Abstract

The literature on solving the shortest path problem is very extensive when it comes to deterministic arc lengths. However, in transportation networks, it is more useful to think of costs on the links as travel times as opposed to distances. Furthermore, while distances are constant between nodes, travel times are random, as a result of drivers preferences, incidents, road conditions, weather and traffic volume among others. And although recent research has developed a variety of algorithms for routing in these stochastic networks, less has been done in identifying how exactly these travel times distributions are. Moreover, most of these algorithms rely on the assumption that travel times are independent between links.

In this thesis, we use data obtained from drivers using in-vehicle navigation systems to analyze the behavior travel times on the US road network. We fit normal, lognormal, gamma and Weibull distributions to these travel times and conclude that the lognormal model provides the better fit except for the right tail. We also explore Generalized Hyperbolic (GH) distributions, which have been used recently on analyzing financial data and have a better fit on heavy tails. We fit GH distributions to the same travel time data and compare its results to the lognormal fit.

We then use our data to test the assumption of independence between arcs. We take a roads segment comprised of eight links and find the correlation between travel times on the links. We compare the correlations for consecutive links, and for links separated by one or more links. We analyze the issue of convoluting travel time distributions when these times are not independent. For this purpose, we use Reciprocal Gamma distributions, which have been proven to represent the infinite sum of correlated lognormal distributions.
Chapter 1

1 Introduction

This thesis looks at the problem of establishing the probability distribution of travel times in some segments of the US road network. To this end, we use data gathered by GPS software Copilot® and fit it to several theoretical distributions. Additionally, we take travel times on several consecutive links of a path and analyze its correlations. This will validate or cast doubts over the hypothesis of travel time independence between links, which is one of the basic assumptions of routing algorithms in stochastic networks.

1.1 The Stochastic Shortest Path Problem

When the costs on the arcs are deterministic, standard algorithms exist to find the shortest path between two nodes. These algorithms play a fundamental role on route planning on transportation, where typical weights on roads are distances, financial costs or travel times. However, the latter are usually derived using speed approximations (like speed limit) and therefore leading to inaccuracies with respect to reality.

Algorithms like Djisktra, Bellman-Ford and its variations have proven to be efficient on calculating the shortest paths on complex road networks of deterministic type, and current travel and in-vehicle navigation software implement them rapidly enough to react in real time. Most of these algorithms rely on dynamic programming, with the assumption that the cost of a path is the sum of the costs of its arcs. Basically, the algorithm starts from the source node and finds the best path to all nodes using best paths that have been previously found. That is, if the best path found from
Finding a shortest path in terms of deterministic costs, like distance, tolls, turning penalties, etc. can be useful when the objective is not to minimize travel time but some other user utility function (financial cost, miles driven). However, when a driver does wants to minimize the time it takes him to go from point A to point B, the deterministic nature of the network becomes unrealistic. Costs on the arcs are now probabilistic because of factors such as traffic volume, weather, road conditions and incidents, among others. Additionally, as a result of traffic volume, costs can also be time-dependent, where traffic patterns during the day affect the travel times on the links.

Classical algorithms are not suitable then to handle time-dependence and randomness of travel times. Furthermore, the notion of a shortest path becomes inadequate, since the additive principle of arc lengths is no longer valid. The problem now has to focus on first establishing an appropriate utility function, which will be user dependent, and then finding an efficient algorithm that can incorporate these new conditions. In the majority of the research made about this topic, the problem is approximated by either the time dependence perspective, or by the random arc cost perspective, although more recent work aims to combine the two.

1.1.1 The Random Arc Cost Perspective

When talking of arc costs as random variables, the assumption is that these costs are not time dependent but rather that a probability density function exists which can describe the travel time on an arc, and a realization of the travel time comes from this function regardless of the time of day. Therefore, on the road network, each link will not have a fixed travel time associated to it but a probability distribution of this travel time.

Accordingly, to talk about a “shortest” or “fastest” path becomes awkward since the problem is now to compare probability distributions instead of plain numbers. A simple and somewhat effective approach is to take the shortest path in terms of expected travel time, where the cost on each link is now the expected time. The network then becomes deterministic and any of the classical algorithms can be used. The key issue here is to notice that defining the shortest path this way implies that the utility function of the driver is the expectation.
Although this objective is quite reasonable, drivers’ utilities are obviously not the same, and other factors come into play that can make path planning more accurate and useful. For example, a user would like to choose the path through which he has the highest probability of arriving at his destination before a specific time. Or a user might prefer a path that has a lower variance in travel time or even a combination of low variance and low expected travel time. Evidently, with these utility functions, reducing the network to a deterministic one is not appropriate and an analysis of the properties of the travel times is necessary.

One simple approach to compare travel time distributions of two paths in order to determine which one is “better” is dominance pruning (Wellman []). This method can be applied in stochastically consistent networks, that is, networks where the probability of arriving at a destination can’t be improved by leaving later. Given two departure times \( s \) and \( t \), \( s \leq t \), travel times \( x_{ij} \) and arrival time \( z \), such a network would follow:

\[
Pr(s + c_{ij}(s) \leq z) \geq Pr(t + x_{ij}(t) \leq z)
\]

It is not difficult to see that when the travel times are not time dependent, stochastic consistency is always achieved.

Dominance pruning relies on the concept of stochastic dominance, which in terms of road networks states that the travel time distribution of a path dominates the distribution of another path with the same origin and destination if and only if the cumulative probability function is uniformly greater or equal to that of the other path. This concept is illustrated by Figure 1.1, where the travel time distribution of path 2 dominates that of path 1. Path 1 would then be pruned as long as the utility of the driver is non-increasing with respect to travel duration, since clearly Path 2 would be preferred.

However, to apply dominance pruning or any of the algorithms developed for stochastic networks (see Miller-Hooks [], Cooper [] among others), a clear understanding of the travel time distribution in a link is needed. Not only this, but in order to make the computational burden more manageable, we need to know how to generate the travel time distribution of a path if only the distributions of the links comprising it are known. Having a system with the probability distribution of all possible paths in the network is clearly inefficient, so we need to understand how to convolute the individual distributions.
The purpose of this thesis is to precisely consider the problem of finding a theoretical approximation of the travel time distributions on different road segments for the country as well as to examine travel time distributions of paths in terms of the distributions of the links that comprise them.

### 1.1.2 The time dependent arc cost perspective

Another perspective to approach the varying nature of travel time is to model it as a function of time of day. This seems as a natural step since travel time between two points on a road or highway changes mainly because of traffic volumes, in what are called “peak” and “off-peak” periods. For example, Figure 1.2 shows how travel time increases on afternoon hours for a segment of Route 601 near Princeton NJ.

A very simple solution to cope with this problem is to divide the day in intervals, where each one will have a travel time (or a distribution) associated to it. On the other hand, a functional form derived from the data can be used to express the travel time on the link as a function of time of...
day. On working with data from the Milwaukee Highway System, Kornhauser and Schrader (see []) developed the following ten parameter function:

\[ TT = f(t) = K + C_1 \eta(\mu_1, \sigma_1) + C_2 \eta(\mu_2, \sigma_2) + C_3 \eta(\mu_3, \sigma_3) \]

where:

\[ \eta(\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(t-\mu)^2}{2\sigma^2}}. \]

This expression comprises a free flow travel time and three bell shaped curves that characterize the three rush hour periods during the day (morning, after school, afternoon). The parameters to be estimated are \( K, C_i, \mu_i \) and \( \sigma_i \).

Regardless of which of these two approaches is used to categorize travel time during the day, dynamic programming algorithms have been the topic of various research efforts to find the shortest path (see Chabini [], Pallotino []). More recently, focus has also been given in considering both the time dependent and random structure of travel times at the same time.
1.2 Motivation

The problem of analyzing the travel time distributions on the road network is specially relevant since most of the work on stochastic road networks has been made from the algorithmic point of view. This work mostly focuses on developing algorithms and properties of the networks in order to route vehicles through them, but the costs on the links have almost always been assumed as independent and with a known distribution, in particular Gaussian.

Therefore, not too much attention has been given to analyzing travel time distributions between nodes, and this is mainly due to not having an extensive amount of data to analyze. However, in the past few years, the development of in-vehicle navigation systems makes millions of average drivers a huge source of potential data. Recent developments in the software that runs these systems have made it possible to gather this data. So, the real value of this thesis is that it is a first attempt to use this type of data in order to infer statistical properties of the travel time distributions on the US road network.

Considering the random nature of road traveling in route planning is one of the major objectives that ITS technologies face today. The challenge lies in transferring to the navigation system the thought process that one makes in choosing a path. Consider the case of choosing a route to go to the airport. Although choosing the path that will yield the minimum time is always natural, one would also like to find the path through which one has the highest probability of arriving before the check in time. Or, instead of picking a path that has a low expected travel time but high variability, one would like a path with a slightly higher expected travel time but a lower variability.

Incorporating all these options into current in vehicle navigation systems means that not only these systems need to handle different drivers utility functions, but also that the network can no longer be considered deterministic but instead a mathematical model of the random travel times must be implemented.

Using data gathered from these navigation systems in order to create these mathematical models and improve their own performance and possibilities is what motivates this research.
1.3 Literature Review

In finding the shortest path on a deterministic network, many algorithms have been developed; being Dijkstra and Bellman-Ford the most popular ones. A comprehensive analysis of algorithms in both static and dynamic (time dependent) networks can be found on Pallotino and Husdal. Additionally, Zhan provides a good comparison of the static algorithms in the Missouri road network.

On dealing with stochastic networks, Frieze considered random arc weights, where distributions between links are assumed to be independent. This work is extended by Cooper. Sigal also considers the problem of random weights by creating optimality indexes for the paths while Wellman analyzes the problem using dominance pruning and a variation of the Bellman-Ford algorithm. Finally, more recent work on transportation networks can be found in Chabini and Miller-Hooks.

For an approximation of the dependency of travel time with the time of day, Kornhauser and Schrader analyze data from the Milwaukee road network and develop a functional form. In more detail, a comprehensive examination of travel time prediction methods can be found on the work by Schblako.

On probability distributions, most of the theoretical work on fittings can be found on Canavos. On the other hand, Carmona develops a framework to explore the data from parametric and nonparametric perspectives. For a theory on Generalized Hyperbolic Functions and its applications to financial data, one can consult Prause. This paper contains the procedure that is applied on this thesis to fit the travel time distributions with the Generalized Hyperbolic model.

Finally, on convoluting lognormal distributions which are correlated, Milevsky and Posner prove that the infinite sum results in the reciprocal gamma distributions. A theory on this distributions is also presented in this paper, and the results are used to price Asian options.

1.4 Objectives of the Research

Current in-vehicle navigation systems can provide pretty accurate shortest paths in terms of distance and reasonable estimates in terms of travel time. To do so, these systems have data on what the
average speeds on the roads are, which are estimated using speed limits and some estimates on actual speeds. These algorithms also take into account penalties for turns, traffic lights, tolls and other obstacles that disrupt free flow. With all these factors, an impedance can be calculated for each road segment in the network and then obtain the required path.

