Using a Gaussian Mixture Model to measure transit time bimodality and its impact on inventory decisions

by

Aravindan Jayantha
Bachelor of Science in Mathematics and Statistics (University of Melbourne, Australia)

and

Didi Dai
Bachelor of Science in Industrial Engineering (Worcester Polytechnic Institute, USA)

SUBMITTED TO THE PROGRAM IN SUPPLY CHAIN MANAGEMENT
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF APPLIED SCIENCE IN SUPPLY CHAIN MANAGEMENT
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2022

Signature of Author: _____
Department of Supply Chain Management
May 6, 2022

Signature of Author: _____
Department of Supply Chain Management
May 6, 2022

Certified by: _____
David Correll
Research Scientist and Lecturer, Supply Chain Management Residential Program
Capstone Advisor

Certified by: _____
Mehdi Farahani
Postdoctoral Associate
Capstone Co-Advisor

Accepted by:_____
Prof. Yossi Sheffi
Director, Center for Transportation and Logistics
Elisha Gray II Professor of Engineering Systems
Professor, Civil and Environmental Engineering

Using a Gaussian Mixture Model to measure transit time bimodality and its impact on inventory decisions

by

Aravindan Jayantha

and

Didi Dai

Submitted to the Program in Supply Chain Management
on May 6, 2022 in Partial Fulfillment of the
Requirements for the Degree of Master of Applied Science in Supply Chain Management

ABSTRACT

Companies make inventory decisions based on well-established safety stock methodologies. In these methodologies, a key assumption is that transit times are normally distributed. Although previous studies have shown a nonnormality in transit time distributions in ocean freight, it is still unclear how transit time is distributed in land freight and how much less inventory a company could hold if transit time estimates were more accurate. Moreover, while safety stock methodologies are accepted practice, the inputs used in them are sometimes sourced from static and unsophisticated transit timetables. To address these limitations, this study conducted a distribution analysis and hypothesis testing on geolocation data captured by the sponsoring company, project44, a supply chain visibility provider. The analysis revealed differences in day-of-the-week transit time distributions. Using a Gaussian Mixture Model, this research also studied day-of-the-week transit time bimodality in detail. It was found that the majority of the first distribution had low dispersion around the mean and the second distribution grouped all long-tail transit times, with typically higher standard deviations as a result. This trend is particularly strong in intrastate full truck load shipping. Furthermore, Monday and Tuesday transit times show lower spread in means and have less variation across transit times. In contrast, the rest of the week has considerably higher spread in transit time distributions. This study shows that the full truck load freight is bimodal. Companies accounting for day of the week and transit time bimodality could reduce safety stock and therefore lower inventory cost by up to16% through forward planning and making orders earlier in the week.

Capstone Advisor: David Correll
Title: Research Scientist and Lecturer, Supply Chain Management Residential Program

Capstone Co-Advisor: Mehdi Farahani
Title: Postdoctoral Associate

ACKNOWLEDGMENTS

We would like to thank both our incredible advisors, **Dr. David Correll** and **Dr. Mehdi Farahani,** for their valuable and constructive feedback throughout this capstone project. We are grateful for their time, expertise, and guidance over the past year.

A special thanks to **Pamela Siska**, for her meticulous attention to detail and tremendous support editing our drafts.

To our sponsor, **Christian Piller**, VP of Research and Sustainability from project44, from whom we had the pleasure of learning the invaluable insights that were used to design our model.

To both of our **amazing partners**, Anushka Pai and Victor Du, who supported our education.

Finally, we would like to thank our **classmates**, the Center for Transportation and Logistics, and the entire MIT community for the amazing experience and the friendships that will last a lifetime.

Table of Contents

LIST OF FIGURES *6*

LIST OF TABLES *7*

1 INTRODUCTION *8*

1.1 Motivation and Relevance **8**

1.2 Problem Statement and Key Research Question **9**

1.3 Hypothesis and Model Development **10**

2 LITERATURE REVIEW *11*

2.1 The impact of transit time on lead times and inventory management **11**

2.1.1 Importance of transit times in inventory management 12

2.2 Evaluating and Quantifying transit time variability and distribution assumptions **13**

2.2.1 Transit time with respect to the total cost of inventory 13

2.2.2 Evaluation of transit time distributions and their underlying assumptions 15

2.3 Impacts of transit time variation on transportation cost **17**

2.4 Summary **18**

3 DATA AND METHODOLOGY *19*

3.1 Data set **20**

3.1.1 Data Cleaning and Simulation 22

3.2 Analysis **22**

3.2.1 Historical lead time performance 23

3.2.2 Baseline Leadtime distributions for each lane 24

3.2.3 Distribution Analysis 25

3.3 Hypothesis testing and Tukey test **26**

3.4 Mixture distributions and Understanding Bimodality impacts **27**

3.4.1 Gaussian Mixture Model (GMM) - Unsupervised learning to identify bimodal clusters 29

3.4.2 Examining Bimodality with the separation of means 30

3.4.3 Defining Bimodality with a separation factor 30

3.5 Safety stock and reorder point using bimodal distributions **31**

4 RESULTS AND ANALYSIS *33*

4.1 Baseline Transit times – Data formation & Distribution Analysis **33**

4.2 Hypothesis testing - Tukey test **36**

4.3 Day of the week bimodality – GMM Model & Separation Factor **39**

4.3.1 Distribution Splits using GMM 39

4.3.2 Transit Time Distribution Spread - Difference in Means statistic 46

4.3.3 Defining Bimodality using a separation factor 46

4.3.4 Interstate, Intrastate Lane and State Bimodality: 48
4.3.5 Bimodality by day of the week: 50
4.4 Numerical Experiment - Bimodality Impact measurement 50
5 DISCUSSION AND CONCLUSION 54
REFERENCES 57
APPENDIX A 59

LIST OF FIGURES

Figure 1 15
Example of normal assumption incorrectly increasing purchasing 15
Figure 2 20
Methodology Process Map 20
Figure 3 23
Scatter plot for full truckload trips outbound from Chicago 23
Figure 4 26
Distribution plot of FL transit time distribution from Chicago, IL to Brooklyn, NY 26
Figure 5 29
Bimodality plot within a mixture distribution 29
Figure 6 35
Transit time distribution by day of the week for 6 sampled zip code pairs 35
Figure 7 36
Coefficient of variation distribution for transits times 36
Figure 8 39
Map visualization for areas with dense traffics and high Difference ratio 39
Figure 9 40
Historical distribution of 3-digit zip code 600 to 181 Monday transit times and GMM distributions 40
Figure 10 42
GMM alpha allocation between interstate and intrastate 42
Figure 11 43
Intrastate/Interstate standard deviation spread by GMM distribution type 43
Figure 12 44
Day of the week standard deviation spread by GMM distribution type 44
Figure 13 45
Day of the week standard deviation spread by GMM distribution type 45
Figure 14 46
Distribution of the difference of means between the two GMM clusters 46
Figure 15 47
Distribution of the difference of means between the two GMM clusters 47
Figure 16 48
Intrastate & Interstate bimodality by Lanes & trip share 48
Figure 17 49
Top 5 Intrastate & Interstate bimodality share by total trip volume 49
Figure 18 49
Top 5 Intrastate & Interstate bimodality share by total trip volume 49
Figure 19 50
Share of Bimodal trips by day of the week 50
Figure 20 53
Safety stock calculation of bimodal distributions for each weekday 53

LIST OF TABLES

Table 1 20
Attributes and definitions in the first dataset provided by project44 20
Table 2 21
Attributes and definitions in the next six datasets provided by project44 21
Table 3 24
Coefficient Variation Table for 3-code zip codes, inbound and outbound from Chicago 24
Table 4 34
Aggregated duration average, standard deviation, and Coefficient of Variation Table for top ten zip code pairs ranked by QTY 34
Table 5 37
P-value heatmap for top ten frequent routes 37
Table 6 38
Difference ratio versus coefficient of variation 38
Table 7 41
Table of GMM distribution clusters 41
Table 8 52
Safety stock calculation of bimodal distributions for each weekday 52
Table 9 52
Holding cost calculation by day of the week for bimodal distribution 52

1 INTRODUCTION

In June 2021, project44 raised $202 million and became a “unicorn”, a privately held startup valued at over $1 billion (“project44 Turns Unicorn after Raising $202M Series E Funding from Goldman Sachs, Emergence Capital,” 2021). Based in Chicago, project44, a supply chain visibility provider, has connections with 850+ ELD and telematics devices and tracks shipments in 165 countries. This big telematic network allows project44 to provide customers with real-time supply chain visibility platforms. Over 600 global customers use project44. With high-fidelity data, project44 monitors all modes in all geographies to estimate time arrival, order and inventory visibility, and condition of the goods. (Project44 | About Project44 - Supply Chain Visibility Leader, n.d., p. 44) Our research paper addressed the opportunity that project44’s customers have to use granular transit visibility data to better inform inventory management decisions.

1.1 **Motivation and Relevance**

To make inventory decisions, shippers consider both the demand for the product and the lead time the product takes to get to the shipper. Decision makers for shippers cannot consider one aspect in isolation to achieve the service levels they need while minimizing the cost to hold the inventory. A variety of different inventory policies may be used by the shipper, with various levels of management priority. Many organizations may use methodologies such as Economic Order Quantity (EOQ) or more sophisticated models, to ensure that they order the optimal order quantity. However, the optimal order quantity alone does not determine inventory cost. Therefore, we must also consider other factors such as lead time variation as part of our analysis.

A key challenge to making inventory decisions is the significant variability in lead time that, in practice, a customer may experience when ordering from a supplier. The lead time can be made up of multiple factors, such as manufacturing time, procurement time, and transit time. Shippers frequently calculate the amount of inventory based on the safety stock equations, in which two inputs are estimated:

lead time to delivery and estimated variability around the lead time estimate (Jacobs, F. R., and Chase, R. B., 2016). The calculation is used to make important inventory decisions for the company. When demand is known, inaccurate prediction of lead time and estimated variance could lead to poor inventory decisions and be reflected in the holding cost.

Of particular interest to customers of project44 is the measurement of transit times and how to better monitor their goods in transit. The transit time is the time needed to move goods from one location to another. Jacobs and Chase's (2016) safety stock equation shows that an accurate and dynamic transit timetable will increase the confidence level of lead time variation. Theoretically, transit time duration and lead time variation could both impact inventory level. Based on the estimated transit time, shippers would make appropriate inventory decisions to minimize out-of-stock risks and meet predefined service level. Therefore, accurate transit time data allow shippers to make better inventory decisions and reduce inventory costs.

project44's real-time shipment geolocation data capture allows shippers to monitor their transit times, lead time variation, and evaluate potential inventory risks due to transit time variance. The company developed sophisticated artificial intelligence and machine learning to improve estimated arrival and to manage pre-transit and in-transit exceptions earlier in the shipment lifecycle, such that their customers can take actions when knowing ahead of time there is a transit delay.

1.2 **Problem Statement and Key Research Question**

The sponsor company's current transit timetables are often inaccurate because the tables are not intelligent enough to reflect correct transit time variance. The inaccuracy results from using static values that are rarely updated, with limited traceability to legacy transit times. To address this problem, the key research question of this capstone is how to identify and measure the impact that a more sophisticated model for calculating transit times would have on improved inventory positions.

Initial investigation revealed that a few key features, such as purchasing order release date and shipper mode, were missing from the current transit timetable. For example, the time of the day and the day of the week when the purchasing order is released may affect the shipment time. An order placed and arranged for shipping during peak traffic hours would have to a longer transit time. Shipper mode is another important feature that has a substantial impact on transit time. Less than truckload (LTL) takes more time than Full Truckload (FTL), which directly serves the origin and destination without additional stops for consolidation (Vega et al., 2021). We narrowed our scope to Full Truckload transit times to isolate the variation resulting from additional stops in less than truckload networks.

Our research examined transit time distributions in land freight and asked how much less inventory a company could hold if the transit time estimate were more accurate. To address the factors driving variation in transit time, our research analyzed the level of transit time bimodality and investigated whether the transit times were different across days of the week. We sought to challenge the normal distribution assumptions through an evaluation of bimodality and its impact on safety stock. Our findings answered the question of how much excess inventory companies hold by using out of date and static transit timetables. Companies could use a sophisticated transit time model, proposed by our research, to reduce inventory costs while keeping the same service level.

1.3 **Hypothesis and Model Selection**

We hypothesized that transit times for the days of the week were not equivalent. We measured similarity between transit times for each day of the week per zip code pair. Our results confirmed that days of the week have unique transit time distributions.

A key assumption made in estimating transit times is that the value follows a normal distribution (Das et al., 2014). Here we relaxed this assumption and examined bi-modal distributions as more reflective of the unique nature of each transit time distribution. We used a Gaussian Mixture model to decouple the normal distribution to a bi-modal distribution, which we used to create a more accurate transit timetable. The improvement in transit time accuracy generated lower safety stock level and reduced holding cost.

2 LITERATURE REVIEW

Many researchers have investigated the value of time in the freight industry. Although some papers focus on the shape of transit time distribution and the impact on inventory management, other papers assess how the variance of transit time impacts marginal utility and marginal cost. Considering project44's customer challenge, to create a more comprehensive transit timetable to improve inventory holdings, this chapter will review past research and build upon the existing knowledge to examine factors that will constitute a more accurate transit timetable.

The review is divided into three sections. The first section explains the area of study for this capstone, which introduces the importance of transit time as a component of lead time and how transit times impact the level of inventory shippers' reorder. The second section evaluates sources of variability in transit times. Finally, the third section reviews research on the distribution of transit times and models investigating the cost impact of underlying assumptions.

2.1 **The impact of transit time on lead times and inventory management**

The lead time in freight refers to the time it takes from placing an order to the time the order is completed. In detail, transportation lead time includes travel time and other logistical operation time between loading goods from origin and unloading goods at destination (Massiani, 2008). Of all the components of lead time, this capstone will focus on transit time, for which project44 tracks telematic data and evaluates the features that can improve a shipper's confidence in transit times and thus improve inventory decision making.

Transit time has two major components: speed and reliability. Speed is measured by mean lead time, and reliability is measured by variance of lead time. Both mean and variance will affect inventory decisions. With discrete and continuous transit time distributions, total logistics cost decreases when mean transit time is shortened and variance of transit time decreases (Allen et al., 1985). With a thorough

understanding of the benefit from reducing variance in lead time, shippers and carriers can make intelligent rate negotiations and draft service improvement proposals.

2.1.1 Importance of transit times in inventory management

The management of inventory can be decoupled into two core segments: cycle stock and safety stock. Cycle stock can be considered the inventory that meets the planned estimated demand and safety stock the inventory level required to meet unplanned demand. This capstone will determine the effects of lead time and lead time variability to meet safety stock inventory levels. A classic formulation to determine the amount to reorder under a normal distribution of uncertainty is the Hadley-Whitin reorder formulation (Hadley and Whitin, 1963).

$$R = \overline{D}\,\overline{L} + Z\sqrt{\overline{L}\sigma_D^2 + \overline{D}^2\sigma_L^2}$$

where

$R = Reorder\ point$

$\overline{D} = E[D] = Expected\ demand\ per\ day$

$\overline{L} = E[L] = Expected\ lead\ time$

$Z = z - score\ of\ a\ normal\ distribution\ for\ required\ service\ level$

$\sigma_D = Standard\ Deviation\ of\ Demand$

$\sigma_L = Standard\ Deviation\ of\ Lead\ Time$

The Hadley-Whitin formulation can be further reduced to:

$$R = E[X_{DOLT}] + Z\sigma_{DOLT}$$

where

$Expected\ Demand\ over\ lead\ time$: $E[X_{DOLT}] = \overline{D}\,\overline{L}$

$Standard\ deviation\ of\ demand\ over\ lead\ time$: $\sigma_{DOLT} = \sqrt{\overline{L}\sigma_D^2 + \overline{D}^2\sigma_L^2}$

This capstone investigates the factors that can improve estimates of the demand variation over lead time (σ_{DOLT}). A key assumption under the Hadley-Whitin formulation is that the demand over the lead time is approximately a Normal Gaussian distribution. (Constable and Whybark, 1978; Das, 2013) This capstone explores whether a normality assumption is sound for project44's underlying data, investigate alternative distributions and develop an approach to improve transit time estimations.

