p \leq 1$.

The precision of the model is also important, since the model should not overestimate transit times for the benefit of lower error rates. Therefore we will also measure the mean absolute percentage error, MAPE, as:

$$MAPE = \frac{1}{n}\left(\sum_{i=1}^{n}(|A_i - T_i|) * \frac{1}{A_i}\right) * 100$$

iii. Benefits of Quantile Regressions

The benefit of using a quantile regression is to allow the user to interpret each of the regressors' influence on the predicted value at specified probabilities. It is possible that certain regressors exhibit a greater influence over the predicted value within specific probability ranges. For example, with transit times, it may be true that Amazon performance factors might exhibit a greater influence on on-time performance at lower service levels, however at higher service levels, these performance factors may have little effect on the predicted transit time value. Quantile regressions compute the coefficients related to the regression based on the specified service level.

Another benefit of using quantile regression is that it does not make any assumptions about the underlying behavior of the data. Therefore, issues of non-homoscedasticity and non-normal error distributions are not an issue with quantile regressions.

iv. Limitations of Quantile Regressions

While quantile linear regressions take into account the varying degrees of influence from the variables at specified service levels, it is not very good at handling outliers. Due to the nature of transit times, large delays can cause huge outliers to occur. These outliers caused very conservative estimates to be produced at high service levels. The quantile regression also did not provide a good estimate for lanes that had a tendency of missing the receiving window at the destination. Due to the miss, the carrier would have to wait up to 12-24 hours until the facility opened before delivering the trailer. Carriers will commonly not report the time of arrival at destination if the facility is closed. Therefore, on certain routes, there were a significant number of deliveries that appeared to need transit of at least 12 hours over their current scheduled transit. In reality, had the truck arrived a few minutes or hours earlier, the prediction would be much closer to the actual transit time, ie: time to depart from origin and drive to the facility, rather than the reported transit time, ie: when the arrival was reported by the carrier. While the quantile regression effectively minimizes the loss function, its MAPE was roughly 33.8%. It provided conservative estimates for many lanes that would be unreasonable to ask Amazon to schedule to. A more robust modeling method for handling outliers was needed to appropriately model transit time.

2.4 Random Forests

Random forests were suggested as a modeling technique that would be able to handle the outliers prevalent in the transit time data set. Below is a formal definition of random forests:

A random forest is a classifier consisting of a collection of tree-structured classifiers {DC(x,t_k), k = 1,...} where the {t_k} are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input x (Breiman 6)

Random forests is a machine learning technique that generates a large number of decision trees, which are used to estimate the dependent variable. Each tree is created by randomly

subsetting the data using the independent variables and their corresponding dependent variables. Within each tree, the data is randomly sampled to generate predictions. These samples are referred to as leaves.

In addition to their ability to handle outliers well, random forests also have other benefits[1] such as:

1. Ability to handle extraneous variables that may not influence the dependent variable
2. Convergence of predictions so over fitting is not a concern
3. Demonstrated high accuracy in predictions

Leo Breiman describes the technique in his paper Random Forest (University of California – Berkley, Sept 1999). This paper is the primary source for the R library that has been developed for random forests and quantile regression forests. My model was developed with these R packages and uses quantile regression forests. The package, quantregForest, is heavily modeled after the R library randomForest. The description of the modeling technique will be described using the terminology from the R packages.

Within the R package, the user may define the minimum number of samples that the tree must have for a given leaf using *nodesize*. For a given *n* set of variables, a random sample is taken of *mtry* variables, defined by a user input. *Mtry* is a setting within the Random Forest function that may be any integer value less than or equal to *n*. It is the number of input variables that will be randomly selected to generate a given tree. It is common to split the data into thirds in order to cross validate the results and prevent over fitting the model. It is assumed that the original data set has n samples. The probability of a randomly sampled data set is missing by sampling n times with replacement from the original data set is:

$$P\,(\,sample\ is\ missing\,) = \left(1 - \frac{1}{n}\right)^n$$

Therefore the limit of this expression as n grows towards infinity is 1/e or approximately 1/3. It is for this reason that a general rule of practice is to set one-third of the n- independent variables equal to *mtry.*

Based on this randomly generated tree, a leaf is created for each unique set of observations, *x*. Let that leaf be denoted by *l(x), where x is a vector.* A prediction is created for every tree *T* in

[1] Breiman, 2001
[2] http://DC.bigrigdriving.com/2010/trucking-industry-debates/is-team-driving-the-future-in-trucking

the forest. The prediction y_i is created by calculating the weighted average over the observed values, Y_i, in $I(x)$ over all trees and leaves.

The weight, w_i, assigned to an observation x_i is a positive constant equal to $\frac{1}{(\#\ of\ observations\ equal\ to\ x_i)}$ if x_i is part of the leaf $I(x)$ and 0, otherwise.

$$prediction\ from\ a\ single\ tree, T:\ \mu_T(x) = \sum_{i=1}^{n} w_{Ti}(x) * Y_i$$

$$where \sum w_{Ti}(x) = 1\ and\ Y_i\ are\ observed\ values$$

To compute the prediction from multiple trees, the weights from all leaves are aggregated and averaged.

$$prediction\ from\ forest: \mu(x) = \sum_{i=1}^{n} w_i(x) * Y_i$$

$$where\ w_i = \frac{\sum_{T=1}^{k} w_{Ti}}{k}, k = total\ number\ of\ trees$$

Ntree is also set by the user as the number of trees to be generated in the forest. Once the entire forest has been grown through *ntree* random samples of the entire dataset, a new prediction can be made by inputting a vector X of n variables into the random forest. The new vector X will be applied across all the trees in the forest and the predictor is calculated by taking a weighted average of all the observed values on all leaves in the random forest for all observations that match X.

2.5 Quantile Regression Forests

i. Background & definitions

Like ordinary least squares regressions, random forests estimate a conditional mean, which is equivalent to the weighted mean of the observed y_i. In contrast, quantile regression forests use a conditional distribution to estimate the weighted distribution of the observed y_i, where the weights attached to observations are identical to the original random forest algorithm.

Combining the concepts from quantile regressions and the methodology of random trees to generate accurate predictions, we can say that the conditional distribution function of DC, given X = x, is:

$$F(y|X = x) = P(Y \leq y|X = x) = E(1_{\{Y \leq y\}}|X = x) \geq p$$

$$where\ 0 \leq p \leq 1,$$

$$1_{\{Y_i \leq y\}} = 1\ when\ Y_i \leq y, 0\ otherwise$$

Therefore predictions of y_i can be stated as:

$$y_i = \hat{F}(y|X = x) = \sum_{i=1}^{n} w_i(x) * 1_{\{Y_i \leq y\}}$$

ii. Implementation in R

The model was constructed in R with the use of the library package quantregForest. The package was developed by Nicolai Meinshausen and the algorithm can be summarized as follows:

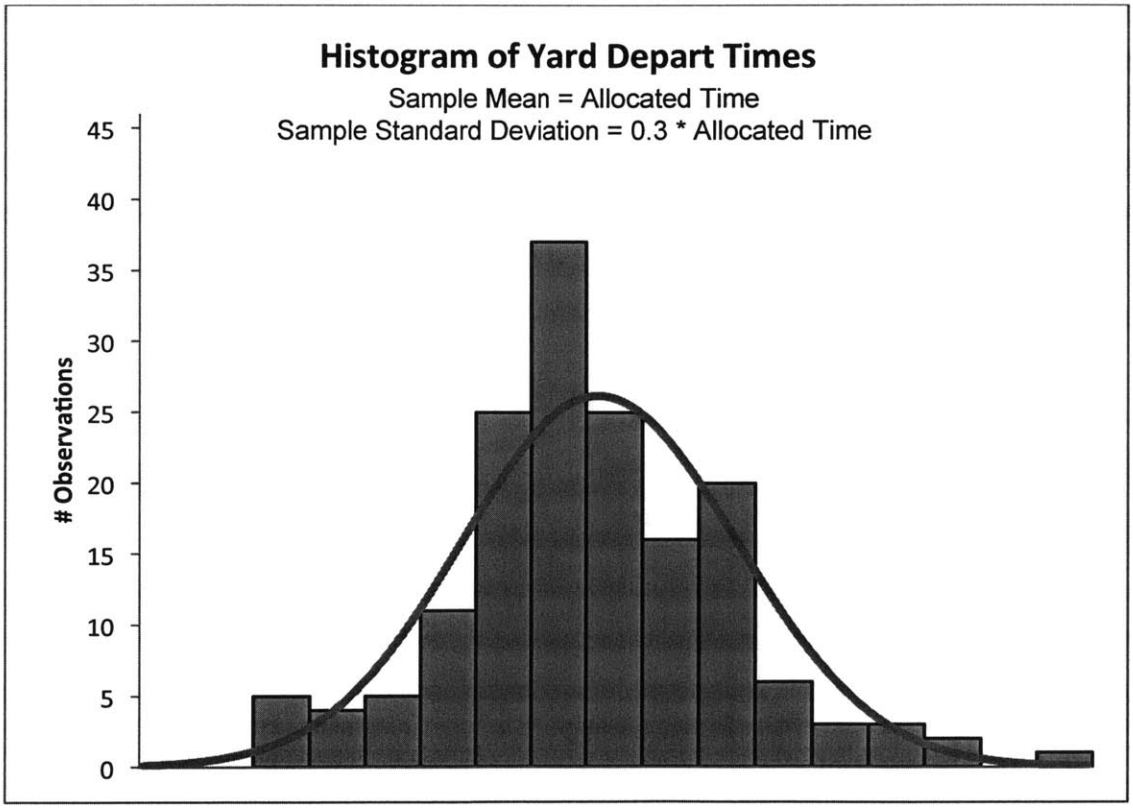
1) Grow ntrees to create your random forest based on the nodesize set by the user. If no value is specified, the default values for ntrees and nodesize are 100 and 10, respectively.
2) For a given vector X=x, to generate a prediction input x and compute the weights w_{Ti} for every observation in the tree and the corresponding w_i for all trees in the forest.
3) Compute the estimate of the distribution function for all y_i using the weights.

The model built using Amazon outbound linehaul data relies on this package to create predictions. Once the forest is built using the function quantregForest(), vectors X must be inputted into the forest to generate the predictions. The vectors inputted are at a linehaul load level. Chapter 4 will discuss the sources of data and the inputs into my model. Chapter 5 will describe stability testing and final results obtained.

3 Understanding Truck Departure Process from Amazon Facilities

Analogous to the gate departure process described by Skaltsas, the first step in outbound linehaul delivery is the truck departure process at the origin. Amazon allocates a fixed period of time from the scheduled transit time to ensure that a truck has a sufficient amount of time to exit the yard. It has been shown in time studies across 12 of Amazon's largest warehouses that while the allocation may seem appropriate, actual performance has a wide variation. On average a truck will take 99% of the allocated time to depart from the site, however the standard deviation is approximately 30% of the average. In response to these time studies, a team of yard personnel, analysts, managers, shipping dock associates and leads were assembled to perform a kaizen on the yard departure process in April 2013. Kaizen is the Japanese term for improvement. The motivation and goal for the kaizen was to develop an understanding of how the variation can be reduced through creating a standard work model that could be applied across the network.

Figure 3-1 Results from Time Study of Yard Depart Times



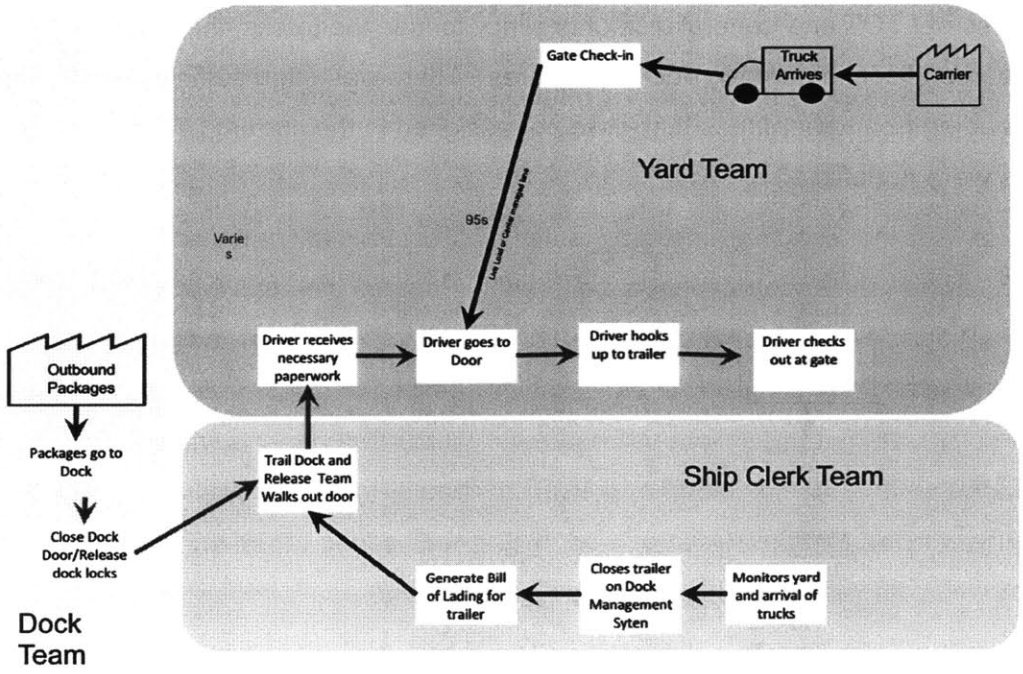
In addition to reducing the variation, there was also a need to understand how much time should be allocated to departing a truck. Thirty minutes seems more than a sufficient amount of time to

exit a truck. In fact, many Amazon Transportation managers believed that it was an excessive amount of time for the process. They believed inefficiencies in the management of the process drove the long right-sided tail of the performance distribution. Therefore, the goals of the kaizen were to reduce variation and to measure and improve cycle time of the departure process. The latter goal was directly related to the needs of modeling performance variation for my transit time forecasting model.

3.1 Overview of Truck Departure Process & the Last Truck Out Time

The truck departure process can be understood as two time periods, before and after last truck out time (LTOT). Prior to the LTOT, the distribution center is still loading the boxes onto the trailers and the tractor assigned to the load will arrive at the origin . After the LTOT has occurred, the truck is closed and initiation of the departure process occurs. The LTOT is somewhat of a misnomer because it is not actually the time in which a truck leaves the dock. It is a scheduled time when a batch of trucks begin to depart. It is also less of a dock metric and more so a warehouse deadline. Process managers are aware of which LTOT they are picking and packing for. While it does refer to when a load will depart from the facility, it has become synonymous with a batch of work. The process flow diagram below outlines the steps that occur in each of those periods.

Figure 3-2 Truck Departure Process at Amazon



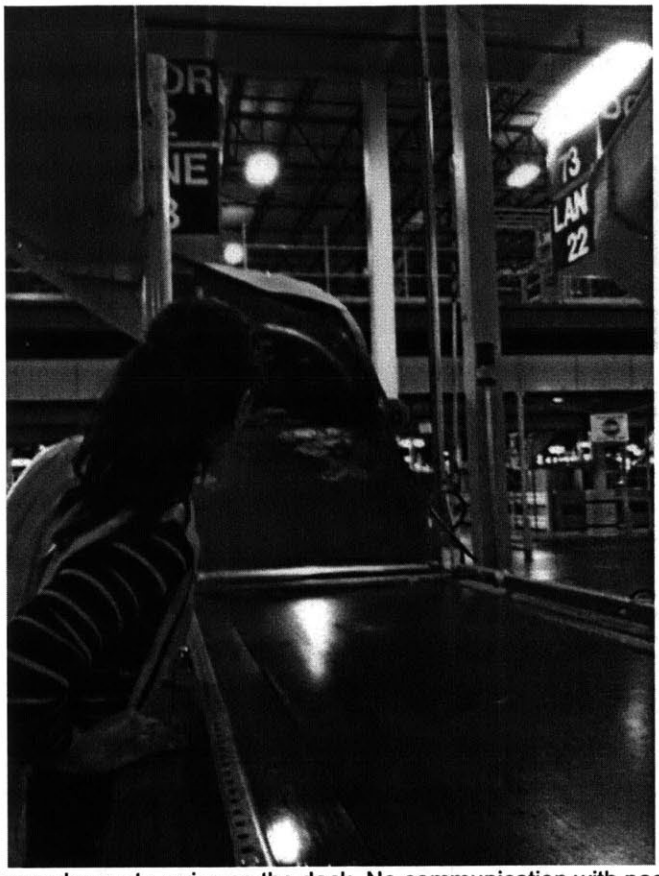
3.2 Kaizen Teams & Findings

A week was spent on-site at a distribution center (DC) to conduct the kaizen event. It was conducted at one of Amazon's largest DCs, both in physical footprint and capacity. The kaizen focused on studying the processes on the outbound dock and yard. The truck departure process can be grouped into three major teams: dock, ship clerk and yard. As a result, the structure of the kaizen was also broken into three major teams that examined the processes of each team. The following subsections will explore operational issues that can influence percent on time performance.

