Vehicle Tracking Data Imputation and Shortest Path Prediction under Stochastic Travel Time Conditions

A Dissertation Presented for the

Doctor of Philosophy

Degree

The University of Tennessee, Knoxville

Michael J Han

August 2025

Copyright © 2025 by Michael J Han

All rights reserved.

DEDICATION

To Mom, Dad, Steven, Grandma, and the memories of Grandpa

ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. Shuai Li, for the privilege of researching and working with him. The past three years have been an extraordinary gift of opportunity and learning, and none of it would have been possible without his mentorship and support.

I am also grateful to Dr. Baoshan Huang and Dr. Weizi Li for their insight and patience, which gave me confidence in pursuing the work documented herein.

Special thanks to Dr. Hyeonsup Lim and Dr. Majbah Uddin of Oak Ridge National Laboratory for the opportunity to work with them at the National Transportation Research Center. Their guidance over the past year has been instrumental in shaping both my research and my career path.

Dr. Chris Cox, Head of CEE, was generous and understanding in his support throughout my graduate studies. I also thank Dr. Hairong Qi of EECS for her patient mentorship during the early stages of my journey.

Above all, I am forever grateful to my parents, grandparents, and brother. Like any meaningful journey, the past three years came with their share of imperfections, but my family has been a constant source of encouragement and strength. Steven, thank you for watching the entirety of *Game of Thrones* with me. Grandpa, I wish I had one more moment to see you light up at the thought of Steven and me singing karaoke versions of hit numbers from *Les Misérables*.

To the many others who lent a hand along the way—if I don't get to thank you in person, please know that your kindness has not gone unnoticed. Thank you all.

ABSTRACT

This study addresses two major challenges in transportation analytics: recovering missing truck-tracking data and predicting shortest paths under stochastic conditions. We first investigated significant data loss in truck-tracking datasets, where nearly one-third of trips were missing due to misaligned data collection windows between origin and destination points. These losses, largely unaccounted for in aggregated statistics, distorted key trends such as average journey time and trip frequency. We identified nine distinct data loss scenarios, with late starts and early ends at destination nodes contributing most significantly to our case study.

To recover the missing data, we expanded and analyzed the raw field data to assess its stochastic distribution, identifying the shifted Negative Exponential Distribution (shifted-NED) as the best fit. Validated through iterative Monte Carlo simulations (MAPE = 4.79%, MAE < 9 minutes), this model enabled the generation of synthetic missing data. By fusing these simulated values with the existing dataset, we created hybrid datasets that corrected misleading trends, replacing artificially decreasing journey times and declining trip frequencies with stable and plausible patterns. Though real-time driver behavior analysis was beyond the scope of this study due to the delayed installation of Dynamic Message Signs, our imputation process lays a foundation for such future work.

Building on this robust, corrected dataset, we turned to the challenge of predicting shortest paths in the face of stochastic travel times. Traditional deterministic algorithms fall short when future travel times are uncertain. We reframed the shortest-path problem

as a stochastic one and introduced the *Degree of Independence (DoI)* to quantify how path interdependencies affect the likelihood of optimality. Through millions of Monte Carlo simulations across varied grid structures, we established a strong correlation between *DoI* and a path's probability of being the shortest.

We developed the *Relative Frequency (RF)* metric to quantify this advantage, showing that high-*DoI* paths can be far more optimal than average paths. We developed a near-perfect regression model ($R^2 = 0.9964$) that allows *RF* estimation across grid sizes, transforming a previously intractable stochastic problem into a simple deterministic one. Together, these contributions enable more accurate modeling of both route performance and traveler behavior under uncertainty.

TABLE OF CONTENTS

INTRODUCTION 1

CHAPTER I LONG-DISTANCE TRUCK TRACKING DATA IMPUTATION FOR IMPROVED TRIP TIME/PATH ASSESSMENT A CASE STUDY 3

ABSTRACT 4

INTRODUCTION 6

Need to Track Trucks 6

Truck Tracking Technologies 7

Problem Statement 9

STUDY SITE AND FIELD DATA 9

Bypass Route Around the Metropolitan Nashville Area 9

High-Resolution Traffic Data 13

ALPR Large Truck Tracking Data 17

ASSESSMENT OF LIMITATIONS AND MISSING DATA SCENARIOS 20

Limitations of NPMRDS or Other Traffic Data 20

ALPR Large Truck Tracking Data 20

DATA IMPUTATION FOR FURTHER STUDIES 30

Imputation to Restore Missing Data Statistics 30

Verification with Monte Carlo Simulations 34

Simulation-Based Imputation for a Hybrid Dataset 37

CONCLUSIONS 41

REFERENCES 44

CHAPTER II LIKELY SHORTEST PATH AMONG VIABLE OPTIONS IN A SQUARE GRID WITH STOCHASTIC TRAVEL TIME 46
ABSTRACT 47
INTRODUCTION 49
Example 1 – Shortest Path Observations in Midtown Manhattan, NYC 49
Example 2 – Deterministic and Stochastic Travel Times 54
Example 3 – Competing Paths with Shared Links 56
Problem Statement 58
ANALYSES 59
Basic Terminology and Assumptions 59
Paths and Shortest Paths on a Rectangular Grid 61
Example 4 – All Viable Paths in a Simple 2×2 Rectangular Grid 63
Example 5 – Simple Scenarios of 3 Paths in a 10×10 Grid 65
The Concept of Fractional Independence among Paths 67
A Path's Degree of Independence and Its Likelihood of Being the Shortest 70
The Relative Frequency (RF) Model 74
So, Which Paths Have the Highest DoI 76
CONCLUSIONS 78
REFERENCES 80
CONCLUSION 82
VITA 84

LIST OF TABLES

Table I-1. Field ALPR data collection results. 18

Table II-1. Viable nodes, links, and paths from (0,0) to (*m*, *n*) with no backtracking..... 62

Table II-2. Winning percentage of each path for several scenarios, including those in Example 5. 66

Table II-3. Degree of Independence of each path in the scenarios in Example 5..... 68

LIST OF FIGURES

Figure I-1. Major U.S. highways and auxiliary routes around selected metropolises 7

Figure I-2. Interstate routing options around Nashville before and after I-840 10

Figure I-3. TDOT signage before I-40E/I-840E split (Google Street View) 12

Figure I-4. Nashville study site with through traffic via I-40E and bypass via I-840E 12

Figure I-5. TDOT Region III (Nashville) RDS station locations in 2016 14

Figure I-6. Sample travel time comparisons on I-40E and I-840E based on NPMRDS .. 14

Figure I-7. Travel time on I-40E and I-840E at the split point based on NPMRDS 16

Figure I-8. Truck tracking study with ALPR units at three sites 18

Figure I-9. Decaying trend of journey time distribution on both routes [Han 2019]..... 19

Figure I-10. Mean journey time and number of trucks matched plotted against the departure time at the origin [Han 2019] 19

Figure I-11. Trends seemingly present in Figure I-10 21

Figure I-12. Individual truck journey time vs time at Origin for I-40E and I-840E..... 23

Figure I-13. Hypothetical patterns of bullet holes on surviving aircraft during WWII 24

Figure I-14. Missing data scenarios, as seen in a journey time vs time at the origin plot 24

Figure I-15. Time-space diagram of data collection windows at the origin and the destinations along I-40E, and several scenarios of trucks not captured 26

Figure I-16. Missing data scenarios, in a journey time vs time at the destination plot..... 27

Figure I-17. General imputation for data loss due to Scenarios 1 and 2..... 31

Figure I-18. Cumulative percentage of all journey times and curve fitting results 31

Figure I-19. Regression summary of the shifted LOG model 32

Figure I-20. Regression summary of the shifted NED model 33
Figure I-21. Visualization of outcomes of one of the runs simulating I-40E truck traffic 35
Figure I-22. Visualization of outcomes of one of the runs simulating I-40E truck traffic 36
Figure I-23. Visualization of outcomes of one of the runs simulating I-40E truck traffic 36
Figure I-24. Sample imputed hybrid data: field data plus simulated missing data..... 38
Figure I-25. Comparisons of sample imputed hybrid statistics vs field data statistics 39
Figure I-26. Comparisons of the number of trips and trip rates vs the time at the origin. 40
Figure I-27. Comparisons of the raw and imputed hybrid data of truck trips at the origin 41
Figure II-1. Observation site and OD pair in Midtown Manhattan. (Google Maps, 06:56:24 AM on 12/03/2024) 51
Figure II-2. Shortest paths for the same OD suggested by Google Maps within 20 minutes 52
Figure II-3. Travel time scenarios for three independent paths. 55
Figure II-4. Three paths on a simplified rectangular grid. 57
Figure II-5. Converging probability of each path winning the shortest path race 58
Figure II-6. Transformation from $[(x_1, y_1) – (x_2, y_2)]$ to $[(0, 0) – (m, n)]$ 62
Figure II-7. All viable paths in a 2×2 rectangular grid 64
Figure II-8. Likelihood of each viable path for a 2×2 rectangular grid 64
Figure II-9. Three simplified scenarios in a 10×10 rectangular grid 65
Figure II-10. All paths from (0,0) passing through link *j* to (*m*,*n*)..... 69
Figure II-11. Likelihood of a path being the shortest vs the path's *DoI* in an 8×8 grid.. 70

Figure II-12. Likelihood of a path being the shortest vs the path's *DoI* in a 7×9 grid.... 72

Figure II-13. Likelihood of a path being the shortest vs the path's *DoI* in a 6×10 grid.. 72

Figure II-14. Relative frequency (RF) of the likelihood for $m = 1..10$ and $n = 1..10$..... 73

Figure II-15. Regression summary of the *RF* model 74

Figure II-16. Visualization of *Log*(*RF*) as a function of *m* and *n* 75

Figure II-17. Number of paths sharing each link in a 20×20 grid..... 77

INTRODUCTION

In his book on Traffic Flow on Transportation Networks, Newell* concerned himself with representing real transportation networks in a mathematical analysis framework. Instead of the common practice of dividing the city or region in question into zones and representing each zone by a single centroid, and hence analyzing trips on a skeletal roadway network, Newell proposed approximations and idealizations in an ingenious fashion. For short trips on the local road network made of numerous roads, typical in major cities, intra-zone trips are not automatically generated or ended at some virtual centroids but at their actual ODs and routed in a dense minor-road network that is considered a continuum. These dense networks can be rectangular, triangular, ring-and-radial, and can have several other configurations. For long trips, on the other hand, once a trip is routed in the most expeditious fashion to the nearby high-speed road, e.g., freeways and major arterials. The number of such highways is not likely to be very large, so they form a graph of quite manageable size for analysis. In that spirit, this dissertation analyzed long-haul trucking trips on major interstate highways in the first chapter and approximated local trips on a dense grid in the second chapter.

Chapter 1 of this dissertation deals with long-distance truck-tracking data on two competing independent routes of over 60 miles. Because of the design of the experiments and the nature of such longitudinal tracking studies, a substantial number of truck trips were "lost," which resulted in aggregated statistics that are significantly biased, but not always discernibly so. We recognized the misleading trends in the aggregated data, identified the source of the biases, generalized the types of data loss scenarios with

longitudinal tracking studies, characterized the stochastic nature of the travel time data, proposed an imputation framework, and verified the methodology with Monte Carlo simulations. The imputed hybrid data sets minimized the bias in the raw data, dispelled the initial misleading trends, and afforded researchers the possibility to conduct further analyses on driver behaviors or trucker activities.

Chapter 2 of the dissertation implemented a dense rectangular grid approximation for city roads for the analysis of the shortest path under stochastic travel time conditions. While prevailing shortest path algorithms are efficient and well accepted, they typically rely on deterministic or even fixed travel time on each link for shortest path calculation. The resultant shortest paths can change within seconds, and again, as demonstrated herein, one is never sure which of all the numerous competing paths was, is, or will be the shortest. This study recognizes the stochastic nature of link and path travel times and establishes in its analyses that the most likely shortest paths are those with the least overlapping or shared links with other paths. A measure of "degree of independence" of a path was defined and was shown to be highly indicative of the likelihood of a path being the shortest. This relationship was verified with Monte Carlo simulations using different stochastic travel times and a wide range of OD configurations.

*Newell, G.F. (1980) Traffic Flow on Transportation Networks, MIT Press, 276 p.

CHAPTER I

LONG-DISTANCE TRUCK TRACKING DATA IMPUTATION FOR IMPROVED TRIP TIME/PATH ASSESSMENT A CASE STUDY

This chapter is being revised and reformatted for publication in a refereed journal.

ABSTRACT

In this study, we identified and addressed significant data integrity issues within aggregated truck-tracking statistics. Initial observations revealed unexplained trends in journey time and truck trip frequencies, which we traced back to a substantial loss of data during the collection process. This loss, we estimated, accounted for nearly one-third of the total dataset and was largely attributable to the misalignment of data collection time windows between origin and destination nodes. When trips were not captured at both ends within the prescribed time windows, those trips were effectively discarded, leaving critical yet inconspicuous gaps in the dataset.

To systematically understand and quantify these gaps, we defined and categorized nine distinct data loss scenarios. These included late starts and early ends at either the origin or destination, as well as station outages and compounded disruptions. Among these, the scenarios involving late starts and early ends at the destination were found to be the most prevalent in our case study. This classification enabled us to better conceptualize the scale and nature of the truncation problem.

We then expanded and aggregated the raw field data to analyze its underlying stochastic distribution—a crucial step toward enabling accurate data imputation. Among the models considered, the shifted Negative Exponential Distribution (shifted-NED) emerged as the best fit. Its suitability stemmed from its ability to capture the decaying frequency of trip durations effectively, aligning with patterns observed in the proportionally adjusted raw data.

To validate this model, we employed iterative Monte Carlo simulations to reconstruct the average journey time from the raw data in reverse. The results confirmed the model's accuracy, yielding a Mean Absolute Percentage Error (MAPE) of 4.79% and a Mean Absolute Error (MAE) of under 9 minutes. Building on this, we simulated multiple sets of missing data and integrated them with the original dataset to create a series of hybrid datasets. These reconstructed datasets are now statistically complete and ready for further analysis.

Notably, the seemingly decreasing journey time trend observed in the original, truncated dataset was corrected in the imputed datasets. Instead, a stable trend with expected stochastic variations emerged. Similarly, the previously observed decline in truck trip frequency over time was replaced by a more plausible and consistent rate of trips per hour.

While this study does not extend to analyzing driver behavior, since the Dynamic Message Sign (DMS) with route travel time information was not installed until four years after the data was collected, our work lays the foundation for such analysis. By designing a robust and transparent imputation process, we enable future research into driver responses and route choices based on complete and reliable datasets.

INTRODUCTION

Need to Track Trucks

Major metropolitan areas in the U.S. are typically connected by Interstate highways and other primary arterials to provide accessibility, mobility, and adequate capacity. With the growth of the Nation's economy around that time, the significant traffic demand and congestion that followed brought need for subsequent parallel routes, beltways, ring roads, bypasses, and other auxiliary highways, see **Figure I-1**, which were added to the initial barebone network and largely completed, over 94%, by 1980 [UTFB 2005]. To maintain mobility and reduce congestion in and around these high-density and built-up areas, it is desirable to route long-haul traffic such that they pivot around spatially irrelevant cities rather than directly enter them. This is especially the case for large commercial vehicles (i.e., utility trucks) that are much bulkier in dimensions, more sluggish in acceleration and deceleration capabilities, less nimble in navigating tighter curves, more liable to generate more pollution, and more likely to carry hazardous materials than passenger cars.

However, the average operator of a large long-haul truck would not necessarily want to take any routes they deem less direct, as it might mean a longer travel time, higher fuel costs, and fewer amenities along these auxiliary routes. To understand the preferences and behaviors of large truck operators, it is crucial to not just count the number of large trucks on the alternate routes but also to track their paths and isolate any cases where both the direct and the alternate routes are viable choices for the specific trips. Other motivations such as traffic control, speed and weight enforcement, electronic

toll collection, emission reduction, international fuel tax collection and refund [IFTA 1998], and freight logistics planning also add to the desirability of tracking large trucks.

Truck Tracking Technologies

Many tracking technologies exist to track corporate inventory or even personal items; they can also be applied to tracking vehicles. Radio Frequency Identification (RFID) is widely used to track shipping containers and by large retail companies like Walmart to track their inventory. Global Positioning System (GPS) based tracking is also widely implemented for tracking the location and movement of cell phones and the person or equipment carrying such devices. Apple's AirTag is a popular Bluetooth technology that can be used to track personal items and valuable goods. Transponders are wireless devices widely used for vehicle identification and tolls. These methods, however, require active installation or embedding of the technology, demonstrate limited coverage or market penetration, and, mostly, lead to potential privacy and data breach issues.

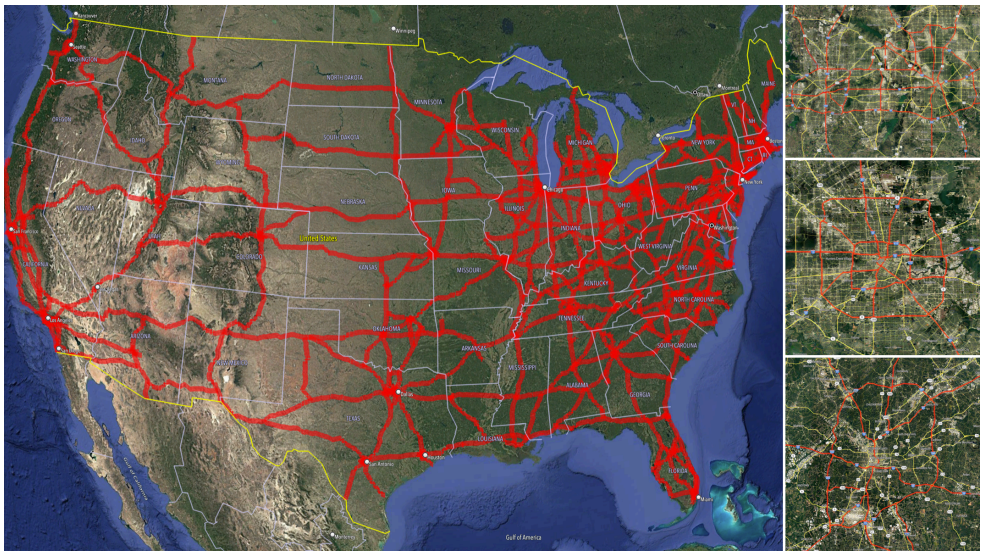


Figure I-1. Major U.S. highways and auxiliary routes around selected metropolises

Government transportation agencies working to track large commercial vehicles find passive or noninvasive technologies with sufficient market penetration to be highly desirable, such as Automated License Plate Recognition (ALPR). License plate recognition was developed as early as the 1970s, extending from optical character recognition (OCR) techniques, which are now mature and widely used in many countries. However, ALPR technologies tend to suffer a relatively low read rate in the U.S. compared to many other parts of the world, mainly because the U.S. does not have a uniform license plate design, format, or template. Thousands of different designs of license plates can be found across the Nation's 50 states and territories [Oliveira-Neto 2009]. If ALPR technologies perform sub-optimally, vehicle tracking requiring the accurate identification of the same plate multiple times at different locations becomes even more challenging. To this end, several ingenious license plate matching algorithms were developed using text-minging techniques and supervised or unsupervised Machine Learning (ML) algorithms [Oliveira-Neto 2009, 2012, 2013, and Hargrove 2016a]. Algorithms like these, following a period of refinement, were implemented successfully in the field for short-distance tracking-based speed accuracy assessment [Hargrove 2016b] and long-distance large-truck route choice assessment [Han 2019]. In that latter study, while the ML-based license plate tracking algorithm performed successfully over real-world long-distance field cases for multiple sites, the resultant statistics indicated certain limitations and potential biases due to missing data beyond the tracking technology itself, be it ALPR, RFID, Bluetooth, or any other multi-point matching mechanisms. This chapter, then, is a follow-up effort to address these issues.

Problem Statement

The objectives of this chapter are to recognize, understand, assess, and impute the data or information that may be missing, even in a latent manner, in matched-pair long-distance vehicle-tracking scenarios to reduce biases in the subsequent statistics. This problem is a special case of any longitudinal tracking study where each subject, be it a patient with a certain illness, an evolving culture phenomenon, a piece of misinformation propagating in social media, or a large truck trundling down the road, needs to be identified and reidentified again at different locations or times. Limitations, and hence latent missing data, could exist due to the truncation of the study duration or the restricted resources of the study. Ultimately, this study aims to minimize the bias in the aggregated statistics due to data loss by data imputation. Once the data is imputed at the individual trip level, additional studies on driver choices, trucking activities, and truck diversion strategies could be carried out. These subsequent analyses would thus be enabled but not engaged in this study.