2.2 Evaluating and Quantifying transit time variability and distribution assumptions

2.2.1 Transit time with respect to the total cost of inventory

In discussing the need to study transit time, Dehayes (1968) found that the cost of material in transit will add to a certain percentage of economic cost of the goods for those with capital values of assets tied up in the transportation system. Transportation time is a significant factor in financing, and it is one of the key determinants of the efficiency of the distribution system. Goods must be delivered to users promptly and reliably to reduce inventory.

Allen et al. (1985) developed two models to represent transit time distribution and presented change in transit cost with respective to mean and variance of transit time distributions in the matrix. First, the paper analyzes transit time in the discrete probability distribution model. The total cost is calculated in the formula below:

$$TC = (QVW)/2 + (A\bar{R})/Q + (\bar{L}\bar{R}VY)/360 + T\bar{R} + eVW + (gk\bar{R})/Q$$

where

$\bar{R}$ = yearly expected demand = 360 $\bar{r}$ (units) (assume 360 day year)
$\bar{r}$ = daily expected demand (units)
A = ordering cost/order (dollars)
V = value per unit of product (dollars)
Q = the ''economic order quantity'', EOQ, (units)
W = yearly carrying cost/item in inventory (percentage)
e = expected excess in inventory per cycle (units)
g = expected shortage of goods per cycle (units)
k = stockout costs/per unit of goods short (dollars)
$\bar{L}$ = expected time in transit (days)
Y = yearly carrying cost/item in transit (percentage)
T = transportation cost/item (dollars)
Var [r] = variance of daily demand (units)
Var [L] = variance of time in transit (days)

This model constitutes 5 terms: inventory carrying cost, ordering cost, in-transit carrying cost, yearly transportation cost, excess cost and shortage cost due to variance in transit and demand. The length of transit time is directly projected in the In-transit carrying cost. If firms pay for the goods in transit, the longer lead time, the more cost will incur. Meanwhile, variance of transit time is indirectly reflected in the cost of excess inventory and cost of stockout. Reorder point is the replenished level, calculated by average demand during lead time. Demand varies depending on mean lead time and lead time variance. When the reorder point is compared to demand, if inventory is less than demand, there is a stockout cost. In contrast, if the reorder point exceeds the demand, there is an excess cost. As a result, a reliable transit time distribution will condense the range of reorder points and demands during lead time, which will in turn lower shortage cost and excess cost.

Second, Allen et al. (1995) evaluated transit time in the continuous probability distribution model, where daily demand and lead time distributions were assumed to be normal. The shortage cost is represented by multiplying the shortage cost and the expected number of shortages. In the case of a continuous distribution, the expected number of shortages is an integral function of the difference between EOQ and reorder point. The total cost formula shown below is analogous to the discrete total cost formula as shown above, but with VW- mean Lr) as inventory, s > mean Lr analogous to excess cost and mean an (s) analogous to shortage cost.

$$Z = (A\bar{R}/Q) + VW[(Q/2) + s - \overline{Lr}] + [k\bar{a}(s)\bar{R}/Q]$$

Allen et al. (1985) found that a lower mean delivery time and greater reliability was the most desired outcome, with most of the saving stems from improvement in reliability, small amount of saving stems from improvement in mean lead time.

Both the Value of Time (VOT) and the Value of Reliability (VOR) are important estimator metrics that provide a uniform value outcome due to lead time variability (Dullaert and Zamparini, 2013). VOR measures the willingness to pay to reduce the variability of travel times, and such a key time conversion

factor into a financial metric. In a paper investigating the impact of lead time reliability on inventory costs across a pool of models, Dullaert and Zamparini, 2013 identified that by evaluating the heterogeneity of VOR estimates, they concluded reduced lead time does not necessarily reduce cost and can cause increased safety stock costs depending on the demand distribution during lead time. This finding is an important outcome in providing a counterintuitive example that simply focusing only on reduction of transit time variability does not always yield intended results. Our capstone will examine proposed transit time models under a range of various demand distributions to ensure the scope of application is also clearly defined.

2.2.2 Evaluation of transit time distributions and their underlying assumptions

As discussed in 2.1.1, a key feature of the Hadley-Whitin reorder amount is that the demand over the lead time follows a normal distribution. Das's paper examined transit time distributions within ocean transportation; exploring features describing the shape of the distributions and inventory outcomes when normality is not the primary distribution. In the examined dataset, Das (2013) found that although 17% of total lanes were not normal distributions, these lanes accounted for 80% of shipment volumes.

Das (2013) used the toy example shown in Figure 1 to highlight the impact that normality of lead time distribution has on incorrectly suggesting too ordering excess inventory.

Figure 1

Example of normal assumption incorrectly increasing purchasing

If we assume demand is normally distributed ~N (100,10) units/day and examine the following 2 cases:

Case 1: Transit time has a 4-day deterministic rate

Case 2: Transit time is stochastic and is 2 days 50% of the time and 4 days the other 50% of the time. Lead time standard deviation is 1.

Applying the above cases to the reduced Hadley-Whitin formula: $R = E[X_{DOLT}] + Z\sigma_{DOLT}$

We find in a 95% service level case:

Case1: $R = (100 \times 4) + 1.645 \times \left(\sqrt{10^2 \times 4}\right) = 400 + 1.645 \times 20 = 433\ units$

Case2: $R = (100 \times (2+4) \times 50\%) + 1.645 \times \left(\sqrt{3 \times 10^2 + 100^2 \times 1}\right) = 300 + 1.645 \times 101.5 = 467\ units$

Note. Assuming normality in transit increases purchasing units. From "Transportation Research Record: Journal of the Transportation Research Board," by L. Das, B. Kalkanci and C. Caplice, 2014, Impact of Bimodal and Lognormal Distributions in Ocean Transportation Transit Time on Logistics Costs, 2409(1), 63–73.

The example highlights that although the probability of being out of stock was 5% when the transit time exhibited stochastic nature in case 2, the underlying lead time of both times was equal to or less than the 4 days in the deterministic case 1. As a result, an incorrect over-purchased of 34 units (approx. +8% would have occurred in case 2, simply due to the normality assumed in the lead time distribution. To evaluate the impact of non-normal distributions, Das simulated bimodal lead times and evaluated outcomes using a range of critical ratios. A critical ratio is defined as the ratio of the cost of understock to the cost of overstocking. Discovered that in cases where the critical ratio was above 0.7, i.e., high cost of understocking, assumptions of normality expressed a noticeable impact to inventory outcomes versus simulated bimodal ones.

Looking to a larger set of continuous distributions, Tadikamalla's (1984) paper compared normal, logistic, lognormal, gamma, and Weibull distributions as estimates of lead time distributions. Tadikamalla (1984) found that when the coefficient of variation was large, the normal distribution was not an appropriate approximation to measure lead time variation. When a demand and lead time distributions were known, the normal distribution would be the most appropriate approximation. However, when demand or lead time distributions were not known, a lognormal, gamma or Weibull distributions would be suitable estimates due to their versatility (Tadikamalla, 1984). In an examination of the robustness of normal approximations of lead-time demand, the normal approximation is the most appropriate under continuous review systems in general business settings (Tyworth and O'Neill, 1997). Echoing Das's finding, Tyworth and O'Neill found the normal approximation of lead time variation becomes more robust at service levels under 80%.

To evaluate the underlying distribution, Lau and Zaki (1982) examined normal distribution features of kurtosis and skewness within a reorder-point/order-quantity (r,Q) policy - whereby the same quantity is

ordered at variable time intervals. Kurtosis, a fourth central moment, references a measure of how large the tails are within a probability distribution, where the larger the tail the higher the variation there is in the underlying data. Skewness, a third central moment, references the density of the distribution being off-center, from left or right of the mean. Lau and Zaki were able to identify that the change in lead time distribution's skewness also changed the kurtosis and vice versa – leading to compounding effects in the required safety stock levels (Lau and Zaki, 1982). Lau and Zaki's findings are important attributes to apply in this research paper to ensure 3rd and 4th central moments are extra features to examine within Project44's data.

2.3 Impacts of transit time variation on transportation cost

An important consideration in this research is assuming that the given transit time and associated distribution is the only option for a particular lane. However, there are ways to improve transit time with a tradeoff in transport cost. As a result, our research should also ensure that the trade-off of transferring the inventory savings benefit onto increased transportation costs is captured, particularly if an improved transit time estimates comes at an outsized cost.

Tyworth and Zeng (1998) applied a non-negative discrete distribution to model transit time and developed an enhanced sensitivity-analysis tool for examining the effects of carrier transit time on both cost and service. Total annual logistic cost is a sum of transportation cost, ordering cost, holding cost and shortage cost:

$$ETAC(s,Q) = F \cdot R\frac{w}{100} + \mu_T\mu_D \cdot V \cdot Y + \left(\frac{Q}{2} + s - \mu_X\right)V \cdot W + A\frac{R}{Q} + ES \cdot B_2 V\frac{R}{Q}$$

Holding cost is constituted by the cost of cycle stock and safety stock. The estimated shortage is calculated using the distribution of demand and lead time. The lead time was defined as transit time and a fixed component '*Y*' represented the time spent on non-transportation activities in the order cycle (Tyworth & Zeng, 1998). To estimate the effects of transit-time performance on total logistics costs, Tyworth and

Zeng (1998) changed the transit-time mean and variance parameters and recalibrated the order quantity decision variables. The findings of Das (2013), Tyworth and Zeng (1998) found the increase of variance in transit time resulted in significantly higher cost in the underlying model. If there is a guaranteed on-time delivery (i.e., the coefficient of variance T= 0), the cost would decrease, and the service level increases. Lastly, examining for holding variance constant and decreasing transit time led to a limited change in total cost and service level. Having a better estimate that is dynamic, regardless of what the transit time is, the reduction in variance will lead to cost reduction.

Distance is one most important variable to model transit time distribution – particularly with respect to the telematics data that Project44 captures. In an investigation of density of transit models, Chiang and Robert argued that neither city size nor inter-city density should be considered as variables (Chiang & O. Roberts, 1980). With an underlying hypothesis that larger cities might have better service and thus lower transit time – Chiang and Roberts rejected this hypothesis by observing no clear pattern from comparing transit time from cities in different sizes. In addition, the hypothesis of lower transit time in dense markets is also rejected by proving dense markets have more carriers and thus less transit time. Revalidating this finding with respect to urban and regional distances across project44's telematic data will help ensure that the features of our recommended model will directly influence the lead time estimates.

2.4 Summary

The examination of transit-time distributions, normality assumptions and the trade-off of transportation cost savings over inventory costs are primary factors that have had significant research. While there are studies on the impact of demand distribution during transit time on inventory management decisions, there is a gap in the research on descriptive factors that impact transit times, such as day of the week bimodality. By developing an advanced model for transit times and simulating it against a range of outcomes, our capstone provides a new approach to a problem that many companies face.

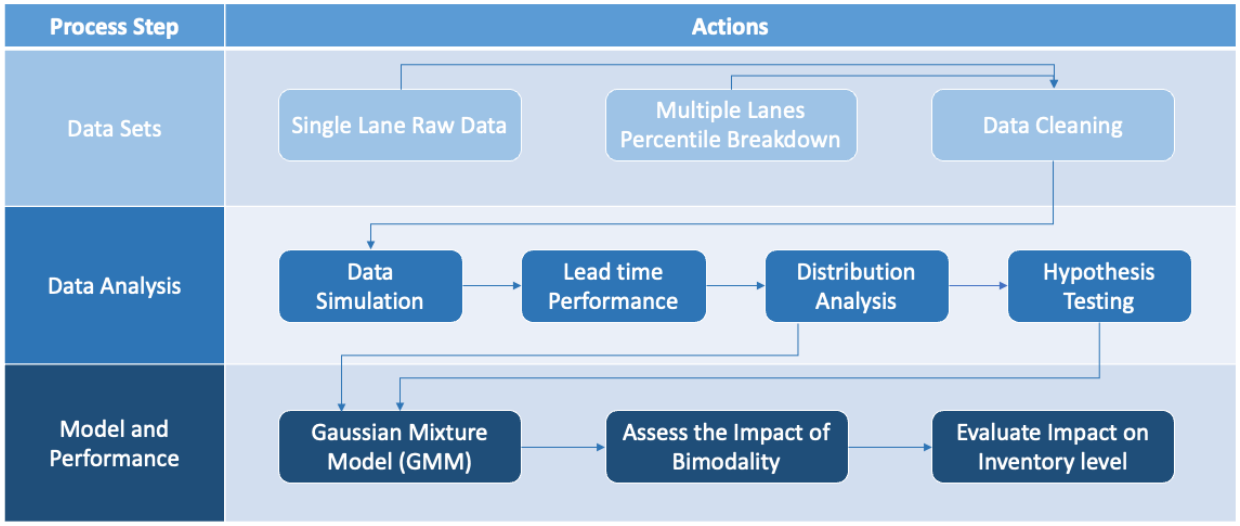
3 DATA AND METHODOLOGY

project44 understands the impact of lead time variation has on inventory management. A key aim of this capstone is to evaluate the underlying assumptions in transit time variability that impact lead-time estimates and subsequently inventory reordering costs. Based on the Hadley-Whitin reorder function mentioned in 2.3.2, we kept all elements but lead time constant. Long lead times and low lead time variation resulted in high pipeline inventory and low safety stock. Inversely, high variability and shorter lead times translated into higher safety stock levels to account for the uncertainty. We conducted a statistical evaluation of the factors impacting transit time and tested the validity of lead time assumptions within the Hadley-Whitin formulation. We also valuated historical lead time performance and calculated the coefficient of variation. For further analysis, we simulated data points based on transit time percentiles provided by the sponsor company and conducted distribution analysis and hypothesis testing to assess the modality of transit time distribution. In addition, we created an unsupervised machine learning model, Gaussian Mixture Model (GMM), to separate transit time distribution by clustering simulated data points. GMM produced the mean and sigma values for each Gaussian model, which are used to calculate separation scores and assess the impact of bimodality. As a final step, we evaluated the performance of the GMM under simulated business cases by using the Hadley-Whitin formulation to compare safety stock level and pipeline inventory outcomes. Recommendations on the most appropriate use cases of the model are made accordingly.

Our research methodology followed a three-phased approach: cleaning and validating our data sets, analyzing the data to understand transit time distributions and test our hypothesis, and finally measuring bimodality and assessing its impact on inventory levels (Figure 2).

Figure 2

Methodology Process Map



Note. This figure shows the overall methodology process flow of this paper.

3.1 Data set

Project44 provided seven datasets. In the first dataset, there are 8 tabs, of which four tabs include Full Truckload (FTL) and Less than Truckload (LTL) transit time data from/to Chicago and another four tabs include FTL and LTL data from/to Boston. A list of attributes, and corresponding data type and description for each attribute, is provided in Table 1.