3.2.1 Dock Team

The Dock team's main responsibility is to ensure that all packages are loaded onto the truck. This is a metric that they are evaluated on. Because of this, there is little awareness on late truck departures. Since transit times are scheduled to start at LTOT, it is critical for on-time departures that the truck be closed as soon as possible. It was discovered that there was no sense of urgency around LTOT to ensure that the truck departs as soon as possible. There were also no clocks around the dock to alert the associates what time it was. It was observed that many associates simply waited around for packages to come down the chutes, as seen in Figure 3-3. The associates' main priority was to ensure that all packages assigned to the truck made it onto the truck, and they were less concerned over how long it took. Because of the lack of awareness of LTOT and general lack of urgency to load the trucks, there was opportunity for better resource allocation among dock associates. There was often sufficient staff to assist with package loading from other lanes that were not assigned to the current LTOT, but these associates were not utilized.

Figure 3-3 Associate at the Dock During LTOT



Associates waiting for packages to arrive on the dock. No communication with pack areas about LTOT

The dock team was most concerned about ensuring that the packages made it onto the truck within 30 minutes of LTOT. This cultural habit made it clear that while transit time is supposed to start at LTOT, the allocated time designed for the departure from the yard was being utilized to finish loading the truck. Associates understood that it does not take that long to drive a truck off Amazon's property and were indirectly trained to use this time period to scramble for last minute packages. The allocated time gave the associates a false sense of excess time, which can cause delays down the road.

Once the packages are fully loaded, the dock process lead will virtually close the truck in Amazon's systems. When there are a large number of trucks departing at once, this step tends to be batched once all physical processes have been completed. Therefore there is often a mismatch between virtual and physical data.

3.2.2 Ship Clerk Team

The ship clerk's main responsibility is to act as the link between the yard and the warehouse. They virtually depart the truck as soon as all packages are loaded and print necessary paperwork to hand off to the truck driver. Since they are the team responsible for communicating between warehouse and yard, the ship clerk must know what is occurring inside and outside the warehouse. In its current state, this cannot be done without physically being in each location.

Many of the inefficiencies that were noted during the kaizen were related to the lack of standard work and misalignment of performance metrics. Each ship clerks had their own process for departing a truck. The ship clerk's performance metric is to ensure that all loads are departed from Amazon within the allotted time from the LTOT. From a ship clerk's perspective the load departed on time if it left the yard within this time from LTOT. Implicit in this metric is the idea that the allotted time is required to depart each truck. Success is defined if a *set* of trucks depart on time, rather than each truck. The batching of truck departures leads to unnecessary queuing and inefficient processes.

Many ship clerks wait on all packages to arrive before initiating any of the steps to depart a truck. Because of this, a lot of wasted time is spent waiting for packages to be loaded. During the kaizen, it was noted that much of the work completed by the ship clerk could be done before LTOT occurred. These included printing the bill of lading for truck drivers and appropriately communicating with yard jockeys about which trailers are to be pulled next.

The longest task required of the ship clerk is delivering the bill of lading to the truck drivers in the yard. Since there is often only one ship clerk staffed per shift, they tend to wait until all of the loads have been completely loaded before walking outside to deliver the bill of lading. This was noted as an inefficiency for the process. This also prevents the yard team from initiating their processes. A take away from the kaizen was to improve the bill of lading delivery process by asking truck drivers to come into the facility or staff at least two ship clerks to expedite the delivery process.

As the intermediary between two teams, the ship clerk's efficiency is highly dependent on how much information it is shared by the dock and yard teams. Information is currently delivered

through handset radios. This method works well but is not effective in delivering messages to truck drivers who also need information. To improve communication, visual cues such as pagers to communicate with drivers, would also help ship clerks understand truck arrivals and departures while not physically in the yard.

3.2.3 Yard Team

The Yard Team analyzed the process from the handoff of the bill of lading (BOL) through the exit of the truck from the yard. The yard process involves three parties: the security guards, truck drivers and yard hostlers. The security guards are responsible for checking in the tractors for their pick up and exit. They also notify the ship clerk when a tractor has arrived for pick up. The truck drivers are responsible for dropping off an empty trailer, picking up the filled trailer and completing their pre-trip inspection before exiting the yard. The pre-trip inspection is a safety check required by law. The yard hostlers are responsible for replacing empty trailers at the dock that have just been cleared. They also pull trailers off docks in cases where the truck driver is late to arrival.

Two key findings were the ambiguity of yard ownership and difficulty communicating across teams during the process. This was highlighted by the disorganization and traffic in the yard. The truck driver would pull out of their trailer slots and move to a convenient location for them to complete the safety check. Due to the random location of trucks staged for pre-trips, this caused congestion and overall disorganized flow in the yard. This disorganization also poses a potential safety issue for the ship clerks traveling in the yard to deliver the BOLs. Simultaneously yard hostlers are also pulling trailers on and off of dock doors. They pull trailers based on the lights by the dock doors controlled by the dock team. There is no prioritization of any specific load. The lack of communication between the teams inside the facility and the yard hostler outside the facility result in a lot of inefficiencies in trailer movement.

Figure 3-3-4 Example of Yard Congestion

Yard congestion due to lack of standard work and communication to drivers

A fact that complicates matters for the yard processes is that the yard hostlers, security guards and truck drivers are all employed by third party contractors. Amazon has no direct influence over these individuals. Therefore, the yard departure process is essentially a third party managed process. One of the main recommendations of the kaizen was to assign yard ownership to the ship clerks. Without Amazon overseeing these teams, it would be difficult to implement process improvement initiatives. Since the yard process is the last step in exiting a trailer, it is important that Amazon is involved in ensuring timely departure.

3.3 Kaizen Results & Implications on Modeling Transit Time

The kaizen highlighted the need to estimate the time it takes for a specific load to exit the facility. I believed that it would considerably influence predictions of short haul transit time estimates. Short haul linehauls are often scheduled to less than one hour and because the prioritization of truckloads does not exist, any delays of exiting Amazon's premise would significantly impact the carrier's ability to arrive on time to destination. Based on my experience, I believed Amazon short hauls suffer from the same scheduling problem highlighted by Skaltsas for short haul flights. Namely, a large buffer is built in to transit time due to performance variability. Additionally, I also expect that at high service levels, gate departure performance would play a significant role in reducing overall transit time.

The operational challenge that most affected reconciling virtual timestamps and physical process was the closing of trucks on the docks. The ship clerks' metric of success is based on their ability to close the loads in the dock management system within the allotted truck departure time of the scheduled pull time. The behavior that resulted from this metric was that the scheduled departure time not being the actual time when a truck departed from the dock but the beginning of the last rush to load the truck.

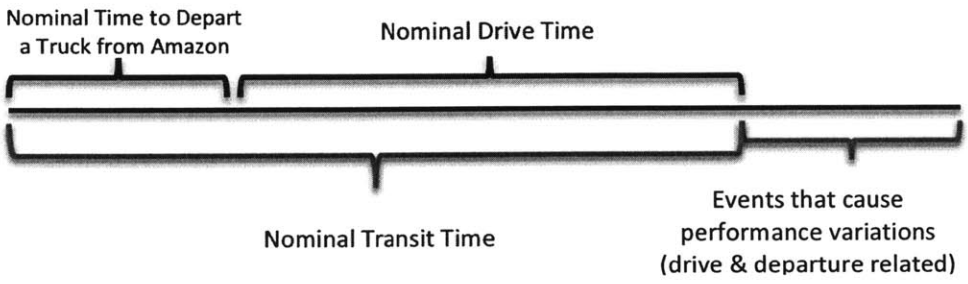
One methodology to estimate how long a truck takes to depart from the yard is to measure the time it spends in the yard. This can be done by using a yard management system (YMS) that utilizes RFID tags. The check in and check out process requires each trailer to be tagged on all incoming trailers and removed from all outgoing trailers. The security guards are responsible for ensuring this process is properly executed. While there has been some shrinkage of RFID tags associated with failures to remove the tags from trailers, this process is believed to be well managed. Amazon has implemented a YMS in its largest distribution centers,. The usage in the largest distribution centers was part of a proof of concept pilot before committing to its wide spread implementation.

The kaizen also offered insight into some of the data discrepancies that I noticed between the three data sources for the yard departure process. The yard management system (YMS), dock management system (DMS) and the electronic data interchange (EDI) data captured events that were in theory simultaneous, however there were often large discrepancies between the timestamps. This is due to the fact that none of these three systems are linked and do not have virtual processes that depend on each other. Chapter 4 will discuss how these discrepancies were reconciled and why EDI data was eventually chosen to be incorporated into the model as the primary source for modeling the gate departure process.

4 Developing An Analytical Understanding of Transit Time

Building on the discussion provided in Chapter 1 and Chapter 2, transit time can be considered as the sum of three components: nominal time to depart a truck from Amazon's premises, nominal drive time and events that cause performance variations.

Figure 4-1 Composition of Transit Time



Developing a new model of transit time began with understanding the current Amazon calculation process. It is also important to understand existing performance under this scheduling technique. This chapter begins with a description of both of these topics. It is followed by a brief discussion of current performance of Amazon's linehauls, segmented by length of haul. The subsequent sections will examine the data sources used to estimate each of the components. The first two components of transit time can be considered intrinsic characteristics, which to some extent, can be predicted and controlled by managing performance by either Amazon or the carrier. Predicting extrinsic characteristics that can cause performance variation proves to be a bit more challenging, since the likelihood of these events may or may not be available prior to scheduling. These types of events include but are not limited to accidents, weather related delays, and Department of Transportation inspections. To narrow the scope of these types of events, an initial list of variables was developed with the help of Transportation carrier managers and the outbound transportation team. The list was pared down to a list of items that could be reasonably measured and then finally to a shorter list of variables that significantly contributed to prediction accuracy.

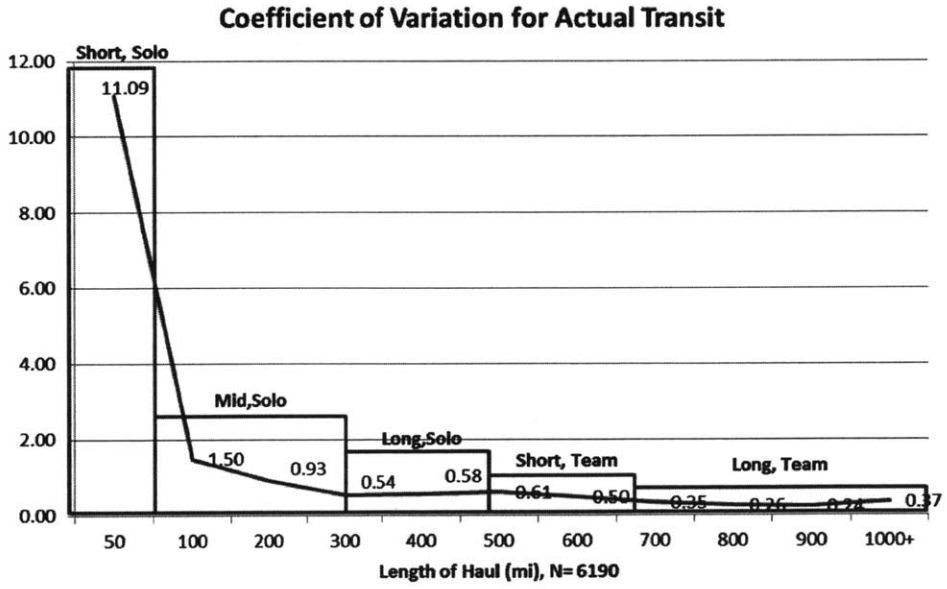
4.1 Current Methodology: Scheduling Transit Time With Respect To Distance

The current methodology of calculating transit time is based on the distance between origin and destination. A common categorization within the trucking industry of linehauls is subdividing into

solo and team drives[2]. Team drives are often assumed to include less breaks because each driver will alternate driving with less stationary breaks. Therefore it is common practice to pad transit times for solo drives slightly less than transit times for team drive. This additional padding is done not only to anticipate delays that could occur during transit. Padding is also done to sufficiently schedule for any breaks that the truck driver may need to take, ie: refueling, lunch, mandatory driving breaks etc.

In order to understand the current performance of Amazon's outbound linehaul network under this methodology, the distribution of linehaul transit time was studied. Figure 4.2 plots the coefficient of variation for North America's outbound linehaul, ie: standard deviation divided by the mean actual transit time, for every group of 100 miles through 1000 miles linehauls. Lanes that exceed 1000 miles were grouped together. From Figure 4.2, it was clear that groups of linehauls shared similar characteristics between actual transit time and the standard deviation of the sample size. These groups were named short, solo; mid, solo; long, solo; short, team; long, team. The groups are overlaid on the graph below.

Figure 4-2 CV of Actual Transit (Feb 18 - March 19, 2013)



There is a larger amount of variability with respect to the mean for lanes fewer than 50 miles. This finding is intuitive since the coefficient of variation is a ratio of the standard deviation of

[2] http://DC.bigrigdriving.com/2010/trucking-industry-debates/is-team-driving-the-future-in-trucking

transit times divide by their average. For any type of delay, the absolute time delay will have a greater effect on short, solo lanes as a percentage of their scheduled transit time. However, this finding implies that if Amazon is to ensure a high percentage of on time delivery for short haul lanes, scheduled transit will be significantly higher than the lane's demonstrated average. This conclusion is also consistent with Skaltsas' finding that the percentage of buffer decreases exponentially with distance (Skaltsas 120).

4.2 Estimating Components of Transit Time

There were several data sources that were available for transit time analysis. The following sections examine the reliability of using specific data sources as a part of the final model.

4.2.1 Validating Nominal Drive Time Estimates Using Transportation Software

Building on the existing practices, the estimation used for nominal transit time was derived from transportation software. As part of due diligence to ensure that these estimates were reasonable, the average speed was calculated by dividing the drive times provided by the software by the distance between the origin and destination. Below are the tabulated results:

Table 1 Lane Types by Average Speed

Lane Type	# of Lanes	*Average Speed (Miles per Hour)
Short, Solo	84	33.30
Mid, Solo	255	50.27
Long, Solo	345	53.19
Short, Team	434	55.08
Long, Team	513	56.34

*Note: Average Speed = Drive Time from Transportation Software/Length of Haul

In general, individual estimates from Transportation Software are reasonable and well under state specified speed limit for 53-foot trailers. A full list of speed limits by state provided by TruckerCountry.com can be found in Appendix B. In the most extreme case trucks are allowed to drive 80 mph and in the most conservative scenario, 55 mph. It should also be noted that as the distance increases, the average speed also increases. This is because as the route gets longer, there is more highway driving, which allows for higher average speeds. Amazon Transportation managers assumed that trucks can drive roughly 45 mph on average. Since

these values represent nominal drive time, defined as driving under optimal conditions, the estimates are reasonably aggressive and therefore are valid to use in the model.

4.2.2 Estimating Time to Depart a Truck From Amazon Facilities

During the kaizen described in Chapter 3, I was able to see how virtual processes tied or did not tie to their physical process. This helped provide insight into which data sources were good approximation of the physical process, as mentioned in section 3.3.

The incongruous communication between the teams operating the dock and yard resulted in a wide distribution of demonstrated time to depart a truck. Because of the wide distribution of yard departure times seen in the time study, the current standard allocated time to depart a truck used by Amazon was not sufficient to use in my model. Through this experience, it became clear that an estimate of yard departure time by load would be needed to estimate transit time.

The wide range of performance is attributed to a series of both controllable and uncontrollable factors that govern the truck departure process at distribution centers. Among the uncontrollable factors are the physical layout of yards and the number of gates available for trucks to enter and depart from. Among the controllable factors are the communication issues, lack of ownership over gate departure process and misalignment of productivity metrics. These operational issues can cause behaviors we observed such as virtually batching physical processes, which inaccurately represent the duration of each step in the process.

The extent an organization can efficiently schedule transit times may be directly correlated to its ability to control and predict the required times of all origin facility controllable processes. Any excess time required because of lack of process standardization introduces additional factors that unnecessarily complicate forecasting transit time. It also effectively requires additional buffer time in order to meet on-time delivery requirements. As Skaltsas notes, "a very important issue is the extent to which this uncertainty is caused by the [airline's] schedule, operational weaknesses, and poor overall performance, rather by external factors that the airline cannot forecast and handle effectively" (Skaltsas 51).

Dock Management Systems & Yard Management Systems

Amazon has two systems that are internally managed that would be the primary sources for understanding the truck departure process and any delays associated with it. The truck departure process can be broken down into two sequential processes: dock departure and yard departure.

To examine dock departure times, the scheduled pull time of each load was compared to the timestamp associated with when the load was virtually closed in the dock management system. Unfortunately, since the physical process of releasing a truck and the virtual process of releasing a truck are not linked, the data was not helpful in analyzing performance. It was also observed that ship clerks would batch these tasks when they were understaffed. Approximately 25% of the data points examined needed to be discarded, which made the data from the dock management system unhelpful for the purpose of estimating transit time. These data points were discarded due to multiple reasons such as departures one hour before scheduled pull time, yard departure events occurring before dock departure events, departures over ten hours past scheduled pull times etc.

Amazon's yard management system is used by the distribution centers to track the location of a trailer in the yard and also the time associated with its exit from the facility. Compliance of the YMS was above 90%. However trailers could often sit in the yard for hours or days before being utilized, so in order to estimate gate depart time, the DMS timestamp would need to be used. Additionally, this system has not been implemented across the entire Amazon network and therefore could only be used for a subset of distribution centers that had the data available. For that reason, Amazon's YMS data was not used since it would have limited the applicability of the model.