STUDY SITE AND FIELD DATA

Bypass Route Around the Metropolitan Nashville Area

Interstate Highway I-840, a bypass route south of Nashville, Tennessee, was completed and opened to traffic in 2012. Before its construction, large trucks in East and West Tennessee heading for Alabama and Georgia would travel to near Nashville city center (see **Figure I-2a**) and switch to I-65S or I-24E, or vice versa. With the opening of I-840, traffic can organically divert away from the city center and take advantage of the shorter travel distance of the new bypass (see **Figure I-2b**). In other words, I-840 is

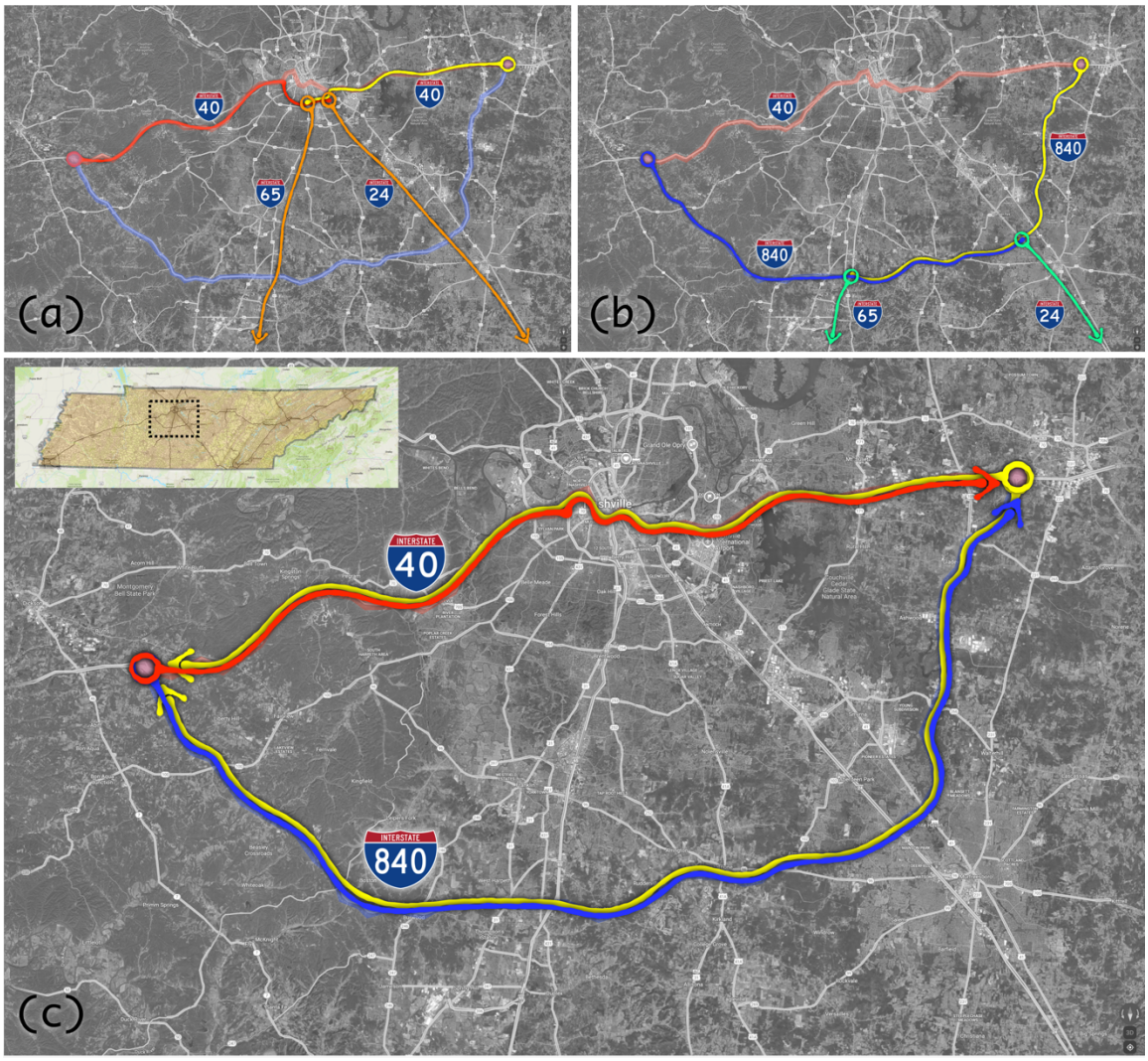


Figure I-2. Interstate routing options around Nashville before and after I-840

clearly intended to function as a bypass route for I-40 traffic from and to I-65/I-24 on the south side of Nashville.

However, it is more than just a curiosity to question whether traffic, be it eastbound or westbound from Nashville, would choose I-840 over I-40. Given that I-80 is longer in distance but less frequent in congestion and delay compared to I-40 (see **Figure I-2c**), what is in question is whether I-840 truly does serve its purpose specifically for long-haul trucks, based on truck drivers' choices.

The distance along I-40 from the I-40/I-840 split point on the west to the remerging point on the east is 59.3 miles. The distance along I-840 in parallel is 78.5 miles. Despite I-40's 20-mile difference, TDOT had previously stationed an exit sign encouraging travelers to use I-840E for Knoxville, TN (see **Figure I-3a**), even though staying on I-40 would also get to Knoxville. However, in 2021, that exit sign was replaced by a new dynamic message sign (DMS), whose electronic display reports the current travel times on I-40E and I-840E to Lebanon, TN (see **Figure I-3b**) just a few miles east of the remerging point on the other side of Nashville. For both cases, the location of each sign is a few miles before the I-40E/I-840E split point as shown in **Figure I-4**.

It should be noted that the DMS in **Figure I-3b** had not yet been placed during the study reported in [Han 2019]. Any travel time comparisons were consequently unable to be drawn from that DMS device for the purpose of the study. More on the data will be presented in the following sections.



(a) before 2021 (b) after 2021

Figure I-3. TDOT signage before I-40E/I-840E split (Google Street View)

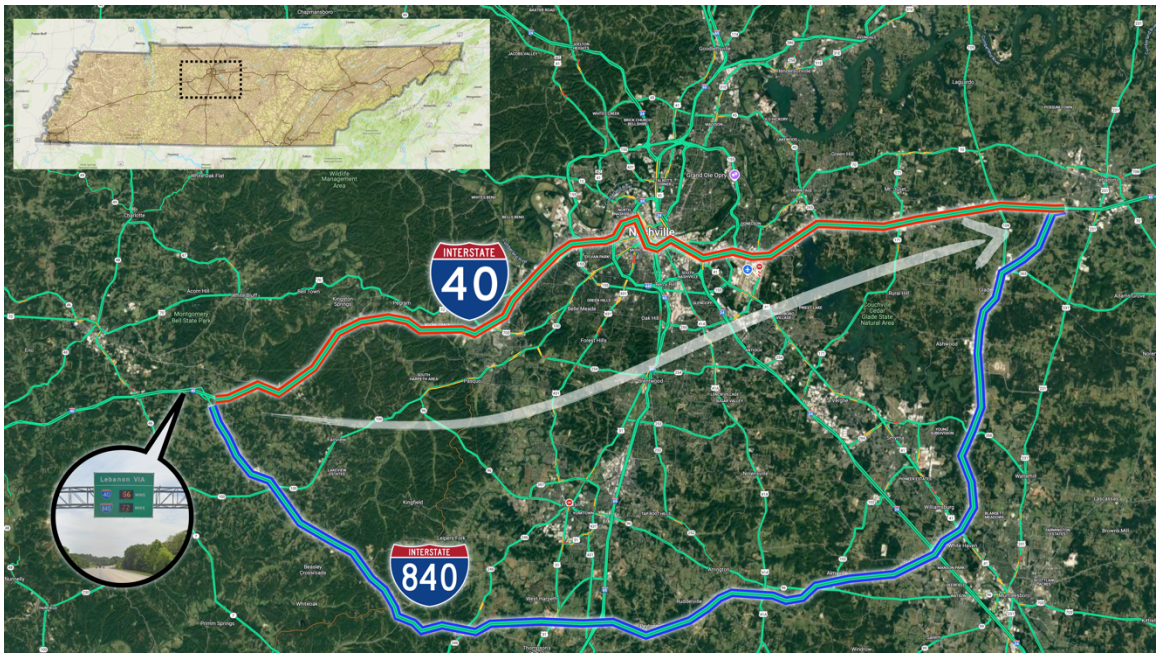


Figure I-4. Nashville study site with through traffic via I-40E and bypass via I-840E

High-Resolution Traffic Data

A 2017 study [Han 19] by the University of Tennessee (UT) documented the attempt to collect high-resolution spatiotemporal traffic data using Tennessee Department of Transportation (TDOT) radar detection stations (RDS), National Performance Management Research Data Set (NPMRDS), and ALPR equipment. While TDOT's RDS data provides the highest temporal resolution at thirty-second intervals with lane-by-lane speed, flow, and occupancy information, it does not provide full coverage of the study site, which can be seen in **Figure I-5**. The background travel time information for the study is based on NPMRDS instead, as it covers the entire study site while using a very short-term probe-based data source.

Figure I-6 compares sample travel time with aggregated NPMRDS segment travel times. This 30-day comparison of eastbound traffic travel times on I-40E (red) and I-840E (blue) at different times of day yields the following observations.

- I-40E travel time is typically around 55 minutes with an AM peak and a longer and more severe PM peak on weekdays. Weekend traffic is better with less delay most of the day.
- I-840E travel time is typically below 80 minutes without significant evidence of AM or PM peaks.
- There are times when peak traffic conditions on I-40E could push the route travel time over that on I-840E, sometimes significantly so. For those cases, using I-840E would reduce delay.

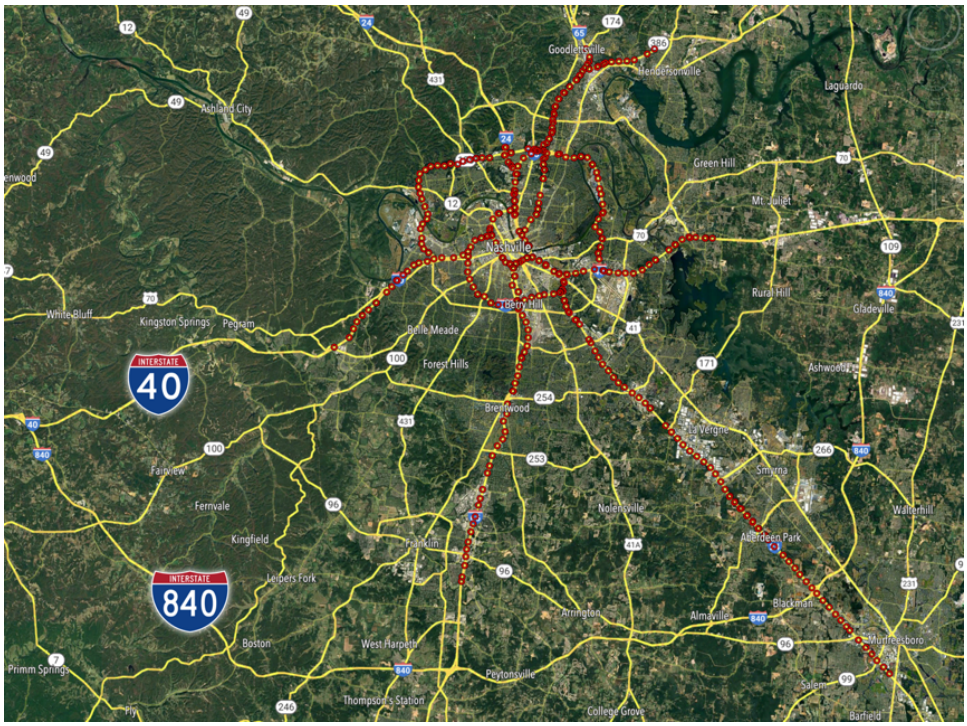


Figure I-5. TDOT Region III (Nashville) RDS station locations in 2016

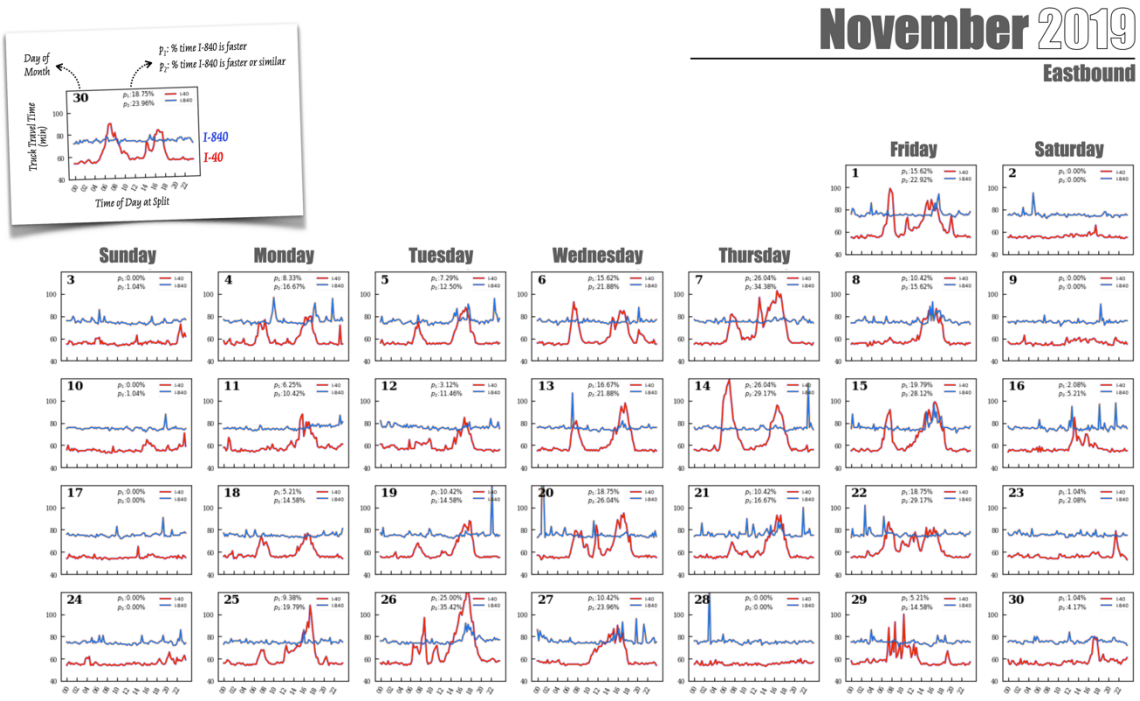


Figure I-6. Sample travel time comparisons on I-40E and I-840E based on NPMRDS

- Thanksgiving Day saw very light traffic and low travel time for both routes. The travel time patterns during the AM and PM peaks in I-40 are somewhat
- expected but not exactly the same every day or precisely predictable for the purpose of forecasting which of the two routes might enjoy shorter travel time. This information on travel time on the two routes was not available to the drivers, via the yet-to-be-installed DMS sign, or even to the researchers in the field during the 2017 study.

For the specific day of the field study, Monday, May 8, 2017, as reported in [Han 2019], the NPMRDS data were not available back then but have been made available now and plotted in **Figure I-7**. Note that while the travel time values are for the entire distance from the I-40E/I-80E split on the west side all the way to the east side where the two routes merge again, they were calculated by aggregating individual segment travel times and not based on trips completed from one end to the other by any specific vehicles. So, the travel time values here assume that all vehicles traveled continuously without stopping.

Like the previous observations made during one month in 2019, this field data collection day in 2017 saw a relatively quiet I-840E with an average travel time of around 70 minutes, which translates into about 67 mph on average. I-40E saw relatively quiet Monday traffic, like some of the days in November 2019, with a tame AM peak and a more pronounced PM peak when the travel time on I-40E exceeded that on I-80E for a couple of hours.

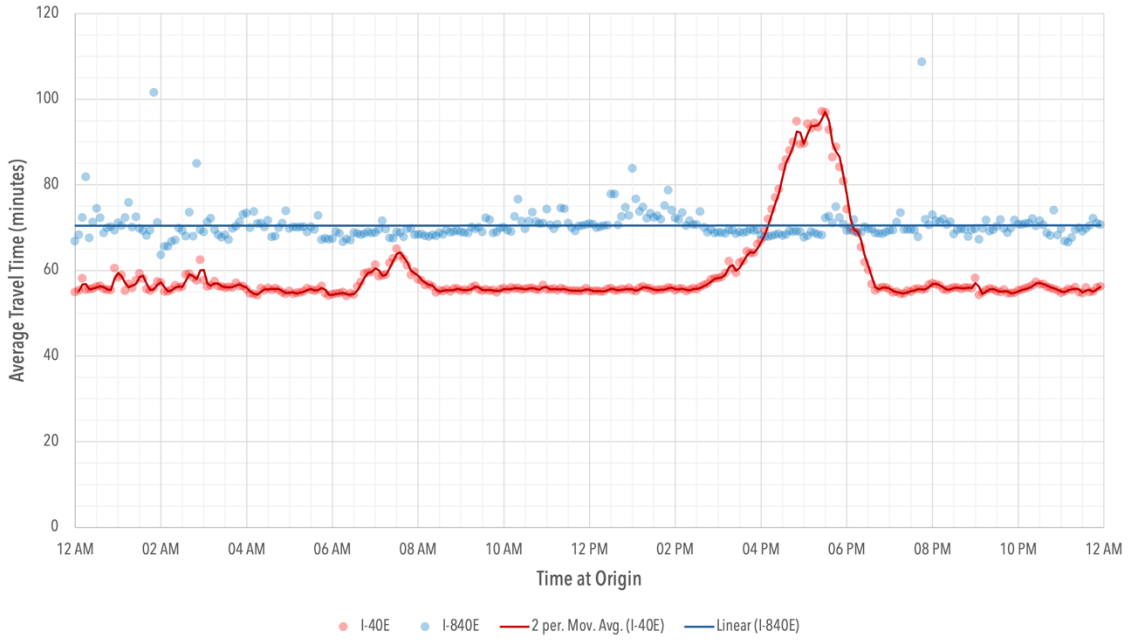


Figure I-7. Travel time on I-40E and I-840E at the split point based on NPMRDS

ALPR Large Truck Tracking Data

The large truck tracking study was conducted on May 8, 2017, using ALPR devices set up by well-trained UT researchers at three strategic locations from roughly 6 AM to 6 PM for the consideration of available daylight and field safety [Han 2019]. Each ALPR station covered two truck lanes. Two stations were placed, one on I-40E and another on I-840E, such that both were just past the split point of the two roadways, with the third station located on I-40E to the east of Nashville just past the merging point of the two roadways (see **Figure I-8**).

The field-collected ALPR plate information was subsequently processed for plate-matching analysis using the previously tested matching algorithm [Oliveira-Neto 2013]. The plate-matching statistics are shown in **Table I-1**. Both routes saw reasonable portions of truck traffic traversing the entire route, be it I-40E or I-840E, from the east end to the west end of the study area. The researchers noticed a decaying pattern of journey time for trips made on both routes, which can be seen in **Figure I-9**.

They also plotted the relationships of mean journey time and number of trucks caught by ALPR units at both ends vs their time at the origin, aggregated to an hourly basis, as seen in **Figure I-10**. Some of these trends are real and can be insightful, while others are artifacts due to the way the data were collected. However, the latter was not obvious at the time of the study and was not explored further. It is important, in hindsight, to recognize their nature so the conclusions are not biased or misleading.

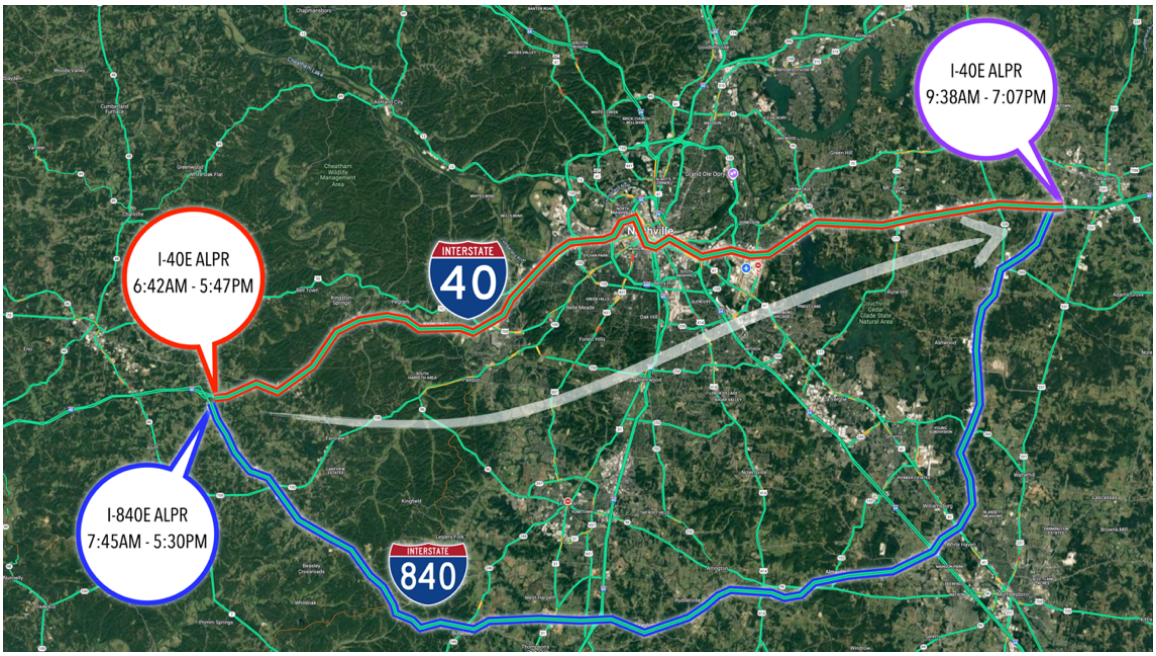


Figure I-8. Truck tracking study with ALPR units at three sites

Table I-1. Field ALPR data collection results.