Table 1

Attributes and definitions in the first dataset provided by project44

Index	Attribute	Data Type	Description
1	Destination Postal	Integer	Destination zip code
2	Qty	Integer	Total number of trips from the same origin to destination
3	AVG Duration	Floats	Average lead time in hours for one trip
4	MED Duration	Floats	Median lead time in hours for one trip
5	STDDEV Duration	Floats	Lead time standard deviation for one trip
6	Estimated Time	Integer	Estimated number of days for one trip
7	Actual Day	Floats	Actual number of days for one trip

Note. This table listed attributes for all FTL routes inbound and outbound Chicago, IL and Boston, MA.

The rest of the six datasets contains aggregated FTL and LTL transit time for eight metro areas: Brooklyn, NY, Chicago, IL, Bell Garden, CA, Aiken, SC, Palestine, TX, Bonny Lake, WA, Lakeland, FL, and Denver, CO. Transit times are aggregated by day and month. The granular data includes trips from origin to destination in a one-year horizon. Note that origin and destination zip codes only display the first three digits for confidential concerns from the company. Details for data covered in these datasets are listed in Table 2.

Table 2

Attributes and definitions in the next six datasets provided by project44

Index	Attribute	Data Type	Description
1	Origin Zip	Integer	Origin zip code
2	Origin Metro	Object	Name of the origin metro city
3	Destination Zip	Integer	Destination zip code
4	Destination Metro	Object	Name of the destination metro city
5	Day	Object	Day of the week
6	Month	Integer	Month of the year
7	QTY	Integer	Total number of transits within a year
8	Duration Average	Floats	Average transit time of a single trip
9	Duration STDDEV	Floats	Standard Deviation transit time of a single trip
10	Duration Min	Floats	Minimal transit time of all trips
11	Duration Max	Floats	Maximum transit time of all trips
12	Duration P10	Floats	10th percentile of the transit time from all transit times
13	Duration P20	Floats	20th percentile of the transit time from all transit times
14	Duration P30	Floats	30th percentile of the transit time from all transit times
15	Duration P40	Floats	40th percentile of the transit time from all transit times
16	Duration P50	Floats	50th percentile of the transit time from all transit times
17	Duration P60	Floats	60th percentile of the transit time from all transit times
18	Duration P70	Floats	70th percentile of the transit time from all transit times
19	Duration P80	Floats	80th percentile of the transit time from all transit times
20	Duration P90	Floats	90th percentile of the transit time from all transit times

Note. This table listed attributes for aggregated transit time history for eight metro areas, including Brooklyn, NY, Chicago, IL, Bell Garden, CA, Aiken, SC, Palestine, TX, Bonny Lake, WA, Lakeland, FL, and Denver, CO.

3.1.1 Data Cleaning and Simulation

We cleaned the datasets before studying the attributes and summarizing patterns. The first step was to transfer attributes to the correct data type and the correct units for analysis. The second step was to drop, impute, or fill in missing values with proper values based on each attribute's unique characteristic. The third step was to cluster adjacent zip codes and similar distances from the origin to destination pairs to calculate the coefficient of variation.

For further analysis, we simulated raw data with given percentile values from metro city data files. The percentile durations were split into ten buckets per zip code pair. Each bucket ranges from the previous duration to the current duration values were generated within the range of a bucket. The number of randomly generated values equals to the total quantity of trips made from each zip code for example, from Chicago, IL (zip code starts with 600) to Brooklyn, NY (zip code starts with 181) had 436 trips in total in the given year. Correspondingly, a total number of 436 random transit time values were simulated based on the given percentiles for the 600 and 181 zip code pair.

3.2 Analysis

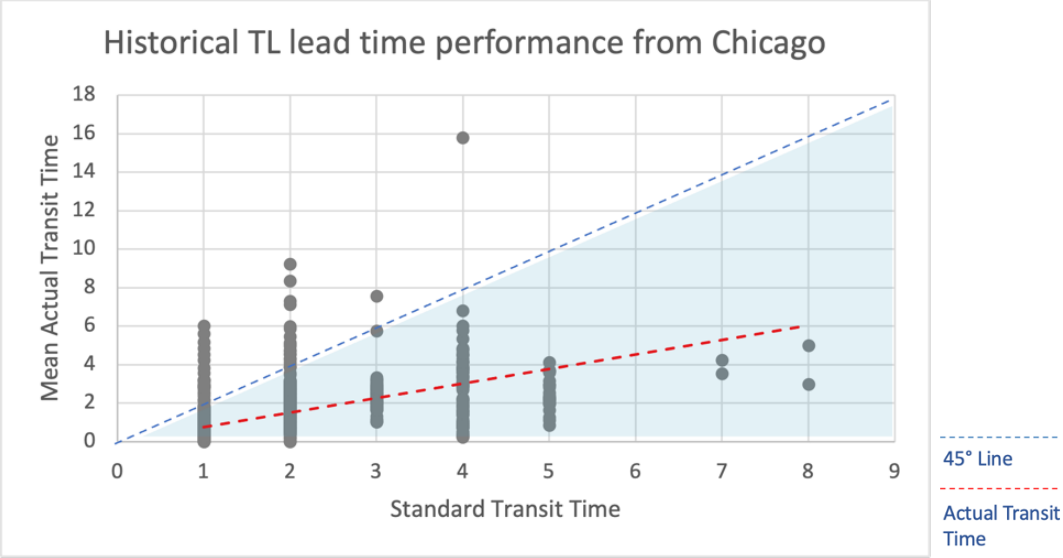
The research problem posed by project44 is to identify factors affecting transit timetable and investigate the impact of transit time variability on inventory cost. To understand lead time variation, we first analyzed historical lead time performance and evaluated coefficient variance for potential factors that may affect transit time over clustered trips. With an understanding of lead time variation in the past, we conducted distribution analysis over past trips and analyzed the distribution of transit time. Hypothesis testing was done after distribution analysis to reject the underlying assumptions. Factors proving to significantly impact transit time variability were further analyzed in the multi-linear regression model.

3.2.1 Historical lead time performance

A scatter plot of lead time variation provides a high-level overview of historical transit time performance. Figure 4 shows the full truckload transit time distribution for all trips made in the last year outbound from Chicago. The 45-degree reference line is the average transit time. Ideally, all data points would converge to the 45-degree line. However, as shown in figure 4, data points are scattered on each side of the diagonal line. Such behavior indicates transit time is not static. The white zone above the 45-degree line is the 'Worse than Contract Zone', meaning that the shipments are delayed and the zone below it is the 'Better than Contract Zone', meaning that the shipments arrived ahead of time. Either case would result in extra logistic costs. Companies will need to coordinate the operation team with the procurement and contract team to update transit time frequently to avoid inaccuracy in transit time.

Figure 3

Scatter plot for full truckload trips outbound from Chicago



Note. This figure shows the lead time variation for full truck load shipments

With an overview of transit time variance in scatter plot in Figure 4, we did further analysis of the magnitude of variation. The coefficient of variation is calculated for transit time with respect to each type when $CV=\sigma/\mu$. Larger CV means more variation and requires more frequent update on transit timetable.

Table 3 shows a sample transit variation which occurs in different payment type and transportation type. Origin to Destination pairs is clustered by region.

Table 3

Coefficient Variation Table for 3-code zip codes, inbound and outbound from Chicago

	Mean transit time (hours)				Standard Deviation				Coefficent of Variation (CV)			
	Inbound		Outbound		Inbound		Outbound		Inbound		Outbound	
3-code zip	FTL	LTL	FTL	LTL	FTL	LTL	FTL	LTL	FTL	LTL	FTL	LTL
605	5.0	13.8	7.4	10.2	7.5	26.8	10.5	22.5	1.5	1.9	1.4	2.2
604	2.4	19.9	7.5	12.3	7.1	36.9	17.6	36.4	3.0	1.9	2.3	3.0
601	6.9	6.5	4.8	21.3	15.0	13.3	2.9	25.3	2.2	2.1	0.6	1.2
600	18.2	24.4	8.7	31.5	30.0	28.4	6.8	38.3	1.6	1.2	0.8	1.2
531	11.9	24.9	13.4	22.7	21.0	18.7	18.9	8.9	1.8	0.8	1.4	0.4
530	15.0	23.0	19.6	35.0	22.7	22.6	19.7	16.0	1.5	1.0	1.0	0.5

Note. This table shows the mean, standard deviation, and coefficient of variation of transit times for sampled zip codes. Coefficient of variation values are color coded for visualization: red means high coefficient of variation and white means low coefficient of variation.

3.2.2 Baseline Leadtime distributions for each lane

To establish a baseline lead time distribution and corresponding data summaries for the percentile data we aggregated daily values into overall lane distributions using weighted means and combined variance calculations. These are as follows:

Combined Arithmetic Mean: $$X_c = \frac{\sum_{i=0}^{j} n_i \overline{X_i}}{\sum_{i=0}^{j} n_i}$$

Combined Variance: $$S_c^2 = \frac{\sum_{i=0}^{j} n_i \left[S_i^2 + (\overline{X_i} - \overline{X_c})^2\right]}{\sum_{i=0}^{j} n_i}$$

where

$n_i = number\ of\ all\ transit\ time\ records\ for\ each\ day\ of\ the\ week\ i$

$\overline{X_i} = mean\ transit\ time\ for\ day\ of\ the\ week\ i$

$\overline{X_C} = Combined\ arithmetic\ mean\ \ transit\ time\ for\ lane\ c\ for\ all\ days\ of\ the\ week\ i$

$S_i = standard\ deviation\ of\ transit\ time\ for\ day\ of\ the\ week\ i$

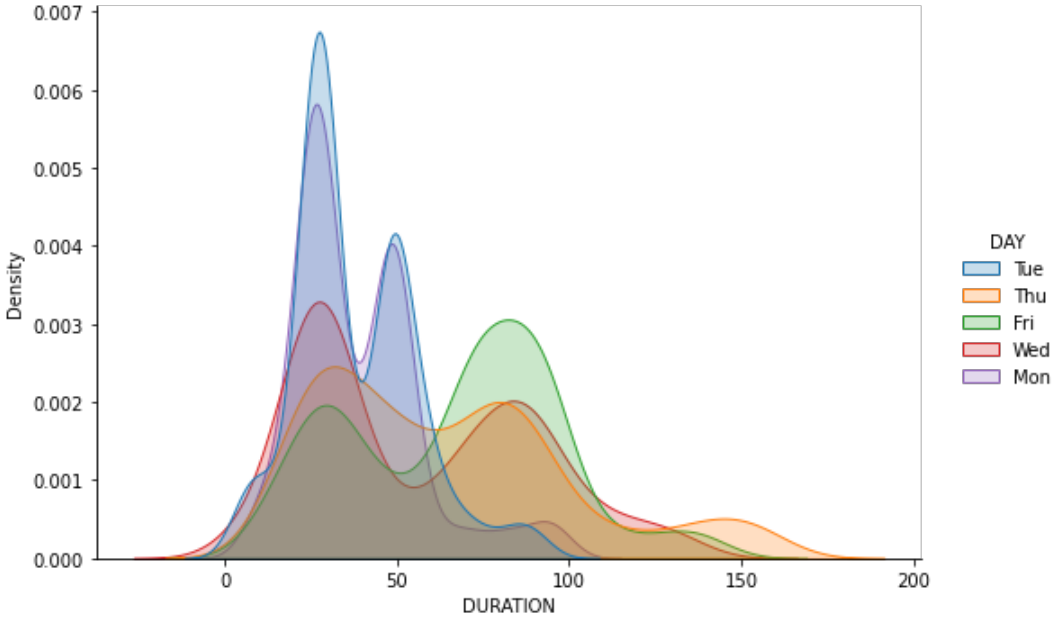
The combined lane transit times and variances defined a mean transit time and variance that are used within the industry transit timetables. These aggregate times and associated standard deviation form a baseline to compare all lane results with and understand the true impact of assuming normality in transit timetables.

3.2.3 Distribution Analysis

This section demonstrates the non-normal distributions in transit time estimates outlined in section 2.3.2, where Das (2014) proved that normality is not the primary distribution for transit time within ocean transportation. With a list of simulated transit times, a histogram was plotted in Figure 6 for distribution analysis, which was performed to understand the average lead time and lead time variation under different circumstances. The distribution plot shows the frequency of transit time from Chicago, IL (zip code starts with 600) to Brooklyn, NY (zip code starts with 181). The distributions are color-coded by day of the week. Saturday and Sunday are missing from the graph because less than 10 trips were initiated on Sunday in the past year. It is obvious to claim a bimodality for the transit time distribution for six days of a week. In Figure 6, Monday and Tuesday have similar distributions, where average transit time μ_1 is close to 28 and μ_2 is close to 51. Compared to Monday and Tuesday, Wednesday, Thursday, and Friday have wider distribution, which indicates larger variance and lower accuracy. Wednesday has lowest mean transit time $\mu_1 = 25$ and $\mu_2 = 70$, among three wider distributions. Thursday and Friday have mean transit time $\mu_1 = 26$ and $\mu_2 = 80$.

Figure 4

Distribution plot of FL transit time distribution from Chicago, IL to Brooklyn, NY



Note. The distribution analysis is created with seaborn plot in python. Chicago, IL zip codes reference start with 600 and Brooklyn, NY zip codes start with 181.

3.3 Hypothesis testing and Tukey test

The next step after the analysis of distributions was to formally verify that chosen features exhibit statistical differentiation from the mean value. The method for doing this was hypothesis testing and a more extensive analysis of variance (Tukey). Hypothesis tests formally examine two mutually exclusive conjectures (hypotheses), H_0 and H_A, and evaluate each against using test statistics. The test statistics used to evaluate the hypothesis can be calculated against non-parametric tests such as the Tukey test. The Tukey test identifies confidence intervals and outputs p values for multiple comparison to indicate the significant difference of paired data points. The results will not prove an alternative hypothesis to be correct but will confirm that a null hypothesis can be rejected. To examine factors that exhibit non-normal distribution of transit times, this section will highlight a broad range of tests within the data set. We constructed the null hypothesis and alternative hypothesis as below:

H0 = Mean transit time for each day of the week pairing is equivalent

HA = Mean transit time for each day of the week pairing is not equivalent

The hypothesis measures similarity between transit times for each day of the week per zip code pair. For example, it compares the transit time between Monday and Tuesday, Monday and Wednesday, Monday and Thursday, and so forth. We performed the Tukey test over simulated data for each zip code pair and recorded p-values in a matrix. Rows of the matrix documents zip code pairs and columns log weekday pairs. The larger the p-value, the weaker the evidence to reject the null hypothesis, meaning that the mean transit time between the weekday pair is equivalent. We set a threshold of 0.05 to evaluate the statistical significance. A p-value lower than 0.05 is statistically significant and indicates strong evidence to reject the null hypothesis.

3.4 Mixture distributions and Understanding Bimodality impacts

Following the results of the hypothesis tests using the Tukey test methodology, the next phase of our methodology was to identify days of the week and months for lanes which had high likelihood for bimodality. As discussed in section 2.3.2, assuming normality in transit time distributions is a hallmark of current inventory management methodology.

