Modelling and inference for the travel times in vehicle routing problems

Christina Wright, MMath(Hons.), M.Res

Lancaster University

Submitted for the degree of Doctor of Philosophy at Lancaster University.

11 November 2019

STOR-i
excellence with impact

Abstract

Every day delivery companies need to select routes to deliver goods to their customers. A common method for the formulation and for finding the best route is the *vehicle routing problem* (VRP). One of the key assumptions when solving a VRP is that the input values are correct. In the case of travel time along a section of road, these values must be predicted in advance. Hence selecting the optimal solution requires accurate predictions. This thesis focuses upon the prediction of travel time along links, such that the predictions will be used in the defined VRP.

The road network is split into links, which are connected together to form routes in the VRP. Travel time predictions are generated for each link. We predict the general behaviour of the travel times for each link, using time series forecasting models. These are tested both empirically, against the observed travel time, and theoretically, against the ideal characteristics of a VRP travel time input, including the resulting prediction uncertainty in the VRP. Small input variations are likely to have little impact upon the optimal solution. In contrast, infrequent and unpredicted large delays, e.g., from accidents, which occur outside the general travel time behaviour can change optimal routes. We study the delay behaviour and suggest a novel model consisting of three parts: the delay occurrence rate, length and size. We then suggest ways to input both the delay and the general travel time models to the VRP, which results in an optimal solution that is more robust to delays.

Traffic moves from one link into the network, so if one link is busier then the same traffic will flow to the connecting links. We extend the single link model to incorporate information from the surrounding links using a network model. This produces better predictions than the single link models and hence better inputs for the VRP.

Acknowledgements

Firstly I'd like to thank all my supervisors, Matt Nunes, Konstantinos Zografos, Nikos Kourentzes and Jon Tawn for all their help and support, and for believing that I could finish this when I wasn't sure. Especially Matt for ensuring that I made progress each week, even if it wasn't very much and Jon for helping me work out how to write. Without the department and structure of STOR-i I'm sure I'd have stopped early on. I'd also like to thank EPSRC for funding my MRes and PhD.

I would like to thank my parents for all their help and support, even when I wanted home to be separate from my PhD. The chocolate after Christmas and Easter was especially appreciated, as were the holidays. I'd also like to thank my brother Jamie for being himself and also liking cycling. It's much more fun to watch cycling with your brother.

Finally I would like to thank my friends for being there over the past four years. Especially Stephen who has somehow managed to live in the same house as me for the past four years, Katherine for lots of meals and cake, Emma for talks and crafting and Lucy for convincing me to cycle socially. And also Emma who reminded me that if everything is going wrong there's always sewing, and if that's also going wrong then it'll probably be better the next day.

Declaration

I declare that the work in this thesis has been done by myself and has not been submitted elsewhere for the award of any other degree.

Christina Wright

The word count is approximately 67,000 words.

Included in the print thesis is a map that is under copyright from and is published under an exception permits acknowledged fair dealing for the purpose of illustration for instruction for a non-commercial purpose. Readers of the electronic version can find the map on page 6 of the report online at https://www.gov.uk/government/publications/highways-englands-strategic-road-network-initial-report

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Outline of problem

1.1.1 What is a vehicle routing problem?

Suppose that a delivery company has a set of customers, who are at different locations, that it needs to deliver items to from a depot. The company wishes to find the best route between them. This problem is called the *vehicle routing problem* or VRP.

Figure 1.1.1 shows the visual representation of a delivery problem for an ice cream company on a map that we will use as an example to illustrate the various aspects of a VRP. There are three customers, represented by different coloured houses, with the ice cream being delivered from the brown factory (depot). The customer requirements are represented by the number of ice cream cones. There are many different routes within the road network (black and blue lines), such that the customers receive their ice creams.

At each junction a vehicle has a choice as to which section of road to take. A link is a section of road between two junctions, where a junction is two or more roads meeting or a customer by the road. By connecting together links a route for the vehicle is created.

One possible route to deliver to the three customers is shown in blue. The vehicle

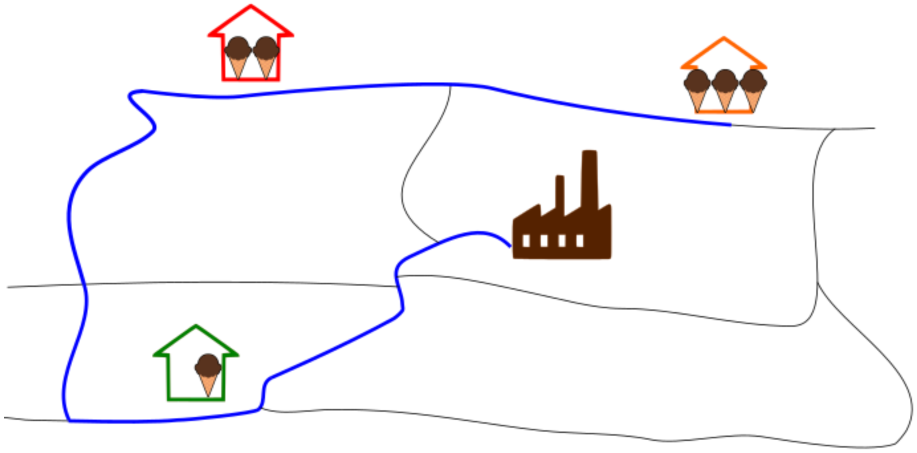


Figure 1.1.1: Ice cream delivery problem.

returns to the depot along the same path. There are many different factors to consider when selecting the best route between the depot and the three customers. The single most important factor is called the *objective* of the problem. To complete the deliveries in the quickest time could be the most important factor. However an alternative route may cost less and save the company money. The best, or optimal, route will be whichever route is best for the objective chosen.

There are additional considerations that limit which routes can be taken. These are called *constraints*. Example constraints include: the red customer may need to have their ice creams within a time slot or there is a maximum acceptable fuel cost.

The VRP consists of the objective with a set of constraints. By defining a set of constraints and the objective with relevant variables the basic vehicle routing problem is transformed to be as close to the problem of the delivery company as possible. Possible components of the problem include using multiple delivery vehicles, considering the capacity of delivery trucks, and selecting a routing based upon two or more competing objectives.

Including some components is simple, while adding others can fundamentally change the problem, requiring different solution techniques. Hence there are many sub-groups of vehicle routing problems that have the same key features, and within each subgroup the problems can be solved in a similar way.

Long term planning decisions, such as the number of vehicles in the fleet, require a VRP but are only run infrequently. Short term planning is much more common. Every day vehicles are sent out to deliver to customers. A plan for these deliveries must be generated in advance. The problem may need to updated throughout the day due to traffic conditions. These adjustments are a separate problem called real time routing. We will focus on the creation of the overall advance plan.

One subgroup is the vehicle routing and scheduling problem. This solves a VRP that is time dependent, such that the time that the vehicle sets off alters the values of the objective and constraints along the route. The vehicle routing and scheduling problem generally returns a schedule of when the delivery driver(s) need to set off from the depot, and the expected arrival and departure times from all of their customers. Any real world delivery problem is time dependent so needs to be formulated as a vehicle routing and scheduling problem. Hence in the remainder of this thesis, we assume that a vehicle routing problem or VRP includes scheduling.

1.1.2 Why the petrol routing problem?

The list of hazardous materials consists of any materials which can cause damage to either humans or the environment. Regulations therefore govern their use, transportation and storage. Hazardous materials frequently require transportation between sites. Department for Transport (2018) estimated that 59 million tonnes of hazardous materials were transported in the UK in 2017. 32 million tonnes of these were flammable liquids and were moved 3,213 million tonne kilometres, with the average journey for coke and refined petroleum products being 97.3 kilometres (Department for Transport, 2019).

Hazardous waste transportation is typically carried out by specialist delivery companies due to the regulations and the high cost of equipment. As the company is in charge of the transportation, the problem will be considered from their point of view rather than regulators or the general public.

The United Kingdom doesn't record the type of flammable liquid. However the equivalent survey in the United States for 2012 found that approximately 50% of hazardous materials transported by truck are petrol, which corresponds to about 25% of the ton-miles by truck (US Department of Transportation, 2012). Due to the frequency with which petrol is transported we will focus upon its transportation.

1.1.3 What is the petrol routing problem?

Petrol is sold to drivers through petrol stations. Alternatively companies with large fleets and specialist storage facilities may choose to source their own petrol. Most petrol stations and storage facilities are only accessible by road and as 58.7% of flammable liquids are transported by road we choose to study transportation by road (US Department of Transportation, 2012). Petrol is normally transported from an oil refinery or oil tanker ship to customers. Hence we have the depot and customers of a VRP.

Various sizes and designs of tanker lorries exist. In the UK tankers must pass certain safety tests and the risk assessment required for the unloading of petrol to a customer is specific to the design of the tanker (Health and Safety Executive, 2014). Thus customers prefer to have only one type of tanker which also makes maintenance and driver training easier. To simplify the problem we will assume that only one type of tanker is used, with a fixed maximum capacity. This tanker is likely to be limited to 56mph by law.

The risk of a serious accident increases when vehicles have higher fuel loads. As transporting excess fuel is expensive, tankers wish to set off with the exact amount of fuel required for their deliveries. Filling up the tankers requires time and must be completed before they can set off. Thus schedules for each vehicle must be known in advance of departure from the depot. This procedure complies with the accepted code of practice of letting the customers know the expected delivery time beforehand. Customers can also specify a time window within which they can accept delivery.

Deliveries of petrol to customers are made every day. However the customers that require petrol and how much they need will be different. The VRP therefore varies slightly every day, resulting in a new schedule for the fleet for each day. This schedule must take into account legal restrictions on drivers working times.

A schedule for the fleet is required every day, but there are often short term changes to the customers requirements. Solving the problem once the day before keeps these changes to a minimum. This assumes that what happens in one day doesn't affect the next day - so drivers are required to complete their journey in the 24 hour period.

Customers keep their petrol in storage tanks which cannot be overfilled by law in the United Kingdom (Health and Safety Executive, 2014). Any petrol that cannot fit in a tank or is refused must remain in the vehicle and as a consequence there is an increased risk for the remainder of the journey. To simplify the problem throughout we will assume that the pre-stated customer demands are accurate and are delivered in full.

Our problem context is therefore: a company with a fleet of tankers of the same type wish to transport petrol from an oil terminal to their customers. They wish to create a schedule that can be given each day to drivers, consisting of: the routes they need to take, when to set off, when they should arrive at each customer and how much to deliver to each customer. This can also be used to fill the tankers to the correct volume before departing. There will also be other requirements that are specific to the delivery company.

Our problem context of petrol delivery will be modelled as a hazardous material vehicle routing problem. This problem has many elements to it and this thesis will focus upon the prediction of travel time as a coefficient in the petrol routing problem. As the VRP assumes that travel times are correct, if they are incorrect then the optimal solution may be wrong.

1.1.4 Networks in VRP

The petrol tankers travel upon roads between the customers. The format of the physical road layout is incompatible with a VRP. Instead we convert it to a network which consists of nodes that are connected together by links. The VRP can then select some of these links to create routes. The links represent roads and the nodes points of interest. Usually junctions and customers are represented by nodes and a link exists between two nodes if a road goes between those two points. The optimisation model requires that customers be nodes so that vehicles can deliver to them, but having a lot of junctions greatly increases the computational time.

The full road network contains many small roads and junctions that a petrol tanker cannot take. Hence we can remove these from the network leading to fewer junctions and a simplified network. All links are directional as travel time depends upon in which direction the vehicle is travelling.

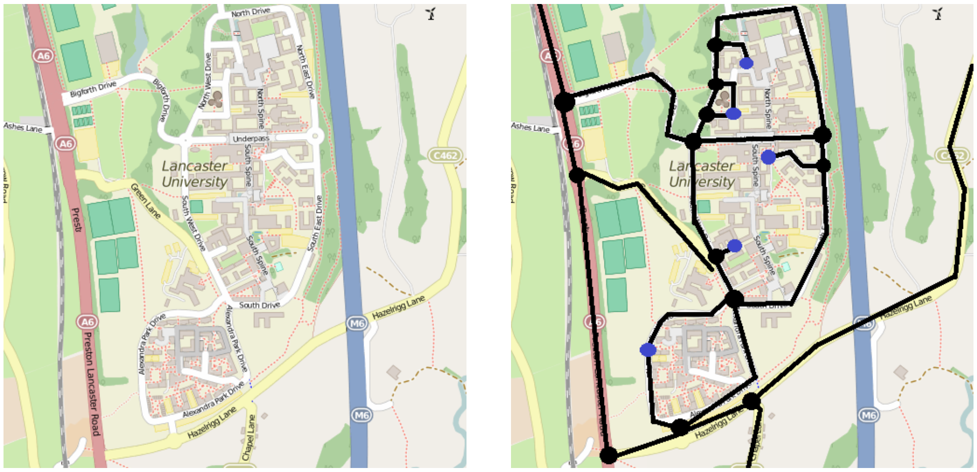


Figure 1.1.2: Transforming original map (left) to network version (right) (OpenStreetMap contributors, 2017).

The pictures in Figure 1.1.2 show how a road map can be converted into a network. To ensure that the network is easy to understand each black line represents two directional links. Only the major roads have been included and some of the more complicated junctions have been simplified to amalgamate any very short links within them into a single node. The black nodes are road junctions while the blue nodes are customers.

The network representation of the road map looks very similar to the actual road layout when it is displayed like this. The network is defined by a set of nodes and a set of links such that link $i.j$ starts at node i and ends at node j. Visual information like the exact positioning of the road is lost in the transition from map to network and different road maps can result in the same network representation.

1.1.5 Mathematical definition of a VRP

A vehicle routing problem finds the optimal route using a linear program. A linear program defines the problem using a set of variables with corresponding constraints. Variables include the links of the network and the volume of petrol on a vehicle. The problem is modelled using a network, where nodes are customers or the depot and links represent roads between them.

There are many different ways to define a linear program for a VRP. A basic, time independent version of a one vehicle VRP model is set out below via equations (1.1.1)-(1.1.4). There are N links in the network. Let $\mathbf{z}$ be the decision vector of size $\mathbb{R}^N$ representing all the edges of the network. Then $z_{i.j}$ takes the value 1 if link $i.j$ is part of the optimal route and zero otherwise. In hazardous routing the risk objective is typically denoted f. This is a function which maps $\mathbb{R}^N$ to $\mathbb{R}$. The VRP needs to solve:

$$\min f(\mathbf{z}) \tag{1.1.1}$$

$$s.t.\ \text{Routing constraints} \tag{1.1.2}$$

$$\text{Other constraints} \tag{1.1.3}$$

$$z_{i.j} \in \{0,1\} \quad \forall z_{i.j} \in \mathbf{z} \tag{1.1.4}$$

The constraints can be split into two types. Firstly we need to ensure that the route is connected, and visits all the customers and the depot. Theses constraints can be referred to as routing constraints.

The other constraints ensure that the linear program is as close to the real life problem as possible. There are many different constraints which can be included, such as limiting the capacity of vehicles and ensuring that vehicles arrive within a set time window. A simple travel time limit constraint is given as an example. Let J_M be the maximum journey for a vehicle and $y_{i.j}$ be the travel time for link $i.j$. Then the optimal route must satisfy

$$\sum_{i.j} z_{i.j} y_{i.j} \leq J_M. \tag{1.1.5}$$

The coefficient and limit values of the constraints are specific to the problem that is being solved. However, in the petrol routing problem the travel times, $y_{i.j}$ are time dependent. This requires the introduction of another variable, t, for time of day, such that the travel time depends upon the time, $y_{i.j}(t)$. The more constraints and variables that are included the more difficult the linear program is to solve, but the constraints ensure that the VRP is close to the real life problem. This issue is discussed further in Section 1.2.

1.1.6 Travel time

Travel time in the problem. Travel time is a coefficient in the petrol routing problem. Each link has a different travel time and for any route the VRP calculates the arrival times for the schedule, based upon the travel times along the links to reach that point. For the solution only the travel times of the links in the optimal route are required, but selecting the optimal route requires the travel times for all links.

The schedule tells the delivery drivers where they need to be and when. As customers need to accept the petrol delivery, vehicles should arrive at customers as close to the arrival time as possible. Sometimes a vehicle may be able to make an early delivery or wait until the arrival time to deliver but this is not always safe. If the actual travel time is much longer than the prediction then a delivery may have to be cancelled, resulting in a loss of money to the delivery company. A decision will be made en-route as to whether to arrive late or go onto the next customer, potentially using a real time VRP.

The vehicles' travel times between customers and the depot are time dependent. In high volumes of traffic, for instance at rush hour, the overall speed of the traffic is slower, resulting in journeys taking longer. There can also be unexpected delays such as accidents or heavy congestion. Thus the travel time varies over the entire day and has considerable stochastic variability.

The travel times of roads for the next day are unknown and hence must therefore be predicted in advance. A VRP assumes that the travel times are correct, hence the travel time estimates along the roads need to be as accurate as possible. This ensures the expected arrival times are close to the true arrival times, and the optimal route is correct for the real world setting.

Modelling travel time. The petrol routing problem requires estimates of travel times for each road individually and at different times of the day. The VRP then combines the travel times for each road within the route to give the overall estimated travel time. They similarly give the arrival and departure times from each customer.

The predictions of travel time must be appropriate for the petrol routing problem. Therefore the travel time must be predicted for the next 24 hours for each road. There is also the question of how detailed the predictions should be. If travel times remain constant for the day then one prediction will be appropriate but if the travel time structure changes every hour then the predictions should change at least every hour.

The road conditions may be such that the travel time predictions can be generated for all the roads for a week and then these forecasts can be repeated whenever the model is run. However there are likely to be residual effects from the current travel time. The current travel time will give a base level for how busy the road is and the previous day(s) can provide additional information. By generating the travel time predictions every time the VRP is calculated, many models can take this into account.

There are many different methods of predicting travel time. Simpler models are based upon other forms of data than the observed travel times, for instance using the speed limit and length of each road. Alternatively models can be created using travel time data themselves, such that the past travel times predict the future. We focus upon those using the observed travel times as they require only one type of data and it is easy to check the predictions are accurate.

1.1.7 Data analysis

Data. To select the best model for the travel times of roads within the petrol routing problem we require travel time data. The data that are used within this thesis are from Highways England. Highways England are responsible for maintaining and managing the major roads within England. These major roads are motorways and some of the most important A-roads and make a large network that spans the country as can be seen in Figure 1.1.3.

To collect such data they have installed a large number of sensors and the travel time between two sensors can be calculated by comparing the times at which a vehicle passes them. The data is collected from automatic number plate recognition (ANPR) cameras. Each direction has sensors so data are direction-specific and are recorded every fifteen minutes. Due to the large volumes of traffic and the possibility of vehicles not being picked up on one sensor the data are recorded as a single average travel time over all observed vehicles on that segment in a 15 minute period. These travel time data are published online, `http://tris.highwaysengland.co.uk/detail/journeytimedata` (Highways England, 2016).

The road network is split into small sections between two sensors, so that the sensor at the beginning of a section is the end sensor of the section immediately before it.

Image removed due to copyright

Figure 1.1.3: Highways England road network (Highways England, 2017). See note on page III.

Sensors are positioned such that the travel times between junctions, on sliproads and at sections of roundabouts can be recorded. Some sectors are further split into multiple sections. These sections cover the entire network and can be modelled individually. However this level of detail may result in too many links in the VRP. Then multiple sections need to be combined to generate the travel time for one link.

As a result of the collection method the travel times are an average of all the vehicles on the road in that segment, which models all the lanes of the road as one. Some vehicle types, such as the tankers used in transporting petrol are limited by speed to 56mph. This means that in free flowing traffic the average may be faster than the tankers can travel at and hence a correction may need to be applied. We will assume that no correction is needed and the potential inclusion of a correction is further work.

Test network selection. To evaluate our proposed methods for travel time estimation we chose a subset of the full network which is shown in Figure 1.1.3. The majority of roads are motorways rather than A-roads and selecting a mixture of the two allows us to examine if they have different characteristics.

The test network has been selected to try and provide relevance to the potential issues with the VRP network choice. There are multiple routes between any two nodes, and these routes can be A-roads, motorways or a combination of the two. This would enable a comparison between choosing a complete graph representation and the road representation. Different road types may need to be modelled in different ways, or exhibited different behaviour.

The depot, located near node 24, is Kingsbury Oil Terminal which supplies petrol from an oil pipeline. Delivery vehicles must return to the depot which limits the size of the network. The focus when delivering petrol is the risk, and in general, the further the tanker has to travel the riskier the journey is. There are many petrol depots located across the country, including several near Birmingham and if the network extends any further then the customer will be closer to another depot.

Figure 1.1.4 shows the selected test network in its geographical setting. This has a depot which is located near node 24 and is made up of 18 nodes and 52 links - 16×2 motorway links and 11×2 A-road links. Nodes exist at any junction where there are more than two links or if the road changes type. There are two directional links, one each way, between each node. This accounts for the travel times varying directionally, for example if the majority of people commute to one location in the morning and leave in the afternoon. This allows us to include up to 18 customers, located at the nodes. Extra links and nodes would need to be added to link the customers of a real life problem to this test network.

