UNIVERSITY OF MARYLAND

National Transportation Center

Project ID: NTC2014-SU-R-06

IMPROVING THE EFFICIENCY AND RELIABILITY OF FREIGHT TRANSPORTATION

Final Report

by

George F. List
gflist@ncsu.edu
919-515-8038
North Carolina State University
909 Capability Drive, Suite 3600
Research Building IV
Raleigh, North Carolina 27606

and

Elizabeth Williams
Jeremy Addison
Atefeh Morsali
North Carolina State University

for

July 2017

ACKNOWLEDGEMENTS

DISCLAIMER

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

EXECUTIVE SUMMARY

This research report is focused on advancing the methodological frontier in the analysis of efficiency and reliability for freight transportation. It deals primarily with truck-related shipments although the tools are applicable to other modes and multi-modal systems. The topic is important because of the economic value that results from minimizing the resource consumption associated with freight activity.

Section 1 presents a perspective on freight-related reliability by describing concepts and assessment tools that pertain to the ensuing analyses. It indicates that freight travel time reliability is about consistency in on-time arrivals and departures, not travel times per se. A carrier's objective is to arrive and depart within on-time windows (OTWs) a very high percentage of the time. In current practice, carriers measure on-time performance in terms of both departure and arrival events, separately and in combination. Shipments (and vehicle moves) are deemed to be "on-time" if they both depart during these specified windows. The objective is to find paths and vehicle tours that conjunctively maximize on-time performance. If a carrier has high percentages for both their arrivals and departures, their service will be perceived as being reliable. Their service is neither late nor early. Regarding the arrival window, if the truck arrives early and must wait, less carrier resources could have been used or extant resources could have been better deployed. If it arrives late, it means the customer's expectations have not been met. The same thoughts pertain to the departure window. If the truck departs early, then the same resource issues exist. And if it departs late, customer resources (dock time and space) have been unnecessarily tied up.

This is not to say that the travel time distributions are unimportant. Rather, it is to stress the fact that reducing the mean or the variance is not the primary objective. It is better to think about managing the mean and the variance to achieve a specific on-time performance objective. This means making intelligent tradeoffs between these two and possibly other measures that affect the total cost. The advantages of a reduced variance is more consistent travel times, regardless of the mean. Visit sequences can be accomplished with more confidence in the reliability of the visits. A reduced mean implies that later departure times can be employed, more customers can be visited within a given span of time, more customers can be visited with a given fleet size, and more customers can be serviced for a given siting decision. Reducing travel times also reduces in-transit inventory costs and fleet size requirements. But the mean should not be reduced at the expense of an increase in the variance.

Section 2 focuses on assessing the reliability of travel time on segments and routes (or paths). This is a fundamental building block upon which all the other reliability analysis depends. If segment and route travel times are reliable, or more precisely, if their reliability is understood, then decision support tools can help improve travel time reliability. The motivation is that a more refined sense of the travel times involved in making trips is helpful in ensuring that reliable freight service is provided. Working with the average travel times is not enough. What is desired is a sense of the individual vehicle travel times that might arise for a given origin-destination (OD) pair, for a variety of operating conditions (e.g., normal and abnormal, with the latter being characterized by weather, incidents, maintenance work, etc.).

In this section, three ways to synthesize the route-level travel time distribution for specific truck types are described. The first method is the least demanding in terms of data, and the most dependent on inference. It is easy to apply for a given operating condition, but the quality of its estimated distributions is highly dependent upon the insights of the analyst. The second method relies on individual vehicle travel time observations for segments that lie along the route of interest. It combines positive, negative and uncorrelated combinations of segment travel time distributions to synthesize a route-level distribution. The process is simple and straightforward yet yields a distribution that very closely matches the one observed. The third method uses segment-specific Monte Carlo sampling to synthesize route-level travel time distributions. The method is intuitively appealing because it capitalizes on natural ideas about how individual vehicle travel times arise on congested and uncongested networks. Each of the methods is described analytically and then illustrated using a case study example.

Section 3 describes how to choose paths and departure times for specific trips. A multi-step process can be used to find optimal departure times and paths. The process is as follows:

1) Solve a deterministic K-shortest path problem working backward from the OTW. In doing this, use the midpoint of the OTW as the nominal arrival time and use the median travel times as the path travel times (or a higher percentile for lower risk tolerance).
2) Develop a relationship between departure time and the probability of arrival during the arrival window for all the K-shortest paths identified.
3) Select the path and departure time that provides the best combination of travel time and reliability based on the risk and travel time preferences of the decision maker.

The section illustrates that there is often a tradeoff between reliability and travel time (and implicitly, cost). A path with a longer travel time may provide better reliability, but the higher travel time may increase cost. Padding a trip with slack time at the destination also improves reliability, but with a cost. Hence, there are likely to be tradeoffs. Anticipating those tradeoffs is important. Using a bi-criterion search is important. (And there may be other criteria such as minimizing the likelihood of exposure to accidents and incidents that should be considered, but those are not examined here.) Once the non-dominated paths have been identified, a utility function or some other evaluation and selection process can be employed to identify the best path to choose.

Section 4 addresses the issue of vehicle routing and scheduling under stochastic conditions. Nominally, the objective is to find an assignment of loads to vehicles and determine routings for the vehicles that optimizes all the performance metrics. In some instances, the loads are full truckloads, in which case the vehicles are assigned to carry loads from one point to another in sequence. Alternately, trucks may pick-up or deliver loads. In a third option, trucks may both pick-up and drop off loads, as with local couriers. The objectives are often to:

1) minimize total cost,
2) maximize on-time deliveries,
3) minimize the fleet size, and
4) maximize vehicle utilization.

Other objectives include:

5) maximizing on-time performance and
6) maximizing the lowest on-time performance among all the vehicles employed.

Two methods are presented for developing shipment assignments, routes and paths for a fleet of trucks operating out of a single depot. The first, a meta-heuristic search technique, uses a simulation-based heuristic based on an initial Clarke-Wright solution followed by merge, insertion, and 2-interchange reduction in order to solve the vehicle routing problem with stochastic travel times and soft time windows. The second is a solution methodology that makes a single pass through the set of customers to be visited and identifies an assignment of trucks which is both efficient and feasible. The method can be described as solving the following optimization problem:

Maximize the on-time performance, and
Minimize the cost of the service provided

Subject to:

1) all shipments are picked up and/or delivered all shipments
2) the capacity of any vehicle is not exceeded, and
3) the maximum tour duration for any vehicle is not exceeded.

An adaptation of the run cutting and scheduling procedure often used to assign buses to transit routes is employed to develop the shipment assignments and routes.

Section 5 focuses on siting analysis. The method presented uses Monte Carlo simulation to assess the reliability of the delivery service quality provided by candidate distribution center sites then identifies the best ones to choose. Because of this, it provides useful and meaningful results that are easy to understand. The method can be described as follows:

1) Specify the location of the customers (sites) to be visited, the locations of the candidate distribution centers (DCs), and the statistical characteristics of the travel times from the DCs to the customer sites.
2) For each DC:
   a. Conduct a Monte Carlo simulation of trips made from the DC to the customer sites for different times of day.
   b. Develop CDFs of the travel time distributions for each of the DCs.
3) Identify the non-dominated DCs.
4) Select the best DC based on the importance of the performance metrics assessed.

An example application studies a hypothetical urban area to be served by one depot where there are 20 customer locations and five candidate distribution center sites. The solution presented indicates that 1) some of the sites are dominated by others and 2) there is a tradeoff among the non-dominated choices in terms of the performance metrics considered.

Finally, section 6.0 presents a summary of the research findings and indicates ways in which the work conducted can be advanced further by future efforts.

Several insights have been derived from this effort. They will have an impact on the way in which freight reliability analyses are performed in the future.

- **On-Time Windows**. Shippers and receivers have expectations about when shipments are going to depart and when they are going to arrive. Hence freight reliability is not about travel times per se, or even the variance in those travel times. Rather, it is about whether shipments arrive and/or depart during these on-time windows.
- **Arrival Times**. For freight, it is always the arrival times and frequently the departure times that matter. Thus, the question becomes: when must the truck leave the depot so that the shipment will be delivered on-time? It is not when it will arrive given a departure time, as is often the case.
- **Searches Backward in Time**. Searches for best paths and departure times must often progress backwards in time, not forwards because the objective is to arrive at a specific time, not estimate when the arrival will occur given a departure time.
- **Doubly-Constrained Path Choices**. Scheduled carriers, like trucking firms, often face doubly-constrained path choice decisions. On-time performance is measured in terms of both departure and arrival events, separately and in combination. This means that the on-time performance in terms of both in terms of departure and arrival events.
- **Measurement Locations**. Timestamps collected at network nodes (intersections or interchanges) tend to be ambiguous. The travel times computed from pairs of sensors located at the nodes include unknown variability due to turning movements at both the upstream and downstream intersections. Collecting timestamps at the midpoints of the links is much better. These midpoints are not locations where processing takes place and/or delays occur. And all pairwise combinations of these adjacent timestamps are then related to vehicles that have followed the same intervening path.
- **Travel Time and Rate Distributions**. There is still a critical need to compute the distributions of travel times (and travel rates) of network segments and paths. Section 2 presented three methods for doing this. They are all useful and apply to different situations depending on the data available.
- **Vehicle Routing and Scheduling**. Section 4 presented two methods for considering reliability in developing solutions to vehicle routing and scheduling problems. The first uses a simulation-based heuristic to search for good assignments of customer visits to trucks. The second solves a bi-objective math programming problem that maximizes the on-time performance and minimizes the cost of the service provided.
- **Site Choice**. A hypothetical case study analysis showed that differences in travel time reliability can exist among candidate sites and hence, this aspect of DC choice should be incorporated into a multi-objective assessment of potential sites. The shows that some DCs are better choices than others in that they dominate poorer performing sites in terms of the combinations of average and maximum travel times that the best sites provide.

Much future work can be carried out based on the analyses conducted so far. Some important examples of these efforts are as follows:

- **Real-World Tests**. As is often the case, the methodological advances presented here have been tested using a blend of empirical data and hypothetical situations. One natural extension for future work is to test these methods based on datasets that are more representative and reflective of real-world conditions.
- **On-Time Windows**. Another opportunity for future work is the further examination of on-time windows and the implication this idea has for freight reliability assessment. Surveys of shippers, receivers, and carriers would be helpful to double-check that on-time windows are, indeed, the current way that reliability performance is assessed.

- ***Backward Searches***. The report asserts that the search for best paths and departure times is one that involves an analysis backward in time. The veracity of this assertion could be checked through a survey of shippers, receivers, and carriers. Also, assuming it is correct, a study of the planning and scheduling practices of these three stakeholders would show how this assessment is presently carried out, what assumptions are made, and what data are employed.
- ***Doubly-Constrained Path Choices***. There is an assertion that path choices for truck-based freight shipments are doubly-constrained, having on-time windows for both departure and arrival. Inquiries with carriers, shippers, and receivers would provide indications as to whether this assertion was true or not; or the extent to which it is true.
- ***Travel Time and Rate Distributions***. Three methods for estimating route-level travel time and travel rate distributions have been presented. Analysis based on more real-world data, especially from trucking firms, would indicate how well these methods work and how they can be enhanced.
- ***Vehicle Routing and Scheduling***. Vehicle routing and scheduling will continue to be a critical element of the reliability analysis. This report presented two methods for developing a vehicle routing and scheduling plan that maximizes the reliability of the service provided.
- ***Site Selection***. The site selection procedure could be enhanced most significantly by linking it to vehicle routing and scheduling. That is, treat the assessment of the reliability of the sites as an evaluation of the best reliability that can be provided by a fleet of trucks given the choice of a specific site.

  Combinations. All these possible extensions to the current work could be done separately or in combinations. For example, the SVRP ideas could be explored in conjunction with the enhancements to the site selection procedures. The enhancements to path characterization could be coupled with SVRP so that the SVRP solutions are more tightly and defensibly tied to the characterization of the performance of the network over which the trips take place. Determining which combinations to select and which options to pursue will depend upon the data available, the interests of the research funding agencies that are involved, and the demands of the stakeholders whose needs are being addressed.

1.0   INTRODUCTION

1.1   RESEARCH FOCUS

The analysis of efficiency and reliability for freight transportation is the motivation of this research report. Although this investigative endeavor focuses on advancing the methodological frontier for truck-related shipments, the developed tools can be applied to other modes and multi-modal systems. These efforts are important as they seek to explore the economic value of the consumption of resources associated with freight activity. Unreliable transport raises costs and diverts scarce factors of production away from other, critically important societal activities. It interferes with the efficiency of the supply chain and increases both monetary and time-related costs and resource requirements (e.g., increased in-process inventory, extra trucks).

As might be expected, the reliability of freight transport has always been important. Military leaders, government officials, and people in general have always been concerned about on-time delivery of goods and services. For example, the Carthaginian army lost to the Romans because of a lack of supplies and manpower in 249 BC and the Germans repeatedly had to revise their battle plans because of a lack of fuel in World War II. Much research has been conducted on investigating transportation logistics throughout history and consequences of unreliable or delayed shipments. Articles on this topic include Stephens (1989), Sakul (2010), Fusaro (2015), and Ronsee and Rayp (2016). Stephens (1989) examined the relationship between the railroads and time awareness in 19th Century America. Each chief engineer of a railroad faced enormous uncertainty for the schedules because of rolling-stock failure, accident, track obstructions, and weather. However, the railroads faced pressure to increase punctuality from the U.S. Post Office and as the result from fatal incidents. The author illustrated the various steps that the railroad industry took to ensure reliable travel times. Such steps included the use of time service based on telegraphed signals and strict scheduled operations. Sakul (2010) investigated the role of military transportation in the Mediterranean Sea trade environment during the War of the Second Coalition. The author examined the supply operation that existed including the military's need to address delayed orders, cancelled orders, and cash payments. Due to the wartime conditions, these operations also faced diplomatic challenges. Sakul provides one such example of military logistics historically documenting logistics in the transportation domain. Fusaro (2015) explored the logistical support that English and Dutch mercantile vessels provided to Venetian naval campaigns during the War on Candia. The use of these ships allowed for the movement of goods and resources under intense deadlines and conditions.. Ronsee and Rayp (2016) studied the impact that advertising by shopkeepers had on the shipping patterns of eighteenth century Ghent. Beyond determining the significant impact that commercial ads had on the shipping industry, the authors subsequently characterized the shipping data in Ghent. As a conclusion, Ronsee and Rayp determined models to assess the impact that the advertisements and the existing shipping market had in Ghent.

Today, the freight industry is still concerned with reliability. To be competitive, companies need to remove inefficiencies in their production functions. Both late and early shipments are problematic. The industry's emphasis on just-in-time manufacturing has squeezed buffer stock out of the logistics supply chain. It has also raised the risk of stock-outs. Because storage space has

been reduced as well, early arrivals are problematic. If reliability suffers, all participants in the supply chain must make extra asset investments to buffer the process and ensure that delivery schedules are met. From a societal perspective, the cost of producing the goods and services increases. Extra scarce resources must be devoted to freight-related activities to make the economic system work.

From a research standpoint, reliability has been of interest for at least half a century. One of the earliest textbooks on reliability is that of Shooman (1968). He focused on the impacts of stochasticity on routing, logistics management, and travel time reliability within specific modes. Shooman applies the concept of a hazard rate from reliability physics to determine the reliability of a system. This allows for a decreased time required to run tests and the use of actual conditions of the system. Shooman then suggests the use of stress-strength models or stress, strength and time models to find the probability that a system will fail. Using these models, a hazard function is used to depict the reliability of the system. This puts into context work on reliability that has occurred more recently, efforts that aims to find solutions for multiple-vehicle routing problems, multi-modal logistics networks, and optimal mode choice and path selection.

This research focuses on creating analytical tools that can help shippers, carriers, and government agencies improve the reliability of the freight system. The tools are aimed at improving reliability-related decision-making in terms of network investments and operating plans, vehicle routing and scheduling, and site selection. In each case, the objective is to identify system treatments that can alter the statistical distribution of travel times and/or delivery times so that variances are reduced and target values and/or windows are achieved with a higher probability. For freight shipments, this means achieving on-time arrivals (OTAs) and on-time departures (OTDs) by making resource investments and operational decisions that maximize the likelihood of arrivals and departures during the on-time windows (OTWs). For servicing customers, this means making fleet investment decisions and routing and scheduling decisions that maximize the OTAs and OTDs. For siting decisions, this means selecting locations that maximize the likelihood of servicing customers within their OTWs while at the same time minimizing supply chain costs.

1.2 GAINING A PERSPECTIVE

Before describing the results of the research, it seems useful to present a perspective about how reliability is perceived in the context of this research. This will help the reader understand the reasoning behind work presented. Moreover, since freight is the focus, the report focuses on the reliability of package shipments and truck movements rather than personal trips and auto movements.

Leemis (2009) has offered the following definition of reliability: "The reliability of an item is the probability that it will adequately perform its specified purpose for a specified period under specified environmental conditions." This is suitable for assessing the mean time to failure for a physical device, but it does not pertain particularly well to the reliability of transport services.

A slight shift in perception makes Leemis' definition better fit with freight reliability. A trip termination or customer visit can be regarded as the "device" and the absolute difference between the on-time arrival window (AW) and the actual time of arrival (ATA) can be the metric monitored. A trip termination or customer visit is considered reliable if the ATA is within the AW. That is, if

t_a is the arrival time, a_b is the beginning of the AW, and a_e is the end, then $d_a = \max(t_a - a_b, t_a - a_e, 0)$ is the deviation from being on-time. If $d_a = 0$, then the delivery was on time, and if $d_a > 0$ then it was not. Evidence that this is the perspective of the freight industry is illustrated by Knowler (2016), who highlights the recent performance of the container shipment industry. Success occurs if an arrival or a customer visit begins within the applicable AW. Actions that increase the likelihood of that occurring increase the reliability of the service.

The same comments pertain to the on-time departure window (DW). If t_d is the departure time, d_b is the beginning of the AW, and d_e is the end, then $d_d = \max(d_b - t_d, t_d - d_e, 0)$ is the deviation from being on-time. If $d_d = 0$, then the departure was on-time, and if $d_d > 0$ then it was not. Hence, freight travel time reliability is about consistency in on-time arrivals and departures, not travel times per se. The focus needs to be on probability density functions (PDFs) and cumulative distribution functions (CDFs) for d_a and d_d.

This notion stresses the fact shorter travel times are not the primary objective, although travel times still are of great importance. Instead, the focus should be on about reducing the variance, managing the mean value, and a making intelligent tradeoffs between these two and possibly other measures that affect the total cost. The benefits of smaller average travel times are that departure times can be employed; more customers can be visited within a given span of time; more customers can be visited with a given fleet size; and more customers can be serviced for a given siting decision. This also results in reduced in-transit inventory costs and fleet size requirements. However, the mean should not be reduced at the expense of an increased variance.

Thus, a carrier's objective is to arrive and depart within the on-time windows (OTWs) a very high percentage of the time. If a carrier has high percentages for both their arrivals and departures, their service will be perceived as being reliably on-time rather than late nor early. Regarding the AW, if the truck arrives early and must wait, less carrier resources could have been used or extant resources could have been better deployed. If it arrives late, it means the customer's expectations have not been met. Regarding the DW, the same two thoughts still pertain. If the truck departs early, then the same resource issues exist. And if it departs late, customer resources (dock time and space) have been unnecessarily tied up.

As is probably obvious, achieving on-time departures and arrivals 100% the time is impossible. Systems do not operate deterministically. Travel times vary. Network conditions change. Loading and unloading times vary. Hence, carrier performance must be measured in terms of the probability that on-time events occur.

When only an AW exists, the prior DW is implicitly identified by tracing backward for paths and their travel time distributions. The search needs to be done backwards because the arrival time window is the constraint. The "best" path maximizes the likelihood that an OTA will occur. There may also be a tradeoff against travel cost or travel time. This may result in the selection of the path with the latest possible departure time window that still meets the objective of achieving an acceptable on-time performance. Otherwise, pathologically, leaving extremely early is the optimal solution as is the idea of arriving very early. Then, an OTA is almost 100% assured.

When only a DW pertains, which is rare for freight shipments, the objective is to find a path that minimizes travel cost. The path needs to maximize the likelihood that an on-time departure will occur and that the travel cost is minimized.

When both a DW and an AW exist, the path search is doubly constrained. A path is then optimal if it maximizes both the probability of an OTA and an OTD. Similarly, multi-vehicle routing and scheduling decisions are optimal if they maximize the likelihood that the trucks will arrive and depart within all AWs and DWs. Shifting shipments among trucks and increasing the fleet size can both have a positive impact on improving performance. Of course, and importantly, the location of the depot also has an impact. If that location is changed, there is the potential to improve the on-time performance.

Scheduled carriers, like railroads, bus companies, and airlines face these doubly-constrained conditions. Trucking firms do as well. They measure on-time performance in terms of both departure and arrival events, separately and in combination. Shipments (and vehicle moves) are deemed to be "on-time" if they both depart during the DWs and arrive within the AWs. A joint density function can be used to track this performance. The objective is to find paths and vehicle tours that conjunctively maximize on-time performance.

Minimizing variance is a similar, but different thought. Minimizing the variance helps reliability, but it must be done in combination with optimizing the travel time and departure time. In addition, there may be no value in minimizing the variance beyond a given point if the desired on-time performance is achieved. In effect, the better thought is to control the shape of the travel time distribution, either viewing it as a PDF, or better yet as a CDF, so that a sufficient percentage of the distribution lies within the AW or DW, or both.

Maximizing the on-time performance is sensitive to the operating conditions under which the trips are made. Different departure times and paths may be considered best for different operating conditions. For example, different routes may be preferable when the weather is inclement, network maintenance is underway, or the network is heavily congested. Consequently, using the right travel time distributions for each operating condition is very important.

A graphical way to think about this is to talk about desired times of arrival (DTA) and actual times of arrival (ATA) as shown in Figure 1.1. If the ATA is within the AW, then an OTA has occurred. The freight transportation system's reliability can be measured by the percentage of trips that have ATAs within their OTAWs.

If it is possible to observe the shipments, as carriers and customers can do, with the AWs being known, then the reliability of a service can be assessed completely in the manner described above. Using this information, customers can assess the percentage of all ATAs that fall within their OTWs. Carriers can do this as well. Customers can nominally adjust their AWs so that the on-time performance is improved. Carriers can adjust their departure times, or add slack to the trip time, so that the probability of arriving within the AW is maximized.
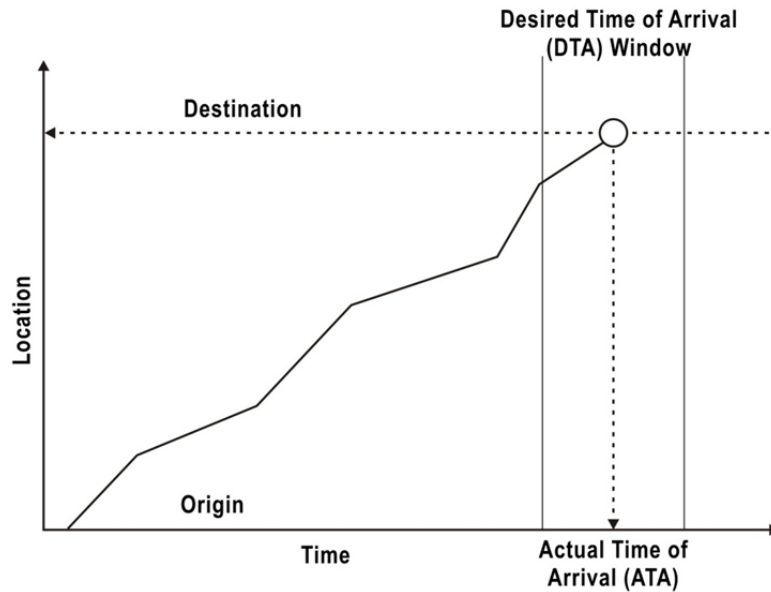
Figure 1.1: Concepts of desired and actual times of arrival

These thoughts have an interpretation that is based in utility theory, as described by Hansson (1994). Each trip has a disutility that reflects the "cost" of making the trip. That cost includes the travel time, tolls, other expenses; and most importantly, here, the "cost" of being either late or early. That delivery-related cost is zero if the ATA is inside the AW. And it becomes non-zero if the ATA is outside the AW either before or after. Moreover, the cost of being late may be different from that associated with being early. This is shown by Figure 1.2 in that the slope of the cost curve indicates the per-unit-time penalties involved. The steeper the slope, the costlier it is to be late or early. In the aggregate, the on-time costs of the trips can be summed to assess the "societal cost" of the unreliability of the system.

If the focus is on when the shipment arrives, not when it departs, then the question is: when should the shipment leave to maximize the probability of arriving within the AW? Or strike a balance between travel time and on-time performance, or minimize the total generalized cost.
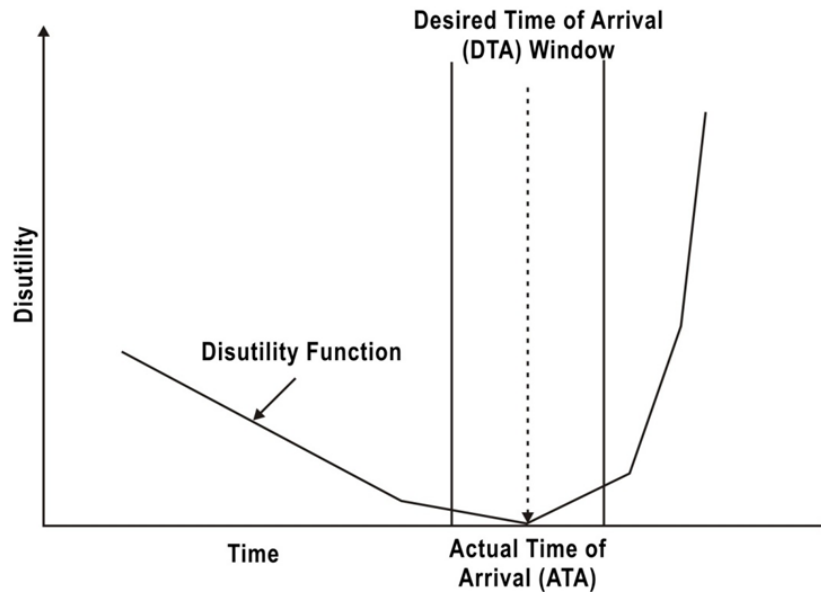
Figure 1.2: Disutility function to characterize desired and actual times of arrival

As Figure 1.3 below shows, reliability of the service is maximized if t^* is employed. t^* maximizes the percent of the travel time distribution within the AW. But even then, there is a non-zero probability that the shipment will arrive outside the AW, either late or early. It is not necessarily true that all the arrival times from the 0th percentile to the 100th will be within the AW.

Maximizing AW performance requires decision-maker actions. The packages cannot move by themselves; they cannot make decisions about when to depart or what route to follow or what truck to use. The shipments must be managed, handled and transported by people. People must take actions. Hence, the objective in this research has been to develop tools that help those decision makers. The tools should facilitate their decision making.

From a public agency perspective, which is part of the focus of this research, monitoring ATAs and arrivals within AWs are not reasonable thoughts. Public agencies do not have access to this information. But they can monitor and "control" the reliability of their networks. They can manage the means and variances of the travel times or set targets for the percentage of time that travel rate targets are achieved (e.g., the percentage of time that the speed is 45mph or greater). This is what the MAP-21 regulations are asking states to do. They are expected to monitor the travel rate distributions for TMC segments and make investment and operational changes that improve the percentage of time that those targets are achieved.
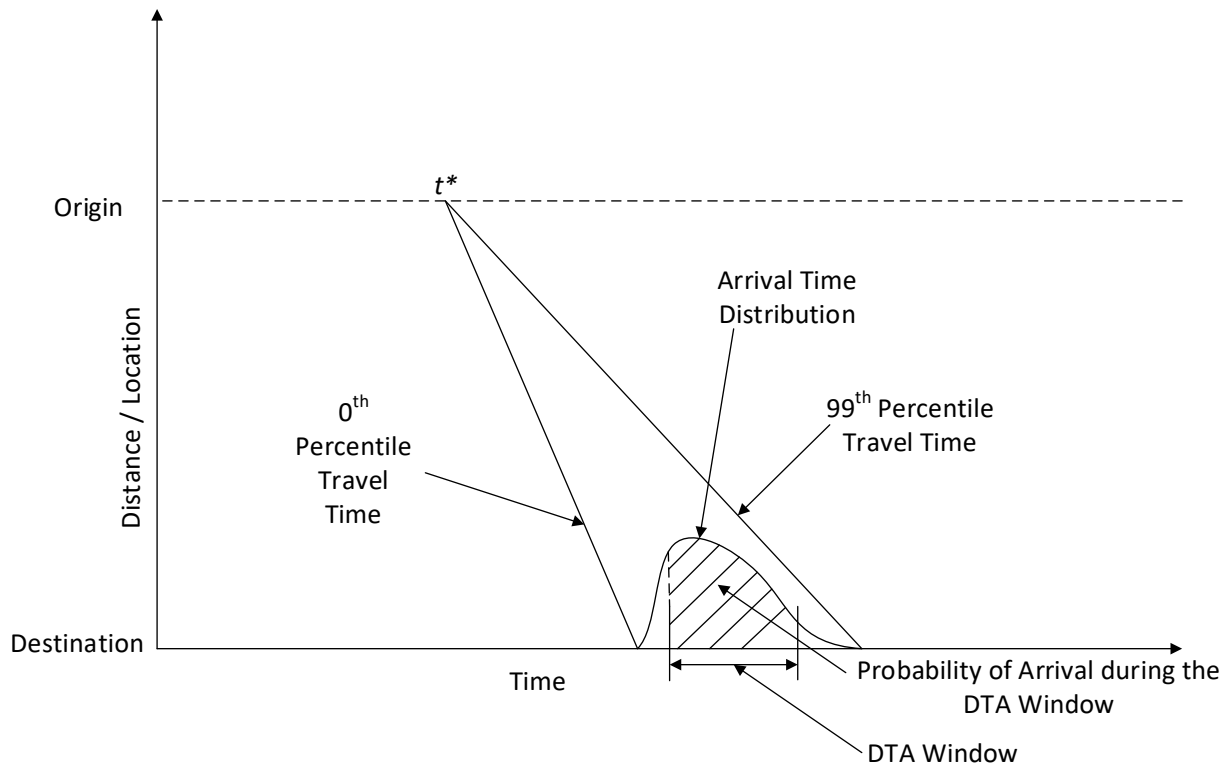
Figure 1.3: Maximizing the probability of arriving during the OTW

As shown in Figure 1.4, the target distribution can vary by network condition, say having different targets of performance for various conditions.
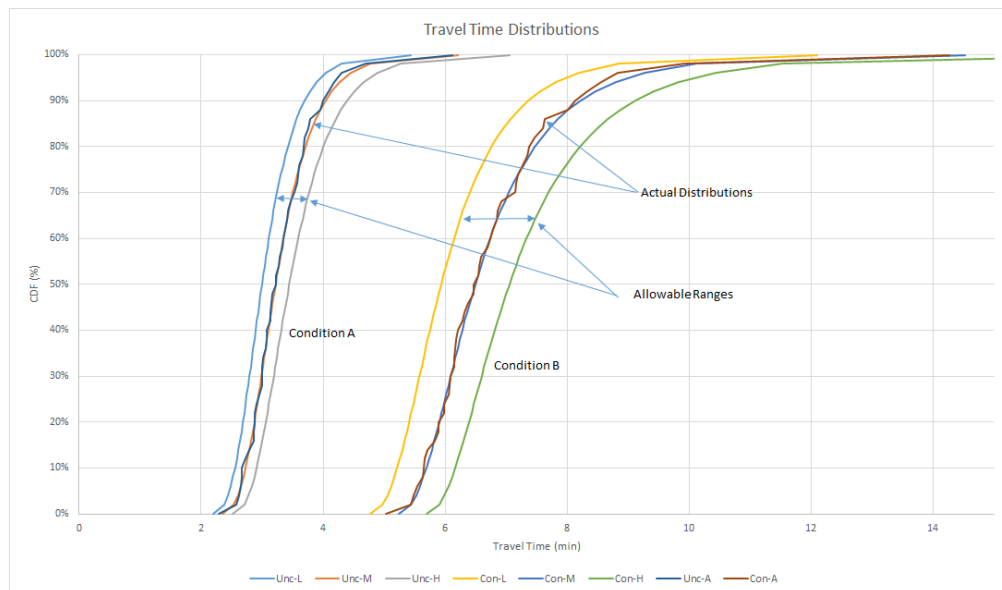
Figure 1.4: Travel time distribution targets by operating condition

If the travel time (or rate) distributions that are observed in the field are close to or better than the policy-based distributions, then the agency can claim that it is providing acceptable service. If not, it ought to take actions that bring the travel time distributions back in line with the targets.

1.3 IS THERE VALUE?

There are many questions surrounding the importance of travel time reliability for freight movements. Does it matter? What is the value of reliability? Does it have a "cost"? How can it be assessed? How does the importance of reliability compare with that of travel time itself? If reliability is valued, then how should it be factored into decisions about departure times, path choice, location decision making, network investment, mode choice, service selection, carrier selection, etc.?

Many transport economists have reached the conclusion that reliability is important. It does have a cost. Experts see that efficiency in the supply chain is critical and the required efficiency is hampered by unreliability and unpredictability. So, if the supply chain is not operating on the efficient frontier because of uncertainty, performance can be improved if uncertainty is reduced, especially in terms of reduced costs. And in return, service quality may improve as well. The value of these improvements depends on what resources become available and what alternate uses for those resources exist. In principle, freed-up resources can be used elsewhere by society and reduce societal costs overall.

The absence of reliability motivates the creation of buffers, safety stock, and distribution channel redundancy to ensure that the supply chain is reliable. In addition, shipments are sent earlier and safety stock is held near the point of demand to ensure that arrivals are not late. Shippers and producers increase the pipeline inventory to ensure that deliveries are always on-time. Trucks are dispatched from depot, earlier, and with fewer stops to make to ensure that pick-ups and deliveries are made on time. If the reliability were higher, less time would be wasted in transit, less inventory would be carried in transit, and more stops could be scheduled to improve efficiency in the system.

Researchers began to examine the value of reliability in the context of freight operations in the late 1960's. The railroad industry elected to invest in a technology called KarTrak, an AVI-like system that made it possible to trace every car moved in interchange service. Cars could be tracked from shipper to consignee and could be checked to ensure that the cars were correctly routed through classification yards. This led to the freight car utilization project conducted by MIT, as described by Lang (1970), that aimed to make improvements to the railroad system that would allow it to overcome delays and improve reliability.

KarTrak and advancing computer technologies made it possible to monitor the reliability of car trips. FRA sponsored the creation of the Freight Car Scheduling (FCS) system (see Shamberger, 1975). FCS used train schedules and blocking plans to make train "reservations" for cars from origin to destination. The plans then used data from the KarTrak system to track actual trips against intended trips. The FCS system was first put in service on the Missouri Pacific (see Sines, 1972). Sierleja, Pipas, and List (1981) examined the benefits that FCS produced. List and Bongaardt (1981) estimated the benefits that FCS would produce for Conrail. Moreover, List, Buchan,

Bongaardt, and Pipas (1981) assessed the benefits from FCS for railroads in New England. It was clear that FCS could help improve on-time performance.

Other work was concentrated on determining the impact of better reliability. Whybark (1974) examined the impact of variations in travel times on supply chain management: reorder points, and order quantities. The objective of this research was to identify transportation alternatives that minimized total transportation and inventory costs for a receiving facility. He developed an effective heuristic procedure that was evaluated over a broad range of conditions. Bevilacqua (1978) examined the relationship between energy conservation and different modes of freight for delivery services. Alexander (1978) reviewed actions taken by ports to better coordinate rail and steamship operations given aspects of the rail travel such as length of haul, speed and reliability of the journey.

Van Der Mede, Palm, and Flikkema (1996) asserted that travel time variability should be a "new" service quality indicator. They measured variations in travel time and the subjective reaction of interviewees to travel time variability. This was done for trips by cars and trucks from door-to-door. The data collection techniques included trip diaries for drivers, black-box data from trucks, and questionnaires. Their finding was that reliability did, indeed, have value. Wigan *et al.* (2000) reported values of travel time and reliability for long-haul and metropolitan freight services. Lam and Small (2001) reported values of time and reliability obtained from a value pricing experiment.

More recent studies include Fowkes and Whiteing (2006), Zamparini and Reggiani (2007, 2010), Nunez *et al.* (2008), de Jong *et al.* (2009), and Fosgerau and Karlstrom (2010). Fowkes and Whiteing (2006) attempted to determine monetary valuations of time for nine commodity groups. The authors determined that no single value can be used in the freight industry, but specific reliability ratios can be used. Zamparini and Reggiani (2007, 2010) explored the value of freight travel time savings. The authors recommended a meta-analytical estimation of this parameter given contextual factors. These contextual factors included the geographic location, mode of transportation, and GDP. Nunez *et al.* (2008) compared the value of time and reliability of rail and road freight transportation in the cross-Alpine and cross-Pyrennean transport domain. The authors used both a classic logit model and a mixed logit model to show the socioeconomic evaluation of these parameters. de Jong *et al.* (2009) proposed values of reliability to measure the benefit of Dutch infrastructure projects. The authors determined that a reliability ratio that can be used to in cost-benefit analysis for passenger travelers and freight transportation. Fosgerau and Karlstrom (2010) investigated the impact that the distribution of travel time has on reliability which they determined to be the result of the standard deviation and mean lateness factor. This allows for the approximation of the value of reliability for any given travel time distribution.

Weigman, Hekkert, and Langstraat (2007) asserted that reliability and costs are the most important aspects of service quality in the intermodal market. For terminal operators, both reliability and flexibility were found to be more important than they were for customers. This suggested that terminal operators could reduce their focus on these aspects without reducing total perceived quality by customers. Moreover, less focus from the terminal operators on flexibility and reliability would offer opportunities for increased attention on other quality aspects (e.g. costs). For the customers, costs and total quality of the service were determined to be more important. Other

quality aspects also matter, but were relatively less important. Moreover, the differences among these less important quality aspects were small.

Researchers also explored the value of reliability in the context of the role it plays in various kinds of freight movement decision-making. Poole (2007) explored its role in the context of truck-only toll lanes. Figliozzi and Zhang (2010) examined its impacts on cost. Ozkaya *et al.* (2010) studied it from the perspective of freight rates in the less-than-truckload sector. McLeod (2012) considered the role it should play as a performance measure for evaluating freeway systems.

Since reliability has been shown to have value from a decision-making standpoint, it should be possible to demonstrate that it influences path and departure time choice decisions. If shippers see a value in reliability, they must select path *A* over path *B* or mode or carrier *A* over mode or carrier *B*, ceteris paribus, if the reliability of *A* is better than *B*. Meixell and Norbis (2008) provide an review of the literature in this area. In their review, they repeatedly observe that reliability is an important factor in the choice of modes and even carriers within modes (especially for trucks).

Swan and Tyworth (2001) looked at the issue from the carrier's perspective. They focused on customer retention, asserting that the US railroads were losing the most profitable share of their business by providing unreliable service. They argued that by choosing to focus on reducing costs, rather than providing better service, they were forcing their customers to shift to other modes, notably to truck services. They asserted that railroads should provide better service and recapture the costs by charging higher rates.

Bontekoning and Priemus (2004) made a similar assertion for intermodal services. They said that the main growth potential for intermodal was in markets for flows that demand speed, reliability and flexibility. They further said that innovations in service offerings will produce a breakthrough in modal split and allow the use of the mode to expand.

On the other hand, Shinghal and Fowkes (2002) presented the results of an empirical study of mode choice for mode choice in the Delhi to Bombay corridor. Travel time, reliability, and service frequency are all found to be important. Service frequency is the most important attribute. The importance of reliability was generally lower than the authors expected. The reliability of transit times was found to be very important for exporters and the auto parts sector because it can affect the production process.

Danielis, Marcucci, and Rotaris (2005) conducted a formal study of freight mode preferences among logistics managers in two regions of Italy. Four attributes were employed to characterize each hypothetical option: cost, time, reliability and damage/safety. Two estimates were obtained of each attribute were obtained: (1) the utility associated with each level of the same attribute, and (2) the attribute utility revealed by an ordered probability model. Both estimates indicated, on average, a strong preference for attributes of quality (time, reliability and safety) over cost. They felt this indicated that modal shift policies needed to focus more on the quality aspects of the modes rather than just their costs.

Fowkes (2007) considered the concepts of freight value of time and reliability in the context of shipments in the UK. He presents findings for nine commodity groups as well as the group overall.

Care was taken in developing the results since the estimated valuation of one attribute can vary depending on the presence of a related variable. The main empirical finding was that, when respondents ignored driver and vehicle costs, for many commodities the valuations of improvements in journey time and its variability were negligible. However, shippers of some commodities did exhibit a willingness to pay for improvements, and occasionally at a great expense.

Fries (2008) reported the results of an effort to develop a freight demand model that could be a comprehensive tool for freight demand forecasting in Switzerland. Fries presented the methodology and results of the project focusing on the development of modal split functions that represent the shippers' demand elasticities. The core part of this project consisted of preparing and executing a survey among shippers and freight forwarders. Stated preference experiments based on revealed preference data were conducted within the framework of the survey to collect the data necessary for the estimation of modal split functions for different commodity groups. Interestingly, reliability was ranked equal to or even higher than transport cost in several commodity groups. Moreover, travel time was generally less important than reliability.

Grosso and Montiero (2008) did a similar study in Italy. They were interested in seeing what factors influenced the decision about choosing a port. A questionnaire was sent to about 30 companies, including shipping companies, freight forwarders and shippers, currently operating in the port of Genoa. They found that port service reliability was among the criteria used.

In the same year, Train and Wilson (2008) completed a study on grain shippers in the upper Mississippi River valley. They sent survey forms to 2,000 shippers and received responses from 480. The survey presented changes in rates, transit times, and reliability, and the respondents were asked to state how their annual volumes shipped by barge as opposed to truck or rail would change given that all other factors remained the same. The basic finding was that, as might be expected, larger declines in reliability increased the likelihood that firms would adjust the volume shipped by barge. For example, if the percentage change in reliability was less than 10%, the elasticity for those shippers that made a change was 2.417. That is, for them, a 1% decrease in reliability would result in a 2.417% decrease in shipment volume. For the survey respondents, including those that did not make a change, the elasticity was 0.619. That is, a 1% decrease in service reliability would produce a 0.619% decrease in the use of barge. In comparison, these same elasticities for a change in rates were -1.407 and -0.075, and for a change in transit time, the elasticities were -1.841 and -0.310 respectively. In these latter two cases, the elasticities are negative because increases in either rates or travel times would result in a decrease in the use of barge. The absolute values are what are important for comparison purposes, and those values show that the sensitivity to reliability was the highest among these top attributes.

Brooks *et al.* (2012) examined the Australian domestic freight transport market with a focus on the decision-making process by which cargo interests and their agents make mode choice decisions between land-based transport and coastal shipping. While their ultimate interest lay in seeing if short-sea shipping could provide a reduced carbon footprint to truck and rail, they nonetheless looked at shipper sensitivities to various service attributes including reliability. The attributes examined were: service frequency, cost (price), transit time, freight distance, direction (headhaul/backhaul), and reliability, measured both by arrival within the delivery window and

delay. The authors concluded that shippers would be willing to pay significant amounts for improvements in the on-time reliability of rail, road, and short-sea shipping.

Spurred by the advent of deregulation in the trucking industry, Bardi, Bagchi, and Raghunathan (1989) conducted a survey of 1,000 transportation shippers randomly selected from the Council of Logistics Management membership directory. Twenty-nine percent of those surveyed responded. Reliability was ranked the first out of 18 criteria by which a carrier could be selected. The next five criteria, in rank order, were door-to-door rates or costs, door-to-door transit time, rate negotiation flexibility, financial stability of the carrier, and equipment availability. It is clear from their research that reliability was important.

Crum and Allen (1997) also examined reliability as a factor in selecting one carrier over another. They reported the results of two surveys, one conducted in 1990 data and another in 1996. Based on the 1990 results, pick-up and delivery reliability was the top ranked criterion and transit time reliability was the second. In 1996, the order was reversed, but the two top measures were still the same.

Kent and Parker (1999) conducted a similar study like Bardi, Bagchi, and Raghunathan focused on the shippers of international containers. They surveyed export shippers, import shippers, and containerized transportation companies and asked for rank order evaluations of the same 18 criteria used by Bardi, Bagchi, and Raghunathan. The most important service attribute again proved to be transit time reliability/consistency. The next five attributes, in descending rank order, were equipment availability, service frequency, rate changes, and operating personnel. Transit time was sixth.

The basic conclusion that these researchers reached is that reliability is important and improving reliability is significant. Higher reliability produces monetary benefits that are important to customers and carriers alike. Therefore, an examination of ways to predict and improve reliability of freight services has value.

1.4   REPORT OVERVIEW

The remainder of the report is divided into five sections. Section 2 focuses on characterizing the reliability of single routes or segments. Section 3 deals with departure time and path choice for specific origin-to-destination (OD) trips. Section 4 focuses on reliability-based vehicle routing and scheduling decisions where multiple stops and multiple vehicles are involved, as in pick-up and delivery schedules for fleets of vehicles assigned to a single terminal. Section 5 is devoted to location choice comparisons for distribution facilities and depots. Finally, Section 6 provides a summary of the material that has been presented and describes work that could be done to further advance the frontier.

2.0   ASSESSING SEGMENT AND ROUTE RELIABILITY

The reliability of travel time on segments and routes (or paths) is a fundamental building block upon which all the of the other reliability outcomes depend. If segment and route travel times are reliable, or more precisely, if their reliability is understood, then decision support tools can help improve travel time reliability.

2.1   BASIC IDEAS ABOUT TRAVEL TIMES

Trips are comprised of 1) transport across links and 2) processing at nodes. Sometimes the processing is significant. Such is often the case with freight shipments. In Figure 2.1, the nodes are dots surrounded by boxes. Each node has a letter designation (A through H). Connections between the nodes are shown as links. The word "link" refers to these connections when no direction is implied. The word arc implies directionality as in the arc DB in which originates at D and terminates at B. A shipment leaves a node as it passes through the box on the departing arc. It arrives that a node as it passes through the box on the arriving arc. Arc travel times arise between the boxes on the arc. Processing times occur between the arriving and departing boxes at the node.
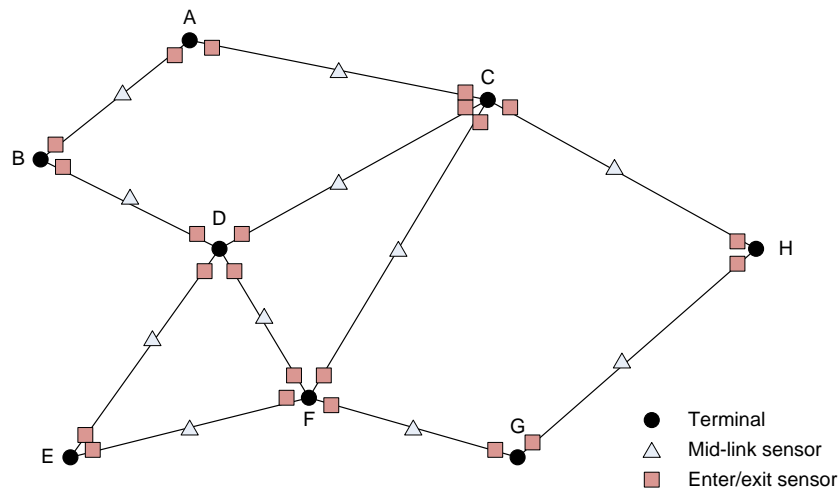
Figure 2.1: A hypothetical network and possible monitoring locations

The network diagram can represent two different situations. The first is as a service network maintained by a carrier. The nodes are depots and/or transshipment locations. The links are the paths between them. In the second, it is a highway network. The nodes are intersections or interchanges and the links are the over-the-road paths between them. Both perceptions are used in this report.

In the case of a trip across a carrier network, there are arc transit times and nodal processing times. If the trip is from B to H, there is an initial processing time at B, a travel time on arc BD, a processing time at D, a travel time on arc DC, a processing time at C, a travel time on arc CH, and a final processing time at H.

19

For a trip across a highway network, there are arc travel times and nodal delays. For example, a truck traveling from B to H has an initial delay at B, a travel time from B to D, delay at D, a travel time from D to C, a delay at C, a travel time from C to H, and then a delivery time at H.

The objective in this section is to understand and improve the reliability of these overall trip times. The challenge is to determine what causes the travel times on the arcs, and processing times at the nodes; and then determine what actions need to be taken to reduce the variability of those times. In a sense, this means creating a function that can predict the trip times and then adjust the parameter values and variable relationships so that the variation in travel times is reduced.