Unfortunately, due to data issues, the Amazon controlled sources for tracking the truck departure process could not be used in the model. A secondary source was used to better estimate the time associated with the truck departure process.

Electronic Data Interchange (EDI) Data: Carrier Reported

Amazon carriers send electronic messages to Amazon about assigned linehauls through an electronic data interchange (EDI). Amazon has worked with their carriers to increase the compliance of EDI messaging. They require the carriers to send origin arrival, origin departure

and destination arrival messages. Compliance metrics as of April 2013 are listed below in Table 2. Their compliance is measured on a weekly score card that is reviewed by the carriers and Amazon carrier managers. Because of this, EDI data, while carrier reported, was the most reliable and network wide reported source for estimating arrival and departure times from Amazon facility, as well as time to destination. Unfortunately, the time to depart a trailer from the dock is not captured in the EDI messaging and could not be estimated from this data source. For this reason, the time to depart a truck was estimated using the difference between fifteen minutes from scheduled pull time and the depart from origin notification from the EDI. Fifteen minutes was chosen because the demonstrated average departure time was twenty-eight minutes with a standard deviation of ten minutes. Therefore fifteen minutes from scheduled pull time was a reasonable estimate to use in the model since actual timestamps were unavailable.

Table 2 EDI Compliance Rates

EDI Message Type	Q1 2013 Compliance Percentage
Arrival at Origin	99%
Departure from Origin	97%
Arrival at Destination	94%

4.2.3 Understanding Performance Variations

Unlike nominal drive time and the truck departure process, performance variations can only be estimated once a load has experienced the delay. Amazon tracks these types of performance issues in an internal report that is reviewed weekly with carriers. The report is compiled based on EDI messaging, distribution center reporting and carrier reported issues. Late is defined as one of three possible events: late to pick up at origin, late to departure from origin, late to delivery at destination. The list was created with the intention of capturing all types of performance aberrations that would help manage carriers. As a result, the report contains reason codes that may be attributed to loads that are late to arrival and late departure from origin. A new metric was developed to track late to destination loads, however was not tied into the report. This is the only source for tracking reasons associated with late to destination. My model ignored any late incidents that flagged late arrival to origins because during the initial development of the OLS model, it was determined to be an insignificant variable.

4.3 Data Cleaning

Despite high compliance of EDI message, there are still errors associated with the messaging because it is a manual process. Data cleaning rules were created, with input from Transportation managers, to ensure that valid load information was being incorporated into the model. The following rules were implemented in order to ensure integrity of the data:

i. Remove anything with depart timestamp earlier than 1 hour from scheduled pull time
ii. Remove anything with depart timestamp later than 10 hours from scheduled pull time
iii. Remove anything with actual transit time less than 0.5 hour
iv. Remove anything more than 2 standard deviations from the mean variation per lane. Variation is defined as the time difference between nominal drive time and actual transit time

Variation was defined so that consistent variation in performance was taken into account when selecting data. There is a cultural belief that poor performing routes, ie: low percentage on time based on existing scheduled transit time, should be discarded so that they would not "negatively influence" the predicted transit time, ie: extend transit. However, this mentality could potentially lead to selecting the data due to bias, rather than absolute rules. It also ignored performance at the extremes, a practice that may misrepresent the required transit time at high service levels. Assuming that each linehaul had a sufficient number of historical data points, the distribution of the variation should be normal and taking two standard deviation greater than the mean was an acceptable boundary that was defined conjointly by myself and Amazon Transportation managers. By defining the variation variable, it ensures that a large amount of data is not discarded based on an intuitive understanding of how long transit "should" take.

4.4 Variables in Model

The variables that were incorporated into the model are a mixture of known factors prior to the load being executed and factors that are reported afterwards. Chapter 6.3 discusses the implementation of the model and the process of how the latter types of variables are dealt with for forecasting purposes. Based on the ability to quantify and reliably measure each of the components of linehaul, the following variables were selected and defined to be in the final quantile random forest model:

Dependent Variables

1. **Actual Transit Time**- Time of arrival defined by EDI destination arrival message less LTOT time from origin. Actual transit time is composed of time to exit the yard plus drive time.

Continuous Independent Variables

1. **Yard/Gate Departure Duration** – Duration from 15 minutes after LTOT to when the truck exits the gate of the DC[3]. Truck exit time is given by EDI origin departure message.
2. **Nominal Transit Time**– Drive time given by Transportation Software plus required 15 minute breaks for truck drivers for every 4 hours of transit and requisite lunch hour for routes with over 8 hours of driving.

Binary Independent Variables

1. **Rush Hour** –1 when the depart hour is between 6 and 9 or between 15 and 18, 0 otherwise
2. **Weekend** –1 for Saturday or Sunday, 0 otherwise
3. **LTOT Truck** –1 if the load's scheduled pull time is equal to the Last Truck Out Time, 0 otherwise
4. **Performance factors** – codes used to explain an instance of lateness. Each historical load has four performance factors variables associated with it, 1 if true, 0 if false
    a. Origin Facility – codes associated with Origin Facility controllable delays
    b. Carrier Controllable – codes associated with carrier controllable delays
    c. Seasonal – traffic and weather
    d. Other – random, difficult to predict events (accidents, DOT inspection etc)
5. **Destination Types** – 1 if destination is Transporter A, Transporter B, Transporter C, Transporter D, Amazon cross dock, or distribution center, 0 otherwise. These destination types were grouped together to limit the total number of variables needed and because of limited operational hours of certain transporters, the destination type was considered significant. There also is a general sense of priority for loads that are customer deliveries versus inventory transfers between distribution centers.

[3] 15 minutes is used as a benchmark because the dock departure timestamp caused 25% of these values to be negative, which implies that the accuracy of the dock departure timestamp is questionable. A static value, 15 minutes after LTOT, was opted to be used for data stability. Theoretically this represents the time it takes after dock door closes to when truck is pulled off the dock.

6. **Origin Distribution Center** – a variable was created for each DC to indicate origin of lane
7. **Adhoc** – 1 if truck is a non-scheduled truck, , 0 otherwise. An Adhoc truck can occur when additional truck capacity is required due to excessive volume of packages during production. This can be caused by poor loading leading to low utilization of trucks, errors in package forecasting, or other operational issues that may lead to another truck being needed.

The next chapter will discuss the development of the quantile random forests, forecast errors and the factors required to provide stable predictions.

5 Quantile Regression Forests

5.1 Selecting Lanes for the Model

The initial data set for modeling transit time consisted of historical loads from ten origins throughout the North American outbound linehaul network. These ten sites were chosen because they were a representative sample of the entire network. These routes consisted of approximately 30% of all North American package volume. Table 3 summarizes the number of loads and number of routes that were included in the initial training data set. There were a total of 117 linehauls in the dataset.

Table 3 Distribution of Load Types Within Test & Training Datasets

Type of Lane	# of Routes	# of Data Points	% of Total Dataset
Short, Solo (< 50 miles)	12	535	6%
Mid, Solo (< 325 miles)	28	3,276	36%
Long, Solo (< 500 miles)	28	1,741	19%
Short, Team (< 725 miles)	24	1,838	20%
Long, Team (> 725 miles)	25	1,665	18%

The time period that was sampled was from January 2013 through April 2013. After the data was cleaned, roughly 45% of the total loads were discarded based on data cleaning rules described in Section 4.3.

5.2 Training & Test Data Set

A training and test data set were created to create a control and test group to develop the model. The training data set was randomly split into two sets. Two-thirds of the data was used to build the initial quantile regression forest. This dataset was known as the training set. The remaining third was then used to test the accuracy of the model generated. This dataset was known as the validation set.

Stability of the model will be discussed later in this chapter to ensure that no specific partitioning of the training and testing data set influenced the random forest's prediction. The accuracy of the model was evaluated by measuring the weighted mean absolute percent error (WMAPE)

defined in Chapter 2. Since the two data sets were from the same time period, the WMAPE values should be relatively close, if not identical, if the predictive model can be considered a reliable tool.

To further validate the predictive ability of the quantile regression forest, a new dataset was created using the historical loads from May 2013 – June 2013 for the same routes. This data set was known as the untouched set. Based on the forest generated, if the predictive model was a reliable method of generating predictions, the weighted mean absolute percent error for the untouched set would be reasonably close to the weighted mean absolute percent error values generated from the two previous data sets.

5.3 Initial Model

The initial model employed a 100 tree forest, an mtry value equal to 8, which was one-third the number of variables and a nodesize of 10. This generated an average WMAPE of 4.8% for the training data set, 4.7% for the test data set and 5.1% for the untouched data set. These are relatively small error values, which indicated the modeling methodology was appropriate for purposes of predicting transit time across the network. It also indicates that the conditions that cause variation in transit time from the untouched data set had changed very little from the time period where the training and validation data set was used. The range for the WMAPE between the validation data set and the untouched data set indicated that the model attributes needed fine-tuning, roughly a 13% percent difference from the mean values. In the next section the process of finding the settings for optimal model stability is presented.

Table 4 Values for the Initial Model Developed
Forest Settings: ntree = 100, nodesize = 10, mtry = 1/3 total number of variables

	Train		Validation		Untouched	
	WMAPE	RMSE	WMAPE	RMSE	WMAPE	RMSE
Mean	4.80	2.43	4.70	6.14	5.05	6.36
Standard Deviation	0.06	0.07	0.54	0.46	0.56	0.35
Min	4.60	2.20	3.69	5.22	3.97	5.68
25%-tile	4.76	2.39	4.34	5.86	4.73	6.10
50%-tile	4.80	2.44	4.62	6.08	4.95	6.33
75%-tile	4.85	2.48	4.90	6.40	5.24	6.54
Max	4.98	2.55	7.39	7.33	8.14	7.80

Range	0.39	0.35	3.70	2.12	4.17	2.12

5.4 Stability of the Model

Since the random forest algorithm relies on a weighted average of individual trees created by random sampling, every forest may produce different predictions. It is necessary to validate if the prediction generated by the model is dependent on the construction of a specific forest. In order to validate that the predictions are independent on the random forest constructed, 100 forests were built to generate predictions, then the WMAPE of each iteration was calculated. From these iterations, the stability of the configuration was determined based on how much the WMAPE fluctuated from iteration to iteration. Since Amazon wanted the model to be able to predict transit times for high service levels, the quantile chosen for the computation of WMAPE was p = 0.95.

In addition to the WMAPE, which measures the overall ability of the model to produce predictions at specified service levels, the root mean square error (RMSE) is also measured. Since the WMAPE may remain relatively unchanged due to the weighting at high values of p, it is important to measure the model's overall deviation from the actual transit times. The root mean square error is defined as:

$$RMSE = \frac{1}{n}\sum \sqrt{(A_i - y_i)^2}$$

where A_i = actual transit time for historical load i,

y_i = predicted transit time for historical load i

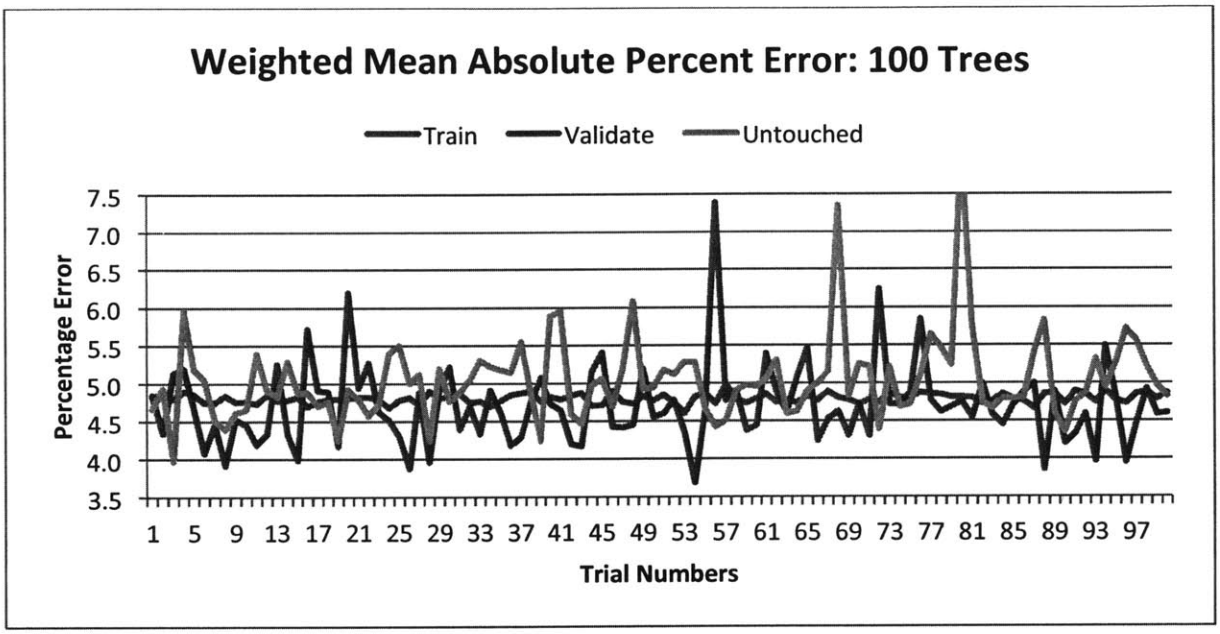
n= total number of loads

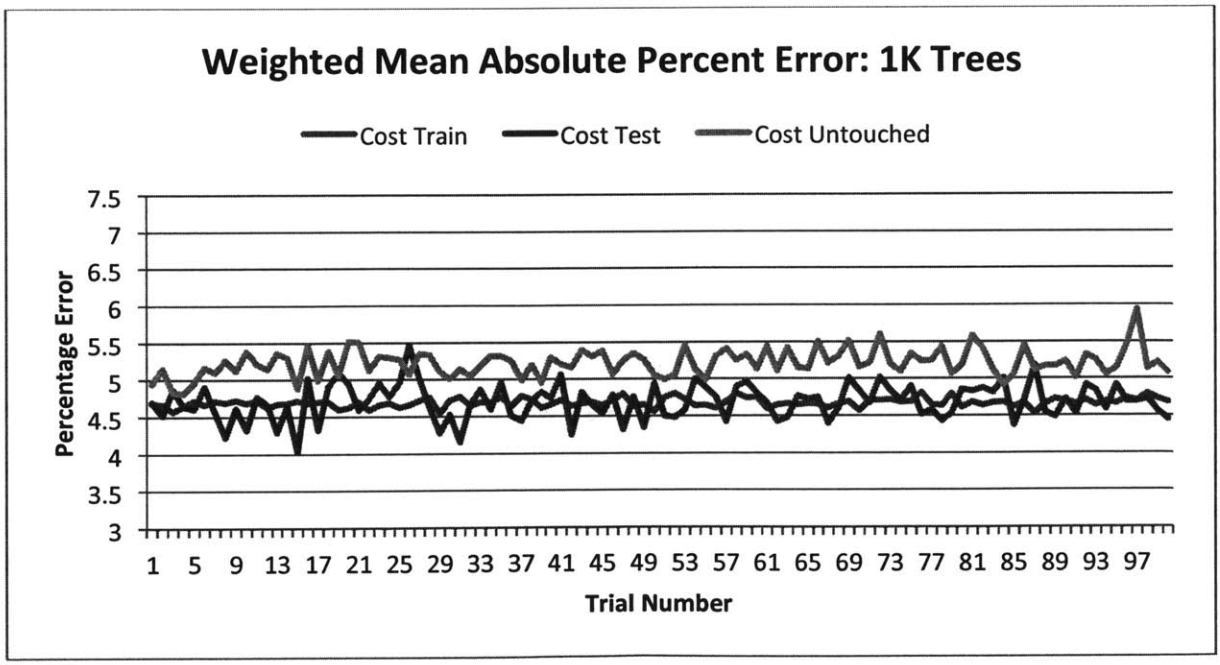
The stability of the forest was tuned using two attributes: the number of trees in the forest and the nodesize of each tree in the forest. The value assigned to mtry should not be used to control the stability of the model since high mtry values may cause a compounding effect where certain variables have an overly emphasizing importance in generating a prediction (Robin Guneur 4). The stability tests were conducted in two phases, first by determining the optimal number of trees in the forest and then using this ntree value and varied the nodesize of the forest to determine optimal values.