Data Analysis. An easy test of the validity of the data is to take the segment length and divide it by 70mph, which is the free flowing speed of the road. We would expect the travel times to be about this value outside of the rush hour periods. If there are large variations then there is likely to be an issue with the data. We can also

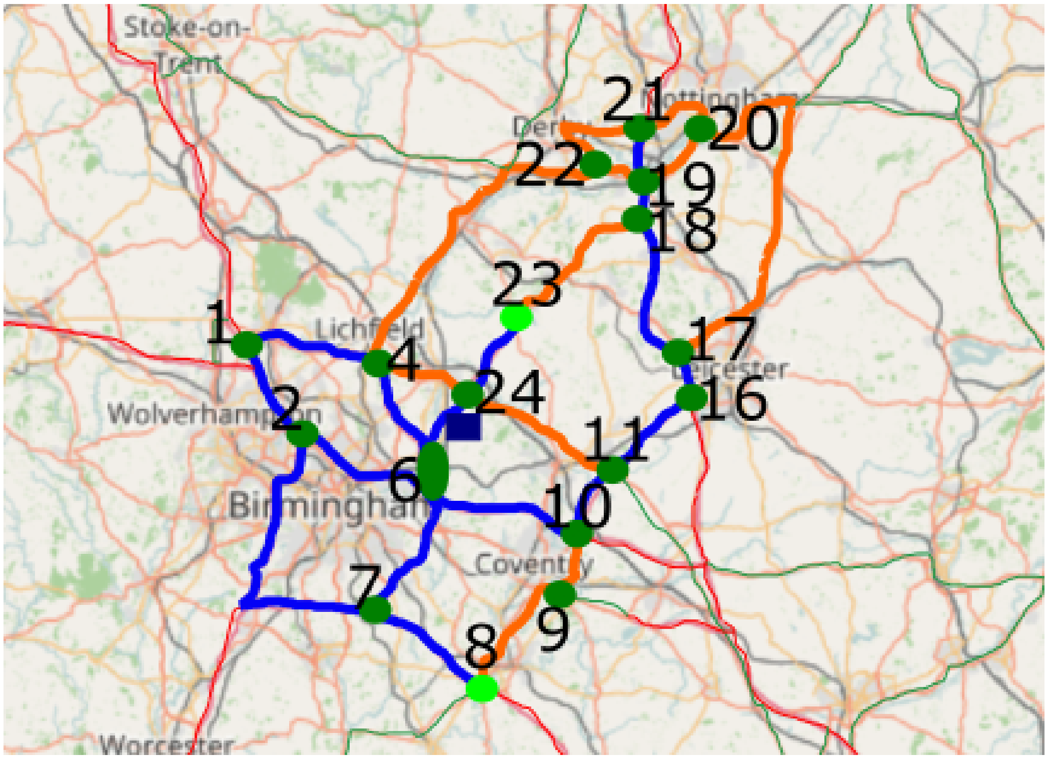


Figure 1.1.4: The road network graph, on top of the geographical road network. Motorway links are in blue, A-roads in orange with all the nodes (circles) labelled. The depot is the indigo rectangle. Nodes exist only where there is a road type change or 3 or more roads converge. Map background OpenStreetMap contributors (2017).

compare from week to week and from month to month to check there are no sudden, unexplained jumps in travel time.

Figure 1.1.5 shows the travel time on one section, on the M69 between the M6 and junction 1, for a week long period. This will be denoted as link 11.10 in our network which is shown in Figure 1.1.4. The travel time data are between Friday January 1st and Thursday January 7th 2016. The week shows much variation from one fifteen minute interval to the next, but there is an underlying structure whereby the early mornings have longer travel times compared to the rest of the day. However as this structure isn't present in all the days, there is more behind the structure than a daily pattern. Detecting patterns that are longer than a day is difficult with only a weeks worth of data, especially when the underlying structure is obscured by noise. As daily life varies over the week there may be a weekly pattern in addition to a daily one.

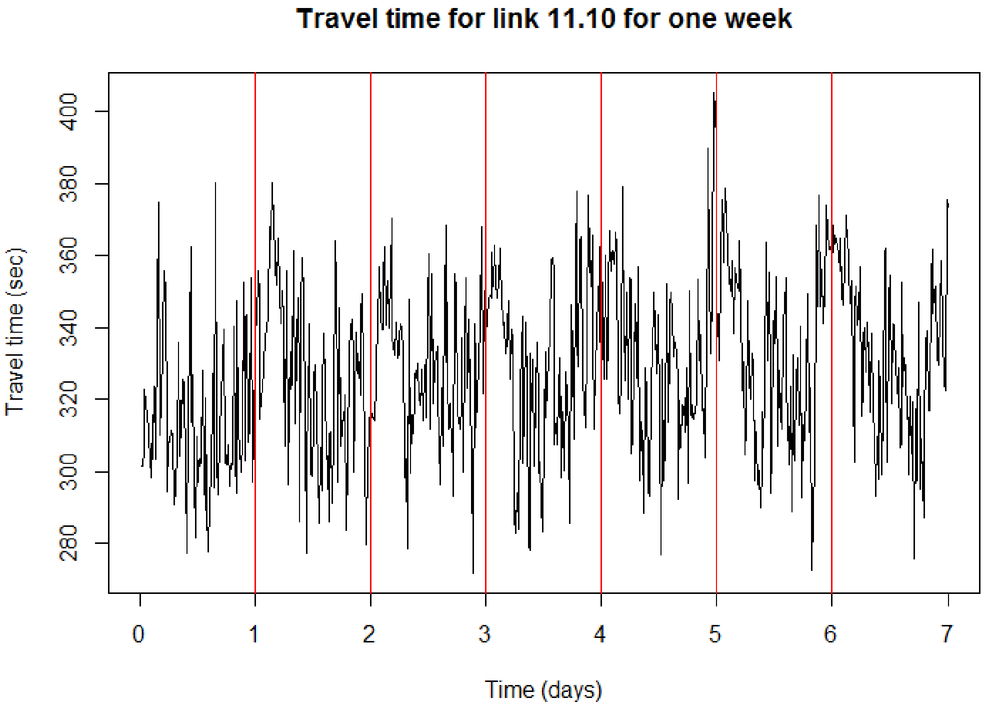


Figure 1.1.5: Travel times for the first week in January for link 11.10.

Figure 1.1.6 plots each week of data on top of each other. Each coloured line represents one week from Monday to Sunday, with the black vertical lines separating the days. This figure allows direct comparison between the coloured lines for each day individually, as well as over the whole seven days. Even with the variation from one time point to the next there is a structure over all seven days. This is particularly clear near the start of day three and day four where there are no values below 350 seconds, indicating there is a consistent increased travel time for all the weeks.

Another feature of the data is that it appears to take longer at night time as opposed to the expected rush hour periods. This pattern occurs for the almost all days within the three month period suggesting that the it is not a temporary sensor issue. The speed limit could have been reduced due to roadworks but it is too long ago easily to check this. Another possible reason is that more of the vehicles are lorries and therefore slower moving. However the data doesn't include the break down of vehicle types.

The whole plot suggests that there is a weekly pattern with the underlying structure remaining the same from one week to the next. Days 6 and 7 are overall lower than the rest of the days, this is not surprising as these are the weekend days. The underlying structure for them is similar to each other. This suggests that it may be possible to model the two weekend days together. Similarly the weekdays follow a similar structure to each other. To do so would simplify the model to two separate day structures. However more detailed analysis in Section 2.3.4 shows this doesn't hold.

Another possible simplification is grouping the data into larger blocks than the fifteen minute intervals. For example modelling the night time as one block and rush hour as another. The plot shows that the underlying structure isn't constant for very long, and there is no clear night time region. Hence the data doesn't support this simplification.

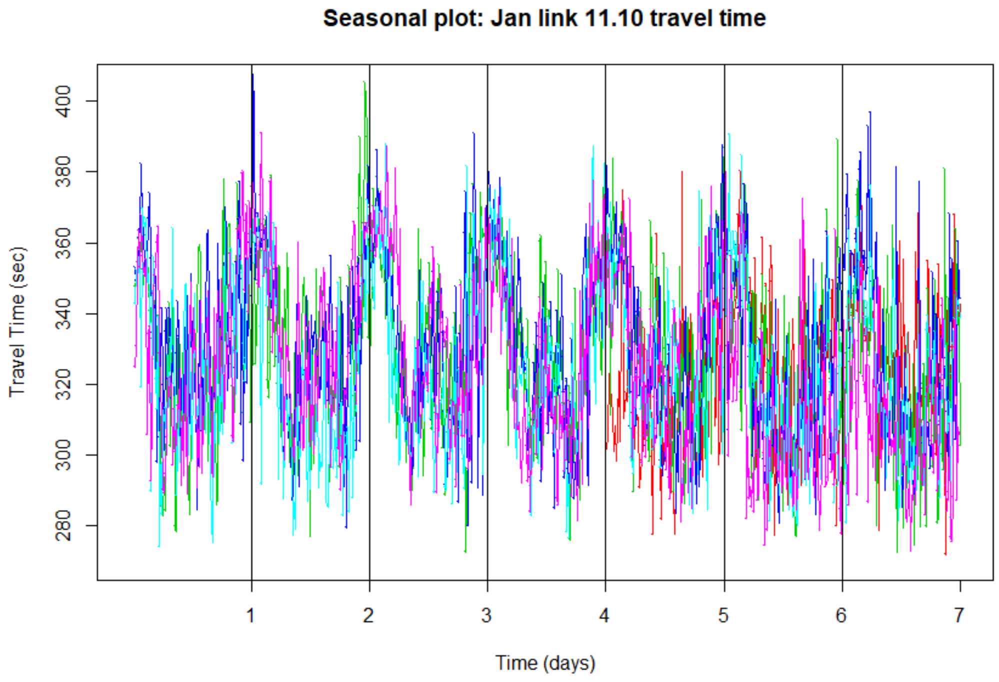
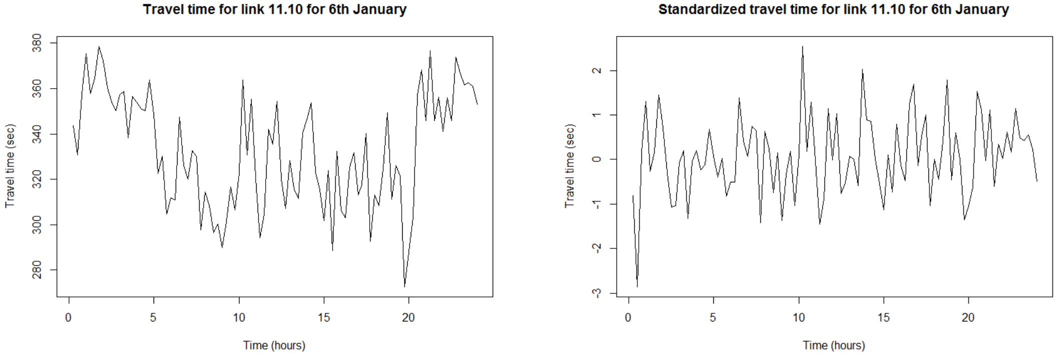


Figure 1.1.6: Travel times for each week January in for link 11.10. Monday is day 1.

Standardization. One of the issues with travel time data is that it has many differing factors that result in the observed travel time. Travel time may vary in congestion, with rush hour being slower. Also there are fluctuations in the travel time from one time point to next. Vehicles all travel at different speeds which is limited by those ahead and hence the average travel time is unlikely to remain constant. Figure 1.1.7a shows the data from just the 6th of January of the time series in Figure 1.1.5. There is a clear pattern in the data with the early morning and evening taking longer than the rest of the day, but with substantial short-term variations around this pattern.



(a) Travel times for the 6th January for link 11.10.

(b) Standardized travel times for the 6th January for link 11.10.

Figure 1.1.7: Travel times for the 6th January for link 11.10 before and after standardization.

Many of the models that are used to predict travel time require assumptions including *strict stationarity* of the data. Strict stationarity states that the probability of observing a set of points from the data at times $t_1, \ldots, t_n$ is the same as observing them at times $t_1 + a, \ldots, t_n + a$, for all lags a and possible combination of time sets. This is a strong assumption that we relax to second order stationarity which requires only that the data must have a constant mean and variance. These two properties are a consequence of the strict stationarity.

To achieve second order stationarity the data are transformed via a process called standardization which is discussed in Section 2.3.5. This removes a pattern from the data, leaving the random changes from one time to another remaining. Here it removes the weekly pattern. Figure 1.1.7b shows the standardized version of Figure 1.1.7a.

The standardized series is centred upon zero, and the variance is approximately the same across the series. Most of the variation is between -1 and 1 with little remaining pattern over the day. This suggests that the data has been successfully transformed to be stationary.

The models can then be applied to the standardized series to generate standardized travel time predictions. These predictions are then unstandardized using the standardization method in reverse to provide the travel time predictions for the VRP model.

Overall aim. Travel time predictions from a model are used as coefficients in the VRP. One of the assumptions of the VRP is that the travel times for each link of the network are correct. If they aren't then the optimal route the VRP selects may be incorrect, as the route may take longer than expected or be infeasible. We therefore need the predictions to be as close to the observed travel times as possible. In this thesis we look how best to model travel time on a link such that the resulting predictions can be input into the VRP.

1.2 VRP Literature Review

Here we conduct a brief literature review of the VRP literature. Further literature reviews of predicting travel time and extreme delays are conducted in the corresponding chapters.

There is a large expanse of literature for the vehicle routing problem which can be split into a very wide range of sub-problems that have similar attributes. An attribute is a feature of the problem, such as considering the capacity of vehicles. Some of these attributes are essential for the petrol routing problem, as defined in Section 1.1.3. The more attributes that are included in a model, the more complicated the model and hence most papers focus upon one main attribute and simplify or neglect the others. The closest sub-problem to the petrol routing problem is the hazardous materials problem, but there are many different attributes that have been focused upon within this literature. For example Abkowitz and Cheng (1988) focus on the capacity of the vehicle and the resulting risk of an accident in relation to this and don't consider travel times at all in their VRP whereas Wijeratne et al. (1993) include stochastic travel times but ignore the vehicles' capacities.

We start by defining all the attributes that are relevant to consider for the petrol routing problem. The attribute for each paragraph is indicated by italics. A collection of papers by these attributes is presented in Tables 1.2.1 and 1.2.2. These have been defined such that the closest model to the real world petrol routing problem, as defined in Section 1.1.3, would have ticks in all columns, with the potential exception of the other objective column. However, any solvable model would need to compromise by not including some of the attributes and all the papers cover only a small subset of them. We first consider the attributes that are time related as these are relevant to the travel time. These are highlighted in bold in the tables.

One key attribute is whether the problem is *dynamic* or static. A dynamic problem includes variables and coefficients that change with time, such as the load. Travel time is dynamic and hence the petrol routing problem is dynamic. Some VRP problems focus upon other attributes of the problem such modelling the risk which results in the overall problem being static (Huang et al., 2010).

Dynamic models fail to take into account the uncertainty in the coefficient values that appear in equations (1.1.1)-(1.1.4). The uncertainty in travel time can be seen in Figure 1.1.6 where a road may be congested or a wide load can disrupt traffic. Thus *stochastic* models are preferred to allow some of the uncertainty to be considered. Stochastic models have different values for each coefficient with associated probabilities. The problem is a lot more difficult to solve as the selection of the optimal route is more complicated. Hence most of the stochastic models aren't dynamic and if they are they are very short term predictions and have very few time periods of predictions (Zhang et al., 2013; Miller-Hooks et al., 1998).

Androutsopoulos and Zografos (2012) includes the majority of the possible attributes for the petrol routing problem - it considers the load, allows scheduling, is offline and plans for the short term. In addition to considering time windows it is also complete and requires that one customer is served by one vehicle and all vehicles are the same. The problem is dynamic both in time for speed and acceleration and load for risk. However the travel time isn't stochastic.

There are a variety of ways that stochasticity can enter a vehicle routing problem. In our petrol delivery problem from Section 1.1.3, the customers and their demands are already known however there is uncertainty in both the travel times and the risk. None of the stochastic papers consider risk as an objective function and all of the stochasticity is in the travel time. With the exception of Laporte et al. (1992) and Errico et al. (2013) the objective is the total travel time, although Wang and Lin (2013) considers the travel time by including it in a cost function which considers the total cost of the route to the company.

Some papers have no *time consideration*, because none of the attributes are time dependent and they focus on other attributes than the travel time. For example Current and Ratick (1995) consider the cost in a five objective hazardous material problem. This is unrealistic when considering travel time within a VRP.

One simplification is changing time dependent variables and coefficients from being continuous to T discrete sets. Set i is defined by a start time t_{i-1} and an end time t_i and contains all times t between these points;

$$t_{i-1} < t \leq t_i; \qquad\qquad i = 1, \ldots, T. \tag{1.2.1}$$

All t in set i are homogeneous so must behave in the same way. A further simplification is to use the model within just one time period such that $T = 1$. However this removes the dynamic nature of the problem. We therefore wish to have *multiple time periods* for the travel time. This allows the travel time to change over the day, as observed in Figure 1.1.5.

Another attribute is how far into the future the models predict. The majority, are short term, finding optimal routes that are completed within a couple of hours of the start time of the model and hence predict travel time at most two hours into the future. However drivers shifts are significantly longer than two hours meaning that we require a *long term* model. Many of the short term models could be extended in length to generate routes for a whole day however this does increase the number of variables (Miller-Hooks and Mahmassani, 2000).

Time windows require a vehicle to arrive between two points in time as discussed in Section 1.1.3. There are two types of time windows - ‘hard’, which cannot be violated and ‘soft’, which can be violated but only at a price. This price is included in the objective function. There is a subsection of the vehicle routing literature that looks specifically at the Capacitated Vehicle Routing Problem with Time Windows or CVRPTW. Time windows can be included in dynamic problems because the dynamic formulation adds another index of complexity (Androutsopoulos and Zografos, 2012; Zografos and Androutsopoulos, 2004; Pradhananga et al., 2010). Including time windows in stochastic problems adds a level of complexity as the exact vehicle arrival time is unknown. Towards the end of the journey the uncertainty from each of the previous links combines so it becomes impossible to ensure that vehicles arrive within the time windows. One suggested method of including hard time windows in stochastic problems is to use chance constraints (Errico et al., 2013). Alternatively Zhang et al. (2013) uses soft time windows to reduce the probability that shipments fail to arrive when requested.

The majority of the models outlined in the papers are *offline* so the solution is run once. Online problems, such as Setak et al. (2015), enable re-routing while the journey is in progress, however this is a separate problem to our defined petrol routing problem and as discussed in Section 1.1.1 would be solved as a separate problem.

The hazardous routing literature also has a simpler version of the VRP whereby the model selects the best route between an origin and a destination. The travel time coefficients are the same as in a true VRP which is *Not Origin/Destination*. If the number of roads needs to be reduced to reduce the computational time an origin-destination problem can be solved between all pairs of customers to reduce the number of links to $c(c+1)/2$ where c is the number of customers. This reduction generates a *complete graph* which connects all customers and the depot by one link between each pair. Hence these models are still relevant to a full VRP. There are only five models that use non-complete graphs (Current and Ratick, 1995; Pradhananga et al., 2010; Zhang et al., 2013; Androutsopoulos and Zografos, 2010; Jula and Dessouky, 2006).

Some of the models only select the best routes without also creating *schedules* for the drivers. If there is no time dimension to the problem then vehicles can set off at any point and still follow the optimal route. Hence there is no need for a schedule to be returned. Similarly in the case of Errico et al. (2013) the model is stochastic but not dynamic. The route chosen will be optimal whenever a vehicle departs and the arrival times are stochastic. However, expected arrival times, and arrival windows, when the vehicle should arrive within, could be inferred from the route and the respective travel times of its links after the route has been selected.

An objective is a mathematical representation of something that should be minimised or maximised to find the optimal solution to the problem (equation (1.1.1)). Some problems are *multi-objective* and try to minimise more than one objective. In the hazardous materials literature the most common and relevant objectives are *risk* and *travel time* but *other objectives* such as cost are also considered. Considering only the dynamic vehicle routing problems, with the exception of Fleischmann et al. (2004), all are multi-objective. Stochastic models are much more complicated to make multi-objective but both Miller-Hooks et al. (1998) and Wijeratne et al. (1993) consider this in origin/destination problems.

One of the attributes in the petrol routing problem is the *capacity*. Customers want a specific amount of petrol and a petrol tanker needs to have that in its tank when it visits them. The subsection of the vehicle routing problem that considers the load is called the capacitated vehicle routing problem. The majority of the vehicle routing models include the capacity, especially those that have multiple vehicles. If there is only one vehicle and the vehicle has the capacity to visit all the customers then the load may be considered to be irrelevant because the initial load will be the same regardless of the route. The approach of Wang and Lin (2013) is both stochastic and dynamic and considers both scheduling and the load while using the same vehicles. In contrast many dynamic and stochastic vehicle routing problems concentrate on the travel time and the load is ignored (Jula and Dessouky, 2006; Carotenuto et al., 2007b).

If there are multiple vehicles then the vehicles can be of different types, e.g. with different capacities. Our petrol routing problem from Section 1.1.3 assumes that they are all the *same vehicles*. Only three papers have considered different vehicles/capacities by having separate capacity constraints for each vehicle (Patel and Horowitz, 1994; Zografos and Androutsopoulos, 2004; Pradhananga et al., 2014). We can convert their models to our context of using the same vehicles by removing the vehicle index from the capacity values of these constraints, such that the capacities are then all the same.

Another simplification is the *one customer to one vehicle* assumption. This requires that each customers' demand be satisfied by only one vehicle. This is a reasonable assumption as customers only want to deal with one delivery if possible. If a complete graph is required as well each link can be traversed only once. Wijeratne et al. (1993) allows vehicles to visit multiple customers but their model isn't a complete graph and doesn't consider the capacity.