Truck Path	Distance (miles)	# of Plates Read at Starting ALPR	# of Plates Read at Ending ALPR	# of Matches	Matching % at Starting Node
I-40E	59.3	2,089	2,298	602	28.9%
I-840E	78.5	568	2,298	287	50.5%

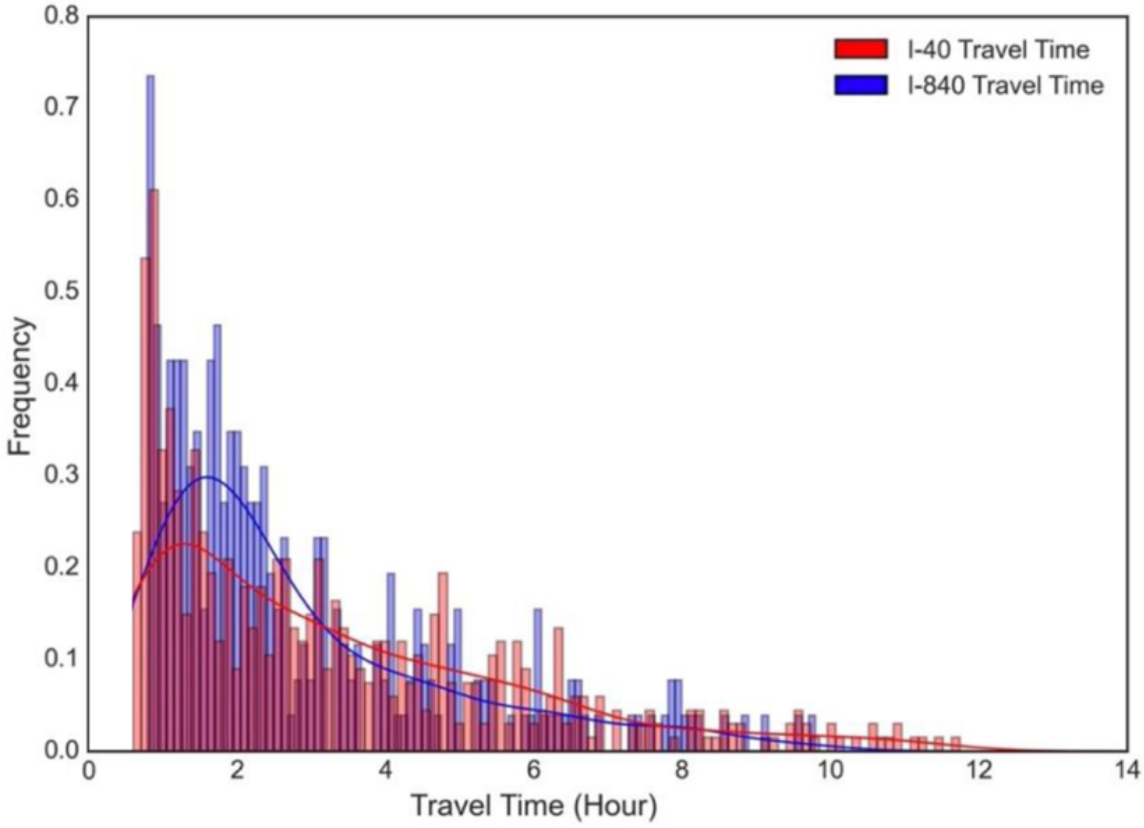


Figure I-9. Decaying trend of journey time distribution on both routes [Han 2019]

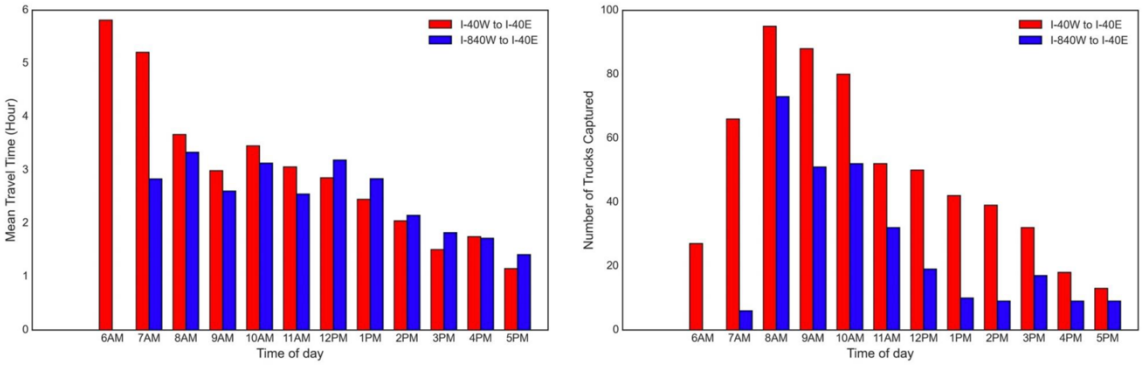


Figure I-10. Mean journey time and number of trucks matched plotted against the departure time at the origin [Han 2019]

ASSESSMENT OF LIMITATIONS AND MISSING DATA SCENARIOS

Limitations of NPMRDS or Other Traffic Data

Despite having a dependable supply of high-resolution data from TDOT's RDS detectors and from NPMRDS, we find that these are point- or segment-based data that do not offer any trip or path information needed for this study. Trip times shown in **Figure I-6** and **Figure I-7** are based on the total sum of the average travel times along all segments, which tend to ignore any durations of time when trucks are off the Interstate for purposes including refueling, dropping off or picking up loads, dining, sleeping, and taking a break in general. In the context of a short-distance study, these activities are uncommon, but for longer-distance tracking studies like this, they turn out to be a rather significant issue to consider. In addition, trucks typically cannot be uniquely identified within these data, which makes them unsuitable as the primary source for this study.

ALPR Large Truck Tracking Data

ALPR-based plate-matching and other tracking technologies can provide trip or path-tracking information. However, this requires data to be collected continuously at each site to avoid truncation issues, which may result in artifacts in the data that could bias or obscure the findings. For example, the ALPR statistics plotted in **Figure I-10a** seem to suggest a decreasing trend in mean travel time on both routes as the day went on, see **Figure I-11a**. At the same time, one could argue that the mean travel time values, ranging from 1 to 6 hours for I-40E and 1.5 to 3.5 hours for I-840E, are significantly higher than the mean travel time of about 55 and 70 minutes, respectively, per NPMRDS in **Figure I-7**.

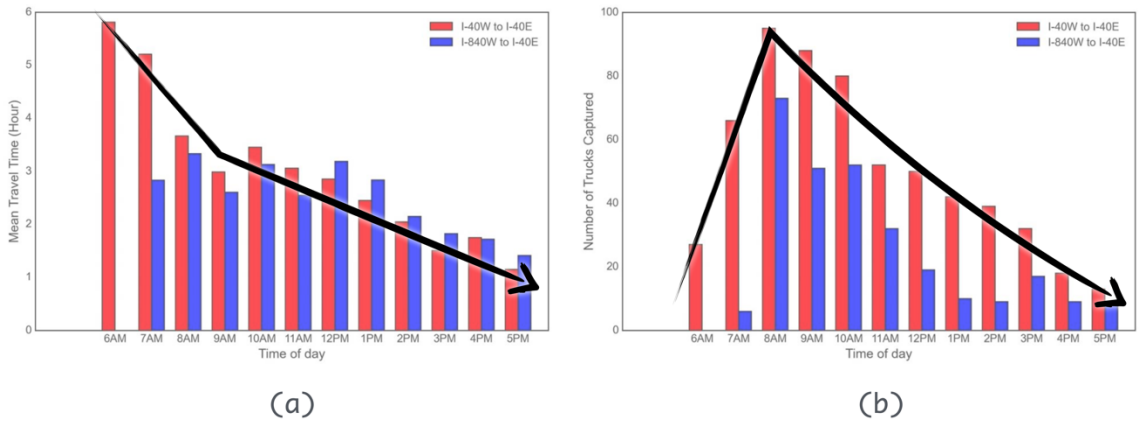


Figure I-11. Trends seemingly present in **Figure I-10**

Likewise, the ALPR statistics plotted in Figure I-10b indicate a rapidly increasing trend in the early morning and a steadily decreasing trend thereafter in the number of truck plates captured and matched at the beginning and the ending ALPR stations on both routes (**Figure I-11b**).

Upon combining these two figures, one might wonder if truck drivers prefer to travel in the morning, even though the average travel time for these trucks is much longer. However, one must also consider that their decision-making is vastly different from that of those who do not necessarily drive trucks. The choices that long-haul commercial truck drivers make are influenced by the fact that they may traverse thousands of miles each week, occasionally from coast to coast. As such, they do not typically arrange their trip in such a way that they would get through Nashville at a particular time of the day. Findings by a Federal Highway Administration (FHWA) study on vehicle volume distributions by classification indicate that straight-through large truck volume typically stays rather steady throughout the day [Hallenbeck]. So, is the ALPR data erroneous in some way?

There are several considerations here. First, the average travel time calculated using NPMRDS is suitable for all vehicles that traverse the entire route without stopping. But the ALPR plate-matching-based trip or path travel time is the more legitimate time duration that could realistically include stops, interruptions, or even side trips between the origin, where an ALPR first captured the truck's license plate, and the destination, where a second ALPR captured the same plate again. Some trucks that took well over 55 or 70 minutes to traverse I-40E or I-840E, respectively, can be attributed to this method. Acknowledging these stops and side trips is important for further studies of the economic activities of these long-haul trucks and is crucial to freight and logistical analysis.

The second but significantly more crucial issue concerns the potential misinterpretation or biases that may result from missing data, particularly the data absent in a latent manner. When plotting the ALPR data of the journey time of individual trucks based on the time of day each of the trucks passed the origin, or the first ALPR station, see

Figure I-12, one can see some blank areas with no data for both cases. For instance, there is no recorded truck passing the origin at around 6:45 AM that traversed I-40E in less than 3 hours. On the other hand, all trucks passing the origin at around 5:45 PM traversed I-40 in less than 1.5 hours. This is because the late starting time and the early termination time of the data collection effort at the ALPR station truncated the data stream that would have filled those blank areas. Huge gaps in what could have been data are responsible for studies falling into "survivorship bias" when drawing conclusions.

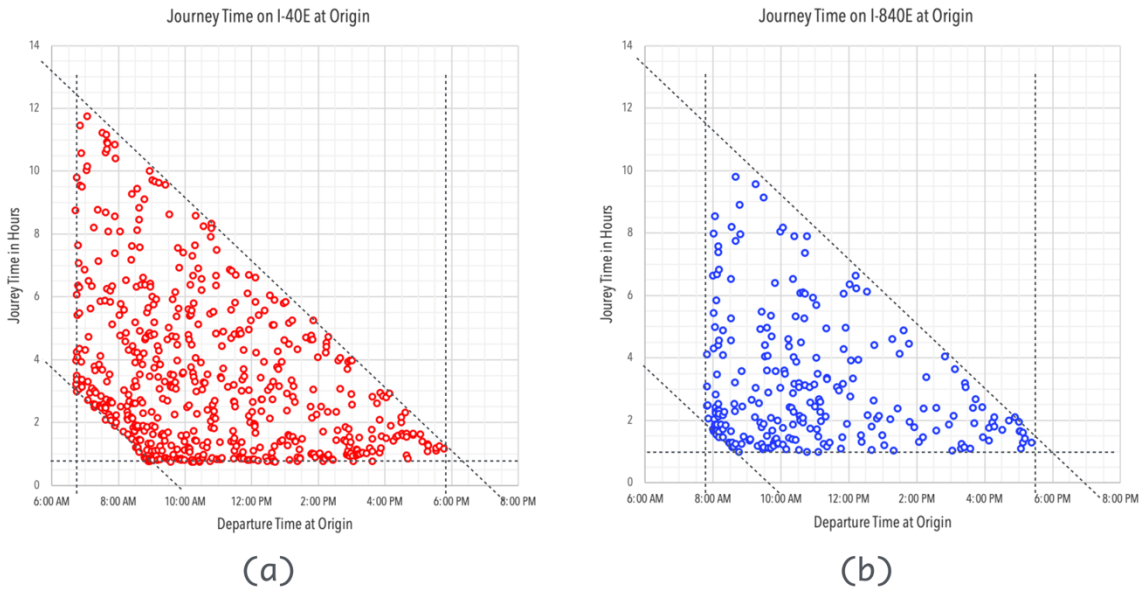


Figure I-12. Individual truck journey time vs time at Origin for I-40E and I-840E

This phenomenon was first observed when engineers noticed the bullet holes on the returning aircraft in World War II (WWII), see **Figure I-13**. Incorrect estimations of journey time and number of trucks traversing one route or the other could result if one did not realize that the missing bullet holes, or the missing trucks, hold important insights about the problem at hand. Further analysis of the latent and patent missing data helps classify these cases into nine scenarios that could affect the findings (see **Figure I-14**).

Scenario 1: Early Stop of Data Collection at Destination – Probably the most obvious, though ironically latent, and most damaging scenario of missing data comes from prematurely stopping the data collection effort at the destination station and not allowing all trucks to complete their journeys within the shrinking time window. Trucks passing the origin station early in the day are afforded nearly 12 hours to complete the trip and still be included in the collected dataset. Each passing hour later in the day saw a steadily decreasing window for trucks to complete the journey and still be counted. Trucks

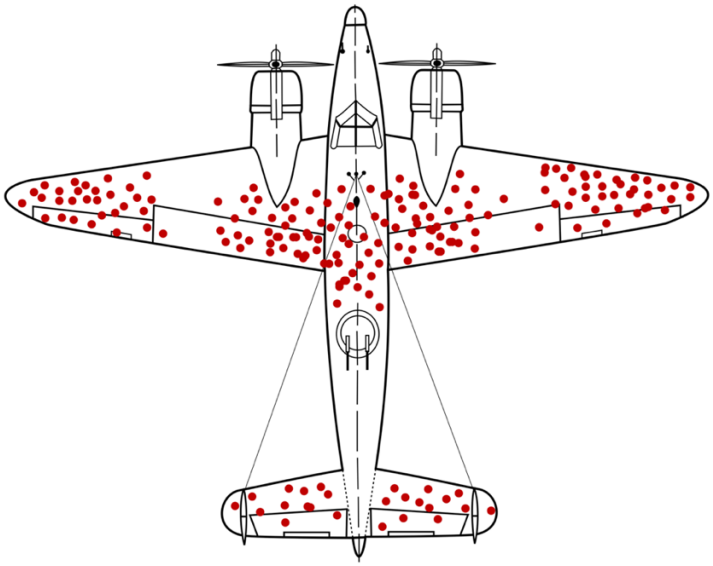


Figure I-13. Hypothetical patterns of bullet holes on surviving aircraft during WWII

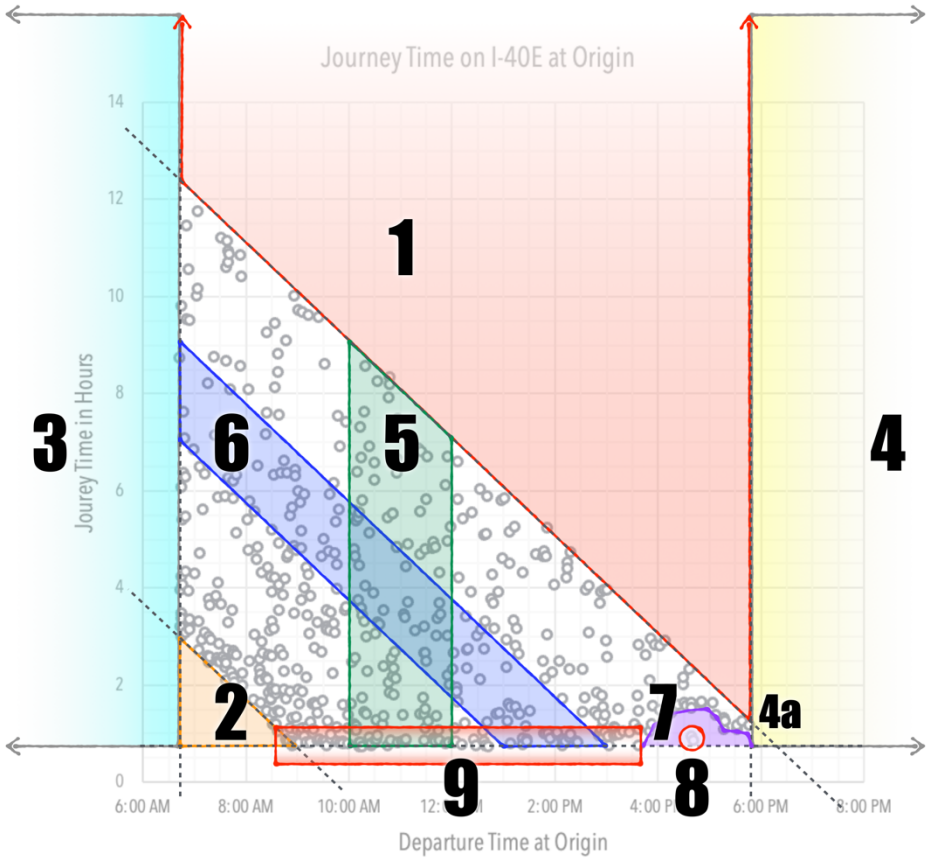


Figure I-14. Missing data scenarios, as seen in a journey time vs time at the origin plot

passing the origin after 5:30 PM would not be included in the study if they did not complete the trip before 7:07 PM, which is most of them. The resultant decreasing journey time was evidenced in Figure I-11a. This scenario is very challenging to overcome because the ALPR station at the destination needs to stay open for 10 hours or more to capture trucks with long journey times like those captured early in the day. This type of missing data would be difficult to avoid unless the ALPR devices were continuously operating instead of a temporary research setup. The concept of this scenario can also be visualized using a time-space diagram like **Figure I-15**.

Scenario 2: Late Start of Data Collection at Destination – The orange triangle to the lower-left of **Figure I-14** is another case of trucks not captured at the destination, resulting from a late start of the data collection activities. Faster vehicles captured at the origin node had traversed the entirety of the route and arrived at the destination before the ALPR devices started recording. In such cases, the lost opportunity resulted in no faster trucks captured and a decided bias towards a longer average journey time, as also evidenced on the left side of **Figure I-11a**, where the already incorrectly depicted reducing average journey time was made even more pronounced by the truncated shorter journeys arriving too early to be counted.

A portion of the late start loss is shown as "2a" in **Figure I-15** and **Figure I-16**. This portion of the loss could be reduced or even avoided by starting the data collection effort earlier at the destination.

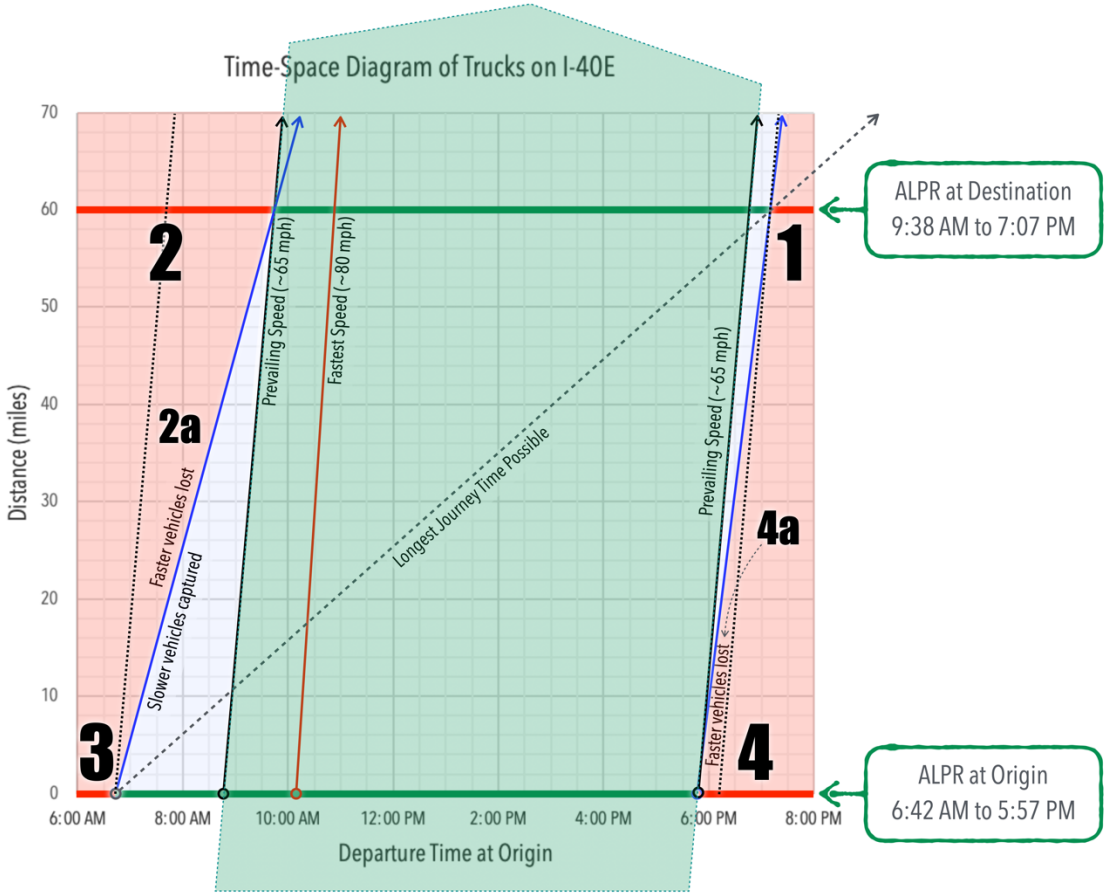


Figure I-15. Time-space diagram of data collection windows at the origin and the destinations along I-40E, and several scenarios of trucks not captured

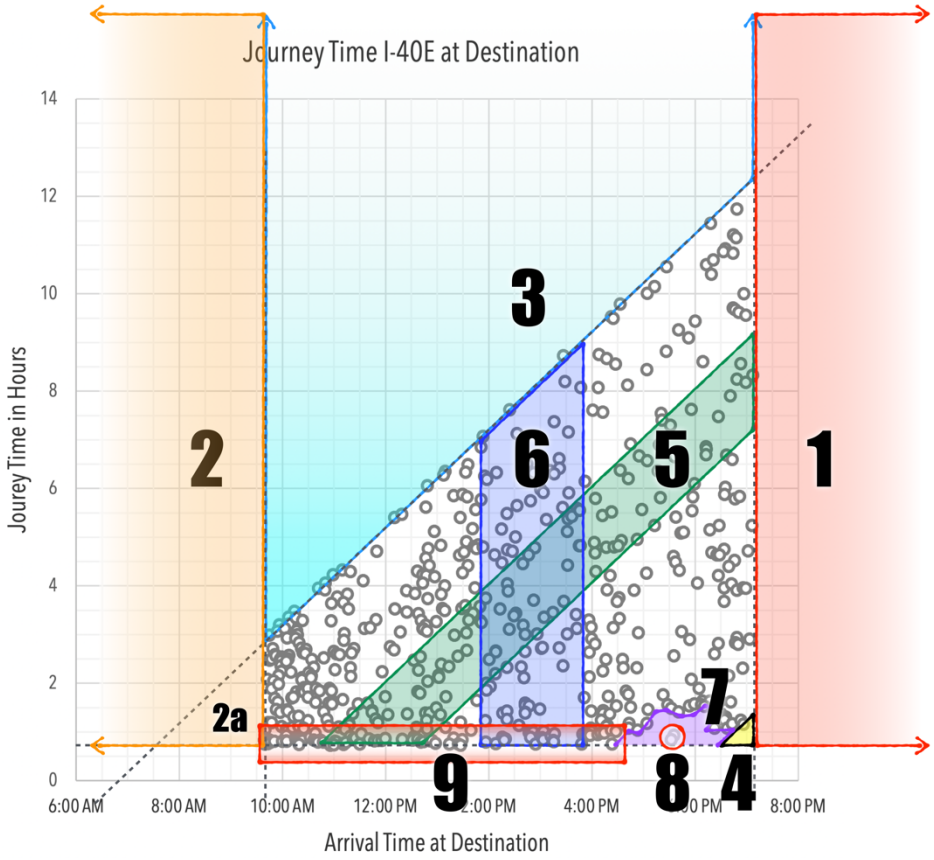


Figure I-16. Missing data scenarios, in a journey time vs time at the destination plot

Scenario 3: Late Start of Data Collection at Origin – Late start at the origin reduces the magnitude of the longest possible journey time to be collected in the matched data set. Obviously, the sooner the data collection effort begins at the origin, the more data can be collected. Without coordinated effort at the destination, however, this would inadvertently increase the size of the 2a data loss zone.