Using the Tukey test described in 3.3, we were able to compare all days of the week with one another, to understand the differences in distributions. We set up a ratio for each lane counting how many comparisons expressed significance in their distribution differences across the various days. While the comparison across days formed a benchmark to identify distribution differences across days, our research led us to also identify bimodality within each day of the week. As highlighted in 3.3.3, bimodality can exist within each day's transit time distribution. To examine this within each day, we broke down days the distribution into a mixture distribution which follows the following function of 2 combined normal distributions:

Mixture Distribution: $f(X) = Y = \alpha_1 X_1 + \alpha_2 X_2$

$let\ \alpha = \alpha_1 = 1 - \alpha_2,$

$$f(X) = \alpha X_1 + (1 - a) X_2$$

$$f(X) = \alpha\varphi\left(\frac{X - \mu_1}{\sigma_1}\right) + (1 - a)\varphi\left(\frac{X - \mu_2}{\sigma_2}\right)$$

where

$f(X) = Y = Mixture\ distribution\ of\ all\ transit\ time\ values\ of\ X$

$X_i = Normal\ distribution\ i\ of\ the\ mixture\ distribution\ f(X), for\ i = 1, 2$

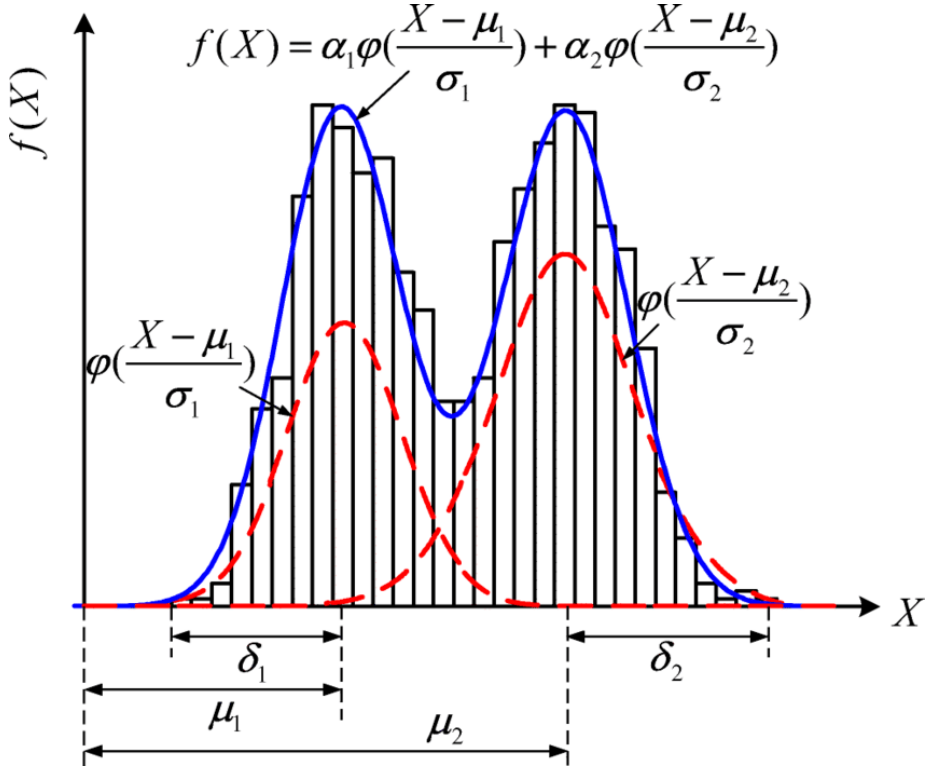
$a_i = weighted\ value\ of\ each\ normal\ distribution\ i\ in\ the\ mixiture\ distribution, for\ i = 1, 2$

$\varphi(z) = \varphi\left(\frac{X - \mu_i}{\sigma_i}\right) = standard\ normal\ distribution\ of\ each\ normal\ distribution\ i, for\ i = 1, 2$

Figure 7 illustrates the mixture distribution as a combination of two normal distributions. By decomposing the daily transit time distribution into a mixture distribution, we will examine not only if they are in fact bimodal transit times, but also which vehicle routes and days are most likely to be bimodal and the overall impact of bimodality is on inventory costs.

Figure 5

Bimodality plot within a mixture distribution



Note. This figure shows the formula to calculate mean and sigma values for bimodal distribution. From "A new uncertainty propagation method considering multimodal probability density functions," by Z. Zhang, J. Wang, C. Jiang, and Z. L. Huang, 2019, *Structural and Multidisciplinary Optimization*, *60*(5), 1983–1999.

3.4.1 Gaussian Mixture Model (GMM) - Unsupervised learning to identify bimodal clusters

To identify the underlying parameters within the mixture distribution, we chose to use a Gaussian Mixture Model (GMM) to cluster the data into two distinct Gaussian distributions. By using an unsupervised machine learning model technique, we were able to group the clusters of transit times into two clear subsets by using the expectation-maximization (EM) algorithm technique within the GMM. Each transit time observation will have a probability of falling within a particular cluster and thus can be assigned to one of two parts of a mixture distribution based on the highest probability. The two distributions that represent the distribution are defined by the clusters, and each cluster's mean and standard deviation is

forming a unique Gaussian distribution shown in 3.4. By using the GMM to cluster the data, our research was able to define the parameters of the mixture distribution and measure the impact when compared to Gaussian only assumptions.

3.4.2 Examining Bimodality with the separation of means

In order to first examine the mean separation of the two distributions found from the GMM, a useful statistic to measure the separation of means relative to their widths we label as D; which is as follows (Muratov and Gnedin, 2010):

$$D = \frac{|\mu_1 - \mu_2|}{[(\sigma_2^2 - \sigma_2^2)/2]^{1/2}}$$

The D statistic represent the relative mean difference, where a D>2 represents a meaningful distribution split. We used the relative separation of means to understand the spread of between the two Gaussian distributions produced by the GMM.

3.4.3 Defining Bimodality with a separation factor

To define if the distribution found is truly bimodal, we explored an approach to calculate a separation factor and evaluate the absolute difference in means of clusters within mixture distributions. Schilling et al. (2002)'s research on human high bimodality defined the generalizable case where $\sigma_1 \neq \sigma_2$ and an unequal mixture of data is present among both X_1 and X_2. It follows that when solving for the setting the mixture distribution to:

$$\alpha f_1'(x) + (1-a) f_1'(x) = 0 \text{ and } \alpha f_1''(x) + (1-a) f_1''(x) > 0\,,$$

Schilling et al. (2002) found that the following could be defined as the separation factor, $S(r)$:

Let $r = \frac{\sigma_1^2}{\sigma_2^2}$, then it follows - $S(r) = \frac{\sqrt{-2+3r+3r^2-2r^3+2(1-r+r^2)^{\frac{3}{2}}}}{\sqrt{r}(1+\sqrt{r})}$

Using the separation factor, we can define a mixture distribution as bimodal if:

$$|\mu_2 - \mu_1| > S(r)(\sigma_1 + \sigma_2)$$

In the case where the inequality is smaller than the absolute difference in means, the distribution would be unimodal. Following the definition, we can identify each daily distribution as either exhibiting bimodality within the day or not. We ranked the data and found key days of the week and lanes which expressed significant deviation and others where underlying normality assumptions were fine to use due to their unimodal classification.

3.5 Safety stock and reorder point using bimodal distributions

The final step in our methodology was to calculate the impact of the advanced transit time model on inventory policy and cost. Inventory cost includes costs for holding cycle stock, safety stock, and pipeline inventory. Lead time variation impacts inventory cost, particularly safety stock cost, as safety stock equals to k multiplied by lead time standard deviation, where k represents cycle service level. To evaluate the impact of the transit time model, we created a numerical experiment that took a demand pattern which followed a normal distribution ~~N (1000, 80) and cycle service level to be 95%. To calculate reorder point for each segment, we used the Hadley-Whitin equation:

$$R = E[X_{DOLT}] + Z\sigma_{DOLT}$$

where

$$E[X_{DOLT}] = \overline{D}\,\overline{L} = Expected\ Demand\ over\ lead\ time$$

$$\sigma_{DOLT} = \sqrt{\overline{L}\sigma_D^2 + \overline{D}^2\sigma_L^2} = standard\ deviation\ of\ demand\ over\ lead\ time$$

Similarly, demand over lead time standard deviation was leveraged to calculate safety stock. Safety stock is calculated by using the formula:

$$SS = K * \sigma_{DL}.$$

Our method focused on comparing the safety stock using normal demand conditions applied to historic transit data evaluate inventory outcomes in using a standard day-agnostic baseline across all days

of the week and a more advanced approach to transit time this research formulated, calculating two safety stock levels for each day of the week. We weighted the two holding costs produced for every day of the week by the alpha value produced by the GMM model, creating a holding cost for each day of the week. Subsequently, we compared the safety stock and holding cost required in both baseline standard approach and each day of the week, to identify opportunities for holding cost savings across the board. Following the numerical experiment, we formed recommendations on the most appropriate use cases of the newly developed mathematical model.

4 RESULTS AND ANALYSIS

4.1 Baseline Transit times – Data formation & Distribution Analysis

As mentioned in section 3.1.1, we cleaned the datasets before studying the attributes and summarizing patterns by simulating raw data with given percentile values from metro city data files. We filtered zip codes with quantities of more than 10 because frequent shipments are more valuable to analyze compared to less frequent shipments. To simulate filtered routes, every route is split into 10 buckets, each with a minimum window value and a maximum window value. For example, the first bucket has minimum duration as minimum window value and 10^{th} percentile duration as maximum window value, the second bucket has 20^{th} percentile duration as minimum window value and 30^{th} percentile duration as maximum window value, etc. With 10 buckets defined, we generated transit time that randomly falls into the 10 buckets with the given quantity per route. For instance, 436 transit times had been recorded in the '8Metros_FTL_by_Day' datasheet for Chicago, IL (zip code starts with 600) to Brooklyn, NY (zip code starts with 181), therefore 436 rows of duration data were simulated for this route. Looping through 1152 unique zip code pairs in '8Metros_FTL_by_Day' datasheet, a total of 145966 rows of data were simulated for further analysis. Please see Appendix A for the simulated data.

With the simulated data, we aggregated daily values into overall route distributions to establish a baseline lead time distribution and corresponding data summaries. We applied the combined arithmetic mean and combined variance formula in section 3.2.2 for duration mean and duration standard deviation. The results were sorted in ascending order by QTY, and the top 10 results are shown in Table 4. We also calculated the coefficient of variation (CV) for each route. From table 5, six out of ten top frequent routes have CV larger than one, meaning a high level of dispersion around the mean. The routes with high CV have large variability in transit time and therefore require more study.

Table 4

Aggregated duration average, standard deviation, and Coefficient of Variation Table for top ten zip code pairs ranked by QTY

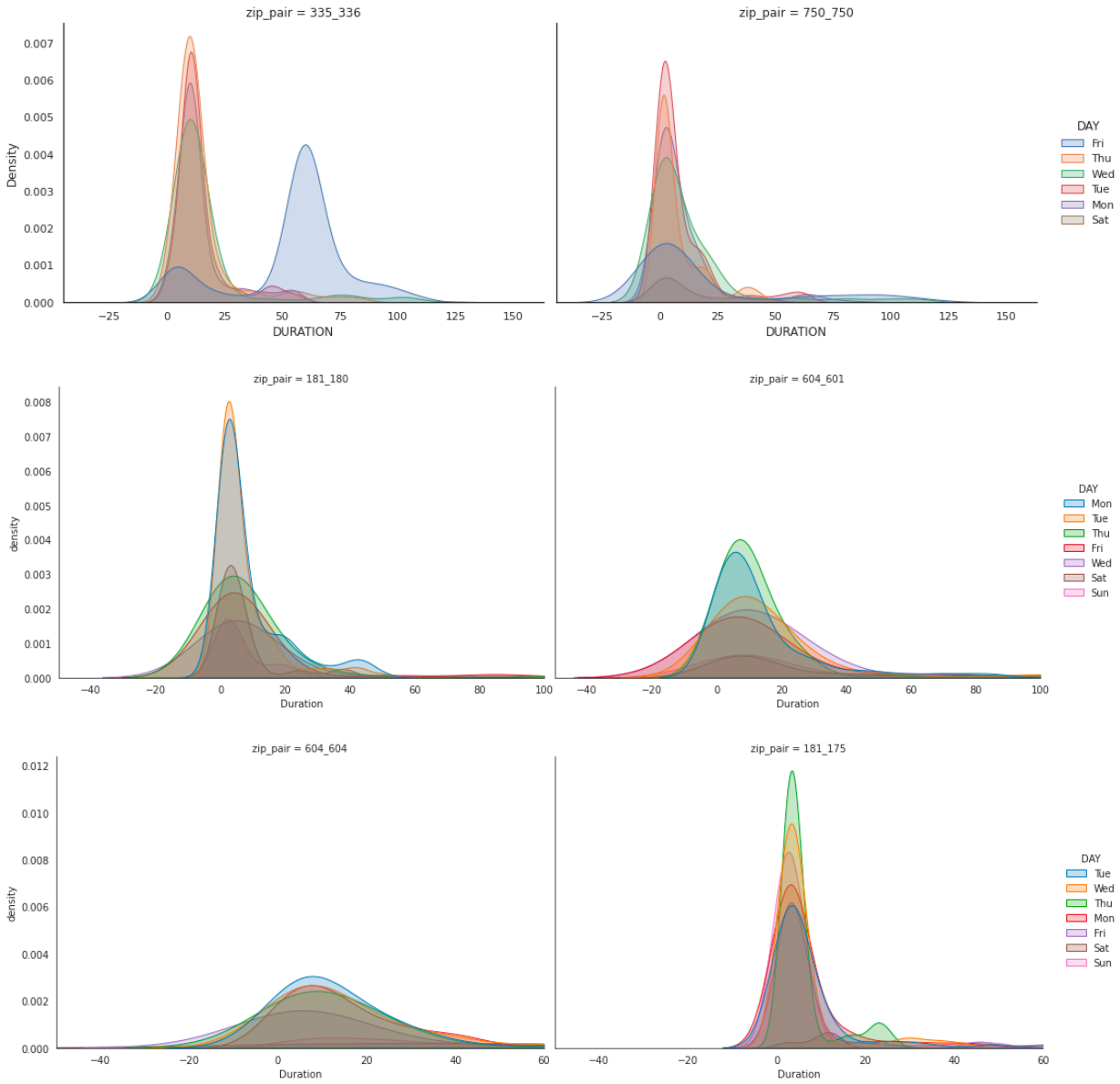
index	ORIGIN - 3ZIP	ORIGIN - METRO	DESTINATION- 3ZIP	DESTINATION- METRO	QTY	DURATION - AVG	DURATION - STDDEV	cv
0	600	Chicago, IL	181	Brooklyn, NY	2118	50.64	27.17	0.54
1	335	Lakeland, FL	336	Lakeland, FL	1738	22.04	24.33	1.1
2	752	Palestine, TX	770	Palestine, TX	1655	27.28	16.42	0.6
3	600	Chicago, IL	750	Palestine, TX	1639	49.66	25.17	0.51
4	750	Palestine, TX	750	Palestine, TX	1453	9.67	13.78	1.42
5	604	Chicago, IL	601	Chicago, IL	1367	16.48	22.22	1.35
6	181	Brooklyn, NY	180	Brooklyn, NY	1354	7.52	12.62	1.68
7	600	Chicago, IL	917	Bell Gardens, CA	1313	87.18	20.49	0.24
8	181	Brooklyn, NY	175	Brooklyn, NY	1257	6.07	8.6	1.42
9	604	Chicago, IL	604	Chicago, IL	1252	22.61	25.32	1.12

Note. This tables shows mean and standard deviation values per zip code pair aggregated by week. Coefficient of variations were derived from the aggregated mean and standard deviation.

We plotted histograms for the top 6 routes with CV larger than one. In Figure 6, the distributions are color-coded by day of the week. Sunday was missing from the first two graphs because less than 10 trips were initiated on Sunday for route 335 to 336 (Lakeland, FL) and route 750 to 750 (Palestine, TX) in the past year. In Figure 6, route 335 to 336 (Lakeland, FL) has a clear bimodality that Friday's transit time distributes separately from the rest of the week; route 750 to 750 (Palestine, TX) has an unnoticeable bimodality on Thursday's distribution; route 181 to 175 (Brooklyn, NY) has a bimodality on Thursday's distribution as well. For the rest of the routes plotted, none of the weekdays have perfectly overlapped mean value and the wideness of distributions are all divergent. We counted the total number of bimodal routes in section 4.3 by using GMM model.