Papers	**Dynamic**	**Stochastic**	**Time consideration**	**Multiple time periods**	**Long term**	**Time Windows**	Offline	Not Origin/Destination	Not Complete graph	Scheduling	Multi-objective	Risk	Travel time	Other objective	Capacity	Same vehicles	One cust-one vehicle
Abkowitz and Cheng (1988)							✓		✓			✓^{c}			✓	✓	
Akgün et al. (2007)	✓		✓	✓			✓		✓			✓				✓	
Androutsopoulos and Zografos (2012)	✓		✓	✓		✓	✓	✓		✓	2	✓	✓		✓	✓	✓
Current and Ratick (1995)					✓		✓	✓	✓		5	✓		✓		NA	
El-Basyony and Sayed (2014)	✓		✓				✓		✓		4	✓	✓	✓		NA	
Karkazis and Boffey (1995)	✓		✓		✓		✓		✓			✓			✓	✓	
Kheirkhah et al. (2015)	✓		✓				✓	✓			2	✓		✓	✓	✓	✓
List and Mirchandani (1991)					✓		✓		✓		2/3	✓		✓		NA	
Patel and Horowitz (1994)					✓		✓		✓			✓				NA	
Pradhananga et al. (2010)			✓	✓		✓	✓	✓	✓	✓	3	✓	✓	✓	✓		
Qiang et al. (2005)	✓		✓	✓		✓	✓		✓	✓	2	✓	✓			NA	
Huang et al. (2010)					✓		✓		✓			✓^{c}				NA	
Marianov and Revelle (1998)							✓		✓		2	✓		✓		NA	
Erkut and Alp (2007)	✓		✓	✓			✓	✓	✓	✓	2	✓	✓			NA	
Nozick et al. (1997)	✓		✓				✓		✓	✓	2	✓^{2}				NA	
Chang et al. (2005)	✓	✓	✓	✓			✓		✓	✓	2	✓	✓			NA	
Zografos and Androutsopoulos (2004)			✓			✓	✓	✓		✓	2	✓	✓		✓		✓
Miller-Hooks and Mahmassani (2000)	✓	✓	✓		*		✓		✓	✓?			✓			NA	
Bowler et al. (1998)	✓		✓	✓		✓$_w$	✓		✓	✓^{d}		✓^{c}				NA	
Miller-Hooks et al. (1998)	✓	✓	✓				✓		✓		2	✓^{c}	✓			NA	
Akgün et al. (2000)					✓		✓		✓			✓				NA	

Table 1.2.1: Papers with their corresponding attributes. Ticks mean that they have that attribute, NA meaning that it's not applicable due to only one vehicle and the numbers are the number of multiple objectives. For the other symbols c is cost, d is departure time, * is either, w is curfew, a accident rate, p is multiple pairs, m is multi paths, r means reduced by model, s is soft time windows and h is hard time windows.

Papers	**Dynamic**	**Stochastic**	**Time consideration**	**Multiple time periods**	**Long term**	**Time Windows**	Offline	Not Origin/Destination	Not Complete graph	Scheduling	Multi-objective	Risk	Travel time	Other objective	Capacity	Same vehicles	One cust-one vehicle
Wijeratne et al. (1993)		✓	✓				✓		✓		3	✓^{a}	✓	✓		NA	
Carotenuto et al. (2007a)							✓		✓			✓				NA	
Carotenuto et al. (2007b)	✓		✓	✓			✓	p		✓			✓			NA	
Lindner-Dutton et al. (1991)					✓		✓		✓			✓				NA	
Kenyon and Morton (2003)		✓	✓		✓		✓	✓					✓			NA	
Laporte et al. (1992)		✓	✓		✓		✓	✓						✓		NA	
Jula and Dessouky (2006)	✓	✓	✓		✓	✓	✓	✓	✓*				✓			NA	
Wang and Lin (2013)	✓	✓	✓		✓		✓	✓		✓			✓^{c}		✓	✓	✓
Gómez et al. (2016)		✓	✓				✓	✓					✓		✓	✓	✓
Pradhananga et al. (2014)			✓	✓		✓	✓	✓	m		2	✓	✓		✓		✓
Fleischmann et al. (2004)	✓		✓	✓		✓	✓	✓	r				✓			NA	✓
Zhang et al. (2013)		✓	✓			✓^{s}	✓	✓	✓				✓		✓	✓	✓
Androutsopoulos and Zografos (2010)	✓		✓	✓		✓	✓	✓	✓	✓	2	✓	✓		✓	NA	
Desai and Lim (2013)	✓	✓	✓			✓		✓	✓				✓			NA	✓
Setak et al. (2015)	✓		✓				✓	✓			2		✓^{c}	✓	✓	✓	✓
Tarantilis and Kiranoudis (2001)			✓				✓	✓				✓				NA	✓
Bertsimas and van Ryzin (1993)		✓	✓				✓	✓					✓		✓	✓	✓
Errico et al. (2013)		✓	✓			✓^{h}	✓	✓						✓		NA	✓

Table 1.2.2: Papers with their corresponding attributes continued. Ticks mean that they have that attribute, NA meaning that it's not applicable due to only one vehicle and the numbers are the number of multiple objectives. For the other symbols c is cost, d is departure time, * is either, w is curfew, a accident rate, p is multiple pairs, m is multi paths, r means reduced by model, s is soft time windows and h is hard time windows

1.3 Thesis structure

Having defined the petrol routing problem in this chapter, we now look at the inclusion of travel time data within this setting. The structure of the remainder of this thesis is as follows.

Chapter 2: Single Link. The first chapter covering novel work analyses how to predict travel time on a single road section such that it can be included in a Vehicle Routing Problem. Different models for travel time from both the travel time literature and the vehicle routing literature are compared theoretically and computationally. These include standard time series models such as the exponential smoothing model and the seasonal ARIMA model using a standardized travel time series. The computational results are compared first over one link and then six links individually. We conclude that the optimal approach for predicting travel time on a single link is the ARIMA model. The optimal orders for the terms of this model are found to be low and to change for each link.

Chapter 3: Extreme Delays. Sometimes travel along a road can be disrupted due to an accident, which causes a delay time longer than the standard model would predict. Chapter 2 models the normal travel time behaviour. As the VRP assumes that travel time predictions are correct then including estimation of the delays will improve the overall model. Every delay is considered to have a cause and delays with the same cause are modelled together via a point process. A point process is used to model events which occur at random, at single points in time, similar to the causes of the delays. In Chapter 3 we first look at the rate of the delays to check if they are random through time before modelling their scale and length with a view to incorporating this information into the travel time model. The delays cause a heavy upper tail which could be modelled using the exponential distribution. However a better fit for both the scale and length of the delays are versions of the generalized

Pareto distribution which has an even heavier tail. These models are time invariant as the data has been standardized. The time invariant model requires unstandardization to be able to estimate the scale and length of a delay. This results in a final model for the probability of a delay on a link, combined with models for the scale and length of that delay.

Chapter 4: Network model. Building from Chapter 2, Chapter 4 studies the possibility of including information from the road network into the prediction model. The petrol routing problem requires forecasts for all the links, and hence using information from the surrounding links doesn't require any more data. Surrounding links could provide more information as traffic passes from one link to another and provide substitute data if a sensor fails on a link but the neighbouring connected links are observed. Models of different complexities are considered, as the network can be large, resulting in models that take too long to run. All models are compared against the single link model of Chapter 2 as a baseline. The optimal model is the link beta model which uses the ARIMA forecasts from neighbouring links in a linear combination generated from a least squares estimate.

The last chapter, Chapter 5, summarises the conclusions drawn in the thesis about the best model for travel time in a petrol routing problem and looks at potential further work including combines the work of the previous three chapters to create one model that produces forecasts for the next day.

Chapter 2

Evaluating methods for modelling and forecasting travel time on single network links

2.1 Introduction

The vehicle routing problem (VRP) chooses the best route for vehicles to take when delivering to customers. Each vehicles' route starts and ends at the depot where the goods are loaded ready for delivery. Multiple vehicles are available to conduct deliveries and vehicles have a maximum capacity which limits how many customers they can deliver to, based upon the customer demands. There are many variants of this problem, such as having multiple depots and modelling the unloading times at customers.

We have defined our specific VRP which we call the petrol routing problem in Section 1.1.3. One of the inputs that is required for this model is the travel time which we introduced in Section 1.1.6. This chapter focuses upon the modelling of travel time such that it can be incorporated in the VRP model.

2.1.1 Model selection process

There are many different models for the travel time of vehicles along roads. This chapter studies the models from the position of a VRP practitioner who wishes to select the best model of travel time so that it can be input into the VRP. We aim to set out a framework of steps that can be followed by anyone who has travel time data, so the selected model will be accurate and appropriate for use in a vehicle routing and scheduling problem.

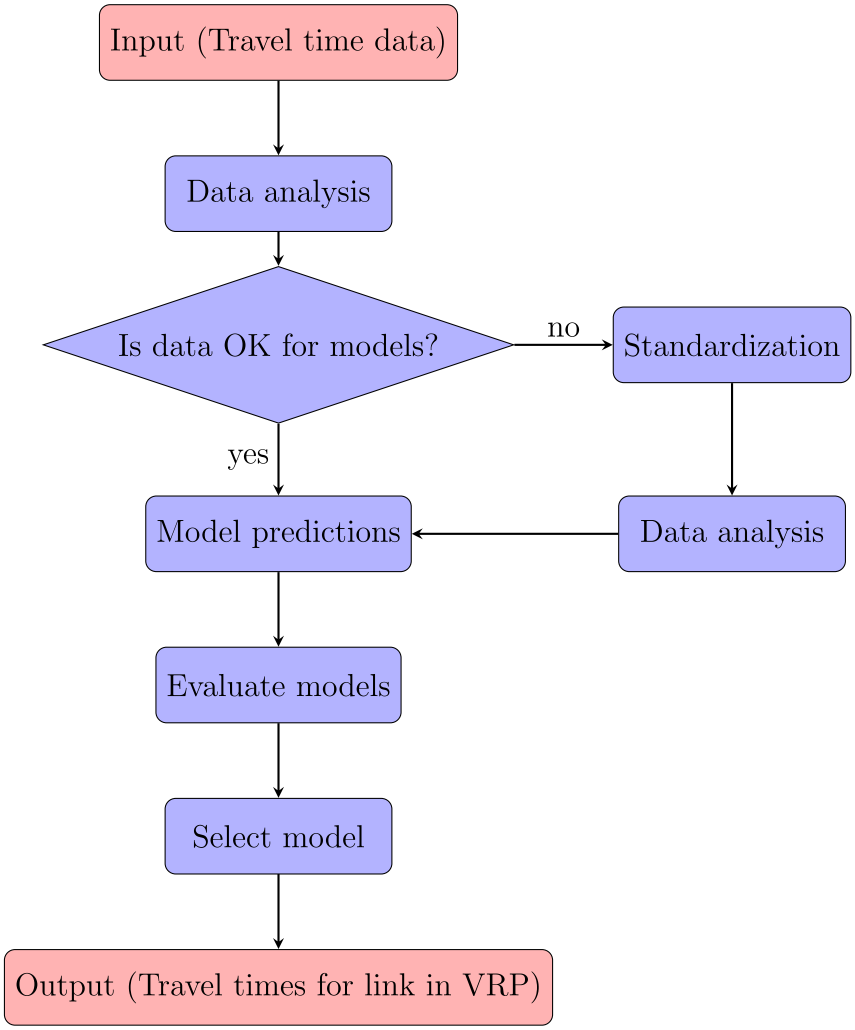


Figure 2.1.1: Travel time prediction model selection process. Flow chart of how to decide the best model which is also the structure of this chapter.

It is assumed that the VRP practitioner already has a predefined VRP and hence we will keep the details of the vehicle routing problem as general as possible. As an example our context refers to the petrol routing problem as defined in Section 1.1.3. VRPs model the future where the travel time is unknown. Hence a prediction must be made.

A prediction requires data of some form. The initial step is to ensure that the data can be used for forecasting. Section 2.3 contains a discussion of currently available datasets, with their advantages and disadvantages, in addition to a variety of different considerations for the data should someone wish to collect the data themselves.

After this initial step we look at the model selection process. Ideally this step would be automated such that a user would enter in their travel time data and receive the travel time predictions as the output. If the predictions are required more than once, the selected model may also be output to speed up the prediction time, since the same model will be used in future predictions. Unfortunately there are several stages that require human evaluation and hence the model selection requires at least some user input. The full process is shown in the flowchart in Figure 2.1.1.

First the data is analysed to identify any key features which will impact the models that can be used. This analysis of the raw data is in Section 2.3.4. If the data isn't of an appropriate form it can be standardized. Standardization, which is the focus of Section 2.3.5, should remove some of the seasonality which can be checked by reanalysing the data.

This chapter evaluates single link models of travel time. The single link model looks at links in isolation, predicting the future travel time for the link using only information from the link. There will be one model for each link of the network. In Chapter 4 we will look at network level models which incorporate information from surrounding links. The evaluations of this chapter are also applicable that chapter and build upon the optimal model we select in this chapter.

Evaluation across multiple days may lead to different models and parameters. We wish to be able to generate the best forecast for the next day using our dataset to select one model with corresponding parameters. The optimal parameters for each model will be found using the relevant training set. The data is repeatedly split such that each full day can be forecast using all the data beforehand as a training set. Each of these optimal models will then give the travel time predictions for the time horizon that the VRP requires. A VRP may require travel times for the entire day or only an hour. The predictions for the travel time on our dataset can be found in Section 2.4. Additional considerations such as missing data may exist, dependent upon the dataset. These are discussed in Section 2.4.1.

These predictions must then be evaluated against the true travel time values of the forecast day to decide which is the best model. The evaluation is in Section 2.5. Finally the selected model will return the travel times for the link so that they can be inputted into the VRP.

Summary of chapter. The format of this chapter is as follows. In Section 2.2, a literature review is conducted of the existing travel time models in the VRP literature, travel time literature in other contexts and the forecasting literature of similar models that may be applicable. Sections 2.3 - 2.5 then implement several of the suggested models with a dataset. This follows the ordering suggested by Figure 2.1.1 to show how the flowchart can be implemented on our dataset and then applied to other datasets.

2.2 Previous models for travel time

There are many different ways of modelling travel time; as an input into the vehicle routing problem, as a separate problem and modelling travel time in other contexts. These methods are usually applied to one link in isolation, rather than across the whole network. We will first define travel time over a link, in the context of a VRP.

Travel time in a VRP. A VRP requires the travel times for each link in the network. Vehicles move along physical roads, often across multiple lanes of traffic with differing speeds. The problem has been simplified such that one link represents the only path between the two nodes and a link $a = (i, j)$ is identified by the two nodes i and j.

The travel time, y_a, for link a is the time it takes to travel from one end of the link to the other and is time dependent. Let $A(t)$ be the arrival time at the start of the link and $D(t)$ be the departure time of the vehicle exiting having entered at time t. Then the travel time of a vehicle entering the link at time t is defined as:

$$y_{a,t} = D(t) - A(t). \tag{2.2.1}$$

This is usually measured in either minutes or seconds depending upon the context. As this chapter considers links in isolation, for simplicity we drop the a. The actual travel time y_t for the future is unknown and hence a prediction must be made. This predicted travel time is referred to as $\hat{y}_t$.

One type of vehicle routing problem considers a single travel time for each link with no uncertainty. In forecasting this is called a point forecast. Alternatively some models consider uncertainty by using distributions for the travel time to account for the uncertainty of the real world. This problem is much more complicated, as the optimal route is only optimal to a probability this gives more realistic models.

Petrol routing problem characteristics. For any VRP we need the model to satisfy several required characteristics. The petrol routing problem, as defined in Section 1.1.3 provides a forecast for the next day so the forecasting method should provide forecasts for this period. The travel times vary across the day so the forecasting model should provide forecasts that are time dependent while still providing accurate forecasts.

The forecasting models can be split into a variety of different sub-categories. These include whether the problem is dynamic or stochastic, if the output is a single travel time or a distribution and if it is real-time. Models may also consider the network structure, level of time dynamic, weather or accidents.

Each day a new petrol routing problem must be solved, requiring new forecasts for each link in the model. This means that the model needs to take into account any changes that may occur in a weekly pattern or potential seasonal differences due to driving slower in wintery conditions.

Section summary. This section is structured by first discussing the travel travel time profile and its use with VRPs, as the vast majority of VRPs make the assumption that the travel times are known. Any models which have been generated using data are then split into distribution and point forecasts, as the two cannot be compared to each other and the type of VRP that can be solved is partly determined by them. Each subsection looks first at the use of each type in a VRP context, before looking at other travel time models and finally any relevant forecasting models.

2.2.1 Travel time functions

Time is continuous, but VRP models require discrete time periods. Time is split into finite chunks to enable the VRP model to potentially consider all the possible options. For each interval, τ, the travel time is modelled as one equation or distribution, and the intervals are then combined together create the *travel time function.* This is of the form:

$$\hat{y}_t = f_\tau(t); \quad t \in \tau, \tag{2.2.2}$$

whereby $\hat{y}_t$ is the predicted travel time at time t and f_τ is a function that is dependent upon the corresponding interval τ to time t. The function can also be dependent upon other factors rather than time alone.

The travel time function is normally used with point forecasts but the same idea can be applied to distributions. For a distribution with a set number K of different values with corresponding probability p_k then:

$$f_\tau(t) = g_\tau^{(k)}(t), \quad p = p_k; \;\; \forall k \in K. \tag{2.2.3}$$

The intervals for distribution estimates are generally wider, as computing accurate distributions is more difficult with less data.

Let N be the number of discrete time intervals. Then the N functions provide the travel inputs to the VRP and hence the function must provide sensible travel time estimates. Carey et al. (2003) study some of the properties that a travel time function requires for a single link in a VRP and Horn (2000) discuss several assumptions that are often made. These were discussed in the context of point forecasts but also apply to distributions. The properties are first summarised below before the assumptions are discussed.

First In First Out. The *first in first out* property or FIFO ensures that vehicles can't miss a delay along the same link by setting off later. This means that if vehicle one enters the link before vehicle two, vehicle one will always exit the link before vehicle two. This issue occurs because of the discretization of the problem from continuous time. Hence

$$t_1 + \hat{y}_{t_1} < t_2 + \hat{y}_{t_2} \quad \forall t_1 < t_2. \tag{2.2.4}$$

This is a reasonable requirement for our problem, as despite travelling upon roads with multiple lanes, delivery vehicles are unlikely to overtake due to their size and speed restrictions and most delays affect all lanes of a road. To reduce the model complexity all lanes are modelled as a single link.

Time dependence. The travel time for time t depends upon the traffic up to and including time t but not later. Time is linear and events that happen further into the future won't impact now. This only refers to the state of the traffic at time $t + 1$ etc. which will be unknown. Extra information such as bank holidays, for which the impact can be estimated in advance, can be included in the model.

Constant reduction. The last property is that if traffic flows through the link at the same constant rate any model should reduce to the simple constant model with no variation through time. The travel time is thus just a constant value.

Constant speed. The speed on a link should be constant for the entire link. If this is not the case then the link should be split into sections where the speed is constant. Otherwise, if an accident occurs on the link, the effect upon the travel time will be directly linked to which part of the link the accident occurs in.

Maximum speed. All vehicles travel are assumed to travel as fast as they can (with respect to the current maximum link speed and other factors.) Many HGVs are limited to a certain speed, hence the link length may be needed to calculate a maximum travel time which is lower than the free flowing traffic.

Time invariant network. The defined network is time invariant so doesn't change. All roads remain in the network and none have their speed reduced by roadworks. This is particularly important in complete graphs which are discussed in Section 1.2.

Positivity. All values must obey the rules of the physical world. The distance, speed and travel time are all positive.

Simple travel time profiles

We now look at the different ways that have been suggested for constructing travel time functions in VRP problems. These start from the very simplest models which are then improved to try and fit the assumptions and properties of the ideal travel time function.

Plotting the travel time function over time gives the *travel time profile* which is much easier to understand. Creating travel time profiles for distributions is considerably more difficult so the following are all point forecasts.

Constant travel time. The simplest way that travel time has been modelled is to assume that each link has a constant travel time for the entire time period under consideration (Pradhananga et al., 2010; Zografos and Androutsopoulos, 2004; Pradhananga et al., 2014). Hence $\hat{y}_t = c$ for some constant c.

Step function. A logical extension to having a constant travel time is to make the travel time depend upon the time of day. The simplest travel time function that is time dependent is a step profile (Malandraki and Daskin, 1992). The day is divided into a given number of time periods and each time period has a constant travel time. Thus $f_\tau(t) = c_\tau$. A five time period step function can be seen in Figure 2.2.1.

However the step function doesn't obey the FIFO property if the differences between two steps are great enough. For example, in Figure 2.2.1, if a vehicle leaves at 19:59:00 it will arrive at 20:03:00, however if it leaves at 20:00:30 it will arrive at 20:02:30, 30 seconds before.

Slanted step function. A slanted step function removes the discontinuities. For a short interval around where each interval τ_i ends and τ_{i+1} begins, the function changes with a constant gradient between the two constant values. Figure 2.2.2 shows how this would look. If the gradient isn't too steep then the FIFO property will hold. Malandraki and Daskin (1992) suggest this function without any advice on how to

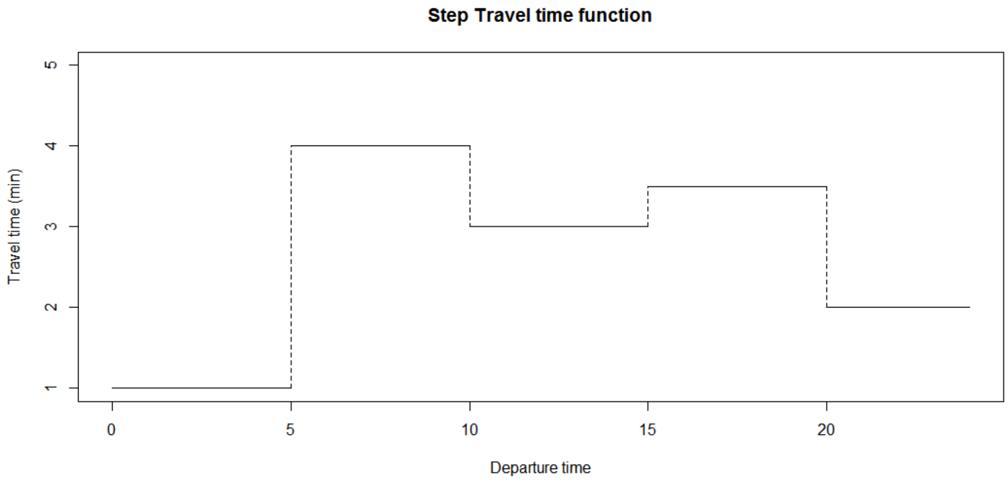


Figure 2.2.1: Step travel time function plot

generate the travel time profile. For a step-function with predefined time intervals the average can be taken for each period to give c_τ, but it is unclear how wide the slanted connecting sections should be.