Scenario 4: Early Stop of Data Collection at Origin – Early stop at the origin does not reduce the magnitude of the longest possible journey time to be collected. It reduces the size of the collection window to cover variations in traffic conditions later in the day. A small portion of the Scenario 4 loss is labeled as 4a in **Figure I-14** and in **Figure I-15** which can be reduced if the data collection activities at the origin lasted longer.

Scenario 5: Station Outage at Origin – In such cases, very noticeable signs of data loss would show for any sizeable station outage at the origin. A blank vertical strip in **Figure I-14** and a blank 45° strip in **Figure I-16** would signify patent data loss. For this example, a two-hour outage, from 10 AM to 12 PM, at the origin station would lead to the size and shape of the strips in those two figures.

Scenario 6: Station Outage at Destination – Similarly, any significant station outage at the destination would also lead to a very noticeable blank -45° strip in **Figure I-14** and a blank vertical strip like the one in **Figure I-16**. For this example, a two-hour outage, from 1:50 PM to 3:50 PM, at the destination would lead to the size and shape of the strips in those two figures.

Fortunately, neither Scenario 5 nor Scenario 6 occurred during the field data collection effort, which is evident in these plots.

Scenario 7: Congestion, Incident, or Delay in Traffic – A lack of data in the lowest journey time region, an example of which can be seen in **Figure I-14**, is not really a case of missing data but all vehicles taking longer to complete the journey for a certain period of time due to recurring or nonrecurring congestion along the route. This increase in minimum travel time due to traffic conditions would render the figure's lowest portion blank. The shape and duration of this could be corroborated with the route travel time from traffic data such as NPMRDS and RDS, see **Figure I-7**, which shows a matched duration when travel time was elevated.

Scenario 8: Error in Matching Data – The plate matching was performed using an ML algorithm tested to be between 95% and 98% accurate [Oliviera 2013]. That means a very small portion of individual truck data may be matched incorrectly, mostly due to a combination of incorrect readings by ALPR devices and a self-learned association matrix always having a margin of error. For example, because of the elevated minimal travel time in Scenario 7, data showing up within that otherwise blank zone is likely incorrect, see **Figure I-14**. In fact, any data point showing up in the blank zones representing the missing scenarios for the aforementioned data could very likely be an error.

Scenario 9: Clock Synchronization Issues – When traffic data needs to be collected at multiple locations and then matched to derive speed or travel time statistics, it is essential that the system clock on all devices is synchronized beforehand and verified again afterward. This can be accomplished readily if GPS signals are used on these devices. A sign of the existence of such errors can be spotted when the lowest portion of journey time in a plot like **Figure I-14** shows a large number of data points that would indicate

very high average speed, e.g., greater than 80 mph, or very low average speed. If synchronization before the fact was not possible, at least one should take note of the differences in clock time on each device and then adjust or correct the time stamp in the data file after the fact.

DATA IMPUTATION FOR FURTHER STUDIES

Imputation to Restore Missing Data Statistics

Based on the assessment in the previous section, the next step is to reduce errors and biases by accounting for the missing data, primarily due to Scenarios 1 and 2. To that end, the available truck journey data for I-40E are aggregated at 15-minute intervals, as seen in **Figure I-17**, proportionally scaled to include zones where data were lost due to late start and early stop at the destination. This yields 46 data points, each representing a journey time interval with a percentage of data falling into that interval. A cumulative distribution function can be plotted, see **Figure I-18**. Regressions were conducted to determine a suitable function to represent the cumulative curve. The best fit comes from a shifted logarithmic function in the form of:

$$F(x) = a \cdot log(x + b) + c \tag{1.1}$$

where x is journey time, a and c are regression coefficients, and the shifted value b is equal to $1 - t_{min}$, where t_{min}, set to 0.73 hours, is the assumed minimal journey time for this data set. The regression results are shown in **Figure I-19**. Given this shifted logarithmic function, the resultant cumulative model is $F(x)$, and its derivative is $f(x)$:

$$F(x) = 0.3753 \cdot log(x + 0.27) - 0.001036 \tag{1.2}$$

$$f(x) = \frac{0.3753}{x + 0.27} \tag{1.3}$$

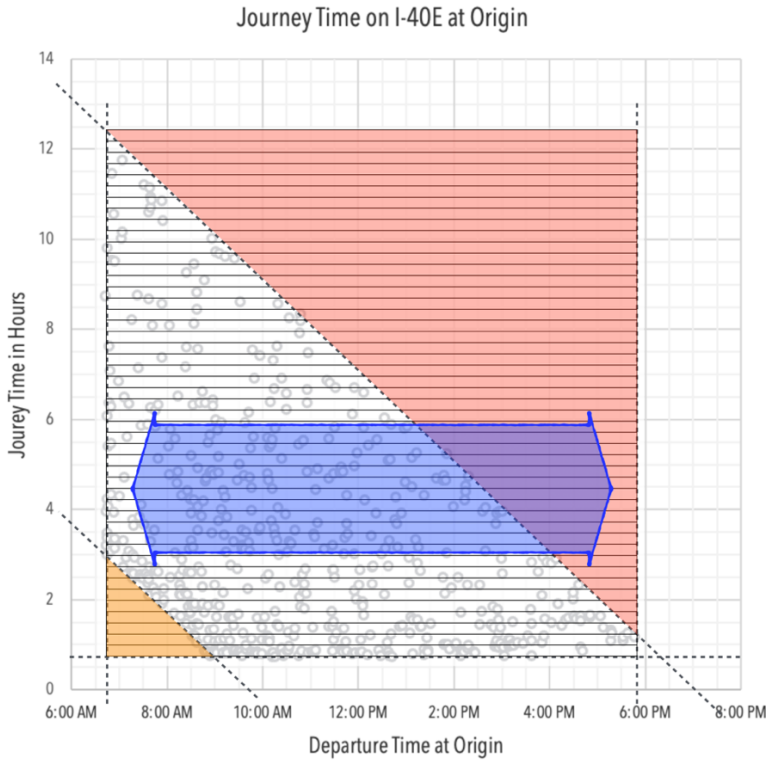


Figure I-17. General imputation for data loss due to Scenarios 1 and 2

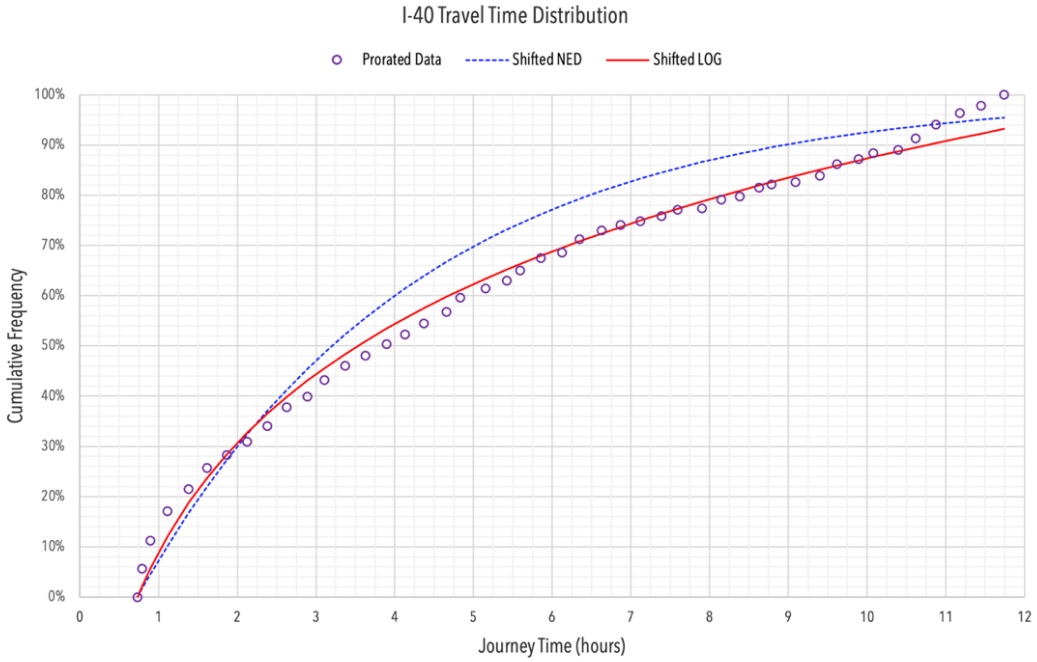


Figure I-18. Cumulative percentage of all journey times and curve fitting results

Regression Statistics	
Multiple R	0.995246118
R Square	0.990514834
Adjusted R Squa	0.990299262
Standard Error	0.026121415
Observations	46

ANOVA

	df	SS	MS	F	Significance F
Regression	1	3.135177592	3.135177592	4594.822542	3.7559E-46
Residual	44	0.030022447	0.000682328		
Total	45	3.165200038			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	-0.001036101	0.009878064	-0.104889072	0.916940526	-0.020944032	0.01887183	-0.020944032	0.01887183
X Variable 1	0.375311467	0.005536782	67.78512036	3.7559E-46	0.364152815	0.386470119	0.364152815	0.386470119

Figure I-19. Regression summary of the shifted LOG model

This best-fitting function with an R^2 value exceeding 0.99 is plotted in **Figure I-18** and labeled as the "Shifted LOG" curve. However, one concern with this function is that it does not converge as x increases, and $P(x)$ would exceed 100% at the same time. Though the model fits the data extremely well, this concern makes the function unsuitable for our purpose unless we truncate the journey time at a point where $P(x)$ reaches 100%. Since the imputation challenge at hand is already severely truncated data, the shifted logarithmic function does not help the cause.

Another function that fits the cumulative distribution exhibited by the ALPR matching data is the shifted negative exponential distribution (NED). NED is a decaying function that is the prevailing distribution for modeling the time until the next event. To account for the minimal travel time, the NED function, which is defined between 0 and ∞, should be shifted by t_{min}. The cumulative distribution function is as follows:

$$G(x) = 1 - e^{-\lambda(x - t_{min})} \tag{1.4}$$

where t_{min} equals 0.73 hours, or about 44 minutes, for I-40E, and λ equals 0.28 based on the regression results in **Figure I-20**. The cumulative distribution function can then be expressed as such:

$$G(x) = 1 - e^{-0.28(x-0.73)} \tag{1.5}$$

and the probability density function is:

$$g(x) = 0.28e^{-0.28(x-0.73)} \tag{1.6}$$

The shifted NED has a decent R^2 value of 0.76, even though it doesn't fit the data as well as the shifted logarithmic function does, as is evidenced in **Figure I-18** and **Figure I-19.** Nevertheless, the NED model is typically used for temporal decay problems and allows convergence to 100%, which is essential for the cumulative distribution function. The NED model also has a more ideal expected value of 1/λ, or 1/λ+t_{min} for a shifted NED, making the model a superior choice for this study.

Regression Statistics	
Multiple R	0.874209095
R Square	0.764241542
Adjusted R Squa	0.74201932
Standard Error	0.98130807
Observations	46

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	140.4709802	140.4709802	145.8733218	1.4519E-15
Residual	45	43.33344874	0.962965527		
Total	46	183.804429			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	0	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
X Variable 1	0.280258103	0.023204395	12.07780285	1.02097E-15	0.233522053	0.326994152	0.233522053	0.326994152

Figure I-20. Regression summary of the shifted NED model

Missing data scenarios 1 and 2 led to a declining trend, see **Figure I-11a**, and an underestimated average journey time of 3.27 hours [Han 2019]. The imputed data and the shifted NED model would instead give us an average journey time of 4.30 hours. Because of the significant number of truck trips lost to scenario 1, it is expected that the naïve mean of all the individual journey times present would underestimate the true average, which is calculated by imputing the truck trip data. Since we do not have the ground truth where the lost data were known in detail, Monte Carlo simulations were employed to verify the results.

Verification with Monte Carlo Simulations

With the parameters associated with I-40E known, we can employ Monte Carlo simulations to generate truck trips based on these parameters and visualize the results of individual trucks completing the entire route via I-40E (see **Figure I-21**). The trips captured at both the origin and the destination, and hence present in the simulated database, are represented by the red dots. The blue dots represent those not captured at both ends of the route, hence lost or missing. One can observe missing data scenarios 1, 2, 2a, 3, 4, and 4a in these plots.

The average data loss rate based on repeated simulations is 31.29%, with a standard deviation of 1.23%. Maintaining the shifted-NED parameter λ at the value of 0.28 according to the regression results in **Figure I-20,** which also means an expected average journey time of 4.30 hours, we can compare the average journey time of the "red dots," based on the simulations, against the field data of 3.37 hours. Again, repeated simulations were performed to yield an average naïve journey time of 3.12 hours, or

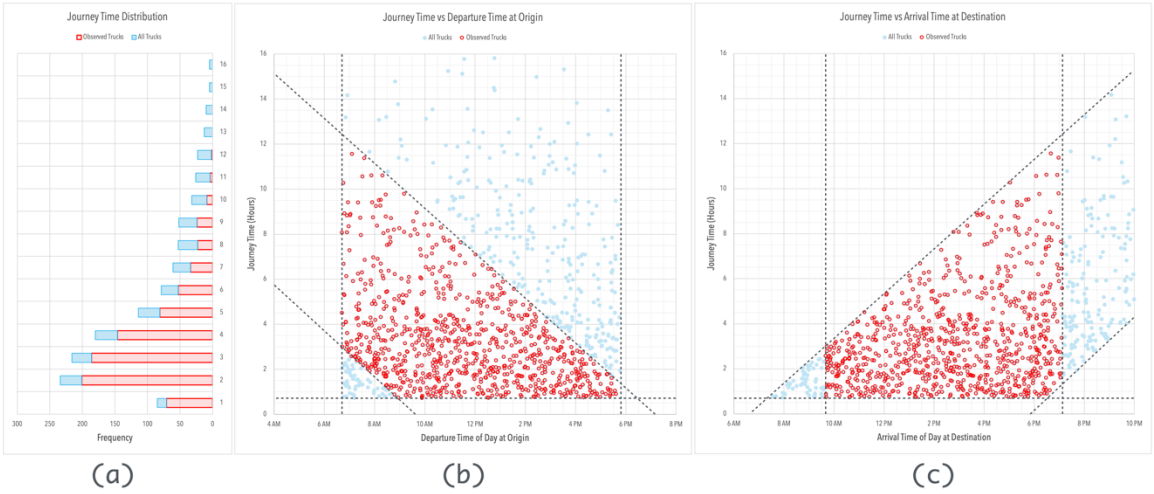


Figure I-21. Visualization of outcomes of one of the runs simulating I-40E truck traffic

about 3 hours and 7 minutes, with a standard deviation of 0.07 hours, or about 4 minutes. The mean absolute percentage error (MAPE) is 4.79%, while the mean absolute error (MAE) is 0.15 hour, or about 9 minutes. These are decent endorsements of the shifted-NED model.

Incidentally, the Monte Carlo simulations were used to simulate and visualize missing data in Scenario 5, see **Figure I-22**, where the ALPR station at the origin was out from 10 AM to 12 PM, and Scenario 6, see **Figure I-23**, where the ALPR station at the destination was out from 1:45 PM to 3:45 PM.

These results provide a promising and easy venue to "cut and paste" the missing data generated by the Monte Carlo simulations into the blank areas to supplement the field data and form an infinite number of hybrid datasets.

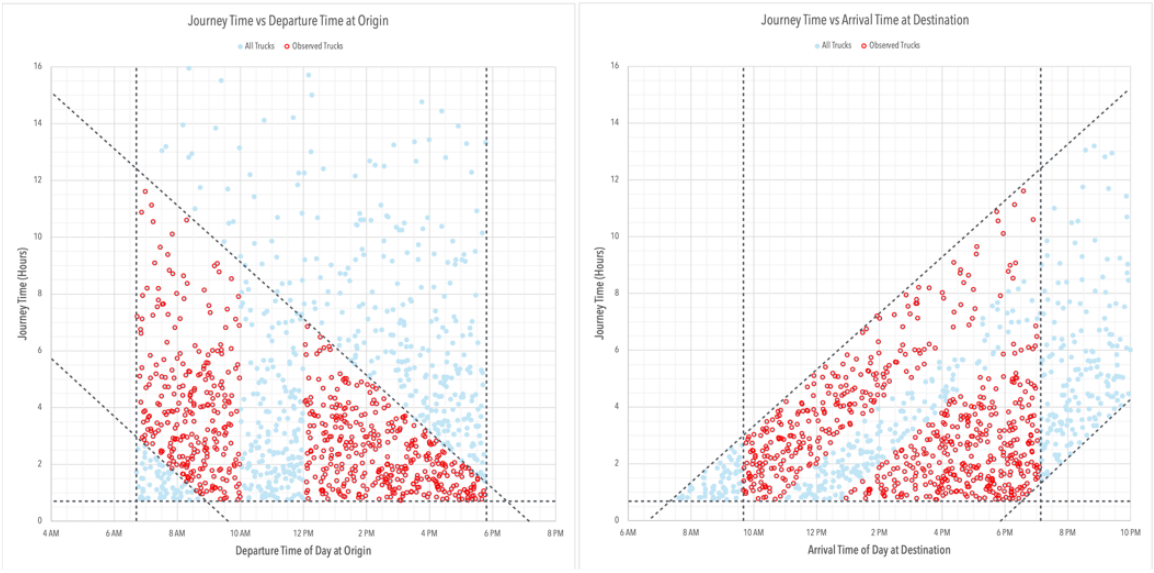


Figure I-22. Visualization of outcomes of one of the runs simulating I-40E truck traffic

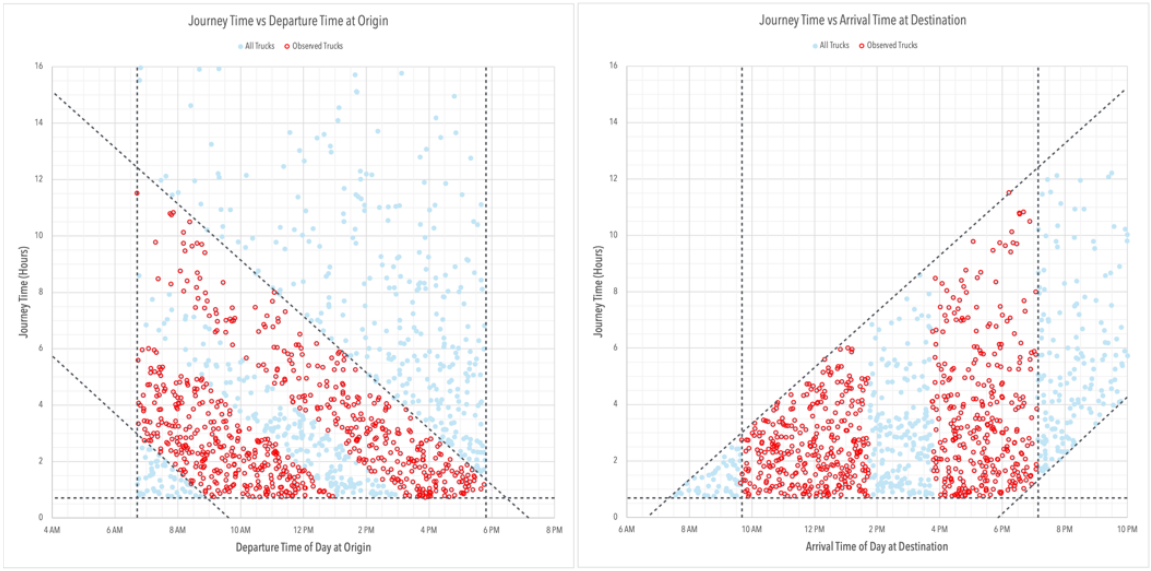


Figure I-23. Visualization of outcomes of one of the runs simulating I-40E truck traffic

Simulation-Based Imputation for a Hybrid Dataset

The use of Monte Carlo simulations to backfill the blank spaces, i.e., data missing due to various scenarios presented previously, is a promising approach to imputation. By supplementing the actual field data (solid red dots) with simulated (virtual) trips (hollow dots) in **Figure I-24**, a hybrid dataset can be assembled for various analyses that would not be biased. Four randomly selected hybrid datasets are shown in **Figure I-24**, and many more could and should be generated for subsequent analyses, as should be the case when Monte Carlo simulations are used.