Insofar as the arcs are concerned, the lengths are important. Longer arcs require more time to traverse. And the travel rates (inverse speeds`) are also important. The travel rates are affected by congestion, weather, incidents, maintenance work, etc. These "operating environment" variables are important because they affect the travel rates that can be achieved. These variables describe the "operating condition."

For the processing times at the nodes, the type of handling is important. At node D, for example, there would be a processing time associated receiving the shipment when it arrives on arc BD, then another for getting it prepared for departure on arc DC, and then a third for loading it for departure. These times are likely to vary depending on the inbound-outbound combination. These handling times may also depend upon when the trucks arrive and depart, how the shipment must be unloaded, how congested the terminal is at that point in time, and the resources available (e.g., forklifts and people). From a highway perspective, trucks traveling from B to H see delays at D and C. If node D is a freeway interchange, then trucks see the time for traveling on a ramp from arc BD to arc DC If it is an at-grade intersection or an interchange that involves intersections, then the trucks experience the delay of making a left turn from arc BD to arc DC, for example.

Other independent variables include the shipment class. Premium shipments are handled more expeditiously than standard ones. Another variable is the type of commodity being transported (e.g., HazMat versus cardboard.) It may be useful to categorize the shipments by class and build separate causal models for each rather than considering these differences as independent variables. If the models turn out to be the same or very similar, then the class categories can be collapsed and combined.

To build a model for carrier shipments, timestamps are needed for shipments and vehicles as they arrive and depart from customer locations and processing nodes. For example, for shipments from B to G for example, timestamps are needed for the pick-up and receiving time at B, the travel time on BD, the processing time at D, the travel time on DF, the processing time at F, the travel time on FG, and the delivery time at G. Data are also required for the physical condition of the network while the trip was underway and the environment in which the system was operating. These pieces of information indicate what the operating condition was when the travel time was observed.

In the case carriers, reliability analysis is relatively easy. Times of arrival and departure are typically sensed. This means depot-to-depot travel times can be computed and in-depot handling times can be observed. From the carrier's perspective, monitoring these times is critical for quality control and cost management.

For trips on the highway network, reliability analysis is more difficult. If an instrumentation strategy like that used by the carriers is employed, then timestamps and vehicle IDs need to be collected upstream and downstream of every node; that is, before the vehicles join the back of queue and then downstream of the intersection. Clearly, this is very sensor intensive.

List *et al.* (2014) recommended collecting data at the midpoints of the links, not at the nodes. They cite two reasons for this. The first is that these midpoints are not locations where processing takes place and/or delays occur. Vehicles are typically moving when they pass these locations, so a clear and meaningful timestamp can be collected. The second is that all pairwise combinations of these adjacent timestamps correspond to vehicles that have followed the same intervening path. They have seen the same processing. For example, trucks going from the midpoint of BD to the midpoint of DC have traversed half the length of arc DB, turned left at D, and then traversed half the length of arc DC. Because the trajectories are the same, the travel times should be similar and predicated on the same sequence of events. If they are different, it is likely that something is different about the servicing received and is causing a decrease in reliability. Moreover, since the processing is the same, it is possible to determine what is causing the variation.

List *et al.* (2014) also suggested using wireless technologies to collect the data (e.g., technologies such as Bluetooth, WiFi, and DSRC) vehicle IDs can be collected along with the timestamps. They also recommend using technologies that allow vehicle IDs to be traced through the network (e.g., toll tags). Technologies like video and radar are less useful because they only provide image information, not vehicle IDs. List *et al.* also proposed collecting information about the operating conditions such as weather, maintenance, incidents, or extraordinarily high demands, because these factors tend to affect the reliability.

List *et al.* (2014) observed that the timestamps collected at the nodes (intersections or interchanges) are problematic. First, they provide ambiguous timestamps. Unless the sensing distances are very short, it is unclear where the vehicle is when the timestamp is obtained. It is only clear that the vehicle was within range of the sensor. Moreover, unless the vehicle's report their paths, it is not possible to tell what turning movement the vehicle was executing. Consequently, since this is true at both the upstream and downstream nodes, the travel times computed from pairs of sensors include unknown variability due to turning movements at the upstream and downstream intersections. Through-through vehicle moves cannot be distinguished from right-lefts, left-rights, right-rights, etc. The variations in these turning movement times completely dominate the causality of the travel time variations. Any variability on the arc is obscured, and variability in the turning movement times at the intersections is impossible to ascertain.

On the other hand, if the sensors are placed mid-block, *all* the timestamp differences for a given pair of sensors reflect similar trajectories. The vehicles do not have to report their trajectories to make the observations clear. The processing (handling) seen by those vehicles is nominally the same. Hence, the variations in their observed travel times must be caused by variations in the turning movement times. And the findings help shed light on what can be done to improve consistency in the travel times.

A comparison of the number of sensors required is also useful. Consider the network shown in Figure 2.2. It has 12 intersections, A through L.
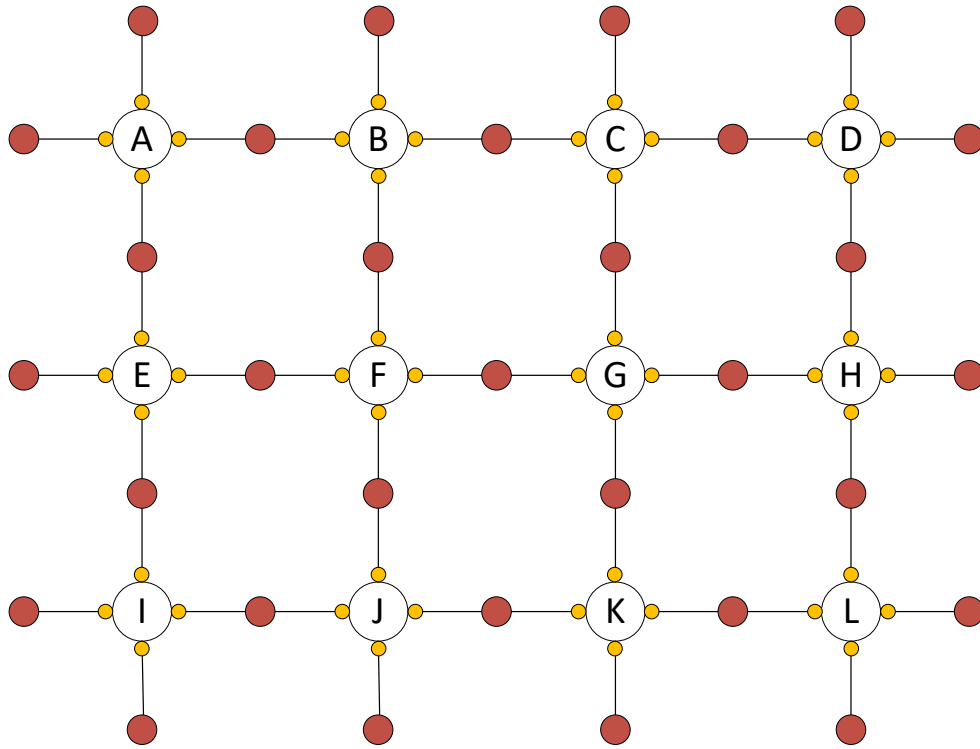
Figure 2.2: Network instrumentation options

For trajectories from F to G, sensors could be placed at the nodes, and only two sensors would be required. But, as indicated above, this is not likely to produce useful information. The travel time observations are confounded by turning movement times that are buried in the observations. Even if the sensors are placed at the dark yellow dots, downstream of F and upstream of G, the data may still be confounded. There would be no way to determine when the timestamps were recorded. Queue delay might be included. This problem is rectified to some degree if all the indicated dark yellow sensors are used, a total of eight. Now the turning movement combinations can be identified and the observations can be classified into turning movement combinations. But using many sensors is expensive. Instead, if the dark red mid-block sensors are used, then only three sensors are required: the ones west of F, between F and G, and east of G. This provides precise and clear information about the travel times for vehicles making through moves. If times for all the turning movement combinations are desired, then the midblock sensors surrounding both F and G would be used, a total of 7. Again, very clear, meaningful information will be obtained.

For the overall network, if sensors are placed at the nodes, 12 are required. But as before, these observations would provide limited information because their travel time measurements would be confounded by the turning movement times at the intervening intersections. If the dark yellow sensors were used, 48 would be required. This has the advantage that the travel time reliability of the intersection-to-intersection links can be observed as well as the turning movement times. But the proximity of the sensors to the intersections means the influence of queueing delay would be unknown. If the mid-block sensors were employed (dark red), then 31 sensors would be required.

This is more than 12, but less than 48, and the travel times from one sensor to the next would be very meaningful because the observations would be based on identical trajectories.

Use of the mid-link monitoring locations creates virtual links. These links can be called "segments" to distinguish them from the links in the physical network. The rest of this section focuses on characterizing the travel time distributions for these segments, not the physical network links. Route-level travel times are combinations of these segment travel times. An important point is that the trajectories followed by vehicles traversing these routes are all the same. The vehicles all experience the same sequence of nodal processing.

To illustrate these ideas of segments and routes, consider Figure 2.3. A sequence of three segments is shown, A, B and C, that form a route from O to D. The nodes R and S are intermediate nodes at the junctures between segments. Each segment has a travel time distribution (and by dividing by the length, a travel rate distribution). The route also has a travel time (travel rate) distribution.
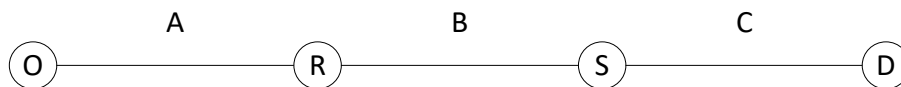
Figure 2.3: A 3-segment route

Many questions arise about the route's reliability: 1) what are the travel time distributions for the segments, 2) what is the travel time distribution for the route; 3) can the segment travel time distributions be combined to create a route travel time distribution; 4) can the route's travel time distribution be predicted based on the segment travel time distributions; and, based on all this, 5) can ways be identified to improve the route's reliability?

Regarding question #1, if the sensors are placed mid-link, then the segment travel time distributions can be observed. And from them, route-level distributions can be constructed. See later discussions about how to do this. If the sensing system can detect vehicle types (e.g., if toll tags are used), then travel time distributions by vehicle type can be developed.

For question #2, the same response pertains. If the sensors are placed mid-link and enough vehicles traverse the entire route, then the route-level travel time distribution can be observed directly. There can be a problem if there are not enough vehicles that traverse the route of interest. An intent of the response to question #4 is to address that type of situation.

For question #3, if the segment travel time distributions are statistically independent, the answer is reasonably simple. The segment-level distributions can be convolved to construct the route-level distribution. But Isukapati *et al.* (2013) and others have shown that the independence assumption does not often pertain. In fact, for networks that are uncongested, positive, serial correlation is likely to exist among the segment travel times. Isukapati *et al.* (2013) show that the travel times for each percentile can often be added across the segments to synthesize the travel time distribution for the route. That is:

$$t_{rp} = \sum_s t_{sp} \quad \forall\, p \tag{2.1}$$

where t_{rp} is the pth percentile travel time for the route and t_{sp} is the pth percentile travel time for segment s. This property is called comonotonicity.

The complexity of this synthesis challenge is illustrated by two routes whose segment and route - level travel times are shown in Figure 2.4. In the case of route #1, adding the percentile-by-percentile travel times, as suggested by equation 2.1, works quite well. In the case of route #2, however, that technique does not work. The significant tail in the distribution for segment C is not evident in the route travel time distribution.

List *et al.* (2016) show that a three-step process can be used to synthesize the route travel time distributions. The first step is to develop the segment-level distributions. The second is to create three hypothetical route level distributions that combine the segment distributions in three ways: positively correlated (PosDis), negatively correlated (NegDis) and uncorrelated (UncDis). In the first case, equation $t_{rp} = \sum_s t_{sp} \quad \forall\, p$ (2.1) is

used. In the second, a variant of equation ($t_{rp} = \sum_s t_{sp} \quad \forall\, p$

(2.1) is used where the percentiles for the segments alternate between p and $1 - p$ every other segment. In the third, the distributions are convolved.

These distributions are then proportionally sampled to develop the route-level distribution:

$$F_r = \alpha F_{PosDis} + \beta F_{NegDis} + \gamma F_{UncDis} \quad where \quad \alpha + \beta + \gamma = 1 \tag{2.2}$$

The values of α, β, and γ are based on minimizing the sum of the squared differences between the observed percentile values of the route travel time distribution, t_{rp} and the corresponding estimated values $\hat{t}_{rp}$ derived from equation $F_r = \alpha F_{PosDis} + \beta F_{NegDis} + \gamma F_{UncDis} \quad where \quad \alpha + \beta + \gamma = 1$ (2.2).

(a) route #1   (b) route #2
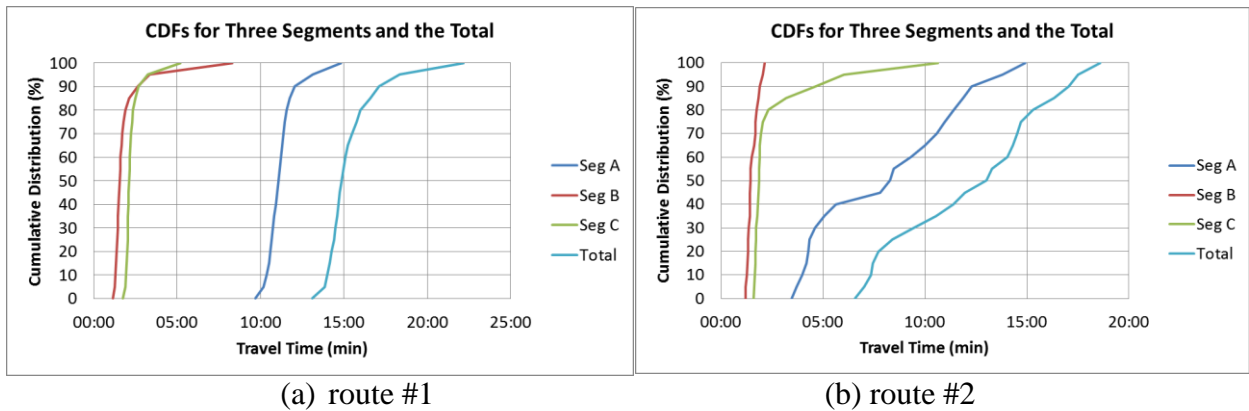
Figure 2.4: Segment and route cumulative distributions (CDFs) for two 3-segment routes

Determining $α$, $β$, and $γ$ leads to insights about the route's operational condition. For example, a high $α$ value pertains when the segments have uninterrupted, uncongested flows. Drivers can pursue their desired speeds and the travel time distributions on successive segments tend to be very similar. A high $β$ value pertains on segments where the travel rates are sequentially high and then low. An example is a signalized arterial where the vehicles see a green signal at every other intersection. A large $γ$ value pertains when the segments involve congested flow. Drivers have little ability to pursue their desired speeds. Travel times tend to be controlled by the congestion.

These insights suggest that look-up tables could be created that store values of $α$, $β$, and $γ$ that pertain to specific routes for specific operating conditions. Given these tables, the predicted distribution of travel times at points in the future could be forecast. This hypothesis, however, has yet to be tested. This is a theorized answer to question #4.

Question #5 may be answered by altering the attributes that produce the segment level distributions (e.g., changes in geometry or operating rules) to improve the reliability of the travel times. Speed harmonization is an excellent example. So is the addition of new capacity and the construction of HOT lanes (at least for the people who use the HOT facility).

2.2   RELEVANT LITERATURE

Node, segment, node, and route (path) level reliability have been topics of considerable research interest for at least half a century.

Within the rail domain, Boysen *et al.* (2013) assessed the reliability of container processing in railway yards. They investigated the problems that railroads face in maximizing reliability and solution approaches that they use. The authors found that improved layouts of the yards and train assignment improved the reliability of train movements. Lang and Reid (1970) examined road train delays between yards. The causes of delay were studied for more than a thousand trains operating over a single main-line division during a two-month period. They found that the variations in travel time were caused by equipment failures (e.g., brakes, couplers, engines), attributes of the train (e.g., length, trailing tonnage) and the alignment (e.g., vertical profile). Martland (1982) created the "PMAKE" function, illustrated in Figure 2.5 to model connection reliability in railroad yards.

The PMAKE function describes the cumulative probability that an arriving car at a classification yard will connect to outbound trains within a specific amount of time. He concluded that small improvements in train and yard operations could translate into significant decreases in delay, which would result in better travel time reliability.

Dejax and Bookbinder (1991) examined the route (OD) reliability of car movements on the French National Railway. They found they could create a reliability index that codified the timeliness and reliability of the services provided. Little *et al.* (1992) similarly reviewed the reliability for boxcar traffic in the US. Kwon *et al.* (1995) studied origin-to-destination (OD) movements for 477 general merchandise trips, 102-unit train trips, and all trips for the 10 largest double-stack corridors. Vromans (2005) examined route-level reliability in the context of a typical European rail network typified by freight services in shared operation with significant passenger services. Yuan (2006)

did a similar study and presented an analytical probability model that estimates the delay for trains caused by route conflicts and late transfer connections. The author found that the variation caused by train events can be approximated as a lognormal distribution or a Weibull distribution. Arcot (2007) considered the problem of modeling uncertainty in rail freight operations and its implications for service reliability by presenting an operational simulation tool to evaluate different rail operational policies. The simulation tool determined that the policies such as priority-based classification, train holding and train cancellation strategies yielded the greatest improvements in shipment connection reliability. Kaplan (2007) examined the reliability of rail-based coal shipments and found that the reliability of transportation of coal yards faces problems such as severe weather, surges in demand, difficulties in integration of railroad mergers, and unplanned maintenance programs.
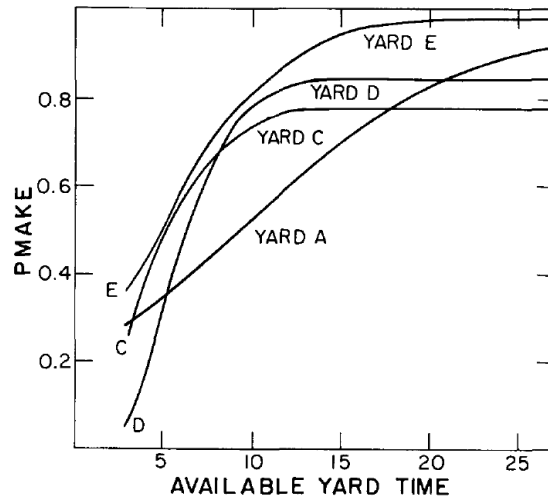
Figure 2.5: The PMAKE function for rail cars passing through railroad yards

For barges, Dai and Schonfeld (1991) studied the reliability of barge trips on a section of the Ohio River. They analyzed delays that accrued as barges passed through locks and interact with other river traffic. For other freight sectors, Wang (2007) examined the reliability of air cargo services in China and Johnson and Dupin (2012) studied the reliability of oceanic trips. Woo and Pettit (2010) examined the reliability of vessel servicing times at ports. Zhao and Goodchild (2011) examined the same issues for drayage at a port. In the latter case, the authors presented a simple method to predict the 95 percent confidence interval of travel time between any OD pair. Their method was validated using global positioning system (GPS) data. Jones and Sedor (2006) studied freight reliability for trucking operations. They used satellite data from trucks traveling on five freight-significant corridors to calculate travel rates and to derive measures of travel time and reliability. Czuch *et al.* (2011) examined the travel time reliability of truck shipments. Bluetooth units were used to measure travel time and reliability. They concluded that the use of Bluetooth readers in combination with simple metrics provided a cost-effective way for municipalities to measure travel time reliability.

From a public agency perspective, highway network reliability has seen attention only recently. In 2001, Lomax *et al.* (2001) began monitoring highway congestion for the major US metropolitan areas. The ideas of buffer time, planning time, travel time index, and other reliability performance metrics were identified in this effort (see Lomax *et al.*, 2003). The group has continued to publish

nationwide performance assessment reports annually, as in the case of the 2012 Urban Mobility Report (Schrank *et al.*, 2012). Van Lint, van Zuylen, and Tu (2008) also examined the issue of how to measure and assess travel time reliability. They reviewed several measures reported in literature. Their most important ones were twofold from comparing the various measures on a large empirical dataset. First, the measures were inconsistent. This was true even when comparing existing commonly used travel time reliability indicators. For example, the results of the misery index differ largely from the results of the buffer time index. Second, a compound measure was suggested. Like Lam and Small (2001), they suggested monitoring both the difference between the 90th and 50th percentile as a robust indicator and the ratio of the difference between the 90th and 50th percentile and the difference between the 50th and 10th percentile as a measure of skew. They interpreted this new measure as the likeliness of incurring a very bad travel time (relative to the median). This new compound measure, in contrast to classical statistical metrics for width and skew, allows a partial reconstruction cumulative distribution function which is useful from a reliability perspective.

Elefteriadou and Ciu (2007) created a model for estimating travel time reliability on freeway facilities. They observed that the commonly held notion of reliability among highway analysts is very different from the one articulated by previous reliability authors such as Ebeling (1997). Ebeling said reliability should be "the probability that a component or system will perform a required function for a given period when used under stated operating conditions. It is the probability of a non-failure over time." Highway analysts have focused instead on the idea of consistency, which must do with the absence of variability.

Chu *et al.* (2010) examined various reliability measures such as the planning time index, the buffer time index, and the reliability index in the context urban freight corridors that provide access to a seaport. The on-board global positioning system (GPS) installed on heavy-duty commercial vehicles was utilized to collect travel time and speed data. Also examined is the validity of using parametric distributions such as Gamma, log-logistic, log-normal, and Weibull to fit the data. Their goodness-of-fit tests indicate that the log-logistic is the best statistical function for freight travel times during the mid-day period. In addition, their travel time prediction models identify relationships between travel time, speeds, and variance-related factors that affect travel time reliability such as incidents, work zones, and traffic signal breakdowns.

Yamamato *et al.* (2006) studied the reliability of travel time estimates based on the frequency of probe observations, focusing on the variability of link /roadway segment travel time estimates for different data frequencies. Their results suggested that higher frequency probe data do not always yield less variance in the link travel time estimates, and lower frequency data have a smaller variance at links just before signalized intersections.

Ramezani and Geroliminis (2012) estimated arterial travel times using data from probe vehicles. Using a heuristic grid and applying Markov chains, the researchers integrated the correlation between successive links in a network in both simulated data and in-field measurements. Jenelius and Koutsopoulos (2013) utilized GPS probes to generate a statistical model for urban road network travel times. The model they developed used spatial moving averages to allow correlation between network links to be shown. The model also considers attributes such as speed limit and trip conditions, such as day of week, season, and weather.

Correlations among link travel times have also been investigated. This is critical for understanding how segment-based distributions can be combined. Zeitsman and Rilett (2003) used automatic vehicle identification to analyze individual and aggregate travel patterns. Significant findings included that link travel times of individuals do not have as high of a positive correlation on a trip-by-trip basis as studies on aggregate trips show. Saberi and Bertini (2010) compared correlated segments to uncorrelated segments to assist in identifying hotspots along a freeway. Seshadri and Srinivasan (2010) developed an algorithm to determine the maximum travel time on a network with normally and correlated link travel times. Ji *et al.* (2011) presented a simulation-based multi-objective genetic algorithm to determine sets of reliable paths in stochastic networks. The researchers considered the uncertainties in the travel times and correlations among link travel times by using Monte Carlo simulation, genetic algorithms, and a Pareto filter module.

Isukapati *et al.* (2013) examined a way to synthesize route travel time probability density functions (PDFs) based on segment-level PDFs. They used real-world data from I-5 in Sacramento, California. Also, List *et al.*, (2014) considered several options for combining segment-level distributions. List *et al.* (2012) and subsequently Isukapati *et al.* (2013) further advanced these ideas. In the work by List *et al.* (2012), it was established that comonotonicity (perfect positive correlation) could be used under a wide variety of conditions to synthesize route-level distributions for freeways. It was observed that the ability to do this was dependent upon the level of congestion extant on the segment.

2.3   ROUTE TRAVEL TIME SYNTHESIS METHODOLOGIES

This section describes three ways to synthesize the route-level travel time distribution for specific truck types. The methods are applicable more generally, for any type of vehicle, but this report is focused on trucks.

The motivation is that a more refined sense of the travel times involved in making trips is helpful in ensuring that reliable freight service is provided. Working with the average travel times is not enough. What is desired is a sense of the individual vehicle travel times that might arise for a given OD (origin-destination) pair, for a variety of operating conditions (e.g., normal and abnormal, with the latter being characterized by weather, incidents, maintenance work, etc.).

2.3.1 Method One: Average Travel Times and Distribution Inference

The first method is the least demanding in terms of data, and the most dependent on inference. It is easy to apply for a given operating condition, but the quality of its estimated distributions is highly dependent upon the insights of the analyst.

The first step is to understand the travel time distributions of the trips that the trucks have already made. It is best to focus on the distribution of the travel rate rather than travel times because then data from different trip lengths can be compared. Data from trips during various operating conditions might also be of interest. The distributions of these travel rates need to be studied to see if the shape is always the same; how the mean travel times compare to the travel times reported by a route guidance system (e.g., Google); if the variance depends on the trip length; and if or to what extent the distribution is affected by adverse operating conditions. Several outcomes are needed:

1) the ratio of the observed truck travel times to the values reported by the route guidance system, 2) a relationship between the length of the travel time and the variance in the truck travel times (or the individual percentiles of the distribution). Both these outcomes might be sensitive to the operating conditions of the network (e.g., normal, adverse weather, incidents, work zones, and combinations of these).

The second step is to collect travel time data for the new route. Such data can be obtained from route guidance services such as Google maps. The method assumes that segment travel times are available at the granularity of about five minutes. This means the observations are not noisy and trends in the temporal variations of the travel times do not confound the observations. The method also assumes that travel time distributions exist for similar truck trip trips that have taken place over segments and routes for other paths through the network.

Having data for a year is ideal. The main value in this is that every seasonal condition has data. Each year has 105,120 five-minute observations (365*24*60/5). 74,880 of those are on weekdays; 29,952 are on Saturdays and Sundays. Given about 250 workdays each year, 72,000 of the observations are on workdays. If the focus is on trips made between 7am and 9am on workdays, then there are about 6,000 observations each year. This is based on 24 five-minute observations per day (or 120 per week) and 250 workdays. Of course, these observations need to be binned based on the operating condition that pertains, with a separation at least into normal and "abnormal" days, with the latter possibly being separated into categories like "incidents", "weather", and "maintenance work", or combinations of these.

The third step is to develop distributions of the five-minute average travel times for the daily timespan of interest and operating condition. An example would be the AM peak on workdays, as was mentioned above. Of course, for the condition of interest, there must be enough observations to create a meaningful distribution (say 50 so, which means there is an observation for every 2nd percentile). This result provides a sense of the variation in the route's average travel times.

The fourth and last step is to combine the results from the first and third steps to create the travel time distribution for the route and condition of interest. Monte Carlo sampling is used. The distribution of the mean travel time is treated as the average travel times that the overall traffic stream experiences and/or the distribution of values that are reported by the route guidance system. The findings from step #1 are used to hypothesize what the truck travel time distribution is based on the average distribution. Monte Carlo sampling is used to sample values from the average and then samples from the hypothesized truck travel time distribution to develop the estimate of the actual truck travel time distribution.

As a postscript, this does mean that longer routes will have wider travel time distributions. The ratios from the travel rate distribution are being applied to the mean travel rate for the route. Aa route with a longer travel time will have a greater spread in the distribution of the travel times than will a route with a shorter travel time. If this seems at odds with the intuitively appealing outcome, then an alternate approach to step #1 is to treat the historical routes separately and compute the differences in times from the mean value to each percentile. These differences, by percentile, can then be averaged to create a composite vector of average differences. This result effectively

assumes that the faster drivers always arrive "a few" minutes earlier than the average drivers and that the slower ones arrive "several" minutes later; and that the length of the trip does not matter.

An example of using this technique is helpful. Figure 2.6 shows an OD pair from Carmel Mountain Road in Torrey Pines, CA where I-5 and I-805 split to Civic Center Drive in National City, CA. Three paths will be considered: Path A is via I-5; Path B is via I-805, CA-163, and I-5; and Path C is via I-805, CA-15, and I-5.
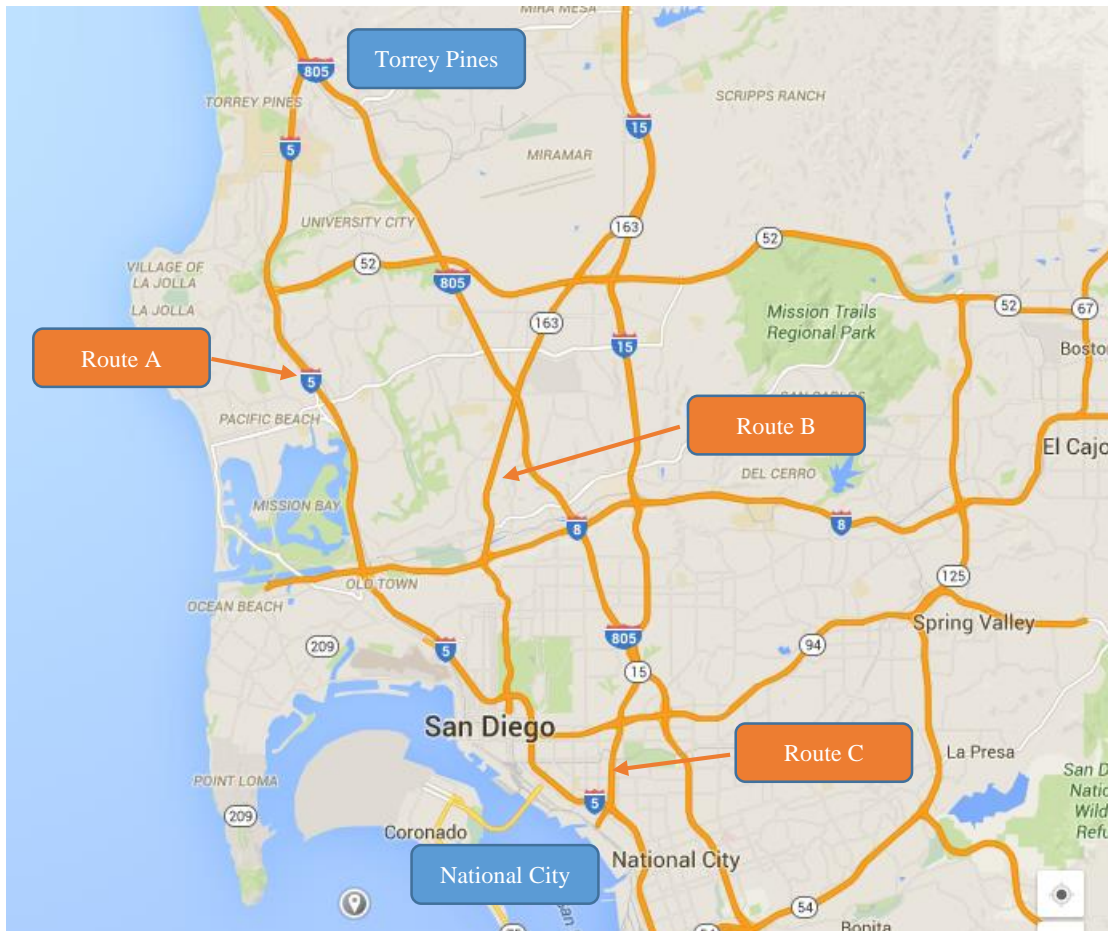
Figure 2.6: Map of three paths from Torrey Pines to National City

The first step is to understand the travel time distributions of the trips that trucks have already made. Hypothetically, Figure 2.7 shows the distributions of travel times when the values have been normalized based on the average travel time. To ensure this figure is clear, assume that existing trips have been analyzed and the observed travel times have been normalized (divided by) the average travel time. Then assume that the observations have been combined into a database and that the distribution of the travel times (relative to the mean) have been created. These hypothetical results are displayed in the figure.

The second step is to collect travel time data for the new route. Guidance about how to do this was given above. List *et al.* (2013) provide further guidance. In this case, 5-minute travel time data were obtained for the workdays in an entire year. The dataset contained 72,000 records. As

originally received, the records contained no information about the environmental conditions under which the system was operating when the travel time observation was obtained. So additional data collection activities followed in which incident, weather, maintenance, flow rate and special event data were obtained. Fields were added for each of these events. A simple severity index was developed and employed for the weather data. The incident, special event, and maintenance work indicators were yes/no plus a duration. The flow data was in vehicles/hour/lane.

(a) Normalized PDF   (b) Normalized CDF
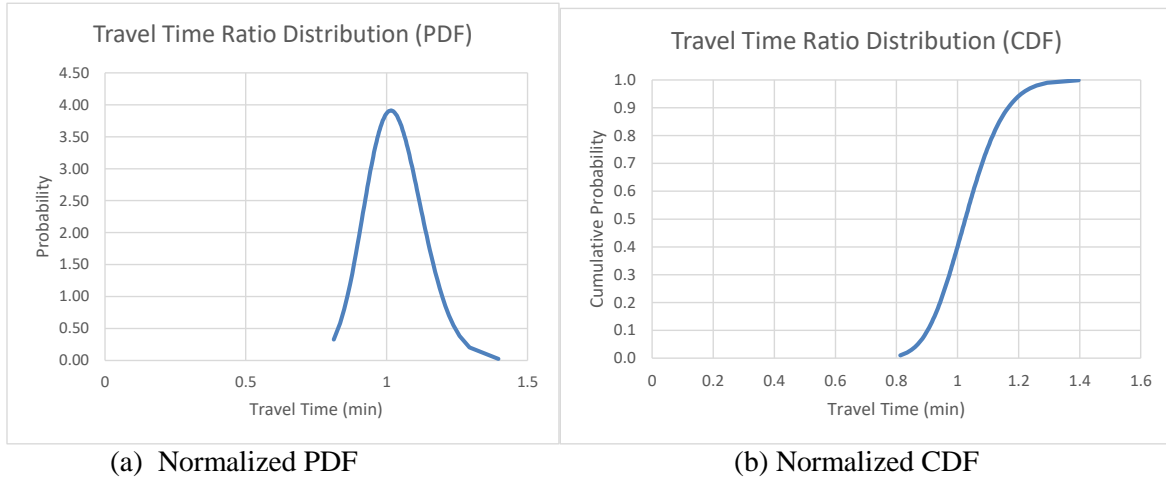
Figure 2.7: Normalized PDF and CDF for truck trips on other routes

The third step is to develop distributions of the five-minute average travel times and travel rates for different congestion levels and environmental operating conditions. For some of the paths, the flow rate data motivated a distinction between low, moderate, and high congestion. For others, low and high were all that seemed appropriate. Plots of the travel rates for the three routes are presented in Figure 2.8, Figure 2.9, Figure 2.10.

As can be seen, the travel rates are sensitive to the operational environment. Weather causes problems as do incidents, special events, and high demand. The impact is more severe when the traffic congestion is high than when it is low. Perhaps surprisingly, for an urban area whose climate is very temperate, weather has a significant impact. The weather events do not occur frequently, but, perhaps, when they do occur, the change in the operating conditions is quite significant.
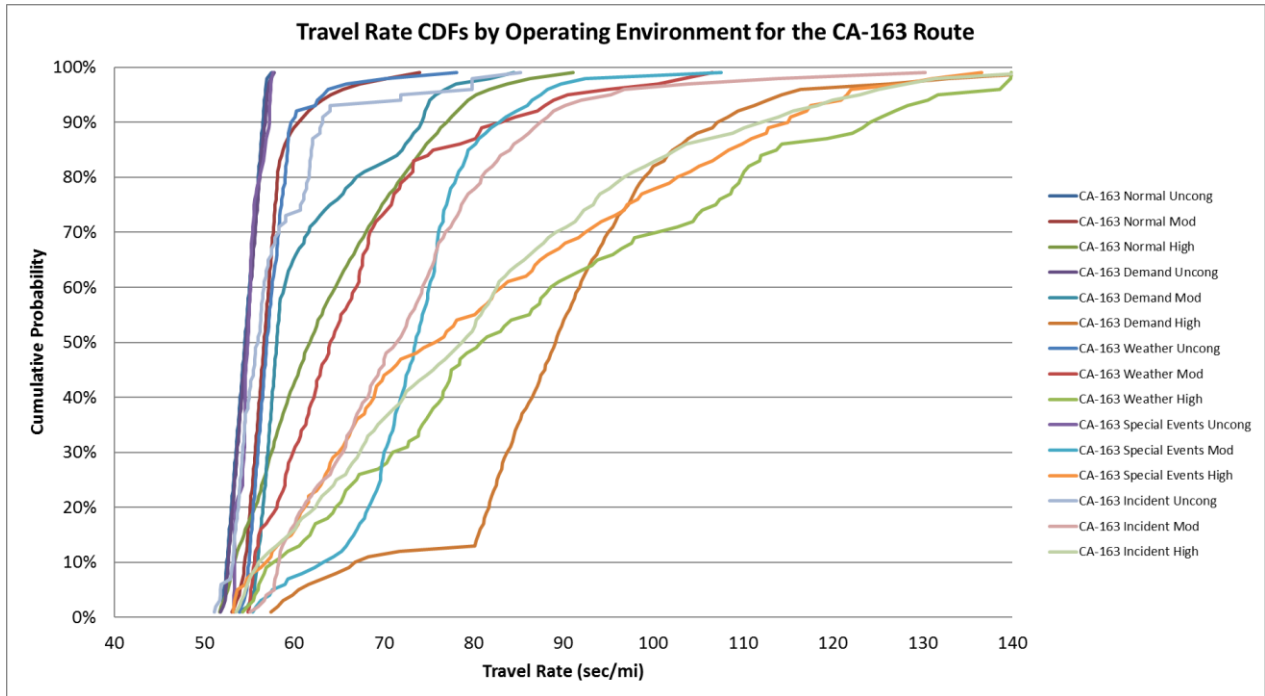
Figure 2.8: Travel Rate CDFs by operating environment for the CA-163 route
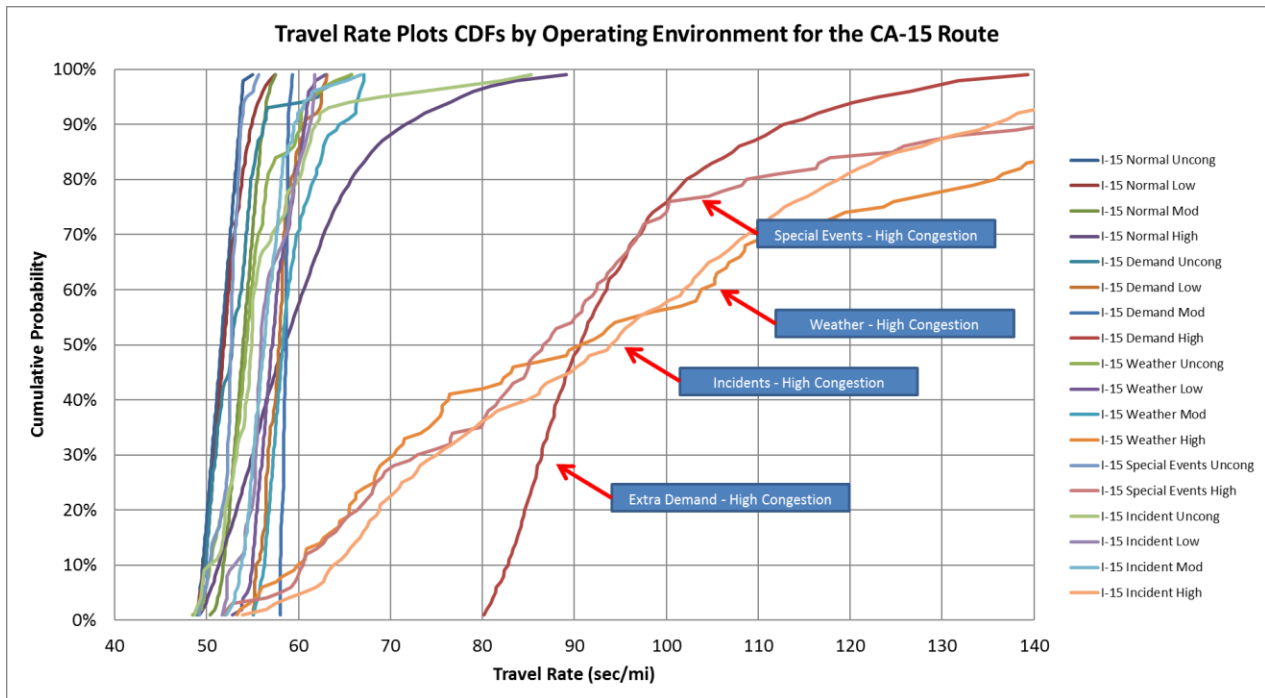
Figure 2.9: Travel Rate CDFs by operating environment for the CA-15 route
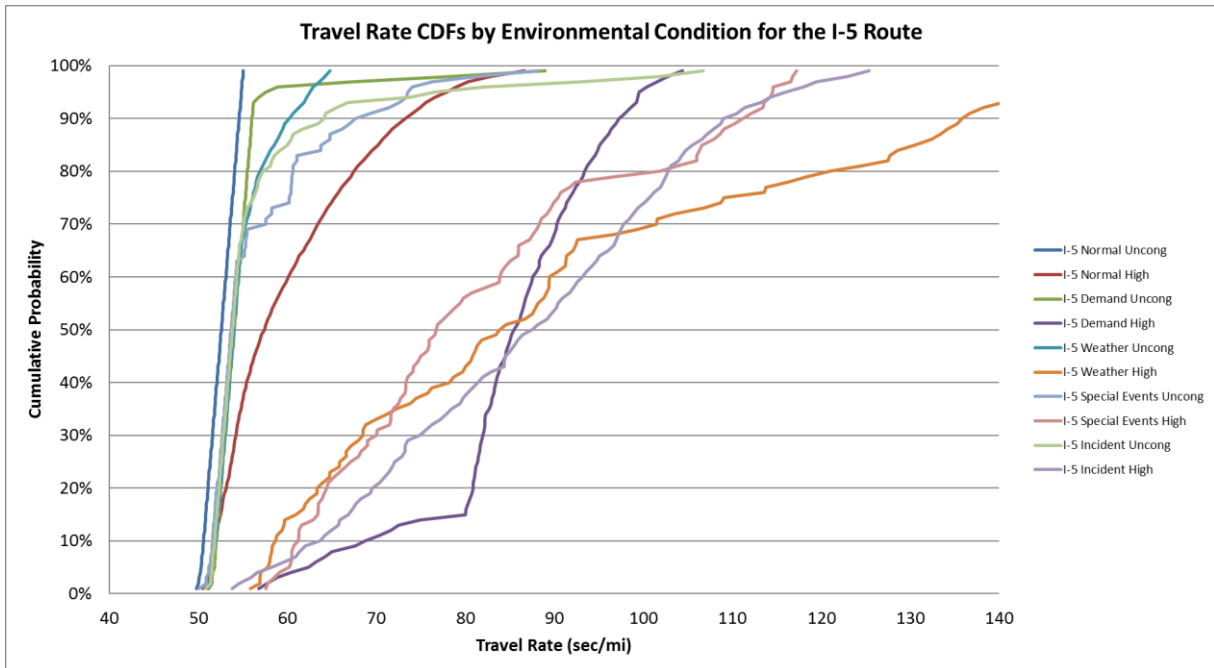
Figure 2.10: Travel Rate CDFs by operating environment for the I-5 route

The fourth and last step is to combine the results from the first and third steps to create a sense of the travel time distribution for the route and condition of interest. Figure 2.11 provides an illustration of how this is done.
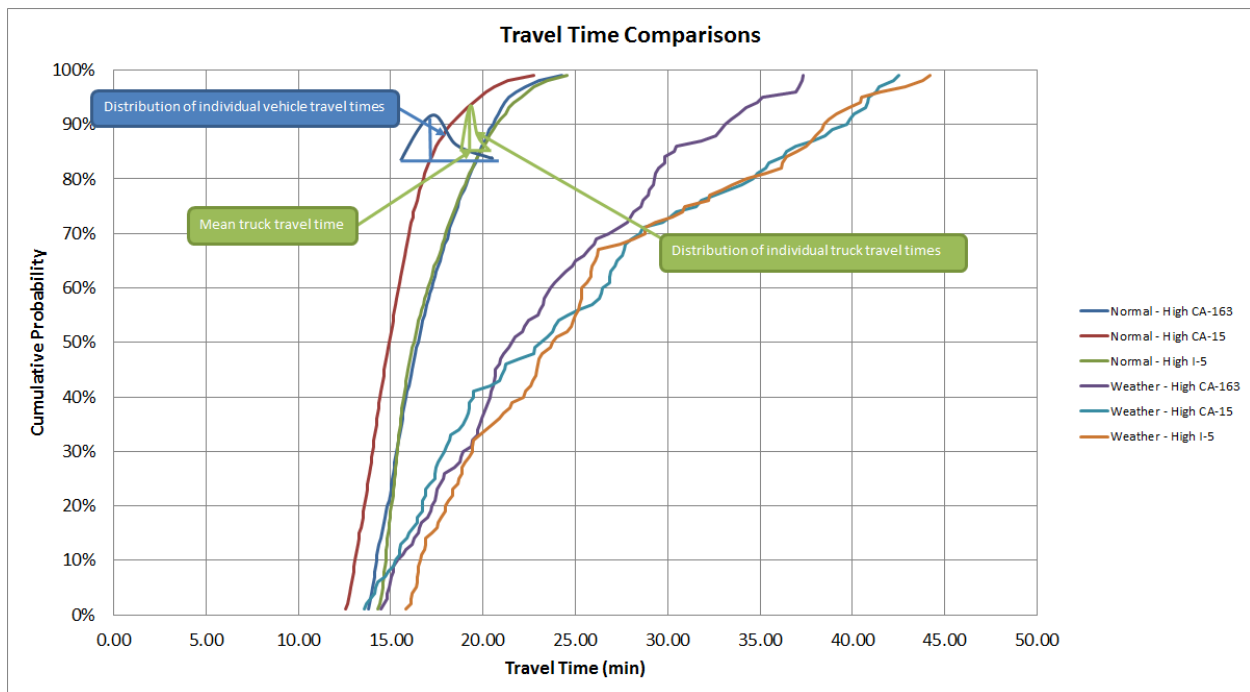
Figure 2.11: Graphical description of how the truck trip times are developed

The figure shows plots of the average travel time CDFs for the three routes under normal and adverse weather conditions. To make sure they are clear, a little interpretation and analysis can be performed. Assume the truck has a travel time that matches the average. Then these CDFs reflect the CDFs that the truck experiences. Hence, the CA-15 path is always the shortest; the CA-163 and I-5 paths have longer travel times and are nearly identical. When the weather is adverse, however, the CA-15 path is not always the best. Sometimes the CA-163 path is better. In fact, if the main interest lies in selecting the path that is most likely to provide the shortest travel time when weather is a problem, the CA-163 path is best. It does not always have the shortest travel times at the lower percentiles, but at percentiles above about 45% it is always best.

The two PDFs that are superimposed on the diagram are interpreted as follows. It is assumed that the average travel time is at the 80th percentile of its distribution. The blue PDF then shows, hypothetically, the distribution of individual vehicle travel times that pertains at the 80th percentile. Said another way, it indicates the distribution of individual vehicle travel times that pertains when the average travel time is at the 80th percentile. Some vehicles have shorter travel times (about half of them since what is shown is the average value, not the minimum); and others have longer travel times. The green PDF shows the distribution of the truck travel times. Since the average truck travels slower than the average vehicle, the PDF for the truck is displaced from the PDF for all vehicles. On days when the average travel time is about 17 minutes (at the 80th percentile) and the distribution of all individual travel times ranges from 15.5 to about 21 minutes, the mean travel time for the trucks is about 19 minutes and the individual vehicle travel times range from 18.5 to 20.5. To be clear, these latter two distributions are hypothetical while the CDF shown is based on real data. The two PDFs have been created to illustrate the ideas being presented. But there is no question that these distributions exist, and the objective of this method is to characterize the distribution of truck travel times.

The distribution of the truck travel times is created by doing Monte Carlo sampling combinations of the average travel time CDF and the truck PDF across the spectrum of both distributions. Random values of the average travel time PDF are sampled to create a representative sample of those values. And then for each of these a second sample is taken from the truck travel time distribution, to create an observation of what the truck travel time would have been. These Monte Carlo samples are then summarized into a CDF for the truck travel time.

The results of these efforts are shown in Figure 2.12. While the plots are like those in Figure 2.11, the CDFs in Figure 2.12 are for truck trips, not average vehicle trips. The values are larger. It is assumed that the truck trips are 30% longer in time than the average trips, and that the truck trip times range from 80% to 140% of the average as shown in Figure 2.6.