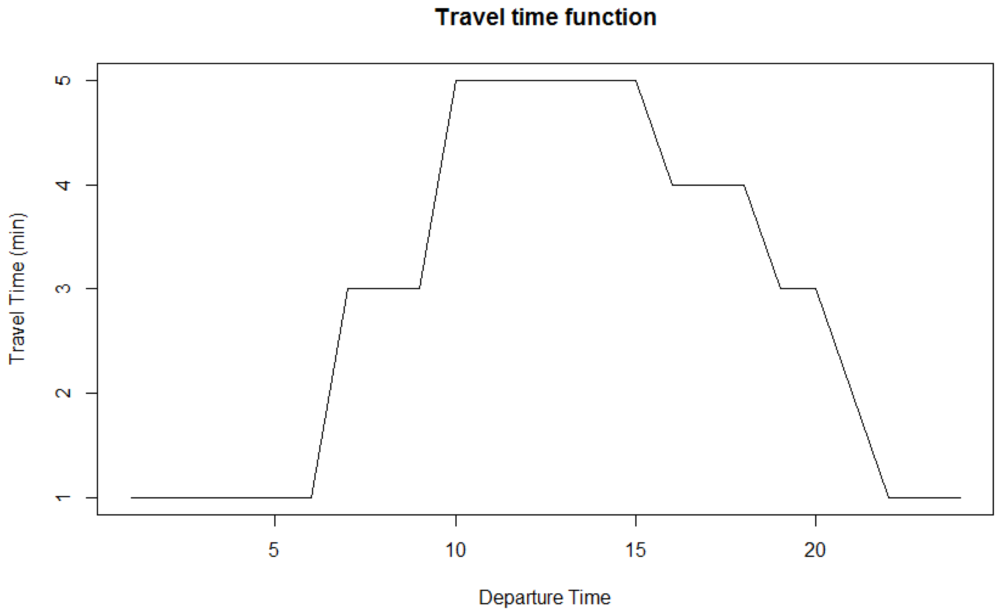


Figure 2.2.2: Slanted step function plot of travel time.

More complicated functions The previous functions have all been linear within each time interval. Relaxing this permits a much wider range of functions.

We next look at how to create travel time functions that accurately reflect the travel time for a given road segment. We will begin by looking at point forecasts which can be used to create an estimate for each time interval.

2.2.2 Point forecasts

Point forecasts are named because they provide at most one value as a prediction of the travel time for a given time in the future. Many of the petrol routing problems require single values for each link of the network and point forecasts are a good way to get an idea about which types of methods are good for forecasting if a distribution is required. We wish to forecast an entire day ahead and hence we require multiple forecasts as the travel time changes over the day. These point forecasts are then combined to make the travel time functions that have been discussed in Section 2.2.1. If the forecast is for an interval this creates a step function, however this can be adapted to create a slanted step function or a more complicated function.

The travel time functions are the inputs to the vehicle routing model which are unique for each link. Then the VRP will find the optimal route, taking into account the time dependent travel time. Point forecasts ignore any uncertainty there is in the future. Confidence intervals are one method used by forecasters to include some form of uncertainty, while still providing a point estimate.

We wish to look at models which provide travel time estimates that can be used in a VRP. There are a wide variety of models that have been proposed for forecasting travel time in many different contexts and we include those that can be adapted.

We consider fourteen relevant models, with their characteristics, which are summarised into two tables. Table 2.2.1 summarises most of the time related aspects, while Table 2.2.2 is the remaining aspects. The tables are composed of the different methods from the travel time and VRP literature. The papers have been grouped by the method that they use to predict the travel time and these are discussed in more detail later in the section. As each paper has different characteristics they each have one line.

The *interval* (width of τ) is how long the travel time function remains the same. This is limited by how frequently the data is recorded. For some of the speed/time rule methods the period is split into intervals of an uneven size. To predict travel time for petrol transportation for an entire day we require an interval that is large enough to show the changes throughout the day but not too frequent that the travel time function fluctuates a lot and there are a lot of points to forecast. This is largely dependent upon the dataset and the dataset we use later in this chapter has an interval of 15 minutes. Both Horn (2000) and Dong et al. (2013) also use 15 minute data. The other time steps vary from 3 minutes to 1 hour which are close enough to 15 minutes that the methods can also be applied to 15 minute data.

The *horizon* is the furthest away travel time prediction. This is recorded in the units of the interval so it is clear how many time steps ahead the models are forecasting. For the petrol routing problem, as defined in Section 1.1.3, we require travel time predictions up to one day or 24 hours ahead which equates to 96 steps on the 15 minute scale. This is considered to be long term and all of the models generate forecasts that are too short term, both in terms of time and the forecast steps ahead. The longer the horizon the greater the uncertainty.

The *time dynamics* show to what level a method deals with the changing travel times throughout a season. For some models there is no consideration of any change (none). When considering real time models the immediate future may be consider to be far more dependent on the immediate past than a daily profile. Other models consider the time of the day (tod), while the most detailed is to consider the day of the week as well as the time of the day (dow). There are also other considerations, such as grouping weekdays as one category. Ideally we want a method that is day of the week as our analysis in Section 2.3.4 shows that travel time varies this way.

Method	**Interval**	**Horizon**	**Reference**	**Time dynamic**
Previous values	1h	10*1h	Taniguchi and Shimamoto (2004)	tod
Simulation	1h	10*1h	Taniguchi and Shimamoto (2004)	tod
Speed/time rules	15 min	30*15min	Dong et al. (2013)	none
Speed/time rules	time window	NA	Ichoua et al. (2003)	tod
Speed/time rules	5 time windows	NA	Androutsopoulos and Zografos (2012)	tod
Speed/time rules	15min	48*15min	Horn (2000)	tod
ARIMA	5min	1*5min	Zhang et al. (2015)	tod
k-nearest neighbours	5min	24*5min	Nikovski et al. (2005)	tod
Support vector regression	3min	1*3min	Wu et al. (2003)	none
Flow models	NA	NA	Carey et al. (2003)	tod
SARIMA	15min	4*15min	Guin (2006)	dow

Table 2.2.1: Summary table time related properties of point forecasting methods. tod stands for time of day and dow is time indexing which is both day of the week and time of day.

Method	Reference	Lit	Context	Data	Future
Previous values	Taniguchi and Shimamoto (2004)	VRP	dynamic VRP with time windows	Y	S/L
Simulation	Taniguchi and Shimamoto (2004)	VRP	dynamic VRP with time windows	Y	RT
Speed/time rules	Dong et al. (2013)	VRP	shortest path	N	S/L
Speed/time rules	Ichoua et al. (2003)	VRP	shortest path	N	S/L
Speed/time rules	Androutsopoulos and Zografos (2012)	VRP	HAZMAT VRP	N	L
Speed/time rules	Horn (2000)	VRP	speed changes in VRP networks	Y	L
ARIMA	Zhang et al. (2015)	T-T	Mean estimates for PI	Y	S
k-nearest neighbours	Nikovski et al. (2005)	T-T	1 link model	Y	S
Support vector regression	Wu et al. (2003)	T-T	1 link model	Y	RT
Flow models	Carey et al. (2003)	T-T	TT link model for VRP	N	S
SARIMA	Guin (2006)	T-T	1 link model	Y	S

Table 2.2.2: Summary table of point forecasting methods for other key features. Travel time has been shortened to T-T while S is short term, L long term and RT is real time.

In the second table, Table 2.2.2 we include both the method and reference for easy comparison. The *literature* column divides the papers into having come from the VRP literature or the travel time (T-T) literature which forecasts travel time independently.

The next column details the *context* in which the travel time forecasts are used. The closer this is to our petrol routing problem the more likely it is that the method will be useful in our context. Only Androutsopoulos and Zografos (2012) has a problem context of hazardous material routing. Speed/time rules assume that the speed and acceleration are known in advance and hence the travel time is known. In our context the travel time isn't known in advance, hence neither is the speed nor acceleration.

Whether or not *data* is used to generate the predictions is also indicated. We will be using data to predict the travel time to ensure that the predictions are appropriate for the roads in the petrol routing problem.

The future column highlights how far into the *future* the predictions are if data has been used to generate forecasts. This also notes if predictions are real time (RT), and the method is therefore designed to be run to update the predictions or routes. The future time frame is split into two categories, short term predictions (S) and long term (L). For our model we ideally need a method that provides long term predictions as one day is 96 time steps when considering 15 minute data.

Method	**Reference**	**Data**	**Time dynamics**	**Future**
ARIMA	Ord and Fildes (2013)	N	dow (implict)	S
ARIMA + Fourier	Hyndman (2010)	N	dow	S
ES	Ord and Fildes (2013)	N	dow	S/L

Table 2.2.3: Summary table of point forecasting methods from the forecasting literature. dow means the method is time indexed over the time of day for every day of the week and S means short term and L long term. ES represents exponential smoothing.

As the forecasting methods from the forecasting literature aren't used to predict travel time we use a separate table more compact table. ARIMA models provide very short term forecasts as the forecasts converge to a single value within a few forecasts. In ARIMA with Fourier models the ARIMA part of the model also rapidly converges but the Fourier terms allow for a slight variation to include some of the seasonality.

ARIMA models shouldn't be used on datasets with seasonal variation. By removing the variation and then forecasting using the altered time series, the day of the week and time of day can be included in the model. When using the altered series in the ARIMA model with Fourier terms, the Fourier terms try to include any remaining seasonality in the data. Standardization which is one way to alter the series is discussed in Section 2.3.5. Use of models of this type can be justified if the current traffic conditions only affect the future travel time predictions in the short term. How far ahead the current travel time has an effect is also studied in Section 2.3.5.

We now look at the different methods of generating the travel time function in more detail. We will first look at methods that create travel time functions without consideration of the current travel times. These are based upon the rules of Physics, assuming that attributes such as the speed are known in advance . We will then look at more complicated models that predict the travel time, using current travel times to form predictions of varying length into the future. These forecasts form the travel time function at a time in the future.

Speed/Time rules. There are several methods that uses the speed-distance-time relationship to calculate the travel time for different intervals in the future to make the travel time function. Dong et al. (2013) use a very simple method to find the travel time for each link when a traffic link is uncongested. They then use these values to find the shortest path in stochastic time dependent networks, taking into account that some of the links are congested. If there is no congestion it is assumed that the travel time is based upon the speed and distance with certain additional conditions. The speed, distance and travel time equation is given in equation (2.2.5), where the speed is v and distance d. This relies upon the constant speed and maximum speed assumptions and is independent of the time of day. Assuming that $v \neq 0$ then the estimated travel time is

$$\hat{y}_t = \frac{d}{v}. \tag{2.2.5}$$

Ichoua et al. (2003) use this equation, letting the speed vary with time to generate a step function for the speed across time to generate the travel time profile. The overall aim is to look at how a vehicle routing between two points can be improved by using time dependent travel times. This function doesn't violate the FIFO principle because the later vehicle will travel at the slower speed until the entire link speeds up. The speeds are assumed to be known in advance so there is no prediction of the speed.

Equation (2.2.5) is also used in Horn (2000) where it is shown that for link speeds that vary with constant acceleration across different time periods the travel time function is smooth and not linear. They used data from an urban region to calculate the maximum and other set speeds for different links for five different times.

Androutsopoulos and Zografos (2012) also rely upon the same concept, using the acceleration and speed for each time period (and those adjacent) to calculate the time dependent travel time function. The travel time function is then used in a bi-objective vehicle routing problem of hazardous materials. This is the most relevant of all the end use applications however the speed and acceleration are assumed to be known for all the future time intervals.

Flow models. The models suggested in Carey et al. (2003) use the number of vehicles on the link to calculate the travel time functions. The travel time functions are for the one link and have been constructed with the aim of conforming to the desirable properties from Section 2.2.1. The first model assumes that the number of cars, $x(t)$, on the link at time t, are linearly related to the travel time through the constants a and b This leads to:

$$\hat{y}_t = a + bx(t). \tag{2.2.6}$$

The number of cars were modelled using the inflow $u(t)$ and outflow $v(t)$ from the link. This results in models that use the number of cars, as well as the inflow and outflow, in a non linear equation. These are expressed as:

$$\hat{y}_t = g(x(t), u(t), v(t)). \tag{2.2.7}$$

This uses only the inflow and outflow at the current time. A further alternative is to estimate the future outflow when the vehicle leaves the link at time $t + \tau(t)$:

$$\hat{y}_t = h(u(t), v(t + \tau(t))). \tag{2.2.8}$$

Future travel times are calculated by forward differencing, assuming the inverse function is known and require the inflow at the current time. However the paper does not demonstrate the effectiveness of the model using real data and makes no mention of how to do this. In addition the model is short term and requires information about the number of vehicles. Our dataset contains only travel time information.

k-nearest neighbours. Nikovski et al. (2005) uses 5 minute data to select the best model for each 5-minute period in the next day using k-nearest neighbours, linear regression and neural network models. This is 266 5-minute intervals from 2am to the end of the day. These models only use the travel time data to compare the prediction for one 15km section of road and then compare how the effectiveness of the model changes as the prediction horizon increases. The model then generates predictions at intervals from 5 minutes ahead up to 2 hours, using the corresponding prediction models to the time period. This model may be too computationally intensive for use in a VRP model as 266 models would need to be selected for each link of the network. The petrol routing problem also requires estimates for the entire next day rather than two hours.

Support vector regression. Wu et al. (2003) use support vector regression to predict one time step ahead, equating to 3 minutes. The real time model is suggested for displaying on traffic information systems. Support vector regression aims to fit a line to the data such that most of the data lies within ϵ of the line, where ϵ is the permitted margin of error. The distance from points that are outside this area to the area is minimised. The model is fitted for three sections of road that are different lengths, all of which are considerably longer than a single link in the network. The travel times are calculated from speed detectors and because the problem has a short prediction horizon and is real time the model isn't time-dependent.

Previous observed values. The simplest way of using past data is to use the previously observed values (Taniguchi and Shimamoto, 2004). This naive approach has been used in a VRP as a comparison to an alternative model. The VRP model is used to assess dynamic vehicle routing and scheduling when the travel times are variable. Thus the optimal route obtained using previously observed values is compared to a route obtained when the current travel times can be used to alter routes. The customers must receive their deliveries within a time window. Travel times are observed for one hour and the maximum forecast is ten hours ahead, however this method can be adapted for any time length.

Simulation. The proposed method of generating the future travel times from the current ones in order to alter the route in Taniguchi and Shimamoto (2004) is to simulate the travel time values using the current flow rates. The aim of the simulation is to ensure that vehicles are constantly on the best route and is rerun whenever a customer is reached. A simulation is run multiple times and the best solution is selected. As each vehicle approaches the end of its route the route choices become more limited and the forecast horizon shortens.

Autoregressive moving average models. Autoregressive moving average models or ARIMA models have two main components - a moving average and a autoregressive term (Ord and Fildes, 2013). The first records where the current series average is, while the second how quickly a trend dies away. The data is also differenced to make it stationary.

Autoregression assumes that the future can be forecast from a linear combination of past values. Only the last p terms are considered to be relevant. Each term contributes by a factor of ϕ_i. The remaining error e_t is white noise while a constant c remains. For consistency with the forecasting literature we use y_t to be the travel time at time t and $\hat{y}_t$ to be the current forecasted travel time for time t. More specifically an AR(p) model can be written as

$$\hat{y}_t = c + \sum_{k=1}^{p} \phi_k y_{t-k} + e_t. \tag{2.2.9}$$

Moving averages are found by considering the past errors as a linear regression rather than the actual values. Similarly only the past q values are considered, which contribute by a factor of θ_i. Let e_{t-i} be the forecast error at time $t-i$ and e_t is again assumed to be white noise. This is an MA(q) model with the forecast for time t being:

$$\hat{y}_t = c + \sum_{k=1}^{q} \theta_k e_{t-k} + e_t. \tag{2.2.10}$$

The final part of an ARIMA model is the differencing. The series is differenced d times to ensure it is stationary. Differencing can remove some non-stationary trends but results in less observations. The first order difference, y' for time t is found as follows:

$$y'_t = y_t - y_{t-1}.$$

These three ideas are combined to make an ARIMA model. First the series is differenced and then the next forecast for this series is found using both moving average and autoregressive terms. This is an ARIMA(p, d, q) model. The forecast $\hat{y}'_t$ must then have the differencing removed, such that the forecast $\hat{y}_t$ is to the correct time scale. The ARIMA forecast model is therefore,

$$\hat{y}'_t = c + \sum_{k=1}^{p} \phi_k y'_{t-k} + \sum_{k=1}^{q} \theta_k e_{t-k} + e_t. \tag{2.2.11}$$

For the ease of notation the backshift operator, B, is introduced:

$$By_t = y_{t-1}. \tag{2.2.12}$$

Equation (2.2.11) then becomes:

$$(1-B)^d\Big(1 - \sum_{i=1}^{p} \phi_k B^k\Big)\hat{y}_t = c + \Big(1 + \sum_{k=1}^{q} \theta_k B^k\Big)e_t. \tag{2.2.13}$$

This method has the advantage of being very good in the fitting stage as it reacts quickly to changes but can perform very poorly when forecasting as it requires the assumption that the time series is stationary. The model converges to a single forecast for all time points in the future which can be a long way from the data when seasonality is present.

ARIMA models are used in Zhang et al. (2015) to estimate the mean but not for forecasting the future travel time values. By its nature ARIMA models produce short term forecasts which rapidly settle down to a mean level however the model requires the seasonality to be removed from the time series before it can be fitted. The final predictions for the travel time require the seasonality to be reinserted and if in the longer term the seasonality is more dominant then the ARIMA model may still be appropriate for longer term forecasts. One method of removing the seasonality is standardization which is discussed in Section 2.3.5.

Seasonal ARIMA. The ARIMA model can't explicitly account for seasonal behaviour which is common in travel time data. However the method can be adapted to contain terms that occur only when a certain condition occurs. For example roadworks will be planned in advance so a roadwork term could be included when appropriate which would be calculated using all previous times there were roadworks. The most common version of this is the seasonal ARIMA or SARIMA model.

The data is assumed to have a seasonal pattern, where one season is defined as the period of time before the pattern repeats. The length of the season decides how many extra terms are required. For a model with season of size S we write ARIMA$(p, d, q)(P, D, Q)_S$. The p, q and d represent the order of the autoregressive, moving average and differencing and P, Q and D are the corresponding orders of the seasonal part. The model can be written

$$\begin{aligned}(1 - B^S)^D(1 - B)^d\Big(1 - \sum_{k=1}^{p} \phi_k B^k\Big)\Big(1 - \sum_{k=1}^{P} \phi_k B^{kS}\Big)\hat{y}_t = \\ c + \Big(1 + \sum_{k=1}^{q} \theta_k B^k\Big)\Big(1 + \sum_{k=1}^{Q} \theta_k B^{kS}\Big)e_t,\end{aligned} \tag{2.2.14}$$

where, as before, B is backshift operator.

There is a moving average of the previous terms that are the length of a season away. For example if a season is a week this could be all Thursdays. This is the same as for the moving average and the differencing. The advantage of using terms within the ARIMA model rather than removing the seasonality and then applying an ARIMA model is that it enables trends between seasons to be modelled appropriately.

Guin (2006) use a SARIMA with 480 time steps, or one working week, with data observed every 15 minutes, as with the dataset we will use later on. Weekends were considered too irregular to be included, due to closures for roadworks. The model was applied to one section of road with the aim of potentially providing short term travel forecasts to motorists. The further ahead the forecasts the less accurate they are. This requires a lot of extra terms in the model.

ARIMA with Fourier terms. One suggested alternative to using SARIMA models is to use Fourier terms to estimate the seasonality and then use an ARIMA model. The Fourier terms are included as exogenous variables. As the Fourier terms are seasonal they will repeat with the maximum seasonality. Let x_τ be a scaled value between such that the sequence has one period over the maximum season length. An infinite Fourier series is given in by

$$f(x_\tau) = 0.5a_0 + \sum_{k=1}^{\infty} a_k \cos(kx_\tau) + \sum_{k=1}^{\infty} b_k \sin(kx_\tau). \tag{2.2.15}$$

This is then incorporated into the ARIMA model forecast as an additional term. Let N_t represent the previous terms in the ARIMA model. Then the forecast for time t is,

$$\hat{y}_t = 0.5a_0 + \sum_{k=1}^{\infty} a_k \cos(kx) + \sum_{k=1}^{\infty} b_k \sin(kx) + N_t. \tag{2.2.16}$$

Inputting an infinite number of cosines and sines series is impossible due to computational tractability. Therefore the number of series is reduced to K, as finding a large number of coefficients for the Fourier coefficients is still computationally intensive. The size of K is recommended to be chosen using the Akaike information criterion or AIC (Hyndman, 2010). If the pattern is consistent across links, this K can be used for all links, reducing the computation time. Else $K_{i.j}$ must be found for each of the links. The forecast using the ARIMA model with Fourier terms is therefore,

$$\hat{y}_t = 0.5a_0 + \sum_{k=1}^{K} a_k \cos(kx) + \sum_{k=1}^{K} b_k \sin(kx) + N_t. \tag{2.2.17}$$

As with the ARIMA model this is much better at short term predictions, particularly as using Fourier series where K is small leads to a poor approximation to the full seasonality.

Exponential Smoothing. An alternative method is exponential smoothing. Exponential smoothing assumes that the data collected further into the past has less influence than that collected sooner (Ord and Fildes, 2013). The weights for each value in the past are exponentially decreasing, hence the name.