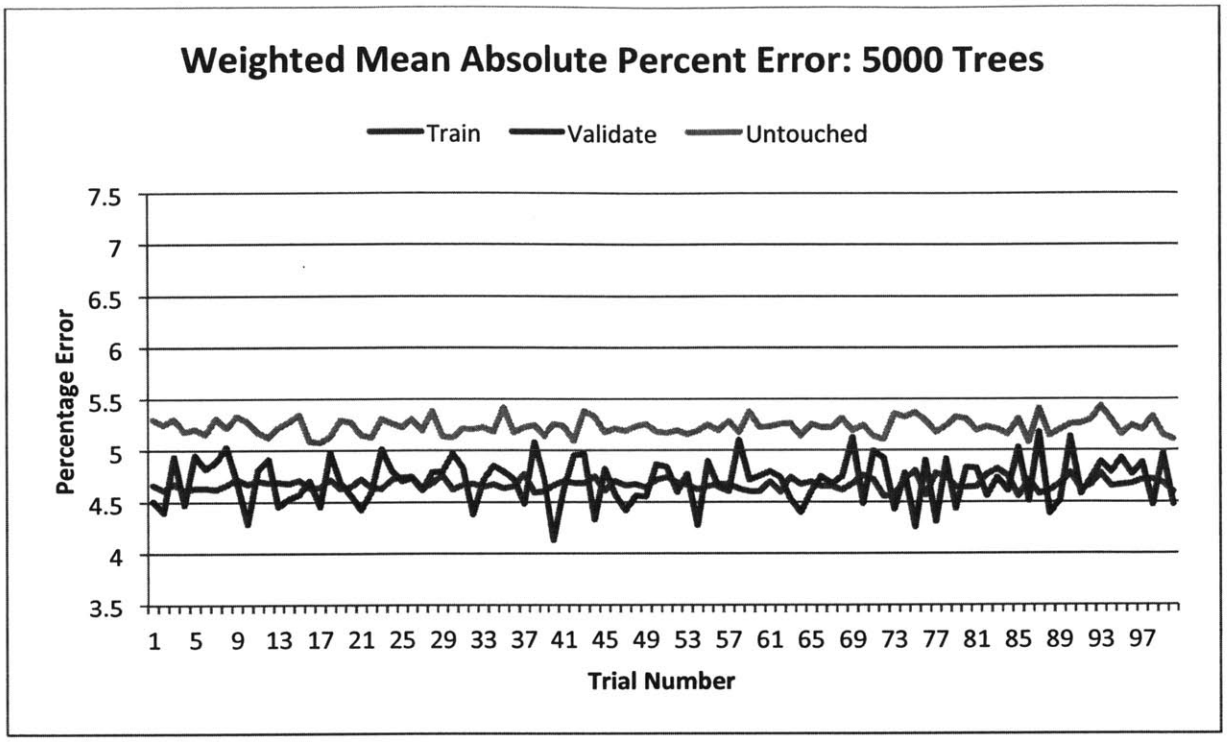
5.4.1 Number of Trees

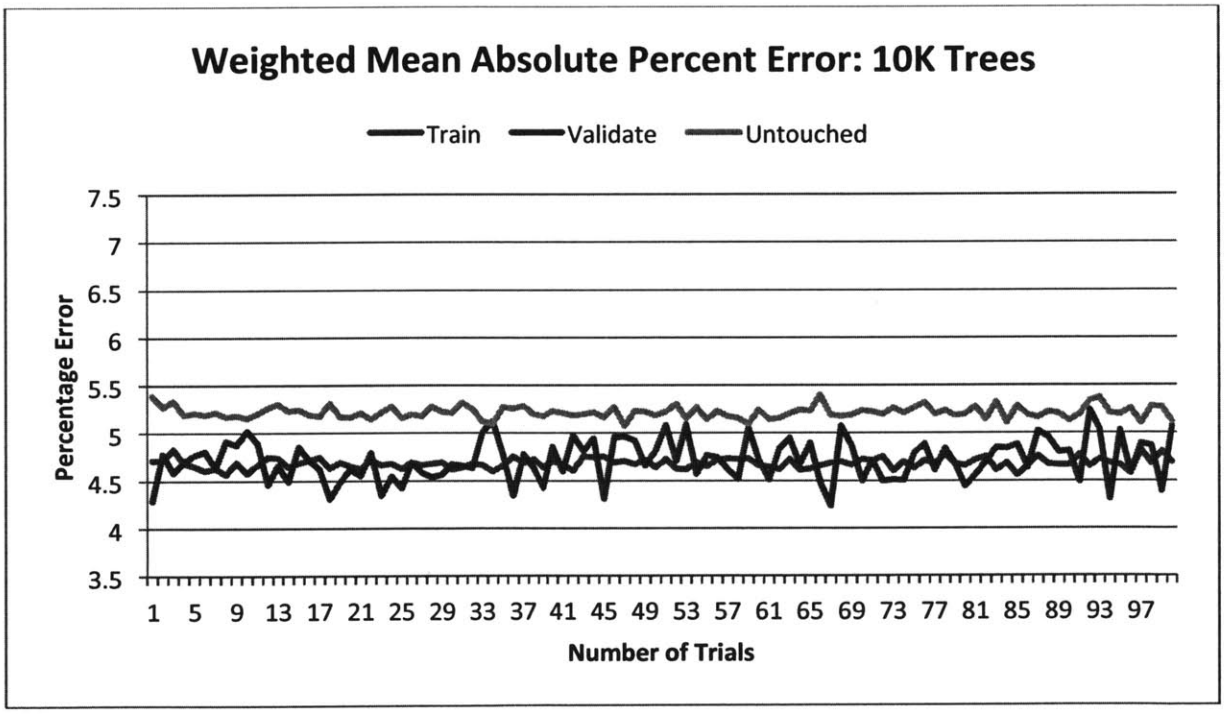
The number of trees in the forest affects the speed in which the model generates predictions. It was noted that as the number of trees increased, the time to compute the predictions increased noticeably. Therefore, it is important to create a model that both minimizes the number of trees and maintains the stability of the forest. The following graphs were obtained for random forests using 100, 1,000, 5,000 and 10,0000 trees:

Figure 5-1 Forest Size Stability Test: Weight Mean Absolute Percent Error from






Weighted Mean Absolute Percent Error: 5000 Trees
Train
Validate
Untouched
Percentage Error
7.5
7
6.5
6
5.5
5
4.5
4
3.5
1 5 9 13 17 21 25 29 33 37 41 45 49 53 57 61 65 69 73 77 81 85 89 93 97
Trial Number


Weighted Mean Absolute Percent Error: 10K Trees
Train
Validate
Untouched
Percentage Error
7.5
7
6.5
6
5.5
5
4.5
4
3.5
1 5 9 13 17 21 25 29 33 37 41 45 49 53 57 61 65 69 73 77 81 85 89 93 97
Number of Trials

The following is a table that summarizes the range of WMAPE for the untouched data set:

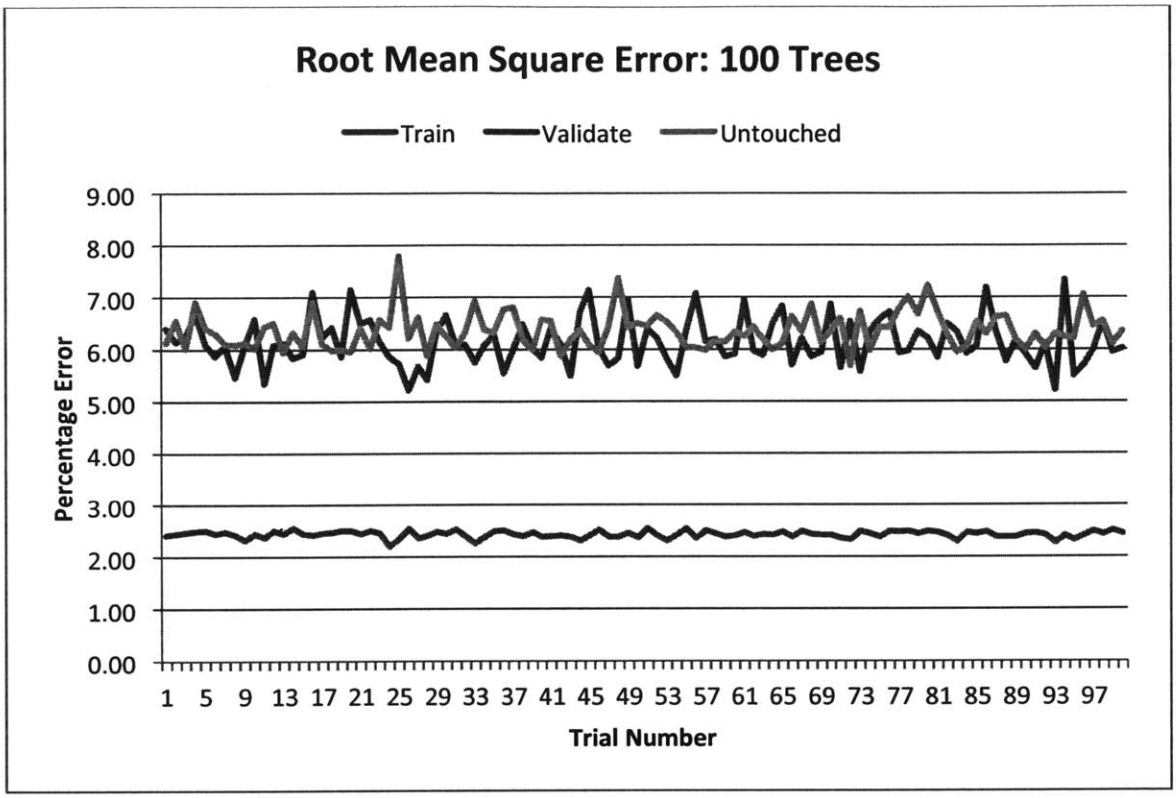
Table 5 Forest Size Test: Weighted Mean Absolute Percent Error of Untouched Dataset

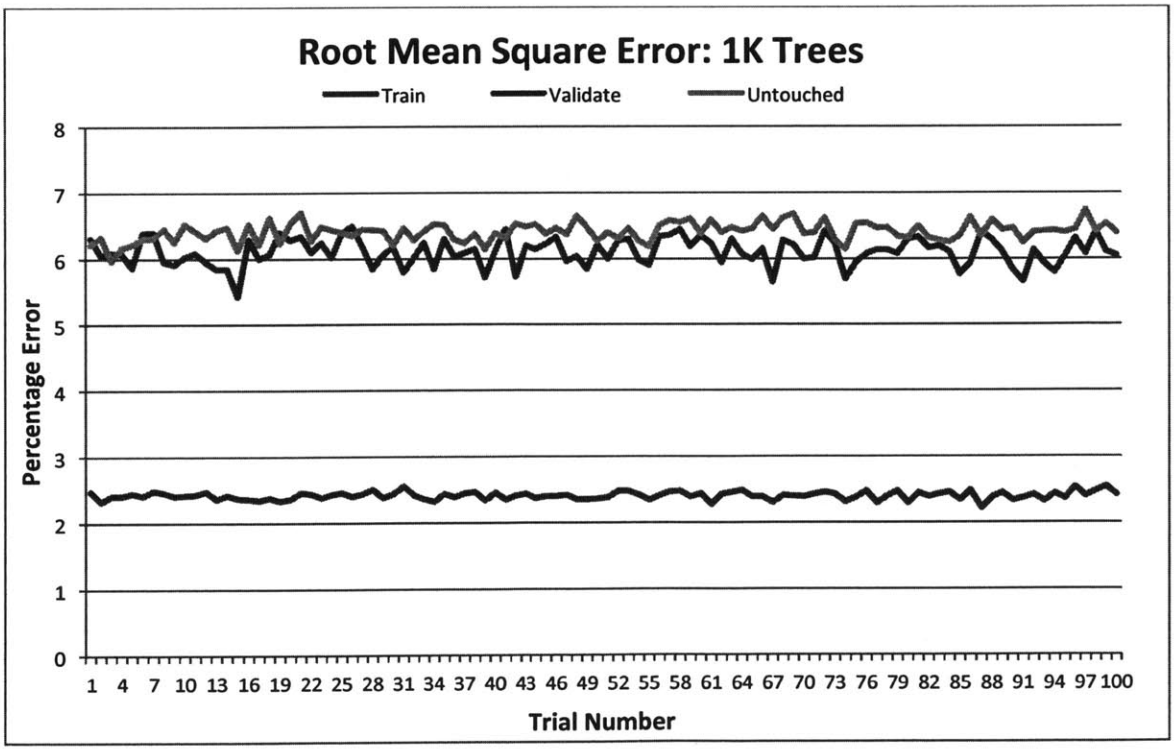
Number of Trees	100	1,000	5,000	10,000
0% (Minimum)	3.97	4.81	5.07	5.08
25%	4.73	5.10	5.18	5.18
50%	4.95	5.21	5.23	5.21
75%	5.25	5.33	5.29	5.26
100% (Maximum)	8.14	5.93	5.43	5.40

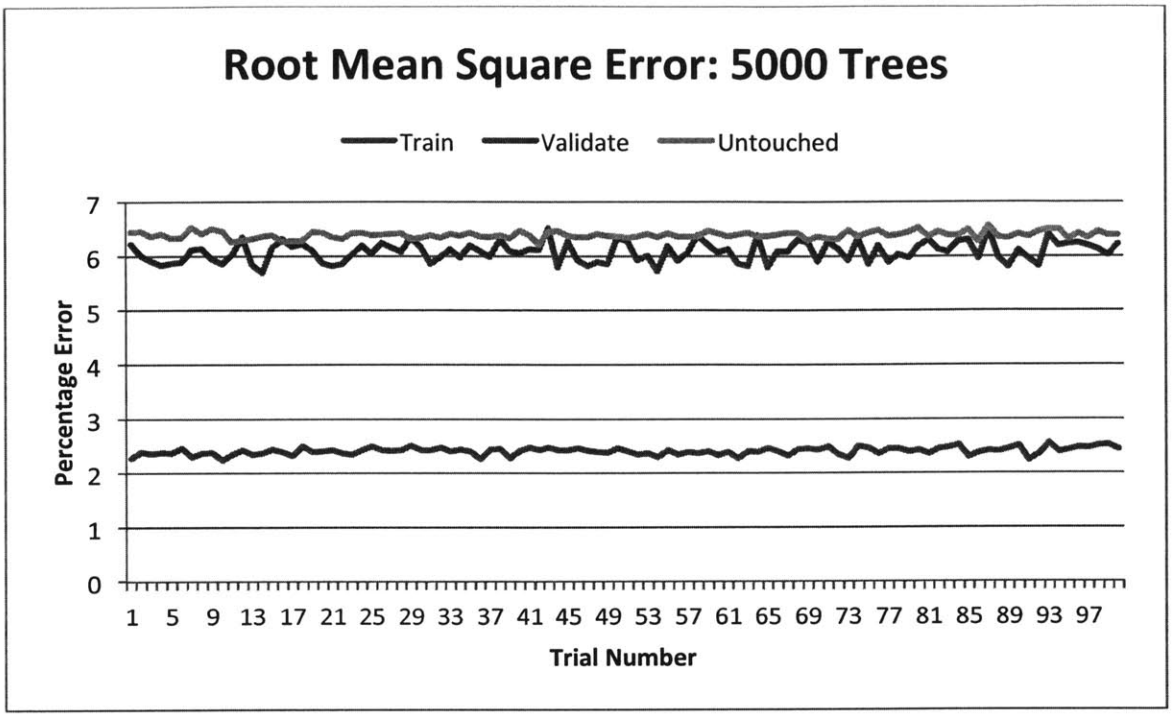
As the number of the trees in the forest increase, there is an increase in the stability of the WMAPE across each iteration of random forest. This is clear from the reduction in the range of values from the test to train to untouched data set, as well as the stabilization of each line as the number of trees increase. This is consistent with expectations since the quantile random forest uses a weighted average of each tree's prediction to generate the final prediction (Meinshausen 5). The model is very unstable when there are only 100 trees. Therefore, while one construction of the forest may produce accurate predictions for transit time, a different construction may produce very different values. Some improvement is seen by increasing the number of trees to 1000, however the range of WMAPE values is still very wide and may produce very different results based on forest construction. At 5,000 trees, the WMAPE range is reduced to 5.07% to 5.43%, which is a great improvement over a 100 tree forest with 3.97% to 8.14%. There is little improvement in WMAPE with 10,000 trees, so 5,000 trees will be the value used in the final model to minimize computational time and resources used.

There is a similar change but less significant change in RMSE over the different numbers of trees. The random subset and sampling does not affect the RMSE nearly as much as it did the WMAPE. Again, as the number of trees increase, there is a reduction in the range of RMSE values, which indicates an increase in stability. Additionally, there seems to be little improvement from 5,000 trees to 10,000 trees that would warrant building a forest bigger than 5,000 trees due to the trade off in computational time.