Statistics like those shown in **Figure I-10** were generated for the four sample hybrid datasets in **Figure I-24** to compare mean journey time and the number of truck trips captured on I-40E (see **Figure I-25**). The solid red bars are based on the field data alone, which exhibits a decreasing trend of mean journey time as well as the number of truck trips captured over time, while the lighter-colored striped bars are based on the imputed hybrid data, which suggests a more steady mean journey time throughout the day as well as more truck trips captured during the later part of the day by comparison.
It should be noted that the first and the last bars in the number of trucks captured vs time at the origin plot in **Figure I-10** and **Figure I-25** (right side of the doublet plots) are misleadingly low. This is because only 18 minutes (6:42 – 7:00 AM) of data were collected in that first hour, and only 47 minutes (5:00 – 5:47 PM) of data were collected in the final hour at the origin. So, a normalized trip rate (trucks per hour) would be far less misleading than the number of trucks captured. A comparison of this is shown in **Figure I-26**.



Figure I-24. Sample imputed hybrid data: field data plus simulated missing data



Figure I-25. Comparisons of sample imputed hybrid statistics vs field data statistics

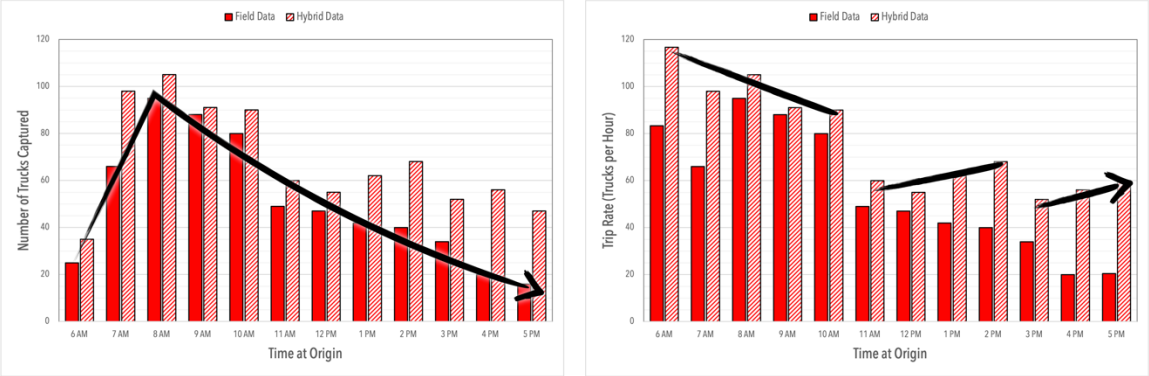


Figure I-26. Comparisons of the number of trips and trip rates vs the time at the origin

In summary, the original raw dataset, shown as green crosses in **Figure I-27**, exhibited a noticeable and significant data loss that led to a substantial underestimation of the number of trucks traversing the I-40E corridor and a biased mean journey time for these trucks. A prorated expansion of the raw data can approximate the amount of data lost due to various scenarios that caused data truncation and form an aggregated dataset for a better estimation of journey time distribution, shown as red dots in **Figure I-27**. The shifted NED model, dashed blue line, is the most suitable function for fitting these data and for the subsequent imputation efforts.

Because of the stochastic nature of the raw data, as well as the missing data, presumably following the same stochastic distribution, Monte Carlo simulations were employed to verify the imputation functions, as presented in the previous section, and to generate sample imputed missing data to form a series of hybrid datasets, shown as the purple dots in the plot.

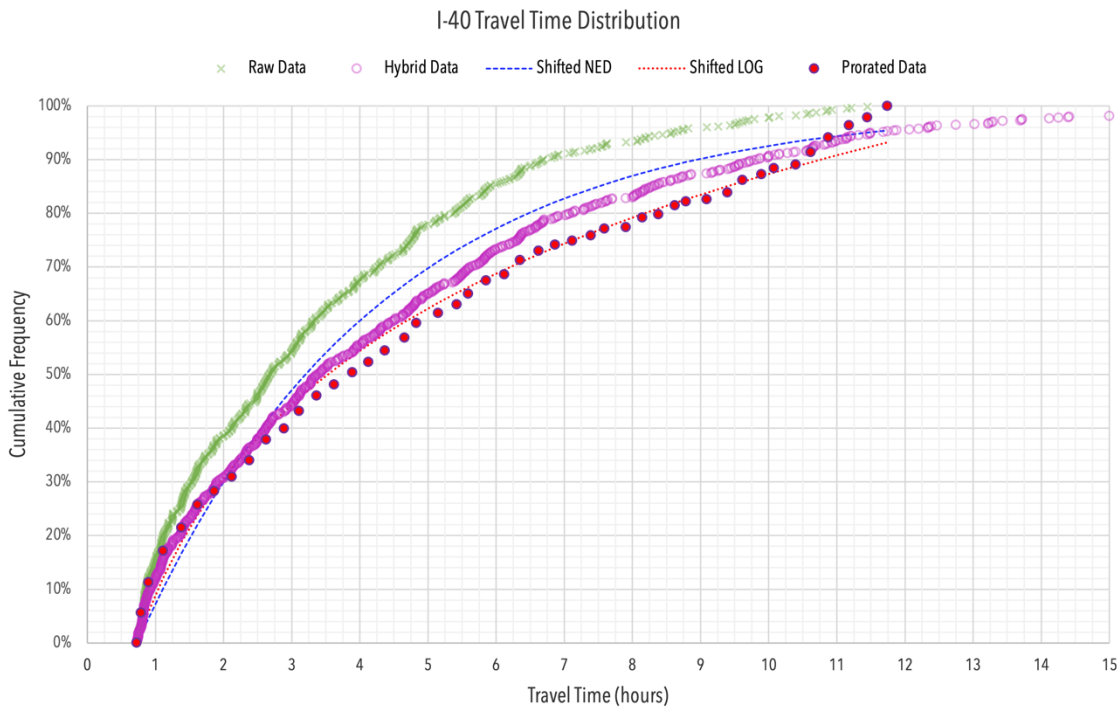


Figure I-27. Comparisons of the raw and imputed hybrid data of truck trips at the origin

CONCLUSIONS

The research documented in this paper sought to provide a procedure for imputing a common but not often acknowledged loss of data when tracking vehicles or events over an extended time. The following is a summary of the thought process and actions taken to arrive at the completion and contributions of the study.

- We observed unexplained trends in aggregated truck-tracking statistics and recognized these as the artifacts of a significant loss of data.
- We estimated a loss of nearly 1/3 of the overall data due primarily to data truncation when two or more data collection time windows interfered or were not in sync with each other so that only trips captured at both locations within the respective time

windows were preserved. Everything else was lost and unaccounted for in the statistics.

- We generalized these into nine data loss scenarios, including late start and early end at the destination node, late start and early end at the origin node, station outages at the origin or the destination nodes, and other compounded scenarios. The first two account for most of the data loss in the case study.
- We aggregated and expanded the raw data to assess the stochastic distribution of the data, which would serve as the foundation for the imputation efforts.
- We found the shifted-NED model to be the most suitable for our purpose because it provides a decent fit to the proportionally expanded raw data, and it is considered the prevailing model for decaying frequency functions.
- We verified the shifted NED model by using iterative Monte Carlo simulations to derive the average journey time in the raw data reversely. The MAPE is 4.79%, and the MAE is less than 9 minutes.
- We subsequently employed Monte Carlo simulations to generate multiple sets of missing data, which were fused with the field data to create a series of hybrid datasets. These datasets are now "complete" and can be used for further analyses beyond this study.
- We observed the misleading trend in the decreasing journey time caused by the truncated data, which is now corrected by a more stable trend with normal stochastic fluctuations in each imputed hybrid dataset.

- We also observed that the declining number of truck trips over time is now replaced by a more reasonable trend of truck trips per hour.
- This study does not intend to perform a full analysis of driver choices since the Dynamic Message Sign with route travel time information was not installed until four years after the field data was collected. However, by designing and carrying out this elaborate imputation process, this kind of driver behavior analysis is now possible.

REFERENCES

1. Urban Transport Fact Book, Wendell Cox Consultancy, 2005: https://www.publicpurpose.com/hwy-intmiles.htm (accessed March 2025)
2. International Fuel Tax Agreement, 1998
3. Oliveira-Neto, F.M., L.D. Han, and M.K. Jeong (2009) "Tracking Large Trucks in Real Time with License Plate Recognition and Text-Mining Techniques," Transportation Research Record 2121(1), 121-127.
4. Oliveira-Neto, F.M., L.D. Han, and M.K. Jeong (2012) "Online License Plate Matching Procedures using License Plate Recognition Machines and New Weighted Edit Distance," Transportation Research Part C, Emerging Technologies 21(1), 306-320.
5. Oliveira-Neto, L.D. Han, and M.K. Jeong (2013) "An Online Self-Learning Algorithm for License Plate Matching," IEEE Transactions on Intelligent Transportation Systems 14(4), 1806-1816.
6. Hargrove, S.R., H. Lim, and L.D. Han (2016a) "Expanding License Plate Matching Capabilities with Secondary Self-Learning Algorithm," Transportation Research Record 2594(1), 51-60.
7. Hargrove, S.R., H. Lim, L.D. Han, and P.B. Freeze (2016b) "Empirical Evaluation of the Accuracy of Technologies for Measuring Average Speed in Real Time," Transportation Research Record 2594(1), 73-82.

8. Han, L. D., Z. Zhang, S. Hargrove, and M. Burton (2019) Automated Plate Recognition and Truck Trip Tracking. RES2016-32 Project Report. Tennessee. Department of Transportation.
9. Hallenbeck, M.E., M. Rice, B. Smith, C. Cornell-Martinez, and J. Wilkinson (1997) Vehicle Volume Distributions by Classification, FHWA-PL-97-025, US DOT, Federal Highway Administration.

CHAPTER II

LIKELY SHORTEST PATH AMONG VIABLE OPTIONS IN A SQUARE GRID WITH STOCHASTIC TRAVEL TIME

This chapter is being revised for publication in a refereed journal.

ABSTRACT

In this work, we address the inherent stochasticity of roadway travel times and challenge the limitations of traditional shortest-path algorithms that rely on deterministic assumptions. Such algorithms, while effective in hindsight, falter in predictive contexts where travel times are uncertain and variable. Recognizing this, we reframed the shortest-path (SP) problem as a stochastic challenge, demanding methods that consider the probabilistic nature of link travel times and their effects on overall path selection.

A key insight emerged from our observation that paths within a network are rarely independent; they often share common links. This interdependence varies based on network structure and the extent of link-sharing among competing paths. To capture this nuanced relationship, we introduced the concept of *Fractional Independence* and formalized it through a novel metric we call the *Degree of Independence (DoI)*. This measure quantifies how independent a given path is by accounting for the number of paths that share its constituent links.

To evaluate the implications of stochastic travel times across all viable paths, we utilized extensive Monte Carlo simulations. Rather than rely on heuristic algorithms with static parameters, we conducted millions of simulations, enabling us to derive converged probabilities for each path being the shortest in a given network. These simulations spanned various grid configurations, including both symmetrical and asymmetrical rectangular layouts.

A breakthrough came with our discovery of a strong correlation between a path's deterministic *DoI* and its likelihood of being the shortest in a stochastic setting. This finding offers a foundational step toward predicting the most probable shortest path without requiring knowledge of actual future conditions. To express this predictive advantage, we introduced the *Relative Frequency (RF)* metric, which compares the success rate of the most likely path against the average path. In practical terms, for an 8x8 grid, the most probable shortest path is about 10 times more likely to be optimal than any random path; for a 10x10 grid, this advantage surges to approximately 36 times.

To generalize this relationship, we developed a near-perfect regression model (R^2 = 0.9964) that links *RF* values with the grid dimensions (*m, n*). This model allows us to extrapolate *RF* values for larger grid sizes where exhaustive simulation is computationally infeasible, significantly broadening the practical utility of our findings.

Ultimately, our research transforms the previously intractable stochastic SP problem into a manageable deterministic prediction task. By leveraging the structural tendencies of grid networks, we show that the shortest path is most likely to traverse the outer edges of the grid. This marks a substantial advancement in predicting optimal paths under uncertainty, grounded in both probabilistic rigor and network theory.

INTRODUCTION

Once considered very challenging, the Shortest Path (SP) problem and its practical applications are now used by millions daily thanks to efficient algorithms, such as Dijkstra's algorithm [Dijkstra] and the A* algorithm [Hart], the superior calculation capabilities of modern computers, and the wide availability of cellphones. However, the so-called shortest paths recommended to the user, typically in real-time, are often not really the shortest by the time the user follows the path from the origin to the destination. This is not due to any deficiency in the heuristic algorithms themselves, but because one could not foretell the future travel condition and must formulate and solve the SP problem based on deterministic or even stationary link travel time values obtained in the past. In the real world, the travel time fluctuates continuously on each link in the road network. Whichever path a particular heuristic algorithm might declare as the solution to the deterministic SP problem is likely the perfect shortest path for the wrong problem. That is, while the network geography stayed the same, the travel time on each link would be different from the original values used by the SP algorithm initially. These travel times will continue to change as the traveler traverses along this recommended path, or any path for that matter, from the origin to the destination. This is the nature of the stochastic SP problem.

Example 1 – Shortest Path Observations in Midtown Manhattan, NYC

For about a month in late 2024, we monitored and recorded the shortest paths recommended by Google Maps for a pair of origin and destination (OD), both located in midtown New York City. The OD pair is selected based on the expansive rectangular

grid roadway network between the origin, at the intersection of 2nd Ave and 57th St. in midtown Manhattan, and the destination, in Chelsea at the intersection of 9th Ave and W 23rd St (see **Figure II-1**). This rectangular grid includes roughly 35 east-west streets and 10 north-south avenues. Many of these roadways are one-way pairs, so the actual viable routes in the general direction for the OD pair are 20 E-W streets and 7 N-S avenues, which affords a solution space of over 800,000 solutions, backtracking not considered.

Google Maps was employed for these observations to take advantage of the near-real-time travel information they continue to amass via cellphone tracking. Once a query is submitted for an OD pair, Google Maps reports up to three best routes recommended for this trip for whichever the desired transportation mode, be it driving, transit, walking, biking, or even flying. For this study, driving will be the mode of choice unless otherwise specified. Here in **Figure II-1**, the highlighted path that takes an estimated 17 minutes was found to be the "fastest route" and recommended by Google Maps at 06:56:24 AM on December 3, 2024. Two other paths are also provided as good alternatives.

We continued to monitor the recommended path between the same OD pair for a time interval comparable to the time it would take to traverse the path and captured 20 maps showing the recommended shortest path at different times sequentially in **Figure II-2**. During the time it would take the traveler to traverse the initially recommended shortest path travel time, about 18 minutes, see **Figure II-2a**, the recommended shortest path changed multiple times, some with mid-way detours, while others flipped back and forth from one side of the network to the other. All these changes happened once per

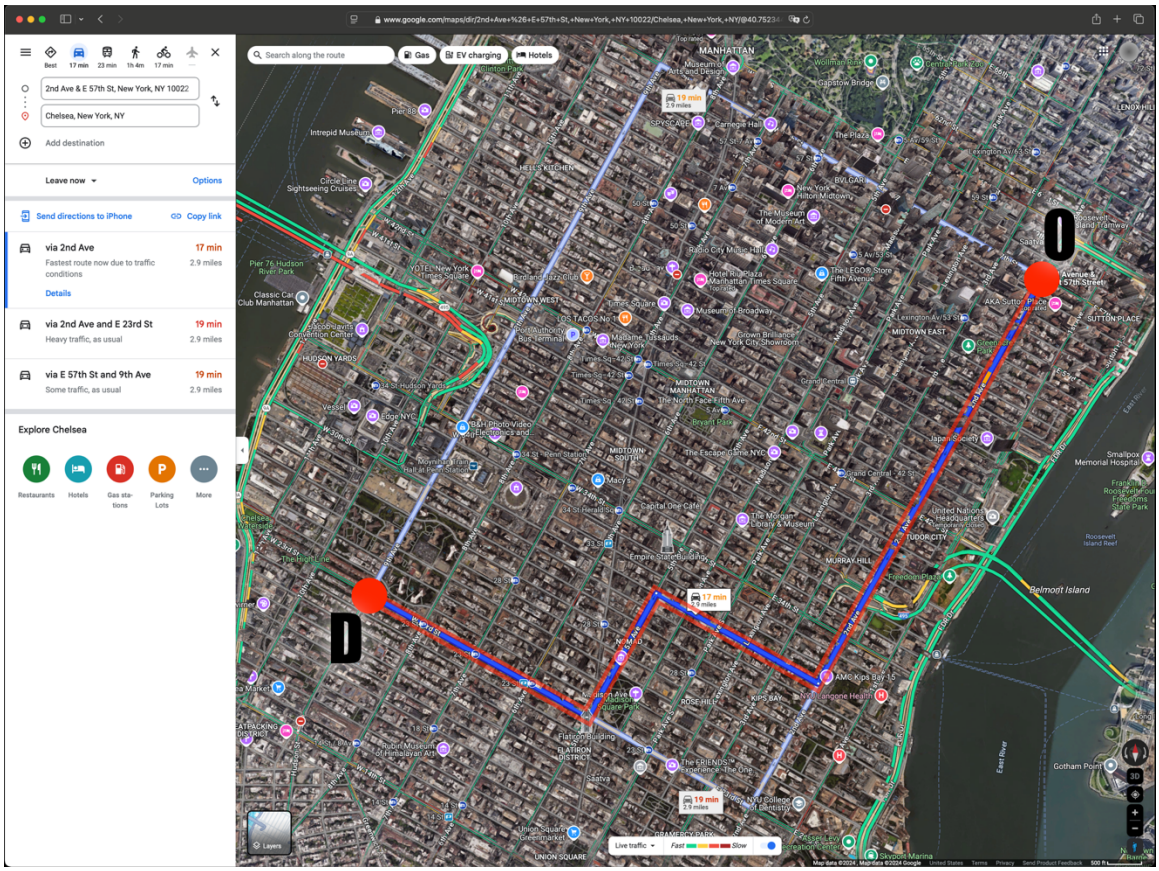


Figure II-1. Observation site and OD pair in Midtown Manhattan. (Google Maps, 06:56:24 AM on 12/03/2024)

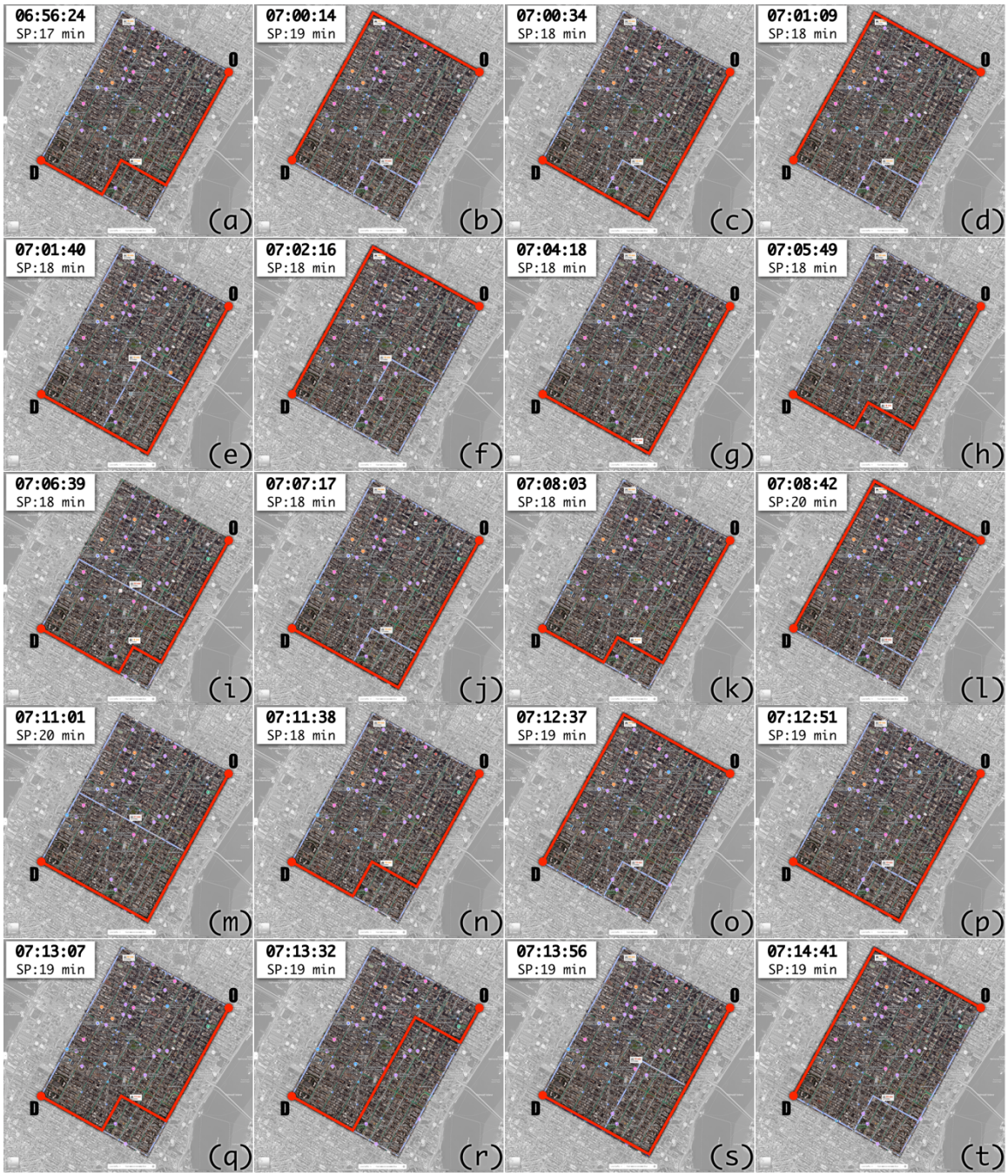


Figure II-2. Shortest paths for the same OD suggested by Google Maps within 20 minutes

minute on average and sometimes more frequently. A note about the real shortest path is that no one can be certain what the true shortest path would be unless one knows the travel time continuously on every link throughout the entire duration. This information is impossible to know before the trip begins and is ignored, by and large, after the trip has ended. In a network with a large number of competing paths with similar lengths and travel times between an origin-destination (OD) pair, the path reported to be the shortest is typically the bona fide solution to the shortest path problem based on an array of stationary link travel times obtained from the past, e.g., the average time for (probe) vehicles to traverse each link or the link length divided by the average vehicle speed within the past few minutes, on each link in the network. These travel time values are typically obtained from deployed technologies (e.g., GPS tracking and traffic sensors) or historical traffic records of the same time of day and day of week previously to populate the array of travel time for all links in the network. This deterministic, or even stationary, travel time assumption works serviceably well under conditions such as:

- The link travel time fluctuation is minuscule or nonexistent.
- No incidents, vehicular accidents, special events, inclement weather, or other factors would affect the deterministic assumption of the traffic conditions on all pertinent network links.
- The actual shortest path is singularly and decidedly much faster than all its competitors, such that the stochastic fluctuations and data sampling errors in travel time would be unlikely to affect the resultant SP choice.