Let L_t and Y_t be the current mean and observation at time t and α the smoothing constant. In exponential smoothing the contributions of previous observations to a forecast decreases as they become older. The current mean for the next time step is found by:

$$L_t = \alpha Y_{t-1} + (1 - \alpha) L_{t-1}. \tag{2.2.18}$$

Any prediction that is past t will be L_t. This is because the model has no evidence of it going up or down in the future and therefore predicts it will remain the same. Thus the forecast for time t is

$$\hat{y}_t = L_t. \tag{2.2.19}$$

Summary. In this section we studied the models that can used in VRP problems that require point forecasts. The models are then summarised by a variety of different characteristics. None of the models have exactly the right characteristics of our vehicle routing problem. However several can be adapted to suit our requirements, by applying the model to a longer time horizon or adjusting the problem context. The models which we can easily adapt are the ARIMA model, ARIMA with Fourier terms model, Seasonal ARIMA model, exponential smoothing and previous observed value model and these are analysed using data in Section 2.4.

2.2.3 Travel time distributions

The previous section provides a single estimate for one point in time, however travel time data has clear variations within it. A distribution of travel time allows some of this uncertainty to be captured. These variations can be due to vehicles travelling at slightly different speeds, different weather conditions or accidents. This should lead to more accurate predictions.

A distribution assigns probabilities to different travel times occurring. The distribution can either be continuous or discrete. Han et al. (2014) use a discrete set of travel times for each link, one for each scenario. All VRPs which allow travel time uncertainty have a travel time distribution. Robust optimisation uses one via a limited number of scenarios across the whole network. This can be useful if busy days are more likely if the previous day has been busy.

The distribution can either be the same for the entire time period or be time dependent. If it varies it is most likely to do so over discrete time periods. Otherwise the distributions will be inaccurate due to a lack of data.

The VRP calculates the best route, which requires the combination of travel times for individual links. With a point forecast this can be done by addition, but combining distributions together can be impossible. For most distributions to combine just two links is very difficult because for every possible travel time on link one the travel time for link two is another distribution, which may be dependent upon the travel time for link one. We also need new criteria to distinguish between the different route distributions and choose the best one. Hence the VRP problem is more complex if the uncertainty is included.

Distribution	**Interval**	**Horizon**	**Reference**	**Time dynamic**	**Size**
Normal	time windows	NA	Jie (2010)	None	8c
Normal	time windows	NA	Li et al. (2010)	None	>8c
Normal	NA	NA	Sumalee et al. (2006)	None	5l
Log-Normal	4 time windows	NA	Westgate et al. (2014)	Rush hour	4/68272l
Truncated	NA	NA	Miranda and Conceicao (2016)	None	3l
Gamma	NA	NA	Ta et al. (2013)	None	Multiple
Alpha-Discrete	NA	NA	Zhang et al. (2013)	None	5c
Erlang	NA	NA	Russell and Urban (2008)	None	100c
Phase-type	NA	NA	Gómez et al. (2016)	None	1l
Triangular	NA	NA	Ando and Taniguchi (2006)	None	22c/218l
Log-Normal	NA	NA	Gajewski and Rilett (2004)	None	2/3l
Truncated	NA	NA	Cao et al. (2014)	None	3l
Log linear	30min	2h	Huang and Barth (2008)	tod	2l
PI	5min	5 min	Khosravi et al. (2011)	tod	2l
GARCH	5min	5min/5h	Zhang et al. (2015)	tod	5l
Burr XII	NA	NA	Guessous et al. (2014)	density	2l

Table 2.2.4: Summary table time related properties of point forecasting methods. tod is time of day, c is customers and l is links.

Method	Reference	Lit	Uses	Data	Time frame
Normal	Jie (2010)	VRP	stochastic VRP with time windows	N	S/L
Normal	Li et al. (2010)	VRP	stochastic VRP with time windows	N	S/L
Normal	Sumalee et al. (2006)	VRP	network design	Y	L
Log-Normal	Westgate et al. (2014)	VRP	shortest path	N	S
Truncated	Miranda and Conceicao (2016)	VRP	stochastic VRP with time windows	N	S
Gamma	Ta et al. (2013)	VRP	stochastic VRP with time windows	N	S
Alpha-Discrete	Zhang et al. (2013)	VRP	stochastic VRP with time windows	N	S
Erlang	Russell and Urban (2008)	VRP	stochastic VRP with time windows	N	S
Phase-type	Gómez et al. (2016)	VRP	stochastic VRP	Y/N	S
Triangular	Ando and Taniguchi (2006)	VRP	stochastic VRP with time windows	Y	S
Log-Normal	Gajewski and Rilett (2004)	T-T	estimating correlation	Y	S
Truncated	Cao et al. (2014)	T-T	4 link with signals	Y	S
Log linear	Huang and Barth (2008)	T-T	2 link model	Y	S
PI	Khosravi et al. (2011)	T-T	prediction interval evalutation	Y	RT
GARCH	Zhang et al. (2015)	T-T	seasonal patterns	Y	L/RT
Burr XII	Guessous et al. (2014)	T-T	different traffic levels	Y	S

Table 2.2.5: Summary table of point forecasting methods for other key features. S is short term, L long term and T-T is short for travel time.

If the travel time distributions are additive the travel time can be calculated for the entire route. Gamma, normal, Weibull, Burr XII, log-normal distributions have been used as well as queuing models (Gómez et al., 2016; Guessous et al., 2014). Which distribution is most relevant, and its parameters, will depend upon the characteristics of the road section, as well as the method of collection. Aron et al. (2014) compare six distributions for one section of road, at different times, to conclude that the distribution depends upon the traffic conditions.

In this section, due to the importance of these combinations, we use $\hat{y}_a(t)$ as the prediction of link a for time t and $\hat{y}_r(t)$ as the prediction of travel time for route r at time t. Calculating each possible route a priori is one way to avoid adding distributions together in the VRP model but this is impractical for any networks that are used in a VRP.

Tables 2.2.4 and 2.2.5 provides a summary of the different distributions that have been used to model travel time. The characteristics are the same as for the point forecasts, but because the majority of distributions are from the VRP literature then a measure of the network size has been included.

The *size* of the network has been included because calculating distributions for one link is usually more computationally intensive than to estimate point forecasts. Estimates must be made for each link and hence, with the size of network required for the petrol routing problem, this can easily become computationally intractable. Our proposed network has 52 links and up to 18 customers. The vast majority of the methods are illustrated with small networks.

Most methods are used to provide a distribution for the road for all time. This means that they don't use time intervals or have a prediction horizon. Of the three methods that have a prediction horizon the horizon is fairly short - at most 5 hours. Instead of using a fixed size of time intervals they split the time period up into time windows which then have different travel time distributions.

As a result of using distributions for short term predictions the majority of methods have no time dynamic as the travel time variation is considered to be part of the distribution, rather than a daily or weekly structure. Many of those that do are focused on modelling the distribution rather than travel time prediction. Westgate et al. (2014) allows for a change in distribution around rush hour, but the three methods have a time of day dynamic that have explicit prediction horizons.

Ten of the models are from the VRP literature, as opposed to six in the travel-time literature. Regardless of the literature the majority of models are either short term or real time methods.

All of the travel time methods use data to calculate their distributions. Sumalee et al. (2006), Gómez et al. (2016) and Ando and Taniguchi (2006) all use a dataset to generate the travel time which is then used in a VRP. Most of these methods are very simple.

The vast majority of the papers are used in stochastic VRP with time windows. The petrol routing problem may contain time windows, but the roads would require more than a single distribution over the day.

In conclusion none of the models exactly fits our requirements for the petrol routing problem. We will now look at the methods in more detail to see if they can be adapted to better fit the petrol routing problem.

Prediction intervals. A simple extension from using point forecasts is to use prediction intervals around the point forecast. Khosravi et al. (2011) look at using a genetic algorithm which is computationally intensive to generate better prediction intervals. Prediction intervals are calculated in real time given the previous five minutes of data. These are used to provide estimates for a sections of a single route or link such as for bus routes. Using this when combining links together would be difficult.

GARCH. Zhang et al. (2015) use component-GARCH (generalised autoregressive conditional heteroskedasticity) models for the variance and an ARIMA model for the mean. This assumes that the variance changes throughout time and a component-

GARCH model estimates this change, while allowing either trend or seasonal behaviour. A prediction interval for each forecast is produced and the paper creates box-plots for each of the point forecasts which can be plotted to enable analysis. This is the long term part of the forecast for a single link. The same link is also analysed when producing one step ahead forecasts (five minutes). These real-time estimates are produced across the day and when the travel time is more volatile the component-GARCH model performs better. Combining links would also be problematic.

Normal Distribution. If the travel time function is effectively constant with small variations from the mean a normal distribution has been suggested. The normal distribution has several useful properties including scaling easily when combining links together. The travel time for link a is normally distributed with mean μ_a and standard error σ_a such that:

$$\hat{y}_a(t) \sim N(\mu_a, \sigma_a). \tag{2.2.20}$$

Jie (2010) and Li et al. (2010) combine links by adding the mean travel time for each link to create the overall route mean and taking the square route of the sum of the variances. Both find the best route for vehicles when considering customer time window limitations with stochastic travel times. The travel time distributions remain the same across time so the model only considers the variation from the mean, without rush hour effects.

Let r be the route we are considering and E_r the set of links that are in route r. Then

$$\hat{y}_r(t) \sim N\left(\sum_{a\in E_r} \mu_a, \sqrt{\sum_{a\in E_r} \sigma_a^2}\right). \tag{2.2.21}$$

Sumalee et al. (2006) finds the total travel time by combining the distributions for each link which is assumed to have a normal distribution. They use this to decide where it is best to improve the network, such as by decreasing journey times. This is a very different application to routing vehicles through a network but both require accurate estimates of the travel time. The total travel time is therefore a Multivariate normal distribution. This can be further extended to a network level by including the covariances between roads in the variation matrix Σ.

Let $\hat{\mathbf{y}}(\mathbf{t})$ be the vector of travel times at time t for the network. Then

$$\hat{\mathbf{y}}(\mathbf{t}) \sim MVN(\mu, \Sigma). \tag{2.2.22}$$

It is relatively easy to add or remove links from the Normal distribution model. However the normal distribution is symmetric and travel times have a lower end point. With relatively little data the normal distribution is likely to be highly inaccurate for one link and this will only expand as more links are added.

Log-normal. Gajewski and Rilett (2004) adapt the Normal distribution by using a natural log transform on the travel time to remove some of the instability of the travel time variance. This is in the context of estimating correlation in the travel times which is a network level context, but the log transforms are performed upon each link individually.

Westgate et al. (2014) also use a log-normal approach, which includes the calculating travel times from the distance and speed, in their modelling of ambulance travel data. This approach models whole trips rather than links but still requires the travel time at link level. The travel times are thus

$$\hat{\mathbf{y}}(\mathbf{t}) \sim MVN(\ln(\mu), \Sigma). \tag{2.2.23}$$

Log linear. Huang and Barth (2008) look at short term prediction for travel on motorways, up to 2 hours ahead. This uses a log linear model, with both historical and real-time information. The method is used mainly for a single route but can be expanded. The travel time for the entire route is calculated by finding the historical mean for each link individually and then using the real-time information. Each link relies upon the travel time for all previous links being known. Thus the final link gives the total travel time for the route. This method means that the stochasticity is lost and that it only works on entire routes rather than links, except where the travel time across a link is calculated for all possible arrival times. As one of the main benefits to using distributions is to include stochastic routing, a method which loses stochasticisty across the network is not appropriate.

Truncated distribution. The previous distributions fail to account for several of the features observed in travel time data. There is a specific travel time below which it is impossible to physically reach, and there is also the travel time associated with the speed limit of the road. This leads to a very skewed distribution. A truncated normal is a normal distribution which is bounded below (and/or above) so it can't drop below a certain level (Miranda and Conceicao, 2016). The authors created an algorithm to compute successive distributions, eventually leading to the generation of the entire route. They use the expected travel time in their objective function to find the best route for vehicles, in the context of customers having hard time windows and stochastic service times.

Cao et al. (2014) use truncated distributions to set minimum and maximum travel times and to eliminate some of the bias when calculating the mean and variance. As is noted in Section 2.3.4, travel time data has large outliers and this distorts the mean and variance. Using the median and inter-quartile range (IQR) is a possible improvement. This method is used in the context of three short connected sections of road which contains traffic lights. Hence there is sometimes waiting. Our links are much longer and have far fewer traffic lights to road length, so the impact of traffic lights is lessened.

Gamma distribution. Ta et al. (2013) use the Gamma distribution for the travel time when using soft time windows in a vehicle routing problem. Gamma distributions combine such that the arrival times at each customer are also Gamma distributed. No data is used to fix the parameters and they focus upon the vehicle routing problem solution rather than the travel times.

Alpha-discrete distribution. Zhang et al. (2013) use both normal and log-normal travel times along the links. The model was created to look at how the uncertainty in travel times can cause logistic companies to miss time windows, and how to model this to assess the cost to the logistics company. If a vehicle arrives too early it must wait until the start of the time window to begin its delivery, which means that the overall distribution of a route isn't normal.

An alpha-discrete distribution is used for the arrival time distribution at a customer, given the departure time from the previous customer and the travel time distribution between them. The arrival distribution back at the depot will give the route travel time distribution, which depends upon the arrival and departure time distributions for the previous customers. Each alpha-discrete arrival distribution is specified using a set of L discrete values, hence the discrete name. The L values are chosen such that their cumulative probabilities are evenly spaced between 0 and 1. Thus while the individual links are modelled using a normal/log-normal distribution the inclusion into the cost function is via the alpha-discrete representation. The travel times of the individual links have no time of day dependence, but the overall route does due to the time windows.

Erlang distribution. Russell and Urban (2008) fit an Erlang distribution to historical data. The authors use this to look at soft time windows inside a vehicle routing problem. An Erlang distribution was chosen because a previous study into truck travel times found a shifted-gamma distribution was most suitable. Erlang distributions have the additional requirement that the shape parameter must be integer.

Links are joined by combining the Gamma distributions from the start of the link. The α and δ parameters are assumed to be constant over the entire network and are scaled proportional to the distance such that the distribution depends upon the distance along the network between any two points. When considering route r, with the distance along link a being d_a, we therefore have:

$$\hat{y}_r \sim \text{Gamma}\left(\sum_{a\in E_r} d_a\alpha, \sum_{a\in E_r} d_a\delta\right). \qquad (2.2.24)$$

Phase-type distribution. Distributions are considered together in families that have similar properties. Phase-type or PH distributions are distributions that arise from Poisson processes, including exponential and Erlang distributions. Gómez et al. (2016) justify using a PH distribution to model links because it permits a minimum travel time and doesn't require a closed form probability distribution which is unknown. Data for one link is used as justification. However the vehicle routing problem uses simulated distributions for each link.

PH distributions can be derived from the minimum, maximum, mean and standard deviation of a given dataset. They use the distributions in the context of finding the best route between customers, whereby PH distributions are found for every link of the route and then combined to make a PH distribution for the whole route.

Triangular distribution. Ando and Taniguchi (2006) look at travel time reliability in VRP problems with time windows, using a triangular distribution fitted by the minimum, maximum and average values. Distributions are calculated for six road types using limited data collected from vehicles doing delivery routes. The probability distribution forms a triangle when plotted, hence it's name. The edges of the triangle are at the minimum and maximum travel time and the point is at the average time, with probability $\frac{2}{t_{max}-t_{min}}$. To combine the links together into routes they use simulation to generate the link values, which add to make the entire link travel time.

Burr XII distribution. Guessous et al. (2014) use a Burr XII distribution which is generated from average speeds, flow and density. This is often used for modelling lifetime data which is skewed and has limits like travel time. The data are the single road section used in Aron et al. (2014) so the transferability to a whole network is unclear. Combining multiple links requires combining Burr XII distributions which is complicated, especially given each links distribution depends upon the arrival time which is dependent upon the previous link.

Distribution summary. There are a wide range of distributions that have been considered for forecasting travel time. The majority of these have no consideration of how traffic may vary over the day and use the same distribution for the entire period of the model. Those that do use the time of day as a consideration have fairly short prediction horizons.

The vast majority of the methods use travel time data as a means to form a single distribution. They therefore don't consider the need to recalculate every day as we would most likely need to for the petrol routing problem.

A lot of the distributions are linked in some way. This is because travel time doesn't follow any of them precisely and they are therefore approximations. Normal distributions have very nice properties and are easy to fit if they are appropriate but as shown in Section 2.3.4 the travel time values have a lower end point and aren't symmetric.

Finding and fitting a good distribution which is computationally difficult due to the large number of links in the petrol routing problem. Each link needs a forecast distribution for each of the time periods that the day has been split into. Hence we choose to focus upon point forecasts in this Chapter. If point forecasts prove to be an inadequate estimation when used in the VRP they can be made into a distribution by either considering prediction intervals around the point forecast, or a discrete distribution around the point forecast.

2.2.4 Section summary

In this section we have reviewed a wide range of travel time prediction methods. The majority of these methods provide forecasts in the short term, at a far shorter length than the day in advance the petrol routing problem requires.

There are two main types of forecast which will depend upon the setup of the vehicle routing problem. Ideally we would be able to use travel time distributions but these are much more complicated to model, both analytically and within the VRP and there is no clear distribution which is the best for travel time data. We discuss this further in Chapter 5 because the long delays we model in Chapter 3 occur stochastically, hence distributions are more appropriate.

We will focus upon point forecasts for the remainder of this Chapter, and Chapter 4. This is to enable us to solve a simpler petrol routing problem. It is much more difficult to generate long term forecast distributions because they require a lot more data than a point forecast for predictions at the same time interval. This is why the majority of distributional forecasts use the same distribution for the entire prediction period.

As a result there is a clear need to identify a method that can accurately use the previous travel time data to generate the forecast for the entire day ahead. This can then be used for each link in the vehicle routing problem, to ensure that petrol deliveries can be planned for the next day.

We wish to analyse and compare models using travel time data for links that could form part of a petrol routing problem. These models need to perform well across an entire day, with different travel times for different times of the day. In addition the forecasts would be required every day, so the model needs to cope well with changing the day of the week and any longer term seasonal changes such as between summer and winter.

2.3 Data

This chapter provides an evaluation of travel time models such that a VRP practitioner can follow our approach to select the best model. They will therefore have a predefined vehicle routing problem with a set network. A brief discussion on how our test network has been selected was conducted in Section 1.1.7 should this not be the case.

2.3.1 Dataset characteristics

The practitioner needs travel time information for each link of the network in order to run the VRP and find the optimal route. We will first look the ideal characteristics of a dataset, before discussing possible datasets and their availability. The characteristics of a dataset partly depend upon the method used to collect the travel time data. One main consideration is that some links may only have one type of data available.

To generate the travel time values from the dataset a forecasting model is used. This produces the input into the VRP for each link and should be in the form of a travel time function. The travel time data for each link is generated by combining multiple sections of the road network as defined in Section 1.1.7. The format of the function is limited by the dataset and the VRP requirements.

One key consideration from the VRP requirements is how far in advance the predictions are needed. This is determined by the context of the VRP and if the VRP is time dependent.

The properties of a dataset that affect the quality of the predictions are summarised below. This includes the preferred characteristics for each property, with the justification including how this affects models.

Time Period. The travel time predictions will be for the period relevant to the VRP, which varies depending upon the application. Our VRP problem requires the prediction of every link for the whole next day, such that petrol deliveries can be planned. The petrol routing problem of Section 1.1.3 is time dependent and hence the data and forecasting model should also be time dependent.

The dataset that is used to forecast the results for the VRP must be relevant to the travel time period that is required for the VRP. If there is a weekly seasonal pattern and the next day is Tuesday, then a dataset with only Saturdays in will miss the pattern for Tuesday. Other information may be contained in the period directly preceding the forecast period and older datasets are usually less relevant to the current conditions.

Dataset length. Datasets vary in the length of their history. Some go back for years, while others contain only a day of observations. If there is only one day's worth of observations the methods suggested here are unlikely to provide good forecasts, unless the problem is real time as checking the forecasts will be difficult. The dataset must be large enough to detect any weekly, daily or other patterns. To detect a seasonal pattern the dataset must be at least twice as long as the season, preferably more.

The further into the future the predictions are the longer the dataset needs to be. We wish to predict a day into the future and as identify in Section 2.3.4 our data has a weekly seasonality. We will therefore need at least two weeks of data, preferably at least four.

Even intervals. The travel time along a link changes continuously but datasets only record the travel time discretely. Depending upon the method of collection data may be recorded at set intervals or infrequently. The second case occurs when the a fleet of vehicles are used in collection. Most models prefer regular data recording, as it is easier to predict what would have happened if a vehicle had been slightly earlier or slightly later.

Frequency. The frequency is how many observations there are in a set period of time. A forecasting model can only output at most the frequency of the observations it is input. Thus the model requires at least one observation in every interval of the VRP. Regular collection intervals vary from data every minute to every day.

In a network with many links, calculating the travel time for a high frequency takes a lot of computational time. The more intervals in a travel time function, the harder it is to ensure that the FIFO property is obeyed. The longer the interval the lower the frequency required.

Our petrol routing problem requires one day of forecasts which is 24 intervals of hourly data, 96 intervals at 15-minute data or 288 intervals of 5-minute data. 288 intervals is too many when forecasting multiple links.

Missing Data. Collection of travel time data relies upon sensors which fail, leading to gaps within the time series. Missing data falls into two main categories - a sensor failing for a long period of time or failing intermittently for short periods. Many forecasting models struggle to cope with missing data, especially when a lot of data is missing. This is a similar issue to requiring even intervals, as the spaces in-between the non-standard records could be considered as missing data.

2.3.2 Collection methods

The coverage of the network geographically and in time depends upon the way the dataset has been collected. In the UK the major roads have a lot more data available as there is a lot more demand and the travel time is likely to change over time. A small country lane may only have 10 vehicles traverse it each day and most have local knowledge from past use. Collecting detailed information on the very large number of smaller roads is impractical. However if a logistics company needs to deliver along the road some estimate is required. The practitioner may therefore need more than one dataset to cover the entire network. If datasets overlap, the one with the highest frequency of collection and longest history should be chosen.