As was the case with the average travel times, the CA-163 route has the best travel times for normal operating conditions. And when the weather conditions are adverse, this path is still very good, but not always the best. At low percentiles, it's travel time is competitive with the best, and at high percentiles it is clearly better than the other two paths.
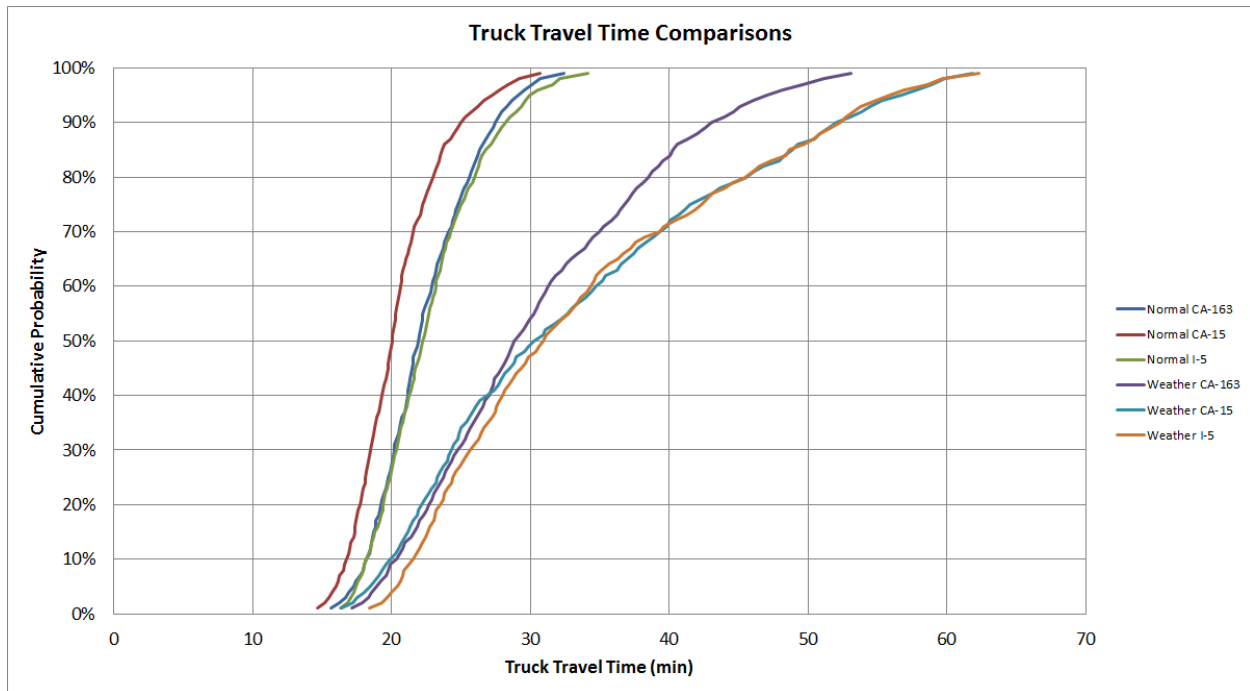
Figure 2.12: Truck travel time CDFs for two operating environments and three routes

Once these distributions have been created, a variety of questions can be answered: how long will the trip take; what is the best path to choose; what is the shortest travel time that might arise; what is the longest travel time? Of course, since the average travel times vary from day-to-day, as shown by any one of the CDFs in Figure 2.12, the truck travel time distribution will shift with it. This means that, on a given day, it is important to have a sense of what the average travel time is, or will be when the trip takes place; and then the departure time window can be determined.

2.3.2 Method 2: Proportional Sampling

The second method relies on individual vehicle travel time observations for segments that lie along the route of interest. As described in Section 2.1, it combines positive, negative and uncorrelated combinations of segment travel time distributions to synthesize a route-level distribution. The equation employed is 2.1. The process is simple and straightforward and yields a distribution that very closely matches the one observed.

The context of the assessment is shown by Figure 2.13. The origin node is O and the destination is node D. The trip involves a short segment from O to A for which only average travel time data are available. For segments AB, BC, CD, and DE, individual vehicle travel time observations are available, but the number of vehicles that travel from A to E may be very limited. Most of the traffic may be going from G to J. There is also traffic that goes from F to J and from H to K. The CDF that can be developed directly from the traffic that goes from A to E is sketchy and coarse (i.e., A, F, or O to D, E, or K). But the other traffic provides significant observations on some of the segments. The objective of the method is to make use of all segment-level observations available to develop a more refined distribution of trip travel times from A to E. Once that

35

distribution has been created, it can be augmented with Monte Carlo-based estimates of the travel time distributions for segments OA and ED to prepare the overall OD distribution.
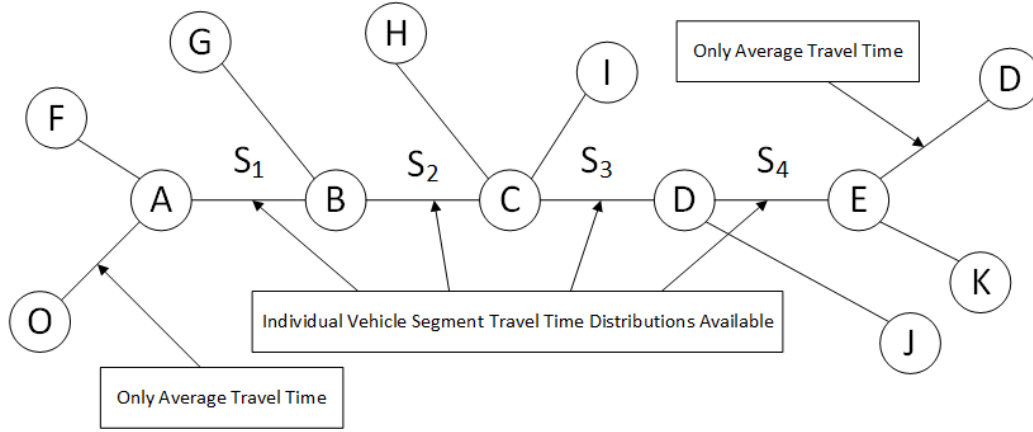
Figure 2.13: Data availability for a route of interest

The first step in this method is again to understand the travel time distributions of the trips that the trucks have already made. The procedure described in Section 2.3.1 is used to develop: 1) the ratio of the observed truck travel times to the values reported by the route guidance system, 2) a relationship between the length of the travel time and the variance in the truck travel times (or the individual percentiles of the distribution). As before, these results are sensitive to the operating conditions (e.g., normal, adverse weather, incidents, work zones, and combinations of these).

The second step is to develop estimates of three mix coefficients: α, β, and γ. These coefficients indicate how the three synthesized distributions F_{PosDis}, F_{NegDis} and F_{UncDis} need to be combined to create the estimated travel time distribution from A to E. Equation 2.2 is employed. For explanation purposes, let X_s be the travel time random variable associated with segment s, $s = 1, \ldots N$. Also, let $\underline{X}$ be a vector representation of these random variables. In the context of the example shown in Figure 2.13, $\underline{X} = [X_{AB}, X_{BC}, X_{CD}, X_{DE}]$. F_{PosDis}, F_{NegDis}, and F_{UncDis} are the positively, negatively, and uncorrelated distributions formed by combining F_{AB}, F_{BC}, F_{CD}, and F_{DE}.

F_{UncDis} is created by convolving X_1 through X_N under the assumption that the X_s random variables are independent. That is, $F_{UncDis} = F_1 \otimes F_2 \otimes \ldots \otimes F_n \otimes \ldots \otimes F_N$. Mathematically, this is done by repeatedly evaluating the following integral for successive segments:

$$(f_u \otimes g_v)(t) \equiv \int_{-\infty}^{\infty} f_u(\tau) g_v(t-\tau) d\tau \tag{2.3}$$

Variables u and v are two random variables whose distributions $f_u(t)$ and $g_v(t)$ are being convolved. Initially, this can be the combination of X_{AB} and X_{BC}. Then that result is convolved with X_{BC} and

so on until X_{DE} has been included. For variables that only have positive values (as is the case with travel times), equation 2.3 simplifies to:

$$(f_u \otimes g_v)(t) \equiv \int_0^t f_u(\tau) g_v(t-\tau) d\tau \tag{2.4}$$

Instead of doing this integration explicitly, Monte Carlo sampling can be used. Segment travel times can be randomly drawn from X_{AB}, X_{BC}, X_{CD}, and X_{DE} and added together. This can be done a countably infinite number of times and then percentile values then drawn from an ordered sequence of the values.

F_{PosDis} is formed by drawing samples at the exact same percentile of each distribution, i.e., from X_{AB}, X_{BC}, X_{CD}, and X_{DE}. Then:

$$X_R^p = X_{AB}^p + X_{BC}^p + X_{CD}^p + X_{DE}^p \quad \forall\, p \tag{2.5}$$

Doing this follows a property called comonotonicity. Isukapati *et al. (13)* demonstrated that this technique can be used to synthesize route-level travel times for uncongested freeways. The proportional sampling produced route-level travel time distributions which were consistent with those \observed from direct measurements.

F_{NegDis} is similar except that, alternately, complementary percentiles are selected on subsequent segments. For this example:

$$X_R^p = X_{AB}^p + X_{BC}^{1-p} + X_{CD}^p + X_{DE}^{1-p} \quad \forall\, p \tag{2.6}$$

Or more generally:

$$X_R^p = \sum_s X_s^n \quad \forall\, p \quad \text{where } n = p \text{ if } s \text{ is odd and } n = 1-p \text{ if } s \text{ is even} \tag{2.7}$$

The ability of this technique to synthesize observed route-level distributions has been illustrated using toll tag data that were collected during the 2007 New York State Fair. Descriptions of the twelve locations are as follows:

- **Fair-1(ART):** On Bear Street at the on / off Ramp for I-690. This sensor captured the left lane for both the on and off ramps.
- **Fair-2, Fair-5 (FWY):** On I-690 westbound on the western end (downstream) of the work zone. Two sensors were employed to ensure that as much data as possible could be collected. Duplicate reads were frequently recorded. For this analysis, the data from these two sensors were combined and the duplicate observations were removed.
- **Fair-3 (FWY):** On I-690 eastbound at the western end of the work zone (upstream). The sensor captured the right most lane.
- **Fair-4 (ART):** On Hiawatha Boulevard, just before Spenser Street. This sensor captured

37

the right most lane.

- **Fair-6 (ART):** On I-690 at Exit 7. This sensor was on the westbound ramp at the base of the entrance to the Orange parking. This sensor captured the right most lane.
- **NY-2 (PKG):** In the Brown parking lot.
- **NY-3 (PKG):** In the Orange parking lot at about the same location as I-690, Exit 7.
- **NY-4 (PKG):** In the Orange parking lot at about the same location as I-690, Exit 6.
- **34-A (TPL):** At NYS Thruway exit 34A (not visible in the map)
- **36-X (TPL):** At NYS Thruway exit 36 (not visible in the map).
- **39-X (TPL):** At NYS Thruway exit 39 (not visible in the map).

Some of the readers were positioned on arterials (ART), others are on freeways (FWY), and still others are in parking lots (PKG) or at toll plazas (TPL). The locations of readers Field-1 through Field-6 are shown in Figure 2.14.
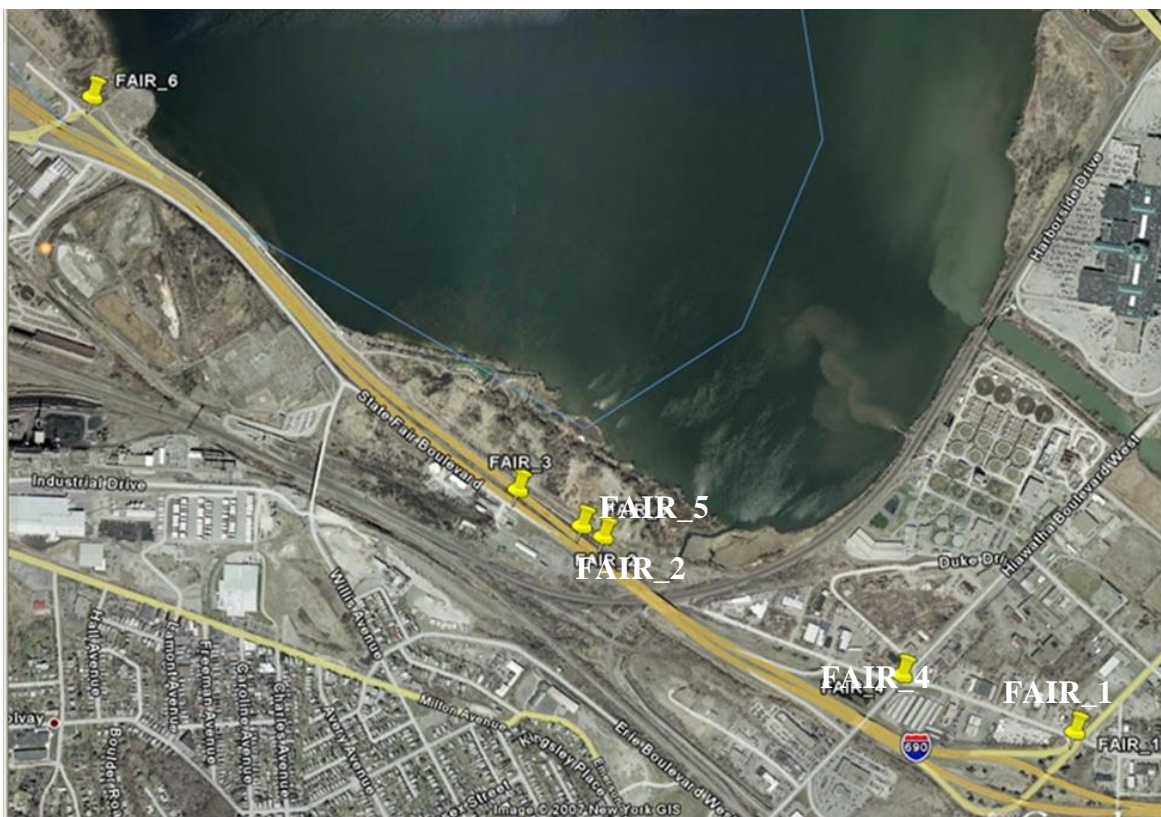
Figure 2.14: Location of readers Field-1 through Field-6

A total of 249,895 readings were recorded from 79,248 unique transponders. The maximum number of observations from a single toll tag was 230 and the minimum was one. 345 toll tags had 30 or more observations; 1,136 had 20 or more; 4,869 had 10 or more; and 33,979 had only one. The average was 3.15 observations.

The toll tags used for analysis had at least three data readings. In addition, they had time stamps that reflected realistic travel times between the monitoring locations and not double-reads at the same location. There were 15,457 such toll tag reads.

Trips were created by chaining together sequences of sequential reads where the differenced in times was greater than 15 seconds and less than an hour. For example, if a transponder was observed at locations *a*, *b*, *c*, *d*, *e*, and *f* and the time differences were 0, 20, 50, 3700, 3750, and 3800 seconds, then two trips would be formed. The one would be *a*, *b*, and *c* and the second would be *d*, *e*, and *f*. Where these resulting trips had three or more segments, subtrips were created. For example, if there were a trip involving the location sequence *r*, *s*, *t*, *u*, *v*, and *w*, then subtrips *r-s-t ˛ r-s-t-u, r-s-t-u-v, r-s-t-u-v-w, s-t-u, s-t-u −v, ... s-t-u -v-w, ...* and so on through to *v-w* were created. This results in 8,107 subtrips based on 4,948 transponders. There are 5,443 subtrips with two segments, 1,252 with three, 125 with four, and 63 with five. The maximum number of subtrips generated by a single transponder is 81; the minimum is one; 38 transponders have 10 subtrips or more; and 130 have 5 or more. The total number of routes (sensor sequences) is 2,465; and 51 of these have 20 or more observations. The sequence with the most observations is Fair-3/Fair-4/Fair-2 with 586 observations. These appear to be vehicles traveling eastbound on I-690 that go past Fair-3, get off at the next exit, turn left, go by Fair 4 while turning left again to get back on I-690, and then pass Fair-2 on their way traveling west on I-690. It is likely that many of these subtrips are for people going home from the fair. The next most common sequence is Fair-3/Fair-1/Fair-2, which is the same U-turn pattern only using the next interchange further east.

An illustration of the travel time distributions for one of the two-segment routes is shown in Figure 2.15. This is the sensor sequence Fair-1/Fair-2/Fair 6. In this case, the vehicles pass Fair-1 when getting on I-690 westbound, then Fair-2 while traveling on I-690, and then Fair-6 when exiting I-690 to go to one of the fairground parking lots. The cumulative distribution functions (CDFs) for the two segments and the overall route are shown. The optimal values of α, β, and γ are 0.2, 0.0, and 0.8, which means 20% contribution from F_{PosDis}, 0% from F_{NegDis}, and 80% from F_{UncDis}.
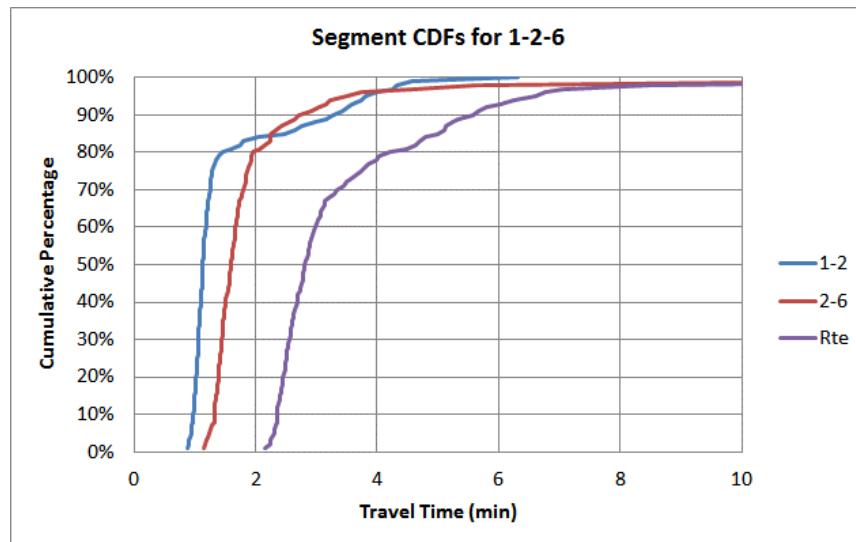
Figure 2.15: Segment and route-level CDFs for route Fair-1 – Fair 2 – Fair-6

For the 51 subtrips that had more than 20 observations, Table 2.1 shows the mix combination values that were obtained.

nSeg	nObs	L1	L2	L3	L4	L5	L6	a (Pos)	b (Neg)	c (Unc)	ksPass
5	32	FAIR_2	FAIR_3	FAIR_2	FAIR_3	FAIR_2	FAIR_3	75%	15%	10%	1
5	31	FAIR_3	FAIR_2	FAIR_3	FAIR_2	FAIR_3	FAIR_2	60%	35%	5%	1
4	54	FAIR_2	FAIR_3	FAIR_2	FAIR_3	FAIR_2		95%	5%	0%	1
4	46	FAIR_3	FAIR_2	FAIR_3	FAIR_2	FAIR_3		65%	5%	30%	1
4	25	FAIR_2	FAIR_3	FAIR_4	FAIR_2	FAIR_6		10%	50%	40%	1
3	369	34A	FAIR_2	FAIR_6	NY3			15%	5%	80%	1
3	217	34A	FAIR_5	FAIR_6	NY3			20%	0%	80%	1
3	125	FAIR_3	FAIR_4	FAIR_1	FAIR_2			25%	60%	15%	1
3	116	FAIR_3	FAIR_4	FAIR_2	FAIR_6			45%	0%	55%	1
3	97	FAIR_2	FAIR_3	FAIR_2	FAIR_3			55%	15%	30%	1
3	84	FAIR_3	FAIR_2	FAIR_3	FAIR_2			75%	25%	0%	1
3	67	FAIR_2	FAIR_3	FAIR_4	FAIR_2			5%	80%	15%	1
3	39	FAIR_1	FAIR_2	FAIR_3	FAIR_1			35%	25%	40%	1
3	37	FAIR_1	FAIR_2	FAIR_3	FAIR_4			15%	40%	45%	1
3	31	FAIR_2	FAIR_3	FAIR_1	FAIR_2			5%	0%	95%	1
3	25	FAIR_3	FAIR_1	FAIR_2	FAIR_3			10%	5%	85%	1
3	23	FAIR_3	FAIR_4	FAIR_2	FAIR_3			15%	25%	60%	1
3	22	FAIR_4	FAIR_1	FAIR_2	FAIR_6			85%	10%	5%	1
2	586	FAIR_3	FAIR_4	FAIR_2				0%	0%	100%	1
2	473	FAIR_3	FAIR_1	FAIR_2				0%	75%	25%	1
2	377	FAIR_2	FAIR_6	NY3				35%	0%	65%	1
2	369	34A	FAIR_2	FAIR_6				10%	0%	90%	1
2	365	FAIR_1	FAIR_2	FAIR_3				5%	15%	80%	1
2	263	FAIR_1	FAIR_2	FAIR_6				20%	0%	80%	1
2	245	FAIR_4	FAIR_1	FAIR_2				15%	5%	80%	1
2	235	FAIR_2	FAIR_3	FAIR_4				15%	0%	85%	1
2	230	FAIR_2	FAIR_3	FAIR_2				60%	0%	40%	1
2	218	FAIR_5	FAIR_6	NY3				15%	10%	75%	1
2	217	34A	FAIR_5	FAIR_6				45%	0%	55%	1
2	204	FAIR_3	FAIR_4	FAIR_1				0%	60%	40%	1
2	190	34A	FAIR_2	NY3				15%	0%	85%	1
2	183	FAIR_3	FAIR_2	FAIR_3				35%	5%	60%	1
2	182	FAIR_4	FAIR_2	FAIR_6				35%	5%	60%	1
2	176	34A	FAIR_6	NY3				35%	25%	40%	1
2	133	34A	FAIR_5	NY3				35%	10%	55%	1
2	110	FAIR_2	FAIR_3	FAIR_1				15%	15%	70%	1
2	97	FAIR_3	FAIR_1	FAIR_5				0%	85%	15%	1
2	73	FAIR_5	FAIR_3	FAIR_4				0%	35%	65%	1
2	72	FAIR_3	FAIR_4	FAIR_5				0%	35%	65%	1
2	69	FAIR_3	FAIR_2	FAIR_6				25%	20%	55%	1
2	57	FAIR_1	FAIR_5	FAIR_3				10%	25%	65%	1
2	53	FAIR_4	FAIR_1	FAIR_5				20%	10%	70%	1
2	38	FAIR_2	FAIR_6	FAIR_3				15%	0%	85%	1
2	37	FAIR_3	FAIR_5	FAIR_3				0%	10%	90%	1
2	33	FAIR_5	FAIR_3	FAIR_1				0%	35%	65%	1
2	32	FAIR_4	FAIR_2	FAIR_3				10%	25%	65%	1
2	28	FAIR_1	FAIR_3	FAIR_4				0%	30%	70%	1
2	27	FAIR_5	FAIR_3	FAIR_5				75%	0%	25%	1
2	24	FAIR_1	FAIR_2	FAIR_4				50%	5%	45%	1
2	24	FAIR_1	FAIR_5	FAIR_6				60%	15%	25%	1
2	23	FAIR_5	FAIR_3	FAIR_2				60%	0%	40%	1

Table 2.1: Optimal proportional sampling rates for the 51 routes examined

As can be seen, the resulting synthesized distributions satisfy the KS test for matching the observed distribution. Also, all but one of the synthesized distributions requires sampling from two or more of F_{PosDis}, F_{NegDis}, and F_{UncDis}. In 50% of the routes, provides the largest contribution to the composite distribution, in 31% of the routes, F_{PosDis} provides the largest sampling proportion, and F_{NegDis} has the largest sampling proportion in 19% of the routes.

Figure 2.16 shows in more detail the frequency with which specific proportional sampling rates are chosen for the F_{PosDis}, F_{NegDis}, and F_{UncDis} distributions respectively. It is easy to see that the sampling percentage for F_{UncDis} is the highest and is non-zero for nearly all 51 routes. The sampling percentages for F_{PosDis} are next highest but diminish to zero more quickly. The sampling percentages for F_{NegDis} are the lowest, but they are non-zero for 37 of the 51 routes.
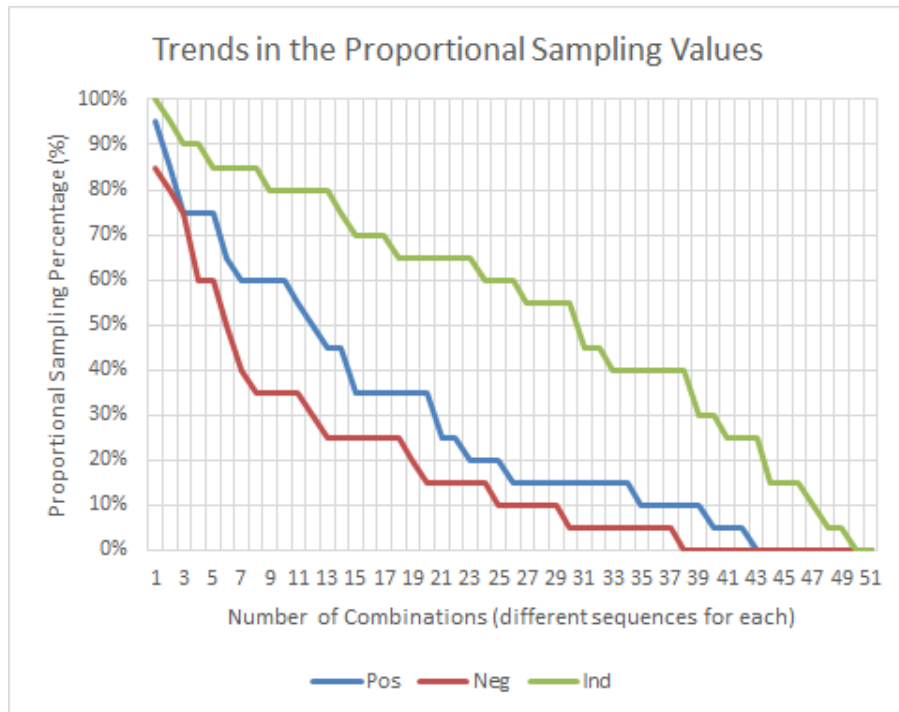
Figure 2.16: Trends in the sampling rates for F_{PosDis}, F_{NegDis}, and F_{UncDis}

Two examples illustrate these results in more detail. The first is for Fair-4, Fair-1, Fair-2 and Fair-6. There are 22 observations of this trip. The α, β, and γ values found as being optimal are 85%, 10%, and 5% respectively. The cumulative histograms for the individual segment travel times and the overall route travel time are shown in Figure 2.17. A cumulative histogram is just a CDF but it is not normalized to 1. This allows the distribution of the observations themselves to be shown. All three segment travel times are about of equal magnitude although their distributions are significantly different as seen in Figure 2.17.

41
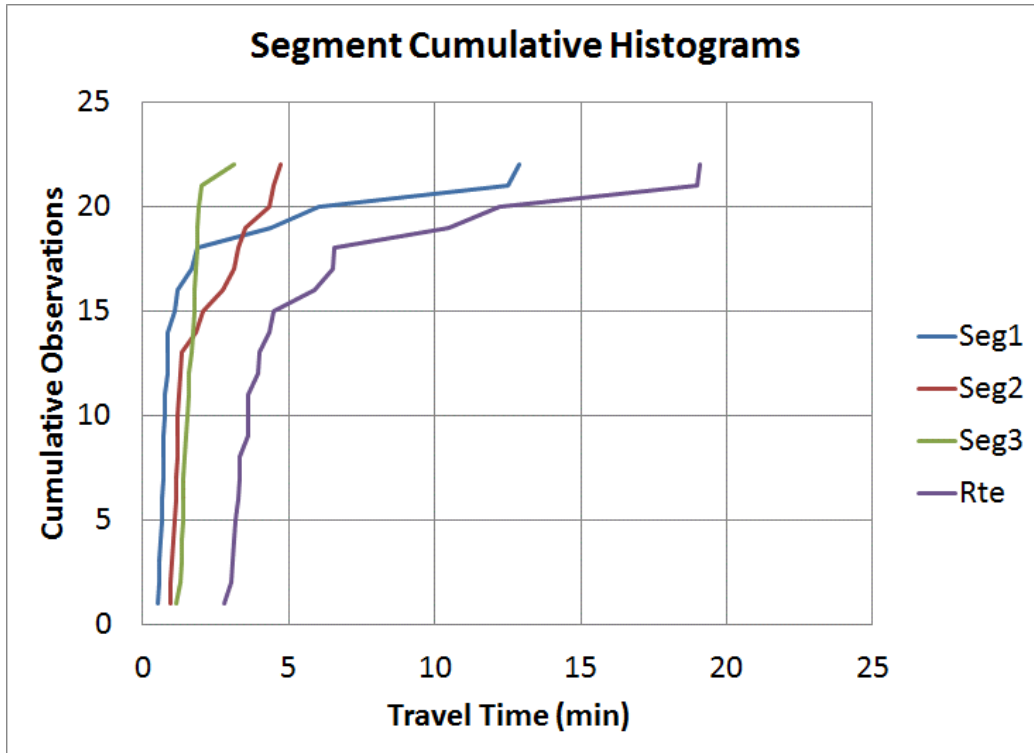
Figure 2.17: Cumulative histograms for route Fair-4, Fair-1, Fair-2, Fair-6

To help with understanding, the F_{PosDis}, F_{NegDis}, and F_{UncDis} distributions are shown in part (a) of Figure 2.18 and a comparison of the synthesized route travel time distribution with the actual distribution is shown in part (b). Neither the F_{PosDis}, F_{NegDis}, nor F_{UncDis} distributions match the observed distribution. A sampling of the three is required. The synthesized distribution shown in part (b) has sampling rates of 85%, 10%, and 5%. By visual inspection, the figure shows graphically that there is a good match between the synthesized and observed distributions.
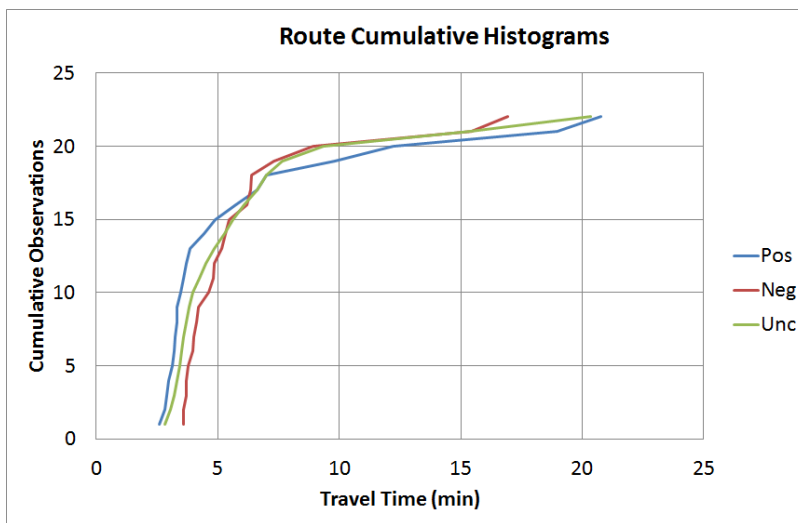
(a) The F_{PosDis}, F_{NegDis}, and F_{UncDis} Distributions
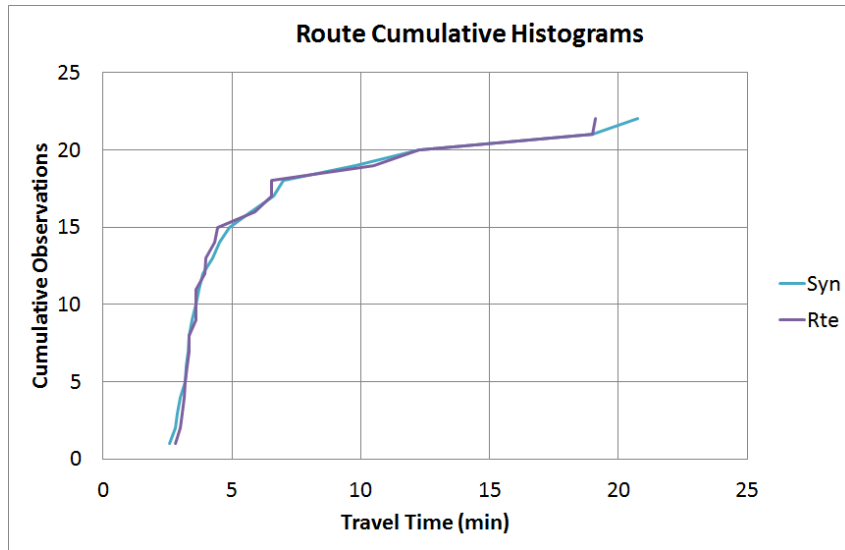
(b)   The Observed and Synthesized Distributions

Figure 2.18: Observed and synthesized cumulative histograms for Fair-4, Fair-1, Fair-2, Fair-6

The second example route is Fair-3, Fair-4, Fair-1. In this case, there are 204 observations. The α, β, and γ values are 0%, 60%, and 40% respectively, meaning the contribution from the F_{NegDis} distribution is quite large. The cumulative histograms for the two segment travel times and the overall route travel time are shown in Figure 2.19. From the figure, three segment travel times are about of equal magnitude despite their cumulative distributions being significantly different.
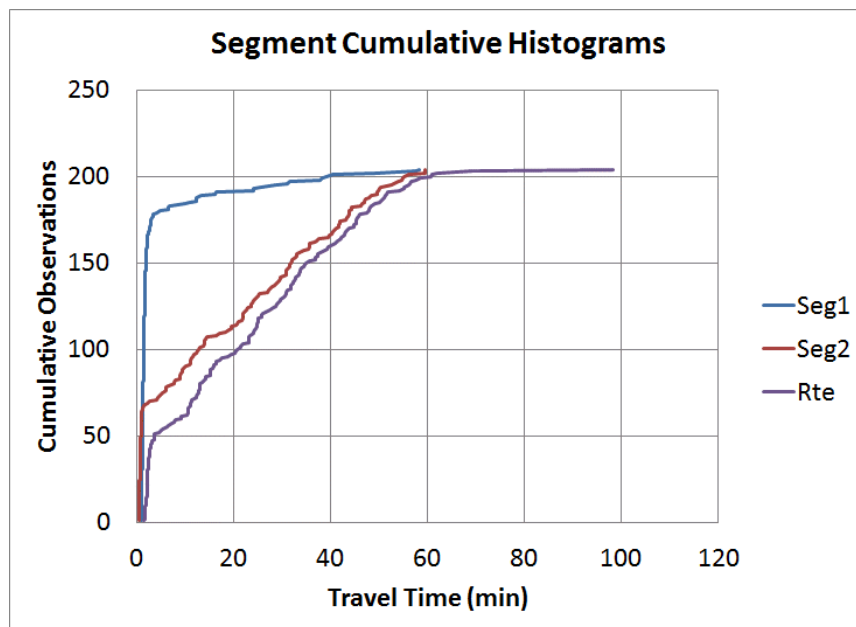
Figure 2.19: Cumulative histograms for the route Fair-3, Fair-4, Fair-1

43

The F_{PosDis}, F_{NegDis}, and F_{UncDis} distributions are shown in part (a) of Figure 2.20. Part (b) presents a comparison of the synthesized route travel time distribution with the actual distribution. It is again clear that there is a good match between the synthesized and observed distributions.
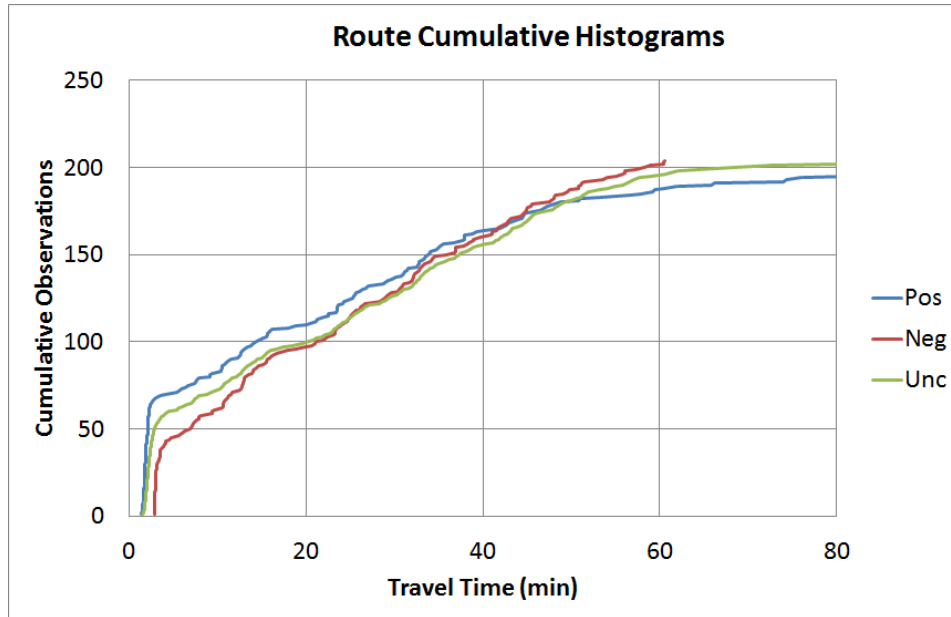
(a) The F_{PosDis}, F_{NegDis}, and F_{UncDis} Distributions
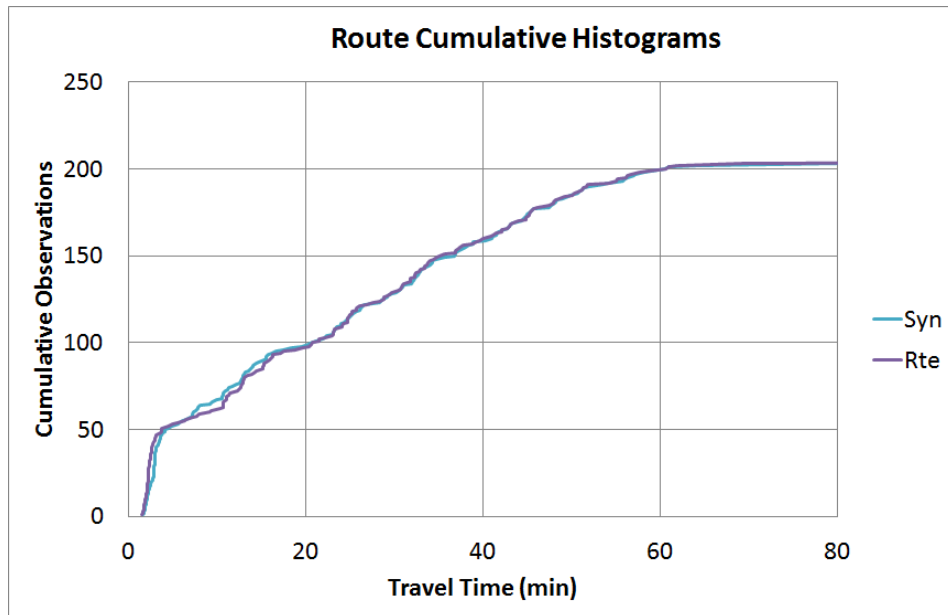
(b) The Observed and Synthesized Distributions

Figure 2.20: Observed and synthesized cumulative histograms for Fair-3, Fair-4, Fair-1

The third step is to develop an estimate of the travel time distribution for the trucks. Of course, in the future, once the trips are being made, they will implicitly a component of the distribution of travel times for vehicles on the route. Thus, the travel times for the trucks will be among the observations upon which the route-level travel time is based. Said conversely, the overall distribution is an amalgamation of distributions for specific types or classes of vehicles. One of these classes will be the trucks for which the travel time distribution is desired. But that is not the case before the trips start to take place. Hence, the distribution must be estimated. The simplest way to obtain this distribution is to use the mean of the distribution obtained from the synthesis exercise in step #2, multiply it by the mean adjustment factor obtained in step #1 and then sample travel times from the normalized distribution also developed in step #1. This is illustrated below.

Let's assume the route of interest is Fair-4, Fair-1, Fair-2, Fair-6. This is the one for which the segment distributions are shown in Figure 2.17 and the overall route distribution is shown in Figure 2.16(b). The average travel time is 6.02 minutes. Assuming, as was done before, that the ratio of the average truck trip time to the average overall travel time is 1.3 and the distribution of truck trip times is shown in Figure 2.6, then the truck trip time distribution for this route is shown in Figure 2.21.
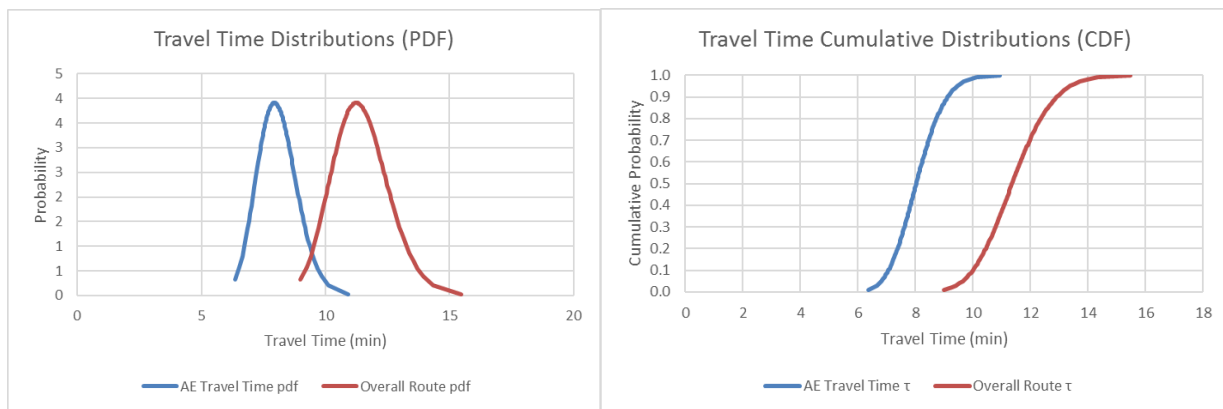
(a) Hypothetical PDF   (b) Hypothetical CDF

Figure 2.21: Hypothetical truck trip time distribution for Fair-4, Fair-1, Fair-2, Fair-6

2.3.3 Method 3: Monte Carlo Simulation

The third method uses segment-specific Monte Carlo sampling to synthesize route-level travel time distributions. The method is intuitively appealing because it capitalizes on intuitively appealing ideas about how individual vehicle travel times arise on congested and uncongested networks.

The method's main assumption is that a vehicle's travel time arises from three behavioral properties. The first is that when vehicles are traversing segments in an uncongested state, the travel time they achieve reflects driving behavior. The second is that when vehicles are traversing congested segments, the travel time is randomly determined. It does not reflect driving behavior. The third is that a mix of these conditions pertains to vehicles on a given segment. That is, the

45

segment travel time distribution is a blend of travel times derived from distributions for the two states. Even though the segment may be labeled uncongested, some vehicle travel times can come from the congested distribution. And although a segment may be labeled congested, some travel times can come from the uncongested distribution.

Extending these thoughts, vehicle-specific route travel times, t_{vr}, can be synthesized by sampling values on uncongested segments that reflect the driver's desired speed. And on congested segments, by random selection. Mathematically, let r_d be a random variable on the interval [0,1] that reflects the driver's driving style, and $t_{vs} \mid r_{vr}$ is the travel time for the vth vehicle and the rth percentile. Similarly, let r_r be a random variable on the interval [0,1] that reflects a randomly drawn travel time, then $t_{vs} \mid r_{vr}$ is a random travel time. Driver behavior has no influence. When the vehicle is in an uncongested state, t_s is the sampled travel time from the uncongested travel time CDF. When the vehicle is in the congested state, t_{vs} is the sampled travel time from the congested travel time CDF. In other words:

$$t_{vr} = \sum_s t_{vs} \quad \text{where} \quad t_{vs} = \begin{cases} F_{su}^{-1}(r_d) \text{ if vehicle } v \text{ is in the uncongested state and} \\ \quad F_{sc}^{-1}(r_r) \text{ if vehicle } v \text{ is in the congested state} \end{cases} \tag{2.8}$$

$F_{su}(t_{vs})$ and $F_{sc}(t_{vs})$ are the uncongested and congested travel time CDFs on segment s. $F_{su}^{-1}(r_d)$ and $F_{sc}^{-1}(r_r)$ are the inverses.

Figure 2.22 shows the congested and uncongested distributions based for a 5-mile stretch of I-5 northbound in Sacramento, just south of downtown. As can be seen, the uncongested travel time distributions have much shorter travel times. And weather or incidents do not seem to have a significant impact. But in the case of the congested distributions, incidents and weather both have significant effects on the travel times, especially for the middle percentiles.
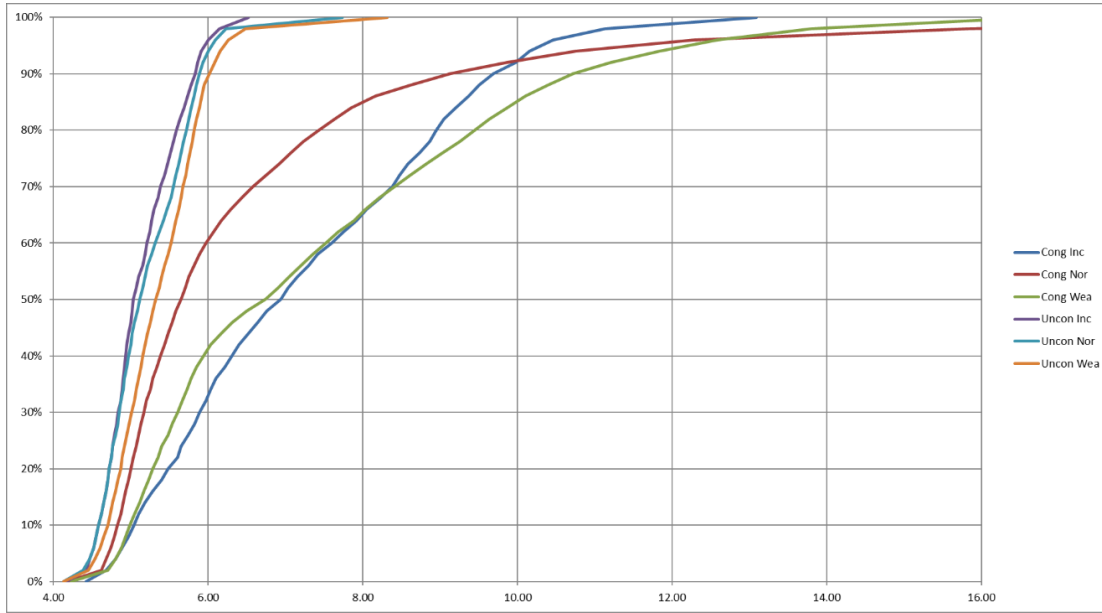
Figure 2.22: CDFs for a 5-mile stretch of I-5 in Sacramento, south of downtown

46

It is important to remember that the CDFs in Figure 2.22 are for individual vehicle travel times not average travel times. This means there is a need to check and see if these CDFs capture day-to-day variations as the average travel time CDFs do, or if they are depicting typical days. Figure 2.23 helps with this determination. It shows the temporal trends for the 5th, 25th, 50th, 75th, and 95th percentiles of the travel times across two days. To obtain these percentiles, temporally successive half-overlapping sets of 50 individual vehicle observations were analyzed. The percentiles are identified by sorting the observations into ascending order and then either selecting a specific observation (in the case of the 50th percentile) or interpolation between two observations (for the others). In the figure, the results are plotted against time using the mid-point timestamp of each set of 50 observations. The interval between the reported distributions varies because the traffic volumes change across the day. At night, when the traffic flow rates are low, it takes time to accumulate 50 probe observations, so the interval between plotted distributions is larger than it is in the daytime. In the peaks, when the flow rates are high, the plotted data points are much closer together.