- It is imperative to have a routing solution promptly, especially when a slight, or not so slight, excess in travel time over the actual, yet unknown, SP, can be overlooked. In other words, whether the shortest recommended path is not crucial.
- The traveler only has moderate expectations and does not demand certainty that the solution is the actual SP as long as the route gets them to the desired destination in a reasonable time.

In the real world, travel time is not stationary or deterministic. So, one has to consider the effect of this stochastic travel time on the shortest problem.

Example 2 – Deterministic and Stochastic Travel Times

Consider a simple case of three entirely separate and independent paths, not overlapping and with no shared link, from an origin to a destination. For the first scenario, all path travel times, TT_1, TT_2, and TT_3, are constant or stationary, though different (see **Figure II-3.a**). Any serviceable shortest path algorithm would be able to identify the shortest path among the three, path 1 for this case, since TT_1 is less than TT_2 and TT_3, in no time. This is the realm where the popular shortest-path algorithms excel and, in reality, how most SP problems are solved: by substituting the stochastic travel time for the links and, hence, paths to a constant but somewhat reasonable travel time.

For the second scenario, where the travel times on these paths are stochastic, which path specifically would come out as the shortest is not as obvious. Each probability density function (PDF) curve in **Figure II-3.b** represents the probability of a path having a particular travel time. If these curves were far apart and hardly overlapping, then the

$$P(TT_1 \leq TT_2, TT_3) \tag{2.1}$$

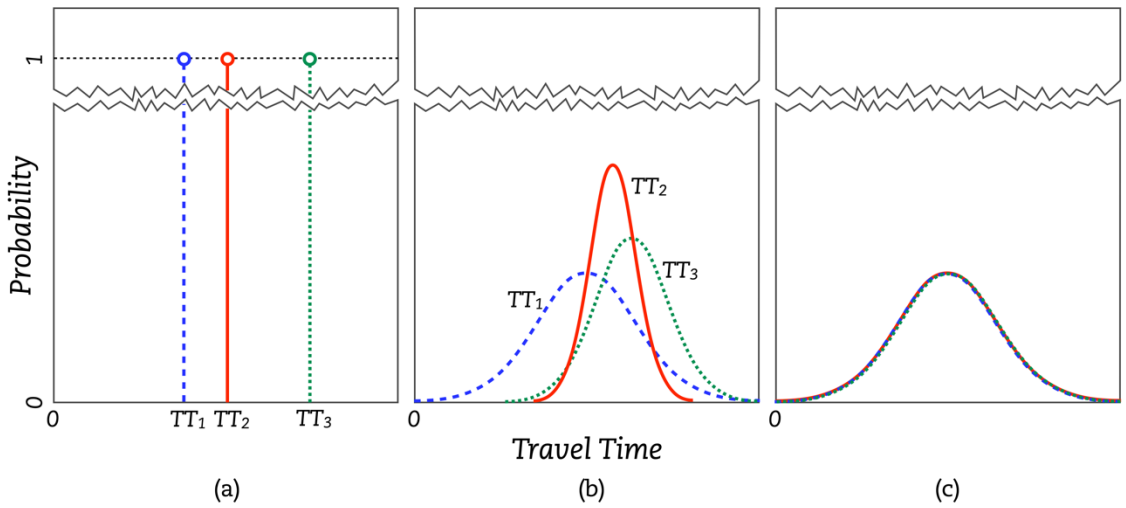


Figure II-3. Travel time scenarios for three independent paths.

PDF on the far left, with the lowest travel time, would most likely come out as the shortest path. For the more common case where the three curves, representing the PDFs of the travel times on the three independent paths, overlap with significant portions, as shown in **Figure II-3.b.** One could analyze the probability of a path, say path 1, being the shortest path by:

If the specific PDFs were known as, say, g_1, g_2, and g_3, for the three paths, the likelihood of path 1 winning can be calculated as:

$$\int_0^{\infty}\int_x^{\infty}\int_x^{\infty} g_1(x)g_2(y)g_3(z)dzdydx \qquad (2.2)$$

where x, y, and z are random variables representing the possible path travel times for TT_1, TT_2, and TT_3, respectively.

For the third scenario, where the three paths have virtually identical travel time distributions (see **Figure II-3.c**), each path is equally likely to end up as the shortest path.

So, the probability of any of the three paths being the shortest path is 1/3, or 1 over the number of these independent paths with identical PDFs.

Note that once we enter the realm of stochastic travel time on links and paths, the exact solution for the SP problem is not deterministic. The probability of a certain path being the eventual winner can be analyzed and understood, but the win is usually not guaranteed. Again, the shortest path algorithms can still be applied when each link's specific travel time value is given. However, unless the algorithms are implemented repeatedly and iteratively throughout the process and the link travel times are updated and accurately predicted for all of the foreseeable future, a critical yet impossible assumption, the initial solution may be irrelevant to the actual and eventual solution.

The likelihood consideration is more complicated for most practical cases, where viable competitive paths are many and mostly not independent. This is the key challenge that this paper seeks to address.

Example 3 – Competing Paths with Shared Links

Consider another example in a simplified rectangular grid network (see **Figure II-4**), where one would like to travel from node A to node I via one of the three available paths: 1) ADGHI in red, 2) ABEFI in green, and 3) ABCFI in blue. The travel times on all links in the network are independent and identically distributed (*iid*) random variables. This, however, does not necessarily mean the path travel times are independent of one another, like the second example above. For this example, paths 2 and 3 share links AB and FI and are, consequently, not independent of each other.

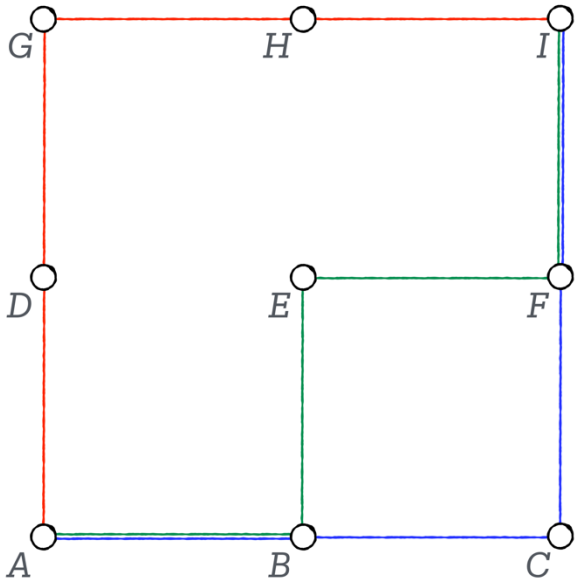


Figure II-4. Three paths on a simplified rectangular grid.

The likelihood of each path being the shortest can be analyzed if the link travel time PDFs are given. For the example here, particularly with the link travel time being identically distributed, Monte Carlo type of simulations could be employed [Metropolis].

Let's assume the links all have a travel time random variable, x, that follows the probability density function $f(x)$ uniformly distributed between a and b, where $a < b$. So, $f(x)$ would have a value of $1/(b\text{-}a)$ in the range of $[a, b]$ and 0 outside of that range. Now, if $a = 0.5$ and $b = 1.5$, in the unit of minutes or some other time units, or even general cost values. The PDF in question is then in the form

$$f(x) = \begin{matrix} 1, & if\ 0.5 \le x \le 1.5 \\ 0, & Otherwise \end{matrix} \tag{2.3}$$

While the three path travel times, TT_1, TT_2, and TT_3, would follow identical probability density functions, much like **Figure II-3c**, the likelihood each path may be the shortest is not the same 1/3. In fact, repeated Monte Carlo simulations would show that Path 1 would be the shortest around 40% of the time, while Paths 2 and 3 each may

win 30% of the time (see **Figure II-5**). After hundreds of trial runs, Path 1 begins to emerge as the favorite. After thousands of trials, the likelihood of all three paths would begin to stabilize.

It should be noted that changes in the values of a and b in $f(x)$ would not alter the results of this converging winning percentage for each path. Even the stochastic function employed, be it uniform, Gaussian, or others, would not alter the results.

Problem Statement

Alternate paths between the same OD pair may be correlated even when all the individual links that make up these paths are independent of one another. The degrees of correlations among competing paths can complicate the shortest path search process and lead to unexpected results where some paths are more likely to win the shortest path race than others, even when they have identical mean travel times. This study seeks to recognize the existence of such an intriguing phenomenon, understand and explain its

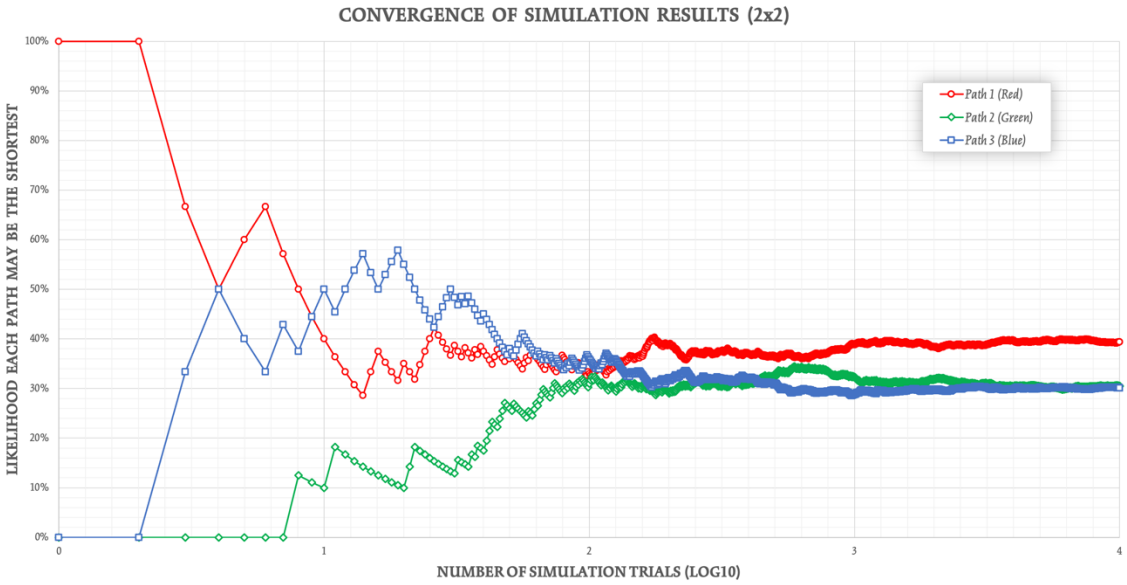


Figure II-5. Converging probability of each path winning the shortest path race

reason, form a theory for analytically predicting the likelihood that any path might be the shortest, and test this theory. The purpose of this study is not to come up with a solution to a particular shortest path problem with a fixed set of parameters. Rather, we seek to develop a sound and tested theory to help "intuit" the solution to large-scale stochastic shortest-path problems.

ANALYSES

The prevailing shortest path algorithms, such as Dijkstra's algorithm and A* algorithm, have been widely studied and used. While these heuristic algorithms are considered efficient and accurate, they are only as good as the travel time information they rely on is a part of the input parameters. For realistic scenarios where the travel time continues to fluctuate, and the future is not deterministic, with or without the presence of incidents, the likelihood of a path deemed as the shortest by these algorithms based on a fixed set of travel times before the trip commences, is most likely to have a low to the infinitesimal probability of being the correct solution. This paper then shifts the focus from the traditional shortest path algorithms to the likelihood that a path may be the shortest under stochastic travel time conditions.

Basic Terminology and Assumptions

To remain consistent throughout this paper, we will use **node**, as opposed to vertex or point, and **link**, as opposed to edge or arc in our terminology, since all these terms are widely used, accepted, and interchangeable. The collection of nodes connected by links then forms a **network**, or graph, through which various kinds of traffic could travel. A unique sequence of nodes and links connecting an **origin** node and a

destination node, or OD, forms a **path**, or route, or tour. We will favor the terminology of link **travel time** over similar but different measures such as edge weight, impedance, cost, or generalized cost as a link and path attribute, which one must overcome to traverse a link or a path.

We make the following assumptions to scope this work for simplicity and clarity purposes.

- The network is a two-dimensional rectangular grid [Fawaz 1976, Newell 1996]. The findings herein can be extended to higher dimensions, nevertheless.
- The nodes in the network can be origins, destinations, or means for connecting links. They do not incur any travel time or cost.
- The links in the network represent connectivity between nodes. They have non-negative travel times. Unless otherwise specified, the travel time on all links in a network is identically and independently distributed (*iid*) random variables.
- A viable path is a series of links providing accessibility from an origin (O) to a destination (D). The path would have a stochastic path travel time, which is the sum of the individual link travel times along that path.
- No backtracking is allowed or considered for the shortest path problem.
- Links have no capacity constraints. So, while link travel times are stochastic, they are not affected by the traffic on these links or the number of paths sharing or going through them.

Paths and Shortest Paths on a Rectangular Grid

In a vast rectangular grid network, each node is connected to four other nodes through direct links, and all nodes are accessible to one another either directly or through a series of links. To make the network easy to represent in this paper, each link would have a unit length, and the location of each node can be represented by a pair of integers (x, y), much like the Cartesian coordinate system, where both x and y are integers. Then, finding the shortest path from a node at (x_1, y_1) to a destination (x_2, y_2) is equivalent to finding the shortest path from (0, 0) to $(x_2 - x_1, y_2 - y_1)$ in a new coordinate system, where the starting node (x_1, y_1) is transformed to the origin (0, 0) of the new coordinate system, see **Figure II-6**. With the assumption of no backtracking, all viable paths between (0, 0) and $(x_2 - x_1, y_2 - y_1)$ would have the exact same length of $|x_2 - x_1| + |y_2 - y_1|$. By defining $m = |x_2 - x_1|$ and $n = |y_2 - y_1|$, both m and n are non-negative integers. The notations hereafter is thus simplified.

Not only are the notations simplified to make it a shortest path problem going from (0, 0), the origin, to (m, n), the destination, but the problem itself is also simplified to a search of the shortest among a finite number of paths going from the origin to the destination. Specifically, if all nodes and links are present and viable, there would exist $C(m+n, n)$ or $(m+n)!/m!/n!$ competing paths going from (0,0) to (m, n). **Table II-1** summarizes the numbers of nodes, links, possible paths, and links per path under a non-backtracking scenario for the rectangular area between (0,0) and (m, n). For cases with reasonably small m and n values, instead of employing shortest path algorithms of a heuristic nature, one could calculate the travel time for every path to find the shortest.

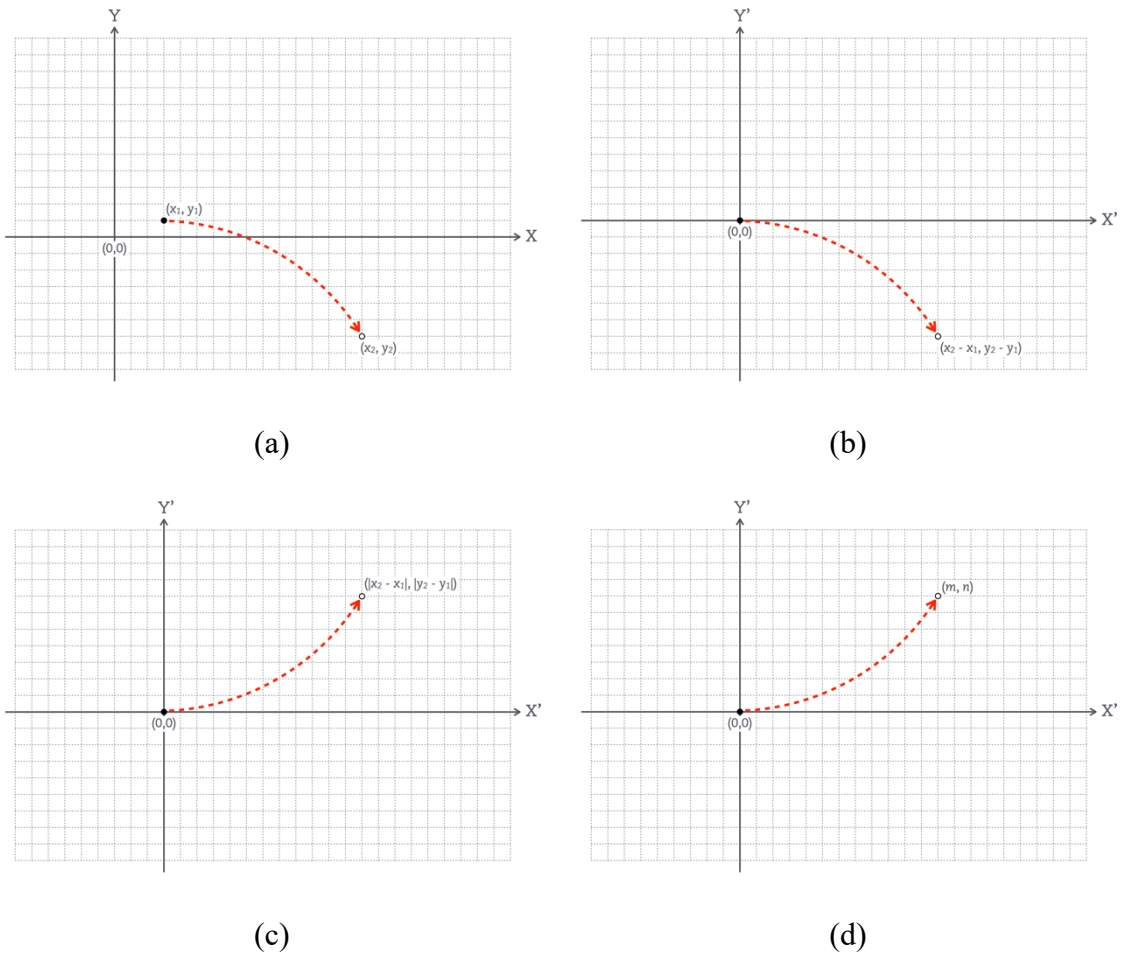


Figure II-6. Transformation from $[(x_1, y_1) – (x_2, y_2)]$ to $[(0, 0) – (m, n)]$

Table II-1. Viable nodes, links, and paths from (0,0) to (m, n) with no backtracking.

For the rectangular area between the origin (0,0) and the destination (m, n)		
Number of Nodes within the rectangular area:	$(m+1)(n+1)$	(2.4)
Number of Links within the rectangular area:	$m(n+1)+n(m+1)$	(2.5)
Number of links in each non-backtracking path:	$m+n$	(2.6)
Number of paths in the rectangular area:	$\frac{(m+n)!}{m!\,n!}$	(2.7)

Example 4 – All Viable Paths in a Simple 2×2 Rectangular Grid

So, for a simplistic 2×2 network, where $m = 2$ and $n = 2$, instead of case with only three paths shown in **Figure II-4,** we now look at all six, (2+2)!/2!/2! = 6, viable paths in **Figure II-7**. Results from tens of thousands of Monte Carlo simulations show that the SP likelihood of each path converges, see **Figure II-8**, with paths ADGHI and ABCFI being the most likely SPs, each at 20%, while the remaining four paths are less likely at 15% each. The "naïve" average probability, assuming that all six paths are equally likely to be the shortest, would have dictated an equal 1/6, or 16.67% probability, for all paths. However, ADGHI (or ABCFI) is 20% (= 20% ÷ 16.67% - 1) higher than the naïve average. The same path is 33.33% (= 20% ÷ 15% - 1) more likely to be the shortest path than any of the less likely paths. Millions more Monte Carlo simulations can be conducted and were conducted to confirm this evident disparity in the likelihood of some paths being higher, sometimes much higher, than other very similar paths with identical expected travel times. The key insight of this study early on is that while the travel times on all links are assumed to be independent, by definition, the resultant path travel times should not be regarded as independent automatically whenever paths share links. The more links a path shares with other paths, the less independence it enjoys.