There are a wide variety of different datasets available, however these are location specific. Some datasets are publicly available, whereas others must be purchased. Since our focus is on scheduling the transportation petrol for the next day we require data that gives an accurate picture of how the travel time varies throughout the day. The majority of datasets with the level of detail required to predict this level of variation come from government sources. They maintain and monitor the roads and it would be very difficult for anyone else to install and maintain a network of automatic numberplate recognition (ANPR) cameras across an entire network.

We exclude less detailed datasets in this analysis which restricts the possible links in the vehicle routing problem network. The removed links are the minor roads and when transporting hazardous materials staying on major roads for as long as possible is preferable as they are better suited to large vehicles. Due to the lack of data for these links, the majority of prediction methods detailed in Section 2.2 are either inappropriate or give predictions on a very large scale.

A report on a variety of considerations for collecting travel time data is provided by Turner et al. (1998). An overview of the main collection techniques, with their characteristics are provided.

A summary of the characteristics for each method presented in Table 2.3.1. This links back to Section 2.3.1 as it summarises the frequency and the even intervals. Of the other three properties, missing data occurs in all of the collection methods and the time period and dataset length depend upon the dataset not the collection method. The other two characteristics in the table depend upon the method of collection, rather than the dataset. Pre-processing is the level of processing on the data that is required before it is in the format to use in the models and collection is who collects and stores the data.

ANPR cameras. One common method is using two ANPR at the start and end of a section of road. The travel time of a vehicle is calculated by taking the time between the two photos. Unfortunately not all vehicles will be captured by both

	Pre processing	Even intervals	Frequency	Collection
ANPR camera	L	Y	H	Road managment Government
Fleet collection	L	N	L	Companies
Floating car	H	N but Av	L/H	Various
Mobile phones	H	N but Av	H	Phone network

Table 2.3.1: Summary of data collection characteristics. L and H are low and high, Y and N are yes and no and Av is average. The characteristics are the level of preprocessing required, the frequency of collection, who collects the data and if the data is collected at even intervals.

cameras and hence if the information is taken too regularly there may be times with no data. ANPR cameras are usually on all the time, and due to the large volume of cars the travel time calculations are recorded as averages at set intervals. The travel times are therefore for exiting the link not entering the link as the travel time function requires. Cameras failing, or poor weather leads to missing data.

Vehicle collection. Another method of travel time collection is to use the travel times of the vehicles of the VRP company. This can lead to very sparse datasets as there aren't many vehicles and some of the potential links may not be traversed at all. If we consider the problem across multiple time steps this becomes even more problematic. Thus the collection intervals aren't even.

Floating car data. Floating car data is collected by a fleet of vehicles travelling within the network (Rahmani et al., 2015). The vehicles are fitted with GPS and other sensors. The sensor information is recorded at fixed time intervals, then sent to the recording station for real time analysis. This data usually requires a lot of preprocessing to extract travel time information, the locations are unlikely to conform to the VRP links, and cars may even travel over multiple links. *TomTom* collect data from their sat nav users and this is available to purchase (TomTom, 2015). As there are many users, and this has been collected since 2006 then the dataset should have sufficient depth. TomTom clean the data before use. As more people are using internet connected phone apps with GPS enabled rather than a sat nav then there is potentially much more frequent floating car data that is being collected.

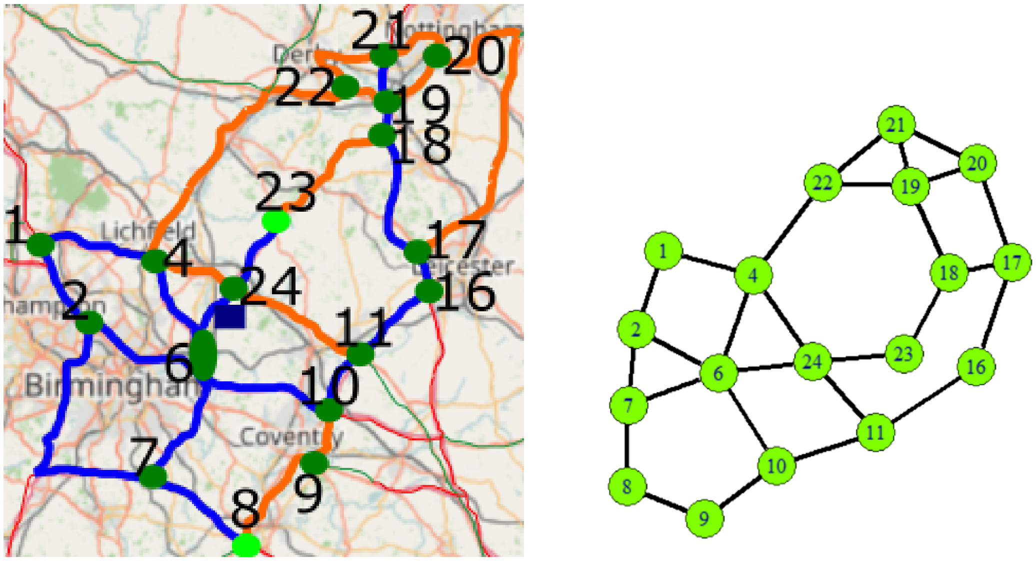
Mobile phone network. Using the mobile phone network is similar to floating car data except that the data collection occurs only under certain conditions. A phone must be either making a call or have a data connection. Locations are determined by triangulation. This means that there is less data available, as few drivers use their phones while driving, and it is unlikely that phones would be in use both at the start and end of the link. Janecek et al. (2015) suggest an alternative method of using idle phone signals. These signals are generated when phones check for new messages and hence are much more frequent. Both types of mobile phone data have data processing issues, and must eliminate phones that are being used by people walking or cycling if the traffic is stationary. The data also relies upon the phone companies so would have to be purchased.

2.3.3 Our dataset

We have introduced the data and the test network we will use in this chapter in Section 1.1.7. This travel time data is published online, `http://tris.highwaysengland.co.uk/detail/journeytimedata` (Highways England, 2016).

We initially presented the road network on the road map for ease of understanding which roads where which but the network is independent of the map. Figure 2.3.1 shows how the network can be represented as a grid layout rather than showing the map overlay which includes features that aren't retained within the network.

Within this chapter we focus upon predicting the travel time of a subset of these links. Initially we look at 7.8, on the M40 between the M42 (junction 17) and junction 15. We then look at a small selection of links to confirm that the models hold across the network. These links form a mini network which we study in further detail in Chapter 4. These links are centred around link 10.11 which is the M69 from the



(a) The road network graph, on top of the geographical road network.

(b) Network as a grid.

Figure 2.3.1: Network on map and independently laid out. There are 2 directional links between all nodes with a line between. Map background OpenStreetMap contributors (2017)

M6 to junction 1 and has a number of surrounding links in the test network. The surrounding links are 11.10 (10.11 in the opposite direction), 11.16 (M69, junction 1 to the M1), 6.10 (M6, junction 3a to 2), 11.24 (A5, from the M69 to the M42) and 9.10 (A46 from the A45 to the M69).

2.3.4 Exploratory data analysis

The first step of the data analysis is to visually inspect the data in order to identify any key elements, such as patterns and outliers, that we must consider when forecasting the travel time. We then examine how we can model these trends by considering the median and the interquartile range. We started this in Section 1.1.7 and we now continue in greater detail.

The link chosen for this chapter is link 7.8 which is the M40 between the M42 (Junction 17) and Junction 15. Its network location can be seen in Figure 1.1.4. The analysis is carried out for the period of time between 1 January 2016 and 25 Mar

2016, due to there being no data recorded from the 26th March to the end of the month.

We plot the travel time values to visually inspect the data. Identifying patterns is easiest when the data is split into repetitions of the pattern which are plotted on the same axis.

Weekly pattern. Travel time depends upon a variety of different factors. These include the weather, the amount of light and the traffic volume. Vehicles make journeys for many reasons - travel to work and school, day trips and business visits. Some of these are dictated in part by external factors, such as the working week. Hence a weekly pattern may be present in the data.

To effectively analyse the travel time patterns the dataset needs to be of a sufficient size. Daily data requires less data than weekly data which requires less data than yearly data. There should be at least 4 points for the largest scale of data being considered to check that the data at that scale follows a similar pattern. Hence for a weekly pattern we need at least four weeks worth of data. More data makes it more likely that changes will be observed and that any patterns can be justified.

The picture in Figure 2.3.2 shows that there are some very severe delays, and less severe delays. Each line represents one weeks worth of data, plotted from Monday to Sunday. The most severe delay is about 5800 seconds, whereas the link normally takes about 450-650 seconds to traverse. At this scale it is impossible to discern any patterns in the data, with the exception that the extreme delays appear to be more common in the morning. Figure 2.3.3 focuses on the region without extreme delays to show the normal weekly behaviour. The plot shows an overall structure of busier periods and free flowing traffic that are followed from week to week. There is also a large amount of variability from one time period to the next. The data is 15 minute averages. Maintaining a constant speed is difficult as vehicles travel at slightly different speeds.

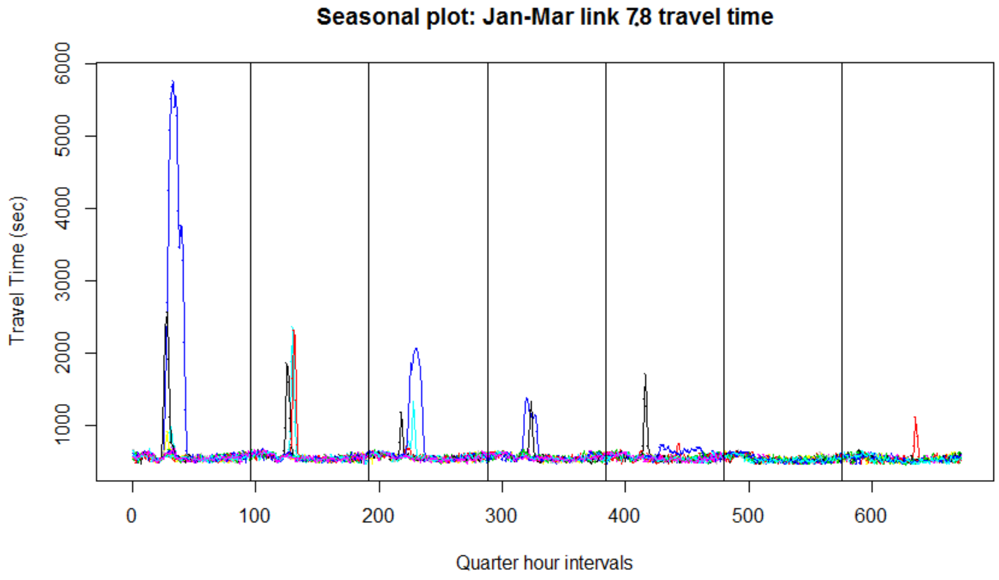


Figure 2.3.2: Plot of each weeks travel time from Monday to Sunday for link 7.8 showing how severe the delays can be. Each week is a different colour.

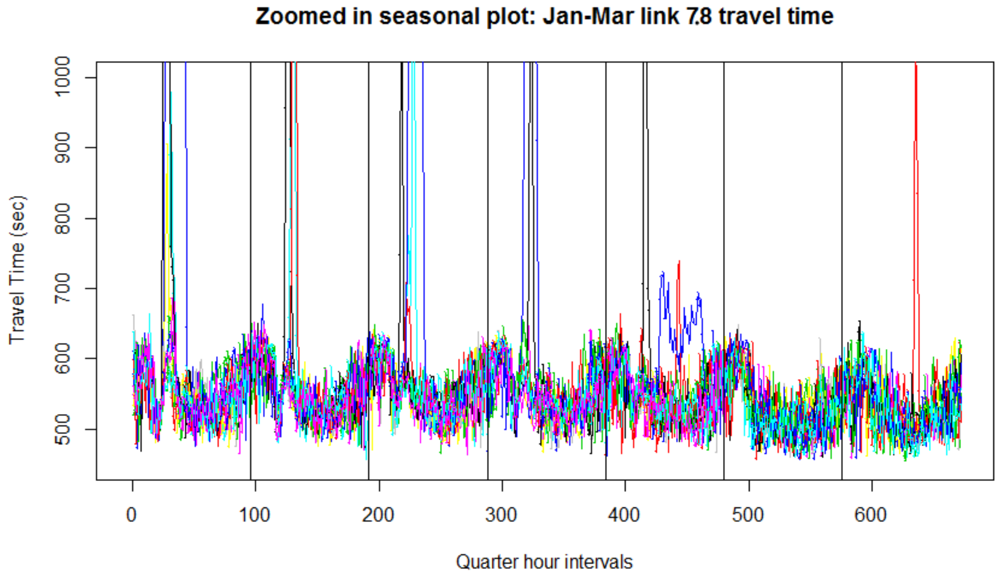


Figure 2.3.3: Plot of each weeks travel time from Monday to Sunday for link 7.8 zoomed in to highlight the weekly structure. Each week is a different colour.

Outliers. There are also several outliers which exist in the data, as highlighted by Figures 2.3.2 and 2.3.3. These correspond to traffic jams. There are different types of traffic jam. Rush hour jams due to large volumes of traffic may be predictable however traffic jams that are due to accidents are impossible to predict.

We therefore treat the data as a mixture of two distributions - the first is the everyday pattern and the second is the outliers which occur at random intervals. The outliers are identified during the standardization process in Section 2.3.5. This chapter focuses on modelling the everyday pattern while a study of the outliers occurs in Chapter 3 using extreme value theory.

Median. The pattern appears to be weekly, however grouping more days together may be possible if some days behave the same. To investigate this possibility we look at the median profiles of each of the seven week days. The median has been chosen due to the extreme outliers which influence the mean far more than the median.

The dataset needs to large enough to calculate the medians. Using only three points for the median isn't very informative. If there are additional non-weekly trends taking the median across the week fails to account for them. A seasonal effect such as traffic travelling slower in winter can be included if the dataset length is small enough. Then the median will lie within the winter travel time rather than in-between the winter and the summer travel times. Figure 2.3.4 shows that the different days should be considered separately.

Interquartile range. Another measure of difference is the variability of the data. To check this we compare the inter-quartile range or IQR, as the variance is also more effected by the outliers. The daily IQR plots can be seen in Figure 2.3.5. There is much less variation in the IQR between each day than for the median.

Summary. The exploratory data analysis has indicated that there is a weekly pattern for link 7.8 with some severe delays. We look at these delays in Chapter 3. In addition the median and IQR are better summaries of the data than the mean and variance. The median and IQR have slightly different values across the day for the different days of the week but the 24-hour profile structure is fairly similar for each

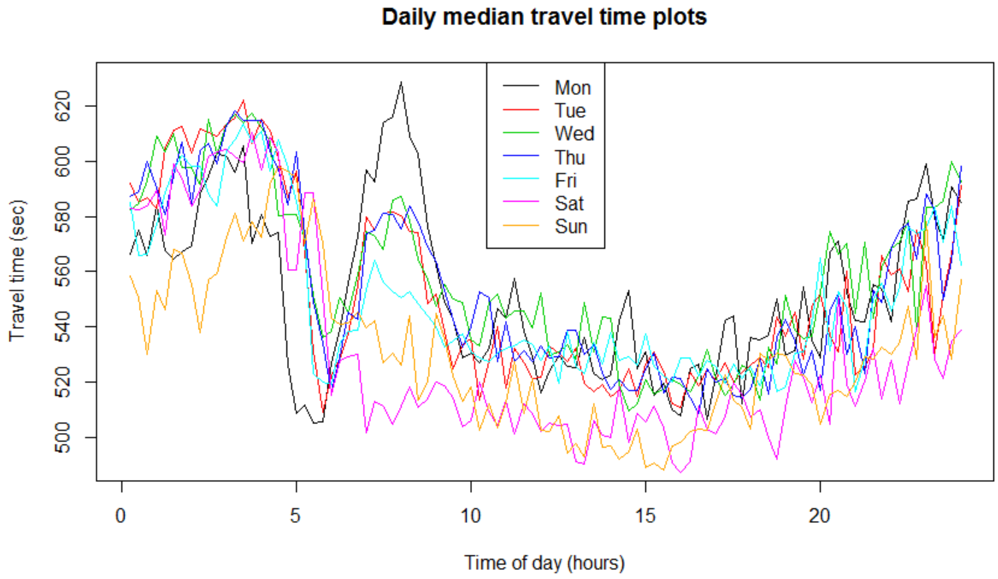


Figure 2.3.4: Plot of median travel times for each day of the week for link 7.8.

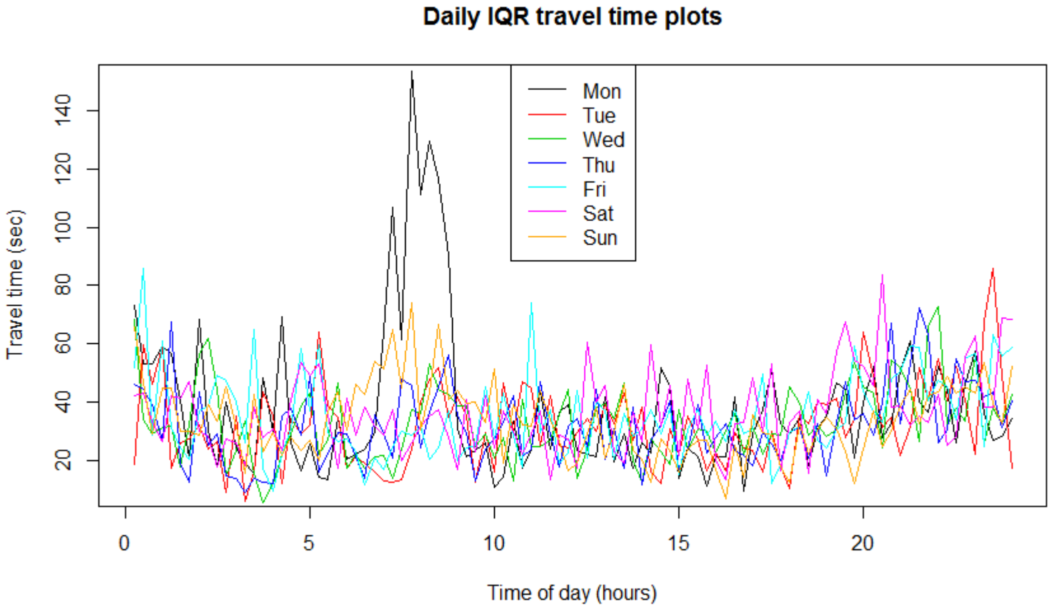


Figure 2.3.5: Plot of the IQR for each day of the week for link 7.8.

day. There is more variation in the medians than the IQRs.

Both plots show that the days of the week behave differently, hence the need to consider an entire week as a season. The differences in the median are more pronounced than in the IQR. The resulting forecasts should therefore alter depending on the day of the week and time of day. The plots show clear time dependence which is incompatible with many of the forecasting methods. One method to remove this time dependence is to standardize the data to generate initial forecasts to which the trends are reintroduced to give the final forecasts. This method is used for the majority of methods tested in the remainder of this chapter, as well as all of Chapter 4 and in Chapter 3 to identify the extreme delays. We now look at the process of standardization.

2.3.5 Standardization

As we have observed a weekly pattern we wish to consider the expected values and variation on a weekly scale. There are thus $7 \times 96 = 672$ different time points in a week that we need values for, one for every 15 minute period of the week. As mentioned in Section 1.1.7 many methods require the removal of this weekly pattern which we will do using standardization.

The normal way to standardize data is to use the mean, μ and the standard deviation, σ. Where the data is seasonal these values will correspond to where in the season a time point is. The weekly time index runs from 1 to 672, such that 1 is Monday morning at 00:15 and thus there are 672 individual means and standard deviations.

Let x_t be the travel time value observed at time t. Then we can calulate $\tilde{x}_t$, the standardized value of x_t, using equation (2.3.1). For data with another pattern, the data should be standardized according to this pattern. Let $\tau(t)$ be the weekly (or otherwise) index for time t. The standardized value for time t is thus:

$$\tilde{x}_t = \frac{x_t - \mu_{\tau(t)}}{\sigma_{\tau(t)}}, \tag{2.3.1}$$

whereby μ_τ and σ_τ are the mean and standard deviation of the weekly time index $\tau(t)$.

However, as we identified in Section 2.3.4 the median and IQR are better measures of the data. We therefore use the median, $\tilde{\mu}_{\tau(t)}$ and the interquartile range or IQR, $\tilde{\sigma}_{\tau(t)}$ which leads to the standardized value of:

$$\tilde{x}_t = \frac{x_t - \tilde{\mu}_{\tau(t)}}{\tilde{\sigma}_{\tau(t)}}. \tag{2.3.2}$$

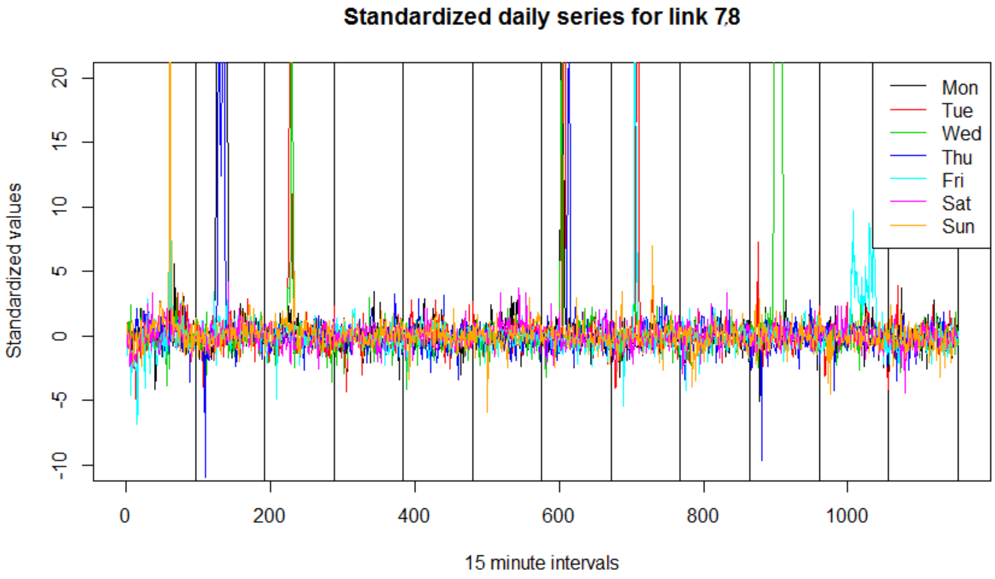


Figure 2.3.6: Plot of the standardized series for each day for link 7.8.