What Figure 2.23 shows is that the peak and off-peak conditions are similar each day. In the off-peak, there is not much variation in the travel time percentiles. And in the peak hours, the percentiles follow a consistent transient where they increase and then decrease.
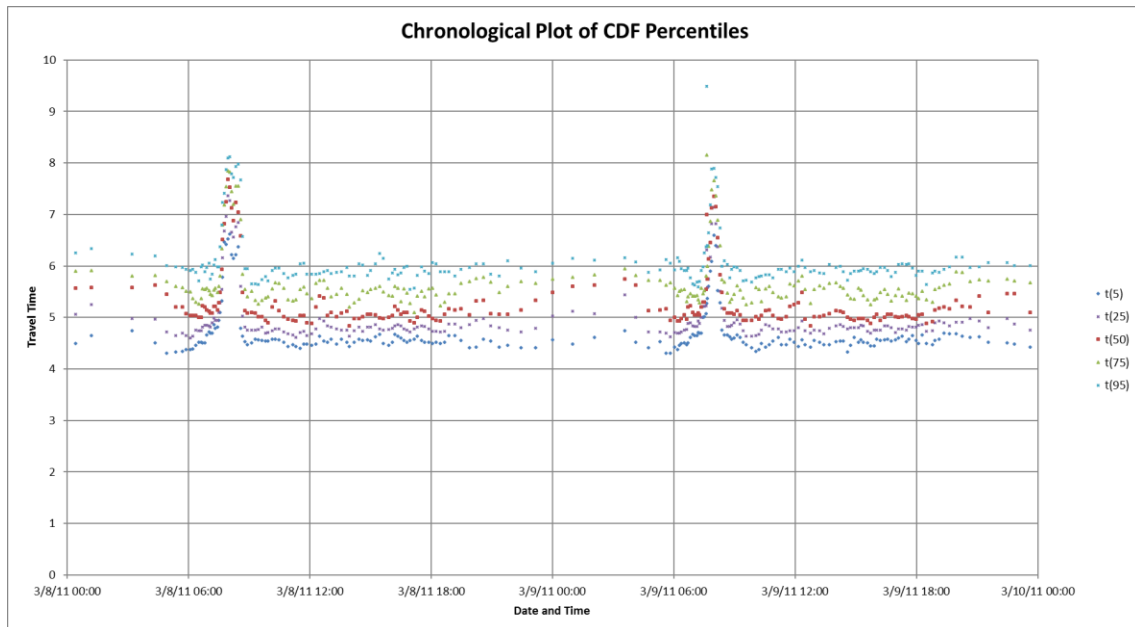
Figure 2.23: Uncongested and congested travel time CDFs for three routes in San Diego

The implication of Figure 2.23 is that the CDFs shown in Figure 2.22 can be treated as depictions of individual vehicle travel time distributions for a typical day, not reflections of day-to-day variations in the travel times as is the case with the CDFs for the average travel times. The figure also suggests (as is intuitively obvious) that different CDFs for individual vehicle travel times exist across a typical day as the congestion conditions increase and decrease.

The method involves two steps to create the predicted truck travel time distribution. The first, which is akin to step #1 in methods #1 and #2, involve creating CDFs of the travel rates

experienced by the trucks when traffic conditions are congested and uncongested (and perhaps for several other operating conditions like weather- and incident-affected). These distributions will look like the ones shown in Figure 2.22 except that they will reflect travel rates rather than travel times (times per unit distance). Ideally, these distributions will be for the segments in the route being examined, but generic distributions or ones from other segments are also suitable to use.

The second step is to synthesize hypothetical trip travel times by summing Monte Carlo sampled segment-specific rates which are then multiplied by the segment lengths. To illustrate, assume a route like the one shown in Figure 2.13 is of interest. Starting at O and working towards D, travel times are sampled for each segment. In the case of segment OA, if no individual vehicle data are available, the same technique used in method #1 can be employed (take the mean multiplied by the truck adjustment factor and then sample from the truck-related travel time ratio distribution. For segments AB through DE, where vehicle-specific data are available, ascertain whether the segment's operating condition is congested or uncongested, and then apply

$$X_R^p = \sum_s X_s^n \quad \forall \ p \quad \text{where } n = p \text{ if } s \text{ is odd and } n = 1 - p \text{ if } s \text{ is even} \tag{2.7.}$$

For segment ED, apply the same technique as used for segment OA. This synthesizing process is repeated a countably infinite number of times and the distribution of these values is ascertained.

As a postscript, if actual distributions of the truck travel times (or rates) exist for the segments in the route (for whatever other operating conditions are of interest), these distributions should be used instead. The distributions can be sampled as appropriate to develop the segment-specific travel times. A separate thought is that whether a segment is operating congested or uncongested is *a different thought* than whether the time of day is a peak period or not. For example, along a freeway, even during the peak period, some segments operate uncongested even though others are congested. The segments with bottlenecks operate congested as do some segments upstream; segments without bottlenecks or queues operate uncongested.

Use of this technique can be illustrated by an example. Assume the path of interest is the one from O to D shown in Figure 2.13. Also, assume individual vehicle travel time distributions are available for all six segments. Table 2.2 provides details.

The top part of the table shows the means and standard deviations for the lognormal distributions of the uncongested and congested travel times for the segments. Parenthetically, these are truck travel times. The cells in the middle of the table show the states of the vehicles as they traverse the route. The clear colored cells in the segment columns show the percentage of trucks in the uncongested and congested states for each segment. The salmon colored cells show the transitions between states from one segment to the next. The numbers in the cells show that in this example, 100% of the trucks experience a congested travel time on segment OA. 50% of these trucks then transition to uncongested operation on segment AB and 50% of them stay in the congested state. This means that on segment AB, 50% of the trucks experience uncongested travel times and 50% experience congested times. In transitioning to segment BC, of the 50% of trucks that experience uncongested travel times on AB, 45% remain in the uncongested state on segment BC (i.e., 90% of the 50%). 5% transition to the congested state (10% of the 50%). Of the 50% of trucks that experience congested travel times on AB, 45% transition to the uncongested state on segment BC. 5% stay in the congested state.

48

		Segment											
		1	xfer	2	xfer	3	xfer	4	xfer	5	xfer	6	
U	Mean	0.75		0.85		1.05		0.90		0.90		0.85	
	StdDev	0.05		0.08		0.06		0.07		0.07		0.07	
C	Mean	1.00		1.05		1.25		1.25		1.20		1.15	
	StdDev	0.10		0.08		0.12		0.12		0.11		0.13	
Probs	U	0%	0%	50%	45%	90%	10%	20%	10%	10%	5%	15%	
			0%		5%		80%		10%		5%		
	C	100%	50%	50%	45%	10%	10%	80%	0%	90%	10%	85%	
			50%		5%		0%		80%		80%		

Veh	rv1	t(AB)	t(BC)	t(CD)	t(DE)	t(EF)	t(FG)	t(AG)
1	0	2.7	2.8	2.8	3.6	3.1	3.5	18.6
2	0	2.6	3.1	2.9	3.3	3.8	2.9	18.6
3	0	2.4	3.0	2.7	4.2	3.0	3.1	18.5
4	0	2.6	3.0	2.8	3.5	3.2	2.9	18.1
5	0	2.7	2.8	4.2	2.5	2.5	4.0	18.6
6	0	2.5	2.5	3.1	3.5	2.9	3.6	18.1
7	0	2.7	2.4	2.6	3.6	2.7	2.9	17.0
8	0	3.1	2.6	3.1	3.0	3.3	2.9	17.9

Table 2.2: Six segment problem description

Figure 2.24 shows the resulting route travel times for 200 hypothetical trucks. The values are plotted one on top of another so that the individual segment travel times can be observed, as well as the overall route travel time. The segment travel times are typically on the order of 3-5 minutes and the overall trip time is about 18 minutes.
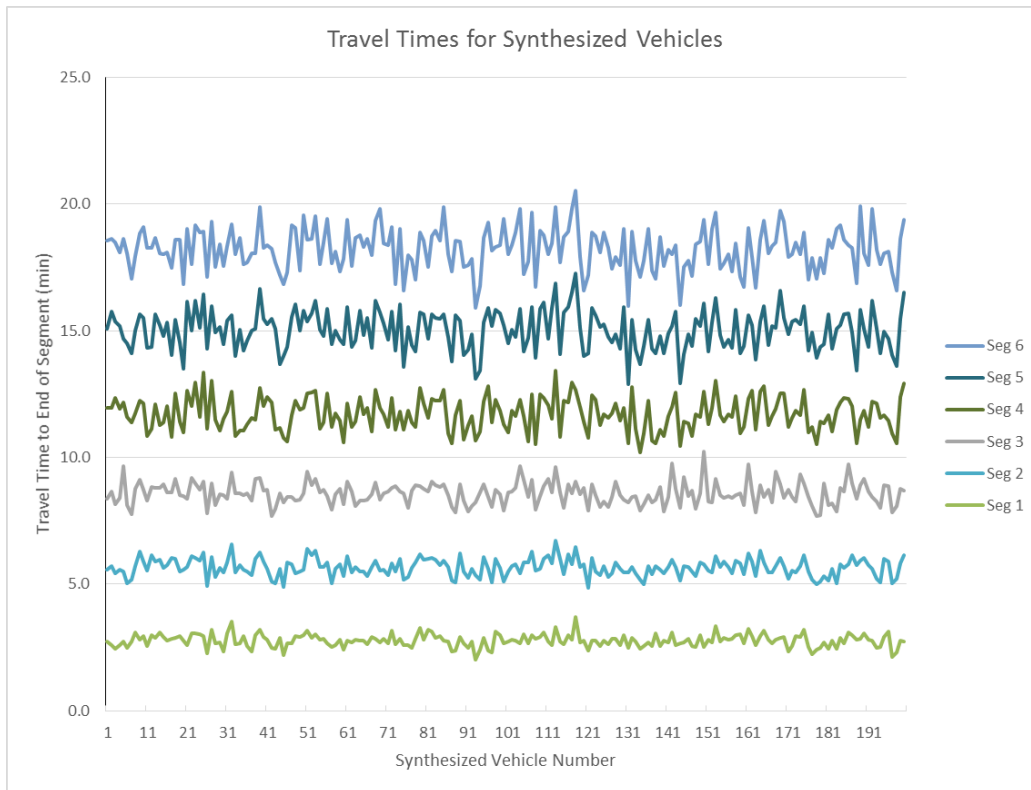
Figure 2.24: Travel time values for 200 synthesized trucks

49

Figure 2.25 shows the cumulative histogram of the truck trip times. It is marked with the designation "Rte". The distributions for positively correlated times, negatively correlated times, and uncorrelated times are also shown.
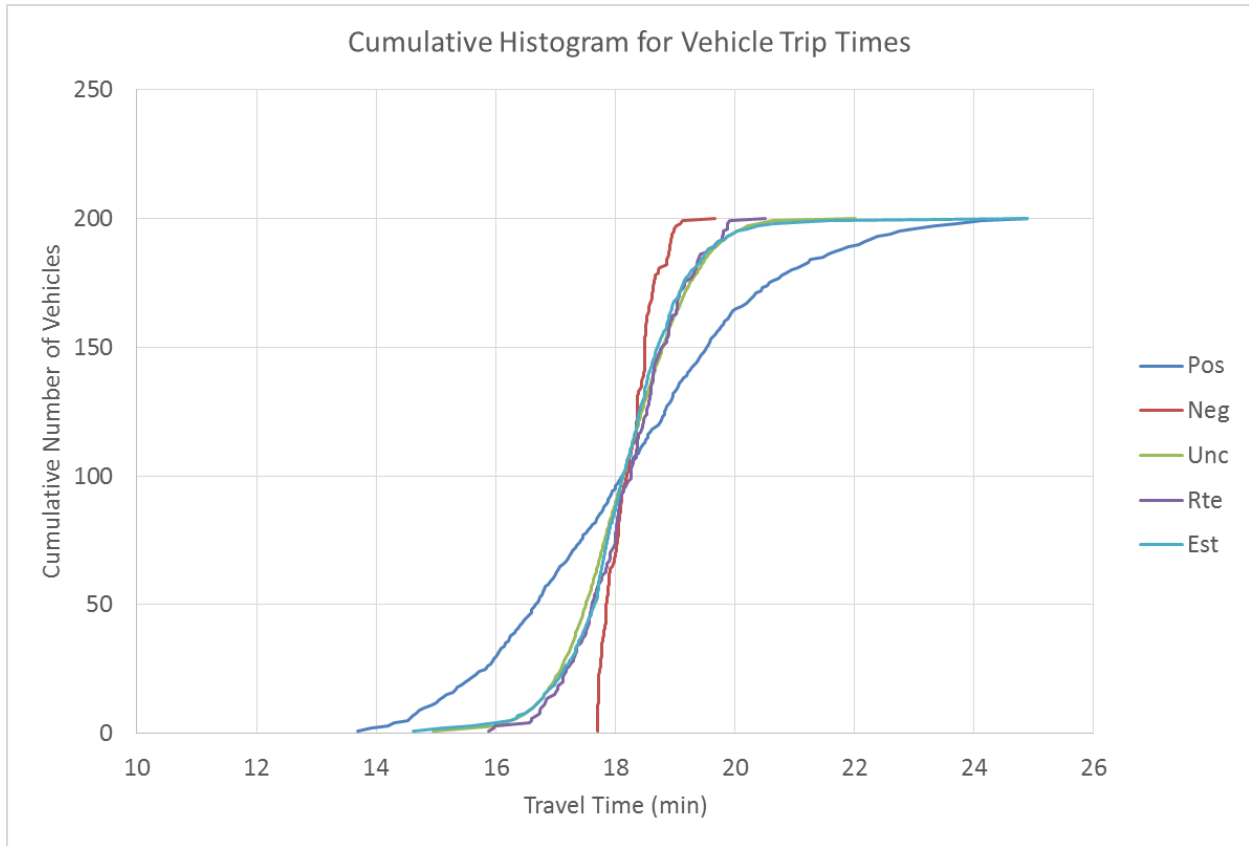
Figure 2.25: Travel time values for 200 synthesized trucks

2.4   POSTSCRIPT: ASSESSING ROUTE STANDARD DEVIATIONS

In some instances, knowing the entire distribution of the travel times is not that important. Planning models, for instance, only need to know the variance of a given route. For route choice, they use a generalized cost functions that applies a weight to the path's variance. This section presents a method for computing the variance of a given route.

Because the segment travel time distributions are often correlated, as indicated previously, the variance of the overall route travel time (i.e., the sum of the segment travel times) is not likely to be the sum of the variances for the route segments. The sum of the covariances must be used. That is, if the distribution of travel times on segment i is given by X_i $i \in \{1,.....,n\}$, then:

$$Var\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} Cov(X_i, X_j) \neq \sum_{i=1}^{n} Var(X_i) \qquad (2.9)$$

To illustrate how the standard deviation can be computed, a few assumptions are needed. First, there is an assumption that there are k travel time samples for each of the n segments in the route

50

$R = \{1, ... n\}$. For the moment, it is assumed each segment has the same number of observations. The sample covariance for each pairwise combination of segments can then be calculated as:

$$s_{X_i, X_j} = \frac{\sum_{m=1}^{k}(X_{im} - \bar{X}_i)(X_{jm} - \bar{X}_j)}{k-1} \tag{2.10}$$

Where $\bar{X}_i$ and $\bar{X}_j$ are the sample means for segments i and j, and X_{im} and X_{jm} are the mth replications for segments i and j respectively. The estimate of the route variance is then:

$$Var\left(\sum_{i=1}^{n} X_i\right) \approx \sum_{i=1}^{n}\sum_{j=1}^{n} s_{X_i, X_j} \tag{2.11}$$

Since these sample covariances are straightforward to compute and provide unbiased estimators of the true covariances, this is an appealing way to proceed.

However, two issues need to be addressed.

The first is that only a subset of all the vehicles observed traverse the entire route. Other vehicles traverse other sequences. For example, if the route contains segments $i = (1,2,3,4,5)$ there may be vehicles that traverse segment sequences (1,2,3,4), (2,3,4,5), (1,2,3), (2,3,4), (3,4,5), (1,2), (2,3), (3,4), and (4,5) as well as single segments (1), (2), (3), (4), and (5). This means some segments will have more observations than others. That is, the fixed k above becomes k_i for each of the n segments and it is likely that $k_p \neq k_q$ for segments $p \neq q$ in the route. While it is appealing to think that all the segment-level travel time observations can be used to estimate the route variance, there isn't a standard way to utilize all available data. Therefore, creating a method that allows all observations to be used justifiably is the focus of this section.

The second issue regards robustness. Outliers can have a large impact on estimation of route variance. A vehicle that is extremely slow or fast can have a significant effect on the estimated value of the covariances. Over-estimated values then have a cascading effect on the total route variance, providing a lower quality estimate of the true value. Ideally such outliers can be removed in preprocessing, but determining which observations to remove requires knowing a priori which distributions the data follow (unknown in most cases). Ways have not yet been identified to address this issue beyond basic statistical analyses in either a frequentist or Bayesian framework.

In numerical experiments, options are first explored by considering a sequence of four segments where the synthesized travel times were known to be independent for all segment sequences traversed. An Excel workbook with VBA code was developed to generate hypothetical data based on known parameters. It is assumed each segment's travel time followed a lognormal distribution with mean 2 and standard deviation 0.5. Sequences of segment observations were isolated to create the set T of all possible subsequences:

$T = \{(1),(2),(3),(4),(1,2),(2,3),(3,4),(1,2,3),(2,3,4),(1,2,3,4) \}$

Sampling with and with and without replacement were then explored. In sampling without replacement, only the travel times for those vehicles that traversed the entire segment sequence (1,2,3,4) were utilized. In sampling with replacement, the desired sample size was arbitrarily set to 1000 and segments with fewer observations were over-sampled (sampled with replacement) to impute values for the voids. In the case of sampling without replacement, the resulting credible interval for the route variance is one whose width is a function of the number samples employed. In the case of sampling with replacement, the credible interval is artificially reduced in size as the number of samples drawn increases.

Two covariance matrices are computed based on the sampled data. The first, C^0, was based on the observations that traversed the entire route. Sampling without replacement was employed. The second, C^1, was calculated based on sampling with replacement. Total route variance was then calculated for each method following equation 2.8 as:

$$s_0^2 = \sum_{i=1}^{n}\sum_{j=1}^{n}C_{ij}^0 \quad , \quad s_1^2 = \sum_{i=1}^{n}\sum_{j=1}^{n}C_{ij}^1 \qquad (2.12)$$

The finding was that sampling with replacement yielded a variance close to the true value. In fact, s_1^2 was often closer to the true route variance than the value obtained by using only the full-route data (i.e., s_0^2).

Observations are next explored where the travel times were positively correlated. It is assumed the correlations decreased by half for each lag-1 advance. That is, segment 1 had a correlation of 1 with itself, 0.5 with segment 2, 0.25 with 3, and 0.125 with 4. Similarly, segment 2 had a correlation of 1 with itself, 0.5 with 1 and 3, 0.25 with 4; etc. Based on data synthesized consistent with these assumptions, covariance matrices are computed without and with replacement as was done before:

$$C^0 = \begin{pmatrix} 19.531547 & 9.214472 & 4.912257 & 2.047562 \\ 9.214472 & 18.420215 & 8.946405 & 4.555028 \\ 4.912257 & 8.946405 & 20.445868 & 8.672634 \\ 2.047562 & 4.555028 & 8.672634 & 18.331467 \end{pmatrix}$$

$$C^1 = \begin{pmatrix} 18.2306275 & -0.3279111 & 0.3938785 & -0.8332173 \\ -0.3279111 & 16.5551669 & 1.7468205 & 0.6830507 \\ 0.3938785 & 1.7468205 & 19.9974095 & -0.5279992 \\ -0.8332173 & 0.6830507 & -0.5279992 & 20.0857689 \end{pmatrix}$$

As can be seen, the sampling with replacement in matrix C^1 does not work well when applied to correlated observations. While the diagonal (variance) elements are relatively close to the true values, the off-diagonal terms are vastly different from the true values. See Figure 2.26 below for a comparison.

52
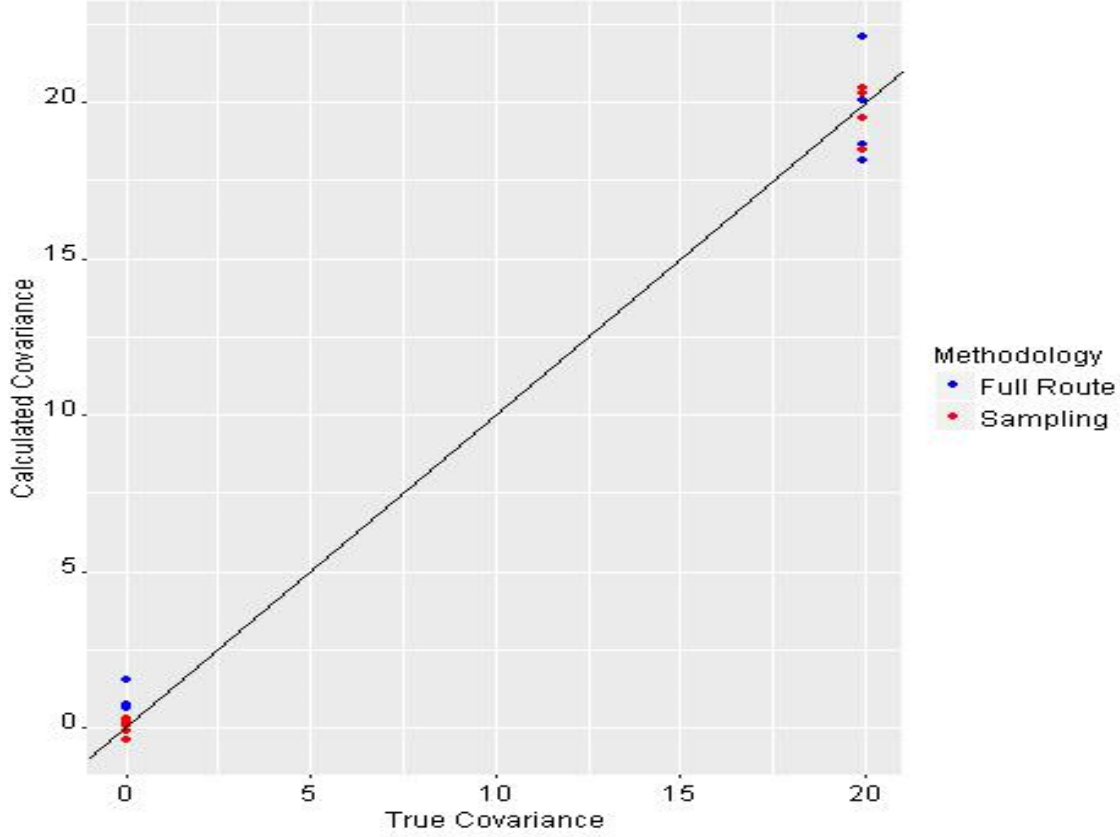
Figure 2.26: Covariance value comparisons for correlated data

An analysis of these results showed that the sampling with replacement from the available segment data served to obscure the correlation among segments, i.e. random draws served to "break" the relationships among the segment times, forcing the $i \neq j$ covariances to tend towards an independent zero.

To address this deficiency, an alternative methodology was explored. It generates a composite covariance matrix $\hat{C}$ by combining covariances computed for all the sub-routes. Three steps are involved. First, the data are sorted into sub-routes. Second, each of the 1×1 up to 4×4 covariance matrices are computed for each of the sub-routes. Third, weighted covariance values are computed based on the number of vehicles associated with each sub-route. For instance, to estimate element $\hat{C}_{12}$ the aggregated covariance is computed as:

$$\hat{C}_{12} = \frac{n_{(1,2)}C_{12}^{(1,2)} + n_{(1,2,3)}C_{12}^{(1,2,3)} + n_{(1,2,3,4)}C_{12}^{(1,2,3,4)}}{n_{(1,2)} + n_{(1,2,3)} + n_{(1,2,3,4)}} \tag{2.13}$$

where $n_{(i,j,k)}$ is the number of observations for route (i, j, k) and $C_{ij}^{(i,j,k)}$ is the sample covariance for segments i and j calculated from only sub-route (i, j, k). This procedure is repeated for all

segment pairs (i,j) to yield the full aggregated covariance matrix $\hat{C}$. As the covariance matrix is symmetric not all n^2 elements are computed, but rather the elements of the upper triangular matrix.

The results from employing this composite procedure have been compared to the prior two methods using an independent set of generated data. Figure 2.27 shows the scatterplots for the three methods (full route C^0, sampling with replacement C^1, and composite C^2) compared to the true covariance values. Clearly method C^1 displays the errant result of independence for the data points under 10 (from correlation to no correlation as its values approach zero). This example has a true route variance of approximately 162.8 while the values of 154.2, 79.1, and 159.79 for C^0, C^1, and C^2 are computed respectively. Clearly, the composite method yields a value close to the actual one with C^0 still performing relatively well, even though it drops many observations in favor of only tracking vehicles that provide observations for all segments.
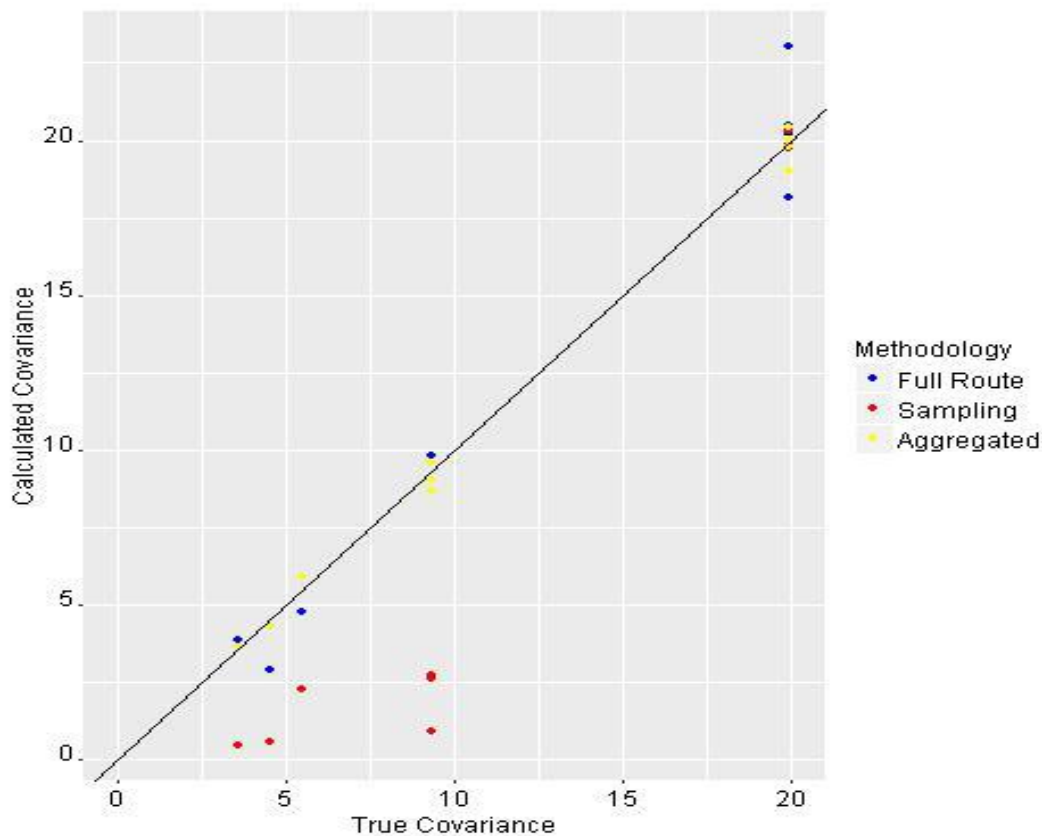
Figure 2.27: Covariance value comparisons for correlated data

The conclusion reached is that the aggregation methodology that generates C^2 provides a reasonable estimate of the true covariance matrix, often a better estimate of the true full route variance when compared to the methodology that only uses full route observations to generate C^0. In sum, the aggregation scheme provides a simple method to effectively estimate route variance, providing insight into a route's overall reliability.

2.5   SUMMARY

This section has focused on the characterization of travel time reliability for truck trips between specific origin-destination (OD) pairs. Three methods have been presented for developing estimates of these distributions for specific operating conditions. The first makes use of data about the average travel times along segments in the path, prior experience with truck travel time distributions on other routes, and inference. The second uses individual vehicle travel time observations among the segments and a synthesizing strategy that combines the segment-level distributions to estimate a route-level distribution. The third uses Monte Carlo simulation, assumptions about the congestion level on each of the segments, and a hypothesis about how the congestion level influences the travel time experienced by the truck. All three methods are useful in different settings depending upon the type of data available. These methods form the basis for the analyses in the subsequent sections because those efforts are entirely reliant upon having valid estimates of the truck travel time distributions on various network paths.

3.0   CHOOSING PATHS AND DEPARTURE TIMES

After the reliability of segments and routes has been assessed, the next challenge is to identify reliable paths. For a trip between a given OD pair, if the AW is specified, the most reliable path maximizes on-time performance. If both an AW and a DW are specified, then the best path maximizes the probabilities of achieving both an OTD and an OTA.

Of course, there is often a tradeoff between reliability and travel time (and implicitly, cost). A path with a longer travel time may provide better reliability, but also a higher travel time, and greater cost. Padding a trip with slack time at the destination also improves reliability, but still with a cost. Hence, there are likely to be tradeoffs. Anticipating those tradeoffs is important. Using a bi-criterion search is important. (And there may be other criteria such as minimizing the likelihood of exposure to accidents and incidents that should be considered, but those are not examined here.)

Once the non-dominated paths fitting the bi-criterion have been identified, a utility function or some other evaluation and selection process can be employed to identify the best path to choose.

3.1   RELEVANT LITERATURE

Chen *et al.* (2011) used the travel time budget as the metric for determining reliability-based user equilibrium (RUE) rather than the typical metric of expected travel time. This allows for inclusion of differing degrees of risk-aversion for distinct user classes evaluated in the system. A column generation technique was used to enumerate over all OD and user class pairs followed by solving the associated restricted subproblems by iteratively shifting flows from the costliest path with maximum travel time budgets to the cheapest path having minimum budget. The cost itself is established using a mean budget (M-B) dominance condition described within the paper.

Chen *et al.* (2013) solved the reliable shortest path problem (RSPP) for α - reliable path wherein the goal is to minimize the travel time budget while ensuring an α level of on-time arrival probability. Link travel times follow normal distributions to allow for an analytical reliability measure when evaluating routes. A stochastic-based dominance condition is described to effectively extend Bellman's Principle of Optimality to the stochastic case. Two algorithms are used to solve the problem, notably a multi-criteria label setting algorithm like Dijkstra's algorithm and the multi-criteria A* algorithm. The A* algorithm inherently favor nodes likely to be on α - reliable paths by assigning higher priority in the search space for these nodes, maintaining a set of eligible non-dominated path sets ordered in descending likelihood of appearing on the α - reliable paths. A case study of Hong Kong shows the computational time advantage of the A* algorithm over the more accurate yet more computationally demanding multi-criterial label setting algorithm.

Huang and Gao (2012) investigated stochastic time-dependent (STD) networks with temporal and spatial correlation among links, using a minimum expected disutility (MED) to evaluate routes. Due to the stochasticity, the problem violates Bellman's Principle of Optimality and the Algorithm CD-Path is designed to find only "pure paths" whose sub-paths are non-dominated. This algorithm iteratively trims the search space of dominated paths until the final solution set only contains non-dominated paths.

Ji, Kim, and Chen (2011) described a simulation-based multi-objective genetic algorithm (SMOGA) that is used to find non-dominant (Pareto optimal) paths. This algorithm consists of a Monte Carlo simulation of correlated travel times, a genetic algorithm to explore the combinatorial solution space, and a Pareto solution filter to maximize the diversity of the Pareto solutions. SMOGA solves the chance constrained multi-objective programming (CCMOP) model for optimal path finding while simultaneously minimizing the travel time budget and satisfying travel time reliability constraints. Numerical experiments on the Chicago Sketch network show feasibility and diversity in explored solution space, while further showing that correlation among link travel times create significant discrepancies in travel time budgets. Without correlation among links, Pareto paths resulting have significant travel time budget bias and provide sub-optimal paths.

Srinivasan *et al.* (2014) solved the most reliable path problem with the added feature of shifted log-normal link travel times (MRP-SLN), a constrained nonlinear integer programming problem. The MRP-SLN algorithm uses lower and upper bounds on path reliability measures to force convergence. A sufficient condition is devised to guarantee that the most reliable path is present in a set of least expected travel time paths. A case study of Chennai city in India was examined, and the generated set of paths contains the true optimum in more than 98% of tested OD pairs with an average relative gap between proposed and true optimum paths less than 0.06%. These sets are generated in under an average of 25 seconds. However, using approximations for normal and lognormal times at link and path level lead to sub-optimal solutions in 14% and 12% of cases respectively, with reliability decreases up to 9%.

Xing and Zhou (2011) sought to answer the most reliable path problem under varying spatial correlation assumptions, with total path travel time variability represented by standard deviation. Lagrangian substitution is used to estimate the lower bound of the most reliable path by solving a sequence of shortest path problems, followed by a subgradient descent to iteratively reduce the optimality condition between primal and dual solutions until a termination condition occurs. Further, when spatial correlation exists among link travel times, a sampling-based solution algorithm is embedded in the above Lagrangian technique. A case study of the Bayshore Freeway between Mountain View and San Jose, California was examined. These experiments showed that utilizing these reformulated models on a large-scale network allows for 10-20 iterations of standard shortest path algorithms to reach duality gaps of about 2-6% for uncorrelated travel times.

3.2   ASSESSMENT METHODOLOGY

A multi-step process can be used to find optimal departure times and paths. The process is as follows:

1) Solve a deterministic K-shortest path problem working backward from the OTW. In doing this, use the midpoint of the OTW as the nominal arrival time and use the median travel times as the path travel times (or a higher percentile for lower risk tolerance).
2) Develop a relationship between departure time and the probability of arrival during the AW for all the K-shortest paths identified.
3) Select the path and departure time that provides the best combination of travel time and reliability based on the risk and travel time preferences of the decision maker.

The first step is to find the K-shortest paths. The K-shortest path problem is a generalization of the shortest path problem. The objective is to find not only the shortest path but also K-1 other paths that are in non-decreasing order of cost. K is the number of shortest paths to find. The problem can be restricted so that the paths have no loops (nodes repeat) or loops can be allowed.

The earliest study of the K-shortest path problem dates from Bock, Kantner, and Haynes (1957). It has been studied extensively as illustrated by the reviews in Yen (1971) and Eppstein (1997). Eppstein (1997) is considered by many to have identified the algorithm that produces the best results.

An extension of Dijkstra's algorithm can be used to find the K Shortest paths. For a single path, the algorithm is as follows:

1) Assign to every node a tentative distance value: set it to zero for the origin node and to infinity for all other nodes. Also, assign a tentative predecessor node and set its value to null.
2) Set the initial node as the *current* node. Mark its predecessor as being the initial node. Mark all other nodes as being unvisited. Create a list of all the unvisited nodes.
3) For the current node, consider all its unvisited neighbors and calculate their *tentative* distances. Compare this *tentative* distance with the currently assigned value (which initially is infinity) and keep the smaller value. Also, update the predecessor node to be the current node.
4) When all the neighboring unvisited nodes for the current have been considered, mark the current node as *visited* and remove it from the list of *unvisited* nodes. Nodes that are marked *visited* set are never checked again. Their shortest distance has been found.
5) If the current node is the destination node, then stop. The algorithm has found the shortest path to that node.
6) Otherwise, select the next unvisited node that has the smallest tentative distance. Set it as the *current* node and go back to step 3.

The extension can be described as follows. It involves keeping a K-long vector of the shortest distances to each node. It also involves keeping a K-long vector of predecessor nodes. In step 1, all the shortest distance cell values are initially set to infinity. Also in step 1, the cell values in the vector of predecessor nodes are set to null. In step 3, the current node is redefined as a node and a cell position in the shortest distances vector rather than just a node. This means the extensions are from the current node and cell position, not just from the node. Moreover, the *visited* label is ascribed to a node and a cell position not just a node. This means a node must be examined K times before all its cell values are marked as having been *visited*. (When this happens, all the K-shortest distances to the node have been identified.) Distances from the current node and cell position are compared to the current vector of shortest distances for all neighboring nodes. If the new distance is shorter than the distance in cell i of the neighboring node's distance vector, then the value in cells i through K-1 are moved to cells $i+1$ to K and the new value is placed in cell i. (The entry in cell K is dropped.) The vector of predecessor nodes in cells i through K-1 are also moved to cells $i+1$ to K. In step 4, the current node and cell position combination is marked as having been visited. In step 5, if all K of the cells in the distance vector are marked as having been visited, the algorithm

has found the K-shortest paths to that node. If that node is D, the search ends. In step 6, the node and cell combination with the smallest tentative distance becomes the current node.

An example of a K-shortest path solution procedure is shown in Figure 3.1.
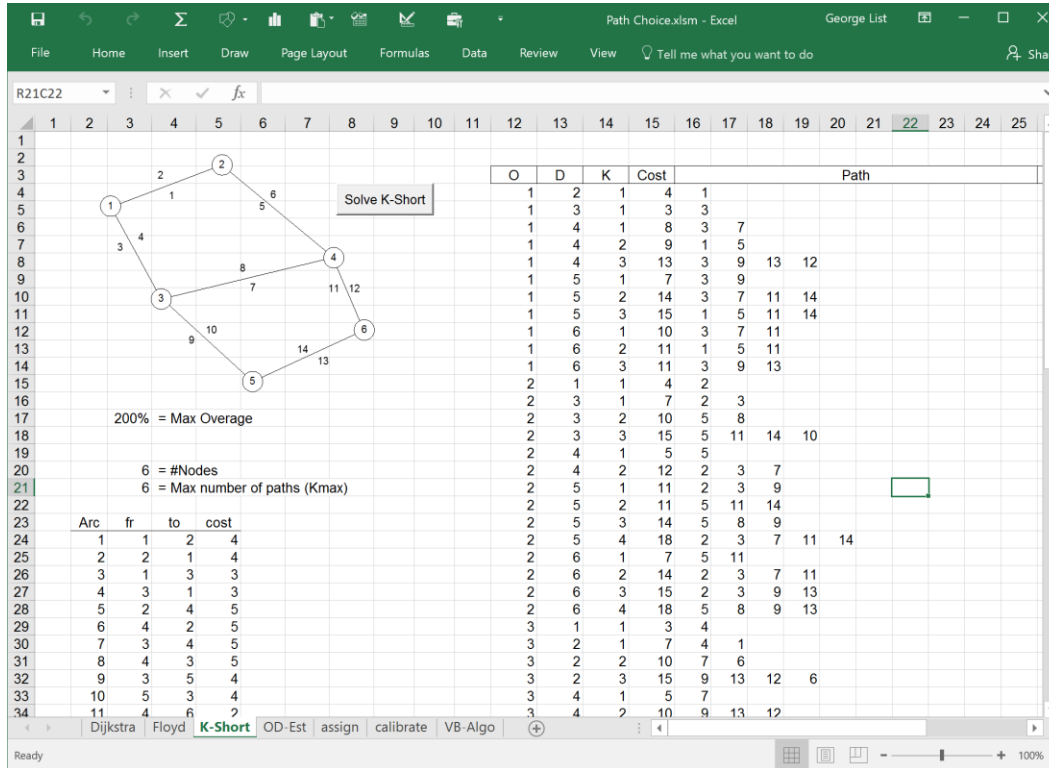
Figure 3.1: K-shortest paths workspace

The network being considered is shown at top left. It has 6 nodes and seven links (14 arcs). The node and arc numbers are indicated. The maximum overage allowed above the minimum cost path is 200%. The maximum number of paths allowed is 6.

The table to the right of the network shows the paths shown for each OD pair. In the case of OD pair (1,4), three paths are found. The first is arc sequence 3, 7 with a cost of 8. The next is 1,5 with a cost of 9. The third is 3, 9, 13, 12 with a cost of 13. Other paths are obviously possible, but they involve greater costs.

The second step is to develop the relationship between departure time and probability of arrival during the AW for all the K-shortest paths identified. To help the reader understand what this means, a few diagrams are helpful. Figure 3.2 shows a hypothetical solution to the first step where the K-shortest paths are found working backward from the median point in the AW. Four paths are found. Of course, there could be a different number of paths.
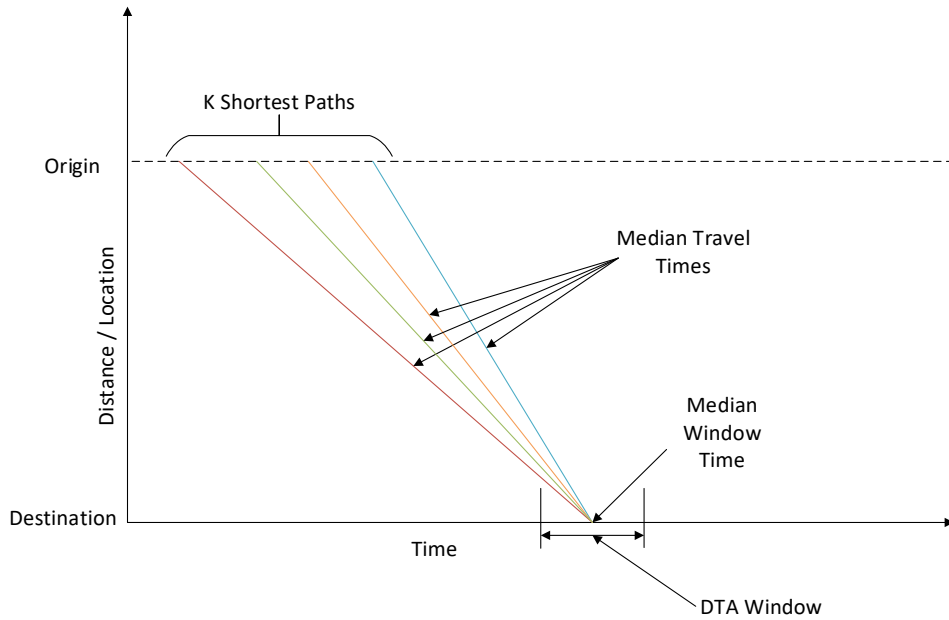
Figure 3.2: K-shortest paths working backward from the median of the OTW

Once the K-shortest paths have been identified, the next step is to evaluate the potential on-time performance of each of the paths. This is illustrated in Figure 3.3.
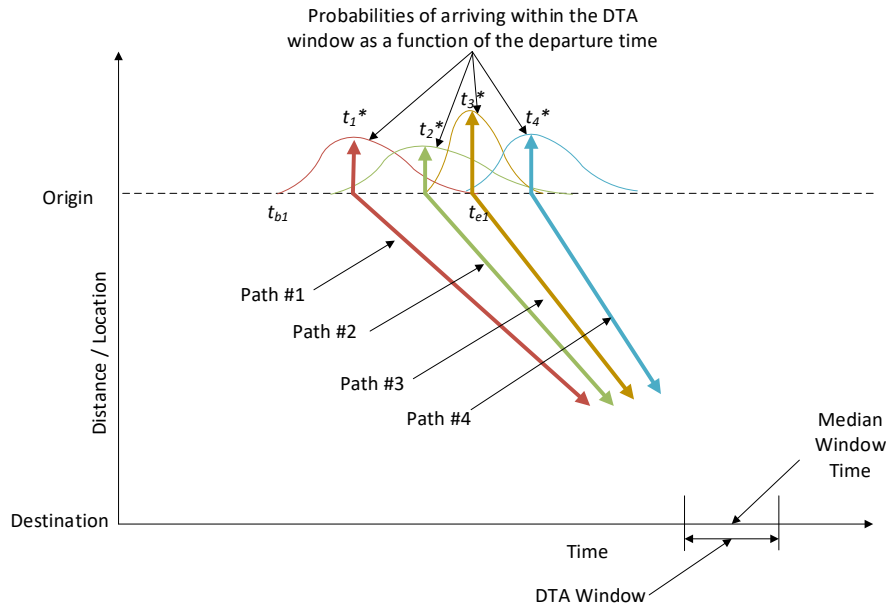
Figure 3.3: Probabilities of arriving during the OTW for paths and departure times

For path #1, for example, there is an earliest possible time, t_{b1}, for which a departure from the origin will result in an arrival during the AW. Of course, from a theoretical perspective, assuming the travel time distribution goes to infinity (it does not have a maximum time), the earliest possible departure time is $t = 0$, but this is an impractical thought. The 99th percentile of the travel time distribution has been used throughout to establish t_{bk}. That is:

$$t_{bk} = t_{DTA}^b - t_k^{99} \tag{3.1}$$

The value of t_{DTA}^b is the beginning of the AW and t_k^{99} is the 99th percentile travel time for path k.

In a similar sense, there is a latest possible departure time, t_{ek}, for which an arrival during the AW is possible. Assuming the travel time distribution has a strict lower bound (minimum travel time), then:

$$t_{ek} = t_{DTA}^e - t_k^0 \tag{3.2}$$

The value of t_{DTA}^e is the end of the AW and t_k^0 is the 0th percentile travel time for path k (i.e., the minimum). If the lower tail of the travel time distribution is unbound (e.g., a normal distribution is employed), then t_k^1, the 1st percentile travel time, can be used.

Figure 3.4 helps to illustrate the ideas. For paths #1 and #4 it shows the range of possible departure times.
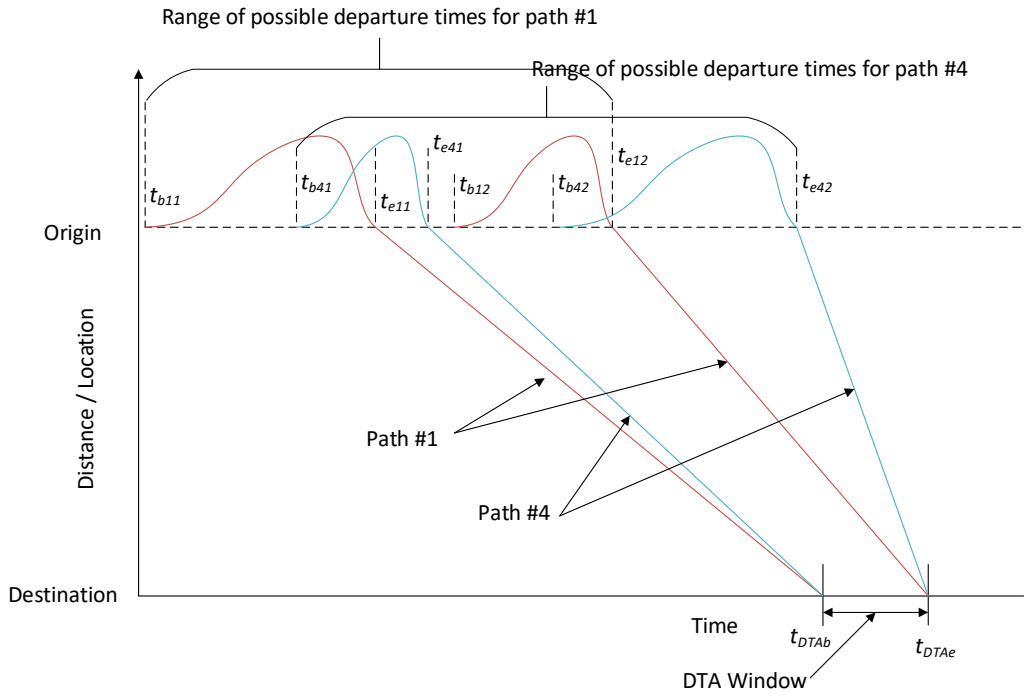
Figure 3.4: Determining the range of departure times to consider

For path #1, the latest possible departure time is t_{e12}. For that time, the truck arrives just at the end of the AW, t_{DTAe}, and it travels at the 0th percentile travel time t_1^0. As depicted by the travel time distribution which extends to the left from t_{e12}, it is possible to leave earlier and arrive at t_{DTAe}, up

until t_{b12}, which corresponds to t_1^{99}. Similarly, the very latest time at which the truck can leave and arrive at the *beginning* of the AW, t_{DTAb}, is given by time t_{e11}. It is possible to leave earlier and still arrive at t_{DTAb}, up until t_{b11}, which corresponds to t_1^{99}. (An astute reader will recognize that the values of t_1^0 and t_1^{99} might be different for the arrivals at t_{DTAe} and t_{DTAb} since the network conditions might change over time.) This analysis leads to the conclusion that the departure times whose performance needs to be assessed start at t_{b11} and end at t_{e12}.

For path #4, the same thoughts pertain. The latest possible departure time is t_{e42}. The truck arrives just at the end of the AW, t_{DTAe}, and travels at the 0th percentile travel time t_4^0. The earliest departure time for which the arrival time is t_{DTAe} is t_{b42}, which involves travel time t_4^{99}. For t_{DTAb}, the latest possible departure time is t_{e41}. And the earliest departure time is t_{b41}, which involves travel time t_4^{99}. The departure times whose performance needs to be assessed start at t_{b41} and end at t_{e42}.

Monte Carlo sampling can be used to determine the probability of arriving on-time within the AW for all the departure times in the range t_{bk1} to t_{ek2} for all paths. These were the results portrayed at the top of Figure 3.3 for all four of the paths that were assumed to have been identified.

The plots look like PDFs, but they are not. Instead they show the trend in the total probability of arriving during the AW as a function of the departure time employed. To illustrate interpretation of the results, path 3 has the highest probability of arriving within the AW. The time at which that occurs is t_3^*. The range of possible departure times is shown in burnt orange. It starts at t_{b3} and ends at t_{e3}. These two times are not labeled per se, but by using path 1 as a reference, they can be identified. Path 4 has a lower maximum probability of arriving within the AW, but it involves latter departure times than those associated with path 3. So, there is an implicit tradeoff here between the likelihood of arriving during the AW and the point in time when the departure time must occur. Path 2 has the lowest maximum probability of arriving during the AW and it has the widest spread in possible departure times. (Intuitively, it does not appear to be a very good choice.) Path 1 has a maximum probability of arrival during the AW that is slightly lower than that for path 4 and it involves a window of departure times that is much earlier. Intuitively, it appears that path 1 is dominated by path 4.