Figure 5-2 Forest Size Stability Test: Root Mean Square Error






Root Mean Square Error: 5000 Trees
Train
Validate
Untouched
Percentage Error
0
1
2
3
4
5
6
7
1 5 9 13 17 21 25 29 33 37 41 45 49 53 57 61 65 69 73 77 81 85 89 93 97
Trial Number

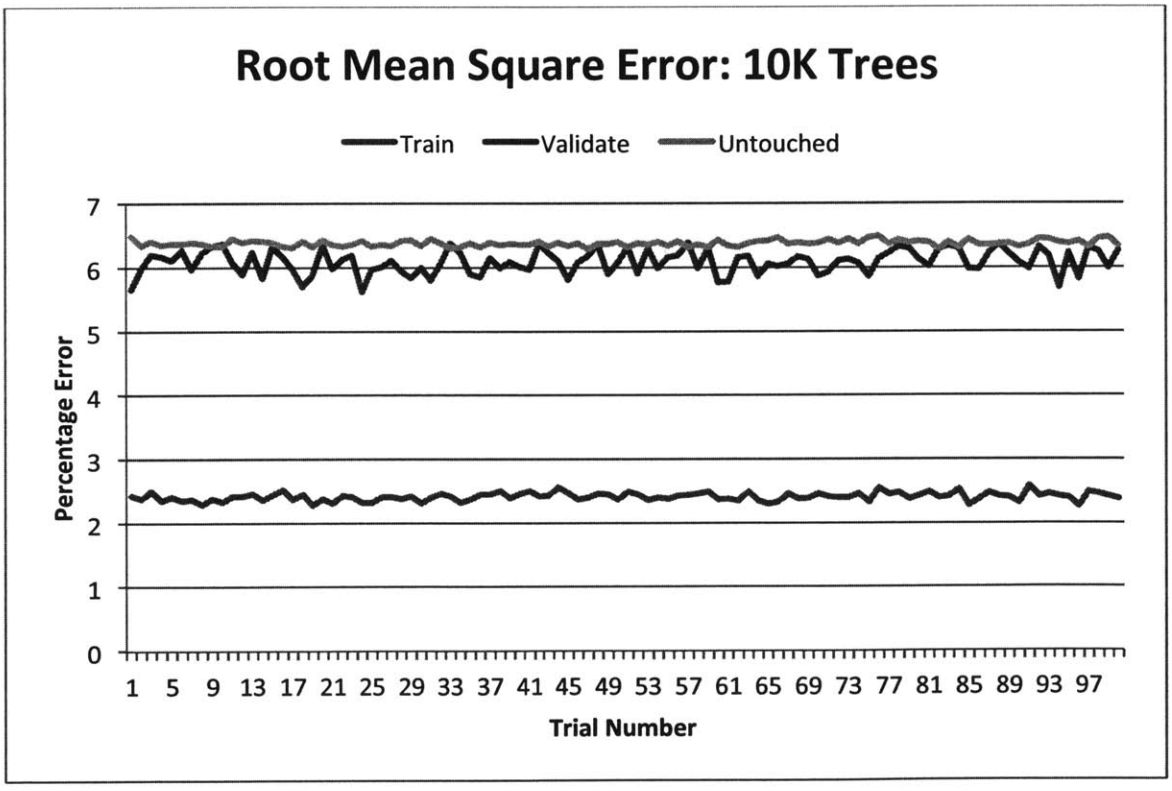

Root Mean Square Error: 10K Trees
Train
Validate
Untouched
Percentage Error
0
1
2
3
4
5
6
7
1 5 9 13 17 21 25 29 33 37 41 45 49 53 57 61 65 69 73 77 81 85 89 93 97
Trial Number

Table 6 Forest Size Stability Test: Root Mean Square Error for Untouched Data Set

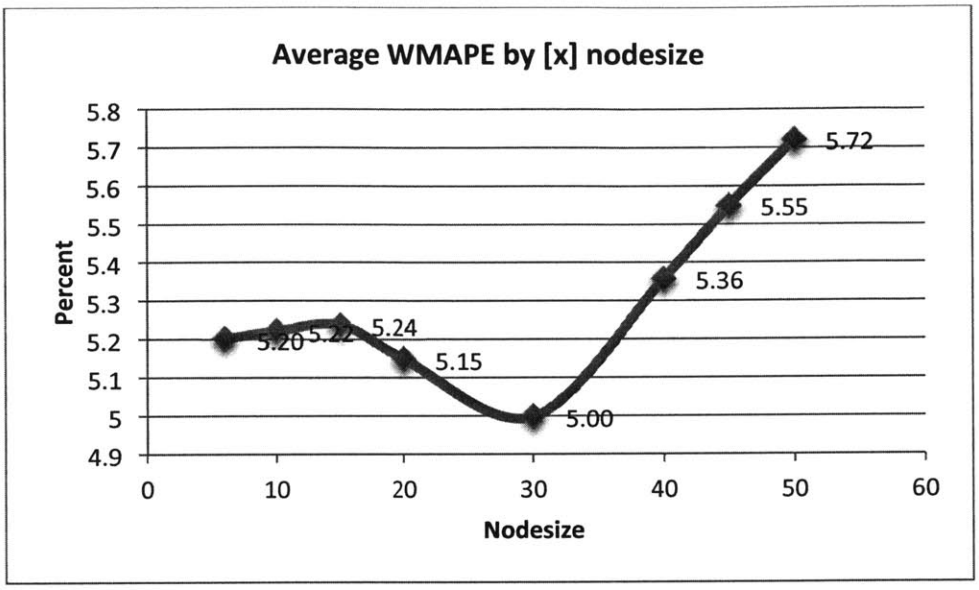
Number of Trees	100	1,000	5,000	10,000
0% (Minimum)	5.68	5.96	6.21	6.29
25%	6.10	6.31	6.35	6.34
50%	6.33	6.41	6.38	6.37
75%	6.54	6.50	6.43	6.41
100% (Maximum)	7.80	6.73	6.56	6.49

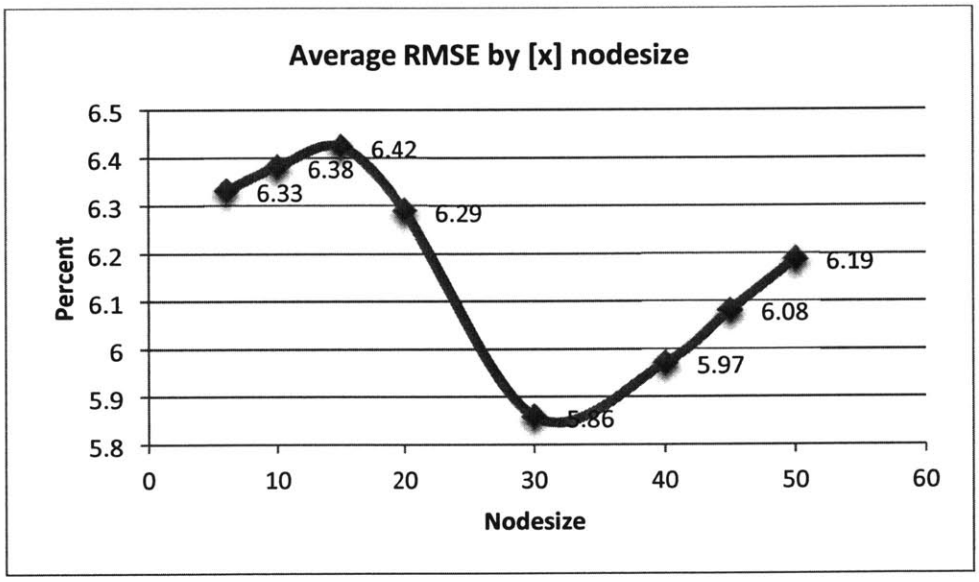
5.4.2 Nodesize

The nodesize of the forest determines the minimum number of observations used before the tree stops splitting and generates a prediction. This value is important for routes with high variance because if it is set too low, the model may tend to create unstable predictions due to the random sampling. Since Amazon's short, solo routes, which make up 16% of all of their network, tend to have high variance, it is important to find an appropriate nodesize.

The quantregForest sets the default nodesize value to 10. The test for nodesize was evaluated at 6, 10, 15, 20, 30, 40, 45, and 50 with a 5,000-tree forest created for each test. Below is a graph of the results of 100 iterations of random forest constructions and the WMAPE and RMSE. There is a clear local minimum around 30 that seemed to reduce both the weighted mean absolute percent error as well as the root mean square error. Therefore to create the most stable model, the quantile regression forest was set to have 5,000 trees, a minimum nodesize of 30 data points, and an mtry value equal to one third of the number of variables (8).

Figure 5-3 Nodesize Test: Weighted Mean Absolute Percent Error by Nodesize





5.4.3 Final Model & Maintenance of Model

The forest is currently set such that there are 5,000 trees, the nodesize is 30, and mtry is equal to 1/3 of the total number of variables in the model. These settings were chosen because they minimized RMSE and WMAPE. The final estimated values for the RMSE and WMAPE for the 117 linehaul model were 5.86% and 5.00% respectively. These values will fluctuate slightly

depending on the sample of the data but because of the testing, fluctuations should be insignificant.

The expansion of the model to include the remaining linehauls in North America required hosting the model on an Amazon Web Services server because of the size of the data set. The final error values for the network was 4.57% and 2.22% for WMAPE and RMSE, respectively.

5.5 Testing for Variable Significance

Random forests generally handle insignificant variables fairly well and the presence of extraneous variables in the model do not deteriorate the quality of the prediction produced (Breiman 25). However, the more variables included in the model, the more computation power and time are required. The development of an importance measure is therefore an area of interest for machine learning scientists in order to balance the trade off between robustness of model and computational speed. The importance measure is best used as just a rough guide to what features can be left out of the model without deteriorating the prediction.

5.5.1 Methodology: Permutation Test for Variable Significance

The most popular and well-documented methodology for testing variable significance for random forests is known as the permutation test (Universite Paris-Sud, 2010). The test, as the name suggests, has the variable permutated within the test set of observations and inputted into an existing random forest to see how the predictions change. Depending on the change of the values predicted in the permutated test set and the non-permutated test set, the importance of that variable is concluded. For the purposes of testing the model, it was assumed that the origin and nominal transit times were considered significant and were not included in the permutation tests. The remaining 16 variables were tested using the permutation test.

The test was performed in R. A forest was constructed using a random subset equal to two-thirds of the data. This random forest would be used to generate all predictions for all data sets. The remaining one-third of the data set was used to generate predictions. These predictions were considered the baseline set of predictions that would be used as a basis of comparison for the permutated predictions. The WMAPE was used to measure the quality of the predictions for the baseline set. Then, using the same one-third test set, one variable was permutated and then inputted into the forest to generate a set of predictions to measure its importance. The

WMAPE was also computed from the permutated prediction set. This permutation step was repeated for each of the variables in the data set. The WMAPE from each of the runs were compared to the WMAPE baseline set of predictions to understand each variable's influence on the predictions in the model.

Additionally, since quantile regressions generate distributions, the predictions and the model's forecasting error rate vary based on the p-value used. Therefore it was important to test variable significance at varying levels of probability on-time delivery. Since Amazon was unlikely to use a p-value less than 85%, I chose to test the variable significance at 85%, 90%, 95%, 96%, 97%, 98% and 99%. Since the loss function minimized by quantile regressions favors higher values of p, there was an expected decrease in WMAPE as the values of p increase.

5.5.2 Results of the Permutation Test

Table 7 Baseline WMAPE Rates by Probabiliity of On Time Delivery

Service Levels	85%	90%	95%	96%	97%	98%	99%
Forest (Not Permutated) Error Rates	8.89%	7.81%	5.59%	4.89%	4.12%	3.37%	2.50%

Table 8 Results from Permutation Test: Percent Error from Baseline Error Rate

Service Levels	85%	90%	95%	96%	97%	98%	99%
Variable Permutated							
Gate Depart Duration	-8.5%	-6.8%	-5.4%	-7.1%	-8.5%	-8.0%	-5.5%
5.5.2.1.1.1 LTOT Truck	-2.1%	-1.7%	-2.3%	-2.8%	-3.1%	-3.3%	-1.6%
Rush Hour	-16%	-13%	-10%	-9.7%	-9.1%	-5.1%	-1.3%
Adhoc Truck	0.1%	0.0%	0.0%	0.2%	0.0%	0.1%	0.1%
Weekend	0.2%	-0.1%	0.1%	-0.2%	-0.6%	-0.1%	0.2%
5.5.2.1.1.2 Transporter A	-11%	-9.7%	-10%	-12%	-12%	-9.2%	-1.5%
5.5.2.1.1.3 Transporter B	-5.9%	-5.9%	-4.8%	-5.4%	-6.8%	-7.6%	-7.3%
5.5.2.1.1.4 Cross Dock	-8.8%	-9.1%	-8.8%	-8.9%	-9.2%	-9.0%	-8.4%
5.5.2.1.1.5 Distribution Center (as destination)	-16.0%	-12%	-13%	-17%	-21%	-20%	-11%
5.5.2.1.1.6 Transporter C	-1.0%	-0.8%	-1.7%	-1.7%	-1.9%	-1.6%	-0.8%
5.5.2.1.1.7 Transporter D	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
5.5.2.1.1.8 Transporter E	-1.3%	-1.3%	-1.8%	-2.1%	-2.3%	-2.8%	-2.2%
5.5.2.1.1.9 Origin Facility Performance Factor	-1.7%	-0.7%	-0.4%	-0.8%	-0.8%	-1.4%	-0.8%
Carrier Controllable Performance Factor	0.0%	-0.2%	-0.1%	-0.2%	-0.6%	-0.4%	-0.1%
Seasonal Performance Factor	-0.4%	-0.4%	-0.4%	-0.3%	-0.3%	-0.2%	-0.2%
Other Performance Factor	0.1%	0.0%	0.1%	0.1%	0.1%	0.2%	0.1%

Overall the test produced very little absolute change in error rates but significant results when comparing the percent change in error. Table 8 displays the percentage error of the WMAPE between the not permutated forest configuration and the forest built respective permutated variable for a selected p-value. The percentage error is defined as the difference between the WMAPE of the not permutated forest and the WMAPE of the permutated forest, divided by the WMAPE of the not permutated forest. The results are from a permutation test for the model with 113 lanes and 5000 trees in the random forest.

The most significant changes in percent error between the baseline case were observed in the Gate Departure Duration, Rush Hour, Transporter A, distribution center as a destination, and cross dock variables. The least significant changes were seen in the performance factors, Weekend and LTOT truck. The percent differences in WMAPE from the set of baseline predictions were less than 0.5%. 1% was determined that this would be the threshold for significance.

5.5.3 Trends Between Variables

It is important to note that since random forests are robust against insignificant variables, the inclusion of these variables did not dilute the model's ability to generate meaningful predictions. The variable significance testing was done to gain further insight into the behavior of the variables as they affect transit time predictions, not necessarily to improve the predictive ability of the model itself.

Significant Variables

Gate departure duration was defined and estimated in the model as the time between fifteen minutes after the scheduled truck departure and when the carrier sent an EDI yard exit message. This was the best approximation based on existing data on how long it would take to successfully depart a truck from the yard. In examining the data, there were also instances where a truck may depart the yard earlier than their scheduled departure time. A full table of early departures categorized by business type can be seen in Appendix E. During the data cleaning stage, any historical data points that indicated the truck departed over an hour earlier than their scheduled departure time was discarded as an invalid data point. When looking at specific predictions in the tested data set, there were significant differences found in the predictions for loads that had early departure values. This intuitively makes sense based on the existing gate departure procedures. It most notably affected predictions for short haul linehauls.

This is most likely due to the fact that the gate departure process makes up a larger portion of the scheduled transit time for short hauls.

Amazon currently departs trucks in batches, defined by the Last Truck Out Time (LTOT). Due to this batching of departures, there was noticeable congestion at the exit gate of the yard. This queue can be exacerbated in inclement weather or when there is a delay in the distribution center. Delays in the distribution center are generally due to waiting on the last few packages to be loaded onto the truck or due to an unbalanced labor on the shipping dock. For any given LTOT, there can be anywhere from 2 to 10 trucks exiting the yard at the same time. Therefore, trucks that depart earlier than the listed scheduled departure time or Last Truck Out Time can avoid the congestion both immediately as they are exiting the yard and presumably any congestion on the immediate roads outside of the Amazon warehouse. This probably contributes to the significance factor of the model's prediction, across all lanes.

Rush hour was classified based on the scheduled departure time and encompassed both morning and afternoon rush hours. For obvious reasons, trucks that depart during these hours may experience more local road traffic to highways and more traffic in general than trucks that depart during non-rush hour periods. In fact, rush hour may be a better proxy in this model for traffic than the traffic late reason code. Reasons for this will be discussed later.

The significance of distribution centers, Transporter A, and cross dock is possibly due to the process of sending EDI arrival to destination messages. Carriers should send their arrival to destination EDI message as soon as the carrier arrives at the destination. However, in some cases, truck drivers will not send the message until the package is delivered to the transporter successfully. If there is a late delivery and the destination facility is no longer open to accept truckloads, the driver will have to hold the trailer until the transporter is ready to receive the load. Because of this, the arrival at destination timestamps are often inappropriately used as successful delivery of package. Amazon has not controlled for this because their primary concern is to ensure that the carrier delivers the load successfully. This phenomenon is most pronounced for transporter destinations, rather than internal destinations such as distribution centers and cross dock. Transporter A and cross dock makes up 43% of the historical external destination type of Amazon's North America network. Therefore the significance of distribution centers, Transporter A and cross dock as variables may be overstated. Additionally, these destination types are used for packages that are non-premium or inventory transfers. While it is

important to get these packages to the destination on time, these loads generally have a lower priority to carriers since there is such a large window for customer delivery.

Insignificant Variables

A surprising result of the variable significance test was the insignificance of the performance factors (origin facility controllable, carrier controllable and Other) on the model. This was surprising since these codes are the closest proxy to explaining performance anomalies. However, the nature of how they were incorporated may have affected their ability to influence or not influence the random forest model. Performance factors were not flagged automatically until March 2013 and required a manual input to update the historical report. Due to this reason, the performance factors may appear to be insignificant due to their scarcity. The performance factors may be insignificant at this point in time however may potentially become more significant as reporting improves.

Unfortunately, performance factors can only act as indicators of the type of delay that caused late arrivals. They do not identify the root cause of late arrivals and to what extent the truck was actually late. A lateness was defined for the purposes of this model as any load that arrived greater than 15 minutes after its scheduled arrival time. Therefore a load that was 20 minutes late was flagged in the same manner as a truck that was over 12 hours late. The lack of granularity essentially diluted the influence of these indicators on the model.