Figure 2.3.6 shows the standardized series for each day of the week across the dataset. The black bars split the series into individual days.

Rolling median and IQR. In some instances, where there isn't a lot of data for each index point, the standardized series can create very large outliers due to the lack of data. This occurs when most of the values are very close and one is slightly further away but within the tolerance for points at either side of the time point. One method to counter this is to use rolling medians and rolling IQRs.

To take a rolling median or IQR while using a weekly index, one considers any values of the time index itself and the ones $(k-1)/2$ either side. This gives the kth rolling median or IQR. The rolling median is:

$$\tilde{\sigma}_{\tau(t)} = \text{median}\left(x_j \in X \mid \tau(j) = \{\tau\left(t - \frac{k-1}{2}\right), \ldots, \tau\left(t + \frac{k-1}{2}\right)\right). \qquad (2.3.3)$$

The collection of points for the rolling IQR is the same as for the rolling median.

If a different pattern is observed the process is similar. The set of possible values include any with the pattern index value, as well as any pattern index values that occur within $(k-1)/2$ time steps in the time series. As pattern indexes could occur next to each other this may not include any more values than before.

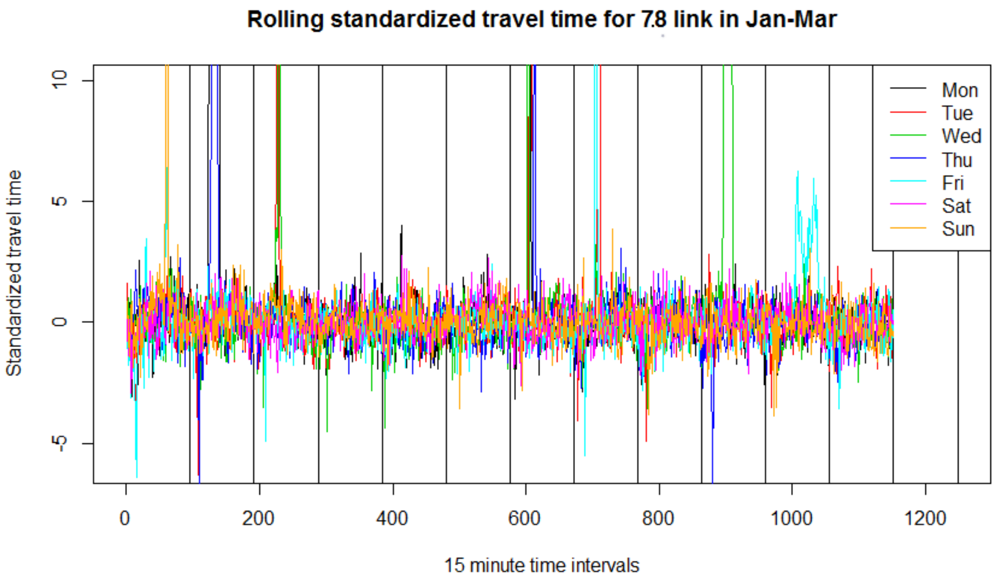


Figure 2.3.7: Plot of the rolling standardized series for each day for link 7.8.

The series in Figure 2.3.7 still have large outliers but we consider these to be true outliers in the data, rather than due to method of standardization. We investigate these outliers in Chapter 3.

Daylight saving differences The standardization is applied across the entire dataset. If the dataset is from a country with daylight saving hours and includes either March or October the clocks will change by an hour. In the UK, in March, the time zone changes from GMT to BST and thus 0100-0145 are non-existent for one day. When standardizing the values need to standardized using the corresponding median and IQR. In a similar way in October the time changes from BST back to GMT and thus results in an extra hour between 0100-0145. The standardization should use the values for the correct time step but for these two hours it is unclear which medians and IQRs should be used. In a similar way the hours directly after may be affected as traffic is still making journeys, although the clocks change at a time when roads are unlikely to be busy and hence the effect is minimal.

The issue is neglected in this chapter as the dataset used is January to March 2016 but the entire day of the 27th March when the clocks change is missing from the data.

Autocorrelation. One of the requirements for ARIMA models is that the data shouldn't have any trends. ARIMA models use the past values to predict the future. However only past terms that provide significant information about the prediction are included. One way to find which values are significant is via autocorrelation plots.

Autocorrelation finds the linear dependence between two points at different points in time. We can then find the autocorrelation between the values, x_t and x_s, observed at these two points assuming that the mean at these points in time is known. The autocorrelation $\gamma_x(t, s)$ between the two points at times t and s such that $t > s$ is:

$$\gamma_x(t, s) = cov(x_t, x_{t-s}) = E[(x_s - \mu_s)(x_t - \mu_t)]. \tag{2.3.4}$$

We are interested in the series as a whole rather than two points separately. The autocorrelation function or ACF looks at the series as a whole. Let t be the time series up to the current time and let s be the lag we are interested in. This is the linear predictability of a series at time t using only x_s. The autocorrelation of the series at a lag of s, given the current time is t, is:

$$\rho(s, t) = \frac{\gamma(s, t)}{(\gamma(s, s)\gamma(t, t))^{0.5}}. \tag{2.3.5}$$

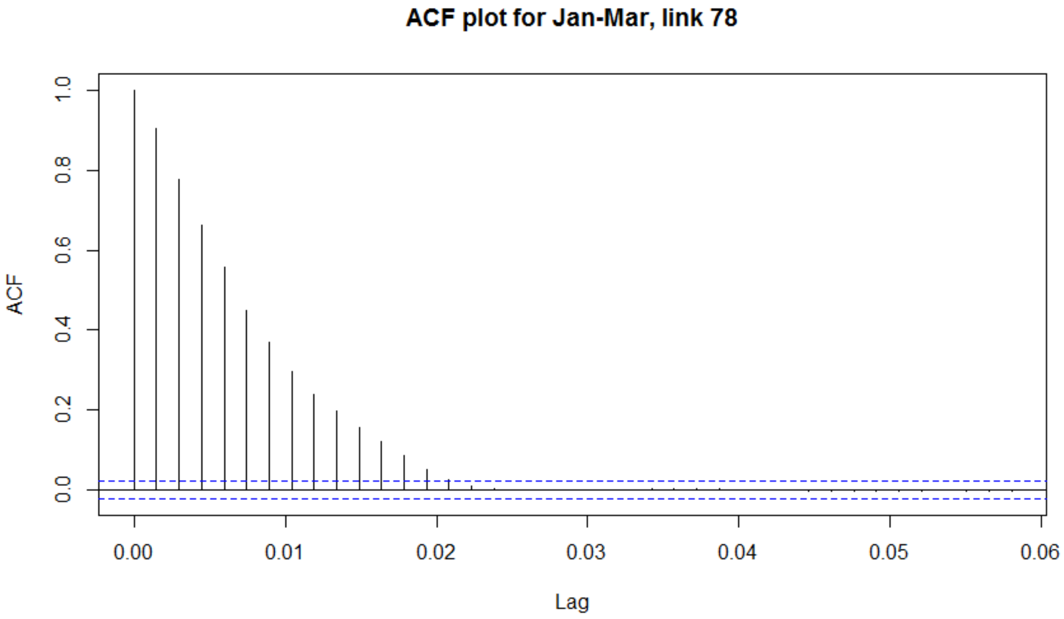


Figure 2.3.8: Plot of the autocorrelation function for link 7.8.

We wish to consider the autocorrelation for different values of s. This will give us an idea of how far into the past the current travel time is relevant. Figure 2.3.8 show the autocorrelation steadily decays away, with only the first thirteen lags being significant.

In addition we consider the partial autocorrelation. The autocorrelation function isn't independent over s. The partial autocorrelation takes into account the smaller lags when finding the linear dependence. The partial autocorrelation function for lag s at time t is:

$$PACF(s,t) = \frac{cov(x_s, x_t|x_{t-1}, \ldots, x_{t-s+1})}{(cov(x_s, x_s|x_{t-1}, \ldots, x_{t-s+1})cov(x_t, x_t|x_{t-1}, \ldots, x_{t-s+1}))^{0.5}}. \quad (2.3.6)$$

From the partial autocorrelation plot in Figure 2.3.9 it is clear that the first lag is significant, the second is slightly significant, the third and fourth aren't significant but the fifth is slightly.

The combination of the auto correlation and the partial autocorrelation plots can be used to identify which past values are of most use when predicting the future.

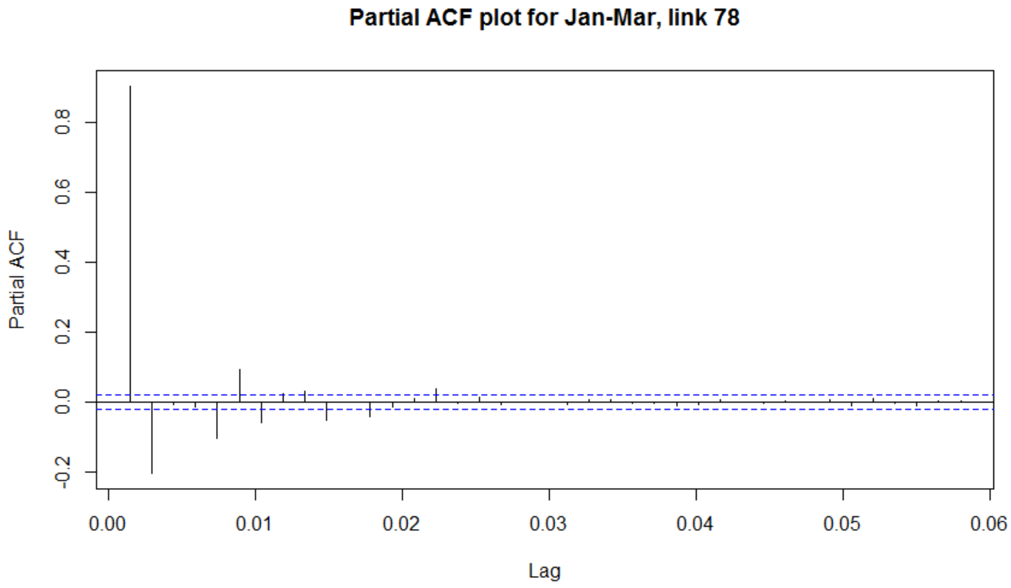


Figure 2.3.9: Plot of the partial autocorrelation function for link 78.

The maximum significant autocorrelation and partial autocorrelation terms indicate the maximum time lag. This equates to the P and Q in ARIMA models, or the maximum number of terms for the AR and MA parts. The plots suggest that we should be looking at ARIMA models with at most 5 terms.

Summary. Standardization of the series using rolling medians and IQRs over the week removes the weekly pattern leaving a series which is centred around zero, with most values between -2 and 2 on the standardized scale. There are some outliers from delays. The autocorrelation in the series is significant to only thirteen time lags and the partial autocorrelation is significant to only five lags. The majority of methods in this Chapter use standardized series which the trends are then reintroduced to give the series. All other Chapters in this thesis use standardized series.

2.4 Forecasting methods

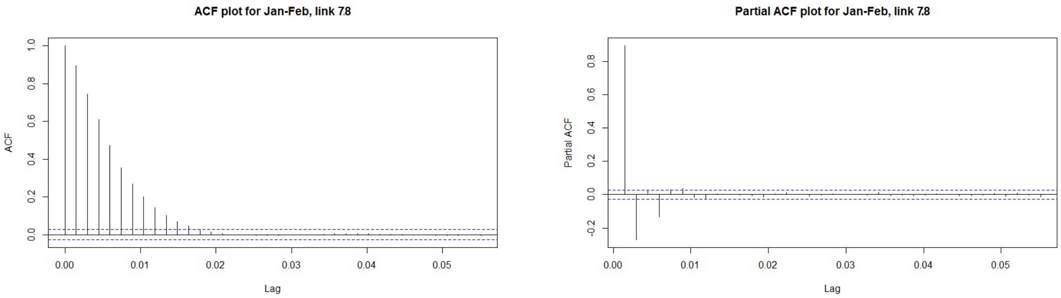
Having analysed the dataset we now move on to applying models to the dataset in order to select the best model to forecast travel time for a link in the network. We look at how they perform in respect to each other as well as the true values, before applying metrics to their performance in Section 2.5. All the models introduced in Section 2.2 that are well defined and practically implementable have been applied to the raw data. The format of this section is - a brief description of each model, then graphs of both the fitted model and the forecast for the 29th February (red) with the observed travel times (black) and a discussion of the results. The forecast plots are for an entire day ahead predicted from the beginning of the day whereas the fitted plot is retuned every time period, hence there are fewer lags than would be expected. Models are presented in order of simplest to more complicated.

The following results are for link 7.8 from Jan 1st 2016 to Feb 28th 2016, forecasting February 29th 2016. We selected the 29th February as it was the last day in the two months of complete data and hence has the largest amount of data to generate forecasts. The previous data analysis includes the month of March. However March includes some missing data so we first analyse the months without missing data.

We therefore recalculate the autocorrelation function and partial autocorrelation plots for the shorter time period which are presented in Figure 2.4.1. Only the first 12 time lags are significant in the January to February auto correlation function plot, as opposed to 13 when March is included. The first and second partial autocorrelation lags are still significant but it is the fourth not the fifth time lag which is significant without March.

We wish to select a model that performs well for fitting but also for forecasting the entire day ahead. In this section we will perform visual inspections of the model fit for the previous days, as well as the forecast for the 29th February. A more detailed comparison is given in Section 2.5. All results are given in terms of seconds, having unstandardized the model outputs.

ARIMA model. The first method we apply to the standardized data is the ARIMA model. This is the simplest model that can be applied from the ARIMA family and


ACF plot for Jan-Feb, link 7.8

Partial ACF plot for Jan-Feb, link 7.8

(a) Plot of the autocorrelation function for link 7.8.

(b) Plot of the partial autocorrelation function for link 7.8.

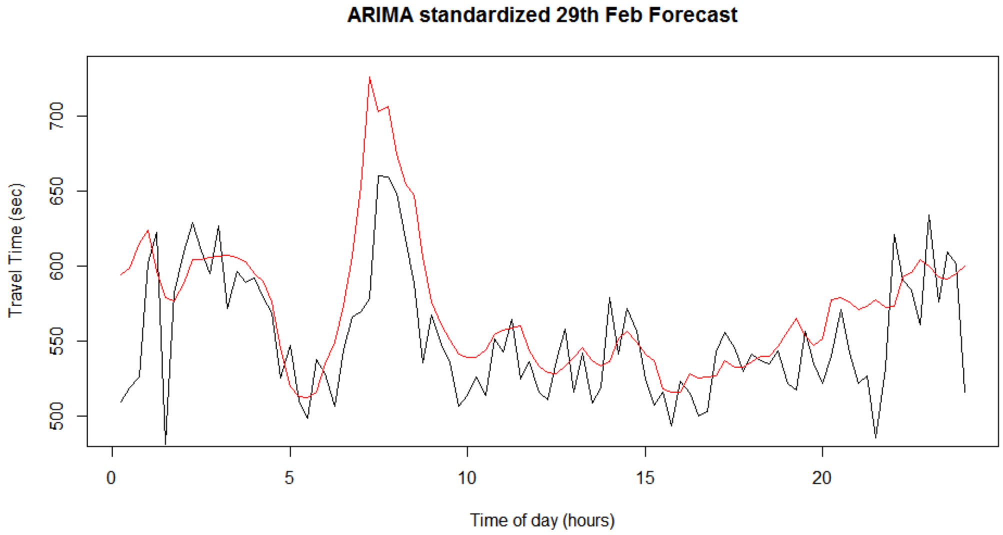
Figure 2.4.1: Plots of the partial and full autocorrelation functions for link 7.8. The significance level is shown in blue.

can be easily implemented using the `forecast` package in R. Applying an ARIMA model to the standardized data produces a forecast that follows the overall pattern of the day well, however the estimates for the peak at about 8am are slightly too high.

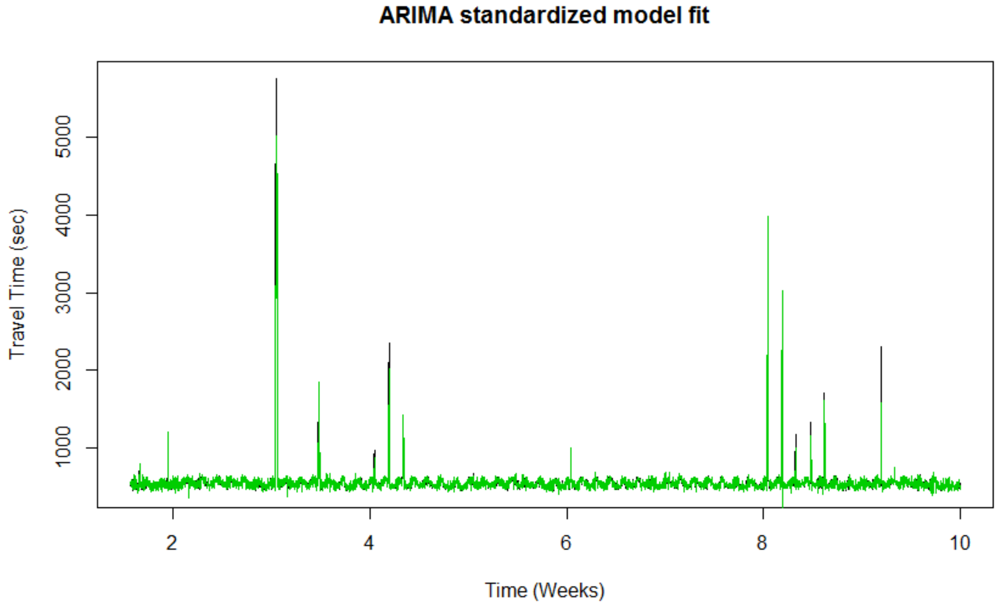
Figure 2.4.2 shows the forecast for the 29th and the fit on the training set. This overestimates the peak at around 8am but approximately follows the observed values throughout the day and correctly locates when the peak occurs. The best ARIMA model was an ARIMA(3,0,4) model. This has no differencing and 3 autoregressive terms and 4 moving average terms. This fits with the partial autocorrelation plot in Figure 2.3.9 as the first, second and fourth lags are significant. The corresponding coefficients are for the AR terms 0.68, -0.22 and 0.23 and the MA terms 0.47, 0.42, 0.27, 0.09 and 0.28.

ARIMA with Fourier terms. ARIMA models can include regression terms to try and reduce the error by modelling extra information such as the weather. Using a Fourier series to a set degree, K, is equivalent to regression terms. The Fourier terms attempt to capture any remaining seasonality in the standardized data and hence each series spans the 672 time steps. When the Fourier term coefficient model was run we first had to check what K was best. The best for this data is $K = 1$.

Figure 2.4.3 shows the forecast for the 29th February and the model fit. As with



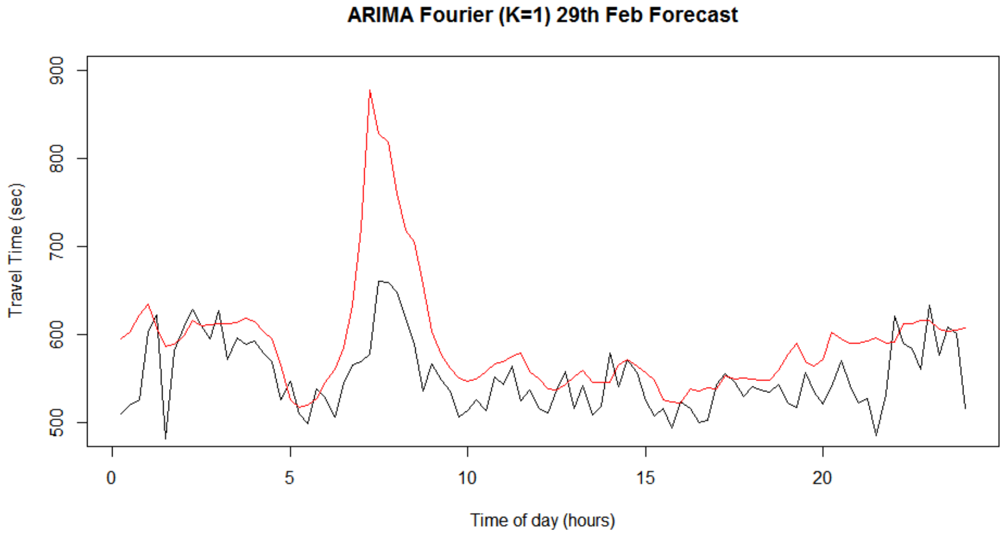
(a) Arima standardized forecast. The observed data are black and the predictions are red.