Consider three paths, each with an identical number of links with *iid* travel times. If the three paths do not overlap with one another, or have no shared links whatsoever, then each path would have a probability of 1/3 of being the shortest path. But if there are some shared links, as is likely in the real world and in a rectangular grid, some of the paths would not be totally independent.

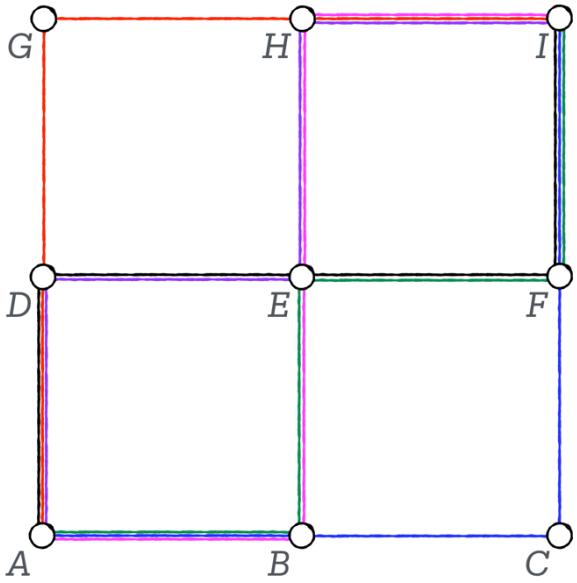


Figure II-7. All viable paths in a 2×2 rectangular grid

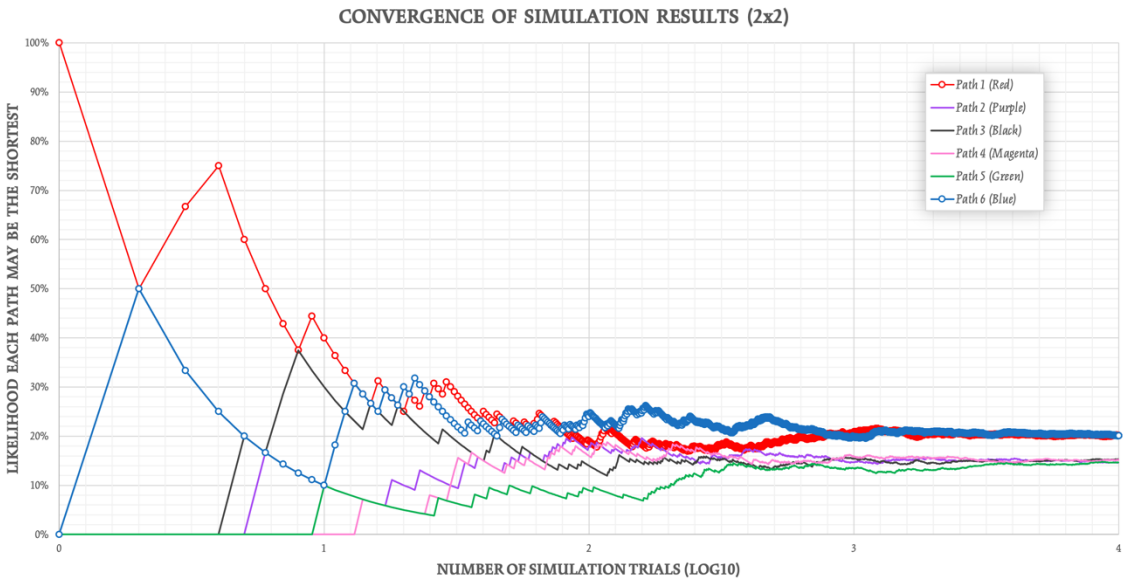


Figure II-8. Likelihood of each viable path for a 2×2 rectangular grid

Example 5 – Simple Scenarios of 3 Paths in a 10×10 Grid

Let's now try to find the shortest path from (0,0) to (10,10) among the three viable paths shown in **Figure II-9**. In all three scenarios, the red path is entirely independent of the other two (green and blue) paths, while the latter share two links, 10 links, and 18 links out of the 20 links along each path.

Again, bear in mind that if the paths are independent of one another, each would have the same 1/3, or 33.33%, chance of being the shortest path. Monte Carlo simulation results are shown in **Table II-2**. The path that is independent tends to have the highest likelihood of being the shortest path. The less independent a path is, the lower the likelihood it would be the shortest path. Paths with a similar number of shared links tend to have similar levels of likelihood of being the shortest path.

This simple example also shows that the correlation between paths is identical to the number of shared links divided by the total number of links in a path. Using $\rho(X,Y)$ to represent the correlation between paths X and Y, we have $\rho(R,G) = 0$ and $\rho(R,B) = 0$ for all three scenarios and $\rho(G,B) \approx 10\%$, 50%, and 90% for scenarios (a), (b), and (c), respectively, which tracks the number of shared links divided by total links in each path.

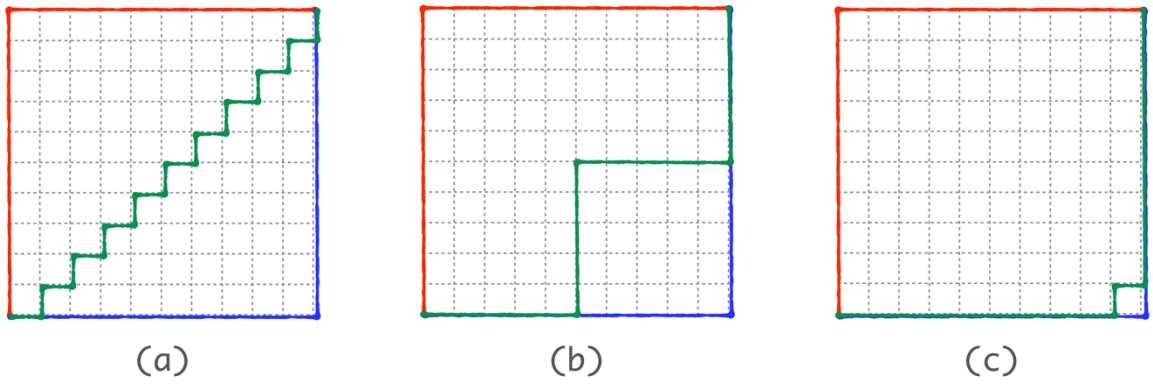


Figure II-9. Three simplified scenarios in a 10×10 rectangular grid

Table II-2. Winning percentage of each path for several scenarios, including those in Example 5.

Path	Independent	Scenario (a)	Scenario (b)	Scenario (c)	$G = B$
Red (R)	33.33%	34.25%	38.52%	44.86%	50.00%
Green (G)	33.33%	32.87%	30.74%	27.58%	25.00%
Blue (B)	33.33%	32.88%	30.74%	27.56%	25.00%
$\rho(G,B)$	0/20	2/20	10/20	18/20	20/20

Note: "Independent" means the three paths have no shared links whatsoever, while "$G = B$" means the green path and the blue path share every single one of their paths.

For the "$G = B$" case, where the two paths share all their links and are, in fact, the same path. In such an extreme situation, Path R would have a winning percentage of 50%, and Path G/B would have the same. But if one insists on treating them as two separate paths, G and B, then each path shares half of that 50% and gets 25% of the winning percentage, just to be fair.

With this in mind, one can now compare scenario (C) with this two-path-only scenario. In effect, scenario (C), where paths G and B share 90% of their links, ought not to be considered as two independent paths but nearly just a single path with a tiny detour in the middle of the journey. The two paths behave nearly identically and would have nearly identical fates in terms of being the shortest or not. At the same time, path R would enjoy a much higher probability of winning, close to 45%, as opposed to 33.33%, when all three paths are truly independent. This, then, is a crucial understanding of the problem at hand: paths that share many links collectively behave like a single path corridor that has its winning probabilities diluted, reduced, and shared by the constituent paths.

The Concept of Fractional Independence among Paths

In a rectangular grid network, whenever l viable paths exist between a pair of OD, where $l \geq 3$, at least l-1 of these paths would be sharing links and, therefore, would not be entirely independent. When a path shares some of its links with other paths, it does not enjoy full independence but still retains certain **Fractional Independence**. Even if every single link of the path is shared with other paths, it will still retain a **Degree of Independence** (*DoI*), or φ, where $0 < \varphi \leq 1$.

The dependency between two paths can be viewed as the correlation of two variables. But when millions of competing paths share some links with others, the complexity of this fractional independence of a certain path to all the rest can be captured primarily by the number of paths each of its links is shared with. So, here we define the Degree of Independence, φ, as:

$$\varphi_i = \frac{1}{k} \sum_{j=1 \ldots k} \frac{1}{Number\ of\ Paths\ Going\ thru\ Link\ j} \tag{2.8}$$

where φ_i is the *DoI* of the path i, which consists of k links; each is used by at least one path, but most likely many more. For an $m \times n$ rectangular grid, the number of links in each path is $k = m+n$ for every viable path within the $m \times n$ grid.

For Example 5 above, the *DoI* for each path in the various scenarios can be calculated using Equation 2.8 and are presented in **Table II-3**. The RED path shares no link with any other paths and will have a 100% *DoI* throughout all cases. For the case where the three paths share no links at all, each would have a *DoI* of 100%. With an increasing number of shared links between the GREEN path and the BLUE path, the *DoI*

Table II-3. Degree of Independence of each path in the scenarios in Example 5.

Path	Independent	Scenario (a)	Scenario (b)	Scenario (c)	$G = B$
Red (R)	100%	100%	100%	100%	100%
Green (G)	100%	95%	75%	55%	50%
Blue (B)	100%	95%	75%	55%	50%
$\rho(G,B)$	0/20	2/20	10/20	18/20	20/20

for each path decreases. In the "$G = B$" case where the two paths overlap entirely, and are essentially the same path, the two paths then split the total independence to each claim a *DoI* of 50%.

For any link j, the number of paths going through it from (0,0) to (m,n) is the product of the number of paths from (0,0) to the near side of link j and the number of paths from the far side of link j to (m,n). Unlike the simple three-pass scenarios in Example 5, if all links are present and all paths between (0,0) and (m,n) are viable, then we can calculate the exact number of paths going through link j. Depending on the orientation of the link j, see **Figure II-10**, this amounts to the following if all paths are present:

$$\frac{(a+b)!}{a!\,b!} \times \frac{(m+n-a-b-1)!}{(m-a)!\,(n-b-1)!} \tag{2.9}$$

or

$$\frac{(a+b)!}{a!\,b!} \times \frac{(m+n-a-b-1)!}{(m-a-1)!\,(n-b)!} \tag{2.10}$$

One could employ equations 2.8 through 2.10 to systematically calculate the *DoI* of every single path for a given set of *OD*. It should be noted that the *DoI* value of a path predicated on m and n, the placement of the path in that $m \times n$ grid, and the existence of other paths that it may share links with. It is not affected by the shape of the travel time

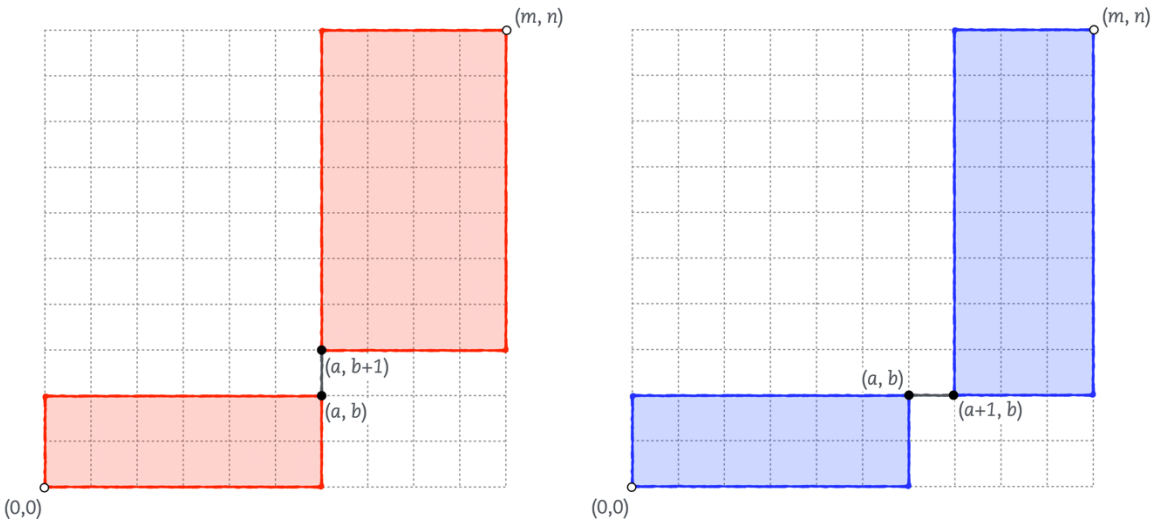


Figure II-10. All paths from (0,0) passing through link *j* to (*m*,*n*)

distribution function for each link or other aspects that change with time. In other words, the degree of independence is not stochastic for any path in a given set of OD, or in a confined space of an *m*×*n* grid, even when the travel time for each link and, for that matter, each path is stochastic. This is a key understanding of the stochastic SP problem with a deterministic solution.

By summing up the *DoI*s of all viable paths, we can obtain the equivalent number of independent paths, N_E, within the confines between (0,0) and (*m*,*n*). For the most common case where all links and paths are present within the *m*×*n* grid, we have:

$$N_E = \sum_i \varphi_i = \frac{2mn + m + n}{m + n} \tag{2.11}$$

This is significantly less than the number of viable paths introduced earlier in Equation 2.7, or $\frac{(m+n)!}{m!n!}$. For the example of $m = 10$, and $n = 10$, while the number of viable paths is a staggering 184,756, the equivalent number of independent paths, N_E, is a mere 11. On

average, each viable path would have a *DoI* of 11÷184,756, or 0.005958%. It should be noted that N_E is a simple m+1, when m=n. N_E can take on non-integer values, which is the case when $m \neq n$. Also, for the simplest case of m=n=1, N_E = all viable paths = 2.

A Path's Degree of Independence and Its Likelihood of Being the Shortest

One may notice that there exists a significant correlation between the likelihood of a path being the shortest in **Table II-2** and the degree of independence of the same path in **Table II-3**. This is not a coincidence. In fact, this relationship is essential for "intuiting" the most likely shortest paths when travel time is stochastic.

Let's look at an 8×8 grid where we can calculate the *DoI* values for all 12,870 paths and estimate each's likelihood of being the shortest using Monte Carlo simulations. The results are plotted in **Figure II-11**. Each of the 12,870 dots in the figure shows the normalized frequency (a surrogate for probability) of a path being the shortest path vs.

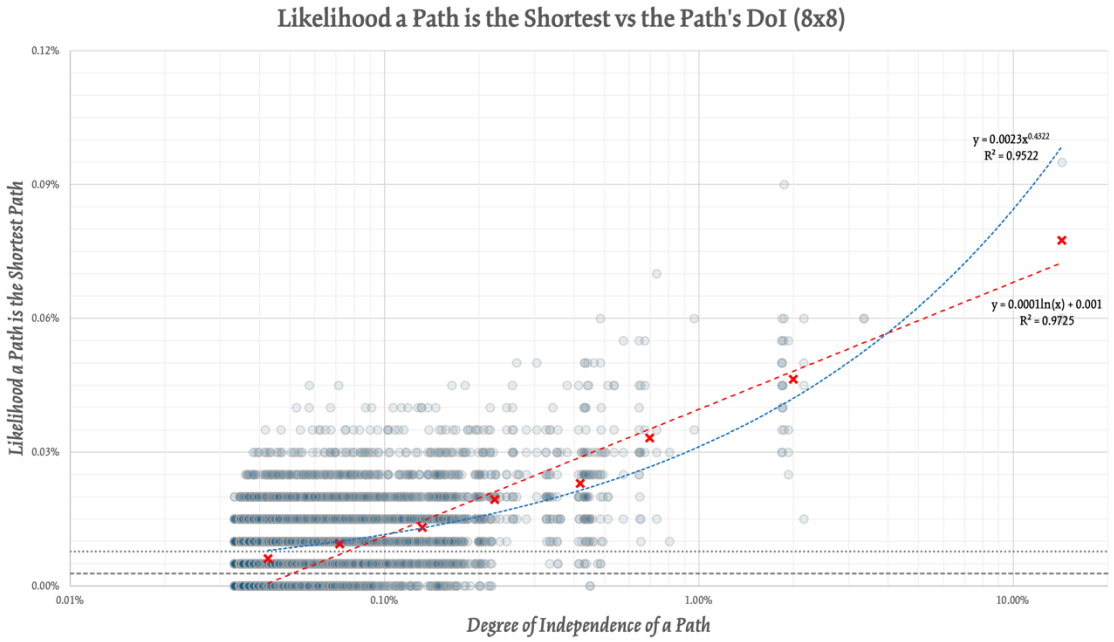


Figure II-11. Likelihood of a path being the shortest vs the path's *DoI* in an 8×8 grid

the same path's *DoI* value. Two models were used to curve-fit the data, both with high degrees of predictive capabilities, and both evidenced a strong correlation between the two variables with an R^2 value of 0.9725 for the simpler log-linear model. The dotted line at the bottom is the naïve average, or the expected likelihood if one thought all paths would have the same probability of being the shortest path.

The key here is that paths with the highest *DoI* have the highest likelihood of being the shortest. Since we have already established that *DoI* values are deterministic, **the most likely path to be the shortest in a stochastic future is actually deterministic**.

For the total of 12,870 viable paths, this expected average likelihood value would come to 1/12,870, or 0.0078%. This is less than one-tenth of the likelihood value of 0.0775% of the path with the highest degree of independence. The relative frequency, RF, (of winning) of the most likely shortest path at 0.0775% compared to the naïve average of 0.0078% is 9.98 times. The RF of winning is analogous to having a certain lottery ticket that is somehow ten times more likely to hit the jackpot than the average lottery ticket.

Holding the $m+n$ value fixed, we also investigated a 7×9 grid (**Figure II-12**) and a 6×10 grid (**Figure II-13**) to demonstrate the strong relationships we observed persist even when grid symmetry, i.e., $m=n$, is eliminated. For all three cases, the most likely shortest paths are those with the highest *DoI* values. The *RF* values are 9.73 and 9.65 for the 7×9 grid and the 6×10 grid, respectively. The *RF* values for grid sizes of m by n, each varying from 0 to 10, were calculated after millions more Monte Carlo simulations and plotted in **Figure II-14**. For smaller grid areas, such as 1x7 and 2x2, the *RF* values

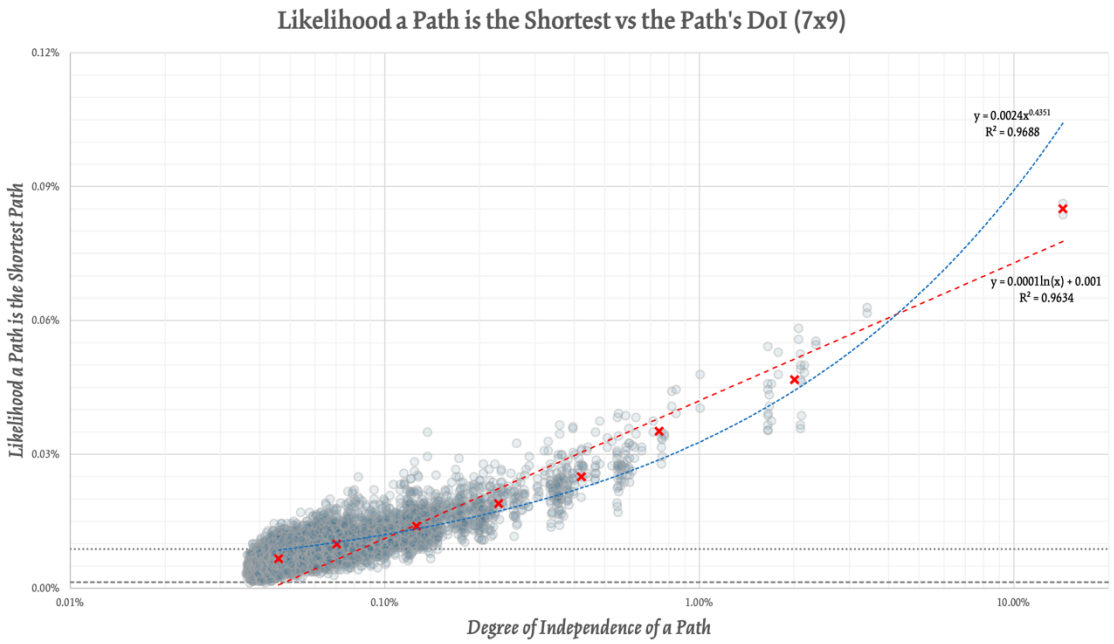


Figure II-12. Likelihood of a path being the shortest vs the path's *DoI* in a 7×9 grid