The lack of influence of the LTOT truck variable was also surprising. The LTOT truck is defined as the last truck to depart from the distribution center in order to make Last Truck In Time at the destination. The incorporation of this variable was based on suggestions from Transportation managers and ship clerks who manage production and scheduling to the LTOT. This bias led me to believe that there were many non-LTOT trucks, ie: sweeper and ad-hoc trucks, that would have significant variation in transit times. In examining the distribution of the types of trucks scheduled, this belief turns out to be false. This can also explain the insignificance of ad hoc trucks. Ad hoc loads make up only about 8% of the total loads in the data set. The scarcity of data probably makes them variables that have little to contribute to the regression.

The least surprising result was the insignificance of the Weekend variable. During my initial conversations with Outbound Transportation managers, they told me they believed Amazon's inability to schedule weekend and weekday trucks differently led to excessively long transit times for the weekend trucks on the same linehaul. They said that they would receive phone calls from carriers who would arrive several hours early to destinations that were not yet open on the weekends due to these inaccurate transit times. The trucks would be able to deliver the loads significantly early because of less traffic and congestion on the road on the weekends, due to the lack of commuters on the road as compared to the weekdays. This made intuitive sense to include this variable in our model. However, when coupled with the issue of delivering EDI messages after the load delivery, these early arrivals are indistinguishable in the data set. Therefore, the weekend variable suffers from a similar issue that Transporter A and cross dock suffer from, which is the inability to distinguish when the truck physically arrives at the destination versus when the packages are delivered. However, in this case, it has caused the Weekend variable to appear to be insignificant.

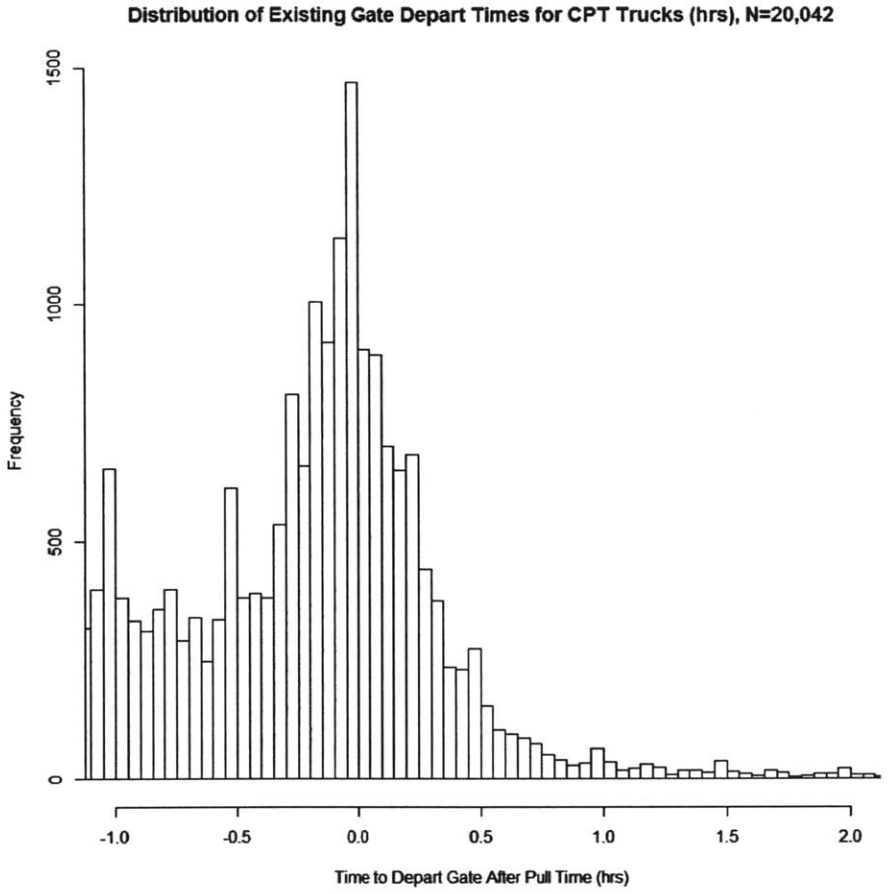
5.5.4 Trend across percentiles

The values of WMAPE decreased in the baseline case as the p-values increased. This was an expected result since as the value of p increases, the WMAPE formula becomes much more heavily weighted towards positive differences between predicted value and actual value. In general, there was a small expected decrease in WMAPE across the permutated variables as well. However, in a few instances, there were some variables where the percent difference between the WMAPE actually increased as the p-values increased. This behavior was found in all the destination types, LTOT Truck, and Gate Departure Duration. This seems to indicate that as the percent of required on-time deliveries value increases, the values they take on have much more significance on the quality of the regression value produced by the random forest.

One possible explanation for this for the destination types is the phenomena of trucks delivering past the destination's operating hours. As described in the earlier section, carriers tend to not deliver the completed delivery EDI message until the load has been delivered to the cross dock. In the most extreme cases, these deliveries may be completed up to 24 hours past their intended delivery time, despite only missing operating hours by a few hours. These types of deliveries are extreme outliers in the data and at higher percentiles, have a stronger effect on the random forest's ability to determine a transit time that would ensure high on-time delivery.

A similar effect can be seen when looking at LTOT Trucks and Gate Departure Duration together. The data cleaning removed any loads that departed from the origin earlier than 1 hour, however loads with excessively long departure times were not removed from the data set. Below is a graph of the distribution of gate departure durations for LTOT Trucks.

Figure 5-4 Gate Departure Estimates Using EDI Data (Q1 2013)



There is a long tail of LTOT trucks that can took over a half hour to exit the yard. These extended gate depart times can be caused by DC load delays, issues with gate security, congestion in the yard and exit gate or carriers forgetting to send the appropriate EDI message once the truck has exited. The latter issue is managed by carrier managers as part of carrier performance evaluations. This long tail explains why at high values of required on time performance, the Gate Departure Duration and LTOT truck become more significant.

5.6 General comments on predictions

A basic sanity check of the recommendations provided by the quantile random forest model would be to check if the historical performance at specified percentages roughly match the predictions. The low values obtained by root mean square error indicate that this is indeed true on average across most lanes. The low values obtained by the weighted mean average percent error also indicate the ability to provide transit times that would ensure a high level of on time performance, given any number of variations. This first pass at the usability of the tool indicates that quantile random forest modeling has succeeded in addressing the issue of forecasting a transit time, given certain factors that may influence a linehaul's ability to deliver its load.

However, there are also shortcomings of the model that should be addressed. The model tends to consistently provide conservative estimates for two specific types of linehauls: short hauls (less than 50 miles) and low volume linehauls. The latter is mainly due to scarcity of data. Each route can generally be identified by its origin, destination type and nominal transit time. It would be rare that any two routes would share exact transit times. Without a proper basis to sample and generate trees for, the random forest does not have an appropriate sample to generate the prediction from. I believe that the random forest determines that these types of routes are to be grouped with outliers. Extremely conservative estimates are provided for infrequently scheduled lanes due to the scarcity of data issue. Often, these lanes will generate predictions that are 3-4 times their nominal transit time. These predictions are generally not consistent with any historical performance on the lane.

Conservative estimates are also generated for short hauls, 50 miles or less. This issue is mainly due to the fat right sided tails of historical transit times. Since short hauls are generally scheduled for roughly 1 hour of transit, there is less time to recover from large unpredictable delays, such as accidents or weather delays. Smaller delays due to the truck departure process also have a greater proportional effect on these types of lanes, thus making consistent performance difficult. The conservative estimates for these lanes are 4-5 times the nominal transit time, however they are generally consistent with historical data provided on these lanes.

6 Implementation of Model & Pilot

6.1 Selection of Lanes

The development of the preliminary model was finalized at the end of May 2013 and 2 two-week pilots were conducted on a total of 27 lanes. There was a net transit time reduction of 10.5 hours. The preliminary model modeled 117 lanes, which were a mix of linehauls of varying distances and ship methods. The selection of the lanes to test were based on the following criteria:

1. Recommended transit time change – For some lanes, the model recommended changes that would have either dramatically reduced or increased the existing transit time. To minimize the risk of disruption to Amazon's business and the challenge to execute from the carriers, I chose lanes whose reductions were no greater than 10% of the existing transit time. There was no limit imposed on increasing transit time.

2. Priority of Package – Packages are divided into two categories at Amazon, high speed and low speed. Low speed packages are orders from customers who have elected a slower ship option. These packages in general have longer total transit times to customers designed into the estimated time to arrive at the customer's door. Therefore, changing them would incur lower risk of disruption to customer expectations. I chose to test lanes with only low speed packages.

3. Buffer between Last Truck In Time & Scheduled Arrival Time – For high volume carrier destinations, Amazon staggers the arrival of their trucks from various distribution centers for carrier defined Last Truck In Time (LTIT). For example, two trucks from two different distribution centers which contain packages for the same transporter may be scheduled to arrive an hour apart from each other so as not to overwhelm the transporter. This staggering implies that there is an additional buffer built into the destination, which allowed transit times to be extended without risk of missing the transporter defined LTIT. By selecting linehauls that did not compromise LTIT, the long process of distribution center negotiations to change transit times was also avoided.

4. Variety of distances with substantial historical data – Because of the preliminary assessment of lanes by length of haul, it was important to test a variety of lanes. In

addition to ensure a good mix of lanes by distance, I also wanted to ensure that each lane selected had a sufficient amount of historical data associated with it. While the random forest should be able to generate predictions for any type of lane, it generally performed much better when there were more than 20 data points for a particular lane.

Based on the above criteria, it was easier to test lanes that had recommended reductions in transit times rather than recommended extension of transit times. Amazon Transportation managers also requested that my initial pilot be contained to one origin. This site handled the highest volume of low speed packages, provided a single point of contact to discuss any issues related to departure processes and had the widest range of linehauls by distance that would provide a good sample to test with. During the second pilot, I was able to add an additional six lanes that were not from this origin. The six lanes were added mainly due to the desire to estimate financial savings of reducing linehauls, in addition to proving that existing transit times for those lanes were excessively buffered. A full list of lanes and the changes are shown in Table 10.

6.2 Selection of Service Level to Test & Expected Error Rates

After determining the appropriate criteria for selecting lanes to test, I needed to select the appropriate service level to use in the model. The model demonstrated low values of error rates for service levels above 85%, specific values shown below. What ultimately drove the decision to set tested service levels for the piloted lanes was the benchmarking with the existing performance of the initial 21 lanes in the first pilot.

Table 9 Error Rates of Model at Specified On Time Probability

Service Levels	85%	90%	95%	96%	97%	98%	99%
WMAPE	8.89%	7.81%	5.59%	4.89%	4.12%	3.37%	2.50%

A 5-week average of on-time performance was used to determine that the initial set of lanes to be tested had roughly a 95% on time performance. Each individual lane ranged in on time performance from 83% to 100%. The model was designed to ensure an average of "p-percent" on-time performance across outputted lanes. Therefore if the model was accurate, the pilot would demonstrate that overall performance would not change. The exact values of each lane's on-time performance can be found in Table 10 below, which includes the additional 6 lanes

added during the second stage of the pilot. Those lanes also had an average on time performance of approximately 95%.

Table 10 Historical Performance Prior to Pilot
(all values are in terms of hours)

Lane	Current Transit Time	New Transit Time	Change	% On Time
A	9.0	8.5	-0.50	94%
B	20.0	18.5	-1.50	95%
C	9.50	8	-1.50	100%
D	18.0	17.5	-0.50	98%
E	10.0	8.5	-1.50	97%
F	20.5	19	-1.50	95%
G	9.0	8.5	-0.50	83%
H	5.0	7	2.00	98%
I	3.5	3	-0.5	100%
J	3.5	3	-0.5	97%
K	8.0	5.5	-2.5	96%
L	11.5	12	0.5	95%
M	12	13	1	91%
N	11.5	11	-0.5	100%
O	11	12	1	95%
P	13	12.5	-0.5	95%
Q	14	13.5	-0.5	89%
R	15.5	15	-0.5	93%
S	18	17	-1	94%
T	20.5	21	0.5	98%
U	22	19.5	-2.5	80%
V	22	24.5	2.5	90%
W	27	28	1	95%
X	45	44.5	-0.5	98%
Y	53	56.5	3.5	99%
Z	54.5	51	-3.5	92%
AA	22	20	-2	96%

Based on the selected 95% on time service level and the projected weighted mean absolute percent error, the pilot would be a success based on two criteria: ability to achieve the selected service level, ie: 95% on-time, and if the recommended scheduled values had a WMAPE of 1.4% of the demonstrated transit times. 1.4% was the WMAPE of the model forest when applied to these 10 specific lanes. The network model predicted a 5.6% WMAPE, however in order to accurately access the pilot, I needed to compare the computed WMAPE for the lanes being

tested. These lanes had a much lower WMAPE because there was more historical data to generate regression values that were better estimates of demonstrated transit times. The WMAPE shown in Table 9 is the value for the forest across the initial 117 lanes used to develop the model.

6.3 Generating Predictions

The quantile random forest model works as a black box forecasting tool. The model requires the user to select the specified service level to predict, input a series of vectors describing the characteristics of a set of linehauls. Based on those inputs, it generates an output that provides a unique prediction for each inputted vector. A lane can be described almost uniquely by its origin, destination and nominal transit time, however in order to get the full benefits of the random forest, other attributes should also be included (ie: weekday, LTOT Truck, etc).

In order to generate predictions for the pilot, I selected 10 baseline scenario vectors to input for each lane and generated predictions for these scenarios. These ten characteristics made up roughly 99% of the lanes inputted to develop the model and nearly 100% of the types of lanes found through the entire network in Q1 & Q2 of 2013. The specific percentages of each scenario are shown in Table 11 below.

Table 11 Scenarios By Percentage in Network & Subset of Lanes in Model

	Total Network	Lanes in Model
Weekday, Adhoc Truck during rush hours	1.52%	1.41%
Weekday, Adhoc Truck during non rush hours	1.52%	1.41%
Weekday, LTOT Truck during non rush hours	56.38%	51.85%
Weekday, LTOT Truck during rush hours	14.01%	17.91%
Weekend, LTOT Truck during non rush hours	10.36%	9.70%
Weekend, LTOT Truck during rush hours	4.47%	5.50%
Weekday, Sweeper Trucks during non rush hours	6.03%	6.05%
Weekday, Sweeper Trucks during rush hours	2.07%	1.90%
Weekend, Sweeper Trucks during non rush hours	1.96%	2.27%
Weekend, Sweeper Trucks during rush hours	1.29%	1.39%
Total	100%	99%

In order to incorporate the performance factors, I used a five-week moving average of the historical performance within each late reason code type to calculate the likelihood of a late

reason code occurring. If the five-week moving average was greater than or equal to 0.50, the late reason code value was 1, otherwise it was set to 0.

Once the vectors were created, they were inputted into the random forest model and a series of predictions were generated for each of the piloted lanes. The recommended transit time to test was based on the highest frequency of the type of scenarios in the schedule. In most cases, it was the LTOT truck, non rush hour, weekday that was used in the schedule.

6.4 Results from Pilot

During the pilot, the lanes performed at 93.6% on-time with a standard deviation of 1.81%. Prior to the pilot, the piloted lanes performed at 94.6% on-time with a standard deviation of 1.80%. If the changes were not made the lanes would have demonstrated 95.6% on-time during piloted period. This was determined by comparing the old scheduled transit time to the actual transit times during the pilot. Below is a tabulated summary of performance five weeks prior to the pilot and during the pilot.

Table 12 Summary of Results from Pilot

Week #	-4	-3	-2	-1	0	1	2	3	4
% of Lateness due to change[4]	PRE-PILOT					33%	58%	24%*	28%
% on Time	94%	96%	93%	94%	96%	92%	95%	92%	95%
% on Time without Changes	PRE-PILOT					94%	96%	94%	97%
% on Time removing Traffic	PRE-PILOT					94%	97%	94%	96%

*Note: Spike in number of lateness during week 2 is due to six mechanical breakdowns, unusually high for one week. Less than 1% of all loads were affected by mechanical breakdown across NA in Q1 & Q2.

Table 13 Reasons for Lateness Due to Reduction in Transit Time
Summarized by major late categories

	Week Number			
Reason	**1**	**2**	**3**	**4**
Traffic/Accident/Construction	2	4	4	4
DOT inspection	1			
Unknown (Pending response Pending Response from carrier	1			
Carrier Controllable	2	1	1	1
6.4.1.1.1.1 Origin Facility Controllable				2
Total	**6**	**5**	**5**	**7**

[4] Lateness due to change are defined as late arrivals to destination that did not exceed the reduction in transit time (ie: if transit was reduced by a half hour but the load arrived 3 hours late, it was not counted as a late that was a result of the transit time change)

During the first week of the pilot, there was a small increase in the number of late on lanes that saw a reduction in transit time. Lanes that had transit times augmented saw an increase in on-time performance. This was expected. Based on the performance of the second week of the pilot, drivers had roughly one week to adjust to the new expectations of the scheduled transit time to achieve an average 95% on-time performance. Therefore, the lanes were able to adjust their behavior to meet the predicted scheduled values within a fairly short period of time. This also demonstrated that the model provided recommended transit times that could successfully predict a specified service level.