The implications of these differences in on-time probabilities and travel times is illustrated by Figure 3.5 which compares a hypothetical path "*a*" with path "*b*". Path "*a*" has a higher probability of arriving during the AW but a larger travel time. This might be a "surface arterial" that has a more consistent but longer travel time. Path "*b*", on the other hand, has a lower probability of arriving during the AW but a shorter travel time. This might be a freeway path. The path which is "better" depends on the perspective of the decision maker in terms of the tradeoff between the importance of shorter travel times and higher reliability.
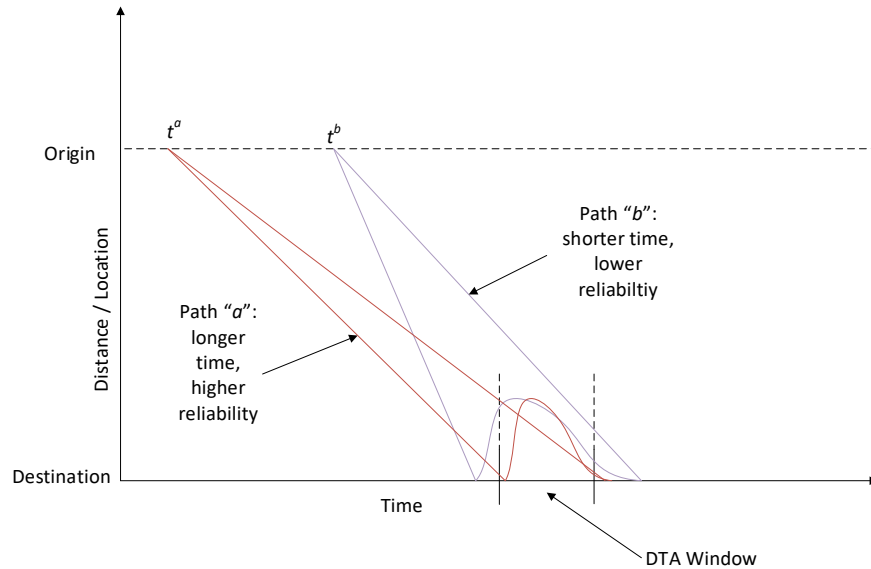
Figure 3.5: A comparison of two paths in terms of travel time and on-time probability

The third step is to select the path and departure time that provides the best combination of travel time and reliability based on the risk and travel time preferences of the decision maker. Since the K-shortest paths are likely to have different travel times and probabilities of arriving during the AW, a tradeoff is likely to exist. As depicted in Figure 3.6, some paths will have shorter travel times but lower on-time probabilities and vice versa. Some of the paths will have non-dominated combinations of these two metrics. Others will be dominated.
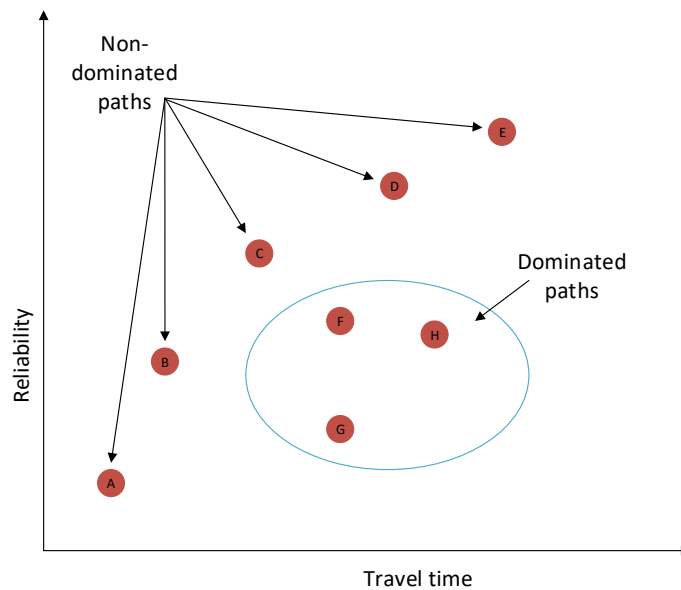
Figure 3.6: Tradeoffs in performance among the K-shortest paths

The non-dominated paths (A, B, C, D, and E) have the best performance combinations. They provide non-dominated combinations of reliability and travel time. For example, path E is the best for reliability, but it also has the longest travel time. If a shorter travel time is desired, path D can

be selected, but the level of performance for reliability must be reduced. The same is true in comparing C with D, B with C, and B with A.

Other paths can exist, such as F, G, and H, but they do not perform as well as the non-dominated paths. For any one of these paths, there is a non-dominated path that does better for one or both objectives. For example, in the case of path F, path C does better. Path C has both a higher reliability and a lower travel time, so it dominates these sub-optimal options.

Selecting the best path among the non-dominated options is a matter of objective importance or weight. If reliability is most important, then path E is best. If travel time is also important, then perhaps path D is better. As the importance of reliability diminishes, the choice switches from E to D, C, B, and then A.

3.3   ILLUSTRATIVE APPLICATION

An example of applying this technique is helpful. The trip being considered is from Carmel Mountain Road in Torrey Pines, CA where I-5 and I-805 split to Civic Center Drive in National City, CA. This is the same trip as the one considered before. The setting is shown in Figure 3.7.
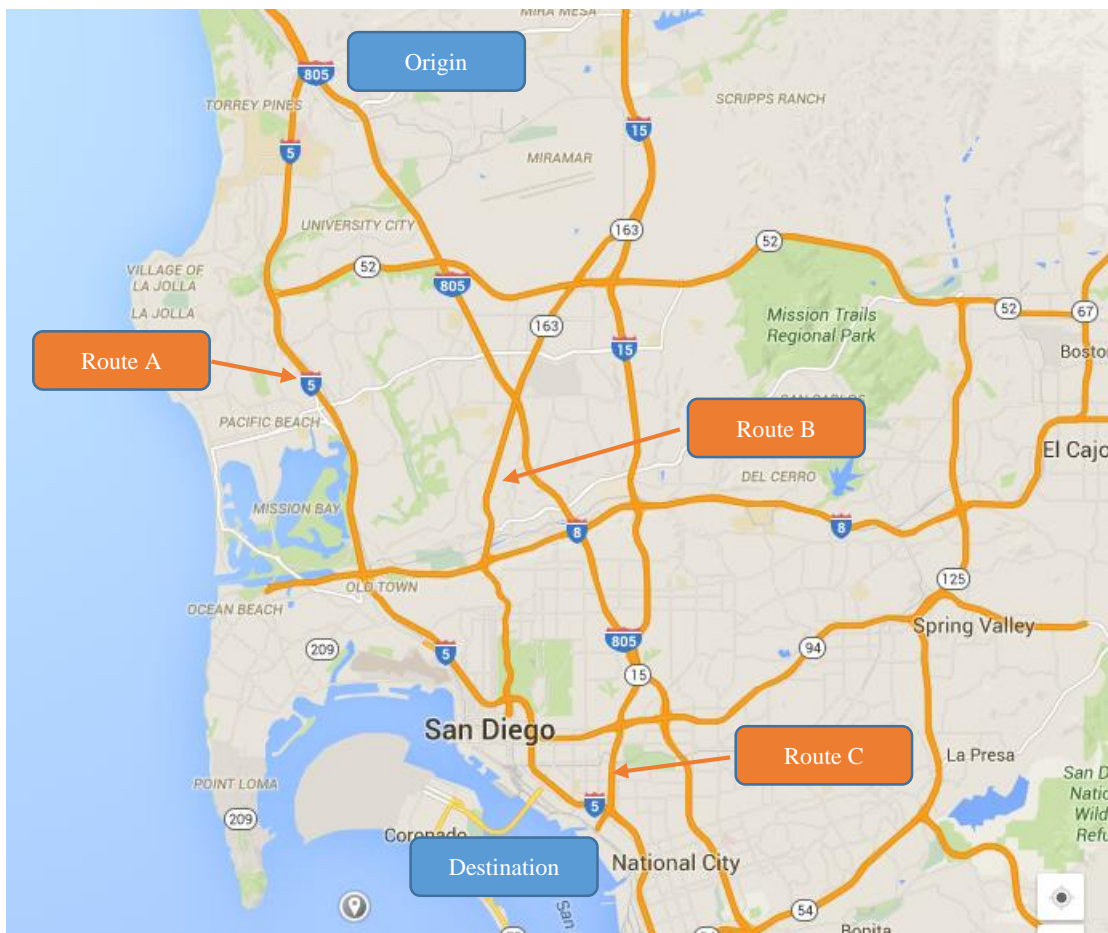
Figure 3.7: Map of three paths from Torrey Pines to National City

64

In step #1, a K-shortest paths assessment needs to be conducted for the origin-destination (OD) pair of interest and for the AWs being considered. Here, this is being done for a trip from Torrey Pines to National City. Two times of day are considered: midday and the PM peak. Since complete data are not available for the San Diego network by time of day and operating condition, an assumption is that the three routes highlighted in Figure 3.7 are the three shortest paths. There is no loss in presentation content to make this assumption.   Path A is via I-5; Path B is via I-805, CA-163, and I-5; and Path C is via I-805, CA-15, and I-5.

Step #2 is focused on understanding the tradeoffs among these paths in terms of their travel times and travel time reliabilities. The objective is to develop characterizations like the ones shown in Figure 3.4 through Figure 3.6 so that the most suitable path can be selected.

Figure 3.8 through Figure 3.10 begin to help us with this task. They show the daily trends in travel times for the three paths. The peak condition occurs in the afternoon. It is also clear that all three paths experience "abnormal" conditions. The travel times that are very different from the typical values, like the high travel times of about 27 minutes at 9:00 am on I-5

In general, it is shown that CA-163 route has difficulty providing consistent travel times during the middle of the day. It is not likely that this path will be selected as being the best.
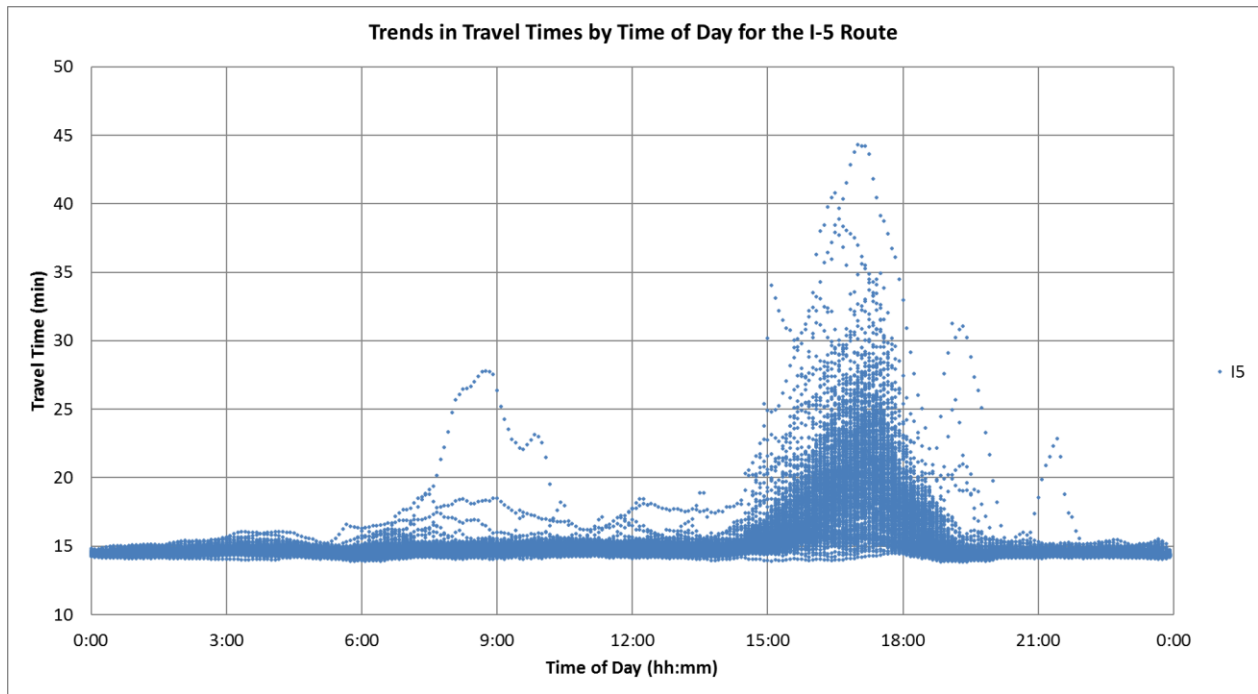
Figure 3.8: Travel times by time of day for the I-5 route
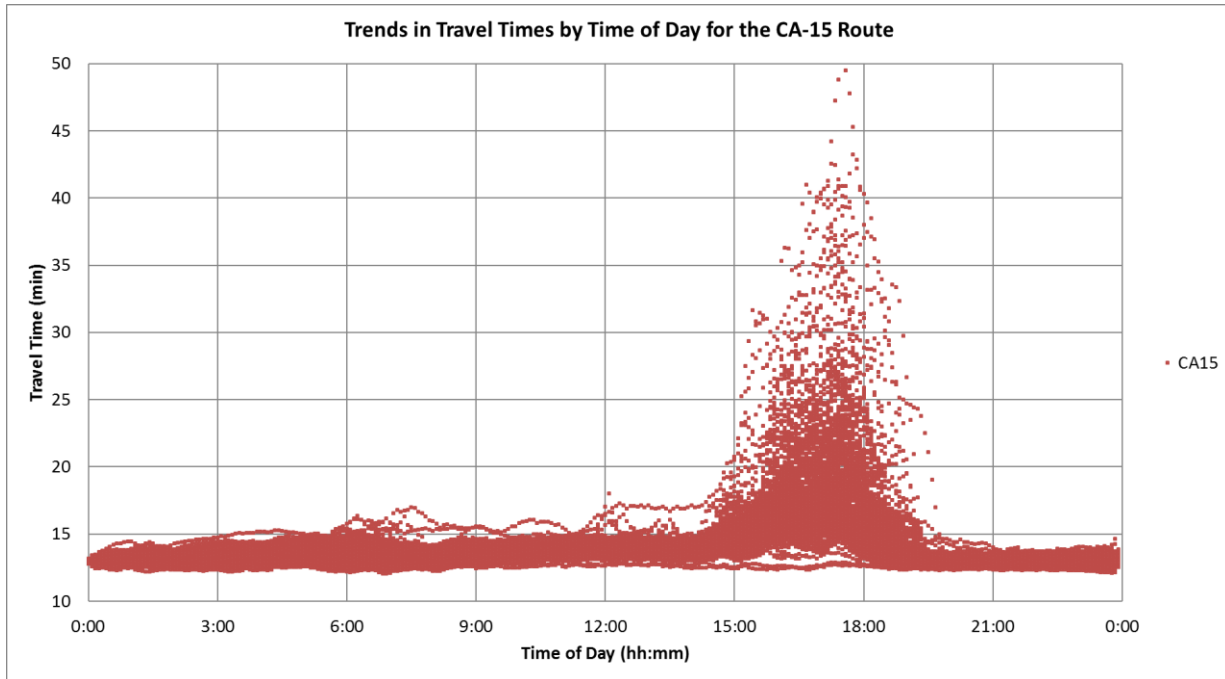
Figure 3.9: Travel times by time of day for the CA-15 route
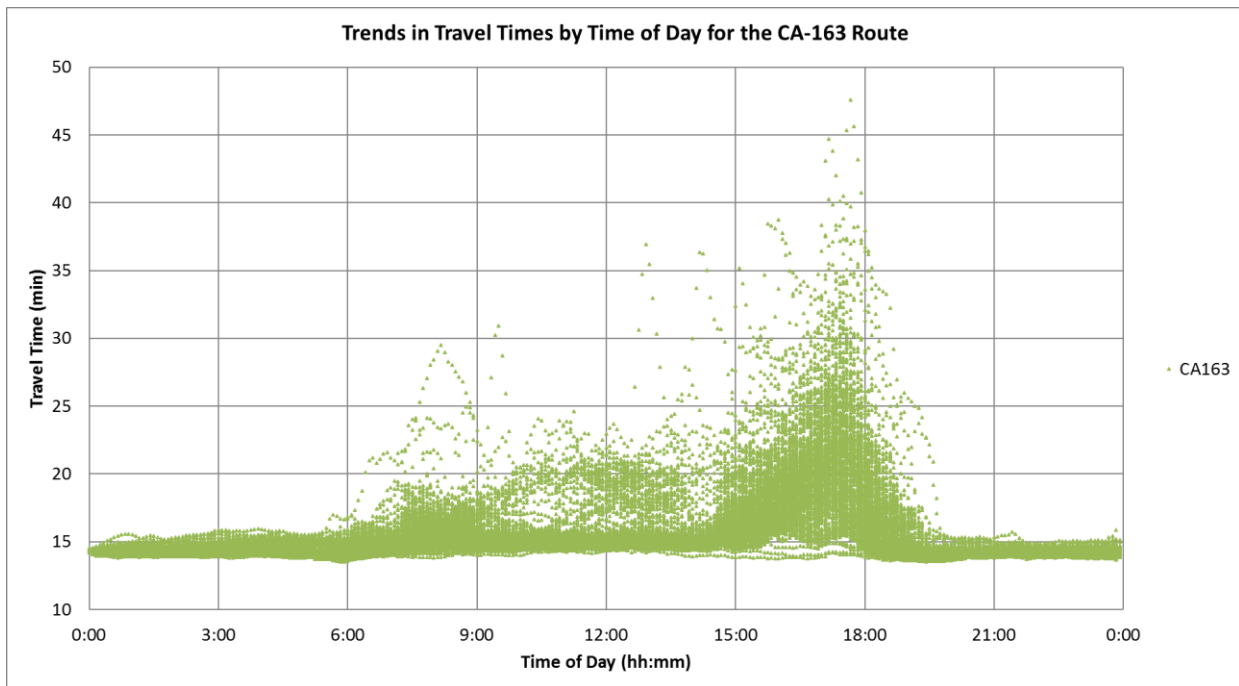
Figure 3.10: Travel times by time of day for the CA-163 route

Another observation is that the CA-163 route seems to have three operating conditions: 1) off peak, 2) AM peak and 3) PM peak. The AM peak appears to be a "moderate" flow condition.

It is possible to get a little more insight by plotting the travel times against a measure of system load, such as the flow rate, as shown in Figures 3.11 through 3.13. This is not something a carrier or customer would do, but an operating agency can and should do it. Moreover, since this research project is examining the issue of freight reliability from an agency perspective it seems useful to present these thoughts. Creating these plots provides a sense of how the travel times increase with traffic. The plots also provide a sense of how the travel time consistency varies.

Of course, these routes have multiple segments, so there isn't a single flow rate that pertains to the route. The traffic varies from one section to another. The option employed here is to use VMT/Hour. (VMT/Hour/Mile can also be used to normalize this measure of system load among routes.) The VMT/Hour captures the total vehicle miles of travel that occurred on all the route segments. The VMT/Hour was associated with every 5-minute data point.

Figure 3.11 through Figure 3.13 show the resulting plots. In every plot, there are 72,000 data points; one for each 5-minute time period during the 250 workdays in the year. The aberrations in travel times due to incidents and other abnormal events are easy to spot. They look like wispy traces that have travel times well above the denser parts of the scatter plot. These abnormal travel times are not the focus of this discussion, but they are useful to note.

For the normal observations, Figure 3.11 shows that the I-5 route has very consistent travel times at low system loads, up to about 60,000 VMT/hour. Above that, the travel times vary more. At about 130,000 VMT/Hour, the variation is greatest and it diminishes up to the highest load of about 145,00 VMT/Hour.
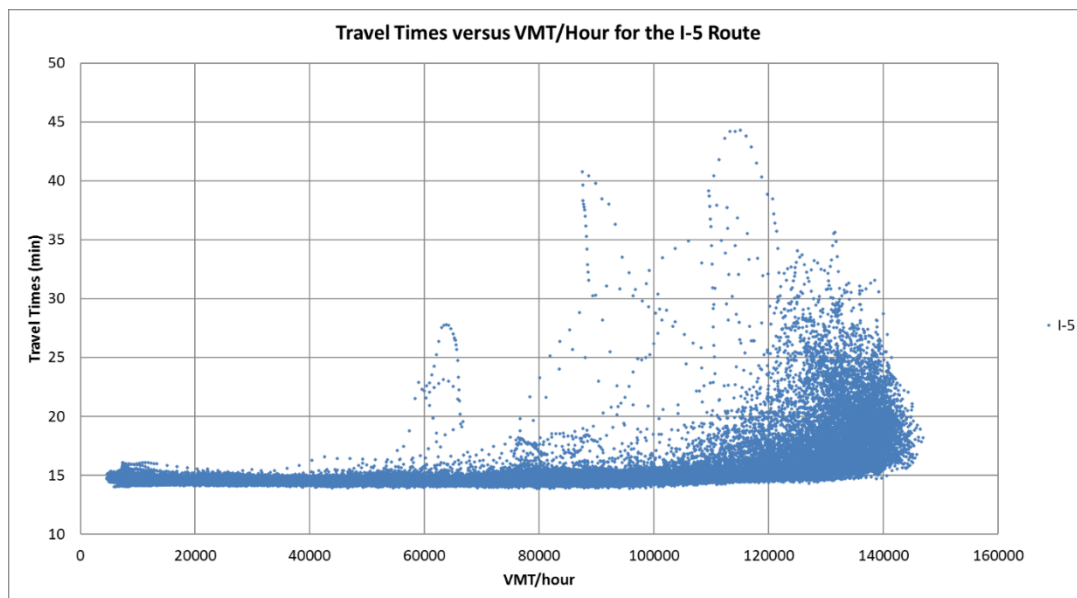
Figure 3.11: Travel times versus system load (VMT/Hour) for the I-5 route

Figure 3.12 shows the data points for the CA-15 route. The same general pattern exists as seen for the I-5 route; and yet it is different.
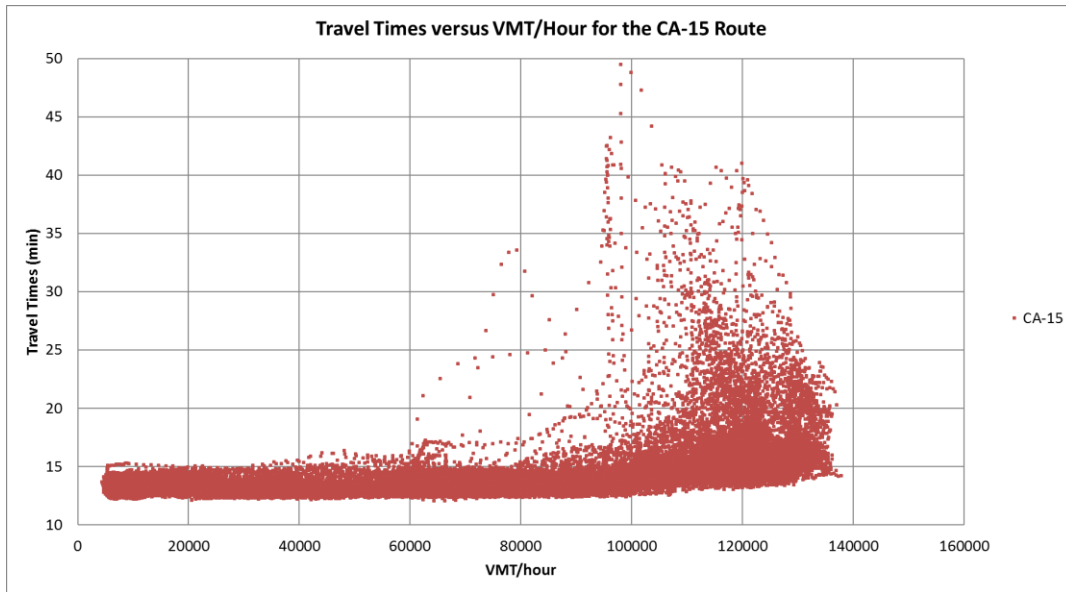
Figure 3.12: Travel times versus system load (VMT/Hour) for the CA-15 route

The rise in travel times is more abrupt, as can be seen around 120,000 VMT/Hour. The variation again diminishes as the maximum VMT/Hour is approached.

The trends for the CA-163 are shown in Figure 3.13. Again, they are similar; but in this case the maximum VMT/Hour is about 115,000, considerably less than for the other two routes. This means there isn't as much traffic on this route as there is on the other two. Also, there is an abrupt increase in the travel time variation at about 75,000 VMT/Hour. This tells us the range of volumes associated with the midday travel times seen in Figure 3.10.
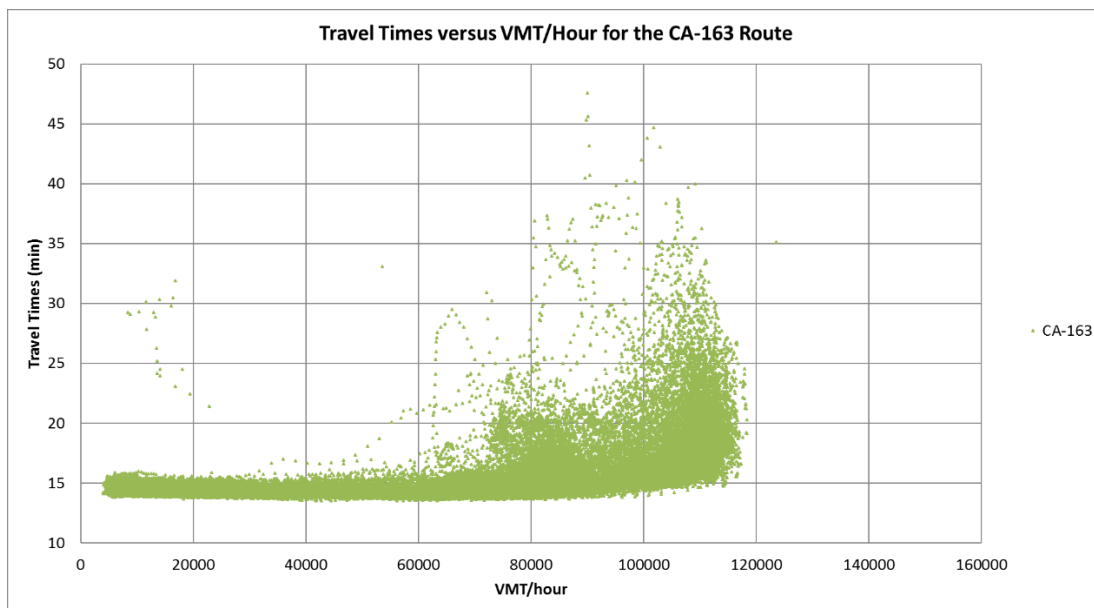
Figure 3.13: Travel times versus system load (VMT/Hour) for the CA-163 route

More specific to the two periods of interest, namely the midday and the PM Peak, it is important to analyze the 5-minute observations that lie within these windows. Without loss of generality, it can be assumed that the midday period is 10:00 am to 3:00 pm and the PM peak is 4:00 pm to 6:00 pm. Also, since the greatest interest is in reliability assessment when the operating conditions are "normal", the "abnormal" observations that involve weather, incidents, etc. can be excluded. If this is done, the focus is on the on-time performance during normal conditions.

Figure 3.14 shows the CDFs of the average travel time for the midday and PM peak conditions for all three routes. During the midday, the CA-15 route has a distribution of travel times that is always less than the other two routes. This does not guarantee that this route will always be the fastest, but it does mean that if this route is always chosen, the resulting *distribution* of travel times experienced will always be better than the distribution of travel times that would have been experienced for the other two. The next best route is via I-5 and the poorest is via CA-163. In the PM peak, the story is very similar. The CA-15 path is again statistically the best, having a distribution of travel times that is less than the other two routes. But the I-5 and CA-163 routes are nearly identical in their travel time CDFs. And the CA-163 route has lower travel times at the higher percentiles, although the differences are only a couple of minutes.
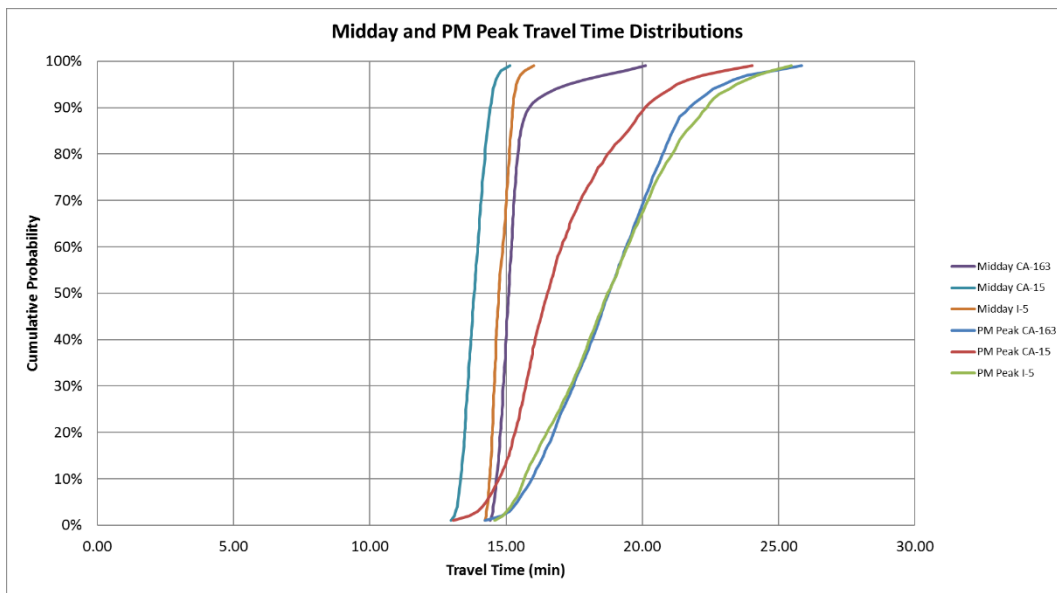
Figure 3.14: Average travel time distributions for the midday and PM peak conditions

Applying the first methodology described in Section 2.0 for estimating the truck travel time distributions results in the CDFs shown in Figure 3.15. The trends in these CDFs are like the average travel time distributions with the very minor exception that in the PM Peak, the relationship between the CDFs for the CA-163 and I-5 route are slightly different. The better performance of the CA-163 for the higher percentiles is no longer present.
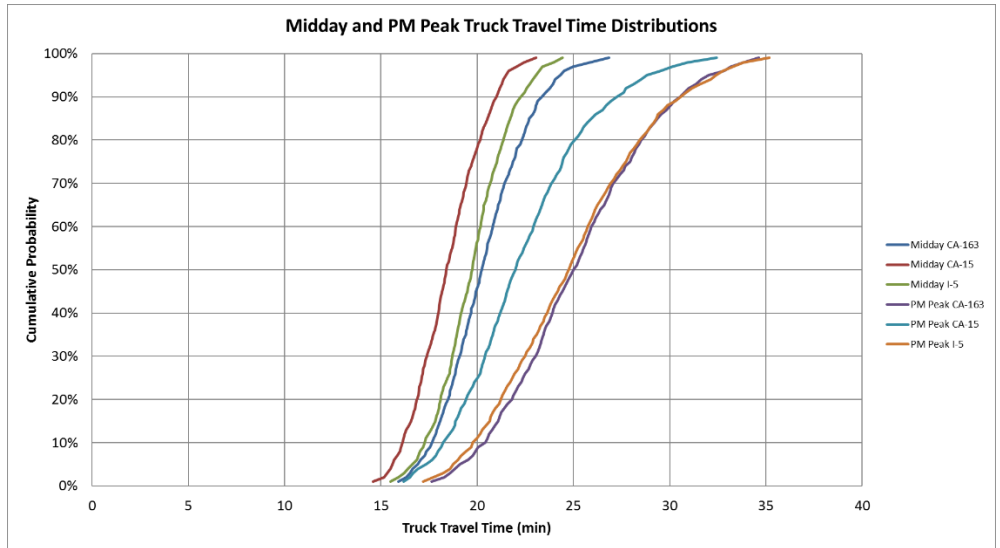
Figure 3.15: Truck travel time distributions for the midday and PM peak conditions

Based on these distributions, the three routes can be assessed in terms of their on-time performance, consistent with the ideas presented in Figure 3.3 through Figure 3.5. Figure 3.16 presents the probabilities of arriving during the OTW for all three routes during the midday and PM peak time periods. The assumed OTW is 10 minutes in duration. Admittedly, this is a small window, but the travel time is relatively short.
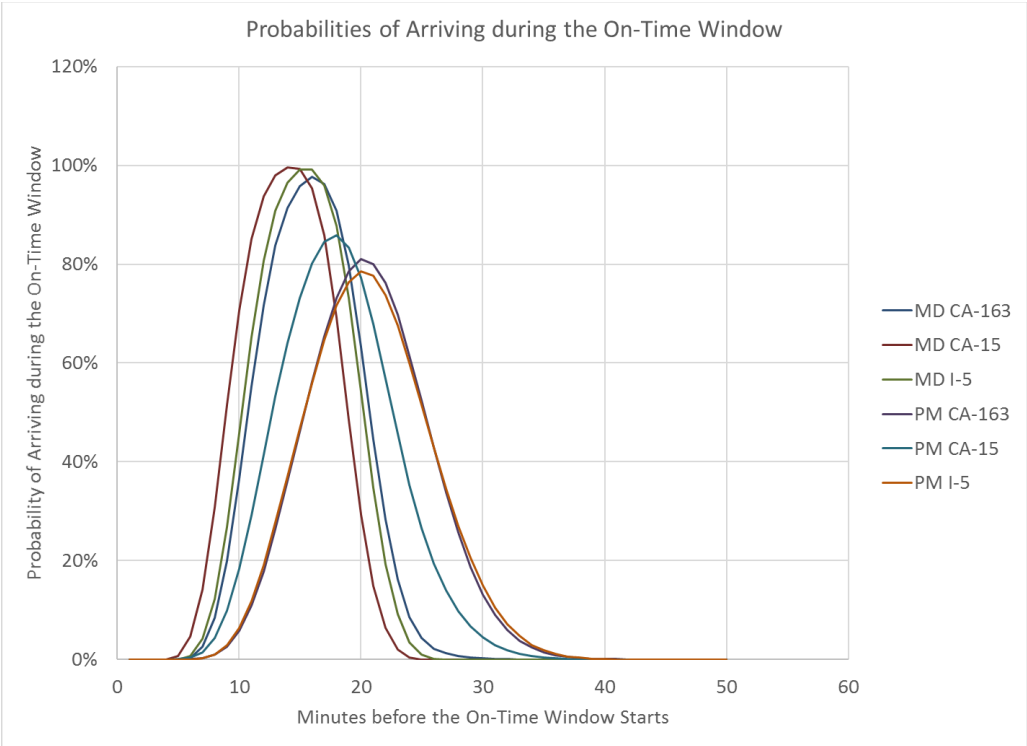
Figure 3.16: OTA probabilities for the midday and PM peak conditions

70

In the midday, the probability of an OTA peaks at 100% for the CA-15 route if the departure occurs 14 minutes before the AW starts. For the I-5 route, the same is true, but the departure time needs to be 16 minutes before the AW starts. The CA-163 route reaches its maximum 16 minutes before the OTW starts, but the probability is slightly lower. In the PM peak, none of the routes can achieve a 100% probability of arriving during the AW. The CA-15 route does the best reaching an 86% probability for a departure time 18 minutes before the start of the AW. The I-5 and CA-163 routes achieve their maximum probabilities for a departure time 20 minutes before the start of the AW and the I-5 route has a slightly higher probability of 81% versus 78%.

These tradeoffs among the performance of the three routes is portrayed in Figure 3.17. Consistent with what was said above, in the midday condition, the CA-15 route dominates the other two. Its travel time is lower and its probability of being on-time is as high or higher than for either of the two other routes. The same is true in the PM Peak; and the maximum on-time probability is significantly higher than it is for the other two routes.
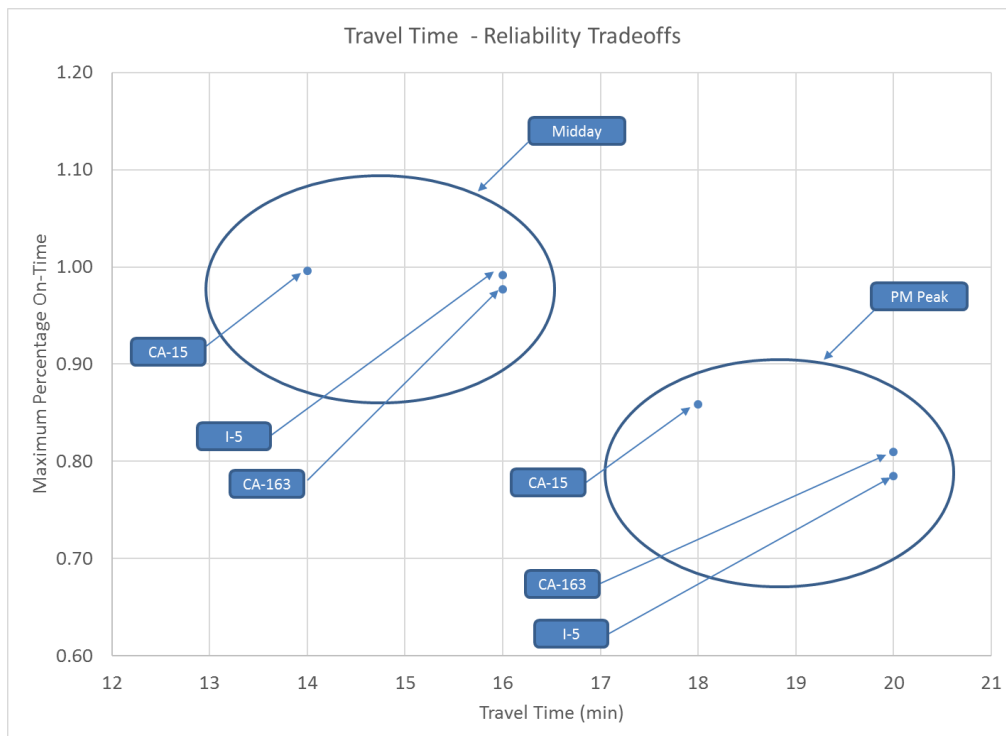
Figure 3.17: Tradeoffs between travel time and on-time performance

The third step involves selecting the path and departure time that provides the best combination of travel time and reliability based on the risk and travel time preferences of the decision maker. In this case, that choice seems relatively clear. Based on Figure 3.17, the CA-15 route is the best in both the midday and the PM Peak conditions.

3.4   SUMMARY

This section has focused on the topics of path and departure time selection. A K-shortest path methodology has been presented that identifies a non-dominated set of path options from which

the most desirable one can be selected. The methodology assumes the vehicles do not arrive early and wait for the AW to begin. (That tactical option is considered in the next section.) An example is presented focused on an assessment of three routes in the San Diego area from Torrey Pines to National City.

4.0   VEHICLE ROUTING AND SCHEDULING

4.1   INTRODUCTION

Vehicle routing and scheduling lies at the heart of operating a freight transportation system. Nominally, the objective of the vehicle routing problem (VRP) is to find an assignment of loads to vehicles and routings for the vehicles that optimizes all the performance metrics. In some instances, the loads are full truckloads, in which case the vehicles are assigned to carry loads from one point to another in sequence. Alternately, trucks may pick-up loads or deliver loads. In yet a third option, trucks may both pick-up and drop off loads, as with local couriers. The objectives are often to 1) minimize total cost, 2) maximize on-time deliveries, 3) minimize the fleet size, and 4) maximize vehicle utilization. Other objectives include 5) maximizing on-time performance and 6) maximizing the lowest on-time performance among all the vehicles employed.
.

4.2   RELEVANT LITERATURE

The body of literature on vehicle routing and scheduling is vast. When Bodin *et al.* (1981) did their review more than two decades ago, over 500 papers were identified. This review, which is broader in scope and almost 25 years later, will not be able to review all this literature in detail.

The first treatment of the topic appears to be Dantzig and Ramser (1959). They presented a formulation of the truck dispatching problem that assigns loads to multiple trucks based on truck capacity. The motivation was refinery trucks delivering gasoline to filling stations. No direct treatment is given to the distances traveled by the trucks. The loads are sorted into a specific order and then assigned to trucks sequentially given the truck capacities. This problem is a generalization of the single Traveling Salesman Problem (TSP), again studied in the initial stages by Dantzig.

A subsequent paper by Clarke and Wright (1964), which describes their savings heuristic, explicitly considered the distances traveled. They state the problem in a somewhat informal manner (by today's standards) by indicating that tours are to be established for *K* trucks such that 1) all loads are carried, 2) the total distance traveled by the trucks is minimized and 3) the capacities of the trucks are not exceeded.

Often, the goal of VRP is to design a set of minimum-cost routes for a vehicle fleet that serves a set of customers at various locations with each route beginning and ending at a fleet depot. Intuitively, the cost attributed to a set of routes includes total travel time or distance across all individual routes in addition to the total number of vehicles required to cover the routing, resulting from vehicle wear and employee payment. As there exist a combinatorial number of routes that can be built, the general VRP falls in the NP-Hard class of problems; therefore, it is unlikely that there exists a polynomial-time algorithm to solve VRP to optimality. Such exact algorithms use problem formulations often from Integer programming, including branch and bound, set partitioning, column generation, and network flow solution methods or formulations from dynamic programming with effective state-space relaxation (see survey paper of Laporte (1992)). These formulations suffer from intractability in practice, thus heuristics have been developed to allow for

often suboptimal although adequate solutions in a fraction of the time required by exact methods. Many popular heuristics for solving instances of VRP rely on a 2-phase route construction and route improvement algorithm. Route construction methods include the Clarke-Wright Savings algorithm described in Clarke and Wright (1964) and the Sweep algorithm in Wren and Holliday (1972). The Clarke-Wright algorithm applied to Euclidean depot-customer locations begins with a number of routes equal to the number of customers and creates routes until no more savings can be achieved by merging routes, respecting direct (i, j) arcs for greedy optimal step saving s_{ij}. As a result, the number of vehicles in the fleet is fixed upon termination of the algorithm. Further reductions in total route cost can often be achieved by limiting the domain of eligible arcs using an upper bound on arc costs as was done in Caccetta, Alameen, and Abdul-Niby (2013). The Sweep algorithm on the other hand uses a fixed number of vehicles when building routes. The solutions to these algorithms are then improved upon using such techniques as k-opt switches of customers, route merging and splitting, and subroute deletion and insertion.

Over time the VRP has been reformulated to include more realistic constraints, such as capacity constraints, time window constraints at customers, nonhomogeneous travel times, stochastic demands and travel times, and pickup and delivery precedences. For instance, in the vehicle routing with hard time windows, each customer has the added restriction that service must occur within a fixed time window specified by early and late times (ℓ_i and l_i respectively for a given customer i). A feasible routing is required to meet all time constraints. On the other hand, the vehicle routing problem with soft time windows allows for the violation of time windows with a resulting penalty cost often proportional to the size of the violation.

One of the more realistic elements to add to VRP problems is the use of stochastic travel times instead of the classical deterministic travel times. In these Stochastic Vehicle Routing Problems (SVRPs), the expected travel times and service times are used as surrogates for the more realistic stochastic variability arguing that a "expected value" problem is being solved. This is expedient, but the resulting solutions provide little or no insight about system performance when the vector of values is significantly different from these mean values. As intelligent transport systems have developed, the integration of real-time information into traffic monitoring systems has allowed for more accurate travel time to be relayed dynamically to interested stakeholders. For example, Ando and Taniguchi (2006) detailed the means of relaying traffic information using vehicle probe GPS devices and Japan's Vehicle Information Communication Systems. Freight transport businesses should then account for the most recent traffic and service information when planning distribution routes of their fleets. However, these travel times are constantly changing and thus prior routing cannot be applied as effectively without compensating for stochastic times. Common distributions used in travel time estimations are the lognormal and gamma distributions (See Tas *et al.* (2013, 2014, 2014)). Notably, Section 错误!未找到引用源。 describes a procedure that generates route-level travel time distributions by sampling distributions of congested and uncongested segment-level travel times. The distributions are lognormal and the likelihood of sampling from the congested distribution depends upon a random number draw.

While VRPs with soft time windows do have specified time windows at customers, this study is interested in investigating the effect of self-imposed time windows at customers. This represents

a determination of delivery windows by supplier rather than the customer, a property that could likely be present when the supply chain is managed by a single operator.

The literature on VRP can be broken down into two subgroups. The first treats the travel times between locations and the servicing times at each location as being fixed or nearly so. This representation of the problem is the classical one and has been studied for a long time. Mathematical programming techniques, like combinatorial optimization, are used to find problem solutions. So are heuristics.

The second subgroup assumes the travel times and servicing times are stochastic. Techniques like stochastic optimization and simulation in combination with optimization (search routines) are used to find solutions. This work is more recent, spawned by the advent of computers that can simulate the movement of large fleets of trucks in reasonable time.

The problem can be stated as follows. Assume there are N loads to be carried to destinations and assume that each has a size given by dem_i. Also, let K vehicles be available and assume each one has a capacity of cap_k. Choosing to use vehicle k is reflected by $z_{0k} = 1$ and the assigning load i to vehicle k is designated by z_{ik}. The sequence for visits to the destinations is captured by x_{ijk} which indicates that load i is to be delivered before load j by vehicle k. If the distance between i and j for vehicle k is given by c_{ijk}, then the problem is:

Minimize:

$$\sum_k \sum_i \sum_j c_{ijk} x_{ijk} \tag{4.1}$$

Subject to:

$$\sum_i dem_i * z_{ik} \leq cap_k * z_{0k} \quad \forall k \tag{4.2}$$

$$\sum_k z_{0k} \leq K \tag{4.3}$$

$$\sum_k z_{ik} = 1 \quad \forall i \tag{4.4}$$

$$\sum_i x_{ijk} = z_{jk} \quad \forall j,k \tag{4.5}$$

$$\sum_j x_{ijk} = 1 \quad \forall i,k \tag{4.6}$$

$$\sum_{i \in S} \sum_{j \in S} x_{ijk} \leq |S| - 1 \quad \forall k \quad where \quad S \subseteq N(z_{ik}) \text{ is the set of all deliveries made by } k \tag{4.7}$$

The objective function specifies that the total vehicle miles should be minimized. Equation

$$\sum_i dem_i * z_{ik} \leq cap_k * z_{0k} \quad \forall k \tag{4.2}$$

ensures that the capacity of each truck used is not exceeded. Equation $\sum_k z_{0k} \leq K$

(4.3) ensures the selected fleet

size is not greater than the fleet available. Equation $\sum_k z_{ik} = 1 \quad \forall \, i$

(4.4) ensures that the loads get assigned, Equations

$$\sum_i x_{ijk} = z_{jk} \quad \forall \, j,k \tag{4.5}$$

and $\sum_j x_{ijk} = 1 \quad \forall \, i,k$ (4.6)

establish the load assignment sequences and equation

$$\sum_{i \in S}\sum_{j \in S} x_{ijk} \leq |S| - 1 \quad \forall \, k \quad where \quad S \subseteq N(z_{ik}) \text{ is the set of all deliveries made by } k \tag{4.7}$$

ensures that the number of arcs traversed by each truck is less than or equal to the number of deliveries made.

The next major focus for VRP was on the dispatching of special purpose vehicles to accommodate the needs of elderly and handicapped individuals. "Dial a Ride" is how it was described. The problem is as follows. A set of requests are made for trips to and from specific locations with specific departure and arrival times, like doctor's appointments and shopping trips. The task is to determine how to assign these trips to the dial-a-ride vehicles, and how many vehicles to use. The solution becomes the pick-up and delivery schedule. Unlike delivering loads, multiple people can be on-board the vehicle at any given point in time.

Bruck (1969) was one of the first to present a formulation. He described a tool called CARS (Computer Aided Routing and Scheduling) that was intended to be a decision support system for solving the dial-a-ride problem. Papers with a similar focus were prepared by Howson (1970), Deleuw, Cather (1971) Arthur D. Little (1971), Roos (1971), Roos and Porter (1971), and Roos and Wilson (1971).