As mentioned before, although this estimate on the fastest path works well when the user wants to get a general idea on what his route should be, the goal is always to provide him with more precise and complete information, which will also allow him to choose from different possibilities according to what his goal is. It is one objective of this thesis to show how the travel time distributions of links in the network can be approximated by theoretical ones, so that the mathematical modeling of the stochastic shortest path can be made in a more realistic fashion.

On the other hand, the amount of data gathered with Copilot is large enough to generate travel time distributions not only on individual links but on paths traveled by individual drivers. Therefore, this thesis will also lever on this data to infer properties on these path distributions as related to the individual links that comprise them. Figure 1.3 shows this problem for a 2-link path.

\[ A \quad (+) \quad B \quad (+) \quad C \]

Figure 1.3: The two link problem
1.5 Organization of the Thesis

The remainder of this thesis is organized as follows. The first chapter describes the in-vehicle navigation system, Copilot®, which was used to gather the data. It also explores the inherent sources of error and bias that comes from using a navigation system of this kind to collect it. In Chapter 3, an initial exploration of the travel times is presented. The nature of their randomness will be explained supported on the data, and special attention will be given to large travel times in their qualification as outliers. Furthermore, the theory of Generalized Hyperbolic distributions will be discussed as well as its usefulness in representing the travel times.

Then, in Chapter 4, the fitting using the actual data is developed, and different theoretical distributions are compared. This is done for several links of different road types so that comparisons can be made in terms of time dependence, expected travel times and variability. Further, in Chapter 5, the issue of independence is addressed by analyzing the distribution of the different paths with the links that comprise them.

Finally, we conclude on the results obtained in the previous chapters and indicate some of the directions where the research can be continued and expanded.
Chapter 2

2 The Road Network

In order to understand how the data was gathered, it is important to understand the in-vehicle navigation system Copilot® and its road network. This product, developed by ALK Technologies, Inc., is a navigation system for PCs and PDAs that contains all road information on North America, Western Europe and Australia. It helps the user find the best path from A to B in the network, using impedance for the travel time approximation, and, when attached to a GPS receiver, reacts in real time when the user deviates from the chosen path.

2.1 Copilot and its network

The road network in Copilot® is naturally represented by a set of nodes and arcs, where a node can be thought as an intersection or branching point, and an arc is a road segment that connects two intersections. Figure 2.1 presents an example of this network near Princeton NJ.

The network contains approximately 5 million nodes and 30 million arcs in the US. However, the arcs are separated in different levels, which makes the system more efficient. For example, if the driver wants to go from Miami to New York, there is no point in checking all the local roads in North Carolina, since it is clear that in this area the safest bet is to drive on an interstate or a highway. Therefore, arcs from these types of roads are on a basic level (Level 3), while the lowest level contains all arcs.

The following is the approximate number of arcs per level:
2.2 The notion of monuments

Before describing the data itself, it is necessary to explain the concept of monuments in the Copilot network. Monuments are midpoints of certain arcs in the Level 1 network, and they were originally created in order to build the network from the bottom up. In total, there are around 280,000 monuments in the US, and Figure 2.2 presents the ones around the Princeton area (marked by red square boxes). Therefore, data in the network was not collected between nodes, as they were defined in the previous section, but rather between these monuments.

Evidently, by using monuments, the network is defined differently than the classical road network, since the nodes in the network are not intersections but the midpoints of road segments, and the arcs are the segments connecting these midpoints. This presents an additional problem, since assigning a cost like distance to these segments will mean that the exact path traveled between the
monuments must be known. From Figure 2.2, it is clear that there are several paths between any two monuments.

However, the data for these monuments is easier to tract and manipulate (because of its smaller size). If we were to take measures at regular nodes (intersections), assigning a time stamp would present a major drawback, since it is not known if the driver is moving or not. When taking the time between midpoints of links, these measurements already have embedded in them whatever happened in the intersections between the monuments. Of course, there is also uncertainty of movement at the intersections, but this error is lower since the probability of not moving at a monument is less than at an intersection.

2.3 Data Collection

The data collected for this thesis comes from individual drivers who are commercial users of the Copilot® software, and who willingly sent the data of their travel trips. Copilot® is able to record every 3 seconds the position of the driver as well as its heading and speed, which is good enough (although not perfect) for our analysis. Figure 2.3 shows how this data is presented graphically in the navigation system.

Each data point is represented by a colored dot with a tail that indicates its heading. Each
measurement contains the following items: Vehicle ID, Location (Latitude and Longitude), Heading, Speed, Date and Time (in Universal Time units). As can be seen in the Figure, it is rarely that a data point (called GGD point) matches perfectly the road traveled since Global Positioning Systems are not 100% accurate. This represents another difficulty, since each GGD point is associated or “snapped” to a link in function of its proximity to the link and heading difference; conflicts will naturally arise near branching points.

In addition to this approximation, the data points for monuments are not recorded exactly when the driver is passing through a monument, since the three second precision is too big to capture this. However, by means of interpolation and using the data points immediately before and immediately after the monument, one is able to obtain an approximation of the time when the driver passes by the monument. Yet, not all links have two data points to interpolate with, but one. In this case, using the speed, time and position of this one point, we can extrapolate and find the time at the midpoint of the link using a simple speed formula.

For the purpose of this study, data was collected between May 1st, 2000 and January 25th, 2004. For this time span, 1,608,430 monument to monument (m2m) were obtained, which were distributed over 171,973 m2m pairs. However, only 1,818 (1.06%) of these pairs had more than
100 observation and 35,236 (20.49%) had more than 10 observations, being 1,118 the maximum number of monuments for a single m2m pair. Figure 2.4 shows how the number of observations was distributed among the m2m pairs (for pairs with more than 10 observations):

![Figure 2.4: Observation spread over m2m pairs](image)

**2.4 Road Type**

For the analysis, it is also important to discuss briefly the road types that are defined in Copilot® for the road network. Table 2.1 shows this categorization. all categories are self explanatory, with examples from the Princeton area which help to emphasize the differences between them.

On the forthcoming chapters, arcs from different road types are going to be examined, since their traffic patterns are inherently different. For example, road type 3 usually has traffic stops that interstates don’t, and this is clearly an important factor that disrupts traffic flow. Road types 5 and 7 are not pertinent to our analysis since they don’t have any monuments (and clearly 5 is not even a road).

Most of the m2m pairs that have sufficient data for the analysis are of road type 1 and 2, since
<table>
<thead>
<tr>
<th>Road Type</th>
<th>Description</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,2</td>
<td>Interstate</td>
<td>I-95, I-295</td>
</tr>
<tr>
<td>3</td>
<td>Divided</td>
<td>US 1</td>
</tr>
<tr>
<td>4</td>
<td>Arterial</td>
<td>Washington Road (County Highway 571), Nassau Street (NJ 27)</td>
</tr>
<tr>
<td>5</td>
<td>Ferry</td>
<td>NA</td>
</tr>
<tr>
<td>6</td>
<td>Primary</td>
<td>Alexander Street, Hamilton Avenue</td>
</tr>
<tr>
<td>7</td>
<td>Ramp</td>
<td>NA</td>
</tr>
<tr>
<td>8</td>
<td>Local</td>
<td>Faculty Road, Prospect Avenue</td>
</tr>
</tbody>
</table>

Table 2.1: Road Types

these are the backbone arcs use to create the network, as explained when discussing the notion of monuments. However, few roads of type 3 and 4 have more than 100 observations, limiting the scope of our analysis on individual links of these types.

### 2.5 Time of Day

Like road type, time of day is another major factor that explains travel time on a given road segment. The denominated “rush-hour” period is the most classical example that illustrates travel time variation during the day. Clearly, this is a direct consequence of volume capacity on the roads, and several mathematical models have been developed to describe their relationship (see []). Take for example 1.2, which corresponds to a m2m pair on Route 601 near Princeton. In the afternoon, it can be seen that many drivers experience low speeds and high travel times. Although this can be used to model the time dependency of travel times as explained on Chapter 1, one might also want to divide the day into different travel intervals and examine the travel time distributions of each one. Serious analysis should then be made on how to categorize these intervals for each specific case. However, the approach on this thesis is simple, using the intervals described on Table 2.2.

Another issue that is addressed when manipulating the data is Universal Time. Data must be converted to local time, in order to categorize it in its correct interval. This is indeed a challenging task due to the complex division of the US territory in time zones, as well as the daylight saving times.
<table>
<thead>
<tr>
<th>Time of Day</th>
<th>Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6 a.m. to 9 a.m.</td>
</tr>
<tr>
<td>2</td>
<td>9 a.m. to 3 p.m.</td>
</tr>
<tr>
<td>3</td>
<td>3 p.m. to 6 p.m.</td>
</tr>
<tr>
<td>4</td>
<td>6 p.m. to 9 p.m.</td>
</tr>
<tr>
<td>5</td>
<td>9 p.m. to 6 a.m.</td>
</tr>
<tr>
<td>6</td>
<td>Weekend 9 a.m. to 9 p.m.</td>
</tr>
<tr>
<td>7</td>
<td>Weekend 9 p.m. to 9 a.m.</td>
</tr>
</tbody>
</table>

Table 2.2: Times of Day

A rather robust approach was taken to do these conversions, using approximate time zones that might not be accurate in some zones of the country but are appropriate for the Northeast area, where the bulk of the data is. Since we have the longitude for every data point, we can convert the Universal Time to local time using Table 2.3.

<table>
<thead>
<tr>
<th>Longitude</th>
<th>Time Conversion (Daylight Savings)</th>
<th>Time Conversion (Winter)</th>
</tr>
</thead>
<tbody>
<tr>
<td>≥-85</td>
<td>UT - 4</td>
<td>UT - 5</td>
</tr>
<tr>
<td>&lt;-85 and ≥-103</td>
<td>UT - 5</td>
<td>UT - 6</td>
</tr>
<tr>
<td>&lt;-103 and ≥-114</td>
<td>UT - 6</td>
<td>UT - 7</td>
</tr>
<tr>
<td>&lt;-114</td>
<td>UT - 7</td>
<td>UT - 8</td>
</tr>
</tbody>
</table>

Table 2.3: Conversion of Universal Time to Local Time for US

### 2.6 Data Bias and Sources of Error

Some sources of error and bias have already been discussed throughout this chapter, but it is appropriate to summarize them here and extend new ones:

1. GGD snapping: As mentioned, GGD points do not match perfectly with road segments, due to the inaccuracies inherent to GPS systems as well as to the inaccuracy on the map layout on the program (since the map is approximated by straight lines and occasionally shape points,
it is not expected to resemble perfectly the shape of the actual road). Therefore, in some cases, GGD points are snapped to an arc to which they do not correspond. This is most critical near intersections and branching points, but since monuments are midpoints of links, this shouldn’t represent a big issue in our analysis.

2. Time interpolation: For monuments for which time is obtained via interpolation, we assume that the speed between the two points of the interpolation is constant. Needless to say, this tends not to be the case. Additionally, some of the monuments do not have two points to interpolate between, but rather the position of the driver and its speed, so using these the time at the monument is calculated using extrapolation also assuming constant speed.