6.5 Evaluating Accuracy of Model & Performance

In addition to evaluating the model's ability to provide transit time predictions at a specified service level, it was also important to evaluate the accuracy of the model compared to actual performance. The results of the pilot were actually quite interesting. The RMSE and WMAPE are tabulated in Table 14. Based on the estimates, I expected the WMAPE for the selected lanes to be roughly 1.4%, however it was higher than that. When computing the WMAPE value against the actual transit times during the pilot, the WMAPE was 1.6%. This implies that the difference between the predicted transit time and the actual demonstrated values were greater than predicted. This was true even after controlling for the time it required for carriers to adjust to the new schedule. Therefore the actual forecast accuracy of the selected lanes was slightly overstated when executed, compared to the predicted forecast accuracy.

However, the demonstrated value prior to the pilot with the original transit times was 1.9%. This signals that there was a slight reduction in variation relative to the previously scheduled transit times. This makes intuitive sense since most of the lanes had a reduction in transit time. Because of the bias implicit in the WMAPE calculation for demonstrated transit times that are greater than the predicted transit time, the RMSE was used to evaluate whether this claim was true.

The demonstrated RMSE prior to the pilot for a 27 lanes is 5.9 hours. The demonstrated RMSE from the pilot was 5.0 hours. This decrease also makes intuitive sense because there was an overall reduction of 10.5 hours across all the tested lanes. Given the new tighter schedule, the carriers were now being forced to perform more consistently to the new lowered transit times. In other words, because time was reduced, the overall perception of buffer was also reduced,

potentially causing drivers to drive more effectively. What is surprising however is that the predicted RMSE was 6.2 hours, which is 24% greater than the demonstrated value. This indicates that the historical loads that were used to estimate the standard deviation from the demonstrated mean are likely much higher than they should be.

From this result, I concluded that given an excess amount of time to drive to a destination, truck drivers may decide to take as long as they want, rather as long as they need. This conclusion also confirmed an anecdotal bias from Transportation managers that carriers will request excess transit, rather than attempt to improve driver performance. This implies that scheduling solely based on historical performance by lane is not a good way of determining the "true" transit time. This also may be an indication that the new scheduled transit time may be closer to the nominal value of an efficiently scheduled transit time compared to the old schedule.

Table 14 Predicted vs. Demonstrated Error Values for Model

	WMAPE - Demonstrated	WMAPE - Predict	RMSE - Demonstrated	RMSE - Predict
Pre-Pilot	1.95%*	1.44%	5.91*	6.18
Pilot	1.63%	-	5.01	-
Wk 4, 6 only	1.64%	1.64%		

*Pre-pilot demonstrated values use the old scheduled transit times, whereas the predicted values use the new scheduled transit time recommendations generated from the model

6.6 Pilot Implications on Supplier Engagement

The discussion of the results from the pilot will mainly be limited to the lateness that were due to transit time reduction. There were two prominent categories that contributed to 83% of these late arrivals to destinations, Traffic and Carrier Controllable issues. Both of these issues are external factors that cannot directly control but can influence through effective supplier engagement.

6.6.1 Carrier & Truck Driver Response to Pilot

The carriers were not very supportive of the transit time reduction. This was expected since their performance evaluations are directly related to their on-time performance. They were given two weeks notice about the pilot. This was done in part to ensure that any carriers that use independent contractors could communicate with their drivers that a scheduling change was going to be made. However, in the first week of the pilot, two of the carriers explained their

lateness as the driver being unaware of scheduling changes. The carriers were quick to claim that the new transit times were impossible to make given traffic in urban areas that the routes would drive through. Without direct visibility into the truck's progress en route to destination, retailers are inclined to take these types of claims from the carrier as truth. However, after the pilot and analysis of the results, it was demonstrated that drivers may not be forthcoming about actual required transit times.

It should also be noted that the second most prominent category of lateness due to the change were Carrier Controllable. These issues contributed to 22% of the total lateness that were caused by the changes in transit time. Specifically, instances of late dispatching and drivers running out of hours causing them to stop driving, were particularly notable. Quality management will be critical to improving on-time performance to destination.

6.6.2 Traffic – Potentially Overused Carrier Reported Issue

Due to the reduction of transit time, the piloted lanes experienced an average 2% reduction in on-time performance during the pilot. However, when the traffic claims were addressed with the carrier by lane, that lane experienced almost no lateness the following week. This leads to the conclusion that once the carrier had drivers adjust their expectations, they were able to perform to the new transit time. If all lateness due to "Traffic" was removed from pilot period, on-time performance would better match percentage on-time without changes. In fact, by the end of the pilot, the demonstrated rate was equal to on-time performance prior to the pilot.

While traffic is understood superficially, there is no formal definition of traffic. It is difficult to define what would qualify as "traffic", ie: an extended wait at an on ramp or one hour congestion of cars during rush hour. This loose definition of traffic does not allow carriers to fully describe the issue. It is also a performance factor that is neither origin facility nor carrier controllable, which may lend to being overused when there are no other appropriate classification codes. Clarity over the use of the code "traffic" or the replacement of traffic with more specific performance factors could also help improve the model, as it drives as the true root cause of the issue.

6.7 Implementation of Model as a Tool

Based on the results of the pilot, the quantile random forest received a lot of attention from Amazon senior Transportation managers, Transportation software engineers and Transportation data scientists. The random forest application was inspired by an internal study that had been conducted by a different internal team that attempted to quantify transit time from distribution center to customer. Unfortunately, the earlier study resulted in unfavorable results so when this model had successfully predicted transit times based on a specified service level, Amazon wanted to turn this model into a tool. At the conclusion of my internship, software engineers and data scientists had been allocated as resources to develop this model into a web interface service tool that Transportation managers could use. The final product is not a direct replicate of the model I built, however the quantile regression forests will act as a framework for developing a similar tool. The estimated start of this project was Q1 2014.

In the interim of software development, I was asked to create a protocol that would enable an analyst to reproduce my model.. The interim tool that I provided to the Transportation managers was an abridged version of the full functionalities that the quantile random forest model was capable of. The model was modified due to the user's requests to simplify the output. To compensate for this request, the model provides a friendly user interface for a non-technical analyst that outputs a minimum and maximum predicted value for the ten scenarios outlined in Table 11.

While the pilot did demonstrate the power of the model, Amazon Transportation managers did not like its black box style forecast. In order to help them understand where these predictions were being generated, the Interim tool provided the total number of historical loads used to generate the prediction by lane. Additionally, they also requested a list of transit times by quantile be provided with the recommendation so they could see the historical performance by lane. I was careful to explain that these historical distributions should not be used as the primary source of determining transit time, due to carrier bias. By providing the historical distribution, they were much more likely to accept conservative or aggressive estimates. Appendix D shows an example of the output of the tool I provided using hypothetical data.

6.8 Updating and Maintaining the Model

In order to continually improve on transit time estimations, it is necessary to continually feed data into the model periodically. The advantage of using a machine learning algorithm is that it can handle large datasets and incorporate additional data to update the regression values as

performance improves. I advised Amazon to set up a monthly schedule to update the model. A month was chosen because updating the model is a manual process that requires input from multiple sources. A month seemed like a reasonable amount of time so that any unnoticeable changes in either seasonal trends that could impact transit time (weather, traffic pattern changes, etc) would not be missed.

I also advised that the model be updated if a significant event has occurred that may affect how predictions are made by the model. A significant event would be defined as:

- New origin is introduced into network
- New destination type is introduced into network
- Process improvements that may affect gate departure duration
- Any changes to carrier management (performance factor updates, EDI messages etc)
- Significant changes to government regulation that may affect Nominal Transit Time
- Any changes to transportation network not already listed above

6.9 Predicting Transit Times for New Linehauls

The quantile random forest is effective for predicting transit times for existing lanes with historical data points. However, in order to provide a methodology for calculating transit time, I must also determine a methodology for predicting transit times for new linehauls as they are introduced into the network. New linehauls can be classified into the following categories:

1. New Linehaul Connecting Existing Nodes (origins and destination)
2. New Linehaul Connecting to a New Destination but an Existing Destination Type
3. New Linehauls for a New Distribution Center or new Destination Type

The quantile random forest model relies on two types of information: intrinsic characteristics of a lane (ie: origin, destination, nominal transit time) and historical performance data (ie: gate departure duration and frequency of performance factors). Intrinsic characteristics are available when the linehaul is created. The lack of historical performance data however can be estimated or assumed when assigning values to the input vectors. However, since gate departure duration and performance factors were shown to have little influence on transit time predictions, the initial estimates are more or less unimportant.

Generating transit time predictions that connect existing nodes is the easiest case to deal with. Since the origin and destination already exist as factors in the input vector, the random forest will simply use the attributes associated with the origin and destination to generate a reasonable prediction for the new linehaul. The same is also true for the second scenario where the destination type already exists in the model (ie: another cross dock or existing transporter). Estimates from these two scenarios will provide transit times that are similar to other lanes in the network that share these origin and destination types and nominal transit times. It is intuitive that this would occur since these attributes were found to be important variables that define transit times.

The last scenario proves to be most challenging. If the origin or the destination type does not already exist in the model, a new origin and destination type will have to be added. However, there is no way of doing this since there is no historical data to rebuild the forest with. Therefore, in order to estimate transit times for these types of lanes, the user must make several assumptions of the new linehauls. The user must choose a distribution center that they believe is similar to the new one being introduced to the network. There is also may be sufficient similarity in many surrounding areas that it would be reasonable to assume a distribution center opening nearby an existing one would share its tendencies to have traffic, weather delays etc. The same must also be done for new destination types, however this task may be more challenging as there may not be a comparable one to choose from

This process provided reasonable guesses at what the transit times would be. During my time at Amazon, several new origins were introduced to the network. When employing this process, the model struggled with predicting mid-short haul lanes (under 100 miles). This is believed to be due to the large variation associated with these types of lanes. For linehauls that exceed 100 miles, it provided estimates similar to what would be used in Amazon's current methodology.

7 Recommendations and Conclusion

7.1 A Perfect Schedule?

Due to the numerous sources of variation associated with outbound linehaul transit, the construction of a perfect schedule would require the ability to anticipate delays and the duration of each delay. This thesis attempted to find attributes that could be collected in a large data set and employ a machine learning technique that would use statistical techniques to anticipate delays based on historical trends. It was discovered that these key indicators have varying levels of influence by lane. Therefore it was impossible, using this methodology, to make generalized statements on how much impact a particular delay would have on a non-specified lane. The trade off between accuracy and visibility was made based on the needs of Amazon. Operationally they understand that there is room for reductions in transit times from process improvement. In the interim of these process improvements occurring, I was asked to deliver a tool that they could use to manage day-to-day operations. The quantile random forest provides them the capability to select a service level and predict the required transit time to a high degree of accuracy. Any adjustments in transit times will impact on-time performance and can be quantified through the differences in predicted transit time by the model. Unfortunately, it can be difficult to understand which attributes have caused the change in performance between service level values.

7.2 Recommendations

The quantile random forest gives Amazon an effective tool for calculating transit times. However, it currently provides conservative estimates for high percentage on-time performance and short hauls due to high variability with these scenarios. While the pilot exemplified that there are linehauls where transit time reduction can occur at 95% on-time performance, almost all of the recommendations to achieve 99% on-time performance required extending the existing schedule transit. Therefore realization of this level of on-time performance in today's system will be challenging. To effectively achieve this goal, Amazon will have to ensure that they have strict control and alignment from all stakeholders of the delivery process. However, operational opportunities exist to reduce variability and drive the actual transit times down without compromising SLA.

Based on learnings from the development of the random forest model, the following recommendations were created to help Amazon achieve its on time arrival to destination goal. These recommendations have been grouped into two categories: operational and data.

Operational recommendations aim to achieve a more efficient process management of the linehauls. These recommendations aim to improve the existing system. Data recommendations have been listed to help Amazon better quantify the system. They stem from needs realized during model development and would help improve the quality of the model developed. While the quantile random forest is an effective model, it is a black box type model. The data recommendations attempt to move away from this type of forecasting towards a more transparent methodology that can also help operation managers drive process improvement.

7.2.1 Operational Recommendations

Hold distribution centers accountable for yard departure process – Accountability was one of the major needs identified at the kaizen. Today,origin facility controllable variability and carrier performance are inherently connected since carriers may use distribution center departure delays as a reason for late arrival to destination. While carriers do play a part in the departure process, an organization should hold its distribution centers accountable for the time to depart a truck. It should be noted that this will be challenging to manage for distribution centers since 3 of the 4 parties are outsourced vendors (guards, yard hostlers and truck drivers). With some process redesigns that were explored during the kaizen, I believe there could be significant operational improvements made over the yard departure process.

Give truck departure teams the ability to prioritize departures of certain types of lanes – Variability among linehauls of differing lengths is one factor that was apparent during the initial assessment. This also became apparent when the quantile random forest would produce estimates 3-4 times the average transit time for short haul lanes. Since short haul lanes are more impacted by departure delays and have less time to recover during its drive time, the ability to prioritize short haul departures in the yard may provide better performance for these lanes. Additionally, best practices may require deprioritizing inventory transfers since they generally do not impact customer orders. I was told this is informally done today but is not standardized. A structure that will allow dock teams to determine which trucks need to be prioritized during the LTOT when many trucks are leaving at once will enable them to lower the variability from load to load.

Hold carriers accountable for variability of arrival in addition to on-time percentage - Carriers have been trained to only be concerned about on-time performance. However good carrier management must also try to control for variation associated with linehaul performance. This

can begin with holding carriers to a metric that measures their reliability to perform to the schedule. RMSE can be helpful with managing consistent performance. Monitoring some level of performance "accuracy" relative to the schedule may help drive performance of suppliers, rather than simply looking at percentage on-time. The adoption of this metric may be challenging. Carriers may push back, claiming uncontrollable variables for inconsistent performance or worse, be dis-incentivized to drive efficiently so that transit times may be inflated. However, it was shown during the pilot that carriers are already dis-incentivized to drive efficiently so it changes nothing from today's circumstances.

7.2.2 Data Recommendations

Creating a scheduling system that allows for different transit times for each scenario- One of the scheduling limitations identified by Amazon's Transportation managers is the inability to schedule a weekday truck differently from a weekend truck on the same route. This need exists because the traffic conditions vary significantly on weekdays versus weekends. This issue was also apparent when developing and testing the model for variable significance. The ability to calculate transit time by lane, under specific circumstances (ad-hoc, weekday, etc), will enable flexible scheduling that should result in cost savings due to reduced transit, without compromising service levels. It will also be favorable to carriers as there will be greater flexibility to assign appropriately long transit times.

Remove "traffic" as a performance factor code – The current method of reporting traffic forces carriers to use traffic if they are late and all other reason codes are not applicable. One reason this performance factor exists today is because it is not possible to take into account variation in traffic patterns along the same lane. As more sophisticated scheduling capabilities are developed, ie: scheduling by time of day and lane, specific transit times can be set to incorporate traffic without over scheduling other loads in the same lane. Traffic and road congestion can therefore be considered a seasonal, predictable attribute that naturally occurs at certain hours of the day. Other instances of traffic would be captured by codes such as "accident" or "weather". The quantile random forest is able to provide estimates based on time of day of departure and lane to take the normal congestion into account. When this is achieved, traffic should never be a relevant code and can be removed. In general, any performance factor that lacks transparency or is difficult to quantify should be removed.

Clarify ownership of each transit time process by detaching yard departure process from the calculation of transit time – While the process of yard departure and drive are inherently linked, the calculation of transit time would be simplified by removing the yard component. Linking them creates an accumulation effect to occur with performance factors. Trying to attribute more than one reason introduces other complications such as trying to determine what percentage of the delay should be attached to each code. Ultimately, since these two processes have their own independent set of possible variabilities, they should be treated as two independent calculations.

Link virtual trailer depart from dock time to the need to physically depart – The estimate used by the model assumes that 15 minutes from the scheduled depart time is actual depart time from the dock. Based on observations during the kaizen, this is clearly not true. Without being able to accurately measure the time it takes from physical trailer departure from dock to exit the yard, it will be very difficult to measure improvements in implementing standard work of yard departure process. If the dock departure process signaled to each driver in the yard electronically which door to proceed to, it would not only solve a data quality issue but improve channels of communication between distribution centers and drivers.

Implement a system wide yard management system to understand yard movement – The implementation of a yard management system would allow management of the movement trailers and tractors that interact in the yard. An RFID tagging system was implemented in several distribution centers in North America as a pilot, however it has not been fully implemented through the network. Implementing this yard management system enable monitoring of the traffic in the yard, but to own the reliability of the information of when trucks exit and enter its yard. During the time of the project, the only network wide data source that collects this information is through carrier controlled EDI. The incentive for carriers to ensure data quality is low and was seen while data cleaning. By owning the data, Amazon will be enabled to not only schedule better but also manage its yard more effectively.

Change process of sending EDI arrival at destination among carriers – In the case of carriers arriving at their destination, Amazon has incentivized its carriers to ensure the data quality is high by closely managing on-time performance. However, there needs to be a distinction between the time the carrier arrives at the destination and the time the load is delivered. This issue was highlighted when looking at high service levels. Due to the constraint that transporters will have limited operating hours, if the truck arrives at the destination while the

transporter is closed, using the load delivery time is an inaccurate measurement of the transit time. Amazon can incentivize carriers to make this distinction by explaining it would allow Amazon to appropriately allocate transit time so they are not late to delivery.