As stated by Cordeau (2006), the problem is as follows. Paraphrased slightly, let n denote the number of users (or requests) to be served. The problem can be defined on a complete directed graph $G = (N,A)$ where $N = P \cup D \cup \{0, 2n + 1\}$, $P = \{1, \ldots, n\}$ and $D = \{n + 1, \ldots, 2n\}$. Subsets P and D contain pick-up and drop-off nodes, respectively, while nodes 0 and $2n + 1$ represent the origin and destination depots. Thus, for each user i there is an origin node i and a destination node $n + i$. Each vehicle $k \in K$ has a capacity Q_k and the total duration of its route cannot exceed T_k. With each node $i \in N$ are associated a load q_i and a non-negative service duration d_i such that $q_0 = q_{2n+1} = 0$, $q_i = -q_n + i$ $(i = 1, \ldots, n)$ and $d_0 = d_{2n} + 1 = 0$. A time window $[e_i, l_i]$ is also associated with node $i \in N$ where e_i and l_i represent the earliest and latest time, respectively, at which service may begin at node i. For each arc $(i, j) \in A$ there is a routing cost c_{ij} and a travel time t_{ij}. Finally, L represents the maximum ride time allowed by policy for a user. For each arc $(i, j) \in A$ and each vehicle $k \in K$, $x_{ij}^k = 1$ if vehicle k travels from node i to node j. For each node $i \in N$ and each vehicle $k \in K$, let B_i^k be the time at which vehicle k begins service at node i, and Q_i^k be the load (number of people) on vehicle k after visiting node i. Finally, for each user i, let L_i^k be the ride time of user i on vehicle k. The formulation is as follows:

Minimize:

76

$$\sum_{k \in K} \sum_{i \in N} \sum_{j \in N} c_{ij}^k x_{ij}^k \tag{4.8}$$

Subject to:

$$\sum_{k \in K} \sum_{j \in N} x_{ij}^k = 1 \qquad \forall\, i \tag{4.9}$$

$$\sum_{j \in N} x_{ij}^k - \sum_{j \in N} x_{n+1,j}^k = 0 \qquad \forall\, i \in i,k \tag{4.10}$$

$$\sum_{k \in K} \sum_{j \in N} x_{0j}^k = 1 \qquad \forall\, k \tag{4.11}$$

$$\sum_{j \in N} x_{ji}^k - \sum_{j \in N} x_{ij}^k = 0 \qquad \forall\, i \in i,k \tag{4.12}$$

$$\sum_{i \in N} x_{i,2n+1}^k = 1 \qquad \forall\, k \tag{4.13}$$

$$B_j^k \geq (B_i^k + d_i + t_{ij}) x_{ij}^k \qquad \forall\, i,j,k \tag{4.14}$$

$$Q_j^k \geq (Q_i^k + q_j) x_{ij}^k \qquad \forall\, i,j,k \tag{4.15}$$

$$L_i^k = B_{n+i}^k - (B_i^k + d_i) \qquad \forall\, i,k \tag{4.16}$$

$$B_{2n+1}^k - B_0^k \leq T_k \qquad \forall\, k \tag{4.17}$$

$$e_i \leq B_i^k \leq l_i \qquad \forall\, i,k \tag{4.18}$$

$$t_{i,n+i} \leq L_i^k \leq L_i \qquad \forall\, i,k \tag{4.19}$$

$$\max\{0,q_i\} \leq Q_i^k \leq \min\{Q_k,Q_k+q_i\} \qquad \forall\, i,k \tag{4.20}$$

$$x_{ij}^k \in \{0,1\} \qquad \forall\, i,j,k \tag{4.21}$$

The objective function (4.8) minimizes the total routing cost. Constraints (4.9) and (4.10) ensure that each request is served exactly once and that the origin and destination nodes are visited by the same vehicle. Constraints (4.11) - (4.13) guarantee that the route of each vehicle *k* starts at the origin depot and ends at the destination depot. Consistency between the time and load variables is ensured by constraints (4.14) and (4.15). Equalities (9) define the ride time of each user which is bounded by constraints (4.19). The latter also act as precedence constraints because the non-negativity of the L_i^k variables ensures that node *i* will be visited before node *n + i* for every user *i*. Finally, the inequality (4.17) bounds the duration of each route while (4.18) and (4.20) impose time windows and capacity constraints, respectively. This formulation is non-linear because of constraints (4.14) and (4.15) but there are ways to convert it to a mixed integer LP. Those techniques are discussed by Cordeau (2006).

The research on VRP saw a diversity of applications while the dial-a-ride problem was being addressed. Nussbaum (1975), Field (1976), and Hinds (1979) described a software package for

routing and scheduling transit buses (The program is called RUCUS for Run Cutting and Scheduling.) Fielding (1977) examined shared-ride taxis (like dial-a-ride). Bodin *et al.* (1978) studied the routing and scheduling of street sweepers. Ghoseiri, Ghannadpour, and Seifi (2010) examined the problem of dispatching railroad locomotives.

VRP has also been applied to other domains. Ronen (2002) described the use of VRP in the context of cargo ships. Zografos and Androutsopoulos (2002), Meng, Lee, and Cheu (2005), and Androutsopoulos and Zografos (2012) examined the domain of hazardous materials transport.

Attention has also been given to finding procedures that can solve very large VRP problems. Agin (1975) described a large number of algorithms. Buxey (1979) explored the possibility of using Monte Carlo simulation to find solutions. Baker and Rushinek (1982) examined large-scale implementation issues.

More recently, AVI (automatic vehicle identification) and AVL (automatic vehicle location) technologies have been used to gain additional insights into travel time reliability. Kwon, Martland, Sussman and Little (1995) examined reliability in the context of samples of railroad freight car movements collected by the Association of American Railroad's Car Cycle Analysis System. Clear differences were found in reliability between general merchandise, unit train, and double-stack container services. Fu and Rilett (2000) explored the estimation of time-dependent, stochastic route travel times by using artificial neural networks. Ichoua, Gendreau, and Potvin (2000) studied dynamic vehicle routing and scheduling options in exploitation of real-time information about vehicle location. Taniguchi, Thompson, Yamada, and van Duin (2001) examined city logistics issues in light of the information provided by global positioning systems. An examination of reliability and the related issues of routing and scheduling for urban pick-ups and deliveries followed (see Taniguchi, Yamada, and Tamagawa, 2001). Kwon *et al.* (1995) used the American Railroad's Car Cycle Analysis to determine reliability of freight car movements. Ichoua, Gendreau, and Potvin (2002) utilized real-time vehicle location data to examine dynamic vehicle routing and scheduling options. In this same period, there was an emphasis on supply chain logistics. Armacost, Barnhart, and Ware (2002) discussed the use of composite variables to describe the logistics of UPS. Kim, Mahmassani, and Jaillet (2004) investigated dynamic truckload routing, scheduling and load acceptance using simulation to evaluate the relative performance of various decision-making policies. In 2004, Armacost, Barnhart, Ware, and Wilson explored the optimization of UPS in regard to aircraft movements and cost efficiency.

Treatment of the problem from a stochastic standpoint starts about 1990. Laporte *et al.* (1992) addressed the problem of finding solutions to the vehicle routing and scheduling problem when stochastic travel times are present. A chance-constrained programming formulation is presented along with two stochastic optimization formulations and a branch-and-cut algorithm for solving all three formulations. The chance constrained formulation performs well as should be expected since it is a variant on the mixed integer LP formulation. Of the two stochastic optimization formulations, the one that more explicitly represents the problem formulation does much better. The authors conclude that such problems can be solved for significant size problems in reasonable time. Powell (1988) describes algorithms that can be used to solve the dynamic (time-based) routing of vehicles in response to known and anticipated, but unknown loads.

Many papers focused on solving stochastic vehicle routing problems followed Laporte *et al.* (1992). There is the notable paper by Bertsimas *et al.* (1995) and the proceedings paper by Taniguchi *et al.* (1999). Campbell (2004) described heuristics for considering a variety of complicating constraints not typically included in the traditional formulations. Yamada, Yoshimura, and Mori (2004) is an interesting paper because it endeavors to use VRP procedures to study and assess road network reliability.

Other papers focused on making decisions about re-routing vehicles in real-time in response to evolving network conditions. Taniguchi, Yamada, and Tamaishi (2001) presented a formulation of the problem. Taniguchi and Nakanishi (2003) gave another. Slater (2002) provided an approach to the problem as does Kim (2003). Dessouky, Ioannou, and Jula (2004) examined strategies that can build partial tours (in time) and then update those tours as more information becomes available. Mitrovic-Minic and Laporte (2004) explored the use of waiting strategies. Hejazi and Haghani (2009) inestigated ways for less-than-truckload services to optimize their services considering evolving conditions in the highway network. Kanturska, Trozzi, and Bell (2013) presented the idea of hyperpaths to help drivers select optimal delivery routes and schedules in response to evolving patterns of network travel times. The hyperpaths are sets of possible paths plus a path selection logic.

The advent of optimization schemes such as genetic algorithms, simulated annealing, and tabu search has motivated explorations of ways to use these techniques to solve VRP problems. The earliest investigation appears to be Garcia and Arunapuram (1993) who explored the use of tabu search. Potvin (2007) provided a survey of evolutionary algorithms that have been applied to VRP. Included in the review are genetic algorithms, evolutionary strategies, and swarm optimization. Weise, Podlich, and Gorldt (2009) completed a similar, newer review.

Subsequent research efforts focused on using a variety of techniques to address specific problems. Cordeau and Laporte (2003) used tabu search to solve a multi-vehicle dial-a-ride problem that is dynamic. Bell (2004) explored the use of game theory to address VRPs where the occurrence of incidents is of concern. Vidal *et al.* (2012) presented a hybrid genetic algorithm for solving multi-depot and periodic VRPs. Lin, Yu, and Chou (2011) employed simulated annealing and Xu, He, and Li (2009) explored a hybrid procedure that integrates genetic algorithms, stochastic simulation, and neural networks.

4.3 HEURISTIC SEARCH

This search procedures uses a simulation-based heuristic based on an initial Clarke-Wright solution followed by merge, insertion, and 2-interchange reduction in order to solve the vehicle routing problem with stochastic travel times and soft time windows. Unique to this method is the fact that time windows are set at runtime based off a lookup table of previously simulated ordered customer pairs, corresponding to supplier-side control of a supply chain. It is able to solve in reasonable time on a relatively weak computer a problem instance with 200 customers for a given parameters in a few minutes, and this resulted in at least 23% cost reductions for 10 or greater customers. This could easily be extended to larger problem instances on larger RAM computers, as storage space is the critical element in the algorithm.

4.3.1 Related Work

Most recently, Wang and Lin (2013) developed a similar algorithm as used in this paper by using Monte-Carlo simulation of the travel and service times in setting a cost for a given route. They used fixed time windows unlike this paper and used only a deletion-and-insertion based improvement scheme from the earliest to the latest route, rather than trying from late to all earlier routes in order until a reduction is found; this method will be discussed in later sections.

Earlier work by Laporte, Loveuax, and Mercure (1992) solved the VRP with stochastic travel times using a chance-constrained model that can relatively easily solve problems with additive probability distributions for travel and service times, in addition to two- and three-index recourse models. Kenyon and Morton (2003) sought to answer small-scale VRP with stochastic travel times using two solution methods: (1) solve an integer programming exactly using branch-and-cut as a deterministic equivalent of a stochastic problem by including probability mass function terms in the model and (2) solve the problem approximately by using simulation to generate lower and upper bounds on candidate solution routes until a solution of a given problem falls within ε of what would be the exact solution in (1). Metaheuristic methods using tabu search and adaptive large neighborhood search have been developed to solve the VRP with stochastic travel times with soft time windows by Tas *et al.* (2013, 2014, 2014) and Li *et al.* (2010). Jabali *et al.* (2013) investigated the VRP with self-imposed time windows and developed a heuristic algorithm that embeds the timing decisions as modified buffer scheduling problems (common to industrial engineering applications) within a tabu search metaheuristic that generates the routes. Lecluyse, Van Woensel, and Peremans (2009) use a 95th percentile approximation for travel times to solve the VRP with time windows embedded within a 2-interchange tabu search. Similar to this paper, they use a lognormal distribution to model travel and service times.

4.3.2 Description of the Formulation

The problem can be formulated as follows. Let N be the number of customers to be served and k be the fleet size. The transportation network is given by the undirected graph $G = (V, A)$. With the depot included as vertex 0, G is then a complete $K_{|V|} = K_{N+1}$ graph. Therefore $V = \{0, 1, \ldots, N\}$ and $A = \{(i, j)\}_{i, j \in N}$. The lognormal parameters of scale squared and location are provided for travel times on all arcs (i, j) among the N customers and the depot node 0 in $(N+1) \times (N+1)$ matrices $\Sigma^2 = (\sigma_{ij}^2)$ and $U = (\mu_{ij})$, respectively. These will be used during simulation to produce traveling times t_{ij} for $i, j \in V$. Additionally the lognormal parameters of scale squared and location are given for service times for each of the N customers in 2 vectors $\vec{\sigma}_s^2$ and $\vec{\mu}_s$, respectively. These will be used during simulation to produce service times $\int_i$ for $i \in V - \{0\}$. Each customer $i \in V - \{0\}$ will be assigned a soft time window (e_i, l_i) upon realization of a given route.

Each route of the solution starts and finishes at the depot vertex 0. Let $\vec{r}^i$ be the vector of the route for vehicle i specified as customer vertices, not including the depot as the first and last elements, and $R = \{\vec{r}^1, \ldots, \vec{r}^k\}$ be the list of routes. Then $\vec{r}^i \cap \vec{r}^j = \varnothing$ given that each customer is in exactly one

route and that $\bigcup_{i=1}^{k}\vec{r}^i = V-\{0\}$ as each customer must be included in one route. Also, $\vec{r}^i$ is an ordered vector for all i; thus $\vec{r}^i_j = m, \vec{r}^i_{j+1} = n$ implies that the arc (m,n) is included in route $\vec{r}^i$.

For simulation, the 2D coordinates of the depot and customers are set ranomly as $(x, y)_j$ for $j = 0, 1, \ldots, N$ within a 90×90 grid. The dimensions represent minutes of travel across a hypothetical urban area. The mean time between customers i and j is calculated as the Euclidean distance as follows:

$$d_{ij} = \sqrt{(x_j - x_i)^2 + (y_j - y_i)^2} \tag{4.22}$$

This means the maximum expected travel time is across the grid diagonal for a total travel time of $90\sqrt{2} \approx 127.28$ minutes. Symmetric travel times are assumed, hence $d_{ij} = d_{ji}$. The location parameters of travel times in the lognormal distribution were set as a function of the distance d_{ij}; notably, the exponential decay-based function described in Westgate *et al.* (2016) to allow for reasonable real-life inspired variance. Their paper explored ambulance travel times in the Toronto area and used Bayesian methods to determine hyperparameters in lognormal-based travel times across a network. The following location parameter function is employed:

$$\sigma_{ij}^2 \triangleq M e^{-\lambda d_{ij}} + \delta \tag{4.23}$$

where $M > 0, \lambda > 0, \delta > 0$. The mean values they obtained were $M = 0.2064, \delta = 0.0576$, and $\lambda = 0.00097$. However as these values corresponded to an emergency vehicle that was assumed to have lower variance than common freight vehicles, the variance here was increased by doubling both M and δ while keeping the rate of variance descent in λ. With the mean travel time and location parameters, the scale parameters can then be calculated as below:

$$d_{ij} = \mathbb{E}(t_{ij}) = e^{\mu_{ij} + \frac{\sigma_{ij}^2}{2}} \tag{4.24}$$

$$\mu_{ij} = \ln(d_{ij}) - \frac{\sigma_{ij}^2}{2} \tag{4.25}$$

The lognormal distributions for service times adhere to the Euclidean nature of travel times, so a mean service time is sampled from $\mathbb{E}(s_i) \sim Unif(10, 60)$ for each customer i with an assumption that the standard deviation is half the mean service time. In reality, if the same truck, the same deliverer, and the same load is unloaded, this standard deviation should be much lower. Let the variance of a service be v_i and mean service time be $m_i = \mathbb{E}(s_i)$ for a given customer i. By assumption, $v_i = \frac{m_i^2}{4}$ and, by definition, the mean and variance on the non-logarithmic scale is calculated from the scale and location parameters σ^2 and μ of a lognormal distribution:

$$m = e^{\mu + \frac{\sigma^2}{2}} \tag{4.26}$$

$$v = (e^{\sigma^2} - 1)e^{2\mu + \sigma^2} \tag{4.27}$$

The log-scale parameters μ_i and σ_i^2 are then derived using this relationship between the logarithm and non-logarithmic parameters for a specific customer i as follows:

$$\mu_i = \ln\left(\frac{m_i^2}{\sqrt{v_i + m_i^2}}\right) \tag{4.28}$$

$$\sigma_i^2 = \ln\left(\frac{v_i}{m_i^2} + 1\right) = \ln\left(\frac{1}{4} + 1\right) \approx 0.223144 \tag{4.29}$$

Note that since it is assumed that $v_i = \frac{m_i^2}{4}$, then σ_i^2 becomes a fixed value.

For a given routing schedule R there is a given set of time windows $\vec{e}, \vec{l}$. Let a_i and d_i be the arrival and departure times at and from some customer i; as a feasible routing requires that only one customer is assigned to each route, the index for vehicle number is omitted. It is assumed that service times must begin within a customer's time window, hence if $a_i \leq e_i$ there is an associated waiting time $w_i = e_i - a_i$. Similarly, it is assumed that service finished within a time window must wait for the end of the window before departure, hence if $a_i + w_i + s_i < l_i$ then $d_i = l_i$. Let d_0^i and a_0^i be departure and arrival time at the depot, corresponding to starting and ending times of a route $\vec{r}_i$. An operating cost rate of o_i is used for vehicle i proportion to total travel time $a_0^i - d_0^i$.

Penalty costs are applied for violating the time windows of customer i as p_i^e and p_i^l. Each customer has a rate of early penalty application c_i^e and late penalty application c_i^l. An overtime penalty cost is similarly defined for vehicle i as p_i^o with rate c_i^o. Overtime hours for a given vehicle i occur when vehicle i is in service longer than some regular hour time $h_{i,regular}$. Also, there is a absolute maximum allowable time for a given route h_i, where $h_{i,regular} \leq h_i$.

A simple linear penalty is applied, hence:

$$p_i^e = c_i^e \max\{e_i - a_i, 0\} \qquad \forall i \in \{1, \ldots, N\} \tag{4.30}$$

$$p_i^l = c_i^l \max\{d_i - l_i, 0\} \qquad \forall i \in \{1, \ldots, N\} \tag{4.31}$$

and an overtime vehicle penalty is employed:

$$p_i^o = c_i^o \max\{a_0^i - d_0^i - h_{i,regular}, 0\} \qquad \forall i \in \{1, \ldots, k\} \tag{4.32}$$

The objective function consists of four parts. The first is the traveling cost representing the operating costs such as employee pay and vehicle maintenance defined by:

$$T(R,\vec{a},\vec{d}) = \sum_{i=1}^{|R|} min\{a_0^i - d_0^i, h_{i,regular}\} \qquad (4.33)$$

The second is the total cost associated with fleet size defined by

$$F(R,\vec{a},\vec{d}) = |R| \qquad (4.34)$$

The third is the total penalty cost associated with violating time windows defined by

$$P(R,\vec{a},\vec{d}) = \sum_{i=1}^{N} p_i^e + p_i^l \qquad (4.35)$$

The fourth is the total penalty cost associated with entering overtime hours defined by

$$O(R,\vec{a},\vec{d}) = \sum_{i=1}^{k} p_i^0 \qquad (4.36)$$

The complete objective function is then:

$$Z(R,\vec{a},\vec{d}) = T(R,\vec{a},\vec{d}) + F(R,\vec{a},\vec{d}) + P(R,\vec{a},\vec{d}) + O(R,\vec{a},\vec{d}) \qquad (4.37)$$

4.3.3 Algorithm

The essential algorithm breaks down into several steps.

First, the relative time windows are determined for each customer. Travel and service times are simulated for adjacent locations i, j to set relative time windows locations. Notably, the travel time from location i to j is simulated along with the service time of j. The total time (travel + service) is the sum of two lognormal distributions and has no analytic form. To set the relative early and late time windows for this ordered i, j a percentile forward across the travel and service times is estimated as well as a percentile back across only the service time, called $\alpha_{forward,ij}$ and $\alpha_{back,ij}$ respectively, to determine the relative time windows. The late time window is set as the $\alpha_{forward,ij}$ percentile of all simulated travel and service cumulative times, and a backtrack from this point in time is used to set the $\alpha_{back,ij}$. A large number of samples are simulated (say 500 travel and service pairs) and the percentile marks are examined to see how they affect a given multi-objective function for a variety of possible α values by finding the average cost of a given pair of $\alpha_{forward,ij}$ and $\alpha_{back,ij}$ for simulated *free-flow times*. That is, if service were to immediately begin upon arrival. An approximate solution is of interest, so only a set number of percentiles is examined for both $\alpha_{forward,ij}$ and $\alpha_{back,ij}$ (e.g., some discrete $0 < \alpha < 1$ set likely greater than 0.5 and less than 0.95). The cost function is piecewise lienar having an early penalty if simulated travel time allows arrival before the servicing window, a service time penalty (p_w) which may be zero

(identical for all simulated runs for a given pair), and a late penalty if the simulated service completes after the servicing window ends. It is assumed further that the linear penalty rates are ordered strictly descending as late penalty (p_l), early penalty (p_e), and service time penalty (p_w), that is, $p_l > p_e > p_w$. Therefore each simulated travel and service time free-flow pair has a given associated cost and it is determined which of the discrete $\alpha_{forward,ij}$ and $\alpha_{back,ij}$ pairs has the lowest average cost. The simulated times are kept the same across all $\alpha_{forward,ij}$ and $\alpha_{back,ij}$ pairs, as this represents one set of realized times. The above procedure is repeated for each i,j pair (excluding $i = j$ pairs) to generate a table of relative percentiles and their associated times. Note that a restriction is also imposed on the late time ($\alpha_{forward,ij}$) and the maximum travel time across the diagonal cannot violate the daily hours otherwise the routing would likely be rendered infeasible.

Figure 4.1 illustrates the above concepts. The anchor points are set as $\alpha_{forward} = \alpha_{back} = 0.9$ for some arbitrary sequential i,j visitation sequence.
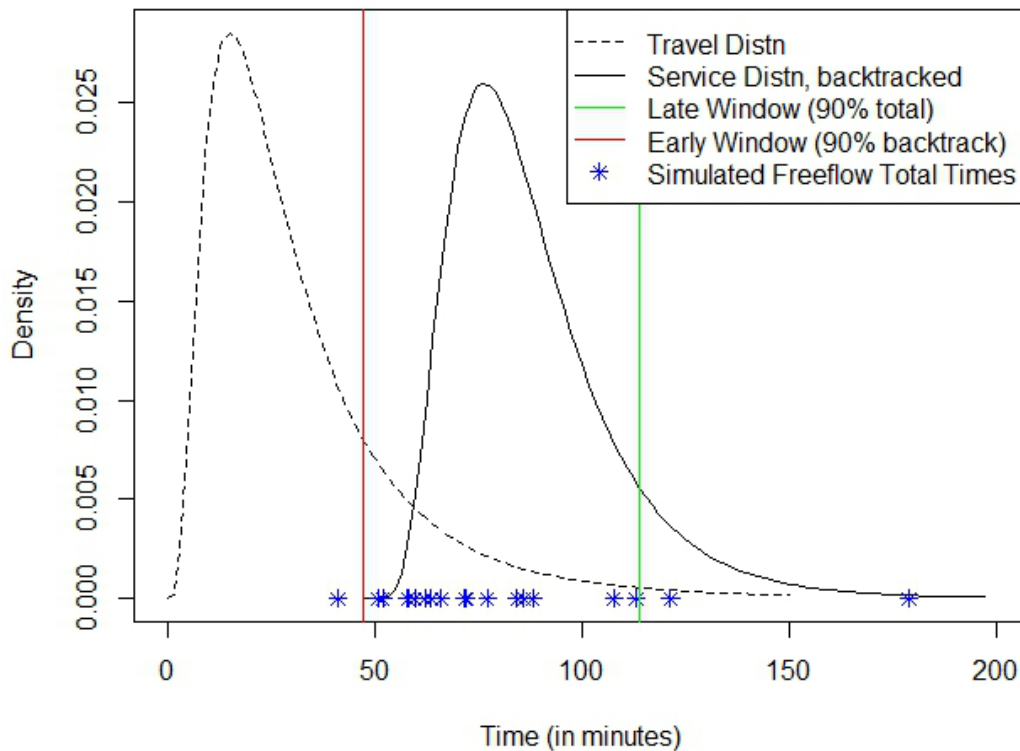
Figure 4.1: Simulated window setting example

The free flow travel + service completion times are represented by the blue asterisks. The 90th percentile of these order times corresponds to the $\alpha_{forward}$ value, with the exact time plotted as the vertical green line. The algorithm then backtracks by the exact 0.90 cumulative probability of the service time to set the early window corresponding to α_{back}. This time is plotted at the red vertical line. In this manner, a single i,j time window is set with known relative early and late window times. The dashed line distribution represents the travel time distribution, and the solid

line represents the service time distribution starting at the early time window. When a route is built in simulation, these relative time windows are found so as to set the ordered time windows in a given route.

As a caveat on the time windows, it is assumed that a given customer j must have identical service times for any i, j location pair. Determined by $\alpha_{back,ij}$ for some predecessor customer i, the optimal $\alpha_{back,ij}$ must be found across **all** predecessor customers i to force equal width service windows. To do this, travel times and service times from customer i to given customer j are simulated and the lowest cost penalty is found assuming only the late and window width penalties are assessed. Since the starting point varies (predecessor customers), the early penalty is not assessed; service is assumed to start immediately upon arrival. Therefore each $\alpha_{back,ij}$ is fixed for a given customer j based on a common service time and the above simulation can be performed by altering only $\alpha_{forward,ij}$ values.

Now with these, the windows can be set and the costs of a given routing evaluated. In the multi-objective function used to evaulate given α pairs the values of $p_l = 6, p_e = 3, p_w = 1$ are used for late, early, and window width penalties respectively. Clarke-Wright is used to establish an initial feasible solution followed by local search updates to improve the solution given simulated travel times; this includes setting the time windows iteratively using the lookup tables followed by a realized testing period. The parallel Clarke-Wright procedure is employed with travel times between customers approximated using the expected travel time as:

$$T_{ij} = \mathbb{E}(t_{ij}) = \mathbb{E}(x \mid X \sim \ln N(\mu_{ij}, \sigma_{ij}^2)) = e^{\mu_{ij} + \frac{\sigma_{ij}^2}{2}} \qquad \forall i, j \in V, i \neq j \tag{4.38}$$

Service times follow a similar derivation to get

$$\mathbb{E}(s_i) = e^{\mu_s^i + \frac{(\sigma^2)_s^i}{2}} \qquad \forall i \in V - \{0\} \tag{4.39}$$

A time-restricted upper bound ia placed on the route lengths as a proportion β of $h_i, 0 < \beta < 1$ for each vehicle i. It is assumed that the max hours are the same across vehicles so let $h = h_i$. In particular, the maximum allowable truck hours set by Federal Motor Carrier Safety Administration (FMCSA) is 11 hours of driving in a given 24 hour period when applied to long-haul freight and so this value is treated as the upper bound on the local delivery routing. If any given route exceeds 11 hours, the full routing is rendered infeasible and cannot be used. The relatively common regular 8 hours is employed to set $h_{i,regular} = 8$ for all vehicles. Therefore overtime penalties are only assessed when a route arrives between 8 and 11 hours after departure. If beyond 11 hours, the route is assessed an infeasible penalty cost of $+\infty$, making the full routing infeasible in the algorithm. To make sure that a feasible solution is used initially, $\beta = 0.8$ initially and this value is decremented by 0.1 until a feasible initial solution is reached. This allows for a somewhat lower the likelihood of overtime hours during later simulation, and as a result, the likelihood that the following improvements will be made. To simplify the presentation of the algorithm let

$$T_{serv,k} = \sum_{i=1}^{|\vec{r}^k|} \mathbb{E}(s_{\vec{r}_i^k}) \tag{4.40}$$

$$T_{trav,k} = \mathbb{E}(t_{0,\vec{r}_1^k}) + \sum_{i=2}^{|\vec{r}^k|} \mathbb{E}(t_{\vec{r}_{i-1}^k,\vec{r}_i^k}) + \mathbb{E}(t_{\vec{r}_{|\vec{r}_i^k|}^k,0}) \tag{4.41}$$

where equation $T_{serv,k} = \sum_{i=1}^{|\vec{r}^k|} \mathbb{E}(s_{\vec{r}_i^k})$ (4.40) is the total expected service time of a

given route k and equation $T_{trav,k} = \mathbb{E}(t_{0,\vec{r}_1^k}) + \sum_{i=2}^{|\vec{r}^k|} \mathbb{E}(t_{\vec{r}_{i-1}^k,\vec{r}_i^k}) + \mathbb{E}(t_{\vec{r}_{|\vec{r}_i^k|}^k,0})$

(4.41) is the total expected travel time of a given
route k. The Clarke-Wright Algorithm with these requirements and definitions is outlined in
Algorithm 1 found in Figure 4.2.

To test whether the given β value produces a feasible initial routing, the customer time windows
are set to test for any violations. For a given routing R, windows are set by iteratively based off
the relative early and late window lookup tables of i,j ordered location pairs. In particular, from
the depot to the first customer i the customer's late window is set as the relative late time window
from location zero (the depot) to customer i (element $0,i$ in the lookup table of late windows).

Algorithm 1 Clarke-Wright Algorithm

1: **function** CW(Graph G)
2:     **for all** $i,j \in V - \{0\}, i \neq j$ **do**
3:         Savings $s_{ij} \leftarrow T_{i,0} - T_{ij} + T_{0,j}$
4:     **for all** $i \in V - \{0\}$ **do**
5:         Build route $\vec{r}^i = \{i\}$
6:     Order *savings* in decreasing order in Queue Q
7:     **while** Q Not Empty **do**
8:         $maxSavings \leftarrow Q.dequeue$, where $maxSavings = s_{ij}$ for some i,j
9:         **if** i is last customer of route $\vec{r}^a$, j first customer of route $\vec{r}^b$ **then**
10:            **if** $T_{serv,a} + T_{serv,b} + T_{trav,a} + T_{trav,b} + T_{ij} - T_{0j} - T_{i0} \leq \beta h$ **then**
11:                Merge Routes $\vec{r}^a, \vec{r}^b$ by deleting arcs $(i,0),(0,j)$ and adding arc (i,j)
12:     **end while**

Figure 4.2: Clarke-Wright Algorithm

Then using this time point as an anchor, the early window is set by backtracking the appropriate
percentile by using the early window lookup table. The anchor time is then the late window and

86

determine the next customer's windows relative to the first customer, again by using the lookup tables. This procedure of setting the relative late window, backtracking to find the early window, and anchoring to the late window for traveling to the next customer repeats until each customer in a given route receives a time window. This procedure is applied to each route so that every customer receives a time window.

With the time windows established, a test simulation can be run to calculate the cost of a route given by equation $Z(R,\vec{a},\vec{d}) = T(R,\vec{a},\vec{d}) + F(R,\vec{a},\vec{d}) + P(R,\vec{a},\vec{d}) + O(R,\vec{a},\vec{d})$ (4.37). With repeated test simulations the mean route cost $\bar{Z}(R,\vec{a},\vec{d})$ is found. This value serves as the iteration score with which the score of a possible route improvement can be compared. Again if the initial Clarke-Wright solution for a given β yields an infinite cost (hence infeasible route), decrement β is decremented by 0.1 until a finite cost initial routing is found.

The improvements explored in reducing the route score are a merge-based reduction, an insertion reduction, and a 2-customer exchange reduction. Again, all time windows are set by using the lookup tables. The merge-based reduction is attempted for each ordering pair of routes; therefore a new merged route is simulated $\vec{r} = \{\vec{r}_i, \vec{r}_j\}$ for all $i, j \in \{1, 2, \ldots, |R|\}, i \neq j$. To limit the number of simulated tentative routes, Wang and Lin's (2013) metaheuristic is employed. Notably, in an effort to balance the travel times, the insertion and 2-exchange reductions are employed from the latest arriving vehicle to earliest first; if this does not work, the second earliest, the third earliest, etc. options are explored. To save some computational time, the aim is move from the single latest route instead of simultaneously decreasing from latest, to second latest, and so on.

If any of the merge, insertion, or 2-customer exchange reductions has a lower score than the incumbent, the incumbent route can be changed to that of the route of maximum reduction (lowest cost routing). If no reduction can be made, the algorithm terminates. See Algorithm 2 found in Figure 4.3 for the pseudocode.

4.3.4 Specifications

The algorithms are coded in Java and runs on a Dell OptiPlex 740 Enhanced desktop with an AMD Athlon(tm) 64 X2 Dual Core Processor 5800+ CPU @ 3000 MHz, 2 Core(s), 2 Logical Processors on a 64-bit operating system with 4.00 GB of RAM. Due to the limitations in memory and the fact that tables of generated statistical distribution objects are used (not primitives) from the Apache Commons Mathematics Library in addition to the α forward and back lookup tables, instances of only up to 200 customers were evaluted. Data files were generated containing travel and service time information in Microsoft VBA and saved these data into text files to be read at runtime by the Java program.

4.3.5 Results

Using the above specifications, 100 random location mappings were generated for 10, 20, 30, 50, 75, 100, and 200 customers. This yielded a total of 700 different maps. Each of these maps was studied three times and the average simulated cost after reductions was found. The average

simulated cost after reduction was compared to the initial Clarke-Wright solution to find a percentage reduction. A daily fixed truck cost of 8 was applied, an early penalty rate of 2, a late penalty rate of 5, an overtime penalty rate of 2, and a travel cost rate of 1.

To compare the reductions across the number of customers, the score was treated as the average percentage reduction in cost from initial Clarke-Wright solutions to a reduced solution. For the 100 randomized maps of customer sizes, the interest lay in the averages across maps for a given customer size. Let X_n be the score for n customers as the average percentage reduction from the Clarke-Wright initial solution to the reduced solution. This measure allows a comparison of the results across a varying number of customers to determine if the algorithm can produce similar reductions regardless of the number of customers with locations randomized across the given plane. It also shows differences in the overall running time of the algorithm itself for a given number of customers so that average running times can be determined. This allows testing of whether the algorithm represents a time-manageable solution for reducing overall routing costs.

Algorithm 2 Reduction Algorithm

1: **function** Reduction (Graph *G*, Initial feasible routing *R*)
2:     **repeat**
3:         *bestRouting* ← copy of R
4:         **for all** *i,j* routes ∈ *R, i≠j* **do**
5:             Try merging routes *i,j* into new routing $R*$
6:             Reset windows for new routing based off lookup table
7:             Score simulated routing cost
8:             **if** Merged route $R*$ has lower cost than current *bestRouting* **then**
9:                 *bestRouting* ← $R*$
10:         r_e ← route predicted returning earliest to depot in *R*
11:         r_l ← route predicted returning latest to depot in *R*
12:         Boolean flag *reduced* ← *FALSE*
13:         **while** $r_e ≠ r_l$ AND *reduced = FALSE* **do**
14:             **for all** customers $i ∈ r_l$ **do**
15:                 **for all** positions $j ∈ r_e$ **do**
16:                 Remove customer *i* from r_l
17:                     Insert customer *i* in r_e at position *j* as new routing $R*$
18:                 Reset windows for new routing based off lookup table
19:                 Score simulated routing cost
20:                 **if** New routing $R*$ has lower cost than current *bestRouting* **then**
21:                   *bestRouting* ← $R*$
22:                   *reduced* ← *TRUE*
23:             **for all** customers $i ∈ r_l$ **do**

```
24:                    for all customers j ∈ r_e do
25:                        Swap customers i,j in routes r_l,r_e respectively for new routing R*
26:                        Reset windows for new routing based off lookup table
27:                        Score simulated routing cost
28:                        if New routing R* has lower cost than current bestRouting then
29:                            bestRouting ← R*
30:                            reduced ← TRUE
31:                    if reduced = FALSE then
32:                                    r_e ← route predicted returning next earliest to depot in R after current r_e
33:                end while
34:                Boolean flag continue ← FALSE
35:                if bestRouting ≠ R then
36:                    R ← bestRouting
37:                    continue ← TRUE
38:        until continue = FALSE
```

Figure 4.3: Reduction algorithm

As seen in Table 4.1: Mean score reductions across customer number from initial solution, it seems that the procedure could reduce costs a minimum of about 23% across all customer numbers. As expected the running times grow super-linearly, mainly from how the reduction algorithm explores the current route's possible local search neighborhood. With 200 customers, this time is only about 190 seconds. This means that the algorithm could easily be applied to more customers while maintaining a reasonable amount of computation time, with the limit on number of customers instead being restricted by lack of local memory on the test computer.

Number of Customers n	X_n	Mean Running Time (seconds)	Mean # Reductions
10	27.634	0.027	1.181
20	24.104	0.415	3.378
30	25.536	1.253	5.109
50	24.584	3.966	7.755
75	23.221	9.818	11.454
100	25.988	20.174	15.751
200	30.577	187.291	35.038

Table 4.1: Mean score reductions across customer number from initial solution

To establish baseline comparisons, an assessment is done to see how the α values affect the cost of routing. This allows a comparison of the costs associated with the more computationally demanding optimization scheme with those of the simple static window setting that requires a simple CDF calculation for chosen percentiles.

4.3.5.1 Fixed Alpha

Instead of optimizing the windows at runtime, the windows are fixed a priori. In particular, there will be no simulation to find heuristic optimal $\alpha_{forward,ij}$ and $\alpha_{back,ij}$ for ordered customer visitation i,j, resulting in improved computation times. The forward and back times are ignored in favor of directly using the travel and service distributions, providing instead $\alpha_{travel,ij}$ and $\alpha_{service,ij}$. The i,j sequences are fixed to the same α_{travel} and similarly for $\alpha_{service}$, i.e. the algorithm is given a fixed α_{travel} and $\alpha_{service}$ such that $\alpha_{travel,ij} = \alpha_{travel}$ and $\alpha_{service,ij} = \alpha_{service}$ for all i,j. Both parameters are varied over the percentiles 0.5, 0.6, 0.7, 0.8, 0.9, 0.95, 0.99. To simplify analysis, only the 100 different 50 customer maps are tested for a total of $100 \times 7 \times 7 = 4900$ scenarios. Each of these scenarios are simulated 2 times. These fixed conditions are compared to the optimized routing from part 1 of the analysis in which windows are set via simulation, notably looking at simulated costs prior to and after reductions as the metric. Upon completion of the simulation, the reduced simulated costs were compared by finding the percent change in reduced costs from the optimal value to the new static window value as follows:

$$\text{Percent Change in Reduced costs} = \frac{(Cost_{static} - Cost_{optimal})}{Cost_{optimal}} \times 100 \qquad (4.42)$$

This means that a positive percentage is an increased cost for the static from the base optimal and vice versa. With this value calculated for each map and for each (α_{travel}, $\alpha_{service}$) pair, the average percent change is then calculated in reduced costs for the 100 maps evaluated at a given (α_{travel}, $\alpha_{service}$). This is the metric for final evaluation. Figure 4.4 below shows the resultant percent change surface.
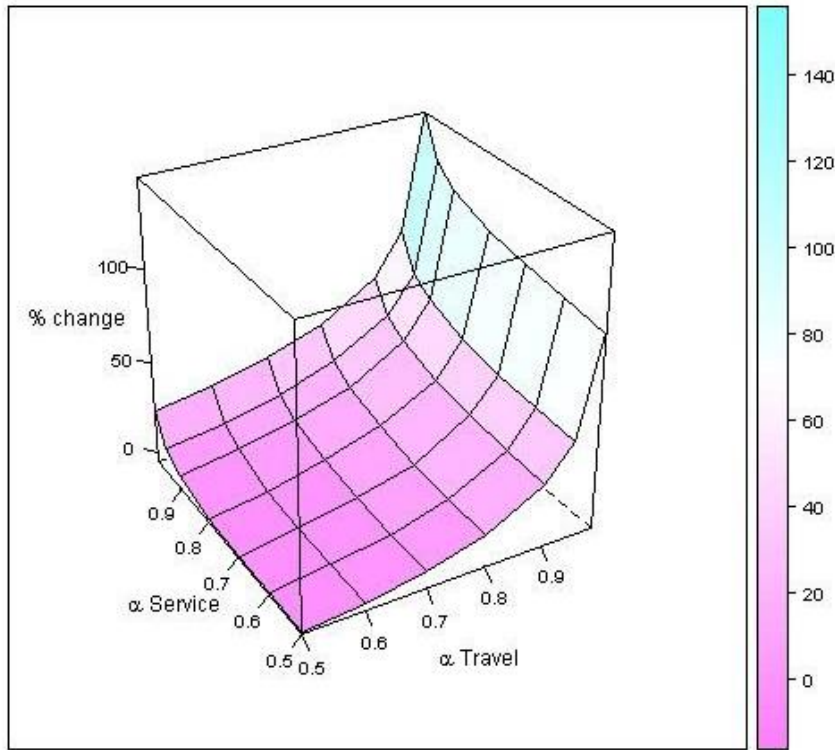
Figure 4.4: Average percent change in reduced costs from optimal to static window setting

For the given penalty rates and parameter choices for travel and service distributions, it appears that a highly aggressive window setting with $\vec{\alpha}$ = (0.5, 0.5) provides a slight reduction in costs compared to setting the windows heuristically at runtime. These values limit potential delay buffers that higher α values generate implicitly for the sake of assigning more customers to a given vehicle route. Of note is that the more conservative α values above about 0.9 lead to large increases in costs, especially for travel time window adjustment as seen in the cost spike along the α Travel axis.

4.3.6 Observations

While the procedure shows promise, only a hypothetical application has been examined. Most importantly, real-world data on travel times and service times would better assess performance.

Another thought is that the procedure could be extended to a Bayesian statistics framework in which times are estimated via prior distributions (not necessarily lognormal as used in this paper) that can be updated upon realized routes. Ostensibly this would allow for any given distribution to be used to set relative time windows and test routes. Notably, a business with repetitive deliveries could benefit from using the same routing for a given period of time to obtain new samples and update the routing accordingly. Additionally, with customer penalty assessment data, customer-specific penalties could be introduced. This would likely result in higher late penalties for larger and/or more demanding clients due to phenomena such as customer loyalty and product obsolescence. Lastly, it would be worth investigating some of the natural extensions to the uncapacitated VRP, such as putting capacities on trucks or implementing precedence constraints as in pickup-and-delivery customers.

4.4   BLOCK-BUILDING METHODOLOGY

A second solution methodology makes a single pass through the set of customers to be visited and identifies an assignment of trucks which is both efficient and feasible. The method is based on List *et al.* (2003) and List and Turnquist (1993). These formulations are based on Goeddel (1975) as enhanced by Ball, Bodin, and Greenberg (1985) and described by Wren and Rousseau (1995).

4.4.1 METHODOLOGY

The problem formulation can be stated verbally as follows:

Maximize the on-time performance
Minimize the cost of the service provided

Subject to:

> Picking up all shipments and/or delivering all shipments
> Not exceeding the capacity of any vehicle
> Not exceeding a maximum tour duration for any vehicle

In creating the solution procedure, several assumptions are made.

- The fleet size is fixed.
- The vehicles are always available. (Maintenance spares exist.)
- There is a single depot where trucks originate and terminate their tours.
- Both pick-ups and deliveries can be made by the same vehicle during a given tour. But delivered shipments must originate at the depot and picked-up shipments must be carried to the depot. Said another way, the depot must be one end of the trip for each shipment.
- Vehicles have a limited capacity and each shipment uses some of that capacity.
- The total time for each tour is the sum of the random variables that describe the travel times between stops and the amount of time spent in pick-up or delivery. This means the travel times are stochastic as are the pick-up and delivery times.
- An upper limit exists for the length of any tour, measured in time.
- Every customer visit (stop) has an AW at the beginning of the service window and a DW at the end.
- If vehicles arrive earlier than the AW, they wait until the AW to commence loading or unloading.
- Vehicles can depart as soon as they are finished with the loading/unloading task.
- The cost equation has five components: a) the number of vehicles used, b) vehicle hours, c) vehicle miles, d) the penalty cost for arriving early, and e) the cost for departing late.
- On-time performance for arrivals is measured by the probability of arriving within the AW. The same pertains to the probability of departing during the DW.
- The probability of an OTA can be improved by adding slack time to the schedule; that is, by arriving early, off-site, near a customer's location, at the cost of adding time to the tour and potentially increasing the fleet size.
- The probability of an OTA can also be improved by increasing the fleet size. A larger fleet reduces the number of customer stops per vehicle and adds more resource availability.
- All shipments are accommodated, either directly, or by outsourcing (at a significant cost).
- Vehicles are interchangeable.
- Drivers are always available.
- A load and a shipment are the same thing. The two words can be used interchangeably. The same is true for the words customers, consignees, and stops.

The AWs indicate when the loading or unloading dock is available for use. A vehicle is "early" if it arrives before the AW begins. It is "late" if it arrives after the AW ends. It is "delayed" if it leaves after the DW ends. When the truck arrives early, it waits until the beginning of the AW. There are costs for being early or delayed, but not late. (This could be changed. The lateness is monitored.) The cost of being early is less than the cost of being delayed.

The AWs are of two types. AWs of Type 1 are set by the customers. Examples include AWs set by warehouses, retail stores, or manufacturing plants that belong to external entities. The carrier and/or shipper can request a specific AW, but cannot control it. An AW of Type 2 is set by the carrier. This happens when the shipper also operates the truck fleet and controls the operation of the customers being served. Illustrations are supplier managed filling stations, grocery stores, or big box retail stores. The start times for the Type 2 AWs are choice variables. For both Type 1 and Type 2 AWs, the on-time performance is optimized by assigning loads to vehicles, sequencing the visits to customers, and in the case of Type 2 AWs, specifying their start times.

Two objectives apply. The first is to maximize the on-time performance for the customers. This is done by minimizing the combination of the average delay for all customers (a mini-avg or mini-sum objective which is an L-1 norm) and the greatest lateness among all customers (which is an L-∞ norm). The second objective is to minimize total cost. Other "objectives" such as minimizing the fleet size and the duration of time required to complete all deliveries (the makespan) are treated parametrically.

The choice variables are: 1) the assignment of loads to vehicles, 2) the sequencing of visits to customers, and, 3) in the case of the Type 2 AWs, the start times. The fleet size is an input as well as the start times for the Type 1 AWs. The sequencing and assignment introduces slack to minimize late arrivals, ensure early or OTAs. Adding slack lengthens the tour durations. Larger fleet sizes improve on-time performance but also increase cost.

The search procedure tracks the performance of various routing and scheduling solutions. The run-cutting procedure develops the solutions. The procedure can deal with large-scale problems and is easy to understand. Optimally cannot be assured, but the procedure's performance can be compared to optimal solutions for small problems.

The pseudo-code representation of the algorithm is shown in Figure 4.5.

Algorithm 3 Assign load sequences to vehicles

1: **function** Assign and Sequence(Graph G, Stop info array $Stop$, Fleet sizes array M, Simulation specifications S)
2: **for all** Fleet sizes $k \in M$ **do**
3:       **for all** $n = 1,...,S.numRuns$ **do**
4:            **for all** i,j node pairs in G **do**
5:                Sample for travel time $t[i,j]$
6:            **for all** i Stops in $Stop$ **do**
7:                Sample for service time $Stop[i].tSvc$
8:            Create array VI containing $Stop$ indices ordered by earliest arrrival window
9:            **for all** vehicles $j \in 1,...,k$ **do**
10:                $veh[j].arrive,veh[j].begin,veh[j].early,veh[j].depart,veh[j].delay \leftarrow 0$
11:                $veh[j].loc \leftarrow 0$ where location 0 is depot
12:            $maxDelay,aveDelay,timesDelayed \leftarrow 0$
13:            **for all** Indices $i \in V I$ **do**
14:                $vehTemp \leftarrow veh$
15:                **for all** vehicles $j \in 1,...,k$ **do**
16:                    $vehTemp[j].arrive \leftarrow t[veh[j].loc,Stop[i].node]$
17:                    $vehTemp[j].early \leftarrow \max(Stop[i].bWin - vehTemp[j].arrive,0)$
18:                    $vehTemp[j].begin \leftarrow \max(vehTemp[j].arrive,Stop[i].bWin)$
19:                    $vehTemp[j].depart \leftarrow \max(vehTemp[j].begin+Stop[i].tSvc,Stop[i].eWin)$
20:                    $vehTemp[j].delay \leftarrow \max(vehTemp[j].depart-(Stop[i].eWin+OTW),0)$
21:            Select $jBest$ such that $vehTemp[jBest].arrive \leq vehTemp[j].arrive$ and $vehTemp[jBest].delay \leq vehTemp[j].delay \ \forall j \in \{1,...,k\},j \neq jBest$
22:                **if** $veh[jBest].delay > 0$ **then**
23:                    $aveDelay \leftarrow aveDelay + veh[jBest].delay$
24:                    $timesDelayed \leftarrow timesDelayed + 1$
25:                    **if** $veh[jBest].delay > maxDelay$ **then**

26:	$maxDelay \leftarrow veh[jBest].delay$
27:	$vehTemp[jBest].delay \leftarrow veh[jBest].delay + vehTemp[jBest].delay$
28:	$veh[jBest] \leftarrow vehTemp[jBest]$
29:	$load[i] \leftarrow veh[jBest]$
30:	$load[i].veh \leftarrow jBest$
31:	**for all** Indices $i \in V I$ **do**
32:	$delayStops[n,i] \leftarrow load[i].delays$
33:	**for all** vehicles $i \in 1,...,k$ **do**
34:	$delayV ehicles[n,i] \leftarrow veh[i].delay$
35:	$maxDelays[n] \leftarrow maxDelay$
36:	$aveDelays[n] \leftarrow aveDelay/timesDelayed$
37:	Provide summary of results through $aveDelays, maxDelays, latenessStops$, and $latenessVehicles$ before iterating to next fleet size

Figure 4.5: Block and truck tour building algorithm

Based on the results for each fleet size provided by the algorithm, determine which solution is optimal: the trade-off between delay performance and fleet size. The truck tours are developed using a greedy heuristic that assigns the best available truck to the loads in chronological order, seen prominently in lines 15-21 of code.