3. Geographical bias: Since the data that is being analyzed is sent willingly by Copilot® users, the sample is going to be biased towards these users and not the whole population of users. Furthermore, since ALK, Inc. is Princeton based, a great percent of these users are Princeton students and costumers that have a personal of professional liaison with the company. Therefore, most of the data received comes from the Princeton area, and in a more general sense, from the Northeast. This limits the analysis towards other parts of the country, since most of the m2m pairs that have sufficient data are from the mentioned region.

4. User bias: Clearly not every American uses Copilot®, since it is a PDA product, and most of its users are business executives, professional workers and students. Then, the analysis will end up giving the travel behavior of this sample of the population, which might differ from other type of drivers.

5. Speed calculation: Further on, in order to compare travel times between links, they need to be normalized. To do this, the distance between two monuments is needed. However, since monuments are not on every link on the network, we would need to determine the exact path that connects the monuments. Since there might be more than one path, we assume a path that is a straight line between them. This of course will yield a speed (normalized time) that will be lower (higher) than the actual one, since the path chosen is the shortest path possible. To obtain the distance, we can again use interpolation to find the latitude and longitude at
the monuments, and then use the following approximation:

\[
D = \arccos(\cos(Lat_1) \cos(Lon_1) \cos(Lat_2) \cos(Lon_2) \\
+ \cos(Lat_1) \sin(Lon_1) \cos(Lat_2) \sin(Lon_2) \\
+ \sin(Lat_1) \sin(Lat_2)) \ast R
\]

where \( D \) is the distance, \( R \) is the earth’s radius and \( Lat_1, Lon_1, Lat_2 \) and \( Lon_2 \) are the latitudes and longitudes of monuments 1 and 2. For cases where there is only one data point to obtain the monument, we can not interpolate, so we use the latitude and longitude of the monument.
Chapter 3

3 Random Arc Costs

In this chapter, the data collected is analyzed from a descriptive point of view, understating the behavior of the travel times with respect to the road type and time of day factors. The methodology to fit distributions is going to be explained, with special attention to heavy tails. For this purpose, a summarize on the theory on Generalized Hyperbolic distributions is presented. Finally, we present the Reciprocal Gamma distribution and its relation to the sum of lognormal distributions.

3.1 The Nature of Random Arc Costs

As mentioned previously, it is reasonable to expect that travel times between two nodes on a road network are of a probabilistic nature. Moreover, since incidents on the network like accidents, severe weather conditions and traffic jams augment the travel time and there will be always be a lower bound on how fast the driver can go, it is expected that this distribution is going to be heavy tailed on the right.

Take for example the distribution of travel times shown in Figure 4.28. It corresponds to a segment in I-295 near Trenton NJ (road type 1). We must first point out that this distribution was obtained after removing data points with extremely high values (This issue is considered in the next section). Without these points, we can see that the distribution is heavy tailed to the right and, at first glance, it does not seem to have a Gaussian behavior.

Figure 4.26 presents the distribution between two monuments on arterial roads near Princeton NJ. Although similar to the previous distribution, it exhibits a higher concentration of data points
near the tail in the form of bumps.

These different patterns are a result of the distinct traffic characteristics that operate on each road type, as explained before. The heavy tail on the distribution for road type 4 can be described as composed of several “bells”, which might correspond to different volume on the roads throughout the day. Figure 3.3 shows this dependence for the first m2m pair and Figure 1.2 shows it for the second one.

From these figures we can conclude that time of day does not affect particularly the travel time on the particular segment of the interstate, while it does for the Princeton roads. The big bulk of data around 30,000 seconds from midnight (between 8 and 9 am) indicates that this is a route that gets highly congested in the morning, and this accounts for the heavy tail on the histogram.

This then leads to the problem of fitting a theoretical distribution to these patterns. The challenge seems more difficult for the Princeton arteries because of its different bell shapes, while the distribution for the interstate presents a distinguishable bell shape with exponential decay to the right that suggests considering distributions like lognormal, Weibull and gamma.
However, the heavy tail is a limitation on these distributions and there is the need to consider it. Two aspects then should be examined: how can the heavy tails be fitted, and which points should be considered as outliers. For the first issue, Generalized Hyperbolic distributions come into the play and they will be presented further on in this chapter. For the second one, in the following section we develop a procedure to remove points with high values. The characteristics of the heavy tail will depend on how many points we remove with this procedure, so the two issues are closely interrelated and will affect each other’s results.

### 3.2 The treatment of high travel times

On categorizing a data point as an outlier, one should examine the causes that lead that specific driver to have such a high travel time, meaning that its value deviates largely from the mean of the sample. One should not separate one of these points from the sample just for having such a value, since they can actually represent a realization of a trip from one monument to another.
Figure 3.3: Time dependency for a segment of road type 1

Consider the case that there is a traffic jam between the two monuments A and B. Clearly, the travel time measured during this situation is meaningful, since the driver did incur on this travel time. On the other hand, consider the case when a driver is pulled over by a policeman. The amount of time the driver is without moving is not representative of the travel time between the monuments, since it is unrelated to traffic conditions on the segment. A more complex case arises when the driver did travel from A to B by taking a path which is longer to the average path (the one that the bulk of the drivers took). Clearly, the travel time of this driver is not representative of the m2m pair, since we are assuming as our arc the path that the average driver takes (in most cases the shortest one).

A major challenged is faced then here. There is nothing in our data that will indicate what happens to the driver in between the two monuments; it can be that he was delayed by traffic or that he stopped at some place. And while data points that belong to the first case are clearly pertinent to the distribution analysis, points that belong to the second one are true outliers and should be removed from the sample. Unfortunately, there is no quantitative way to differentiate
between them.

Nevertheless, some of these points should be removed to make the distribution analysis more relevant. Evidently, trying to fit a distribution to the sample of times of the m2m pair in I-295, represented by the histogram in Figure 3.4 is extremely difficult and not useful.

![Histogram of travel times](image)

Figure 3.4: All travel times for a segment of road type 1

The approach chosen is to divide the sample into two groups, the outlier candidates and the bulk of the data, by doing the following:

1. Take the sample and find its median.
2. Obtain the standard deviation with respect to the median (not the mean).
3. Remove the highest travel time value and repeat step 2.

Figure 3.5 shows the behavior we obtain repeating step 2 until 20% of the points are removed. It is characterized by a steep decrease of the standard deviation when the first points are removed, followed by an asymptotical behavior towards a constant value. We want to remove the points before the start of this asymptote, in what is defined as the “knee” of the curve.

The second derivative of the curve is taken and we approximate the knee by the point where this
derivative is zero (a tolerance of 0.005 is used). This criterion follows from the fact that where the asymptote begins, the slope of the curve becomes constant. Of course, one major exception would be when the curve shown above is a straight line. However, in the tests made none of the links selected presented this behavior. Furthermore, if some point does have it, it would mean that the shape of the distribution would not follow the pattern of the distributions that are being analyzed, characterized by a bell shaped curve and a heavy tail.

### 3.3 Generalized Hyperbolic Distributions

Generalized Hyperbolic (GH) distributions were developed by Barndorff-Nielsen [] in order to model the distributions of grain sizes of wind blown sands. They possess the properties required for the modeling of travel time data, i.e., heavy tails and a high concentration of points around the median (this is known as leptokurtosis, see Prause []). For these same reasons, they have also been used to model increments of financial price processes, for the purpose of assessing risk[].

![Figure 3.5: Standard Deviation as points are removed](image-url)
GH distributions are given by the following density function:

\[
gh(x; \lambda, \alpha, \beta, \mu, \delta) = a_{\lambda}(\delta^2 + (x - \mu)^2)^{\lambda-1/2} K_{\lambda-1/2}(\alpha \sqrt{\delta^2 + (x - \mu)^2})e(\beta(x - \mu))
\]

\[
a_{\lambda} = a_{\lambda}(\alpha, \beta, \delta) = \frac{\alpha^2 - \beta^2)^{\lambda/2}}{\sqrt{2\pi\alpha^\lambda - \beta^2}} K_{\lambda}(\delta \sqrt{\alpha^2 - \beta^2}); \quad x, \mu \in \mathbb{R}
\]

\[
\delta \geq 0, |\beta| < \alpha \quad \text{if } \lambda > 0
\]

\[
\delta > 0, |\beta| < \alpha \quad \text{if } \lambda = 0
\]

\[
\delta > 0, |\beta| \leq \alpha \quad \text{if } \lambda < 0
\]

\[K_{\lambda}\] is a modified Bessel function which solves the differential equation

\[
z^2 \frac{d^2 y}{dz^2} + z \frac{dy}{dz} + (z^2 - v^2)y = 0
\]

A major drawback of this distribution is that the meaning of the parameters is not clear. Nevertheless, one can get a broad sense of its shape with \(\mu\) and \(\delta\), which describe respectively the location and scale of the distribution. Moreover, the normal distribution is obtained as a limiting case for \(\delta \to \infty\) and \(\delta/\alpha \to \sigma^2\).

Normal Inverse Gaussian (NIG) distributions and Hyperbolic (HYP) distributions are subclasses that are obtained for specific values of \(\lambda\). For \(\lambda = -1/2\), the NIG is defined by:

\[
nig(x; \alpha, \beta, \mu, \delta) = \frac{\alpha \delta}{\pi} e(\delta \sqrt{\alpha^2 - \beta^2} + \beta(x - \mu)) \frac{K_1(\alpha \sqrt{\delta^2 + (x - \mu)^2})}{\sqrt{\delta^2 + (x - \mu)^2}}
\]

\[x, \mu \in \mathbb{R}, \quad \delta \geq 0, \quad |\beta| < \alpha\]

while for \(\lambda = 1\), the hyperbolic distribution is obtained:

\[
hyp(x; \alpha, \beta, \mu, \delta) = \frac{\sqrt{\alpha^2 - \beta^2}}{2\delta \alpha K_1(\delta \sqrt{\alpha^2 - \beta^2})} e(-\alpha \sqrt{\delta^2 + (x - \mu)^2} + \beta(x - \mu))
\]

\[x, \mu \in \mathbb{R}, \quad \delta \geq 0, \quad |\beta| < \alpha\]

### 3.3.1 Parameter Estimation

For the travel time data, the parameters of the GH distributions are estimated by maximizing the log-likelihood function. This optimization is performed sequentially on each of the parameters \((\mu, \delta, \lambda, \beta, \alpha)\), using the fminsearch function in Matlab. Bessel functions are also evaluated in Matlab and, for the samples chosen, all the results satisfy the constraints on the parameters.
Key to the optimization algorithm is a good choice of the starting value. Since the travel times on the links analyzed are close in value for different links, the same starting value is used for all the links, always obtaining convergence of the maximization. Mr. Ernst August Freiherr von Hammerstein, from the Institute for Stochastic Mathematics in Freiburg, Germany, provided this initial value for the data, using specialized programs developed at this institute for the analysis of financial returns using GH distributions.

3.3.2 Goodness of Fit

The results are presented for the GH distributions as well as for several other classical distributions (lognormal, beta, gamma, normal). Density and cumulative plots are presented to obtain a visual assessment of the fits. Additionally, goodness of fit measures are necessary to evaluate numerically the fits. For this purpose, $L_1$, $L_2$, Kolmogorov-Smirnov and Anderson-Darling statistics were chosen.