7.3 Conclusion

The development of the quantile random forest enables Amazon to calculate transit times of linehauls based on a desired service level. This understanding began with looking at historical performance, however there was no systemic methodology to incorporate historical data into the scheduling. The historical performance was generally limited to the percentage of on-time performance.

The quantile random forest allows for a large body of information, based on EDI carrier messages, to be incorporated into the development of transit time. It also builds on historical performance, allows for updating forecasted transit times as processes improve, and allows the user to understand service level relative to allocated transit times. However, as discovered through the pilot, historical transit times may not be the best methodology to compute the transit times. The model, as with most historical data based model, assumes that the system is stagnant. Relying only on historical transit times to schedule efficiently is therefore misguided. Operational improvements must be made to drive progress in on-time delivery to destination.

These operational improvements will require aligning incentives for carriers and the teams responsible for the truck departure process. This process will be challenging, as there are multiple teams who are contracted. Specifically, the yard departure team has two externally managed contractors who must work efficiently with warehouse personnel to ensure communication from warehouse to yard is clear. The operational improvements that can be achieved by ensuring the design of this process is standardized will result in impactful financial savings.

The advantage of using quantile random forests, and machine learning algorithms in general, is that consistently updating the structure will appropriately adjust the transit times as operational improvements are made. The implementation of the quantile random forest provides a dynamic and consistent methodology that will adapt as distribution networks grow.

8 Appendix

8.1 Appendix A: Performance factors

Reason Code	Other	Origin Facility Controllable	Carrier Controllable	Seasonal
[Not Late]	0	0	0	0
CARP Issue	0	1	0	0
Late Sweep Request	0	1	0	0
Pending Investigation	1	0	0	0
Loaded Overweight	0	1	0	0
Schedule Error	0	1	0	0
Vendor Delay- Pickup Scheduling	0	1	0	0
Vendor Delay-Loading	0	1	0	0
Accident	1	0	0	0
Border Delay	1	0	0	0
DOT Inspection	1	0	0	0
Traffic	0	0	0	1
Weather	0	0	0	1
Capacity	0	0	1	0
Dispatch Error	0	0	1	0
Driver Error	0	0	1	0
Mechanical	0	0	1	0
Medical	0	0	1	0
Prior non-amazon load	0	0	1	0
Rail Delay	0	0	1	0
Carrier Not Responding	0	0	1	0
DC Load Delay - Truck Utilization	0	1	0	0
DC Load Delay - No Response	0	1	0	0
DC Load Delay – DC Disputed	0	1	0	0
DC Load Delay - DC Admin Issues	0	1	0	0
DC Load Delay - TOC Admin Error	0	1	0	0
DC Load Delay - Physical Loading Issues	0	1	0	0
DC Load Delay - Pending	0	1	0	0
DC Load Delay - Late Depart	0	1	0	0
DC Load Delay - Carrier Reported	0	1	0	0
Previous stop	1	0	0	0
Carrier disputed	1	0	0	0
Buffered Arrival	0	0	0	0
Driver Late	0	0	1	0

8.2 Appendix B: Speed Limits For Trucks by State

State	Trucks (MPH)		State	Trucks (MPH)
Alabama	70		Rhode Island	65
Alaska	65		South Carolina	70
Arizona	75		South Dakota	75
Arkansas	65		Tennessee	70
California	55		Texas	75-80
Colorado	75		Utah	75-80
Connecticut	65		Vermont	65
Delaware	65		Virginia	65-70
District of Columbia	55		Washington	60
Florida	70		West Virginia	70
Georgia	70		Wisconsin	65
Hawaii	55		Wyoming	75
Idaho	65			
Illinois	65			
Indiana	65			
Iowa	70			
Kansas	75			
Kentucky	70			
Louisiana	70			
Maine	65			
Maryland	65			
Massachusetts	65			
Michigan	60			
Minnesota	70			
Mississippi	70			
Missouri	70			
Montana	65			
Nebraska	75			
Nevada	75			
New Hampshire	65			
New Jersey	65			
New Mexico	75			
New York	65			
North Carolina	70			
North Dakota	75			
Ohio	65-70			
Oklahoma	70-75			
Oregon	55			
Pennsylvania	65			

8.3 Appendix C : Pearson Correlation Matrix

	Nominal Transit Time	Day 6	Day 5	Day 4	Day 3	Day 2	Day 1	Seasonal	Carrier Controll able	Origin Facility Contrl.	Rush Hour	LTOT Truck	Miles	Gate Depart. Duration
Gate Depart Duration	-0.03	-0.03	-0.01	0.01	0	0.04	-0.03	0.08	0.28	0.3	0.05	-0.08	-0.03	1
Miles	1	0	0	-0.02	-0.01	0.03	-0.04	0.07	0	-0.13	-0.04	-0.04	1	-
LTOT Truck	-0.04	0.02	-0.01	0.02	0.03	0	-0.06	0.03	-0.04	-0.01	0.07	1	-	-
Rush Hour	-0.05	0.01	0	-0.06	-0.02	-0.06	0.02	-0.01	0.04	0.04	1	-	-	-
Ori. Fac. Controlla ble	-0.12	-0.04	-0.01	0	0.04	0.04	-0.07	-0.1	-0.13	1	-	-	-	-
Carrier Controlla ble	0	0.03	0.02	-0.03	-0.04	-0.02	0.01	-0.05	1	-	-	-	-	-
Seasonal	0.07	-0.01	0	0	0	0.02	0.01	1	-	-	-	-	-	-
Day 1	-0.04	-0.13	-0.17	-0.17	-0.16	-0.16	1	-	-	-	-	-	-	-
Day 2	0.03	-0.15	-0.2	-0.2	-0.19	1	-	-	-	-	-	-	-	-
Day 3	-0.01	-0.15	-0.2	-0.21	1	-	-	-	-	-	-	-	-	-
Day 4	-0.02	-0.16	-0.22	1	-	-	-	-	-	-	-	-	-	-
Day 5	-0.01	-0.16	1	-	-	-	-	-	-	-	-	-	-	-
Day 6	0	1	-	-	-	-	-	-	-	-	-	-	-	-
Nominal Trnslt Time	1	-	-	-	-	-	-	-	-	-	-	-	-	-

8.4 Appendix D: Sample Output of Tool

Predicted performance level given the current transit scheduled

Historical performance for lane with exceptions removed. See Section 2 for list of exceptions

LANE	Current Transit	Current Estimated %On Time	Min Prediction (hours)	Max Prediction (hours)	H.100%	H.95%	H.90%	H.85%	# of obs
A	22	90%	19.83	20.67	28.58	25.29	21.90	21.71	84.00
B	8.5	100%	7.72	7.80	8.15	7.80	7.63	7.58	102.00
C	9	100%	8.32	8.5	10.63	8.93	8.57	8.32	91.00
D	20.5	100%	19.00	20.37	26.93	22.18	21.55	20.49	91.00
E	3.5	100%	3.33	3.37	2.97	2.73	2.53	2.42	82.00
F	3.5	100%	3.33	3.37	3.10	2.73	2.65	2.50	63.00
G	19.5	94%	17.9	19.40	20.97	19.50	19.33	18.81	86.00
H	22	98%	21.17	23.00	23.50	21.87	21.70	21.55	43.00
I	9	96%	6.57	8.98	9.42	9.17	8.93	8.50	83.00
J	19	97%	17.25	19.00	19.48	18.53	17.97	17.71	91.00
K	22	93%	21.15	22.78	23.48	22.87	21.89	21.47	97.00
L	53	81%	31.88	53.00	67.00	64.46	53.28	53.00	70.00
M	9	82%	8.5	9.42	12.90	10.64	9.56	9.10	66.00
N	18	90%	17.25	18.88	23.50	19.42	17.94	17.83	63.00
O	7	92%	7.68	8.00	8.27	7.02	5.51	5.00	37.00
P	12	97%	12.97	13.82	12.08	10.99	10.57	10.36	56.00
Q	14	92%	13.90	16.50	16.32	14.88	14.76	14.67	70.00
R	18	92%	17.50	19.77	17.72	17.39	17.21	16.73	57.00

The range is based on the 10 scenarios that that tool takes into account. Estimates may be somewhat inconsistent with historicals since estimates are based on similar lanes across entire network

8.5 Appendix E: Note on Linear Models Attempted

For reasons discussed in Chapter 2, the original linear models did not sufficiently suit the needs of the project. An artifact that carried over into the final model from the linear models were variables that were eliminated from the linear models based on F-tests and correlation tests. Since the random forest uses the quantile regression to generate its regression value, it is assumed to be insignificant in the random forest implementation as well.

Correlation tests

While random forests are capable of handling correlated variables, to simplify the model, I eliminated highly correlated variables from the final list of variables in the random forest model in order to simplify the number of inputs. Significant correlation is defined as greater than .5 or less than -.5 (citation?). Miles was eliminated from the model since it is almost perfectly correlated to nominal transit time. No other variables have any significant correlation to each other. The full Pearson correlation matrix can be found in Appendix C.
The following variables were eliminated through the development of the linear model:

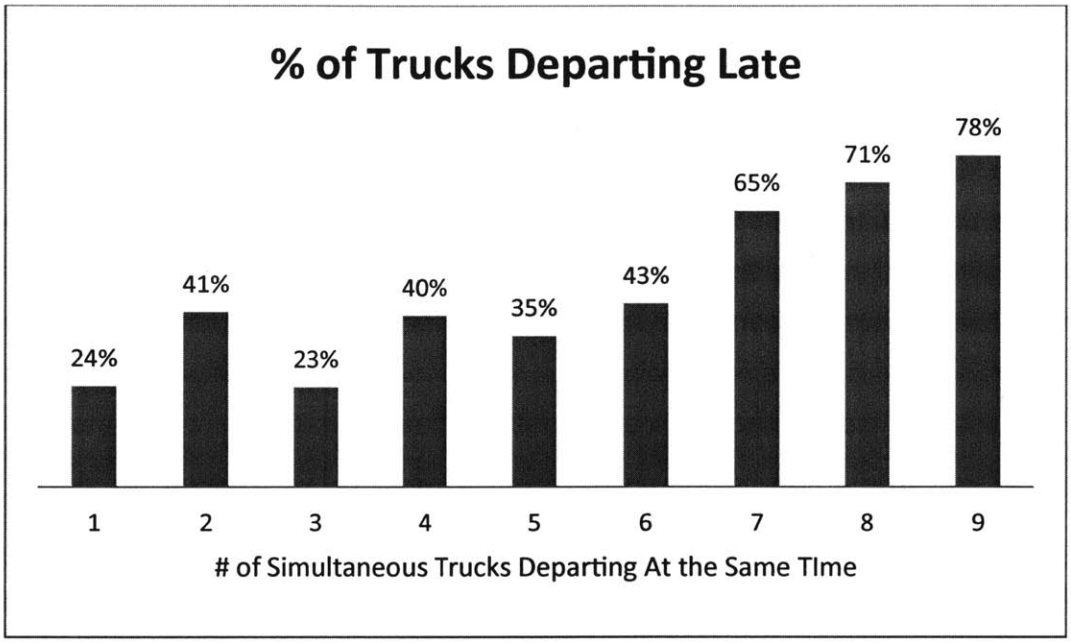
1. **Destination Arrival Delay** – Difference between scheduled arrival time of truck at destination and actual arrival time at destination. Actual arrival time provided by EDI 214 X1 NS message. Scheduled arrival time to destination provided by Amazon schedule.
   Rationale for Exclusion: This variable assumes that there is a fixed schedule to adhere to and including it in the model produced results that were very close to scheduled transit time. It seemed to force an over fitting to a predetermined transit time, which was undesirable.

2. **Miles** – distance between origin and destination pair given by transportation software
   Rationale for Exclusion: This variable was perfectly correlated with Nominal Transit Time, which makes sense since Nominal Transit Time is derived from length of haul. Including correlated variables in a model increases the risk of over fitting, so it was discarded (Breiman 20).

3. **Number of Trucks Departing** – the total number of trucks departing at the same scheduled depart time
   Rationale for Exclusion: The motivation for including this variable was to try to determine some relationship between the number of trucks departing and ability to depart within 30

minute window. A time study conducted by Amazon showed that the number of trucks departing simultaneously increase the percent of late departures, as shown in Figure 8. Unfortunately this variable does not play an important role in the larger calculation of transit time since time to depart from the gate is such a short portion of most transit times. The F-tests from the quantile linear model also determined this input was not significant in the regression.

Figure 8-1 Late Depatures vs. Number of Trucks Departing Simultaneously



4. **Day of Week** – 7 binary variables, 1 to indicate true, 0 to indicate false for each day of the week
   Rationale for Exclusion: This was simplified into weekday or weekend. After consulting with transportation managers, they believed that little variation occurred any given weekday but significant variation may exist between the weekdays and weekends. This was also helpful in limiting the total number of variables

5. **Scheduled Depart Hour** – the hour the truck was scheduled to depart
   Rationale for Exclusion: This value was eliminated due to the need to introduce 24 unique categorical values into the model. To keep the data set simple, I created a variable called Rush Hour that would indicate whether the departing truck was going to

be leaving the DC between the hours of 6-9 or 15-18, which could capture any variations due to traffic congestion.

6. **Origin Arrival Delay** – Difference between scheduled arrival time of truck to origin and actual arrival time at origin. Actual arrival time provided by EDI 214 X3 NS message. Scheduled arrival time to origin is provided by Amazon schedule.
   Rationale for Exclusion: While this is an important metric to track to ensure carrier performance, it is outside of the defined transit time process. It was therefore defined as an extraneous value, despite being tracked by the Transportation team for carrier management.

8.6 Appendix F: Gate Departure Duration by Business Type

Loads from January – April 2013 in NA

The table below shows a significant number of loads departing earlier than pull time. This is mainly due to transfers, outbound from B and tote loads. A full breakdown of the distribution by business type can be found below. Without accurate timestamps of when the truck pulled off the dock, we cannot estimate the actual duration to depart the yard. An estimate that used the timestamp associated with dock door close provided too many exceptions and was disregarded.

Negative values highlighted in red

	Gate Departure Duration by Business Types						
	0%	**25%**	**50%**	**75%**	**100%**		**% of Loads**
INBOUND TO A	(0.98)	0.03	0.25	0.45	25.50		13%
INBOUND TO B	(0.98)	(0.07)	0.20	0.37	5.27		3%
INBOUND TO C	(0.98)	(0.10)	0.18	0.43	24.18		6%
LOCAL TRANSPORTERS	(0.98)	(0.68)	(0.07)	0.25	12.57		8%
NON-INVENTORY	0.87	0.87	0.87	0.87	0.87		0%
OUTBOUND FROM B	(0.98)	(0.72)	(0.45)	(0.10)	22.92		10%
POSTAL INJECTION	(0.98)	(0.07)	0.20	0.40	43.87		30%
TBD	(0.10)	0.08	0.25	0.50	1.17		0%
TOTE LOAD	(0.98)	(0.58)	0.12	0.93	26.75		0%
TOTE RETURN	(0.98)	(0.44)	0.15	1.29	38.00		3%
TRANSFER - A	1.78	1.86	1.93	2.01	2.08		0%
TRANSFERS - B	(0.40)	(0.40)	(0.40)	(0.40)	(0.40)		0%
TRANSFERS - C	(0.98)	(0.52)	(0.10)	0.22	24.08		7%
TRANSFERS - D	0.60	0.75	0.89	1.04	1.18		0%
TRANSFERS - E	(0.97)	(0.58)	(0.20)	0.11	2.78		1%
8.6.1.1.1.1 RETURNS	(0.77)	0.25	2.84	4.32	6.85		0%
TRANSFERS - F	(0.98)	(0.23)	0.17	0.73	23.02		0%
TRANSFERS - G	(0.22)	4.52	7.27	14.59	23.88		0%
TRANSFERS - H	(0.98)	(0.30)	0.02	0.30	8.57		17%
8.6.1.1.1.2 OTHER	(0.97)	(0.27)	0.20	0.38	2.88		1%
WAREHOUSE DEALS	0.32	0.52	0.72	0.76	0.80		0%
						Total	**100%**

9 References

Breiman, L. (2001). Random Forests. *Machine Learning* , 45:5-32.

Meinshausen, N. (2005). Quantile Regression Forests. *Journal of Machine Learning Research* .

Green, Heather. How Amazon Aims to Keeps You Clicking. February 18, 2009. Retrieved from: businessweek.com/stories/2009-02-18/how-amazon-aims-to-keep-you-clicking.

Robin Guneur, J.-M. P.-M. (2010). Variable Selection Using Random Forests. *Pattern Recognition Letters* , 2225-2236.

Skaltsas, G. (2008). *Analysis of Airline Schedule Padding on U.S. Domestic Routes.* Cambridge: Massachusetts Institute of Technology.

Thomas,Owen. Here Are the 10 Cities Where Amazon Offers Same Day Delivery (And Why We'll See More Soon). August 6, 2012. Retrieved from: http://businessinsider.com/amazon-local-express-delivery-2012-8.