The equivalent math programming formulation is a stochastic version of Bender's decomposition. Each sub-model is a realization of the travel times and load-unload times. It has a probability that the scenario arises. These probabilities are used as weights in computing the overall objective function. The overarching choice variable is the fleet size. The assignment of loads to vehicles and the tour sequences vary by sub-model. The overriding purposes are to 1) identify a fleet size that can accommodate the stochasticity in an acceptable manner and 2) seek general patterns in the assignment of loads to vehicles.

4.4.2 EXAMPLE APPLICATION

The example application is hypothetical. The network comprises 10 nodes. Node #1 is the depot. New networks are created using stochastic equations to generate the travel times. The equation is $t_{ij} = a_{ij} + b_{ij}r_{ij}$ where t_{ij} is the travel time, r_{ij} is a uniform random variable on the interval [0,1] and a_{ij} and b_{ij} are constants. Table 4.2 shows a sample set of values. The unit of travel time is one minute. The travel times in the network are symmetric. That is, t_{ij} is the same as t_{ji}.

					Times					
	1	2	3	4	5	6	7	8	9	10
1	1	20	10	13	15	21	16	15	28	19
2	20	1	21	22	18	7	17	17	22	21
3	10	21	1	19	17	14	17	20	5	30
4	13	22	19	1	22	28	30	29	30	7
5	15	18	17	22	1	12	16	27	16	22
6	21	7	14	28	12	1	28	14	20	10
7	16	17	17	30	16	28	1	6	28	23
8	15	17	20	29	27	14	6	1	27	15
9	28	22	5	30	16	20	28	27	1	12
10	19	21	30	7	22	10	23	15	12	1
10	= #Loc				1 Time unit = 1 minute					

Table 4.2: Network travel times

Twenty visits are to be made, as shown in Table 4.3. Each one has a location (2 .. 10), a beginning time for servicing/visitation (*bWin*), an ending time for servicing (*eWin*), and a service time (*tSvc*). The locations are the nodes also shown in Table 4.3. The unit of time is one minute. (An 8-hour day is 480 minutes long.) New problems are created using a stochastic equation to compute the service times. The equation is $t_s = a_s + b_s r_s$ where t_s is the service time for stop s, r_s is a uniform random variable on the interval [0,1] and a_s and b_s are constants. (Besides the twenty actual stops there are two more, stop #0 which is departure from the depot at the beginning of the tour and stop #21 which is the return to the depot at the end of the tour.)

Trucks that arrive before bWin must wait until bWin to start their servicing. They are considered late if they start servicing after bWin + OTW where OTW is the duration of the on-time window. OTW is set to 10 minutes. Trucks cannot leave until eWin. They are considered delayed if they depart later than eWin + OTW. The same value of 10 minutes is employed.

Stop	Loc	bWin	eWin	tSvc
1	8	135	165	18
2	10	240	270	20
3	2	195	255	23
4	7	15	45	20
5	4	180	225	23
6	3	240	255	16
7	3	105	150	23
8	6	195	225	20
9	8	285	315	23
10	9	435	480	21
11	8	195	240	16
12	6	240	255	21
13	10	285	345	20
14	3	360	390	23
15	5	405	450	18
16	9	300	360	15
17	4	255	300	16
18	8	345	375	15
19	6	255	285	24
20	2	300	315	24

Table 4.3: Visits and their characteristics

A business day of 10 hours is assumed, which is equivalent to 600 minutes. For each realization of the problem setting, a single day is examined.

The fleet size can range from 1 to 10 trucks. It is set before each solution is obtained and its impact on the objectives is explored parametrically. All truck tours originate and terminate at the depot (node #1). An upper bound can be imposed on the capacity of each truck.

For a given problem (set of nominal travel times and service times), stochasticity is introduced by pre-multiplying the nominal times t_{nom} by a random variable r_k. That is, $t_k = t_{nom} r_k$. And r_k follows a four-point discrete distribution. Specifically, the possible values of r_k are 0.8, 1.0, 1.5, and 2.0 and they have probabilities of 10%, 50%, 20% and 10%. To illustrate, this means there is a 20% chance that r_k will be 1.5 and therefore t_k will be 1.5 times the nominal value. The travel times and load/unload times are both treated this way. And they are assumed to be independent random variables.

A set of stochastic realizations are analyzed for a given set of nominal travel times and service times so that the scope of performance variations can be understood. So far, the number of realizations studied is 100. Based on these, plots and tables are created to portray the results. Figure 4.6 presents the distributions of delays for the 20 customer visits when three trucks were employed in the problem setting described earlier. (Visit #0 is departure from the depot at the beginning of the tour and visit #21 is the return to the depot at the end of the tour.) As can be seen, for about 10 customers, the delay is almost always zero. For four others, visits 8, 16, 17, and 20, the delay is considerably different, being 40 minutes or more. Further, visit 18 typically has delays of 10 minutes or less, but the value can reach up to 30.
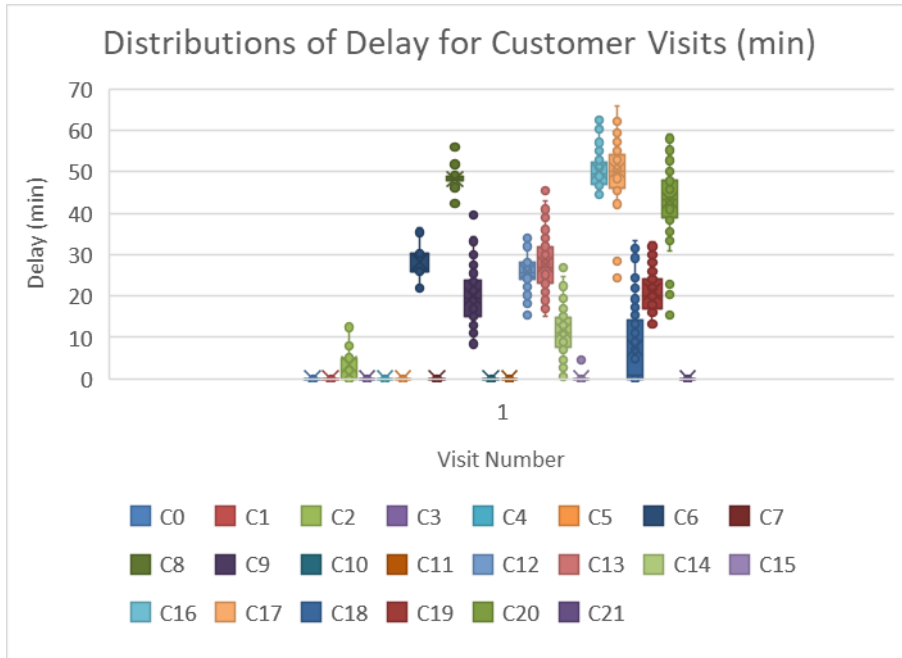
Figure 4.6: Distributions of delays for customer visits – 3 trucks employed

Figure 4.7 shows the distribution of tour times for the three vehicles in this same analysis. All three vehicles have tour durations of about 400 minutes (about 5 hours) although the range is from 380 to 440. Consequently, a solution involving three trucks is feasible although the customer delays may be greater than desirable.
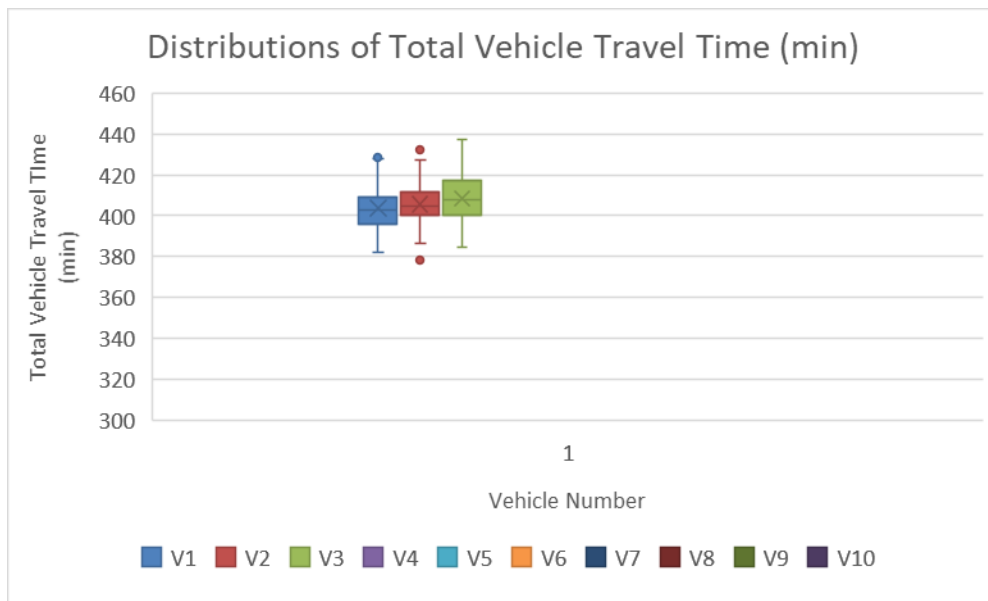
Figure 4.7: Tour Durations for the three vehicles in the scenario analyzed

As would be expected, the fleet size has a major impact on the quality of the solution. Figure 4.8 shows the performance that results from fleet sizes ranging from 1 to 6 trucks. The figure presents data for 1) the early arrivals and 2) delays, both average and maximum values. A visit to a customer is considered early if the truck arrives before the beginning of the servicing window. The visit is

97

"late" if the truck arrives later than 10 minutes after the beginning of the service window. The visit is "delayed" if the truck leaves later than the end of the service window.
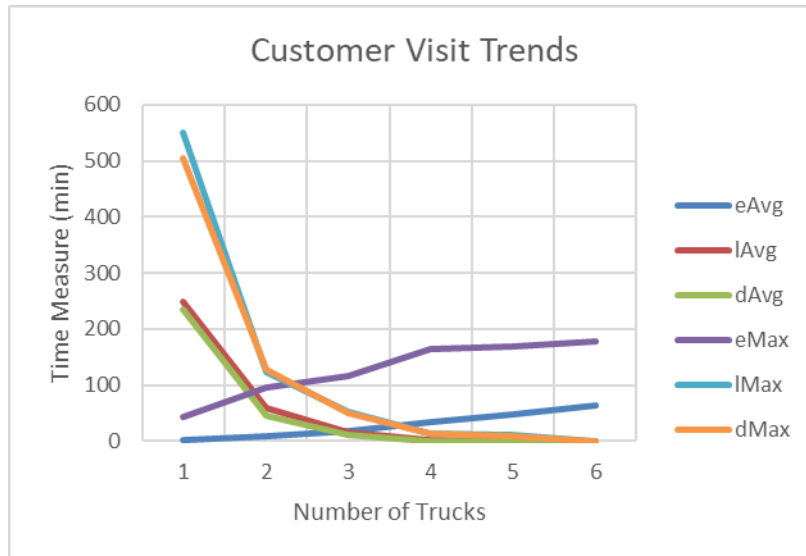
Figure 4.8: Trends in early arrivals and delays

It is easy to see that the delays fall toward zero rapidly as the fleet size increases from 1 to 3 trucks. At 6 trucks, the delays are all zero. Hence, for this setting, a fleet size of 4 or more trucks ensures that reasonably high-quality service will be provided.

It is also important to view these trends in performance from the perspective of the trucks providing the service. Figure 4.9 presents trends in 1) the number of visits assigned per truck, 2) the extent to which the trucks are early and 3) the extent to which they are delayed; both average and maximum values.
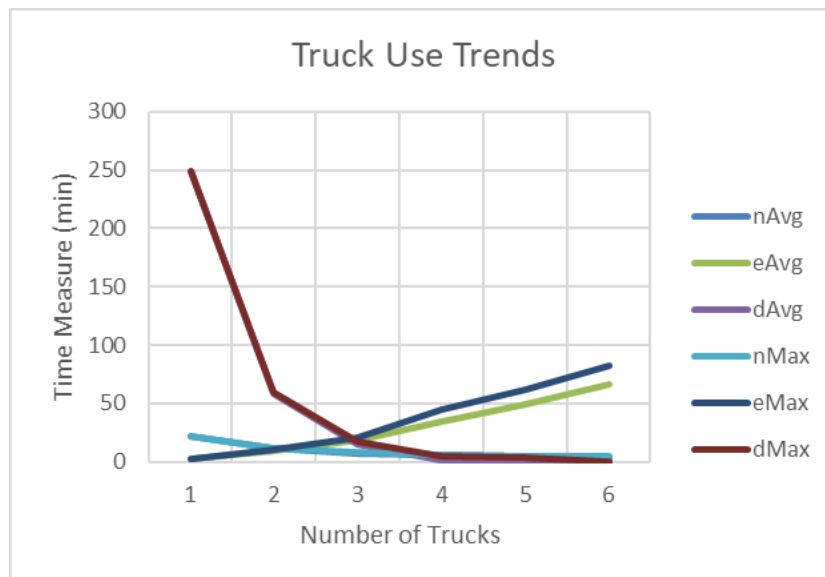
Figure 4.9: Trends in truck performance as affected by fleet size

98

It is easy to see that truck performance improves dramatically as the fleet size grows from 1 to 4 trucks. Beyond that, the improvement is very gradual. The delay performance improves most dramatically, from an initial value of 250 minutes down to nearly zero. It is also easy to see that the reason why this happens is because the early arrivals continue to increase, starting from zero and reaching up to 60 minutes. The implication is that to achieve the objective of minimizing delays (providing reliable service), at least 4 trucks are required and early arrivals of 50 minutes are more are involved.

4.5   SUMMARY

This section has presented methods for solving the stochastic vehicle routing and scheduling problem. This problem lies at the heart of operating a freight transportation system. Nominally, the objective is to find an assignment of loads to vehicles and routings for the vehicles that optimizes all the performance metrics. In some instances, the loads are full truckloads, in which case the vehicles are assigned to carry loads from one point to another in sequence. Alternately, trucks may be picking-up loads or delivering loads. In yet a third option, trucks may both picking-up and dropping off loads, as with local couriers. The objectives are often to 1) minimize total cost, 2) maximize on-time deliveries, 3) minimize the fleet size, and 4) maximize vehicle utilization. Other objectives include 5) maximizing on-time performance and 6) maximizing the lowest on-time performance among all the vehicles employed.

The first method uses a simulation-based heuristic starting from an initial Clarke-Wright solution followed by merge, insertion, and 2-interchange reduction in order to solve the vehicle routing problem with stochastic travel times and soft time windows. Unique to this method is the fact that time windows are set at runtime based off a lookup table of previously simulated ordered customer pairs, corresponding to supplier-side control of a supply chain. It is able to solve in reasonable time on a relatively weak computer a problem instance with 200 customers for a given parameters in a few minutes, and this resulted in at least 23% cost reductions for 10 or greater customers. This could easily be extended to larger problem instances on larger RAM computers, as storage space is the critical element in the algorithm.

The second method makes multiple single passes through the set of customers to be visited based on Monte Carlo simulations of the location-to-location travel times and the loading/unloading times and identifies assignments of the trucks to the customer visits which are both efficient and feasible. The objectives are to 1) maximize the on-time performance and 2) minimize the cost of the service provided subject to: a) picking up all shipments and/or delivering all shipments, b) not exceeding the capacity of any vehicle, and c) not exceeding a maximum tour duration for any vehicle. One of the main insights provided by the method is the relationship between the reliability of the service provided and the size of the fleet employed. As would be expected, it shows that higher reliability is provided by larger fleet sizes. But more importantly, it quantifies the extent of that improvement through the stochastic analysis presented. It also allows the analyst to see if there are consistent patterns in the assignment of customer visits to trucks across the problem realizations examined.

5.0   SITING ANALYSIS

5.1   INTRODUCTION

Building a distribution center, a terminal, or any other type of large facility is a major investment decision. Ceteris paribus, it makes sense to cite these facilities at locations where the travel times will be reliable to all major trip origins or destinations. On the manufacturing side, it makes sense to locate plants where the inbound and outbound travel times are reliable. Of course, the travel times are only part of the overall stochastic process. Hence, solutions that focus on the travel times are only myopically optimal, but it still an interesting area on which to focus research.

5.2   PREVIOUS WORK

As early as 1979, efforts regarding siting analysis have been completed. Mirchandani and Odoni (1979) may have been the first to consider the location of facilities on networks where the travel times were stochastic. They examined problems where the travel times on the network links were random variables with discrete probability distributions. They demonstrated that solution algorithms for such problems can be developed and reasonable size problems can be solved if the number of states of the system (considering the stochastic travel times) is small. More from the standpoint of siting emergency response services than logistics facilities, Mirchandani (1980) again considered the problem of locating facilities when the travel times are stochastic. He shows that realistic and reasonable size problems can be formulated and solved using a variety of solution techniques.

Daskin (1985) reviewed the location decision-making literature and indicated that "both demands and link travel times should, in principle, be treated as random variables" as should the demands. But his review does not identify location models where the network travel times are stochastic.

Owen and Daskin (1998) provided a second review in which stochastic location problems are considered. They indicated that "any number of system parameters might be taken as uncertain, including travel times, construction costs, demand locations, and demand quantities. The objective is to determine robust facility locations which will perform well (based on the defined criteria) under several possible parameter realizations."

Despite these assertions that it is important to treat the travel times as stochastic, it appears that only limited work has been done to advance this frontier. Wang and Ma (2008) appear to be the next authors to explore ways to solve this problem. They use a mixed genetic algorithm to solve problems of various sizes. The results are compared with two greedy heuristic algorithms which have been shown to be good at solving set covering location problems. The computational experiments showed good performance for the mixed genetic algorithm.

More recently, Fazel-Zarandi, Berman, and Beck (2013) have considered stochastic facility location / fleet management problems where the travel times are random variables. They use stochastic programming to solve the problem. Two-level and three-level logic-based Benders'

decomposition models are employed. The computational experiments showed that these developed models can substantially outperform the integer programming model the authors also presented to determine optimal siting solutions.

5.3   METHODOLOGY

The solution methodology employed is relatively simple and straightforward. It uses Monte Carlo simulation to assess the reliability of the delivery service quality provided by candidate distribution center sites and then it identifies the best ones to choose. Because of this, it provides useful and meaningful results which are easy to understand.

The method proceeds as shown in Figure 5.1.

---

1) Specify the location of the customers (sites) to be visited, the locations of the candidate distribution centers (DCs), and the statistical characteristics of the travel times from the DCs to the customer sites.
2) For each DC
   a. Conduct a Monte Carlo simulation of trips made from the DC to the customer sites for different times of day.
   b. Develop CDFs of the travel time distributions for each of the DCs
3) Identify the non-dominated DCs
4) Select the best DC based on the importance of the performance metrics assessed.

Figure 5.1: Site selection algorithm

5.4   EXAMPLE APPLICATION

The example application is predicated on a hypothetical urban area. As shown in Figure 5.2, there are 20 customer locations to be served and five centrally located candidate distribution center (DC) sites.

The hypothetical travel times are developed in the following manner. First, it is assumed that there are four time periods during typical workdays, AM, Midday (MD), PM, and nighttime (NT) In addition, the distribution of trips from the DC to the customer sites is assumed to be 25% during the AM period, 40% midday, 25% during the PM period, and 10% at night. Second, for each of these time periods there are three parameters that describe the travel times involved: 1) circuity measures that convert the Euclidian distance into an over-the-network distance, 2) travel time multipliers which convert the over-the-road distances into travel times by time period, and 3) coefficients of variation that allow computation of standard deviations. The travel times are assumed to be lognormal. The circuity values are (1.3, 1.2, 1.4, and 1.1) for the (AM, MD, PM, and NT) periods respectively, reflecting the effects of congestion on route choice – i.e., more circuitous routes during time periods with higher congestion. The travel time multipliers are (2.4., 2.1, 2.5, and 2.0), again reflecting the impacts of congestion. The coefficients of variation are (15%, 10%, 20%, and 5%), again because of congestion. These are all hypothetical, but perceived to be reasonable values. For example, urban truck trips tend to occur predominantly during the daylight

hours and during the midday especially (Rodrigue, 2017). Network circuity is typically in the range of 1.2 (Levinson, 2012). Travel speeds are often 20-30 mph, which implies travel time multipliers of 2.0-3.0.

Figure 5.2: Customer sites (red) and potential distribution center locations (blue)

For a specific realization of the problem, Figure 5.3 presents the CDFs for travel times to all the customers from the five candidate DCs.
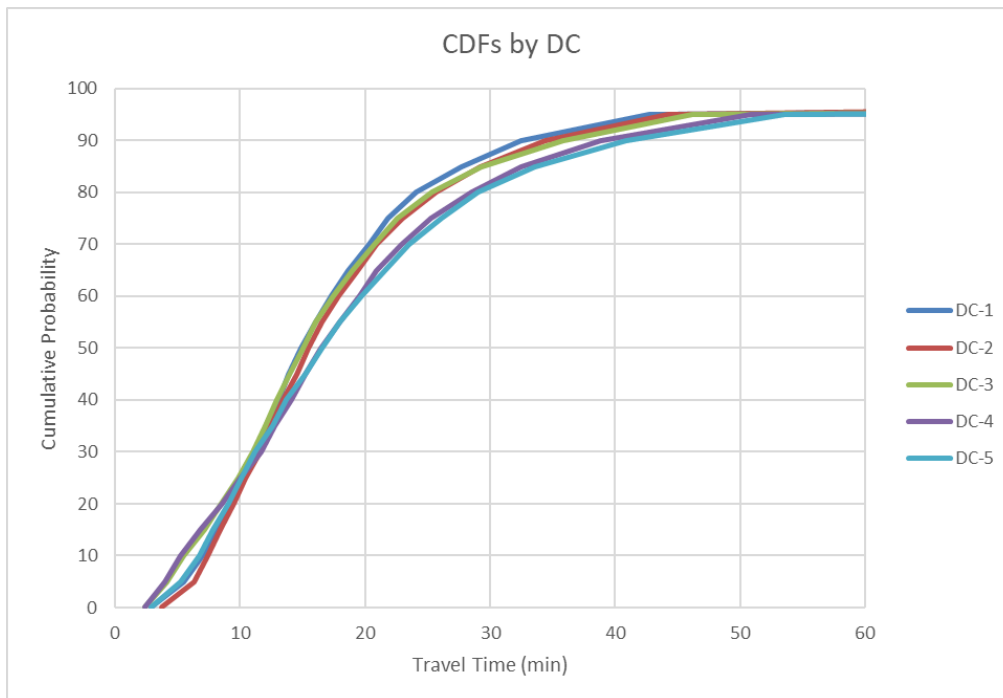
Figure 5.3: Travel time CDFs to all customers

It is interesting that the CDFs are very similar. DC-1 appears to provide the most reliable travel times, but its performance is nearly matched by DC-2 and DC-3. DCs 4 and 5 have poorer performance especially for the likelihood of longer travel times. But DC-4 does have the shortest overall travel times.

However, travel time reliability is not the only important metric to consider, Figure 5.4 presents the bi-objective performance of the DCs in terms of average and maximum travel times to the customer sites. It addresses the question of reliability in the sense of focusing on the average and maximum travel times, in combination. Here it is clear that DCs 1 and 2 are "the best" choices, dominating the other three in terms of combinations of average and maximum travel times. And if the tradeoff between these two performance measures is known, then the best DC can be selected. If minimizing the maximum travel time is relatively important, then DC-2 is the best choice. If the average travel time is more important, then DC-1 is best. Under no conditions (reflected in the average and maximum travel times) are DCs 3, 4, and 5 good solutions. They are dominated by DCs 1 and 2.
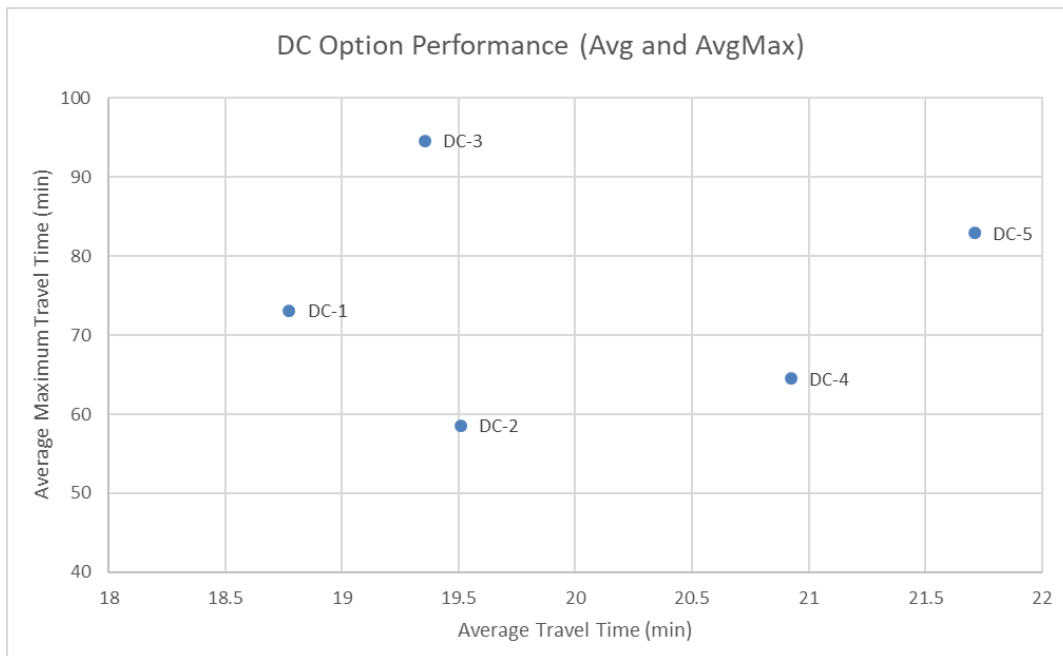
Figure 5.4: DC option performance in terms of average and maximum travel times

Figure 5.5 considers reliability from a slightly different perspective, focusing on the tradeoff between average travel time and the standard deviation of the travel times. Here again, DC candidates 1 and 2 are the best choices, dominating candidates 3, 4, and 5. DC-1 has an average travel time which is significantly shorter than that for DC-2 and the standard deviations are nearly identical. Hence, it might be, that for nearly all preferences between these two metrics, DC-1 would be the best choice. The importance of the standard deviation would have to be very high to allow DC-2 to be preferable to DC-1.
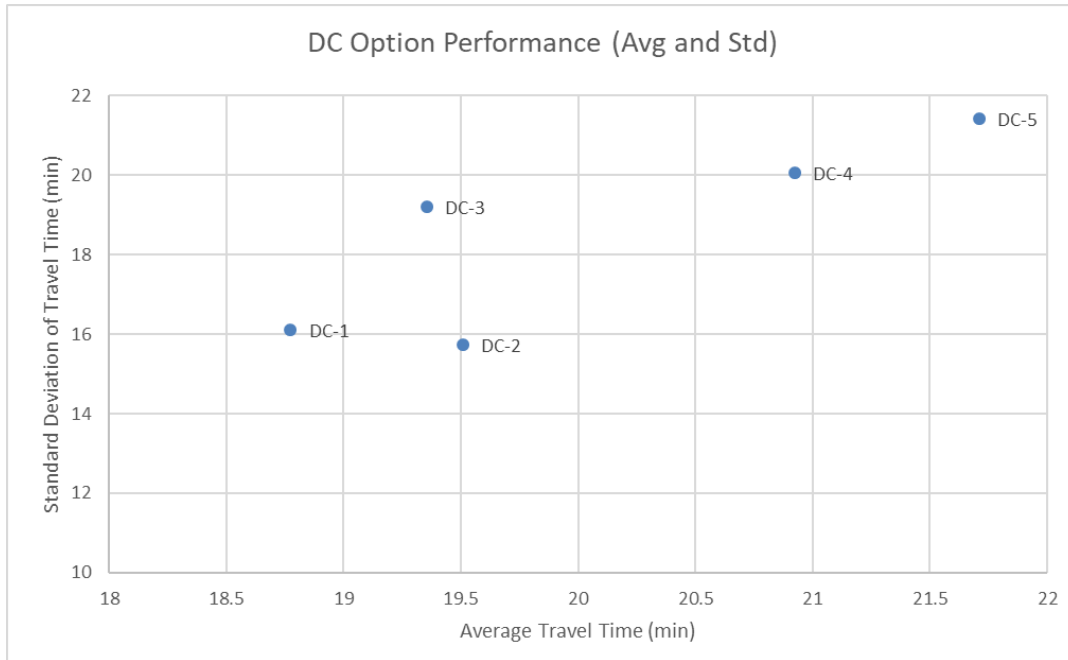
Figure 5.5: Average travel time versus the standard deviation of the travel times

As shown in this example problem, the question facing siting analysis z would be, based on these two assessments of performance, which DC is best. It could be that DC-1 is the better choice despite the larger maximum travel time it has. The important point here is not which choice is better, but that the methodology provides a mechanism for arriving at that conclusion.

5.5   SUMMARY

This section has focused on assessing the reliability performance of distribution centers. Building such facilities is a major investment decision. Hence, it makes sense to choose locations where the travel times to customers will be reliable. There are tradeoffs against other performance metrics like the average travel time and cost. Consequently, having a methodology that allows assessments to be performed is important. One method for doing that has been presented.

6.0   SUMMARY AND FUTURE WORK

6.1   SUMMARY

This research report has focused on advancing the methodological frontier in the analysis of efficiency and reliability for freight transportation. It deals primarily with truck-related shipments although the tools are applicable to other modes and multi-modal systems. The topic is important because of the economic value that results from minimizing the resource consumption associated with freight activity. Unreliable transport raises costs and diverts scarce factors of production away from other, critically important societal activities. It interferes with the efficiency of the supply chain and increases both monetary and time-related costs and resource requirements (e.g., increased in-process inventory, extra trucks).

The freight industry continues to be concerned with reliability. To be competitive, companies need to remove inefficiencies in their production functions. Both late and early shipments are included in the set of problematic inefficiencies facing freight companies. The industry's emphasis on just-in-time manufacturing has squeezed buffer stock out of the logistics supply chain. It has also raised the risk of stock-outs. Because storage space is reduced as well, early arrivals are problematic. If reliability suffers, all participants in the supply chain must make extra asset investments to buffer the process and ensure that delivery schedules are met. From a societal perspective, the cost of producing the goods and services increases. Extra scarce resources must be devoted to freight-related activities to make the economic system work.

Several insights have been derived from this effort. They will have an impact on the way in which freight reliability analyses are performed in the future.

On-Time Windows. Shippers and receivers have expectations about when shipments are going to depart and when they are going to arrive. People see similar on-time windows when they travel by commercial carriers in the context of published departure and arrival times. They may not be aware of the details, but the carriers measure their on-time performance based on whether the vehicles (trains, planes, buses) depart and arrive at times that are consistent with the published timetable. The same is true for freight. Shippers perceive that packages have left on-time if they are picked-up by the carrier within a specific window. The window might be wide (a couple of days) or narrow (less than an hour), but a window exists. The same is true for the receiver. There are expectations that shipments will arrive at specific times. Or more precisely, within given windows.

Hence freight reliability is ***not*** about travel times per se, or even the variance in those travel times. Rather, it is about whether shipments arrive and/or depart during these on-time windows. In other words, reliability is assessed not based on travel time distributions but rather whether the arrival or departure was on time or not. That is, reliability is the probability of arriving (or departing or both) during the on-time window.

However, minimizing the variance in the travel times is still an important thought. But there may be no value in minimizing the variance beyond a certain value once the desired on-time performance is achieved. In effect, the better thought is to control the shape of the travel time distribution, either viewing it as a PDF, or better yet as a CDF, so that a sufficient percentage of the distribution lies within the AW or DW, or both.

Arrival Times. For freight, it is always the arrival times and frequently the departure times that matter. This is different from most personal trip-based analyses where the focus is on the reliability of travel times based on a departure time. The focus of fright movements is on delivering or picking up packages on-time. And thus, the question becomes: when must the truck leave the depot so that the shipment will be delivered on-time? It is ***not*** when it will arrive given a departure time, as is often the case.

Searches Backward in Time. An implication of the focus on arrival times is that the searches for best paths and departure times must often progress *backwards* in time, not forwards. This means they must often be performed based on projections about travel rates that will pertain in the future when the truck will traverse the highway network. Estimates of future travel times, based on past performance, become extremely important. Path search algorithms will need to make assumptions about what the network operating conditions will be in the future and work backwards (in time and space) to determine what path should be employed and when to depart.

Doubly-Constrained Path Choices. Scheduled carriers, like trucking firms, often face doubly-constrained path choice decisions. On-time performance is measured in terms of both departure and arrival events, separately and in combination. Shipments (and vehicle moves) are deemed to be "on-time" if they both depart during the DWs and arrive within the AWs. A joint density function can be used to track this performance. The objective is to find paths and vehicle tours that conjunctively maximize on-time performance. This means that the path choices are doubly, not singly constrained. On-time performance in terms of both in terms of departure and arrival events. Shipments (and vehicle moves) are deemed to be "on-time" if they both depart during the DWs and arrive within the AWs. A joint density function needs to be used to track this performance.

Measurement Locations are Critically Important. Timestamps collected at network nodes (intersections or interchanges) tend to be ambiguous. Unless the sensing distances are very short, it is unclear where the vehicle is when the timestamp is obtained. Moreover, unless the vehicle's report their paths, it is not possible to tell what directional movement the vehicle was executing. Consequently, since this is true at both the upstream and downstream nodes, the travel times computed from pairs of sensors include unknown variability due to turning movements at both the upstream and downstream intersections. Especially for truck trips, there are nodal delays. Hence, it is not wise to place the measuring locations at the physical network nodes.

Collecting timestamps at the midpoints of the links is much better. Two reasons exist for this. The first is that these midpoints are not locations where processing takes place and/or delays occur. Vehicles are typically moving when they pass these locations, so a clear and meaningful timestamp can be collected. The second is that all pairwise combinations of these adjacent timestamps are then related to vehicles that have followed the same intervening path. And because of that, they are very likely to have seen the same processing.

Estimating Travel Time and Rate Distributions. Although the focus is on on-time performance, there is still a critical need to compute the distributions of travel times (and travel rates) of network segments and paths. Section 2 presented three methods for doing this. The first method used average travel times and assumptions about how the truck trip travel times are related to these averages. Monte Carlo sampling was used to sample values from the average and then samples from the hypothesized truck travel time distribution to develop the estimate of the actual truck travel time distribution. This process is the least demanding in terms of data, but the most dependent on inference. It is easy to apply for a given operating condition, but the quality of its estimated distributions is highly dependent upon the insights of the analyst.

The second method used proportional sampling from three hypothetical distributions derived from segment-level travel time distributions: positively correlated, negatively correlated, and uncorrelated. Coefficients α, β, and γ indicate the relative percent to which the three distributions should be sampled to synthesize the overall route travel time. Empirical analyses suggest that the value of α (for the positively correlated distribution) is high and predominant when the segments are uncongested; the value of β (for the negatively correlated distribution) is high when an oscillating pattern in the travel times exists (short travel times on one segment followed by long travel times on the next, as sometimes occurs on signalized arterials), and the value of γ (for the uncorrelated distribution) tends to be high and predominant when the segments are operating at or near capacity. The process is simple and straightforward and appears to yield distributions that very closely match the ones observed.

The third method used segment-specific Monte Carlo sampling to synthesize route-level travel time distributions. The method is intuitively appealing because it capitalizes on ideas about how individual vehicle travel times arise on congested and uncongested networks. The method's main assumption is that a vehicle's travel time arises from three behavioral properties. The first is that when vehicles are traversing segments in an uncongested state, the travel time they achieve reflects driving behavior. The second is that when vehicles are traversing congested segments, the travel time is randomly determined. It is less reflective of driving behavior. The third is that a mix of these conditions pertains to vehicles on a given segment. That is, the segment travel time distribution is a blend of travel times derived from distributions for the two states. Even though the segment may be labeled uncongested, some vehicle travel times can come from the congested distribution. And although a segment may be labeled congested, some travel times can come from the uncongested distribution.

Vehicle Routing and Scheduling Can Address Reliability. Section 4 presented two methods for considering reliability in developing solutions to vehicle routing and scheduling problems.

The first method used a simulation-based heuristic to search for good assignments of customer visits to trucks. It starts from an initial Clarke-Wright assigment of visits to trucks and then modifies that solution through merge, insertion, and 2-interchange reduction analyses. Unique to this method is the fact that time windows are set at specified relative times along a vehicle route provided by a lookup table of previously simulated ordered customer pairs; in particular, given percentile values for travel and service simulated distributions, along with this corresponds to supplier-side control of a supply chain. It often results in at least a 23% cost reduction for 10 or more customers.

The second method provided a formulation of a bi-objective problem that involves maximizing the on-time performance and minimizing the cost of the service provided. This is subject to: picking up all shipments and/or delivering all shipments, not exceeding the capacity of any vehicle, and not exceeding a maximum tour duration for any vehicle. It assumes the fleet size is fixed, the vehicles are always available (that is, maintenance spares exist), and there is a single depot where trucks originate and terminate their tours. It also assumes both pick-ups and deliveries can be made by the same vehicle during a given tour, but delivered shipments must originate at the depot and picked-up shipments must be carried to the depot. Hence, the depot must be one end of the trip for each shipment. It assumes the total time for each tour is the sum of the random variables that describe the travel times between stops and the amount of time spent in pick-up or delivery. This means the travel times are stochastic as are the pick-up and delivery times. It conducts the evaluation by simulating the system's performance multiple times. It samples random variables to establish each realization; assigns loads to the trucks and develop the tours; sequences the loads based on the beginning of their windows; and for each stop, selects the truck that can arrive earliest and has the least delay. The reliability is improved by adding slack time to the schedule; that is, by arriving early, off-site, near a customer's location, at the cost of adding time to the tour and potentially increasing the fleet size. It is also improved by increasing the fleet size. A larger fleet reduces the number of customer stops per vehicle and adds more resource availability.

Site Choice is Critically Important. Section 5 presented a method for considering the reliability of different sites for distribution centers (DCs). Technique employed uses Monte Carlo simulation to assess the reliability of the delivery service quality provided by candidate distribution center sites and then it identifies the best ones to choose. It provides useful and meaningful results which are easy to understand.

A hypothetical case study analysis showed that differences in travel time reliability can exist among candidate sites and hence, this aspect of DC choice should be incorporated into a multi-objective assessment of potential sites. The example portrayed this thought in the context of a bi-objective assessment in which the average and maximum travel times to the customer sites is considered. It is clear from this analysis that some DCs are better choices than others in that they dominate poorer performing sites in terms of the combinations of average and maximum travel times that the best sites provide. Moreover, if the tradeoff between these two performance measures is known, then the best DC can be selected.

6.2   FUTURE WORK

Much future work can be carried out based on the analyses conducted so far. Some important examples of these efforts are as follows:

Real-World Tests. As is often the case, the methodological advances presented here have been tested using a blend of empirical data and hypothetical situations. One natural extension for future work is to test these methods based on datasets that are more representative and reflective of real-world conditions. This pertains to all of the methods presented, from the assessment of reliability for segments and routes to the selection of plans for vehicle routing and locations for distribution centers.

On-Time Windows. Another opportunity for future work is the further examination of on-time windows and the implication this idea has for freight reliability assessment. Surveys of shippers, receivers, and carriers would be helpful to double-check that on-time windows are, indeed, the current way that reliability performance is assessed. Such surveys could also indicate the way in penalties are assessed for degradations in on-time performance and how expected performance is specified contractually. The knowledge gained would further enhance the proposed methods.

Backward Searches. The report asserts that the search for best paths and departure times is one that involves an analysis backward in time. The veracity of this assertion could be checked through a survey of shippers, receivers, and carriers. Also, assuming it is correct, a study of the planning and scheduling practices of these three stakeholders would show how this assessment is presently carried out, what assumptions are made, and what data are employed. If this perspective is not embraced, then such surveys would help improve the understanding of how on-time performance and/or reliable service is planned and scheduled for when on-time windows are part of the overall picture.

Doubly-Constrained Path Choices. There is an assertion that path choices for truck-based freight shipments are doubly-constrained, having on-time windows for both departure and arrival. Inquiries with carriers, shippers, and receivers would provide indications as to whether this assertion was true or not; or the extent to which it is true. Assuming it is an issue and major constraint, such inquiries would also provide indications of how carriers choose paths and departure times to ensure that these doubly-constrained solutions are obtained. For example, it could be seen if is there a conscious effort to add slack time to the trips to ensure that arrival windows are achieved.

Travel Time and Rate Distributions. Three methods for estimating route-level travel time and travel rate distributions have been presented. Analysis based on more real-world data, especially from trucking firms, would indicate how well these methods work and how they can be enhanced. Especially for the third method, which has been evaluated entirely based on hypothetical data, such an analysis would be very fruitful. This ability to synthesize travel time distributions for paths that lack direct observations is a critical need, especially for planning activities. Having a method which works reliably under a wide variety of conditions would be very helpful. Also, being able to see how the correlations among travel time distributions can and should be addressed is very important.

Vehicle Routing and Scheduling. Vehicle routing and scheduling will continue to be a critical element of the reliability analysis. This report presented two methods for developing a vehicle routing and scheduling plan that maximizes the reliability of the service provided.

The meta-heuristic method could be further improved in several ways. One of which is to incorporate a Bayesian statistics framework in which times are estimated via prior distributions (not necessarily lognormal as used in this paper). Ostensibly this would allow for any given distribution to be used to set relative time windows and test routes. Notably, a business with repetitive deliveries could benefit from using the same routing for a given period of time to obtain new samples and update the routing accordingly. Additionally, with customer penalty assessment

data, customer-specific penalties could be introduced. This would likely result in higher late penalties for larger and/or more demanding clients due to phenomena such as customer loyalty and product obsolescence. Lastly, it would be worth investigating some of the natural extensions to the uncapacitated VRP, such as putting capacities on trucks or implementing precedence constraints as in pickup-and-delivery customers.

The building-block method could be enhanced by adding a search procedure to the current heuristic. Presently, it sorts the customer visits into ascending order by the beginning of the on-time arrival window and assigns the earliest available truck to each visit in sequence. The production of sub-optimal solutions has been demonstrated preliminarily by taking specific realizations of the problem settings and finding solutions by both the heuristic and explicit optimization using a mixed-integer linear programming (MILP) representation of the problem. The MILP is almost always able to find a better solution than does the heuristic. (However, the MILP formulation can only be applied to small-scale problems.) One way to improve the solution provided by the heuristic is to link it to a search procedure (e.g., a genetic algorithm, tabu search, or simulated annealing).

Site Selection. The site selection procedure could be enhanced most significantly by linking it to vehicle routing and scheduling. That is, treat the assessment of the reliability of the sites as an evaluation of the best reliability that can be provided by a fleet of trucks given the choice of a specific site. This means solving the SVRP for each of the sites and then comparing the results.

Combinations. All these possible extensions to the current work could be done separately or in combinations. For example, the SVRP ideas could be explored in conjunction with the enhancements to the site selection procedures. The enhancements to path characterization could be coupled with SVRP so that the SVRP solutions are more tightly and defensibly tied to the characterization of the performance of the network over which the trips take place. Determining which combinations to select and which options to pursue will depend upon the data available, the interests of the research funding agencies that are involved, and the demands of the stakeholders whose needs are being addressed.

REFERENCES

Abkowitz, M., Josef, R., Tozzi, J., & Driscoll, M. K. (1987). Operational Feasibility Of Timed Transfer In Transit Systems. *Journal of Transportation Engineering, 113*(2), pp. 168-177.

Administration, U. S. M. (2007). The Maritime Administration and the U.S. Marine Transportation System: A Vision for the 21st Century, pp. 48.

Agin, N. I., & Cullen, D. E. (1975). Algorithm For Transportation Routing And Vehicle Loading. *Logistics*, pp. 1-20.

Alexander, N. J. B. (1978). Railways And Their Interaction With Ports. *Cargo Processing Abstracts*(3), pp. 5.

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, 19B*(5), pp. 445-447.

Anderson, C. M., Opaluch, J. J., & Grigalunas, T. A. (2009). The Demand for Import Services at US Container Ports. *Maritime Economics & Logistics, 11*(2), pp. 156-185.

Anderson, J. E., & Kraus, M. (2014). Quality of Service and the Demand for Air Travel. *63*(4), pp. 533-540.

Ando, N., & Taniguchi, E. (2006). Travel Time Reliability in Vehicle Routing and Scheduling with Time Windows. Networks and Spatial Economics, 6(3–4), pp. 293–311.

Androutsopoulos, K. N., & Zografos, K. G. (2012). A bi-objective time-dependent vehicle routing and scheduling problem for hazardous materials distribution. *EURO Journal on Transportation and Logistics, 1*(1-2), pp. 157-183.

Arcot, V. C. (2007). *Modeling uncertainty in rail freight operations: Implications for service reliability*: ProQuest.

Armacost, A. P., Barnhart, C., & Ware, K. A. (2002). Composite variable formulations for express shipment service network design. *INFORMS*, *36*(1), pp. 1-20.

Armacost, A. P., Barnhart, C., Ware, K. A., & Wilson, A. M. (2004). Ups optimizes its air network. *INFORMS*, *34*(1), pp. 15-25.

Arunapuram, S., K. Mathur, and D. Solow. (2003). "Vehicle Routing and Scheduling with Full Truckloads". *Transportation Science,* 37 (2), pp. 170–182.

Arvis, J.-F., Raballand, G., & Marteau, J.-F. (2007). The Cost of Being Landlocked: Logistics Costs and Supply Chain Reliability, pp. 81.

Assad, A. A. (1981). Analytical Models in Rail Transportation - An Annotated Bibliography. *Infor, 19*(1), pp. 59-80.

Assad, A A. (1988)"Modeling and Implementation Issues in Vehicle Routing", in B.L. Golden and A.A. Assad (Eds.), *Vehicle Routing: Methods and Studies,* Elsevier Science Publishers (North-Holland), Amsterdam, The Netherlands, pp. 7-45.

Baker, E. K., & Rushinek, S. F. (1982). Large Scale Implementation of a Time Oriented Vehicle Scheduling Model, pp. 115.

Ball, M. O., Bodin, L. D., & Greenberg, J. (1985). Enhancements to the RUCUS-II crew scheduling system. In J. M. Rousseau (Ed.), Proceedings of the Second International Workshop on Computer-Aided Scheduling of Public Transport (pp. 279–293). North-Holland.

Ballou, R. H., Agarwal, Y. K., Koskosidis, Y. A., & Powell, W. B. (1990). A Performance Comparison of Several Popular Algorithms for Vehicle Routing and Scheduling. *Journal of Business Logistics, 11*(1), pp. 101-128.

Barber, D., & Grobar, L. (2001). Implementing a Statewide Goods Movement Strategy and Performance Measurement of Goods Movement in California. pp. 48.

Barcelo, J., Grzybowska, H., & Orozco, J. A. (2008). *A Simulation Based Decision Support System for City Logistics Applications*.

Bardi, E. J., Bagchi, P. K., & Raghunathan, T. (1989). Motor carrier selection in a deregulated environment. *Transportation Journal*, pp. 4-11.

Barnhart, C., Belobaba, P., & Odoni, A. R. (2003). Applications of Operations Research in the Air Transport Industry. *Transportation Science, 37*(4), pp. 368-391.

Beattie, D. (1989). Should dangerous goods be moved by rail rather than by road? Reliability on the Move: Safety and Relability in Transportation. Proceedings of the Safety and Reliability Society Symposium. Bath, 11-12 October 1989. *Elsevier Applied Science Publishers Limited*, pp. 149-159.

Bechhofer, R., Brown, M., Dafermos, S., Eisner, M., Little Feb, J., Nemhauser, G., . . . Prabhu, U. (1969). Analytical Methodology and Optimal Control in Urban Traffic Networks. Vol. 1-4, pp. 287.

Bell, M. G. H. (2004). Games, heuristics, and risk averseness in vehicle routing problems. *Journal of Urban Planning and Development, 130*(1), pp. 37-41.