Figure 6

Transit time distribution by day of the week for 6 sampled zip code pairs



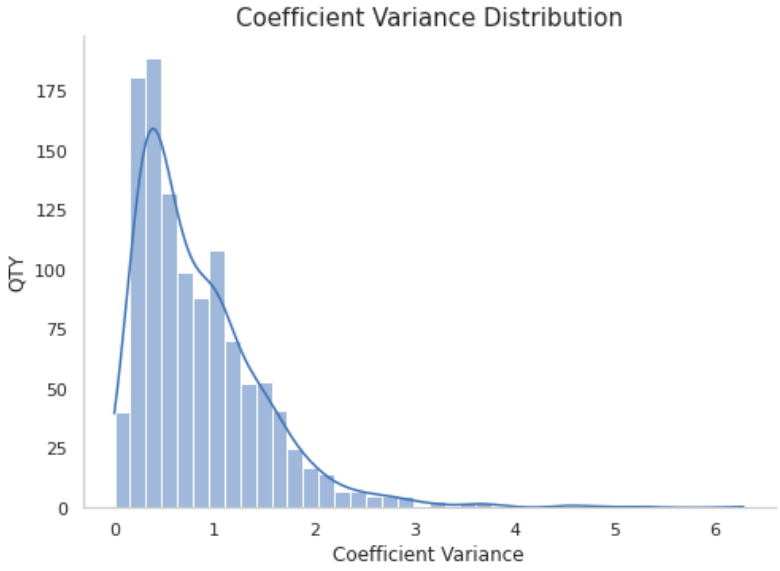
Note. The transit time distribution figures were plotted using seaborn package in python. The figures were graphed for the use of this research to perform distribution analysis.

To examine the number of high CVs in a bigger picture, we plotted a coefficient variance distribution in Figure 7. Out of all routes, 778 are with low CVs and 374 are with high CVs larger than 1.

Routes with high CVs have more variation on transit time and require further analysis to improve accuracy and reduce additional operation cost.

Figure 7

Coefficient of variation distribution for transits times



Note. This figure shows the coefficient of variation distribution for aggregated transit times per route recorded in the eight metro area transit datasets.

4.2 Hypothesis testing - Tukey test

We conducted Tuckey test for extensive analysis of variance to prove the hypothesis established in section 3.3 and formally verified that the day of the week exhibits statistical differentiation from the mean value. In detail, we looped Tuckey test over simulated data to compare all possible pairs of means, except for routes with only one or two days of transit. For example, we compared similarities between Monday's and Tuesday's distributions for route 335 to 336 (Lakeland, FL) and output a p-value. The smaller the p-value, the stronger the evidence that we should reject the null hypothesis, meaning that Monday's and Tuesday's distributions are different. For routes with one weekday transit record, like route 806 to 801 (Denver, CO), Tuckey test will not be able to operate and outputs a coding error.

Table 5 is a heatmap of p-values for top 10 frequent routes sorted in ascending order regarding to the origin zip codes. Dark blue represents low p-values and white indicates high p-values. Due to the nature of zip codes, the codes started counting from northeast, and the more southwest, the higher the zip codes. A large white area in the middle of Table 5 shows that the transit time is far from unimodal in south and mid US, particularly around Lakeland, FL and Chicago, IL. More white numbers on the right-hand side of the table than left means that Friday's transit time is very different from the rest of the weeks.

Table 5

P-value heatmap for top ten frequent routes

zip_pair	Mon_Tue	Mon_Wed	Mon_Thu	Mon_Fri	Tue_Wed	Tue_Thu	Tue_Fri	Wed_Thu	Wed_Fri	Thu_Fri
181_180	0.900000	0.156100	0.016500	0.004800	0.130300	0.013700	0.003900	0.900000	0.900000	0.900000
335_334	0.900000	0.086200	0.218200	0.001000	0.104400	0.182500	0.001000	0.001000	0.001000	0.001000
335_336	0.879400	0.015200	0.900000	0.001000	0.169600	0.900000	0.001000	0.077700	0.001000	0.001000
600_181	0.825900	0.001000	0.001000	0.001000	0.001000	0.001000	0.001000	0.001000	0.001000	0.900000
600_750	0.900000	0.772100	0.001000	0.001000	0.590300	0.001000	0.001000	0.001000	0.001000	0.001000
600_917	0.667900	0.001000	0.001000	0.001000	0.001000	0.001000	0.001000	0.002400	0.001000	0.798200
604_601	0.615800	0.001000	0.505800	0.025500	0.150000	0.900000	0.705400	0.163500	0.900000	0.748600
604_604	0.900000	0.900000	0.900000	0.757000	0.900000	0.900000	0.608900	0.900000	0.900000	0.900000
750_750	0.900000	0.368500	0.900000	0.001000	0.442400	0.900000	0.001000	0.191100	0.061600	0.001000
752_770	0.900000	0.192800	0.469200	0.049100	0.083900	0.695700	0.023700	0.001000	0.833600	0.001000

Note. This table records p-values output from Tukey tests. The values are arranged in the order of zip code pairs and weekday pairs.

Other than top 10 frequent routes, we performed Tukey Test over each unique zip code pair available and produced p-values for every zip code pair. We set a threshold of 0.05 and calculated the ratio of p-values lower than the threshold to the total number of p-values for each route. High ratio means that more days of the week pairing have different transit time and low ratio means less difference in day of the week. Besides, we ranked the routes with ratio over 0.5 in descending order regarding QTY. According to Table 7, Chicago, IL to Brooklyn, NY is the most frequent route and has the highest difference ratio.

Chicago, IL to Palestine, TX and Chicago, IL to Bell Gardens, CA are ranked number two and number three in terms of frequency and difference ratio. All three directions outbound Chicago have a high difference ratio and it makes Chicago the most volatile metro area in terms of transit time. Other than Chicago, Lakeland, FL, Bell Gardens, CA and Palestine, TX also have relatively high transportation frequency and different transit times depending on the day of the week.

Table 6

Difference ratio versus coefficient of variation

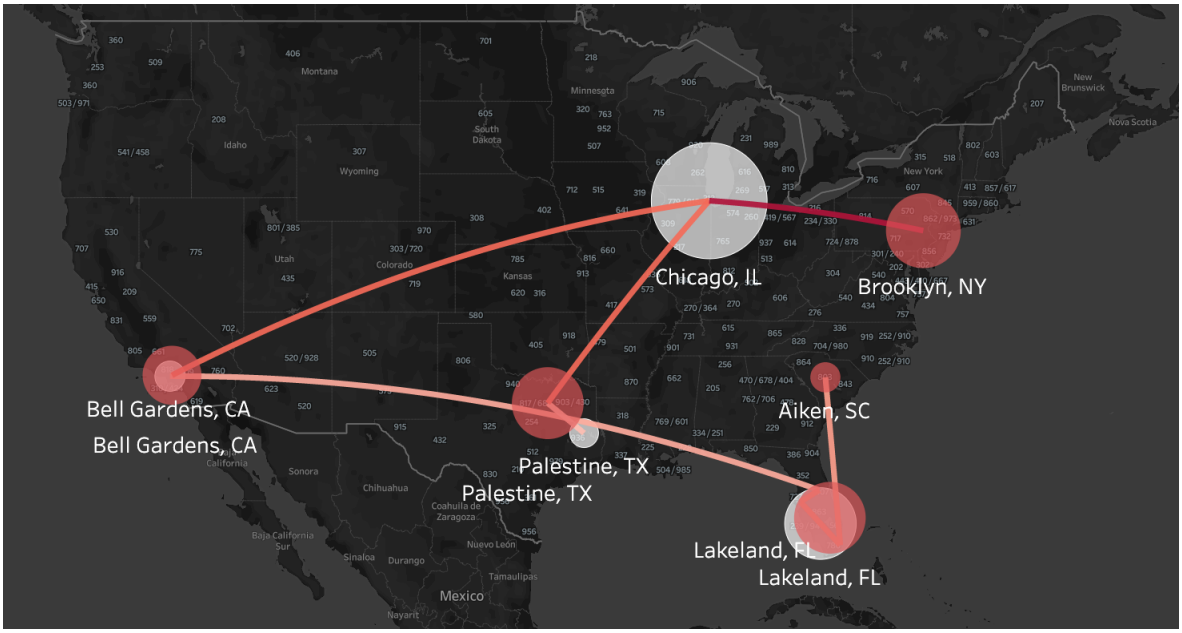
zip_pair	ORIGIN_METRO	DESTINATION_METRO	QTY	CV	Ratio
600_181	Chicago, IL	Brooklyn, NY	2118	0.54	0.90
600_750	Chicago, IL	Palestine, TX	1639	0.51	0.70
600_917	Chicago, IL	Bell Gardens, CA	1313	0.24	0.70
335_328	Lakeland, FL	Lakeland, FL	959	0.94	0.60
335_330	Lakeland, FL	Lakeland, FL	634	0.97	0.60
900_328	Bell Gardens, CA	Lakeland, FL	345	0.45	0.57
330_291	Lakeland, FL	Aiken, SC	343	0.71	0.60
759_752	Palestine, TX	Palestine, TX	311	1.18	0.53
984_982	Bonney Lake, WA	Bonney Lake, WA	277	0.89	0.57
924_923	Bell Gardens, CA	Bell Gardens, CA	229	1.62	0.60
334_297	Lakeland, FL	Aiken, SC	213	0.51	0.83
605_750	Chicago, IL	Palestine, TX	213	0.69	0.60
775_802	Palestine, TX	Denver, CO	204	0.3	0.83
917_760	Bell Gardens, CA	Palestine, TX	199	0.37	0.53
908_917	Bell Gardens, CA	Bell Gardens, CA	191	1.47	0.70
750_601	Palestine, TX	Chicago, IL	177	0.65	0.60
907_902	Bell Gardens, CA	Bell Gardens, CA	176	0.7	0.73
193_174	Brooklyn, NY	Brooklyn, NY	152	1.08	0.60
917_802	Bell Gardens, CA	Denver, CO	151	0.62	0.60
181_071	Brooklyn, NY	Brooklyn, NY	136	1.45	0.60
170_173	Brooklyn, NY	Brooklyn, NY	128	0.82	0.67
752_802	Palestine, TX	Denver, CO	123	0.36	0.70
604_296	Chicago, IL	Aiken, SC	109	0.57	0.60
907_601	Bell Gardens, CA	Chicago, IL	100	0.28	0.67

Note. This table shows the difference ratio and coefficient of variation in parallel for sampled zip code pairs. The values are sorted by QTY.

Figure 10 shows top 10 frequent routes with high difference ratio on a map. White circles are origin points and red circles are destination points. The size of each circle represents the total number of trips recorded in the datasets, the larger the circle, the busier the region. Paths are colored based on difference ratio levels. Dark red paths' transit times highly fluctuated by day of the week. The level of variation decreases as the path color is lighter. From Figure 10, although some routes with big variation in transit times are in-state trips, most of the routes are out of state.

Figure 8

Map visualization for areas with dense traffics and high Difference ratio



Note. This map shows the metro areas with frequent shipments and highly variate transit times by day of the week. The map was graphed using Tableau public.

4.3 Day of the week bimodality – GMM Model & Separation Factor

Following the findings of the Tukey test in section 4.2, we next investigated identifying and quantifying within distribution bimodality. To evaluate the lane and day of the week bimodality, we chose to use the Gaussian Mixture Model (GMM) to cluster distributions into two groups. As discussed in section 3.4, the GMM model is used to create a clear distribution separation that can be further examined to examine bimodality within the day of the week.

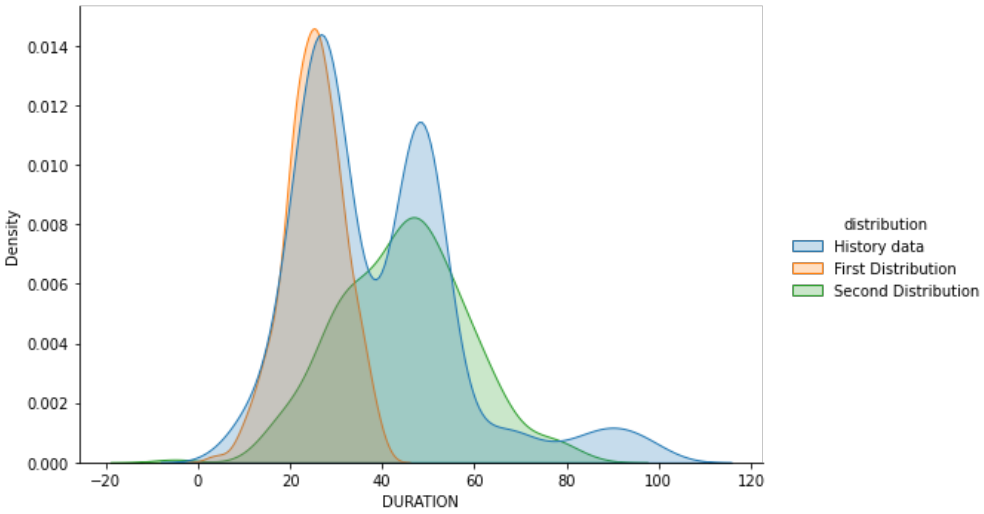
4.3.1 Distribution Splits using GMM

We next applied the GMM to zip code pair data, filtering for lanes with greater than 10 trips across the dataset. The output generated clusters of 2 gaussian distributions, with parameters: mean 1, mean 2, standard deviation 1, standard deviation 2 and alpha. Together, the two gaussian distributions form a mixture distribution. The distributions were ordered, such that mean 1 < mean 2. A visual example

of this can be seen in Figure 9, illustrating the 2 output distributions from the GMM on 600_180's Monday transit time distribution. The first distribution captures the data around the first hump, with the second distribution capturing the second hump and the wider long tail of longer transit times.

Figure 9

Historical distribution of 3-digit zip code 600 to 181 Monday transit times and GMM distributions



Note. This distribution was created using seaborn within Python. zip code 600 contains regions within Chicago, IL zone and zip code 181 contain regions within the Brooklyn, NY zone.

The GMM output of the 10 highest volume lanes can be seen in Table 7.