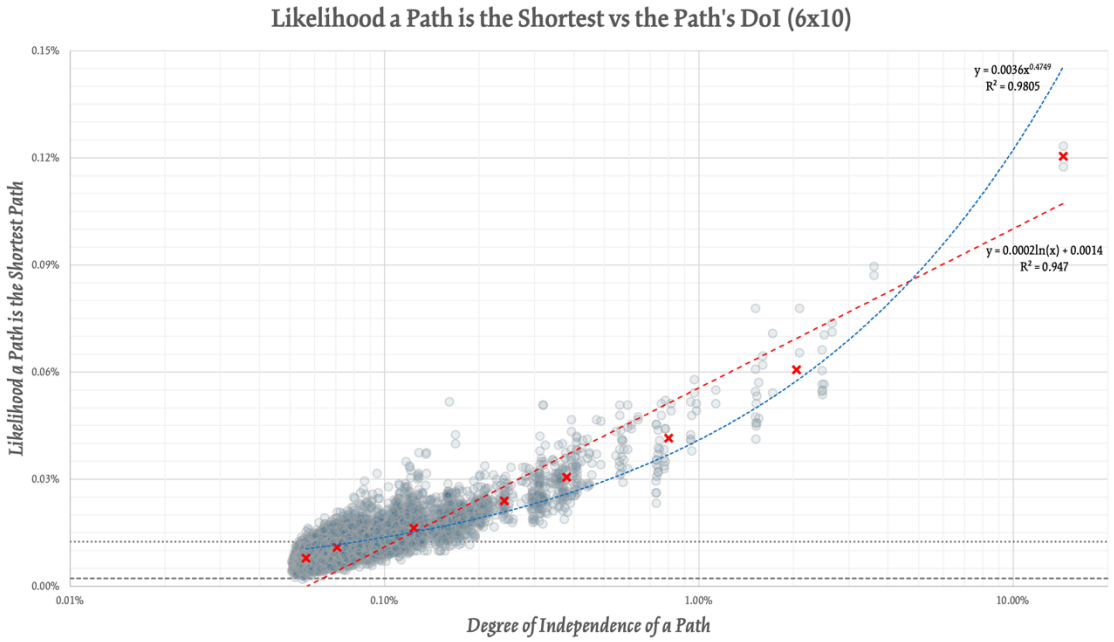


Figure II-13. Likelihood of a path being the shortest vs the path's *DoI* in a 6×10 grid

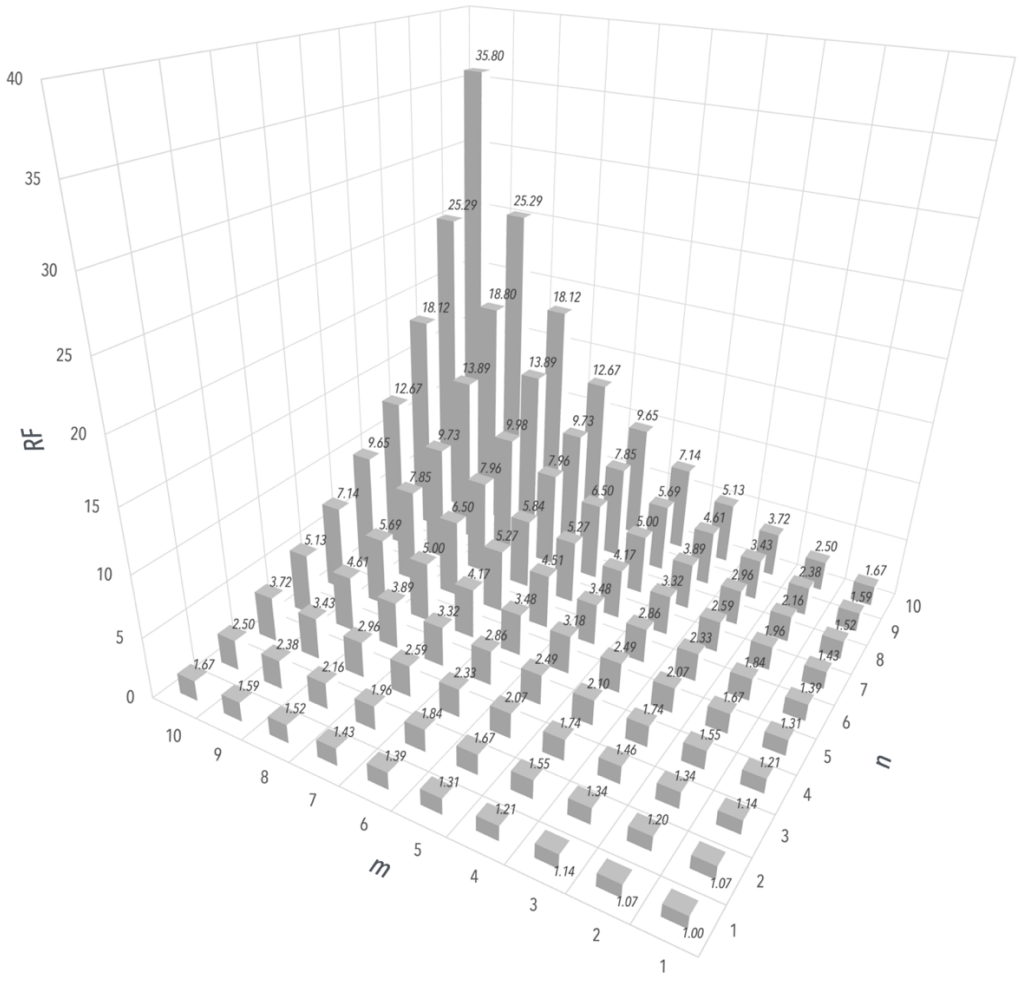


Figure II-14. Relative frequency (RF) of the likelihood for m = 1..10 and n = 1..10

indicate a less than 50% increase in the likelihood of winning than the average. But as *m* and *n* increase, the advantage increases. For the 10x10 case, the path with the highest *DoI* is 36 times more likely to win than the average path.

The Relative Frequency (RF) Model

The *RF* value for each given (0,0) to (*m*,*n*) pair in **Figure II-14** exhibits a clear trend of symmetrical two-dimensional exponential growth. The regression summary in **Figure II-15** presents the best model for *Log*(*RF*) with a near-perfect R^2 of 0.9964.

$$Log(RF) = -0.0608 + 0.0365(m+n) + 0.0291mn \tag{2.12}$$

So, the *RF* model can be written as:

$$RF = 0.9410e^{0.0365(m+n)+0.0291mn} \tag{2.13}$$

Exhaustive simulation runs would become prohibitive for *m* and *n* values larger than 10. For example, for the case of a 30x30 grid, we would have to exhaustively calculate the travel time for over $1x10^{17}$ viable paths for about that many times or more.

Regression Statistics	
Multiple R	0.99819147
R Square	0.99638622
Adjusted R Squ	0.99627329
Standard Error	0.05117327
Observations	100

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	69.3144082	23.1048027	8822.99263	4.763E-117
Residual	96	0.25139555	0.0026187		
Total	99	69.5658037			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-0.06083257	0.02388086	-2.54733599	0.01244329	-0.1082357	-0.01342944	-0.1082357	-0.01342944
m	0.03650713	0.00384875	9.48545376	1.897E-15	0.02886742	0.04414684	0.02886742	0.04414684
n	0.03650713	0.00384875	9.48545376	1.897E-15	0.02886742	0.04414684	0.02886742	0.04414684
mxn	0.0290504	0.00062028	46.8341768	6.3211E-68	0.02781915	0.03028165	0.02781915	0.03028165

Figure II-15. Regression summary of the *RF* model

Here, Equations (2.12) and (2.13) become very useful for estimating the magnitude of the advantage of the paths with the highest *DoI* values. As shown in **Figure II-16**, the *Log*(*RF*) value increases linearly with *m*, *n*, and *m*·*n*. For a 20×20 grid area, the paths with the highest *DoI* value would have *RF*(20,20) = 451,176. Using the lottery analogy before, it's like a ticket that is nearly half a million times more likely to win than an average ticket.

Looking back at Example 1, with 20 viable E-W streets and 7 N-W avenues in Midtown Manhattan, we obtain *RF*(20,7) = 147. That is, the path with the highest *DoI* value would be about 150 times more likely to be the shortest than an average path.

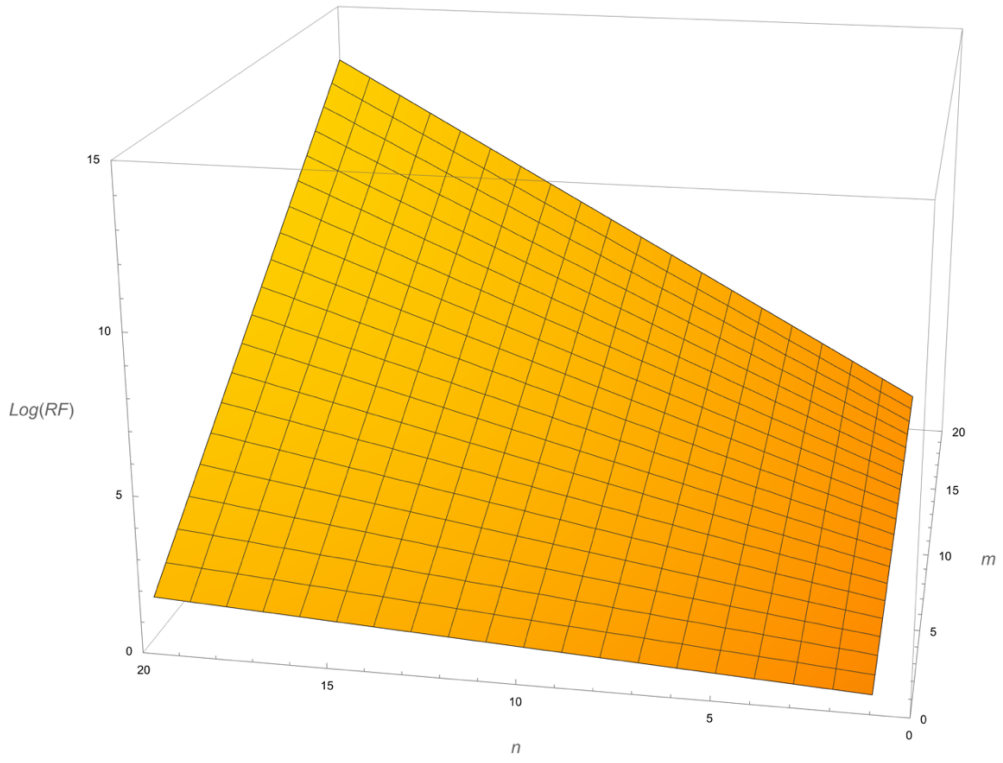


Figure II-16. Visualization of *Log*(*RF*) as a function of *m* and *n*

The question, then, is how do we identify these paths with the highest *DoI* without all the tedious calculations?

So, Which Paths Have the Highest DoI

Examining the schematics of simpler cases and the simulation results of more complicated cases reveals a pattern: the two paths at the opposing edges of the $m \times n$ grid possess the highest *DoI* values, which means they share the least number of links with all other competing paths. For the example of the 20×20 grid, see **Figure II-17**, the number of paths sharing each horizontal link is shown in pink above the link, and the number of paths sharing each vertical link is shown in blue to the left of the link. The links in the middle of the upper-left to lower-right patch are each shared by billions of paths, which results in a low degree of independence and should be avoided. The links in the upper-right and lower-left corners are shared by far fewer paths, which results in a significantly higher degree of independence, and are preferred.

So, without exception, the paths with the highest *DoI* values and are most likely to be the shortest are the two paths hugging the edges of the $m \times n$ grid on opposite sides, see **Figure II-17.** This is true for any m and n larger than zero. Even in the cases where the two most likely paths do not win, the paths mostly hug the edges, with only minor deviations in the green areas, would be the strongest contenders to win. Paths that spend a significant portion in the central red band have low *DoI* values and tend to fare very poorly in their bid to be the shortest.

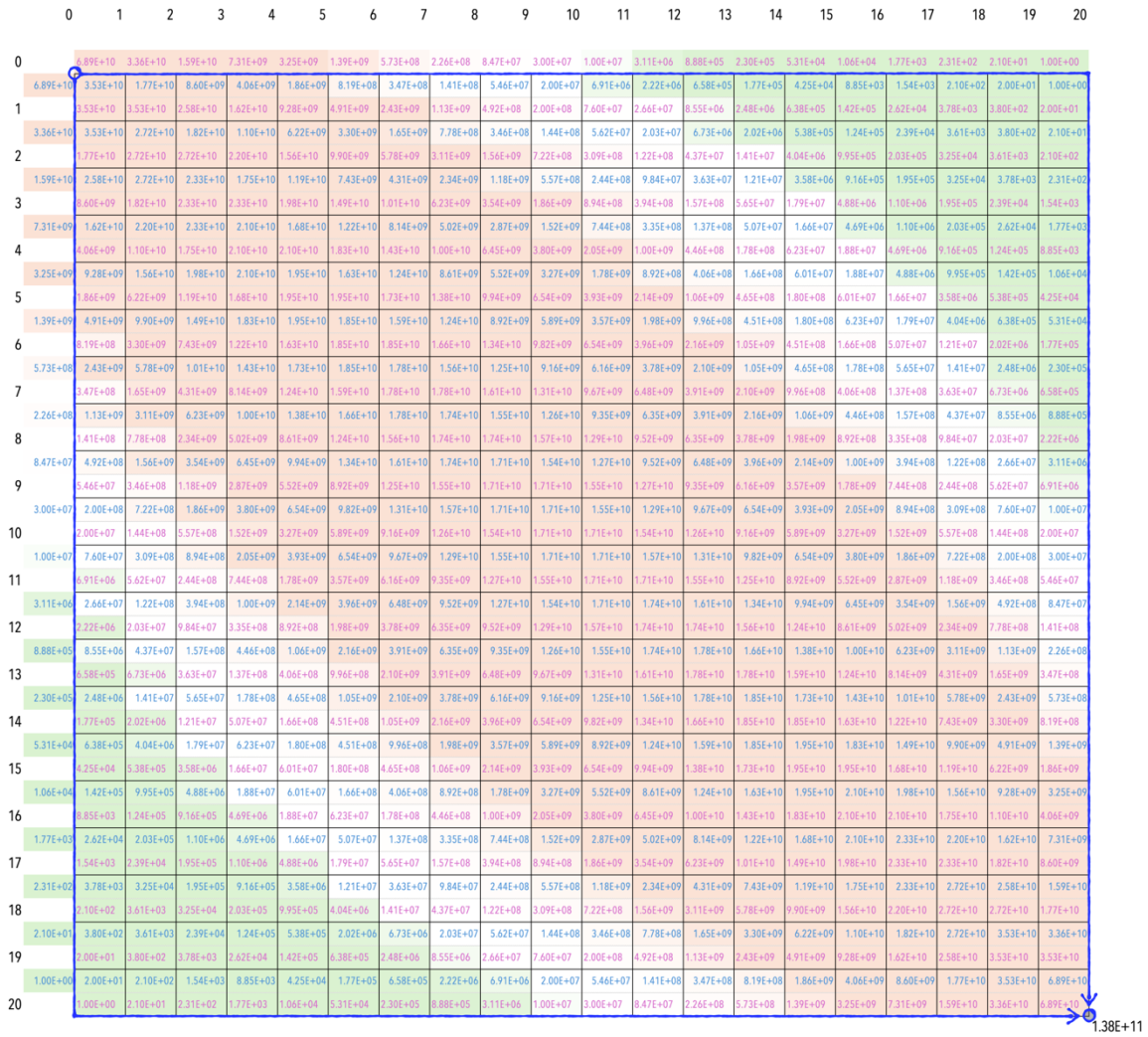


Figure II-17. Number of paths sharing each link in a 20×20 grid

CONCLUSIONS

The research documented in this paper sought to provide an intuitive solution to the shortest path problem with stochastic travel time in an expansive rectangular grid network with numerous similar and competing paths. The following is a summary of the thought process and actions taken to arrive at the completion and contributions of the study.

- We recognized that roadway travel times are stochastic in nature and cannot be predicted with accuracy beforehand. Traditional shortest-path algorithms that work well after the fact with deterministic travel time data would fail because they do not treat the link travel time as stochastic and lack the ability to predict future travel times.
- We observed that paths often share common links and are not independent. Depending on the network structure, the number of shared links, and the number of paths these links are shared with, different paths can be more independent or less so than others. This Fractional Independence in paths is a key factor affecting the likelihood that a path might be the shortest compared to numerous otherwise similar paths.
- We introduced a measure of the Degree of Independence (*DoI*) to gauge how independent a path is based on the number of paths sharing each of its constituent links.

- We employed Monte Carlo simulations to examine the stochastic travel time for every viable path to evaluate and solve the SP problem exhaustively, as opposed to using heuristic algorithms with fixed travel times.
- We conducted millions of Monte Carlo simulation runs to obtain the converged likelihood values of each viable path, which is the shortest for a wide number of grid sizes with symmetrical and asymmetrical rectangular areas.
- We successfully established the strong correlation between any path's deterministic *DoI* and its stochastic likelihood of being the shortest among numerous competing paths. This is a key step towards predicting the shortest path while lacking the ability to know the stochastic future.
- We devised Relative Frequency (*RF*) to measure the ratio of the frequency of the most likely shortest path actually winning over the average frequency of any path winning. The greater the *RF* value, the more likely the path with the highest *DoI* is to be the shortest compared with any other random path. For an 8x8 grid, this advantage is about 10 times, and for a 10x10 grid, this advantage increases to about 36 times.
- We developed a near-perfect regression model, with an R^2 of 0.9964, to fit this relationship between *RF* and (*m,n*) and expanded our ability to estimate the *RF* value for any given *m* and *n* well beyond the range where we could comfortably simulate the winning frequencies of all viable paths in an exhaustive manner.
- We, thereby, reduced the impossible stochastic shortest path problem to a simple deterministic prediction where the eventual shortest path is most likely to follow the outside edges of the rectangular grid.

REFERENCES

1. Dijkstra, E. W. (1959). “A Note on Two Problems in Connexion with Graphs.” *Numerische Mathematik,* 1(1): 269–271.
2. Hart, P. E., Nilsson, N. J., & Raphael, B. (1968). “A Formal Basis for the Heuristic Determination of Minimum Cost Paths.” *IEEE Transactions on Systems Science and Cybernetics,* 4(2): 100–107.
3. Metropolis, N., & Ulam, S. (1949). “*The Monte Carlo Method.”* Journal of the American Statistical Association, 44(247): 335–341.
4. Fawaz, M.Y. & G.F. Newell (1976). “Optimal Spacing for a Rectangular Grid Transportation Network – I: A Hierarchy Structure.” Transportation Research, 10(2): 111-119.
5. Newell, G.F. (1996) “Non-Convex Traffic Assignment on a Rectangular Grid.” Transportation Science, 30(1): 32-42.
6. Loui, R. P. (1983). “Optimal Paths in Graphs with Stochastic or Adversarial Costs.” *Communications of the ACM, 26*(9): 670–674.
7. Rubinstein, R. Y. (1999). “The Cross-Entropy Method for Combinatorial and Continuous Optimization.” *Methodology and Computing in Applied Probability, 1*(2): 127–190.
8. Long, K., L.D. Han, & S. Wang (2011). “Shortest Path Search in Road Network with Incomplete Information.” *Journal of Transportation Applications*, 31(3): 6-11.

9. Lu, W., L.D. Han, C. Liu, & K. Long (2014). “A Multiple-Source, Nearest Destination, Shortest Path Problem in Evacuation Assignments.” CICTP 2014: Safe, Smart, and Sustainable Multimodal Transportation Systems.
10. Han L.D., H. Wang, & W.F. Mackey Jr (2006). “Finding Shortest Paths under Time-Bandwidth Constraints by Using Elliptical Minimal Search Area.” Transportation Research Record 1977(1): 225-233.
11. Hobeika, A.G., S.A. Ardekani, & L.D. Han (1988). “Logistics Decisions Following Urban Disasters.” Computer-Aided Civil & Infrastructure Engineering, 3(1): 13-27.
12. Han, L.D. (1987). A Decision Support System for Mass Evacuation and Emergency Management, Master’s Thesis, Virginia Polytechnic Institute and State University.

CONCLUSION

This dissertation addresses two interrelated but distinct challenges in transportation systems analysis: the imputation of significant data loss in vehicle tracking and the prediction of optimal paths in a network under stochastic travel time conditions.

The first part of the study focused on identifying and remedying a common but often unrecognized form of data loss in long-distance vehicle-tracking datasets. The issue arose from misaligned or truncated time windows across data collection points, resulting in nearly one-third of trips being unintentionally discarded from aggregated statistics. By analyzing the patterns of missing data, the study identified nine distinct loss scenarios, most notably late starts and early ends at both origin and destination nodes. To address this, we expanded the raw data and modeled journey times using the shifted Negative Exponential Distribution (shifted-NED), which was found to be the most suitable for capturing the decaying frequency nature of truck trips. Through iterative Monte Carlo simulations, the model was validated with low error margins (MAPE of 4.79%, MAE under 9 minutes). These simulations enabled the reconstruction of the missing data, resulting in hybrid datasets that corrected previously misleading trends, such as declining journey times and trip counts, thereby restoring the integrity of the data and opening the door for more accurate downstream analyses, including future studies on driver behavior.

The second part tackled the stochastic shortest path (SP) problem in expansive rectangular grid networks with many competing and overlapping routes. Recognizing the limitations of traditional deterministic SP algorithms, the study emphasized the inherently stochastic nature of travel time and the interdependence among paths due to shared links.

It introduced the concept of the Degree of Independence (*DoI*) to quantify how unique a path is based on the overlap of its portions with other paths. Extensive Monte Carlo simulations across various grid configurations revealed a strong correlation between a path's *DoI* and its probability of being the shortest. The concept of Relative Frequency (*RF*) was then developed to express the advantage of the most likely shortest path over the average path. This relationship was captured with a highly accurate regression model ($R^2 = 0.9964$), allowing predictions of path performance across a broad range of network sizes. Ultimately, the study demonstrated that, in the absence of precise future travel times, the most efficient paths tend to follow the grid's outer edges—a significant step in simplifying complex stochastic routing problems.

Together, these studies provide valuable frameworks for improving both the reliability of transportation datasets and the predictability of routing decisions in uncertain environments.

VITA

Michael J Han was born in Knoxville, Tennessee, in 1995. After graduating from Farragut High School with honors in May 2013, he enrolled in the University of Tennessee that August. There, he earned the Bachelor of Science in Computer Science in December 2017, the Master of Science in Computer Engineering in December 2021, and the Doctor of Philosophy in Civil & Environmental Engineering in August 2025.