$L_1$ and $L_2$ are simply the sums of the distances between the estimated c.d.f.s and their empirical values, with absolute distance for $L_1$ and distance squared value for $L_2$. Kolmogorov-Smirnov distance is defined as

$$ KS = \max_{x \in \mathbb{R}} |F_{emp}(x) - F_{est}(x)| $$

Where $F_{emp}$ is the empirical c.d.f. and $F_{est}$ is the value obtained from the estimated theoretical distribution. On the other hand, the Anderson-Darling statistic is given by

$$ AD = \max_{x \in \mathbb{R}} \frac{|F_{emp}(x) - F_{est}(x)|}{\sqrt{F_{est}(x)(1 - F_{est}(x))}} $$

The difference between the $AD$ statistic and the $KS$ statistic is the denominator on the first one. This denominator actually gives more importance to the tails of the distribution, and therefore is useful in the sense that this is the particular factor that is being fitted more carefully.

In both the c.d.f. plots as well as in the statistics, the cumulative functions need to be evaluated. For the calculation of these cumulative functions, Reinmann sums were used to approximate the integration of the density function for GH distributions.
### 3.4 Reciprocal Gamma Distributions

A random variable $X$ is gamma distributed if its probability density function is:

$$f(x; \alpha, \beta) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} x^{\alpha-1} e^{-x/\beta} \quad \forall x \geq 0$$

where $\alpha > 0$, $\beta > 0$, and $\Gamma(\alpha)$ is the gamma function defined by

$$\Gamma(\alpha) = \int_{0}^{\infty} x^{\alpha-1} e^{-x} dx.$$

If $X$ is a gamma random variable with cumulative distribution $F(x; \alpha, \beta)$, then the continuous random variable $Y = 1/X$ is a gamma reciprocal random variable described by the following density and cumulative distribution functions:

$$f_R(y; \alpha, \beta) = \frac{f(1/y; \alpha, \beta)}{y^2}$$

$$F_R(y; \alpha, \beta) = 1 - F(1/y; \alpha, \beta) \quad \forall y > 0.$$

On the other hand, a continuous random variable $X$ is lognormally distributed if $Y = \ln(X)$ is normally distributed. The probability density function of $X$ is given by

$$f(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma x} e^{-[\ln(x)-\mu]^2/(2\sigma^2)} \quad \forall x \geq 0.$$
so in this sense, it is difficult to conclude on how the parameters are related. However, reciprocal gamma distributions are an attractive possibility on modeling these distributions since we will see that the distributions on the links are correlated and the best candidate to model each distribution is the lognormal distribution.
Chapter 4

4 Numerical Results for Distribution Fittings

In the last chapter, we hinted at the characteristics that the travel time distributions have. It is now our purpose to dwell deeper on these characteristics and fit theoretical distributions to the empirical histograms. In doing so, we start by selecting several arcs from the network, since we have 25,236 arcs with a reasonable amount of data. We fit the distributions on three of these arcs and analyze their corresponding statistics. We then select groups of arcs by road type and observe their speed distributions, with the objective of deriving common distributions for arcs of the same road type.

4.1 Numerical Results for Individual Arcs

As mentioned on Chapter 2, the data collected using Copilot® is biased towards the Northeast region of the country. Since our goal is to analyze arcs with a large number of observations, most of these arcs will be located in this region. It is also our purpose to perform the analysis for different road types, so the arcs selected were the ones with the most number of observations for road types 1, 3 and 4. Only these road types had m2m pairs with more than 100 observations, which we consider a reasonable threshold for the study to be meaningful.
4.1.1 Case 1: Road Type 1

All m2m pairs that have 1000 or more observations are of road type 1, which is a natural consequence of the definition of monuments. The m2m pair that contains the maximum number of observations corresponds to monuments A and B displayed in Figure 4.1. It is natural to consider the road that connects A and B as our arc. This corresponds to a segment of I-295 in New Jersey, which is an interstate that stretches from the Delaware Memorial Bridge to the Trenton area. It is the interstate that lies closer to Princeton, so it is no surprise that the m2m pair with the largest amount of data lies on it.

![Figure 4.1: I-295 near Trenton NJ](image)

We first start our analysis by presenting in Figures 4.2 and 4.3 the histogram and empirical cumulative distribution function for the data. Although most of it is concentrated around low values of travel time, we still have some points with large values. Since we don’t know their nature, our best hope is that they are outliers that do not affect the travel time distribution, that is, they come from stops, alternate paths or errors in their measurement.

The next step is to apply our heuristic for removing these outliers. To do so, we calculate the
Figure 4.2: Travel Time histogram for I-295

Figure 4.3: Cumulative Distribution of Travel Times for I-295
median and standard deviations of the data (with respect to this median) as we remove one point at a time, obtaining the plot shown in Figure 4.4. We notice a steep descent of the standard deviation, stabilizing after approximately 1% of the points have been removed. Using our second derivative criteria, the number of actual points removed from the set is 12 (1.07%) out of 1118. After removing these points, the histogram showed on Figure 4.5 is obtained.

![Figure 4.4: Removal of high values for I-295](image)

In this histogram we see that the travel time is concentrated with a bell-shaped figure around 75 seconds, but still some high values remain on the right that refrains us from assuming Gaussian distribution. In addition, another major constraint on assuming normality is the required positiveness of travel times, suggesting that a lognormal distribution might actually do the work.

However, before examining the distributions themselves, it is necessary to explore the time dependence properties of the travel time. Our hypothesis is that there can be a certain time frame or rush-hour period in the day where high values appear, and a possible strategy would be to remove these points and fit the distributions to travel times under normal conditions. In addition, we will fit another distribution to the data points belonging to the removed interval. For the shortest path problem, this would mean that both the random and time dependent arc cost perspectives should
be considered.

Figure 4.6 displays this time of day dependency for our I-295 arc. We can see that there isn’t any significant difference of the travel times during the day. Although the highest value comes from a time period that we might tend to consider to be rush hour (6 p.m. - 9 p.m.), we also notice that for the same time period there are very low travel times. Moreover, Table 4.1 shows that the mean and standard deviation for this time period is not significantly different to that of the rest of the sample. Figure 4.7 presents the c.d.f. for the travel time where each color corresponds to a time period, and we can see that all the colors are distributed over the whole range.

Therefore, we take the whole sample of data points and perform the distribution fittings. Performing this fitting for the normal, lognormal, gamma and Weibull distributions is straightforward. On the other hand, we use the optimization procedure described on Chapter 3 to fit a Generalized Hyperbolic distribution to the data. To do this, we use the following initial solution:

\[
\lambda = -0.65527288 \quad \alpha = 0.02841332 \quad \beta = 0.02607272 \quad \mu = 48.80583846 \quad \delta = 2.72620337
\]

The results are shown in Figures 4.8 and 4.9. First, from the histogram we notice that the GH
Figure 4.6: Travel Time during the day for I-295

Figure 4.7: Cumulative distribution of travel times without high values for I-295
<table>
<thead>
<tr>
<th>Time of Day</th>
<th>Mean</th>
<th>St. Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>73.27044</td>
<td>3.893379</td>
</tr>
<tr>
<td>2</td>
<td>74.71077</td>
<td>4.684149</td>
</tr>
<tr>
<td>3</td>
<td>74.60109</td>
<td>4.813776</td>
</tr>
<tr>
<td>4</td>
<td>71.86014</td>
<td>5.696717</td>
</tr>
<tr>
<td>5</td>
<td>72.37489</td>
<td>5.334079</td>
</tr>
<tr>
<td>6</td>
<td>75.61176</td>
<td>5.656150</td>
</tr>
<tr>
<td>7</td>
<td>NA</td>
<td>NA</td>
</tr>
</tbody>
</table>

Table 4.1: Statistics by Time of Day

distribution seems to fit better the high concentration around the mean. However, we should note that the height for the intervals on a histogram depend upon the choice of their width, so a qualitative appreciation of the fit is not accurate. For this reason, we should focus the attention on the c.d.f. plot shown on Figure 4.9.

![Density fittings for I-295](image)

Figure 4.8: Density fittings for I-295

On one hand, we see how the GH distribution has a heavier tail than all the other distributions
Figure 4.9: c.d.f. fittings for I-295

tested. But, on the fit of the body, the distribution deviates largely from the empirical c.d.f. This is indeed the tradeoff that comes with GH distributions: as more mass gets concentrated on the tail, the distribution is more peaked around its maximum. The parameters obtained for this distribution are:

\[
\lambda = -0.7928 \quad \alpha = 0.0583 \quad \beta = 0.0213 \quad \mu = 71.9675 \quad \delta = 3.6278
\]

On the other hand, all the other distributions seem to have a better fit, except for the Weibull distribution. If we look at low values of travel time, the lognormal distribution seems to do a good job, with an almost perfect fit. The gamma distribution behaves very similar, with a fit almost indistinguishable from the one of the lognormal.

Table 4.2 compares all the statistics for the distributions tested. We can see that for this arc the GH distribution has higher statistics, and since we are talking about deviations from the empirical distributions, the fit performs poorly in comparison to other distributions. Even the normal distribution seems to perform better, although the lower AD statistic indicates that the GH distribution does a better fit on the tail. The best fit, in terms of all statistics, is obtained by the
lognormal distribution, with statistics slightly better than those of the gamma distribution. This helps us to choose between the two, which we were unable to do by simply looking at the plots.

<table>
<thead>
<tr>
<th>Distribution</th>
<th>AD</th>
<th>KS</th>
<th>L1</th>
<th>L2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>0.3262475</td>
<td>0.07685269</td>
<td>0.03452969</td>
<td>0.0017115102</td>
</tr>
<tr>
<td>Lognormal</td>
<td>0.1926083</td>
<td>0.06240267</td>
<td>0.02630589</td>
<td>0.0010215136</td>
</tr>
<tr>
<td>GH</td>
<td>0.2582971</td>
<td>0.10015379</td>
<td>0.03790891</td>
<td>0.0024710611</td>
</tr>
<tr>
<td>Gamma</td>
<td>0.2284630</td>
<td>0.06728824</td>
<td>0.02900753</td>
<td>0.001230432</td>
</tr>
<tr>
<td>Weibull</td>
<td>0.9743608</td>
<td>0.11418452</td>
<td>0.06715731</td>
<td>0.005747394</td>
</tr>
</tbody>
</table>

Table 4.2: Fitting Statistics for I-295

Although for this arc we find that the GH fit does not perform better than other distributions, this is not always the case. Take for example the data for this same arc on the opposite direction, that is, from B to A. After removing 18 high values (1.71%) from 1050 observations, and performing the fits, we get the results shown in Table 4.3. The parameters obtained for the GH fit are:

$$\lambda = -0.7941 \quad \alpha = 0.0511 \quad \beta = 0.0210 \quad \mu = 70.7220 \quad \delta = 3.7151$$

<table>
<thead>
<tr>
<th>Distribution</th>
<th>AD</th>
<th>KS</th>
<th>L1</th>
<th>L2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>7.9169128</td>
<td>0.09564398</td>
<td>0.04453154</td>
<td>0.0029402617</td>
</tr>
<tr>
<td>Lognormal</td>
<td>1.3740114</td>
<td>0.07630413</td>
<td>0.03355134</td>
<td>0.0017153656</td>
</tr>
<tr>
<td>GH</td>
<td>0.1988517</td>
<td>0.08189845</td>
<td>0.03702799</td>
<td>0.0018606752</td>
</tr>
<tr>
<td>Gamma</td>
<td>2.0665746</td>
<td>0.08289197</td>
<td>0.03717582</td>
<td>0.002092110</td>
</tr>
<tr>
<td>Weibull</td>
<td>239.1041158</td>
<td>0.17025242</td>
<td>0.08808213</td>
<td>0.010253795</td>
</tr>
</tbody>
</table>

Table 4.3: Fitting Statistics for I-295 in opposite direction

For this direction we get an AD statistic that is significantly lower for the GH distributions than for all the other distributions tested. This tells us that the GH distributions does a better job in fitting the heavy tail than the lognormal, normal and gamma densities. However, the rest of the statistics still suggest a better fit of the lognormal function for the body of the data. In this case, the Weibull fit is so bad at the tails that we get a high Anderson Darling statistic.
These two results can be seen explicitly on Figures 4.10 and 4.11. The relatively higher L1, L2 and KS statistics agree with the sharp peak of the GH density with respect to the histogram, while the lower AD value represents a slower decay of the density to the right as compared to the other distributions. This is more evident on the c.d.f. plot, where the c.d.f.s for these distributions approach 1 faster than the GH and empirical c.d.f.s.