Table 7

Table of GMM distribution clusters

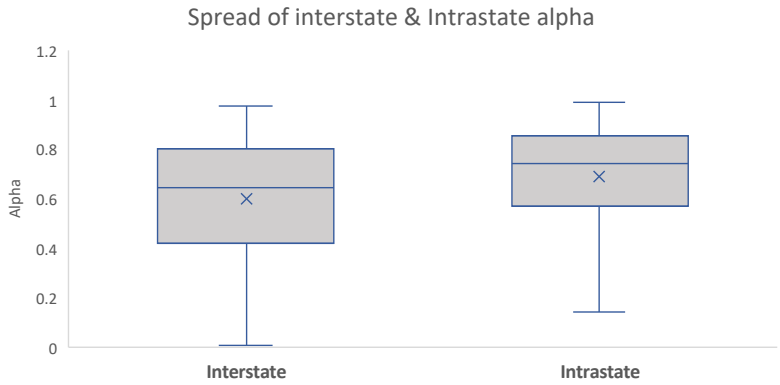
			Lanes: Zip-code Pairs									
	GMM Distr.	Day	181_180	335_334	335_336	600_181	600_750	600_917	604_601	604_604	750_750	752_770
Mean	Distribution 1	Monday	2.5	10.5	9.6	26.5	31.9	66.9	6.1	8.3	5.3	22.8
		Tuesday	2.4	10.1	10.2	27.2	37.2	29.5	10.3	11.1	2.0	21.9
		Wednesday	5.0	9.7	10.2	27.1	37.5	84.0	10.6	7.5	6.4	21.5
		Thursday	2.9	9.5	9.7	30.5	43.2	72.9	9.3	10.2	1.2	17.6
		Friday	4.0	58.3	6.3	27.2	31.1	74.7	5.9	5.3	1.4	13.2
	Distribution 2	Monday	20.3	38.6	34.6	45.5	60.0	93.0	35.8	37.3	49.1	60.6
		Tuesday	17.2	33.0	34.5	44.2	100.7	70.1	81.0	80.0	21.0	67.6
		Wednesday	88.7	44.8	60.9	82.0	110.8	101.8	144.8	51.4	67.1	172.3
		Thursday	41.8	63.2	38.9	80.2	100.7	105.8	80.7	114.1	15.4	76.7
		Friday	70.6	84.3	63.8	83.3	88.0	103.7	88.4	82.6	45.8	73.4
Standard Deviation	Distribution 1	Monday	1.4	3.6	3.4	2.7	9.3	3.3	4.4	6.3	5.6	8.9
		Tuesday	1.3	3.1	3.6	2.1	12.5	3.9	7.9	9.1	1.1	10.4
		Wednesday	3.9	2.8	3.4	8.4	12.7	17.4	8.8	5.3	7.4	9.2
		Thursday	1.8	3.1	3.4	9.0	16.3	17.4	6.9	7.9	0.6	9.8
		Friday	2.9	2.6	5.3	8.5	13.5	7.0	3.7	4.1	0.6	6.5
	Distribution 2	Monday	12.8	21.2	14.2	18.9	15.6	40.0	22.2	20.5	21.2	9.3
		Tuesday	12.5	14.6	14.2	18.5	12.7	17.2	41.9	32.0	16.8	9.9
		Wednesday	48.9	30.4	32.4	21.9	19.8	13.1	53.7	34.2	30.1	69.8
		Thursday	46.4	14.9	22.1	27.4	14.7	11.1	31.2	53.3	11.9	15.0
		Friday	47.4	35.6	13.8	21.4	26.2	12.1	54.9	58.8	39.3	22.2
Allocation	Alpha	Monday	64%	77%	81%	36%	65%	77%	74%	65%	90%	92%
		Tuesday	68%	79%	83%	33%	93%	4%	86%	89%	64%	95%
		Wednesday	86%	77%	86%	50%	90%	35%	90%	66%	89%	93%
		Thursday	74%	92%	81%	40%	74%	19%	93%	91%	55%	92%
		Friday	86%	75%	16%	27%	28%	10%	79%	80%	56%	57%

Note. This table shows the GMM distribution outputs for mean, standard deviation and alpha weighting for the top 10 zip-code pairings within the examined data set.

Each distribution cluster found expressed unique alpha values that combined the two distributions into a mixture distribution. Figure 10 highlights lanes with strong preference for the first distribution – such as 604_601, within Chicago, IL intracity transit. Cross country transit lanes such as 600_917, Chicago to Bell Gardens, CA showed much stronger preference for the second distribution on all days except Monday. Figure 10 confirms the insight with in stronger bias for the first distribution seen in intrastate transit lanes and much wider spread of alphas seen in interstate transit lanes. Our underlying hypothesis to explain higher alpha in intrastate lanes is that short distance transit lanes have longer tails which are more likely to be clustered in the second, larger GMM distribution. There could be many factors driving the longer tail of transit times, such as urban traffic or higher impact of delays on overall time in transit.

Figure 10

GMM alpha allocation between interstate and intrastate

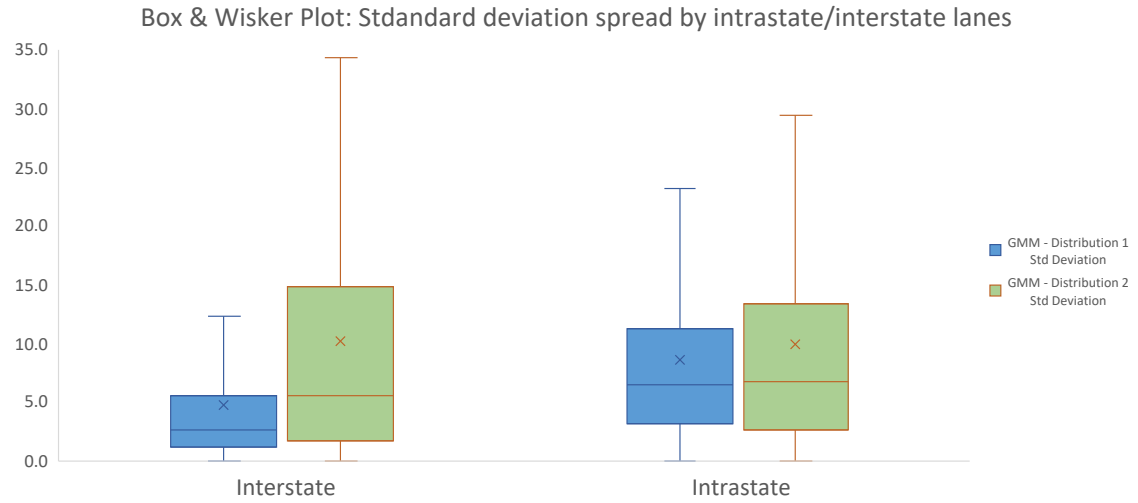


Note. This box and whisker plot shows the spread of the alpha weighting for a mixture distribution for interstate and intrastate lanes

When evaluating the standard deviation across the two distributions, we find that the intrastate transit lanes have higher levels of standard deviation across both distributions (Figure 11). The interstate transit lanes are more likely to have much tighter standard deviation within the first distribution and much wider potential outcomes in the second distribution. Earlier we confirmed in Figure 111 that there is less skew towards the first distribution, given a high alpha weight the first distribution more within a mixture distribution. Pairing the fact that interstate routes have more weighting towards their second distribution, we can infer that interstate transit times have higher spreads in standard deviation and more likely to have their true transit times within the second part of the mixture distribution. Intrastate lanes on the other hand are less likely to have distributions within the second distribution, and thus those that are within the second distribution may have an oversized impact on the overall transit time distribution for intrastate trips.

Figure 11

Intrastate/Interstate standard deviation spread by GMM distribution type

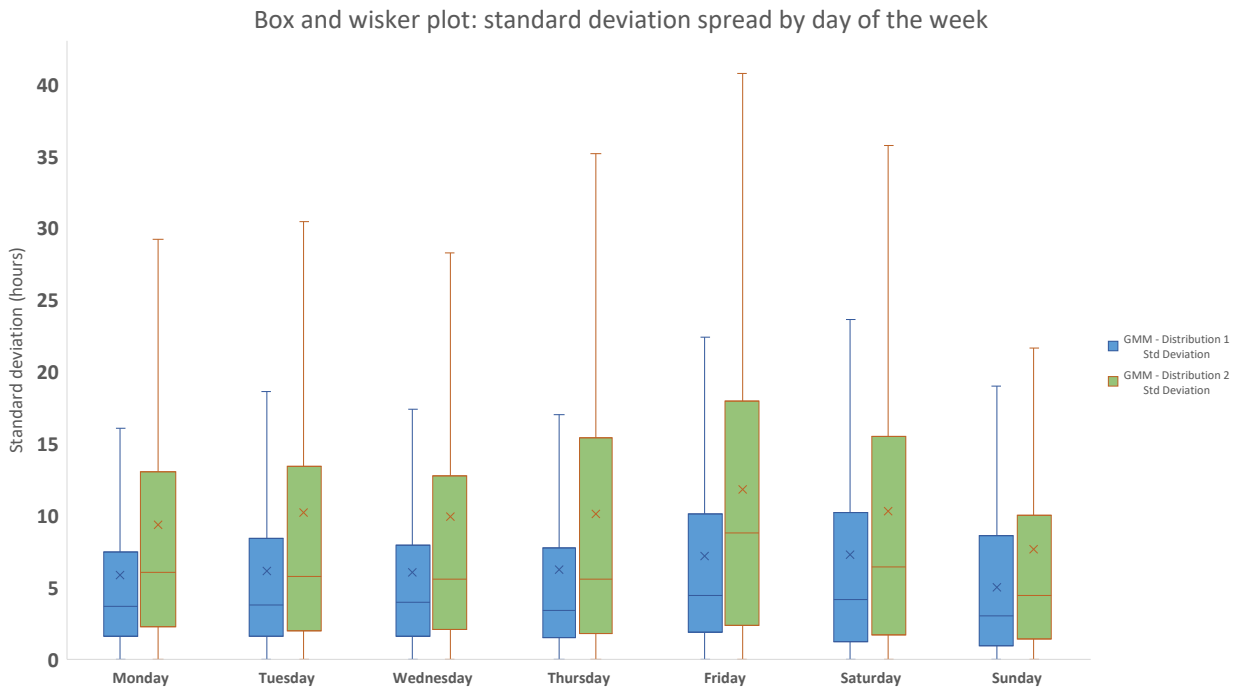


Note. This box and whisker plot shows the spread of the standard deviations for intrastate/interstate lanes for each distribution produced by the GMM.

In addition to intrastate/interstate splits, we also chose to evaluate the spread of each distribution's day of the week standard deviation. We found a similar large spread in standard deviations across the second GMM distribution, with the Thursdays and Fridays significantly wider in transit time variation compared with Mondays and Tuesdays (Figure 12).

Figure 12

Day of the week standard deviation spread by GMM distribution type

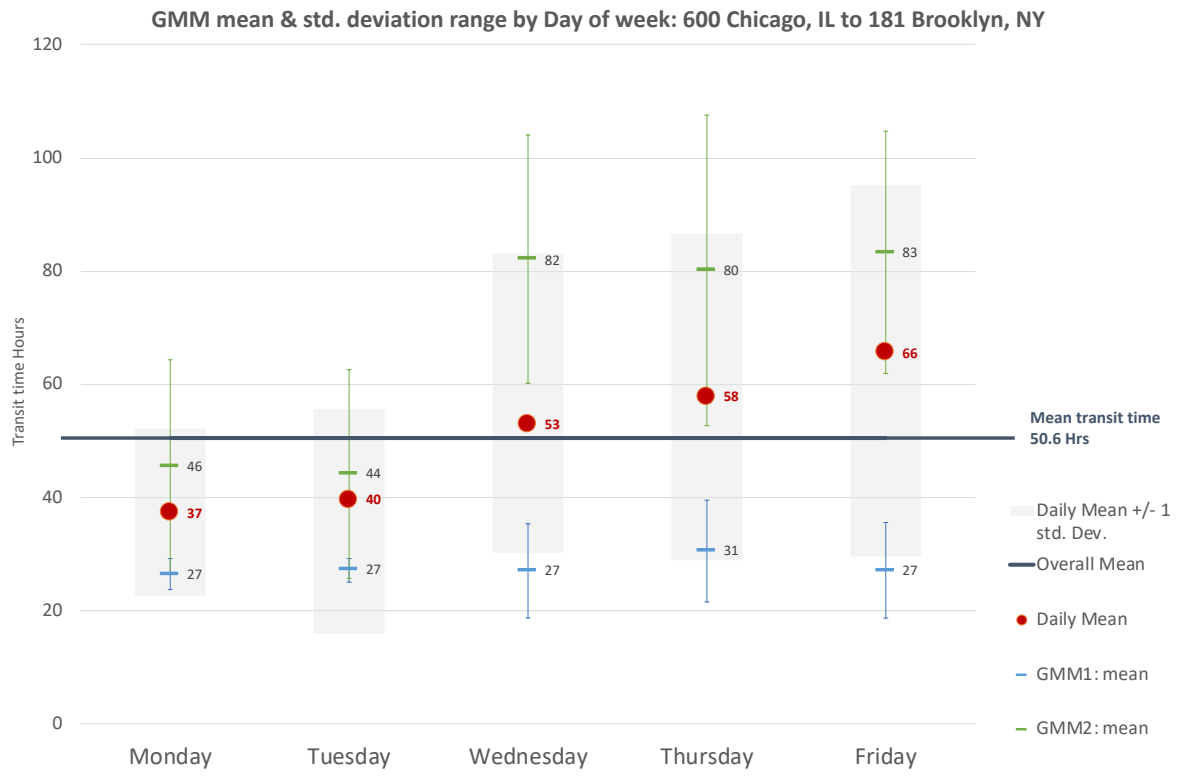


Note. This box and whisker plot shows the spread of each lanes standard deviation by day of the week, for each of the two distribution produced by the GMM.

Figure 13 is a clear example of how spread of transit times impacts a single lane. We plotted the results of the GMM, the original daily and overall transit time averages to understand how the spread of transit times impact one lane, 600 to 181 – Chicago, IL to Brooklyn, NY. The key insight figure d highlights are that Monday and Tuesday's spread in transit times is much tighter than the remaining weekdays. Additionally, there is a step change in transit times from Wednesday onwards, with high levels of spread between the two GMM distributions. The 2nd of the two GMM distributions also exhibit much higher levels of spread in standard deviation on Wednesday, Thursday, and Friday.

Figure 13

Day of the week standard deviation spread by GMM distribution type



Note. This figure shows the spread of each unique distribution produced by day of the week for the 600 to 181 zip code pairing, illustrating the changes in spread when using the distributions produced by the GMM.

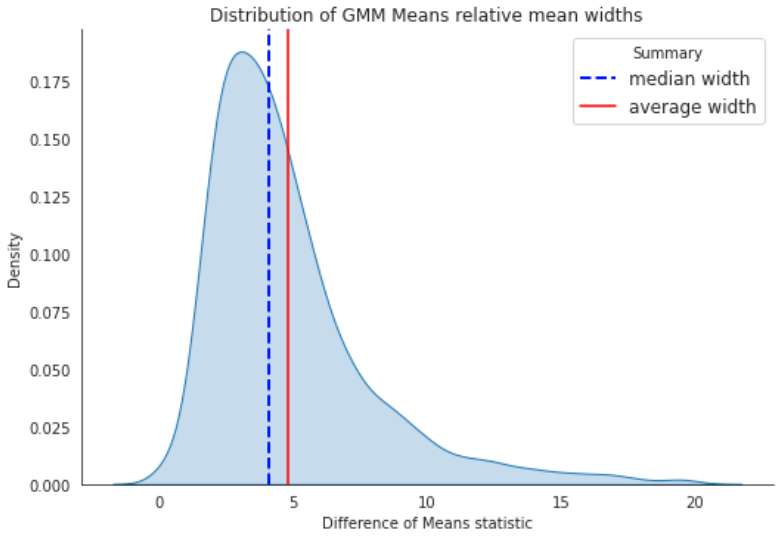
Within the single lane example shown in Figure 13, we can see that there is clear split in transit time distribution by day, with this lane having a large overall standard deviation of 27 hours. Monday and Tuesday's transit times would be significantly under the 27-hour standard deviation, driving large amounts of safety stock to account of non-existent variables. However, the remaining days would have very wide distributions, with negative kurtosis and spread in GMM distributions found. Therefore, as each day of the week has different distributions, using a static transit time for a lane would drive either too much safety stock or understocking depending on which day and other factors that drive within day bimodality.

4.3.2 Transit Time Distribution Spread - Difference in Means statistic

To form an understanding of how spread the means produced from the two GMM distributions, we calculated the difference of means statistic described in 3.4.2. The difference of means statistic is used to understand how means were separated relative to their standard deviations. Figure 14 shows the distribution of the difference of means statistic, illustrating the widespread in relative widths of 4.5-5 hours between the two distributions. As a rule of thumb, distributions that have a difference greater than 2 can be potential candidates for exhibiting bimodality. This, however, is not definitive, as overall distributions with large negative kurtosis (i.e. spread) may also have high difference of means as a result of the output from the GMM.