Belovarac, K., & Kneafsey, J. T. (1972). Determinants of Line Haul Reliability. Studies in Railroad Operations and Economics, Vol. 3, pp. 97.

Bertsimas, D., Chervi, P., Peterson, M., Golden, B. L., Laporte, G., Taillard, E. D., . . . Badeau, P. (1995). Computational Approaches to Stochastic Vehicle Routing Problems. Institute for Operations Research and the Management Sciences.

Bevilacqua, O. M. (1978). Energy conservation and intercity freight transport. *Transportation Planning and Technology*, *4*(4), pp. 227-240.

Bodin, L., Golden, B., Assad, A., & Ball, M. (1981). The State of Art in the Routing and Scheduling of Vehicles and Crews, pp. 413.

Bodin, L. D., & Jursh, S. J. (1978). A Computer-Assisted System for Routing and Scheduling of Street Sweepers. *Operations Research, 26*(4), pp. 525-537.

Bone, I., Wallis, I., O'Fallon, C., & Nicholson, A. (2013). *Reliability and freight: literature and practice review*.

Bontekoning, Y. M., & Priemus, H. (2004). Breakthrough Innovations in Intermodal Freight Transport. *Transportation Planning and Technology, 27*(5), pp. 335-345.

Boysen, N., Fliedner, M., Jaehn, F., & Pesch, E. (2013). A Survey on Container Processing in Railway Yards. *Transportation Science, 47*(3), pp. 312-329.

Brogan, J. (2007). Integrating Freight into Transportation Planning and Project-Selection Processes.

Brooks, M. R., Puckett, S. M., Hensher, D. A., & Sammons, A. (2012). Understanding Mode Choice Decisions: A Study of Australian Freight Shippers. *Maritime Economics & Logistics, 14*(3), pp. 274-299.

Bruck, H. (1969). CARS (Computer Aided Routing System) A Prototype Dial-A-Bus System.

Buchan, A. B., List, G. F., Bongaardt, J. H. L., & Pipas, G. (1981). Expansion of the Use of Freight Car Scheduling: Development of a Strategic Plan for New England. pp. 113.

Burt, Jr., J. M., & Garman, M. B. (1971). Conditional monte carlo: A simulation technique for stochastic network analysis. *INFORMS*, *18*(3), pp. 207-217.

Buxey, G. M. (1979). The Vehicle Scheduling Problem and Monte Carlo Simulation. *Journal of the Operational Research Society, 30*(6), pp. 563-573.

Caccetta, L., Alameen, M., & Abdul-Niby, M. (2013). An improved clarke and wright algorithm to solve the capacitated vehicle routing problem. Engineering, Technology & Applied Science Research, *3*(2), pp. 413.

Caldwell, H., & Sedor, J. (2002). The Freight Story: A National Perspective on Enhancing Freight Transportaiton. pp. 38.

Cambridge Systematics, I. (2004). Traffic Congestion and Reliability: Linking Solutions to Problems, pp. 104.

Cambridge Systematics, I. (2013a). Incorporating Reliability Performance Measures into the Transportation Planning, Programming Processes.

Cambridge Systematics, I. (2013b). Waterborne Freight Transportation: Bottom Line Report. pp. 114.

Campbell, A. M., & Savelsbergh, M. (2004). Efficient Insertion Heuristics for Vehicle Routing and Scheduling Problems. *Transportation Science, 38*(3), pp. 369-378.

Canadian National Railways. (1976). SIMTRAC. *Canadian National Railways*.

Casavant, K. (2013). Developing a Performance Measurement Approach to Benefit/Cost Freight Project Prioritization.

Cassidy, P. J., & Bennett, H. S. (1972). Tramp-A-Multi-Depot Vehicle Scheduling System. *Operational Research Quarterly /UK/, 23*(2), pp. 151-163.

Chen, B. Y., Lam, W. H. K., Sumalee, A., Li, Q., Shao, H., & Fang, Z. (2013). Finding Reliable Shortest Paths in Road Networks Under Uncertainty. Networks and Spatial Economics, *13*(2), pp. 123–148.

Chen, B. Y., Lam, W. H. K., Sumalee, A., & Shao, H. (2011). An efficient solution algorithm for solving multi-class reliability-based traffic assignment problem. Mathematical and Computer Modelling, *54*(5), pp. 1428–1439. https://doi.org/10.1016/j.mcm.2011.04.015.

Chen, C.-C., & Schonfeld, P. (2010). Modeling and Performance Assessment of Intermodal Transfers at Cargo Terminals. *Transportation Research Record: Journal of the Transportation Research Board, 2162*(-1), pp. 53-62. doi: 10.3141/2162-07.

Chen, C.-C., & Schonfeld, P. (2011). Alleviating Schedule Disruptions at Intermodal Freight Transfer Terminals: Real-Time Dispatching Control. *Transportation Research Record: Journal of the Transportation Research Board*, *2238*, pp. 32-43.

Chen, C. F. (2010). Intermodal Transfer Coordination in Logistic Networks. *PhD Thesis, University of Maryland*.

Chu, H.-C. (2010, 2010). *Estimating Travel Time Reliability on Freight Corridors*.

Clarke, G., & Wright, J. W. (1964). Scheduling of vehicles from a central depot to a number of delivery points. *Operations Research, 12*(4), pp. 568-581.

Clinkenbeard, G. R. (1976). Terminal service evaluation. *St Louis - San Francisco Railway Company*.

Coelho, L. C., Cordeau, J.-F., & Laporte, G. (2014). Thirty Years of Inventory Routing. *Transportation Science, 48*(1), pp. 1-19.

Consulting, I., & Decision-Economics, H. (2002). Economics Effects of Transportation: The Freight Story. pp. 24.

Cook, P. D., & Kent, P. (2007). *Measuring the Performance of Transport Logistics Corridors in Developing Countries: The FastPath Approach*.

Cordeau, J.-F. (2006). A branch-and-cut algorithm for the dial-a-ride problem. *Operations Research, 54*(3), pp. 573-586.

Cordeau, J.-F., & Laporte, G. (2003). A tabu search heuristic for the static multi-vehicle dial-a-ride problem. *Transportation Research Part B: Methodological, 37*(6), pp. 579-594. doi: 10.1016/s0191-2615(02)00045-0.

Crainic, T. G., & Rousseau, J.-M. (1986). Multicommodity, multimode freight transportation: A general modeling and algorithmic framework for the service network design problem. *Transportation Research Part B: Methodological, 20*(3), pp. 225-242.

Crum, M. R., & Allen, B. J. (1997). A longitudinal assessment of motor carrier-shipper relationship trends, 1990 vs. 1996. *Transportation Journal*, pp. 5-17.

Czuch, C., & McDaniel, Z. (2011). *Measuring Goods Movement Travel Time and Reliability: Proposed Methodology*.

D'Acierno, L., Gallo, M., & Montella, B. (2011). A Fixed-point Model and Solution Algorithms for Simulating Urban Freight Distribution in a Multimodal Context. *Journal of Applied Sciences, 11*(4), pp. 647-654. doi: 10.3923/jas.2011.647.654.

Daganzo, C. F. (1986). Static blocking at railyards: Sorting implications and track requirements. *Transportation Science, 20*(3), pp. 189-199.

Daganzo, C. F., Dowling, R. G., & Hall, R. W. (1983). Railroad classification yard throughput: The case of multistage triangular sorting. *Transportation Research Part A: General, 17*(2), pp. 95-106.

Dai, M. D. M., & Schonfeld, P. (1991). Simulation of Waterway Transportation Reliability. *Transportation Research Record. 1313*. pp. 98-105.

Danielis, R., Marcucci, E., & Rotaris, L. (2005). Logistics Managers' Stated Preferences for Freight Service Attributes. *Transportation Research Part E: Logistics and Transportation Review, 41*(3), pp. 201-215.

Dantzig, G. B., & Ramser, J. H. (1959). The truck dispatching problem. *Management science, 6*(1), pp. 80-91.

Daskin, M. S. (1985). *Logistics: an overview of the state of the art and perspectives on future research*, UK.

de Jong, G., Kouwenhoven, M., Kroes, E., Rietveld, P., & Warffemius, P. (2009). Preliminary monetary values for the reliability of travel times in freight transport. *European Journal of Transport and Infastructure Research, 9*(2), pp. 83-99.

Dejax, P. J., & Bookbinder, J. H. (1991). Goods Transportation by the French National Railway SNCF: The Measurement and Marketing of Reliability. *Transportation Research Part A: Policy and Practice, 25A*(4), pp. 219-225.

Deleuw, C., & Company. (1971). Tri-City Transportation Needs Study: Analysis of Flexible Transit Service Conceprs and Plans.

Dessouky, M., Ioannou, P., & Jula, H. (2004). A Novel Approach to Routing and Dispatching Trucks Based on Partial Information in a Dynamic Environment, pp. 49.

Detmold, P. J. (1972). Some Thoughts on the Reliability of Railroading, *13*.

Dominicis, R. D., Russo, F., & Quattrone, A. (2012). *National freight multimodal transport system: the Italian project for the ITS integration*.

Dowling, R., List, G., Yang, B., Witzke, E., & Flannery, A. (2014). *NCHRP Report 31: Incorporating Truck Analysis into the Highway Capacity Manual.* Transportation Research Board, the National Academies, Washington, DC.

Duhamel, C., Gendreau, M., Laporte, G., & Potvin, J.-Y. (1994). Heuristic Approaches to the Vehicle Routing Problem with Backhauls and Time Windows.

Ebeling, C. E. (1997). An introduction to reliability and maintainability engineering, 1997: McGraw-Hill, New York.

Eisele, W. L., Lomax, T., Schrank, D. L., & Turner, S. M. (2014). *Lessons Learned for Transportation Agencies Preparing for MAP-21 Performance Management Requirements Related to Mobility and Reliability*.

Eisele, W. L., & Schrank, D. L. (2010). Conceptual Framework and Trucking Application for Estimating Impact of Congestion on Freight. *Transportation Research Record: Journal of the Transportation Research Board, 2168*(1), pp. 94-103.

Elefteriadou, L., & Cui, X. (2007). Travel time reliability and truck level of service on the strategic intermodal system part A: Travel time reliability.

Fazel-Zarandi, M. M., Berman, O., & Beck, J. C. (2013). Solving a stochastic facility location/fleet management problem with logic-based Benders' decomposition. *IIE Transactions (Institute of Industrial Engineers), 45*(8), pp. 896-911. doi: 10.1080/0740817X.2012.705452.

Feo-Valero, M., García-Menéndez, L., & Garrido-Hidalgo, R. (2011). Valuing Freight Transport Time using Transport Demand Modelling: A Bibliographical Review. *Transport Reviews, 31*(5), pp. 625-651. doi: 10.1080/01441647.2011.564330.

Ferreira, L., & Bunker, J. (2003). The Role of Governments in Improving Freight Logistics in Queensland.

Field, J. F. (1976). *Rucus: an overview implementation and operation of a run cutting and scheduling system for buses*.

Fielding, G. J. (1977). Shared-Ride Taxi Computer Control System Requirements Study, pp. 53.

Figliozzi, M. (2011). Freight Distribution Problems in Congested Urban Areas: Fast and Effective Solution Procedures to Time-Dependent Vehicle Routing Problems, pp. 60.

Figliozzi, M. (2014). Analysis of Travel Time Reliability for Freight Corridors Connecting the Pacific Northwest.

Figliozzi, M., & Zhang, Z. (2010). *A Study of Transportation Disruption Causes and Costs*.

Figliozzi, M. a., Wheeler, N., Albright, E., Walker, L., Sarkar, S., & Rice, D. (2011). Algorithms for Studying the Impact of Travel Time Reliability Along Multisegment Trucking Freight Corridors. *Transportation Research Record: Journal of the Transportation Research Board, 2224*(1), pp. 26-34. doi: 10.3141/2224-04.

Folk, J. F. (1972). Models for Investigating Rail Trip Time Reliability. *Studies in Railroad Operations and Economics, 5*, pp. 213.

Fosgerau, M., & Karlström, A. (2010). The value of reliability. *Transportation Research Part B: Methodological, 44*(1), pp. 38-49. doi: 10.1016/j.trb.2009.05.002.

Fowkes, T. (2007). The design and interpretation of freight stated preference experiments seeking to elicit behavioural valuations of journey attributes. *Transportation Research Part B: Methodological, 41*(9), pp. 966-980.

Fowkes, T., & Whiteing, T. (2006). The Value of Freight Travel Time Savings and Reliability Improvements - Recent Evidence from Great Britain. pp. 1-14.

Fries, N. (2008). Modal split functions for a Swiss national freight transport model. *Proceedings from European Transport Conference 2008*, pp. 17.

Fu, L., & Rilett, L. R. (2000). Estimation of Time Dependent, Stochastic Route Travel Times Using Artifical Neural Networks. *Transportation Planning and Technology, 24*(1), pp. 25-48.

Fusaro, M. (2015). Public service and private trade: Northern seamen in seventeenth-century Venetian courts of justice. *International Journal of Maritime History*, 27(1), pp. 3–25.

Garcia, B. L., & Arunapuram, S. (1993). Development of Techniques for Tabu Research for the Vehicle Routing Proble with Time Windows.

Gendreau, M., Laporte, G., & Solomon, M. (1991). Sequence Dependent Routing or Scheduling to Minimize the Number of Missed Deadlines- Revised Edition, pp. 20.

Ghoseiri, K., Ghannadpour, S. F., & Seifi, A. (2010). *Locomotive Routing and Scheduling Problem with Fuzzy Time Windows*.

Goeddel, D. L. (1975). An examination of the Run Cutting and Scheduling (RUCUS) system - A case analysis. *In Automated Techniques for Scheduling of Vehicle Operaors for Urban Public Transportation Services.* Palmer House, Chicago.

Golroo, A., Shariat Mohaymany, A., & Mesbah, M. (2010). *Reliability Based Investment Prioritization in Transportation Networks*.

Gong, Q., Miao, Q., Wang, B. X., & Adams, T. M. (2012). Assessing Public Benefits and Costs of Freight Transportation Projects: Measuring Shippers' Value of Delay on the Freight System, pp. 62.

Grosso, M., & Monteiro, F. (2008). Relevant strategic criteria when choosing a container port? The case of the Port of Genoa. *Proceeding from European Transport Conference 2008,* pp. 21.

Hansson, S. O. (1994). Decision theory. *A Brief Introduction. Department of Philosophy and the History of Tecnology, Royal Institute of Technology, Stockholm*.

Hayya, J. C., Ramasesh, R. V., Tyworth, J. G., Kim, J. G., & Sun, D. (2013). JIT delivery with stochastic lead time. *Journal of the Operational Research Society, 64*(1), pp. 97-105. doi: 10.1057/jors.2012.27.

Hejazi, B., & Haghani, A. (2009). *Dynamic Decision Making for Less-Than-Truckload Trucking Operations Using Path-Based Network Partitioning*.

Hensher, D. A., Puckett, S. M., & Rose, J. M. (2007). Agency decision making in freight distribution chains: Establishing a parsimonious empirical framework from alternative behavioural structures. *Transportation Research Part B: Methodological, 41*(9), pp. 924-949.

Herr, P. R. (2013). Department of Transportation: Key Issues and Management Challenges, 2013. pp. 29.

Hinds, D. H. (1979). RUCUS: A Comprehensive Status Report and Assessment. *Transit Journal*, pp. 17-34.

Hnaien, F., Delorme, X., & Dolgui, A. (2008). Supply planning for two-level assembly systems with stochastic component delivery times: trade-off between holding cost and service level. pp. 1-7.

Holstein, L., Krampe, S., & Inninger, W. (2012). *Safe Navigation for Hazardous Goods Transport: Results from the Project SafeNav*.

Howson, L. L., & Heathington, K. W. (1970). Algorithms for Routing and Scheduling in Demand-Responsive Transportaiton Systems. *Highway Research Record*.

Hsu, C.-I., & Wang, C.-C. (2013). Reliability analysis of network design for a hub-and-spoke air cargo network. *International Journal of Logistics Research and Applications, 16*(4), pp. 257-276.

Huang, H., & Gao, S. (2012). Optimal paths in dynamic networks with dependent random link travel times. *Transportation Research Part B: Methodological*, *46*(5), pp. 579–598. https://doi.org/10.1016/j.trb.2012.01.005.

Ichoua, S., Gendreau, M., & Potvin, J.-Y. (2000). Diversion Issues in Real-Time Vehicle Dispatching. *Transportation Science, 34*(4), pp. 426-438.

Iikkanen, P., Rasanen, J., & Touru, T. (2012). Level of service in travel and transport chains, pp. 42. Finnish Transport Agency.

Isukapati, I. K., List, G. F., & Karr, A. F. (2013). *Synthesizing Route Travel Time Distributions From Segment Travel Time Distributions.* Paper presented at the Transportation Research Board 92nd Annual Meeting.

Jabali, O., Leus, R., Woensel, T. V., & Kok, T. de. (2013). Self-imposed time windows in vehicle routing problems. *OR Spectrum*, *37*(2), pp. 331–352. https://doi.org/10.1007/s00291-013-0348-1.

Jane, C.-C. (2011). Performance Evaluation of Logistics Systems Under Cost and Reliability Considerations. *Transportation Research Part E: Logistics and Transportation Review, 47*(2), pp. 130-137.

Ji, Z., Kim, Y. S., & Chen, A. (2011). Multi-objective α-reliable path finding in stochastic networks with correlated link costs: A simulation-based multi-objective genetic algorithm approach (SMOGA). *Expert Systems with Applications*, *38*(3), pp. 1515–1528. https://doi.org/10.1016/j.eswa.2010.07.064.

Johnson, E., & Dupin, C. (2012). Measuring reliability: Mit research shows ocean transit reliability isn't as problematic as landside delays. *American Shipper*, *54*(1), pp. 9-11.

Jones, C., & Sedor, J. (2006). Improving the reliability of freight travel. *Public roads, 70*(1).

Kanturska, U., Trozzi, V., & Bell, M. G. H. (2013). Scheduled Hyperpath: A Strategy for Reliable Routing and Scheduling of Deliveries in Time-Dependent Networks with Random Delays. *Transportation Research Record: Journal of the Transportation Research Board, 2378*, pp. 99-109.

Kaplan, S. M. (2007). Rail Transportation of Coal to Power Plants: Reliability Issues, pp. 93. Congressional Research Service.

Kent, J. L., & Parker, R. S. (1999). International containership carrier selection criteria: shippers/carriers differences. *International Journal of Physical Distribution & Logistics Management, 29*(6), pp. 398-408.

Kenyon, A. S., & Morton, D. P. (2003). Stochastic vehicle routing with random travel times. *Transportation Science*, *37*(1), pp. 69–82.

Kikuchi, S. (1987). Alternative Routing and Scheduling Schemes for Specialized Transportation Vehicles. *Specialized Transportation Planning and Practice, 2*(4), pp. 303-323.

Killough, K. L. (2008, 2008). *Value Analysis of Truck Toll Lanes in Southern California.*

Kim, C., & Simchi-Levi, D. (2012). Optimal Expediting Policies for a Serial Inventory System with Stochastic Lead Time and Visibility of Orders In Transit. pp. 1-50.

Kim, S. (2003). Optimal Vehicle Routing and Scheduling with Real-Time Traffic Information, pp. xi - 111.

Kim, Y., Mahmassani, H. S., & Jaillet, P. (2004). Dynamic Truckload Routing, Scheduling and Load Acceptance. *Transportation Research Record, 1882*, pp. 120-128.

Kittelson, W. (2014). A Framework for Improving Travel Time Reliability.

Kittelson and Associates. (2013). *Evaluating Alternative Operations Strategies to Improve Travel Time Reliability*: Transportation Research Board.

Kokubugata, H., Itoyama, H., & Kawashima, H. (1997). *Vehicle Routing Methods for City Logistics Operations.*

Koskosidis, Y. A., Powell, W. B., & Solomon, M. M. (1992). An Optimatization-Based Heuristic for Vehicle Routing and Scheduling with Soft Time Window Constraints. *Transportation Science, 26*(2), pp. 69-85.

Kraay, D., Harker, P. T., & Chen, B. (1991). Optimal pacing of trains in freight railroads: model formulation and solution. *Operations Research, 39*(1), pp. 82-99.

Kraft, E. R. (1995). The Link Between Demand Variability and Railroad Service Reliability. *Journal of the Transportation Research Forum, 34*(2), pp. 27-43.

Kraft, E. R. (2002). Priority-Based Classification for Improving Connection Reliability in Railroad Yard. Part I of II: Integration with Car Scheduling. *Journal of the Transportation Research Forum, 56*(1), pp. 93-105.

Kraft, E. R. (2002). Scheduling railway freight delivery appointments using a bid price approach. *Transportation Research Part A: Policy and Practice, 36*(2), pp. 145-165.

Kreutzberger, E. D. (2008). The Innovation of Intermodal Rail Freight Bundling Networks in Europe Concepts, Developments, Performances, pp. 363.

Kwon, O. K., Martland, C. D., Sussman, J. M., & Little, P. (1995). Origin-To-Destination Trip Times and Reliability of Rail Freight Services in North American Railroads. *Transportation Research Record*, *1489*, pp. 1-8.

Lai, C. M. (2011). *Pickup and Delivery Problem with Stochastic Travel Times for Semiconductor Supply Chains*.

Lam, T. C., & Small, K. A. (2001). The value of time and reliability: measurement from a value pricing experiment. *Transportation Research Part E: Logistics and Transportation Review, 37*(2), pp. 231-251.

Lang, A. S., & Reid, R. M. (1970). Studies in Railroad Operations and Economics. *Volume 1. Railroad Car Movement Reliability: A Preliminary Study of Line-Haul Operations*, pp. 33.

Langevin, A., & Soumis, F. (1989). Design of Multiple-Vehicle Delivery Tours Satisfying Time Constraints. *Transportation Research Part B: Methodological, 23B*(2), pp. 123-138.

Laporte, G. (1992). The vehicle routing problem: An overview of exact and approximate algorithms. *European Journal of Operational Research*, *59*(3), pp. 345–358. https://doi.org/10.1016/0377-2217(92)90192-C.

Laporte, G., Louveaux, F., & Mercure, H. (1992). The vehicle routing problem with stochastic travel times. *Transportation Science*, *26*(3), pp. 161–170.

Laporte, G., & Osman, I. H. (1995). Routing Problems: A Bibliography.

Lecluyse, C., Van Woensel, T., & Peremans, H. (2009). Vehicle routing with stochastic time-dependent travel times. *4OR*, *7*(4), pp. 363–377. https://doi.org/10.1007/s10288-009-0097-9.

Lee, E.-S., & Song, D.-W. (2011). Knowledge management systems in maritime logistics. pp. 50-67.

Levy, L., & Bodin, L. (1988). Scheduling the Postal Carriers for the United States Postal Service: An Application of Arc Partitioning and Routing. Vehicle Routing: Methods and Studies. Studies in Management Science and Systems - Volume 16, *Publication of: Dalctraf*, pp. 359-394.

Li, Q., Nie, Y., Vallamsundar, S., Lin, J., & Homem-de-Mello, T. (2013). *Incorporating Environmental Measures into a Reliable Freight Routing Model*.

Li, Q., & Savachkin, A. (2013). A heuristic approach to the design of fortified distribution networks. *Transportation Research Part E: Logistics and Transportation Review, 50*, pp. 138-148.

Li, X., Tian, P., & Leung, S. C. H. (2010). Vehicle routing problems with time windows and stochastic travel and service times: Models and algorithm. *International Journal of Production Economics*, *125*(1), pp. 137–145. https://doi.org/10.1016/j.ijpe.2010.01.013.

Li, Z., & Hensher, D. (2012). Accommodating Risk Attitudes in Freight Transport Behaviour Research. *Transport Reviews, 32*(2), 221-239. doi: 10.1080/01441647.2011.645906.

Lin, S.-W., Yu, V. F., & Chou, S.-Y. (2011). A simulated annealing heuristic for the truck and trailer routing problem with time windows. *Expert Systems with Applications, 38*(12), 15244-15252.

List F, G., & Bongaardt, H. L. J. (1981). Freight Car Scheduling Benefit Project on CONRAIL: The Pillsburge to St. Louis Corridor, pp. 107.

List, G. (2014). Establishing Monitoring Programs for Mobility and Travel Time Reliability.

List, G., Mirchandani, P., and Turnquist, M. (1990) Logistics for Hazardous Materials Transportation: Scheduling, Routing and Siting: First Year Interim Report, prepared for the Region II Transportation Research Center Program. https://books.google.com/books/about/Logistics_for_hazardous_materials_transp.html?id=mI0gAQAAMAAJ.

List, G.F., and M.A. Turnquist, Prototype Crew Assignment Module, working paper, including software, prepared for Sandia National Laboratories, December 1993.

List, G. F., Williams, B., Rouphail, N., Hranac, R., Barkley, T., Mai, E., . . . Khattak, A. (2014). *Guide to Establishing Monitoring Programs for Travel Time and Establishing Monitoring Programs for Travel Time Reliability, Prepared for the SHRP2 program under project L02*. Paper presented at the Transportation Research Board, Washington, DC.

List, G. F., Wood, B., Nozick, L. K., Turnquist, M. A., Jones, D. A., Kjeldgaard, E. A., & Lawton, C. R. (2003). Robust optimization for fleet planning under uncertainty. *Transportation Research Part E: Logistics and Transportation Review*, *39*(3), pp. 209–227. https://doi.org/10.1016/S1366-5545(02)00026-1.

List, G. F., Wood, B., Turnquist, M. A., Nozick, L. K., Jones, D. A., & Lawton, C. R. (2006). Logistics planning under uncertainty for disposition of radioactive wastes. *Computers & operations research, 33*(3), pp. 701-723.

Little Inc, A. D. (1971). Economic Impact of Freight Car Shortages. Executive Summary, pp. 8.

Little, P. D., Kwon, O. H., & Martland, C. D. (1992). *An Assessment of Trip Times and Reliability of Boxcar Traffic*.

Lomax, T., Schrank, D., Turner, S., & Margiotta, R. (2003). Selecting travel reliability measures. *Texas Transportation Institute monograph (May 2003)*.

Lomax, T., Turner, S., Hallenbeck, M., Boon, C., & Margiotta, R. (2001). *Traffic Congestion and Travel Reliability: How bad is the situation and what is being done about it?*

Louly, M.-A., Dolgui, A., & Hnaien, F. (2008). Supply planning for single-level assembly system with stochastic component delivery times and service-level constraint. *International Journal of Production Economics, 115*(1), pp. 236-247. doi: 10.1016/j.ijpe.2008.06.005.

Lyman, K., & Bertini, R. L. (2008). Using travel time reliability measures to improve regional transportation planning and operations. *Transportation Research Record: Journal of the Transportation Research Board, 2046*(1), pp. 1-10.

Ma, X., McCormack, E. D., & Wang, Y. (2011). Processing commercial global positioning system data to develop a web-based truck performance measures program. *Transportation Research Record*, *2246*, pp. 92-100. doi: 10.3141/2246-12.

Macharis, C., & Bontekoning, Y. M. (2004). Opportunities for OR in intermodal freight transport research: A review. *European Journal of Operational Research, 153*(2), pp. 400-416. doi: 10.1016/s0377-2217(03)00161-9.

Markovic, N., & Schonfeld, P. (2011). Scheduling Under Uncertainty for Single-Hub Intermodal Freight System. *Transportation Research Record: Journal of the Transportation Research Board, 2238*, pp. 24-31.

Martland, C. D. (1975). Railroad Reliability and Freight Car Utilization: An Introduction. *Issue 75, Part 8 of CTS report.*
https://books.google.com/books/about/Railroad_Reliability_and_Freight_Car_Uti.html?id=Fw1FvgAACAAJ

Martland, C. D. (1982). PMAKE analysis: Predicting rail yard time distributions using probabilistic train connection standards. *Transportation Science, 16*(4), pp. 476-506.

Martland, C. D. (1992). Rail Freight Service Productivity from the Manager's Perspective. *Transportation Research Part A: Policy and Practice, 26*(6), pp. 457-469.

Martland, C. D., & Smith, M. E. (1990). Estimating the Impact of Advanced Dispatching Systems on Terminal Performance. *Journal of the Transportation Research Forum, 30*(2), pp. 286-300.

Martland, C. D., & Sussman, J. M. (1972). Rail Service Time Reliability, *An Analysis of Operating Data, 13*.

McCormack, E. D., Ma, X., Klocow, C., Currarei, A., & Wright, D. (2010). Developing a GPS-based truck freight performance measures platform, pp. 61.

McKinnon, A., Palmer, A., Edwards, J., & Piecyk, M. (2008). Reliability of Road Transport from the Perspective of Logistics Managers and Freight Operators.

McLeod, D. S., Elefteriadou, L., & Jin, L. (2012). *Travel Time Reliability as a Performance Measure: Applying Florida's Predictive Model on the State's Freeway System*.

Meng, Q., Lee, D.-H., & Cheu, R. L. (2005). Multiobjective Vehicle Routing and Scheduling Problem with Time Window Constraints in Hazardous Material Transportation. *Journal of Transportation Engineering, 131*(9), pp. 699-707.

Meyburg, A., & Mbwana, J. (2001). Conference Synthesis: Data Needs in the Changing World of Logistics and Freight Transportation. *Albany, NY, USA: New York State Department of Transportation*.

Miller, S. B., & Little, J. D. C. (1967). Evaluation and improvement of traffic signal settings by simulation. *Highway Research Record, 170*, pp. 56-59.

Mirchandani, P. B. (1980). Locational decisions on stochastic networks. *Geographical Analysis, 12*(2), pp. 172-183.

Mirchandani, P. B., & Odoni, A. R. (1979). Locations of Medians on Stochasitc Networks. *Transportation Science, 13*(2), pp. 85-98.

Mitrovic-Minic, S., & Laporte, G. (2004). Waiting Strategies for the Dynamic Pickup and Delivery Problem with Time Windows. *Transportation Research Part B: Methodological, 38*(7), pp. 635-655.

Motraghi, A., & Marinov, M. V. (2012). Analysis of urban freight by rail using event based simulation. *Simulation Modelling Practice and Theory, 25*, pp. 73-89. doi: 10.1016/j.simpat.2012.02.009.

Muthuraman, K. (1991). Inventory Management with Stochastic Lead Times. *Mathematics of Operations Research* 40(2), pp. 302-327.

Nunez, A., Berger, L., Morcello, E., & Noleo, M. (2008). The value of time and the value of reliability in cross-Alpine and cross-Pyrennean freight transport.

Nussbaum, E., Rebibo, K. K., & Wilhelm, E. (1975). RUCUS (Run Cutting and Scheduling Implementation Manual). pp. 198.

O'Brien, T., Hicks, G. V., & Lyte, B. (2007). Growth of California Ports: Opportunities and Challenges. A Report to the California State Legislature. pp. 122.

Office, U. S. G. A. (2013). Puerto Rico: Characteristics of the Island's Maritime Trade and Potential Effects of Modifying the Jones Act. pp. 46.

Ogawa, M. J., Poole, A. J., Meekums, R. J., Cooke, D., & Self, S. (2012). *Managed Motorways: Keeping Business Moving*.

Ouyang, L.-y., & Chang, H.-c. (2001). The variable lead time stochastic inventory model with a fuzzy backorder rate. *44*(1).

Owen, S. H., & Daskin, M. S. (1998). Strategic facility location: A review. *European Journal of Operational Research, 111*(3), pp. 423-447. doi: 10.1016/S0377-2217(98)00186-6.

Ozkaya, E., Keskinocak, P., Joseph, V. R., & Weight, R. (2010). Estimating and Benchmarking Less-Than-Truckload Market Rates. *Transportation Research Part E: Logistics and Transportation Review, 46*(5), pp. 667-682.

Pape, V. (1988). Car Transportation by Truck. Vehicle Routing Methods and Studies. *Studies in Management Science and Systems*, 16, *Publication of: Dalctraf*, pp. 425-437.

Pickrell, S. (2007). Cost-Effective Measures and Planning Procedures for Travel Time, Delay, and Reliability.

Plotkin, S. C. (1969). Automation of the highway, an overview. *IEEE: Transactions on Vehicular Technology*, *18*(2), pp. 77-80.

Poe, C. (2010). Geometric Design and Operational Factors that Impact Truck Use of Toll Roads, pp. 78.

Poole, R. W. (2007). The Case for Truck-Only Toll Lanes. *Public Works Management & Policy, 11*(4), pp. 244-249.

Potvin, J.-y. (2007). Evolutionary Algorithms for Vehicle Routing. (November).

Powell, W. B. (1988). A Comparative Review of Alternative Algorithms for the Dynamic Vehicle Allocation Problem. *Vehicle Routing*, pp. 249 - 291.

Prozzi, J., Seedah, D., Carrion, M., Perrine, K., Hutson, N., Bhat, C., & Walton, C. M. (2011). Freight Planning for Texas—Expanding the Dialogue (pp. 169p-169p).

Puckett, S. M., & Rascuite, S. (2010). Freight stakeholders' sensitivities under road user charging: a latent class approach. *Australasian Transport Research Forum (ATRF), 33rd, 2010, Canberra, ACT, Australia*, pp. 19.

Raicu, R., Raicu, S., & Popa, M. (2005, 2006). *The Influence of Transportation Network Reliability on City Logistics*.

Reilly, J., & D'Ignazio, J. (1983). Microcomputer Applications in Transit Agencies. *Transportation Research Record*(932), pp. 9-12.

Ronen, D. (2002). Cargo Ships Routing and Scheduling: Survey of Models and Problems. *Maritime Transport*, pp. 3-10.

Ronsee, S., & Rayp, G. (2016). International shipping traffic as a determinant of the growing use of advertisements by local shopkeepers: a case study of the eighteenth century Ghent. *Business History*, *41*(2), pp. 479–500.

Roos, D. (1971). *Dial-A-Bus System Feasibility*.

Roos, D., & Porter, E. H. (1971). *Dial-A-Ride: Urban Mass Transportation Demonstration Project*, pp. 35.

Roos, D., & Wilson, D. G. (1971). *Dial-A-Ride: An Overview of a New Demand-Responsive Transportation System*.

Ruamsook, K., & Thomchick, E. A. (2009). *Sustainable Freight Transportation: A Review of Strategies*. Brown, pp.1-20.

Rubin, F. (1975). Routing Algorithms for Urban Rapid Transit. *Transportation Research, UK, 9*(4), pp. 215-223.

Saiyed, S., Khaderoo, N., & Kocialek, G. (2012, 2012). *Meeting Freight Challenges Through Effective Regional Planning and Partnerships*.

Sakul, K. (2010). Military Transportation As Part of Mediterranean Maritime Trade: Ottoman Freight Payments During the War of the Second Coalition (1798-1802). *Journal of Mediterranean Studies*, *19*(2), pp. 389–406.

Sauder, R. L. L. (1976). *SIMTRAN*.

Schrank, D., Eisele, B., & Lomax, T. (2012). TTI's 2012 urban mobility report. *Texas A&M Transportation Institute. The Texas A&M University System*.

Sexton, T. R., & Bodin, L. D. (1985). Optimizing Single Vehicle Many-To-Many Operations with Delivery Times: II. Routing. *Transportation Science, 19*(4), pp. 411-435.

Shamberger, R. C. (1975, 1975). *Freight Car Management Systems: A New Way Out of the Car Utilization Dilemma.*

Shinghal, N., & Fowkes, T. (2002). Freight Mode Choice and Adaptive Stated Preferences. *Transportation Research Part E: Logistics and Transportation Review, 38*(5), pp. 367-378.

Sierleja, E. J., Pipas, G., & List, G. F. (1981). Evaluation of MOPAC's Freight Car Scheduling System, pp. 72.

Sines, G. (1972). MOPAC'S Transportation Control System. A Systems Approach to Achieving Service Reliability, *13*.

Slater, A. (2002). Specification for a Dynamic Vehicle Routing and Scheduling System.. *International Journal of Transport Management. 1*(1), pp. 29-40.

Srinivasan, K. K., Prakash, A. A., & Seshadri, R. (2014). Finding most reliable paths on networks with correlated and shifted log–normal travel times. *Transportation Research Part B: Methodological*, *66*, pp. 110–128. https://doi.org/10.1016/j.trb.2013.10.011.

Stephens, C. (1989). "The Most Reliable Time": William Bond, the New England Railroads, and Time Awareness in 19th-Century America. *Technology and Culture*, *30*(1), pp. 1–24. https://doi.org/10.2307/3105429.

Strocko, E., & Schoener, G. (2012). Moving the Nation's Goods. *Public Roads, 75*(6), pp. 42-50.

Sussman, J. M., & Martland, C. D. (1974). Studies in Railroad Operations and Economics. Volume II: Improving Railroad Reliability: A Case Study of the Southern Railway. Executive Summary, pp. 24.

Swan, P. F., & Tyworth, J. E. (2001). Shipper Sensitivity to Unreliable Service in Carload Markets. *Transportation Journal, 40*(3), pp. 16-25.

Taniguchi, E., & Ando, N. (2008). Probabilistic vehicle routing and scheduling based on probe vehicle data. *International Journal of Its Research, 2*(1), pp. 29-37.

Taniguchi, E., & Nakanishi, M. (2003). ITS Based Dynamic Vehicle Routing and Scheduling with Real Time Traffic Information.. *International Journal of ITS Research, 1*(1), pp. 49-60.

Taniguchi, E., & Okamoto, M. (2005, 2005). *Evaluating Urban Dynamic Vehicle Routing and Scheduling with Time Windows Using Traffic Information*.

Taniguchi, E., & Thompson, R. G. (2002). Modeling city logistics. *Transportation Research Record*(1790), pp. 45-51.

Taniguchi, E., Thompson, R. G., Yamada, T., & van Duin, R. (2001). City Logistics. *Network Modelling and Intelligent Transport Systems*, pp. 252.

Taniguchi, E., & Yamada, T. (2003, 2003). *Reliable Vehicle Routing and Scheduling with TIme Windows Towards City Logistics.*

Taniguchi, E., Yamada, T., & Tamagawa, D. (1999, 1999). *Modelling Advanced Routing and Shceudling of Urban Pickup/Delivery Trucks.*

Taniguchi, E., Yamada, T., & Tamagawa, D. (2001). Reliability of Transport Networks. Chapter 6: Probabilistic Routing and Scheduling of Urban Pickup/Delivery Trucks with Variable Travel Times. *Publication of: RESEARCH STUDIES PRESS LTD*, pp. 73-89.

Taniguchi, E., Yamada, T., & Tamaishi, M. (2001). *ITS Based Dynamic Vehicle Routing and Scheduling.*

Taş, D., Gendreau, M., Dellaert, N., van Woensel, T., & de Kok, A. G. (2014). Vehicle routing with soft time windows and stochastic travel times: A column generation and branch-and-price solution approach. European Journal of Operational Research, 236(3), pp.789–799. https://doi.org/10.1016/j.ejor.2013.05.024.

Taş, Duygu, Dellaert, N., van Woensel, T., & de Kok, T. (2013). Vehicle routing problem with stochastic travel times including soft time windows and service costs. *Computers & Operations Research*, *40*(1), pp. 214–224. https://doi.org/10.1016/j.cor.2012.06.008.

Taş, Duygu, Dellaert, N., van Woensel, T., & de Kok, T. (2014). The time-dependent vehicle routing problem with soft time windows and stochastic travel times. *Transportation Research Part C: Emerging Technologies*, *48*, pp. 66–83. https://doi.org/10.1016/j.trc.2014.08.007.

Tavasszy, L. A., Ruijgrok, K., & Davydenko, I. (2012). Incorporating Logistics in Freight Transport Demand Models: State-of-the-Art and Research Opportunities. *Transport Reviews, 32*(2), pp. 203-219. doi: 10.1080/01441647.2011.644640.

Teodorovic, D., Kikuchi, S., & Hohlacov, D. (1991). A routing and scheduling method in considering trade-off between the user's and the operator's objectives. *Transportation Planning and Technology, 16*(1), pp. 63-75.

Train, K. E., & Wilson, W. W. (2008). Transportation Demand and Volume Sensitivity: A Study of Grain Shippers in the Upper Mississippi River Valley. *Transportation Research Record: Journal of the Transportation Research Board*(2062), pp. 66-73.

Transportation_Research_Board. (2013) Innovations in Freight Demand Modeling and Data. *Transportation Research Board.* pp. 4.

Tuler, S., Kasperson, R. E., & Ratick, S. (1989). Human Reliability and Risk Management in the Transportation of Spent Nuclear Fuel. Reliability on the Move: Safety and Reliability in Transportation. *Proceedings of the Safety and Reliability Society Symposium, 1989. Bath, 11-12 October 1989*. 1989. *Publication of: Elsevier Applied Science Publishers Limited*, pp. 169-194.

Turnquist, M. A., & Daskin, M. S. (1982). Queuing models of classification and connection delay in railyards. *Transportation Science, 16*(2), pp. 207-230.

Van Der Mede, P. H. J., Palm, H., & Flikkema, H. (1996, 1996). *Trave Time Variability as a New Quality Indicator.*

Van Dyke, C. (1986). The automated blocking model: A practical approach to freight railroad blocking plan development. *Paper presented at the Proceedings of the Transportation Research Forum.*

van Lint, J. W. C., van Zuylen, H. J., & Tu, H. (2008). Travel time unreliability on freeways: Why measures based on variance tell only half the story. *Transportation Research Part A: Policy and Practice, 42*(1), pp. 258-277. doi: 10.1016/j.tra.2007.08.008.

Van Vliet, D. (1976). Road assignment-III: Comparative tests of stochastic methods. *Pergamon Press, Incorporated, 10*(3), pp. 151-157.

Varma, A. (2008). Measurement Sources for Freight Performance Measures and Indicators, pp. 555.

Vernimmen, B., Dullaert, W., & Engelen, S. (2007). Schedule Unreliability in Liner Shipping: Origins and Consequences for the Hinterland Supply Chain. *Maritime Economics & Logistics, 9*(3), pp. 193-213.

Vidal, T., Crainic, T. G., Gendreau, M., Lahrichi, N., & Rei, W. (2012). A hybrid genetic algorithm for multidepot and periodic vehicle routing problems. *Operations Research, 60*(3), pp. 611-624.

Vromans, M. (2005). *Reliability of Railway Systems*. Netherlands TRAIL Research School. pp. 243.

Walsh, P., & Somers, A. (2011). *Monash-Citylink-West Gate Freeway upgrade freeway management system*.

Wang, D., & Ma, Y. (2008). *A location programming under stochastic travel time through genetic algorithm.* Paper presented at the 2008 International Conference on Wireless Communications, Networking and Mobile Computing, WiCOM 2008, October 11-12, 2008, Dalian, China.

Wang, R.-T. (2007). Improving service quality using quality function deployment: The air cargo sector of China airlines. *Journal of Air Transport Management, 13*(4), pp. 221-228. doi: 10.1016/j.jairtraman.2007.03.005.

Wang, Z., & Lin, L. (2013). A Simulation-Based Algorithm for the Capacitated Vehicle Routing Problem with Stochastic Travel Times. Journal of Applied Mathematics, pp. 1–10. https://doi.org/10.1155/2013/127156.

Washburn, S. S., & Ko, B. (2007). Travel Time Reliability and Truck Level of Service on the Strategic Intermodal System Part B: Truck Level of Service, pp. 303.

Weise, T., Podlich, A., & Gorldt, C. (2009). Solving real-world vehicle routing problems with evolutionary algorithms *Natural Intelligence for Scheduling, Planning and Packing Problems*, pp. 29-53.

Westgate, B. S., Woodard, D. B., Matteson, D. S., & Henderson, S. G. (2016). Large-network travel time distribution estimation for ambulances. *European Journal of Operational Research*, *252*(1), pp. 322–333. https://doi.org/10.1016/j.ejor.2016.01.004.

Whybark, D. C. (1974). The Combined Transportation and Inventory Policy Decision.

Wiegmans, B. W., Hekkert, M., & Langstraat, M. (2007). Can Innovations in Rail Freight Transhipment Be Successful? *Transport Reviews, 27*(1), pp. 103-122. doi: 10.1080/01441640600765091.

Wigan, M., Rockliffe, N., Thoresen, T., & Tsolakis, D. (2000). Valuing Long-Haul and Metropolitan Freight Travel Time and Reliability. *Bureau of Transportiton Statistics.- United States Department of Transportation Journal of Transportation and Statistics, 3*(3).

Williams, J. H. (1972). Some Techniques for Controlling Rail Service Reliability, *13*.

Williamson, W. V. (1977). Freight Car Utilization: A Look at Some Basic Constraints and Possible Solutions. pp. 128-135.

Wilson, D. G. (1971). Automated Guideway Transportation Between and Within Cities, pp. 251.

Winder, A. (2009). Transport management - thematic research summary, pp. 70.

Wong, P., Husing, J. E., & Pfeffer, N. (2007, 2007). *Southern California Goods Movement Challenge, Opportunity, and Solution*.

Woo, S.-H., & Pettit, S. (2010). Port performance in changing logistics environments: measurement development and model testing. pp. 439-488.

Wren, A., & Holliday, A. (1972). Computer Scheduling of Vehicles from One or More Depots to a Number of Delivery Points. *Operational Research Quarterly (1970-1977)*, *23*(3), pp. 333–344. https://doi.org/10.2307/3007888.

Wren, A., & Rousseau, J.-M. (1995). Bus Driver Scheduling — An Overview. In Computer-Aided Transit Scheduling, *430*, Springer, Berlin, Heidelberg, pp. 173–187. https://doi.org/10.1007/978-3-642-57762-8_12.

Wright, D. S. (1986). Freight Information Systems - Implications for Infastructure Policy -Freight. *Proceedings of Seminars G and T held at the PTRC Summer Annual Meeting*, Sussex University, England, 14-17 July 1986. *277. Publication of: PTRC EDUCATION AND RESEARCH SERVICES, LTD*, pp. 53-59.

Wurfel, E., Huan, L., Bai, Y., & Buhr, V. (2008). Development of Freight Analysis Framework for the Sustainable Economic Growth pp. 84.

Xing, T., & Zhou, X. (2011). Finding the most reliable path with and without link travel time correlation: A Lagrangian substitution based approach. *Transportation Research Part B: Methodological*, *45*(10), pp. 1660–1679. https://doi.org/10.1016/j.trb.2011.06.004.

Xu, J., Hancock, K. L., & Southworth, F. (2008). Dynamic Freight Traffic Simulation Providing Real-Time Infimation, pp. 1711-1719.

Xu, W., He, S., Song, R., & Li, J. (2009). Reliability based assignment in stochastic-flow freight network. *Applied Mathematics and Computation, 211*(1), pp. 85-94. doi: 10.1016/j.amc.2009.01.024.

Yagar, S., Saccomanno, F., & Shi, Q. (1983). An efficient sequencing model for humping in a rail yard. *Transportation Research Part A: General, 17*(4), pp. 251-262.

Yamada, T., & Taniguchi, E. (2005, 2006). *Modelling the Effects of City Logistics Schemes*.

Yamada, T., Yoshimura, Y., & Mori, K. (2004, 2004). *Road Network reliability Analysis Using Vehicle Routing and Scheduling Procedures*.

Yang, C. H., & Regan, A. C. (2012). A multi-criteria decision support methodology for implementing truck operation strategies. *Transportation, 40*(3), pp. 713-728. doi: 10.1007/s11116-012-9432-7.

Young, R. (2012). Use of Travel Time, Travel Time Reliability, and Winter Condition Index Information for Improved Operation of Rural Interstates.

Yuan, J. (2006). Stochastic Modelling of Train Delays and Delay Propagation in Stations, pp. 162.

Zamparini, L., & Reggiani, A. (2007). Freight Transport and the Value of Travel Time Savings: A Meta-analysis of Empirical Studies. *Transport Reviews, 27*(5), pp. 621-636. doi: 10.1080/01441640701322834.

Zamparini, L., & Reggiani, A. (2010). The value of reliability and its relevance in transport networks. pp. 97-115.

Zhao, W., & Goodchild, A. V. (2011). Truck Travel Time Reliability and Prediction in a Port Drayage Network. *Maritime Economics & Logistics, 13*(4), pp. 387-418.

Zhao, W., McCormack, E., Dailey, D. J., & Scharnhorst, E. (2013). Using Truck Probe GPS Data to Identify and Rank Roadway Bottlenecks. *Journal of Transportation Engineering, 139*(1), pp. 1-7.

Zhao, Y., & Simchi-Levi, D. (2006). Performance Analysis and Evaluation of Assemble-to-Order Systems with Stochastic Sequential Lead Times. *Operations Research, 54*(4), pp. 706-724. doi: 10.1287/opre.1060.0302

Zheng, F., & Van Zuylen, H. (2013). Urban link travel time estimation based on sparse probe vehicle data. *Transportation Research Part C: Emerging Technologies, 31*, pp. 145-157. doi: 10.1016/j.trc.2012.04.007

Zografos, K. G., & Androutsopoulos, K. N. (2002). Heuristic Algorithms for Solving Hazardous Materials Logistical Problems. *Transportation Research Record, 1783*, pp. 158-166.