![Figure 4.10: Density fittings for I-295 in opposite direction](image)

Finally, for both directions we get similar mean and standard deviations, as shown in Table 4.4. This, along with the results above, suggests that traffic flow behaves similarly in both directions.

<table>
<thead>
<tr>
<th>Direction</th>
<th>Mean</th>
<th>StDev</th>
</tr>
</thead>
<tbody>
<tr>
<td>A to B</td>
<td>73.46438</td>
<td>5.424696</td>
</tr>
<tr>
<td>B to A</td>
<td>72.22047</td>
<td>6.0448</td>
</tr>
</tbody>
</table>

Table 4.4: General Statistics for I-295
4.1.2 Case 2: Road Type 3

We now turn our analysis to a m2m pair where the road types of both monuments is Divided (road type 3). Figure 4.12 shows the arc chosen with our maximum observation criteria (472 from A to B, 396 from B to A), which represents a segment of route US 1 near New Brunswick NJ. We expect a different travel time pattern as compared to the interstate since this segment has traffics lights, which undeniably distort free flow travel behavior.

We start our process by eliminating those data points with high travel time values using the second derivative criteria. Analyzing both directions at the same time, we obtain that 39 points (8.26%) are removed for direction 1 (A to B) while 16 (4.04%) are removed for direction 2 (B to A). It is important to notice the relatively larger amount of points removed for this road type than for I-295 (around 1%). Although a route of this type, with traffic lights, tends to get congested more easily because of the interruptions induced on the flow, it is also true that drivers on these routes have more reasons to stop (restaurants, gas stations, etc.). Therefore, it is not easy to hypothesize on the reasons for having such a larger amount of high values, and we can just hope that the points removed are not representative of the travel times in between the monuments.
Next, we look into the variation of these travel times as the day progresses. Figure 4.13 displays the travel time as a function of time of day for direction 1 while Figure 4.14 displays it for direction 2. We see that there seems to be two different segments of the day where the travel times are significantly different. Between 6 a.m. and 10 p.m. (20000 to 80000 seconds from midnight approximately), the travel times present higher values than the rest of the day. This is rather logical, since between 10 p.m. and 6 a.m. Route 1 presents free flow characteristics, while congestion is characteristic during the day. Moreover, these higher travel times are specific for direction 2 between 12 p.m. and 10 p.m. (40000 - 80000 seconds from midnight), suggesting a rush hour period for this direction on the afternoon.

Unfortunately, we can also notice in these plots that the amount of data for the free flow periods is very small, raising the question on the convenience of performing a distribution fitting for each time interval. If we had more data, this approach could be taken. For the moment, we just fit our different candidate distributions to all the data. The results are presented in Table 4.5 and Figures 4.15 and 4.16 for direction 1, with parameters

\[
\lambda = -0.7231 \quad \alpha = 0.0278 \quad \beta = 0.0248 \quad \mu = 292.7400 \quad \delta = 44.9964
\]
Figure 4.13: Travel time on US1 during the day from A to B

Figure 4.14: Travel time on US1 during the day from B to A
for the GH distribution, and in Table 4.6 and Figures 4.17 and 4.18 for direction 2 with the following parameters:

$$
\lambda = -0.8302 \quad \alpha = 0.0264 \quad \beta = 0.0223 \quad \mu = 288.0703 \quad \delta = 39.9622
$$

Figure 4.15: US1 density fit for direction 1

<table>
<thead>
<tr>
<th>Distribution</th>
<th>AD</th>
<th>KS</th>
<th>L1</th>
<th>L2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>0.3935794</td>
<td>0.13099832</td>
<td>0.06133044</td>
<td>0.005225290</td>
</tr>
<tr>
<td>Lognormal</td>
<td>0.2388515</td>
<td>0.08275059</td>
<td>0.03586091</td>
<td>0.001761431</td>
</tr>
<tr>
<td>GH</td>
<td>0.4236994</td>
<td>0.08742381</td>
<td>0.03252682</td>
<td>0.001613518</td>
</tr>
<tr>
<td>Gamma</td>
<td>0.2771806</td>
<td>0.09938515</td>
<td>0.04435360</td>
<td>0.002718938</td>
</tr>
<tr>
<td>Weibull</td>
<td>0.3028461</td>
<td>0.13289237</td>
<td>0.06941200</td>
<td>0.006549713</td>
</tr>
</tbody>
</table>

Table 4.5: Fitting Statistics for US1 in direction 1

Several things can be concluded. First, the GH distributions do not fit acceptably the data, since all its statistics are lower than those of all the other distributions for both directions. The lognormal seems to be a better candidate.
Figure 4.16: US1 c.d.f. fit for direction 1

Figure 4.17: US1 density fit for direction 2
If we look at the plots for direction 1, we can see that although the GH distribution approximates the empirical cumulative distribution on the right tail, it performs poorly for low travel time values. This case is reversed for the other functions, they do well for low values but have lighter tails than the ones that the empirical data suggest. For direction 2, the situation is worse for the GH distribution, where it doesn’t perform well in general for all values. This is reasonable if we look at the histogram; the empirical data is not as skewed to the left as for the other distributions we have previously analyzed. The GH distribution even has a heavier right tail than the one suggested by...
the data. Needless to say, for all cases, the normal Gaussian distribution is not the better candidate we can find to approximate the data.

4.1.3 Case 3: Road Type 4

For road type 4, we chose the arc displayed in Figure 4.19. The first monument is located on a segment of Route 601, just by Skillman Road, while the second monument is on Route 206 near its intersection with County Highway 630.

![Figure 4.19: M2M pair of road type 4](image)

We first analyze the travel time from A to B, which is not too different from the analysis of the previous monument to monument pairs. The total number of observations gathered was 476, and of these 17 (3.57%) were removed using our derivative procedure. As we can see on Figure 4.20, the travel time seems to behave in the same way throughout the day, so there is no need to divide the data in sets according to time of day.

Therefore, we fit our candidate distributions to all the data, obtaining the statistics displayed on Table 4.7 (with $\alpha = 0.0309, \beta = 0.0209, \lambda = -0.8631, \mu = 194.0154, \delta = 15.6501$ for the GH distribution).
Figure 4.20: Travel time on arc of type 4 during the day from A to B

<table>
<thead>
<tr>
<th>Distribution</th>
<th>AD</th>
<th>KS</th>
<th>L1</th>
<th>L2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>0.12246257</td>
<td>0.05197979</td>
<td>0.01632743</td>
<td>0.00004044510</td>
</tr>
<tr>
<td>Lognormal</td>
<td>0.09140453</td>
<td>0.04183772</td>
<td>0.01177670</td>
<td>0.00002359537</td>
</tr>
<tr>
<td>GH</td>
<td>0.30643179</td>
<td>0.12860229</td>
<td>0.05842737</td>
<td>0.00051317352</td>
</tr>
<tr>
<td>Gamma</td>
<td>0.10005230</td>
<td>0.04562740</td>
<td>0.01272138</td>
<td>0.00002706202</td>
</tr>
<tr>
<td>Weibull</td>
<td>0.26792949</td>
<td>0.06849101</td>
<td>0.03437732</td>
<td>0.0015209699</td>
</tr>
</tbody>
</table>

Table 4.7: Fitting Statistics for road type 4 in direction 1

Once again, the lognormal distribution provides the better fit, with the GH distribution performing worse even than the Gaussian distribution. This comes natural if we look at the histogram on Figure 4.21. The data in this case do not present such a skewed or tilted bell towards any of the sides, which is why we don’t get such a drastic difference in the statistics of the normal, gamma and lognormal statistics as we did with previous links.

Additionally, the histogram does not present a clear bell-shaped concentration of points, so even if the lognormal is the better fit, there is still some doubt if it is a good fit. However, the c.d.f.
plot on Figure 4.22 shows how all the distributions, except for the Generalized Hyperbolic, are very close to each other. As we mentioned before, it is plausible to get a better bell-shaped form on the histogram if we use a different width for the intervals.

Figure 4.21: Density fit on arc of type 4 for direction 1

We now turn to the analysis of the travel time for the opposite direction, where 15 out of 397 points (3.78%) are removed. From Figure 4.23, we observe that the highest travel times all belong to the interval between 6 a.m. and 9 a.m., i.e., the morning rush hour. Fitting only one distribution to all the data would then imply that the heavy tail would also appear for other periods of the day which don’t have this high travel time behavior. For this reason, we proceed to divide the data into two groups: the first one contains all points belonging to the rush hour period while the second one contains all other points.

Figures 4.24 and 4.25 display the time dependence on the correspondent time intervals for the two groups. We notice how on both plots there doesn’t appear to be any time interval where the travel times are significantly higher or lower, except for the early morning or late night (but they have a very small amount of data).

Therefore, we analyze these two groups by first obtaining the descriptive statistics shown on
Figure 4.22: c.d.f. fit on arc of type 4 for direction 1

Figure 4.23: Travel time on arc of type 4 during the day from B to A
Figure 4.24: Travel time on arc of type 4 for time of day 2 from B to A

Figure 4.25: Travel time on arc of type 4 for all other times of day from B to A
Table 4.8. Clearly, the mean difference is large enough to reassure that is better to do the analysis for each interval. The drawback is how to account for this separation when considering the problem of convoluting the distributions on a path.

<table>
<thead>
<tr>
<th>Time of Day</th>
<th>Observations</th>
<th>Mean (secs)</th>
<th>St Dev (secs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>6 a.m. - 9 a.m.</td>
<td>135</td>
<td>209.0655</td>
<td>20.53824</td>
</tr>
<tr>
<td>All others</td>
<td>217</td>
<td>193.228</td>
<td>16.93871</td>
</tr>
</tbody>
</table>

Table 4.8: Descriptive Statistics for arc of road type 4 on different times of day

The analysis of the data not contained in the 6 a.m. - 9 a.m. interval provides results similar to those obtained for other links. Figures 4.26 and 4.27, and Table 4.9, show that the best fit is for the lognormal distribution, although as for the opposite direction, the histogram does not have a nice bell shaped form.