Figure 14

Distribution of the difference of means between the two GMM clusters



Note. This figure was created using seaborn in python. This shows the distribution of the difference of means statistic and the mean and median of this distribution.

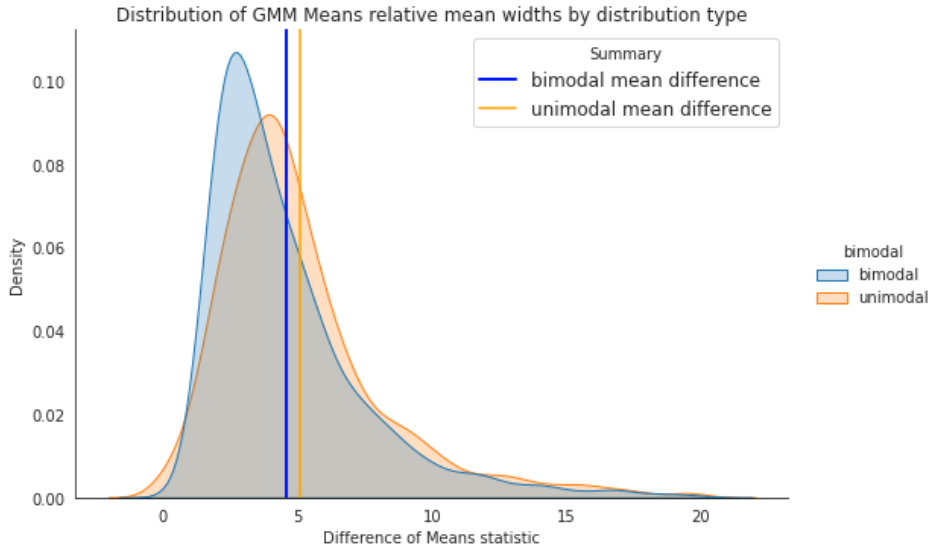
4.3.3 Defining Bimodality using a separation factor

Following our application of the difference of means statistic in section 4.3.2, we applied a separation factor truly classify distributions as bimodal or unimodal. Described in section 3.4.2, the separation factor is a hurdle factor that defines bimodality with respect to two Gaussian distributions ratio

of variances. Applying this to the transit lanes by day of the week, we were able to identify that 1,985 of the 4,013 unique lane and day of the week pairs were bimodal, accounting for 49% reviewed lanes. When weighting each lane by the number of trips made in each unique lane and day of the week pairing, we find that 86,907 of the 145,966 trips were bimodal: accounting for 70% of reviewed trips. The key finding that the separation factor was able to highlight was that bimodality is present in almost half the lanes reviewed, and in particularly high-volume lanes.

Figure 15

Distribution of the difference of means between the two GMM clusters



Note. This figure was created using seaborn in python. This shows the distribution of the difference of means statistic for the GMM output distributions classified as bimodal and unimodal.

An interesting counter pattern arose when investigating the difference of means for lanes classified as bimodal and unimodal (Figure 15). Our intuition suggested that bimodal distributions would exhibit high levels of differences in means, however what the distribution in Figure 15 highlights that bimodal route distribution means are on average closer together. Our hypothesis is that true unimodal lanes that have negative kurtosis, i.e., more spread out and flatter, with larger tails but no distinct two peak distributions.

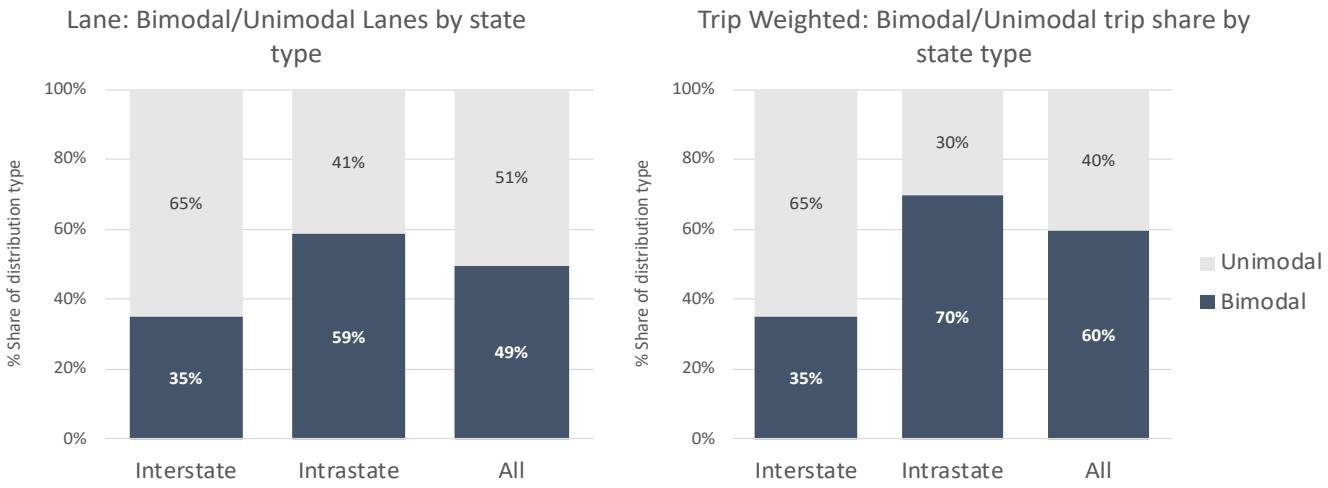
On the other hand, bimodality within the underlying lane would exhibit density around two peaks, with smaller densities in the tails of the distribution.

4.3.4 Interstate, Intrastate Lane and State Bimodality:

After we were able to define which lanes were bimodal and those that were unimodal, we investigated the lane categories to understand if there were patterns in bimodality by location. A key finding in our research was that intrastate lanes that were proportionately more bimodal, with 59% of lanes and 70% of intrastate trips on those lanes classified as bimodal. (Figure 16)

Figure 16

Intrastate & Interstate bimodality by Lanes & trip share

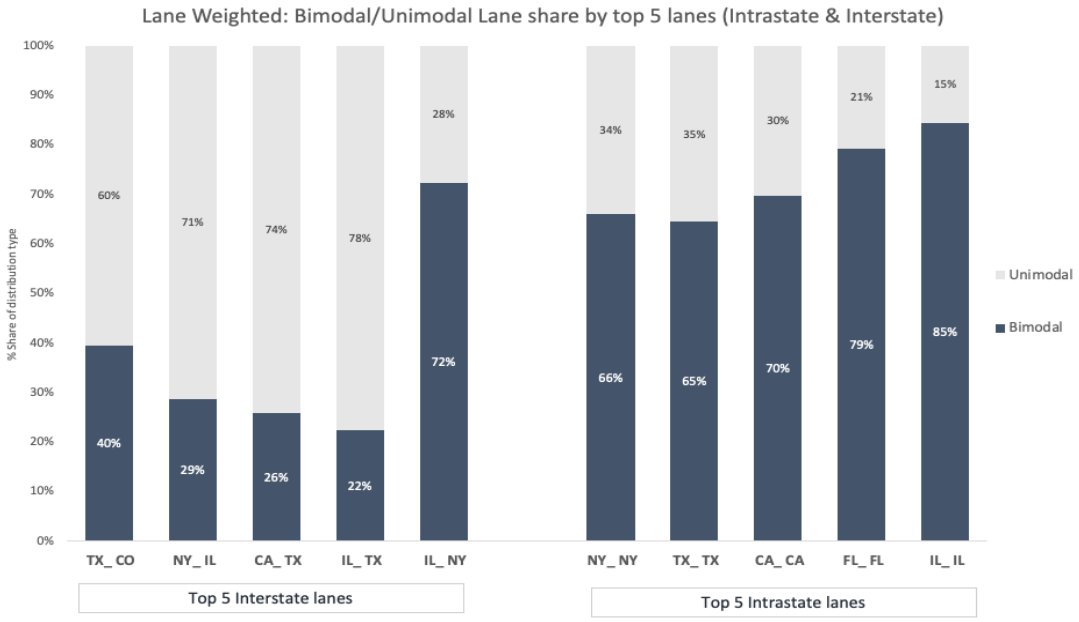


Note. This figure shows the share of bimodality by lane count and weighed by trips within those lanes for both interstate and intrastate transit times.

When evaluating the highest volume lanes for bimodality, we saw that across almost all intra state lanes there was strong bimodality across all lanes, with lanes within Florida and Illinois with the highest intrastate bimodality. (Figure 17). Interstate lanes however were primarily unimodal, with Illinois to New York lanes a rare standout in having higher share of bimodal transit time distributions.

Figure 17

Top 5 Intrastate & Interstate bimodality share by total trip volume

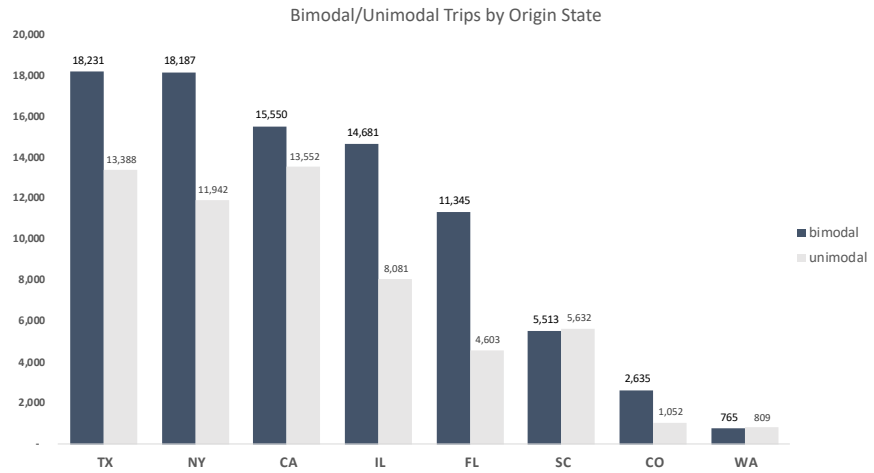


Note. This figure shows the top 5 intrastate/interstate transit lanes volume that is classified are bimodal.

When pooling by lane origin state, we found that the highest differences in trips across lanes that had bimodality were in Florida, Illinois, and Colorado. (Figure 18). In relative terms, low levels of bimodality existed ins California and South Carolina.

Figure 18

Top 5 Intrastate & Interstate bimodality share by total trip volume



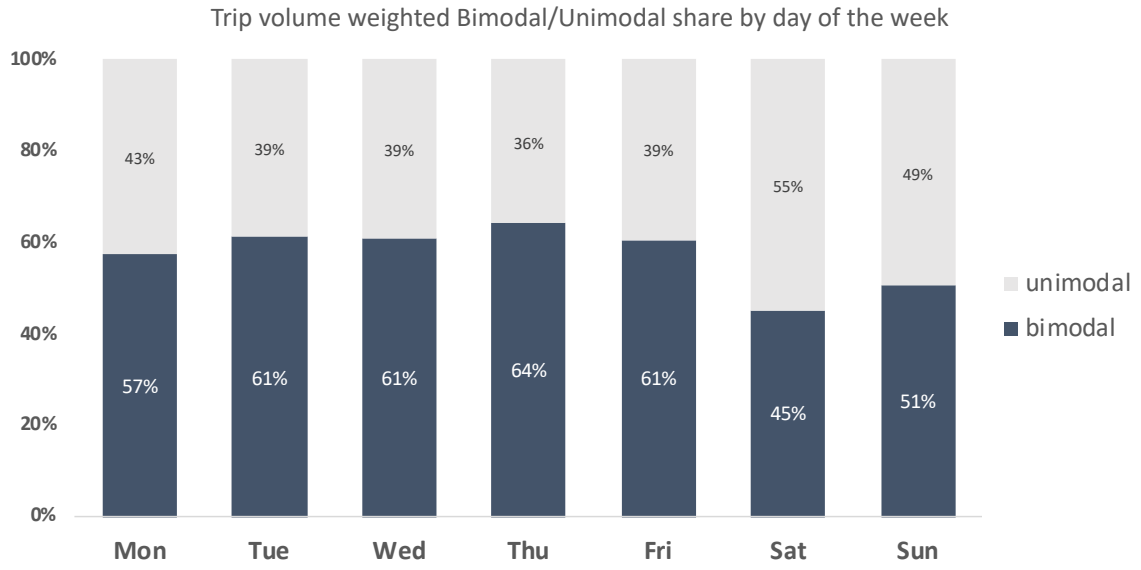
Note. This figure shows the origin state volume of trips that are bimodal and unimodal.

4.3.5 Bimodality by day of the week:

In addition to understanding bimodal lanes by states, we also grouped lanes by day of the week found that there was a clear split in weekday bimodality over a lower likelihood during Saturdays and Sundays. (Figure 19) Multiple factors could be involved in driving this relationship, with weekday traffic or lower weekend volumes helping consistency in transit distributions. Although less apparent in aggregate than in the individual lane shown in figure d, we can see that there is a slight increased share in Thursday's bimodality relative to other days of the week. A key practical takeaway from this could be to save on unnecessary transit time safety stock by avoiding starting trips on Thursday.

Figure 19

Share of Bimodal trips by day of the week



Note. This figure shows the share of bimodal trips across all 4013 unique lane days of the week.

4.4 Numerical Experiment - Bimodality Impact measurement

To measure the impact of bimodality, we created three case scenarios to compare safety stock and its impact on inventory costs by using the safety stock inventory level formula in section 2.1.1 and logistic cost formula in section 2.3.1. We defined the first scenario as the base case, which used a static transit timetable commonly used in industry, to calculate the reorder amount. The treatment cases were to use

different transit times determined from GMM bimodal distribution to calculate reorder amount, while keeping the same demand level and desired service level. The outcome of three cases were used to compare inventory costs, calculate cost savings, and determine the best day of the week for shipping.

The example case that we used was to assume a fulfillment center in Brooklyn, NY (Zip code starting with 181) needs to replenish toilet paper and the toilet paper was dispatched from Chicago, IL (Zip code starting with 600) distribution center. The carrier assumed a lane average transit time of 50.64-hours with 27.17-hour standard deviation. Furthermore, assume the daily demand is normally distributed ~N (1000,80) units. Under a case where an order is placed weekly, therefore the total lead time is 7-day. We calculated the safety stock at 95% service level:

$$SS = Z\sqrt{\overline{L}\,\sigma_D^2 + \overline{D}^2\,\sigma_L^2}$$

$$SS = 1.645\sqrt{(7)\cdot 80^2 + 1000^2\,(\frac{27.17}{24})^2}$$

$$SS = 1{,}895$$

The unit costs per bag of 18-rolls ultra-soft Charmin toilet paper is $18.79 and the cost per order is $1000. The annual holding cost is 20% of unit cost. Keeping the order frequency constant, changing the number of safety stock will only affect the cost of inventory:

$$Holding\ cost = holding\ cost\ of\ cycle\ stock + holding\ cost\ of\ safety\ stock$$

$$Holding\ cost = Unit\ holding\ cost * (\overline{L} * \frac{\overline{D}}{2} + SS)$$

$$Holding\ cost = 18.79 * 20\% * (7 * \frac{1000}{2} + 1895)$$

$$Holding\ cost = \$20{,}274.4$$

For the treatment case, we calculated the reorder amount's safety stock (SS1 and SS2) by using means and standard deviations from GMM bimodal distribution per day of the week. First safety stock (SS1) was calculated for the first distribution and second safety stock (SS2) was calculated for the second

distribution. As shown in Table 8, all safety stock for the first distribution is significantly lower than the safety stock from the base case. In particular, shipments start from Monday and Tuesday require fewer than 400 units for safety stock, which is close to 22% of base case safety stock quantities. Safety stock for the second distribution, in contrast, is much higher than safety stock for the first distribution, but the majority is lower than base case values, except for Thursday.