Figure 4.26: Density fit on arc of type 4 without rush hour period

The fit on the data for the rush period however does not seem to perform well, as can be seen on the histogram displayed on Figure 4.28. This is also a result of the small number of data that we have. Let us recall that while for the arcs on I-295 we had around 1000 data points, for this
Figure 4.27: c.d.f. fit on arc of type 4 without rush hour period

<table>
<thead>
<tr>
<th>Distribution</th>
<th>AD</th>
<th>KS</th>
<th>L1</th>
<th>L2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>0.1425356</td>
<td>0.06787527</td>
<td>0.02959297</td>
<td>0.0012564413</td>
</tr>
<tr>
<td>Lognormal</td>
<td>0.1125514</td>
<td>0.05042851</td>
<td>0.02104273</td>
<td>0.0006664123</td>
</tr>
<tr>
<td>GH</td>
<td>0.2411573</td>
<td>0.09798362</td>
<td>0.04040092</td>
<td>0.0024085030</td>
</tr>
<tr>
<td>Gamma</td>
<td>0.1207381</td>
<td>0.05618040</td>
<td>0.02375083</td>
<td>0.0008357195</td>
</tr>
<tr>
<td>Weibull</td>
<td>0.2306392</td>
<td>0.10586094</td>
<td>0.05267016</td>
<td>0.0037630795</td>
</tr>
</tbody>
</table>

Table 4.9: Fitting Statistics on arc of type 4 without rush hour period

arc, as well as for the arc on road 3, we have half as much data. This helps us understand why for I-295 we get nice bell shaped histograms while here we get funny shaped ones. The case is worse if we only produce the histogram for a certain time interval like this one, since we only have 135 data points.

Nevertheless, the c.d.f. plot on Figure 4.29 presents the fits from the cumulative perspective, and here, as we have seen many times before, the lognormal distribution fits the empirical one for most part of the range, except for values around the mean. Fit statistics are summarized on Table
Figure 4.28: Density fit on arc of type 4 during rush hour

Figure 4.29: c.d.f. fit on arc of type 4 during rush hour
If we now take all the data regardless of time of day, we get the histogram and fits presented on Figure 4.30. We can notice two peaks on the histogram, which actually correspond to our two group separation of the day. Still, this histogram presents a much better form than the previous ones, mostly because of the larger amount of data. The cumulative distribution functions and fitting statistics are presented on Figure 4.31 and Table 4.11.

Table 4.10: Fitting Statistics on arc of type 4 during rush hour

<table>
<thead>
<tr>
<th>Distribution</th>
<th>AD</th>
<th>KS</th>
<th>L1</th>
<th>L2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>0.1628613</td>
<td>0.08027388</td>
<td>0.02625867</td>
<td>0.0011773787</td>
</tr>
<tr>
<td>Lognormal</td>
<td>0.1273965</td>
<td>0.06312671</td>
<td>0.01805288</td>
<td>0.0006445469</td>
</tr>
<tr>
<td>GH</td>
<td>0.3134006</td>
<td>0.13447315</td>
<td>0.04679965</td>
<td>0.0038882783</td>
</tr>
<tr>
<td>Gamma</td>
<td>0.1399808</td>
<td>0.06924155</td>
<td>0.02041600</td>
<td>0.0008083605</td>
</tr>
<tr>
<td>Weibull</td>
<td>0.2470084</td>
<td>0.09825944</td>
<td>0.04813364</td>
<td>0.0029749849</td>
</tr>
</tbody>
</table>

Figure 4.30: Density fit on arc of type 4 for direction 2
4.2 Numerical Results for Arcs of the Same Road Type

There are 280 monuments on Copilot’s network system, and we gathered measurements for approximately 1.6 million pairs. In terms of applying the analysis presented in the previous section to every pair, we are obviously constrained.

On one hand, not every single one of these pairs has sufficient data to perform the analysis, as we mentioned on chapter 2. Therefore, we are unable to obtain distributions for the majority of the pairs. Moreover, most of the pairs with sufficient data are of road type 1 so we would have no
meaningful analysis for the other road types. If we are looking at what we are trying to achieve by finding the distributions, that is, route planning, it is really limited to have paths only in interstates. Other road types are essential to our path planning procedure, specially if we are at a local level.

On the other hand, a system that has distributions for every single m2m pair is in itself highly inefficient. Furthermore, our analysis is made using only 280 monuments from the US road network, but if we want to expand it to include more locations in the actual road network (as opposed to the monument road network), the computational burden increases exponentially. Let us recall that just for this 280 monuments, we have 1.6 million pairs, which means we would have to calculate 1.6 million distributions.

Therefore, our goal in this section is to obtain general distributions for the different road types. Of course, by doing so, we are assuming that the behavior of a certain road type is the same throughout the country. Clearly, this means that a County Highway on New Jersey has similar characteristics as one in Colorado, but this might not be the case, since factors like geography, weather, and speed limit among others, are different. Nevertheless, the analysis serves as a first approach on the characterization of these road types, and, surprisingly, we get similar results to the ones obtained for individual links. However, we must not forget that the data is biased towards the Northeastern region of the country, so it will be most representative of this area.

Two special remarks should be made about the analysis. First, the amount of data is enormous for road type 1, so performing a general analysis for this road type was not computationally viable. Even so, since our purpose now is to look at roads with insufficient data, we focus our analysis on road types 3, 4 and 6. Second, in order to compare different links of the same road type, the travel time should be normalized, since distances between monuments are obviously different. The distances are obtained using the formula described on chapter 2, which assumes that our arc is a straight line connecting the two monuments. This of course induces an error in our calculations of these normalized times (which are simply the inverse of speed). Finally, to do the analysis, we took pairs where both monuments are on the road type being analyzed. But recalling the definition of monuments, we could have that in between two monuments of road type \( x \), the path includes several arcs of road type \( y \). However there is no information in our data that can tell us the path traveled between the monuments.
4.2.1 Case 1: Road Type 3

To start, we have 133,458 observations for travel times between monuments on road type 3, of which, 2271 (1.71%) are removed using the usual procedure. We notice that this is consistent with the percentage of points removed for the individual links analyzed previously.

In this case, there is no real point in analyzing time of day dependency, since this is a factor that affects individual links and grouping them cancels it. If some arcs have high travel times in the morning rush hour and others on the afternoon rush hour period, we will not be able to see this time of day effect when looking at them together.

We rather look at the histogram of the data and try to fit some of the distributions to it. The results are presented on Figure 4.32. In this case, we do not contemplate the Generalized Hyperbolic distribution, since the amount of data demands high computational power on the optimization procedure. However, the heavy tail and sharp peak of the histogram suggest it may be worthwhile to test it in the future.

![Figure 4.32: Density fit for roads of type 3](image)

On the histogram, we can see that the tails on the distribution is heavier than those obtained for individual links. This is a result of the large amount of data now being considered, as well as
the limitations of the removal heuristic. What is worth noticing is that we still have the heavy tail, left skewed distribution that individual links had. Additionally, on trying to fit it to the theoretical distributions we have considered, the lognormal still seems to be the better candidate, as shown in Table 4.12.

<table>
<thead>
<tr>
<th>Distribution</th>
<th>AD</th>
<th>KS</th>
<th>L1</th>
<th>L2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>5.6271694</td>
<td>0.1772641</td>
<td>0.09913586</td>
<td>0.013165886</td>
</tr>
<tr>
<td>Lognormal</td>
<td>0.3176093</td>
<td>0.1142278</td>
<td>0.05956216</td>
<td>0.004832021</td>
</tr>
<tr>
<td>Gamma</td>
<td>0.5922743</td>
<td>0.1363549</td>
<td>0.07295533</td>
<td>0.007204726</td>
</tr>
<tr>
<td>Weibull</td>
<td>0.7741201</td>
<td>0.1796415</td>
<td>0.08967158</td>
<td>0.010726447</td>
</tr>
</tbody>
</table>

Table 4.12: Fitting Statistics for road type 3

Figure 4.33 shows the fit in terms of the cumulative distribution function. As expected, the distributions that better approximate the empirical c.d.f. are the lognormal and gamma distributions. However, it is not as sound as for the individual links. Clearly, this is a result of the larger amount of data in the tails and further research should be focused on trying to fit the GH distributions, as mentioned previously.

Finally, we can compare the results obtained here to the ones obtained for the single link analyzed on Section 4.1.2. For this link, we obtained a mean speed of 46.61374 mph for direction 1 and 50.11419 mph for direction 2, while for all links we get 47.11571 mph. Although this global mean is not exactly the same as those of the individual links, it is still a good approximation in case there is no actual data for the individual link. Unfortunately, we can not say the same thing for the standard deviation; it is 9.225901 and 10.23584 for directions 1 and 2 of the individual link, while it has a value of 14.53153 for the general case. This indicates that although the means of the speeds (and travel times) are very close, the distributions are not, and we will be giving more variability to the travel time of an individual link than it really has if we were to apply to it the distribution obtained for all links of the same road type.
4.2.2 Case 2: Road Type 4

The procedure and results obtained for arcs of road type 4 are very similar to those of road type 3. 87,254 measurements were obtained, of which 2555 (2.92%) were removed. We get the density and cumulative distribution fits displayed in Figures 4.34 and 4.35 and the statistics of Table 4.13. We notice that the distribution of travel times is less peaked than that for road type 3, hence the better fit and statistics for the lognormal and gamma distributions. However, it also presents a heavy tail behavior that these distributions do not grasp.

<table>
<thead>
<tr>
<th>Distribution</th>
<th>AD</th>
<th>KS</th>
<th>L1</th>
<th>L2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>0.6275728</td>
<td>0.13602267</td>
<td>0.07286768</td>
<td>0.007188157</td>
</tr>
<tr>
<td>Lognormal</td>
<td>0.1821176</td>
<td>0.06365670</td>
<td>0.03448341</td>
<td>0.001552227</td>
</tr>
<tr>
<td>Gamma</td>
<td>0.2337587</td>
<td>0.08787588</td>
<td>0.04718219</td>
<td>0.002981016</td>
</tr>
<tr>
<td>Weibull</td>
<td>0.3513240</td>
<td>0.12129106</td>
<td>0.06254498</td>
<td>0.005292584</td>
</tr>
</tbody>
</table>

Table 4.13: Fitting Statistics for road type 4

For this road type, we get a mean of 39.1959 mph and standard deviation of 13.73897 mph for
Figure 4.34: Density fit for roads of type 4

Figure 4.35: CDF fit fit for roads of type 4
the speed, while for the individual link analyzed in section 4.1.3 we get means of 52.90556 mph and 54.54079 mph for directions 1 and 2, and standard deviations of 4.837281 mph and 4.631835 mph. Clearly, the general properties of the road type are not representative of the specific links. This is no surprise, since we are talking of paths on local roads. If for interstates or divided roads it is more likely that the path between two monuments is of the same road type, this is not the case for the arterial roads. Therefore, the paths between monuments of this type might comprise segment of other road types, and therefore it is more difficult to get to general conclusions. In this case, it would be more pertinent to analyze travel times at a more detailed level on the roads, and not at the monument level.