Table 8

Safety stock calculation of bimodal distributions for each weekday

Day	Mean 1	Mean 2	Std. Dev. 1	Std. Dev. 2	alpha 1	alpha 2	SS1	SS2
Monday	26.5	45.5	2.7	18.9	0.4	0.6	395	1344
Tuesday	27.2	44.2	2.1	18.5	0.3	0.7	377	1316
Wednesday	27.1	82.0	8.4	22.0	0.5	0.5	671	1544
Thursday	30.5	80.2	9.0	27.4	0.4	0.6	707	1909
Friday	27.2	83.3	8.5	21.4	0.3	0.7	676	1509

Note. This table shows the safety stock calculated by using four mean and standard deviation output from GMM model per day of the week.

We calculated holding cost with respective safety stock values in Table 8, by using the same demand distribution and service level from the base case. We also calculated weighted holding cost per day of the week by multiplying holding cost by alpha values for each distribution. From the weighted holding cost in Table 9, shipments starting Monday and Tuesday generated more than $3K (or 16%) cost saving compared to the base case. Thursday, which was considered the least favorite day of the week, also reduced $1.7K (or 9%) holding cost compared to the base case.

Table 9

Holding cost calculation by day of the week for bimodal distribution

Day	Holding Cost (Std. Dev. 1)	Holding Cost (Std. Dev. 2)	Total Cost	Saving	Saving %
Monday	$14,639	$18,203	$16,937	$3,336	16%
Tuesday	$14,570	$18,099	$16,934	$3,339	16%
Wednesday	$15,675	$18,956	$17,317	$2,955	15%
Thursday	$15,811	$20,329	$18,521	$1,752	9%
Friday	$15,695	$18,826	$17,980	$2,292	11%

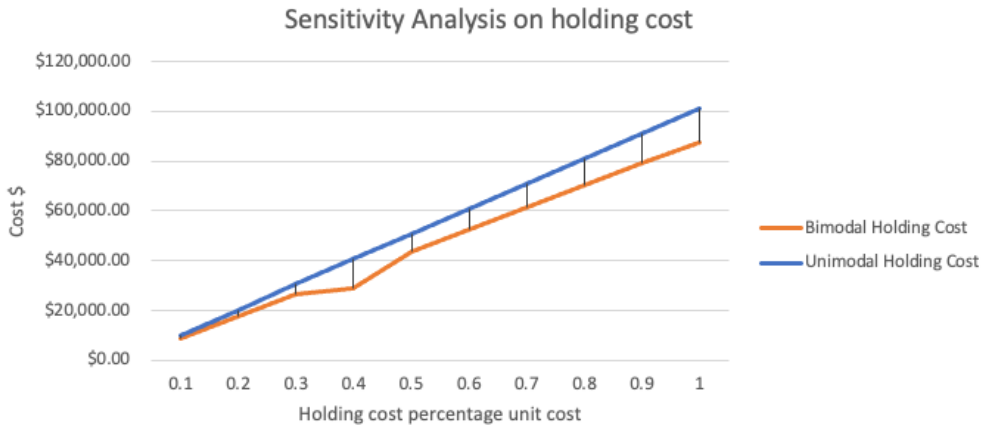
Note. This table shows the cost saving values and cost saving percentages from total cost by shipping from different day of the week.

Treating transit time differently across day of the week could generate as high as 16% cost reduction in the toilet paper case above. Companies looking to reduce inventory cost should add day of the week to the transit timetable and choose the most cost-effective day to start shipment.

We applied the same methodology to other commodities with higher unit prices. Keep the safety stock and unit cost unchanged, the larger holding cost percentage unit cost, the higher the total holding cost and the larger the cost savings. For perishable items with high holding cost rate, the impact of bimodality in transit time is amplified by generating big cost savings.

Figure 20

Safety stock calculation of bimodal distributions for each weekday



Note. This figure shows the change of cost savings by increasing holding cost rate while keeping the safety stock and demand constant.

5 DISCUSSION AND CONCLUSION

This research showed bimodality in transit times and demonstrated impact on inventory decisions. We have developed a Gaussian Mixture Model (GMM) based on geolocation data captured by project44 and analyzed the impact of bimodality on safety stock in order to reduce inventory cost while maintaining the same service level. Shippers could leverage the bimodality discovered from this research to create a more sophisticated transit timetable to improve the measurement of transit times and better monitor the goods in transit.

We conducted distribution analysis for transit duration for each day of the week on the simulated transit time from eight US metro areas' aggregated geo data. The distribution analysis in 4.1 clearly showed differences in distributions for each day of the week, with each day having different means and variance. The day of the week difference created a large dispersion around the mean transit time, which caused shippers to gauge their safety stock inaccurately. Our research shows that 40% of the routes recorded have a high coefficient of variation in transit duration. Lanes with a high coefficient of variance in transit times drive high safety stock requirements to account for their variability. However, our research showed that high variability in some lanes may in fact be genuine and cannot only be attributed to bimodality.

Further exploration on the difference among day-of-the-week transit time distributions confirmed our initial investigation that using transit times static distributions, which ignored day-of-the-week information, would result in excess inventory holding or in shipping delays. We performed Tukey hypothesis tests to evaluate whether mean transit times are equivalent across unique day-of-the-week combinations. Based on Tuckey test results, we rejected the null hypothesis and proved that each day of the week has a unique transit time distribution. Particularly, Chicago, IL, has the highest proportion of unique daily transit time distributions, i.e., difference ratio, regardless of direction, inbound or outbound. This means that, for Chicago, IL, most days of a week have a unique transit time that shippers should account

for. Other metro areas, such as Lakeland, FL, Bell Gardens, CA and Palestine, TX, also expressed a high difference ratio.

Comparing the coefficient of variation to difference ratio in parallel, we found no significant correlation. High coefficient of variation does not necessarily indicate high difference ratio. In other words, a high variation in the distribution does not mean that each day of the week has a different time. Our interpretation of this result is that those distributions with high coefficient of variability have high negative kurtosis with flatter spread in distribution. Therefore, the overall distribution cannot easily be differentiated among each day of the week and thus cannot be classified as truly different distributions.

project44's main goal is to prove that a more accurate transit timetable would make better delivery date predictions and reduce safety stock for shippers and carriers. To create a more sophisticated timetable, we studied day-of-the-week transit time distribution in further detail. Based on the distribution analysis and hypothesis testing results, we created an unsupervised machine learning model, Gaussian Mixture Model (GMM) to define distinct transit time distributions by clustering distribution data points within each day of the week. GMM split the distribution in two humps and produced the mean and sigma values for each bimodal mixture distribution. We discovered that the majority of the first peaks had low dispersion around the mean and the second peaks grouped all long-tail transit times, with typically higher standard deviation as a result; with this trend being particularly strong within intrastate transit lanes. Furthermore, according to the box plot in section 4.3.1, Monday and Tuesday transit times express lower spread in means and have less variation across transit times. In contrast, the rest of the week has considerably higher spread in transit time distributions. Finally, we compared the intrastate and interstate transit time distributions. Intrastate transit routes had higher possibility to be bimodal. We infer that this could be due to intrastate transit routes having shorter lead times in high traffic areas or high-density zones.

The variation of transit time could significantly impact the safety stock level, according to Hadley-Whitin reorder formula in section 2.1.1. We conducted a numerical experiment, evaluating the GMM outputs to reorder formula and discovered that a more sophisticated transit timetable, factorized by day of

the week and bimodality of the transit time distribution, reduced the variation and increased accuracy. Our research proved, within the numerical experiment shown in section 4.4, that accounting for bimodality and day of the week resulted in 38% inventory reduction and up to 16% holding cost saving in toilet paper case study. Testing for sensitivity around the assumptions of the experiment that extend these findings to different commodities unit prices or holding cost rate could result in higher cost saving for the shippers and carriers. For example, perishable items like fruit that have higher holding cost per unit would reduce more inventory cost according to the safety stock formula due to the higher penalty of expirations in this type of commodity. Companies can use their understanding of the bimodality within their key transit lanes to segment their inventory according to their holding cost impact and thus prioritize high holding cost SKUs for transit across days have lower bimodality. A further extension of this work could further enable shippers to reduce inventory holding costs to accommodate costs for routes and transit practices that are more environmentally sustainable.

Our research findings clearly described that accounting for day of the week and transit time bimodality can yield shippers positive returns in reducing safety stock required for inaccurate lead time variability. By forward planning and making orders earlier in the week, shippers can reduce excess safety stock required for end-of-week transit time bimodality. Further, if shippers can identify the drivers of intraday bimodality, they will be able to better segment their transit timetable into each component of the mixture distribution and thereby significantly reduce safety stock requirements.

REFERENCES

Allen, W. B., Mahmoud, M. M., & McNeil, D. (1985). The importance of time in transit and reliability of transit time for shippers, receivers, and carriers. *Transportation Research Part B: Methodological*, *19*(5), 447–456. https://doi.org/10.1016/0191-2615(85)90057-8

Chiang, Y.-S., & O. Roberts, P. (1980). A note on transit time and reliability for regular-route trucking. *Transportation Research Part B: Methodological*, *14*(1–2), 59–65. https://doi.org/10.1016/0191-2615(80)90032-6

Constable, G. K., & Whybark, D. C. (1978). The Interaction of Transportation and Inventory Decisions. *Decision Sciences*, *9*(4), 688–699. https://doi.org/10.1111/j.1540-5915.1978.tb00754.x

Das, L. (2013). *The impact of bimodal distribution in ocean transportation transit time on logistics costs: An empirical & theoretical analysis* [Thesis, Massachusetts Institute of Technology]. https://dspace.mit.edu/handle/1721.1/81117

Das, L., Kalkanci, B., & Caplice, C. (2014). Impact of Bimodal and Lognormal Distributions in Ocean Transportation Transit Time on Logistics Costs. *Transportation Research Record: Journal of the Transportation Research Board*, *2409*(1), 63–73. https://doi.org/10.3141/2409-09

Dullaert, W., & Zamparini, L. (2013). The impact of lead time reliability in freight transport: A logistics assessment of transport economics findings. *Transportation Research Part E: Logistics and Transportation Review*, *49*(1), 190–200. https://doi.org/10.1016/j.tre.2012.08.005

Jacobs, F. R., and Chase, R. B. (2016). *Operations and Supply Chain Management* (4th Edition). McGraw Hill.

Lau, H.-S., & Zaki, A. (1982). The Sensitivity of Inventory Decisions to the Shape of Lead Time-Demand Distribution. *A I I E Transactions*, *14*(4), 265–271. https://doi.org/10.1080/05695558208975239

Massiani, J. (2008). Can we use hedonic pricing to estimate freight value of time? Economics and Econometrics Research Institute, 24(1), 24.

Muratov, A. L., & Gnedin, O. Y. (2010). Modeling the Metallicity Distribution of Globular Clusters. *The Astrophysical Journal*, *718*(2), 1266–1288. https://doi.org/10.1088/0004-637X/718/2/1266

project44 | About project44—Supply Chain Visibility Leader. (n.d.). Project44. Retrieved November 29, 2021, from https://www.project44.com/about

Project44 turns unicorn after raising $202M Series E funding from Goldman Sachs, Emergence Capital. (2021, June 2). *SaaS Industry*. https://saasindustry.com/news/project44-turns-unicorn-after-raising-202m-series-e-funding-from-goldman-sachs-emergence-capital/

Project44 valued at more than $1B after raising more funding. (n.d.). Chicago Inno. Retrieved October 12, 2021, from https://www-bizjournals-com.ezp-prod1.hul.harvard.edu/chicago/inno/stories/fundings/2021/06/01/project44-valued-at-more-than-1b-after.html

Tadikamalla, P. R. (1984). A comparison of several approximations to the lead time demand distribution. *Omega*, *12*(6), 575–581. https://doi.org/10.1016/0305-0483(84)90060-4

Tyworth, J. E., & O'Neill, L. (1997). Robustness of the normal approximation of lead-time demand in a distribution setting. *Naval Research Logistics (NRL)*, *44*(2), 165–186. https://doi.org/10.1002/(SICI)1520-6750(199703)44:2<165::AID-NAV2>3.0.CO;2-7

Tyworth, J. E., & Zeng, A. Z. (1998). Estimating the effects of carrier transit-time performance on logistics cost and service. *Transportation Research Part A: Policy and Practice*, *32*(2), 89–97. https://doi.org/10.1016/S0965-8564(97)00020-7

Vega, D. A. S. D. L., Lemos, P. H., Silva, J. E. A. R. da, & Vieira, J. G. V. (2021). Criteria analysis for deciding the LTL and FTL modes of transport. *Gestão & Produção*, *28*. https://doi.org/10.1590/1806-9649-2020v28e5065

APPENDIX A

zip_pair	ORIGIN_3ZIP	ORIGIN_METRO	DESTINATION_3ZIP	DESTINATION_METRO	DAY	DURATION
802_604	802	Denver, CO	604	Chicago, IL	Tue	28.82714119
802_604	802	Denver, CO	604	Chicago, IL	Tue	38.53547071
802_604	802	Denver, CO	604	Chicago, IL	Tue	26.74298829
802_604	802	Denver, CO	604	Chicago, IL	Tue	35.55495107
802_604	802	Denver, CO	604	Chicago, IL	Tue	28.59024707
802_604	802	Denver, CO	604	Chicago, IL	Tue	50.68087591
802_604	802	Denver, CO	604	Chicago, IL	Tue	40.98050292
802_604	802	Denver, CO	604	Chicago, IL	Tue	37.5344923
802_604	802	Denver, CO	604	Chicago, IL	Tue	32.72773124
802_604	802	Denver, CO	604	Chicago, IL	Tue	29.89655638
802_604	802	Denver, CO	604	Chicago, IL	Tue	29.05793709
802_604	802	Denver, CO	604	Chicago, IL	Tue	29.27361004
802_604	802	Denver, CO	604	Chicago, IL	Fri	18.55417753
802_604	802	Denver, CO	604	Chicago, IL	Fri	62.825
802_604	802	Denver, CO	604	Chicago, IL	Fri	44.7
802_604	802	Denver, CO	604	Chicago, IL	Fri	62.29051116
802_604	802	Denver, CO	604	Chicago, IL	Fri	44.7
802_604	802	Denver, CO	604	Chicago, IL	Fri	67.61741926
802_604	802	Denver, CO	604	Chicago, IL	Thu	32.03826305
802_604	802	Denver, CO	604	Chicago, IL	Thu	35.3666665
802_604	802	Denver, CO	604	Chicago, IL	Thu	24.27158299
802_604	802	Denver, CO	604	Chicago, IL	Thu	24.7833335
802_604	802	Denver, CO	604	Chicago, IL	Thu	43.29754074
802_604	802	Denver, CO	604	Chicago, IL	Wed	144.3714958
802_604	802	Denver, CO	604	Chicago, IL	Wed	29.6583335
802_604	802	Denver, CO	604	Chicago, IL	Wed	22.01926034
802_604	802	Denver, CO	604	Chicago, IL	Wed	29.6583335
802_604	802	Denver, CO	604	Chicago, IL	Wed	52.84163515
802_604	802	Denver, CO	604	Chicago, IL	Mon	18.17148197
802_604	802	Denver, CO	604	Chicago, IL	Mon	23.25693091
802_604	802	Denver, CO	604	Chicago, IL	Mon	42.3583335
802_604	802	Denver, CO	604	Chicago, IL	Mon	27.7416665
802_604	802	Denver, CO	604	Chicago, IL	Mon	40.44411079

Appendix A. Sample Simulated Data from Denver, CO, to Chicago, IL.