4.2.3 Case 3: Road Type 6

We do not have individual links on road types 6 that have sufficient data for an analysis to be performed on them. In this case then, we have to rely on the results that we get for all the m2m pairs of this type grouped together. Of course, this has the same complications and inaccuracies that the previous analysis for the other road types had. Additionally, the amount of data gathered is rather small (3543 observations) as compared to these other road types, a natural result of the nature of monuments.

Nevertheless, the results obtained are similar to all our previous analysis, with density and cumulative distribution fits on Figures 4.36 and 4.37. We get a mean speed of 31.59544 mph and standard deviation of 12.7467 mph. We can notice that although there are clear differences between the mean speeds among road types, standard deviations are similar. Finally, the fitting statistics are presented on Table 4.14, where again, the lognormal distribution presents the lowest statistics.

<table>
<thead>
<tr>
<th>Distribution</th>
<th>AD</th>
<th>KS</th>
<th>L1</th>
<th>L2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>2.3308121</td>
<td>0.17006226</td>
<td>0.09395807</td>
<td>0.011751308</td>
</tr>
<tr>
<td>Lognormal</td>
<td>0.2363851</td>
<td>0.08619486</td>
<td>0.04501607</td>
<td>0.002722616</td>
</tr>
<tr>
<td>Gamma</td>
<td>0.3734158</td>
<td>0.11722447</td>
<td>0.06170871</td>
<td>0.005083171</td>
</tr>
<tr>
<td>Weibull</td>
<td>0.3924406</td>
<td>0.14909186</td>
<td>0.07315037</td>
<td>0.007146493</td>
</tr>
</tbody>
</table>

Table 4.14: Fitting Statistics for road type 6
Figure 4.36: Density fit for roads of type 6

Figure 4.37: c.d.f. fit fit for roads of type 6
Chapter 5

5 Random Costs for Paths

In the last chapter, we fitted different theoretical distributions to the travel times on the US road network for links defined by two monuments and an arc connecting them. We found that the best distribution to model these times, after removing several high values, is the lognormal distribution.

In this chapter, we turn our attention on establishing a relationship between the travel times on the individual links and the travel time on a path that comprises them. To do so, we will examine a path of 8 links on I-295, which has enough data in order to obtain a meaningful analysis. On the first part, we will explore the issue of independence between the links of the path by using the actual data. Then we will fit theoretical distributions for the path and compare the results to the fits on the links. We will also test the reciprocal gamma distribution for the path, as the individual links are lognormally distributed.

5.1 The Issue of Independence

Since the objective of analyzing the travel time distribution on a path is to obtain this distribution by means of the individual distributions of the links comprising it, it is necessary to begin by exploring the relationship between the distributions of these links. It has already been established that the lognormal distribution does the better job at approximating these distributions, but given this, the question on how to “add” or convolute these distributions still remains. This indeed is an open question on the field of statistics, since closed formula have only been derived to convolute some distributions, like an exponential and a normal distribution, two Poisson distributions, two
normal distributions with mean zero, among others (see []).

On the other hand, if we have variables $X$, $Y$ and $U$, where $U = X + Y$, and we know that the distributions of $X$ and $Y$ are lognormal, there is no closed formula that gives us the distribution of $U$, which is the convolution of $X$ and $Y$. Closed formula exist for normal random variables when they are independent, but clearly here two things are different. First, what is considered normal is not the travel time but its logarithm and if $x + y = z$, which is clearly the case of travel times, $\log x + \log y \neq \log z$. Second, it is not known if the travel times are independent or not.

On this second issue, it is rather obvious that the distributions will not be independent. If a driver is a fast driver, his travel time will be low on all segments of the path, while if the driver is a slow driver, the travel time will be high. Therefore, for each driver the travel time on one link will be undeniably correlated to the travel time on the next link, unless there is some exogenous factor that makes him change his travel behavior. For example, if there is a link where all drivers are forced to slow down because of congestions or weather conditions, then the travel time on the link is going to be very similar for both the slow and fast drivers. In this case, the correlation between the travel time on link and the previous link will be lower.

A question also arises when the links examined are not consecutive. Taking an initial link $l_1$, its travel time is correlated to the travel time in link $l_2$ of the path, as explained before. But what is the correlation between $l_1$ and $l_{100}$, this latter being the hundredth link on the path? Intuition would say that as the driver is moving through the path, the correlation decreases. If the driver is 100 links away from the link he started on, it is not expected that his travel time on this link is correlated to the travel time on the initial link.

However, with a closer look, we find that this last statement is not necessarily true, but the correlation might actually depend on the path. A fast driver will be a fast driver along any path no matter where he is, unless the same exogenous conditions mentioned before do not let him. This is more valid if the path is along an interstate, since congestion is less probable on this type of road, making travel times measured more dependent on driver behavior than on actual travel conditions.

Our data only allows us to examine this latter case for reasonably long paths. The longest one we obtain is comprised of eight links along Interstate 295 and it is displayed on Figure 5.1. This interstate surrounds the Trenton area in New Jersey and only 386 drivers used it, which is
significantly lower than the number of observations we obtain for the first link analyzed on chapter 4 (1118), which is in fact link 2 on the path.

The analysis on independence is done by first setting link 1 as the initial link and obtaining the correlation of travel times between this link and all the others. This is displayed on Figure 5.2. On the plot we see that all the correlations are positive, which agrees with our hypothesis of driver behavior. And we also notice that this correlation is high, above 0.5 for all the links except for 4 and 8. This last result is puzzling, since it doesn’t seem reasonable that the correlation is low between links 1 and 4, but high between links 1 and 5. However, on the I-295 path, link 4 is the closest one to the Trenton urban area. Our guess is that in this link, speed is drastically reduced for all vehicles, probably due to congestion, and therefore a low correlation appears. After traversing this link, fast drivers speed up again and slow drivers remain at low speeds. Link 8 is even more puzzling, since it is not close to an urban area like link 4 is.

Nevertheless, the crux of the matter is that positive correlation exists, and therefore independence between the time distributions on the links can’t be assumed. This leads us to consider reciprocal gamma distributions.

Figure 5.1: Path on I-295
Figure 5.2: Correlation between links for a path on I-295

5.2 Distribution Fittings for a Path

On obtaining a distribution for the path, we follow the same procedure as for regular links. First, we plot the standard deviation from the media as we remove points with high values, obtain the plot displayed on Figure 5.3. Our heuristic removes 10 (2.59%) points of the original 386, which in the plot agrees with the point at which most of the standard deviation has decreased.

Next, we plot on Figure 5.4 the travel time versus time of day, to see if we can identify any rush hour period. Although there are a couple of high values for late evening, the lowest values also belong to the same period and so it can not be said that dividing the data into two or more time of day intervals is helpful. This agrees with the result of the individual analysis of link 2 presented earlier. Additionally, the amount of data is too small to get meaningful results for each interval, as we saw for the link of road type 4 on chapter 4.

Let us recall too that our purpose is too obtain the best fit of a theoretical distribution on the travel times for the fit. With this perspective, and from the results obtained previously for links, we skip using the Weibull distribution since it has been proven that lognormal and gamma distributions
Figure 5.3: Removal of high values for a path on I-295

Figure 5.4: Travel time during the day for a path on I-295
are better candidates. In addition, the lognormal did the best job on fitting the distributions of the individual links (we only presented results for link 2), and therefore, we also want to test reciprocal gamma distributions, since they approximate the sum of dependent lognormal distributions. These results are presented on Figures 5.5 and 5.6, and Table 5.1.

![Figure 5.5: Travel time density for a path on I-295](image)

<table>
<thead>
<tr>
<th>Distribution</th>
<th>AD</th>
<th>KS</th>
<th>L1</th>
<th>L2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>0.2621094</td>
<td>0.05059800</td>
<td>0.01950737</td>
<td>0.0005874036</td>
</tr>
<tr>
<td>Lognormal</td>
<td>0.1955283</td>
<td>0.04068330</td>
<td>0.01506347</td>
<td>0.0003559525</td>
</tr>
<tr>
<td>GH</td>
<td>0.6392344</td>
<td>0.11240763</td>
<td>0.04878410</td>
<td>0.0034277715</td>
</tr>
<tr>
<td>Gamma</td>
<td>0.2181898</td>
<td>0.04396567</td>
<td>0.01633686</td>
<td>0.0004202561</td>
</tr>
<tr>
<td>Reciprocal Gamma</td>
<td>0.1774471</td>
<td>0.03736379</td>
<td>0.01372900</td>
<td>0.0002957556</td>
</tr>
</tbody>
</table>

Table 5.1: Fitting Statistics for a path on I-295

We obtain the best fit from the reciprocal gamma distribution as expected. The statistics are slightly lower than those of the lognormal distribution, which indicates a better fit both at the
body and at the tail of the distribution. As with the cases for the individual links, the Generalized Hyperbolic does not present satisfying results.

So if based on this evidence it can be said that the distribution of the travel times for a path is reciprocal gamma, how can $\alpha$ and $\beta$ be obtained? At this point of the research, it is difficult to say since the lognormal distribution for each link has different parameters. Table 5.2 summarizes them for the different links considering all points and not just the ones from drivers who traverse the whole path.

If we take only the 376 drivers that traverse the whole path, we get the means and standard deviations presented on Table 5.3. Comparing both tables we see that the means of the travel times are only a few tenths of a second different between the reduced and the whole samples, which is not the case for the standard deviations. Since the lognormal distributions for the individual links are defined by the mean and the standard deviation of the logarithm of the time, the distribution for the path would be determined by the parameters of the second table. In practice however, we would like to use as much data as possible and not just simply the sample from drivers who traverse the whole path. This issue would then need further exploration, since the reciprocal gamma distribution we
Table 5.2: Descriptive Statistics for links on a path on I-295 for all drivers

<table>
<thead>
<tr>
<th>Link</th>
<th>Observations</th>
<th>Mean Time</th>
<th>St Dev</th>
<th>Mean log(Time)</th>
<th>St Dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>996</td>
<td>94.10921</td>
<td>9.740906</td>
<td>4.54008</td>
<td>0.09017714</td>
</tr>
<tr>
<td>2</td>
<td>1107</td>
<td>73.48673</td>
<td>5.472986</td>
<td>4.294429</td>
<td>0.07259892</td>
</tr>
<tr>
<td>3</td>
<td>1099</td>
<td>118.2433</td>
<td>9.585741</td>
<td>4.769753</td>
<td>0.07585458</td>
</tr>
<tr>
<td>4</td>
<td>990</td>
<td>102.9832</td>
<td>8.555382</td>
<td>4.631293</td>
<td>0.08002494</td>
</tr>
<tr>
<td>5</td>
<td>735</td>
<td>97.8279</td>
<td>6.288397</td>
<td>4.581192</td>
<td>0.06324897</td>
</tr>
<tr>
<td>6</td>
<td>753</td>
<td>116.9538</td>
<td>8.135697</td>
<td>4.759515</td>
<td>0.06639429</td>
</tr>
<tr>
<td>7</td>
<td>719</td>
<td>91.23436</td>
<td>6.013444</td>
<td>4.511348</td>
<td>0.06402061</td>
</tr>
<tr>
<td>8</td>
<td>636</td>
<td>218.6982</td>
<td>14.3407</td>
<td>5.385685</td>
<td>0.06250484</td>
</tr>
</tbody>
</table>

Table 5.3: Descriptive Statistics for links on a path on I-295 for drivers that traverse the whole path

<table>
<thead>
<tr>
<th>Link</th>
<th>Mean Time</th>
<th>St Dev</th>
<th>Mean log(Time)</th>
<th>St Dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>93.92853</td>
<td>7.418414</td>
<td>4.539561</td>
<td>0.07641678</td>
</tr>
<tr>
<td>2</td>
<td>73.74722</td>
<td>6.016167</td>
<td>4.297685</td>
<td>0.07513368</td>
</tr>
<tr>
<td>3</td>
<td>118.50459</td>
<td>8.102919</td>
<td>4.772738</td>
<td>0.06582741</td>
</tr>
<tr>
<td>4</td>
<td>102.24822</td>
<td>17.484168</td>
<td>4.620991</td>
<td>0.09568686</td>
</tr>
<tr>
<td>5</td>
<td>98.47402</td>
<td>7.332563</td>
<td>4.587355</td>
<td>0.06804023</td>
</tr>
<tr>
<td>6</td>
<td>117.36300</td>
<td>7.732509</td>
<td>4.763203</td>
<td>0.06374867</td>
</tr>
<tr>
<td>7</td>
<td>91.59588</td>
<td>6.018731</td>
<td>4.515346</td>
<td>0.06317888</td>
</tr>
<tr>
<td>8</td>
<td>218.42972</td>
<td>13.263637</td>
<td>5.384704</td>
<td>0.05883480</td>
</tr>
</tbody>
</table>

At this point, we can say that using the total sample of drivers to calculate the mean of the path is not far fetched. Table 5.4 displays the means obtained for the travel time on the subpaths of path 8 as obtained from the data as well as if we add the mean times of the links comprising them. Although as we increase the number of links this difference also increases, the relative error is stable (with a maximum of 8 seconds for a travel time of 8 minutes between links 1 and 5).
<table>
<thead>
<tr>
<th>Subpath (Links)</th>
<th>Obs.</th>
<th>Mean (from data)</th>
<th>Mean (from sum of links)</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-2</td>
<td>957</td>
<td>167.5744</td>
<td>167.59594</td>
<td>0.01%</td>
</tr>
<tr>
<td>1-2-3</td>
<td>927</td>
<td>284.6296</td>
<td>285.83924</td>
<td>0.42%</td>
</tr>
<tr>
<td>1-2-3-4</td>
<td>800</td>
<td>387.3487</td>
<td>388.82244</td>
<td>0.38%</td>
</tr>
<tr>
<td>1-2-3-4-5</td>
<td>567</td>
<td>479.8318</td>
<td>486.65034</td>
<td>1.42%</td>
</tr>
<tr>
<td>1-2-3-4-5-6</td>
<td>482</td>
<td>598.8283</td>
<td>603.60414</td>
<td>0.80%</td>
</tr>
<tr>
<td>1-2-3-4-5-6-7</td>
<td>447</td>
<td>691.0926</td>
<td>694.8385</td>
<td>0.54%</td>
</tr>
<tr>
<td>1-2-3-4-5-6-7-8</td>
<td>376</td>
<td>908.8343</td>
<td>913.5367</td>
<td>0.52%</td>
</tr>
</tbody>
</table>

Table 5.4: Mean Travel Times for Subpaths on I-295

The question on how to add these lognormal travel time distributions to obtain a gamma reciprocal distribution still remains, specially on the calculation of the parameters $\alpha$ and $\beta$ of the latter. However, the mean analysis hints that a relationship between the parameters should exist and can be a topic of further study. Moreover, if we select links that have equal travel time distributions on the path (say links of equal distance), the analysis can be simplified, in terms of the equal length intervals of the Asian options. However, deep analysis should also be made in this case, since the lognormal distribution of these options is determined by Brownian motion.
Chapter 6

6 Conclusions and Further Research

In this chapter we summarize the results obtained for the distribution fittings that we have previously discussed. The arcs, roads and paths chosen and presented in this thesis are the most representative of a pool of monument to monument pairs we were able to analyze, so these results are not exclusive to just the cases presented here. Although the large amount of data gathered allowed us to perform the analysis on a substantial amount of links, most of them belonged to interstates and presented similar behaviors. The analysis can certainly be extended to other road types and links not covered, but to do this, the first step is to implement a system that can gather a sufficient amount of data for them.

First we will analyze the results for the individual arcs, then for the groups of arcs of the same road type, and finally, we will talk about the path case, which is the biggest milestone that should be surpassed in order to make the travel time analysis applicable.

6.1 Results for Individual Arcs

Arcs for road types 1, 3 and 4 were presented since only these road types had sufficient data and each of them presents different traffic characteristics. On the first hand, we showed that several distributions display better fitting statistics than the Gaussian distribution, which is the most commonly used distribution when talking about the stochastic shortest problem.

In particular, the lognormal distribution presented the best results for all road types, except for some links on interstates where the Generalized Hyperbolic distribution had better fits on the tails.
This is a very important issue, since high travel times cannot be arbitrarily discarded but they might actually represent effective travel times which are meaningful for the travel time estimation. However, for the cases where the GH distribution does the better fit on the tail, it does not perform so well on the body, due to its high leptokurtic nature.

Related to the tail fit, the heuristic devised to remove high values before performing the fit presented similar results for each road type, that is, the percentage of values removed was similar for each. However, its value was higher as the road types became more local, and it is difficult to say if this is a result of an increasing volatility of the travel times as more interruptions (traffic lights, intersections) are added to the traffic flow, or simply because as roads become more local, drivers stopping possibilities that do not affect traffic flow (gas, restaurants, shops) increase. We do as much as we can with the data that we have, and nothing in it can tell us in which of these two cases each high value is.

On comparing the fits for the road types, as the disrupting conditions increase, the fit seems to behave more poorly. Although for all the road types the lognormal distribution does the better job, a chi square test would tell us that we cannot assure that the distributions are lognormal (p-values of zero are obtained). Nevertheless, of all the family of distributions that we possess it is the one that most resembles it, as it was seen with cumulative distribution plots.

Time of day was another issue that became more relevant as the road types became more local. As a result of more likely congestion on arterial roads than on interstates, travel time behavior varied during the different time of day intervals, and we saw that it was more convenient to divide the data into two groups according to this. This implies that on a local level it is more difficult to obtain travel time distributions for all measurements but that rather it is better to obtain different distributions for the different time of day intervals. This would complicate further our goal of using these travel time distributions on the stochastic shortest path problem, since time of day also becomes a factor and the two perspectives described on chapter 1 should be considered.
6.2 Results for Groups of Arcs

The analysis of all the arcs on the US road network for different road types was made to assess the problem that most road types, except for interstates, do not have sufficient data for their individual links. Therefore, we tried to derive speed properties for them that can be applied to the individual links. As with the individual arcs, the distribution that represented more accurately the data was the lognormal. However, the empirical distributions were more peaked on the left hand side, that is, there were a large number of high values that weren’t removed. Additionally, performing the analysis for general road types has two major drawbacks.

First, we can not incorporate time of day effects since a rush hour period on one link is canceled out but different rush hour periods on other links. This represents a strong disadvantage, since for the road types that we do not have enough data for individual links (3, 4, 6, 8), traffic congestion is more likely than for interstates. Second, between two monuments of the same road type, there could possibly be actual road links that belong to other road types. Therefore, how correct it is to generalize this distributions is an issue that should be addressed.

Finally, for the road type 3 case that was tested, we found that the mean speed approximation obtained from analyzing all road types is a good approximation for the mean speed of the individual link. This was not the case for road type 4, and this is related to the road links that might appear in between the two monuments as we mentioned previously. Drivers that travel through roads of type 3 tend to stay on these roads (US 1 for example) while in between arterial road it is more likely that there are roads of different type (secondary and local roads).

6.3 Results for a Path

Our final analysis focused on how to obtain the travel time distributions for a path if we know the distributions on the individual links. For this purpose, it was appropriate that the travel times on the links were more closely modeled by lognormal distributions, since the sum can be approximated by a reciprocal gamma distribution. The path chosen agreed with this notion, although the relationship between the parameters of the links and the paths still remain a question to be extended. We also showed that in general the sum of the mean of travel times on the links can approximate (with an
error below 1%) the mean time on the path.

At this point a nice topic arises. We have seen that indeed the reciprocal gamma distribution is the best fit for the travel time of the path, and we define the path as a succession of links. However, we can define links anyway we want. In fact, the links described on this thesis are not the classical links on a road network but actually segments between midpoints of links. So, in that sense, we can also say that one of these monument links is a path. Now, we can also divide this links into several sublinks. There is no reason that these sublinks should not have lognormal distributions, since the traffic flow on them behaves the same way as in the link. Therefore, we can continuously divide each link into infinitesimally sublinks that have lognormal travel time distributions and the problem of finding the path travel time distribution becomes continuous rather than discrete we would then think that a reciprocal gamma distribution is in fact a better representation of the travel time on the link than the lognormal one. In fact, this happens for the first arc analyzed on I-295, as shown on Table 6.1.

<table>
<thead>
<tr>
<th>Distribution</th>
<th>AD</th>
<th>KS</th>
<th>L1</th>
<th>L2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lognormal</td>
<td>0.1926083</td>
<td>0.06240267</td>
<td>0.02630589</td>
<td>0.0010215136</td>
</tr>
<tr>
<td>Reciprocal Gamma</td>
<td>0.1752167</td>
<td>0.05857176</td>
<td>0.02443341</td>
<td>0.0008821929</td>
</tr>
</tbody>
</table>

Table 6.1: Fitting Statistics for Lognormal and Reciprocal Gamma Distributions on I-295

### 6.4 Further Research

Different topics were covered on this thesis where each by itself requires deeper and extended research. The analysis of travel time distributions for individual links still requires to try several other distributions that might seem good candidates to approximate them. An incredible amount of theoretical distributions exist, and of these, it is difficult to try them all in just one thesis. However, it is important to keep in mind that the objective is to convolute these distributions so care should be taken in making them as simple as possible or choosing distributions that can generate closed form convolutions. Still, if more suitable distributions are found, interesting research can be made in how to produce these convolutions.
On the other hand, a lot of work is also needed on obtaining a procedure to gather and manipulate data for travel time between nodes on local roads, and that this data is sufficient enough to perform analysis like the one presented in this work. An approximation was made in this thesis knowing that the data for these roads was enough to aggregate all the links and analyze the travel time as a whole, but clearly it is more precise doing the analysis for each individual link. However attention should be paid on how efficient this would be, given the huge amount of links on the US road system.

Finally, the most interesting theoretical research that the results suggest is the one regarding the convolution of lognormal distributions and the continuity that was implied in the previous section. This is a major topic on stochastic calculus, and research on finance and option pricing can lead the way to obtain some nice properties on the stochastic nature of travel times between nodes.