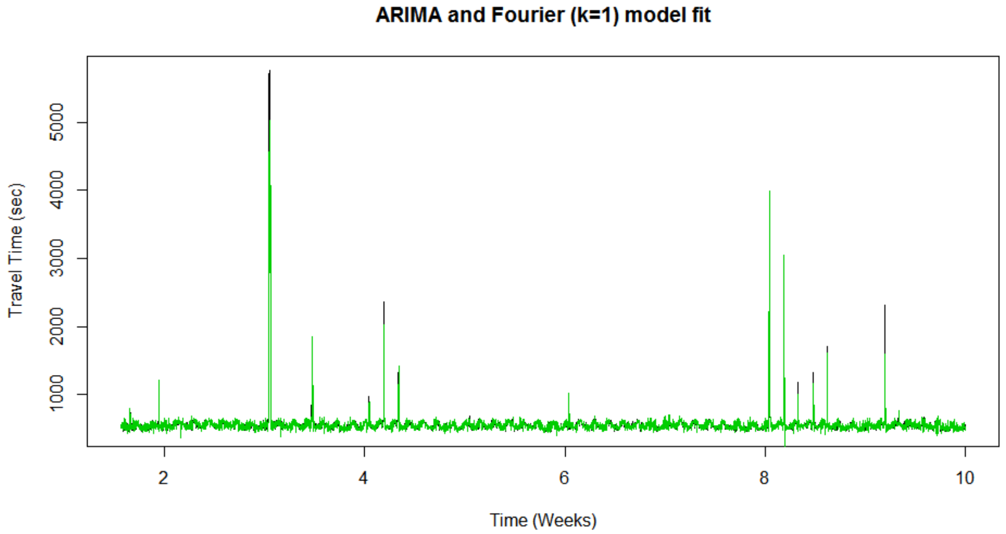
(b) Arima standardized forecast fit. The fitted values are in green while the observed values are in black.

Figure 2.4.2: The fit and forecast for the Arima standardized model.

all ARIMA type models the fit is good because it reacts quickly to change. The Fourier terms cause a large overfitting of the peak at 8am, much larger than the over-prediction of the ARIMA(3,0,4). The rest of the days prediction is also slightly



(a) Arima Fourier standardized forecast. The observed data are black and the predictions are in red.



(b) Arima Fourier standardized forecast fit. The fitted values are in green while the observed values are in black.

Figure 2.4.3: The fit and forecast for the Arima Fourier standardized model.

higher than the observed values for 29th February. The ARIMA part of the model is an ARIMA(3,0,5) which is an extra moving average term than expected from the partial autocorrelation. The AR coefficients are 0.25, -0.05 and 0.34, the MA terms 0.90, 0.74, 0.38, 0.19 and 0.68, the intercept 0.30 and the Fourier terms -0.078 and

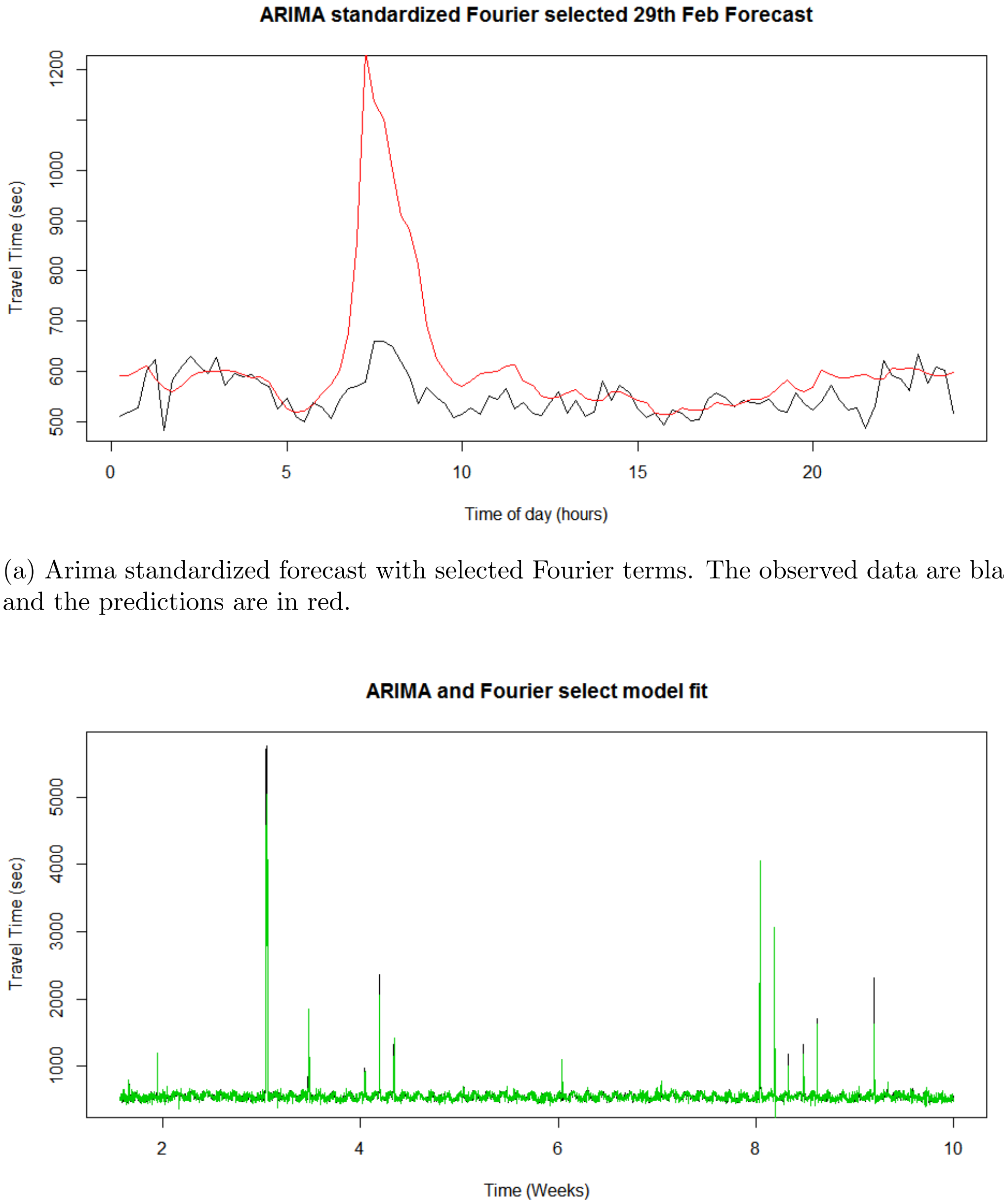
-0.35.

ARIMA with selected Fourier terms. Instead of selecting the first K terms, an alternative is to pick certain parts of the Fourier series. This series is made up of Fourier terms with $k = 1, 7, 14, 168$ which correspond to weekly, daily, half daily and hourly.

Figure 2.4.4 shows the ARIMA with selected Fourier terms on the standardized series. The Fourier terms have caused the slight peak to be greatly overestimated. The ARIMA part of the model is ARIMA(3,0,5), the same as the ARIMA with Fourier model. The coefficients of these are 0.22, -0.04 and 0.34 for the AR terms and 0.93, 0.76, 0.39, 0.20, 0.07 for the MA terms which are also similar to the ARIMA with Fourier coefficients. The selected Fourier terms were 0.297 for the intercept then the sin and cosine terms from weekly to the hourly are -0.08, -0.35, 0.30, 0.50, -0.52, 0.01, 0.04, 0.00.

ARIMA with day effect and selected Fourier terms An different model that can be tried is to use a day of the week term in addition to the selected Fourier terms. The ARIMA with day effect and selected Fourier terms also performs poorly, greatly overestimating the peak, although not quite as much as without the day effect. This can be seen in Figure 2.4.5 where the ARIMA part is also an ARIMA(3,0,5). The coefficients for the AR are 0.24, -0.05 and 0.34, the MA are 0.91, 0.75, 0.39, 0.20, 0.07. The Fourier coefficients 0.05, -0.57, 0.30, -0.50, -0.52, 0.01, 0.04, -0.00 and the Day coefficents are Tue 0.35, Wed 0.24, Thu 0.61, Fri 0.21, Sat 0.36 and Sun -0.18.

Exponential Smoothing. An alternative method is using exponential smoothing. The forecast for exponential smoothing can be seen in Figure 2.4.6. The forecast underestimates the peak slightly but provides a good estimate for the rest of the day. The model fit is also as good as the others. The ES model has an alpha parameter of 1.

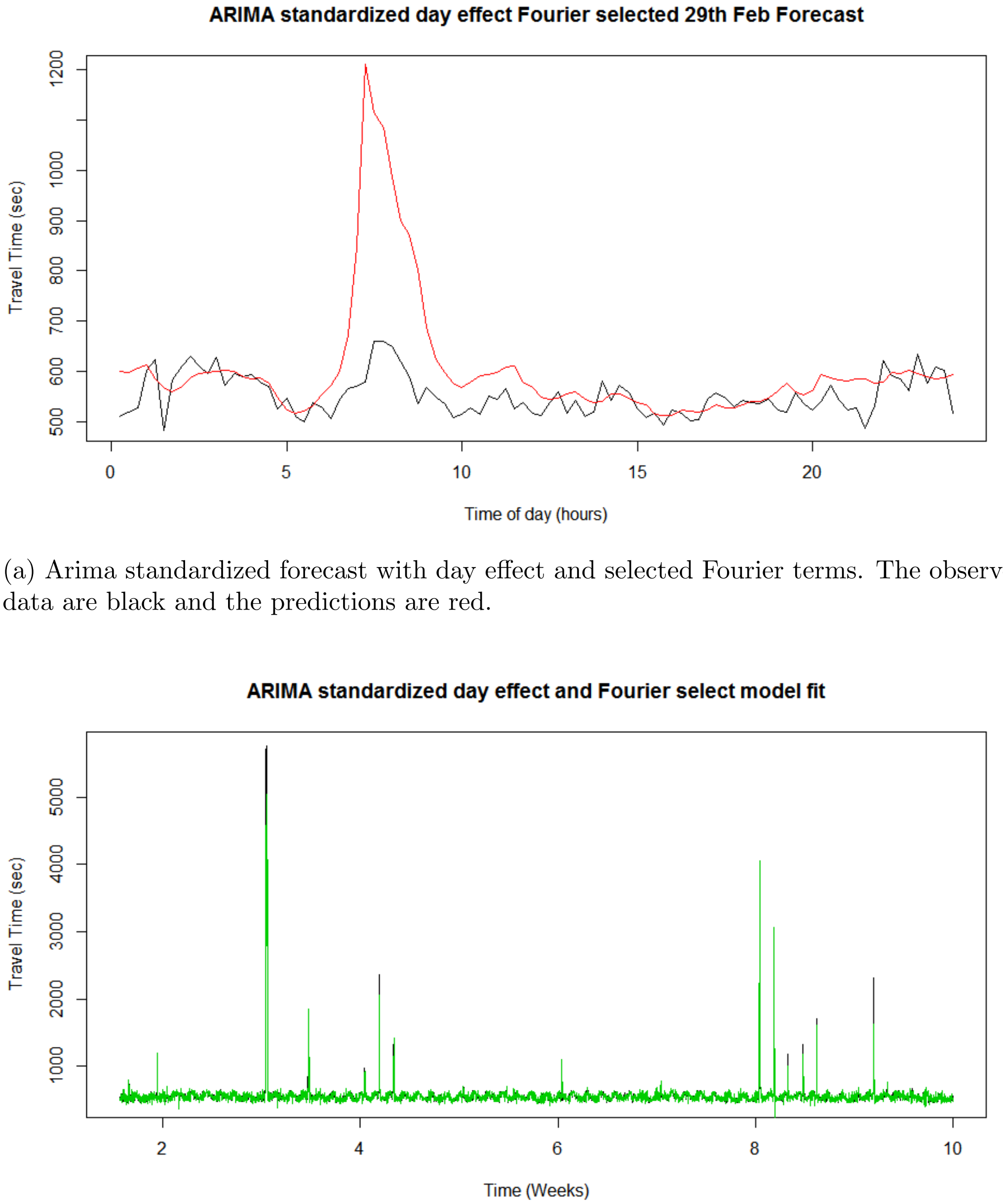


(a) Arima standardized forecast with selected Fourier terms. The observed data are black and the predictions are in red.

(b) Arima standardized fit with selected Fourier terms. The fitted values are in green while the observed values are in black.

Figure 2.4.4: The fit and forecast for the Arima standardized with selected Fourier terms model.

Comparison with other existing methods. One of the methods suggested in VRP the literature is to simply take the values of the previous day (Taniguchi and Shimamoto, 2004). Figure 2.4.7 shows the forecasts created using both the previous

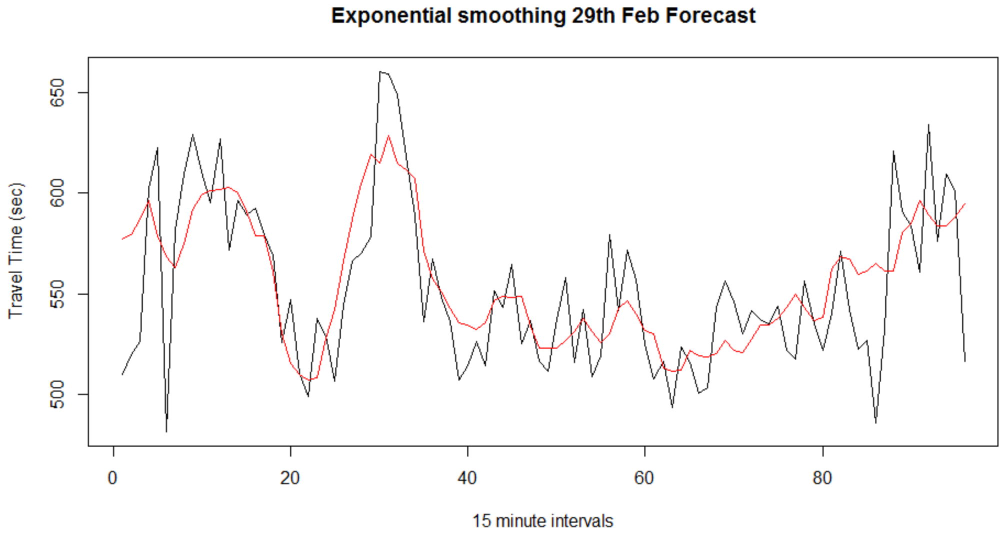


(a) Arima standardized forecast with day effect and selected Fourier terms. The observed data are black and the predictions are red.

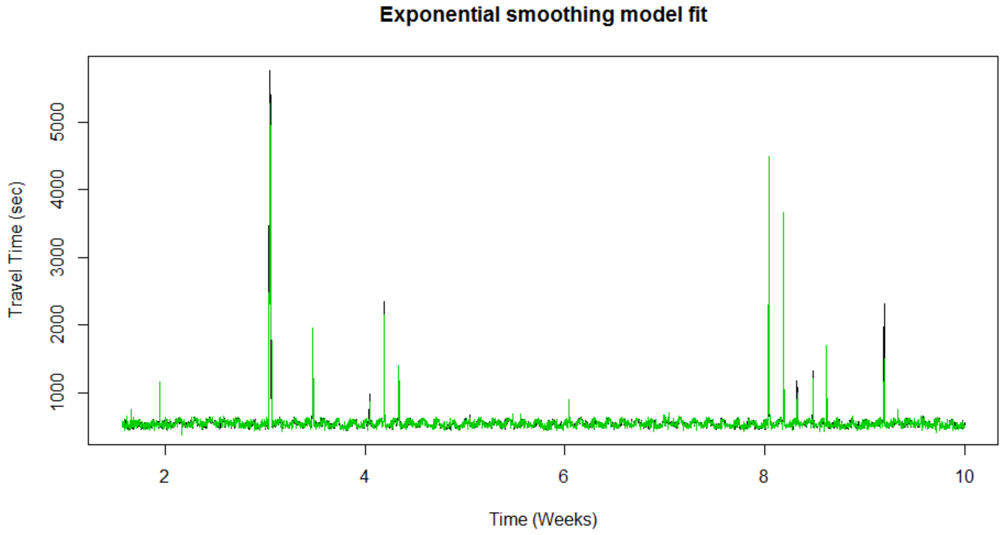
(b) Arima standardized fit with day effect and selected Fourier terms. The fitted values are in green while the observed values are in black.

Figure 2.4.5: The fit and forecast for the Arima standardized model with day effect and selected Fourier terms.

day (28th February) and the last Monday (22nd February). The forecast using the 22nd is much better than the one using the 28th. In the paper this method is used for a maximum of 10 hours in the future.



(a) Exponential smoothing forecast for 29th Feb. The observed data are black and the predictions are red.



(b) Exponential smoothing fit. The fitted values are in green while the observed values are in black.

Figure 2.4.6: The fit and forecast for the exponential smoothing model.

The same paper also proposed using a simulation with flow rates to generate the travel times but this isn't reproducible. All the other methods are used for short term forecasts only or rely upon additional information such as the speed.

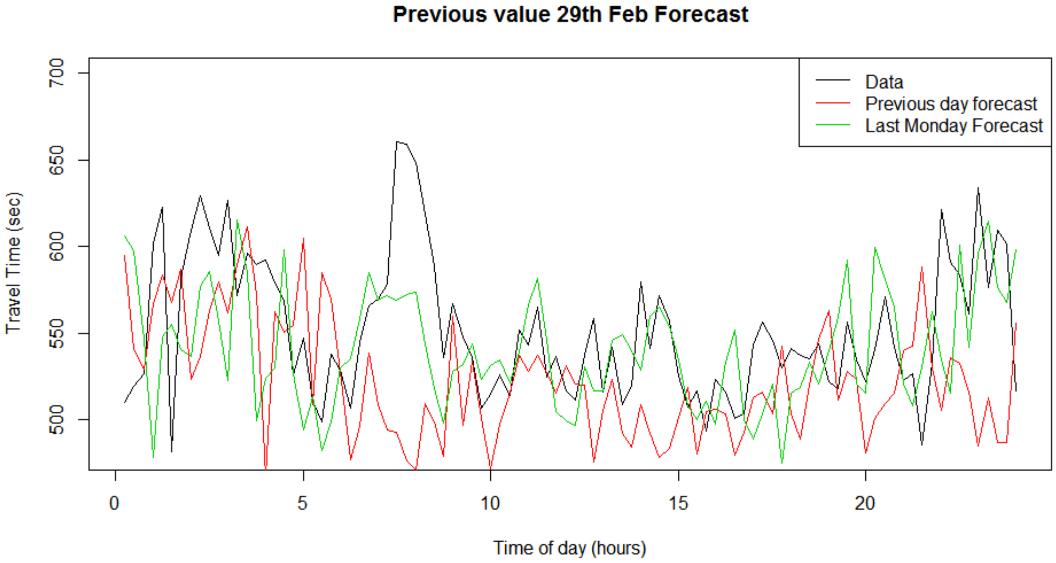


Figure 2.4.7: Forecasts using previous day and previous Monday.

Summary of methods. All of the model forecasts for the 29th February can be seen in Figure 2.4.8. It can be clearly seen that the ARIMA forecasts create too high a peak, whereas the other ones are too smooth for the data structure.

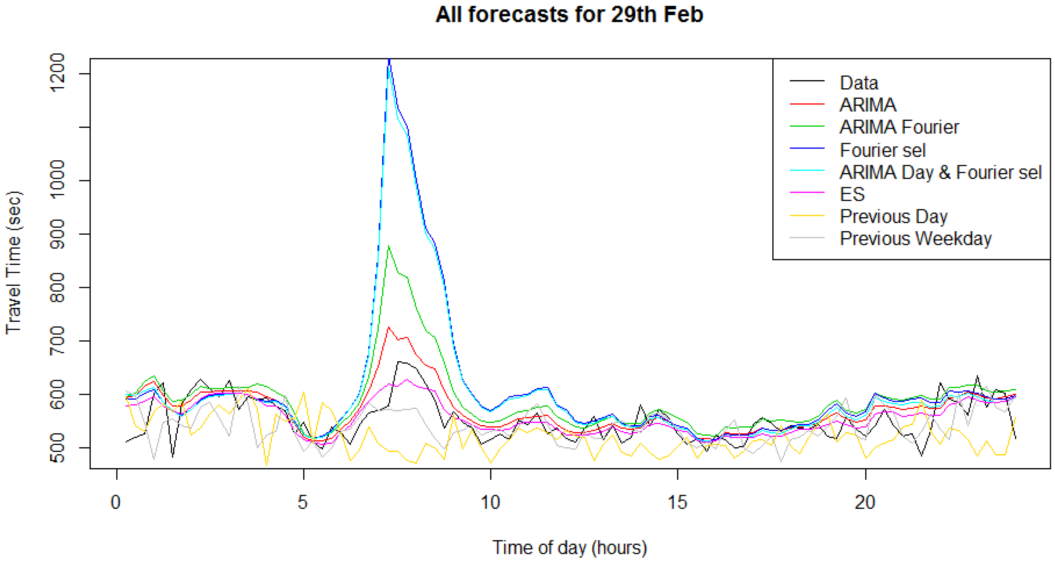


Figure 2.4.8: All travel time forecasts for 29th February.

The standardized ARIMA performs best at capturing the data patterns, but the naive previous value and ES methods also give close forecasts. We will now look at expanding the amount of data that is available to use to forecast by adding an additional month.

2.4.1 Robustness to missing data

Missing data. The January/February 2016 period for link 7.8 was chosen as it was the only two consecutive complete months. Also including March means that there is now missing data, which by convention are coded as NA. As mentioned in Section 2.3.1 missing data is common within any method of recording travel time. This can be due to sensor failure, bad weather conditions or recording errors.

To check the reliability of the forecast we require a complete day of observations to compare the forecasts with. The last complete day is 25th March hence Friday 25th of March is forecast using the data up to 24th March.

Bank holiday. Friday 25th March is a bank holiday which may cause some difference in the data due to less people driving to work. The plots are shown for the 25th of March only to demonstrate their effectiveness. The error metrics in Section 2.5.1 therefore also consider other days, starting with the 23rd and the 24th of March. We present the fit and forecast plots for each method in a similar format to the first part of Section 2.4. We focus upon a single model for all days but further work could be conducted upon improving the model with respect to atypical days such as bank holidays that we know about in advance as opposed to the unpredictable delays we look at in Chapter 3.

Models with missing data

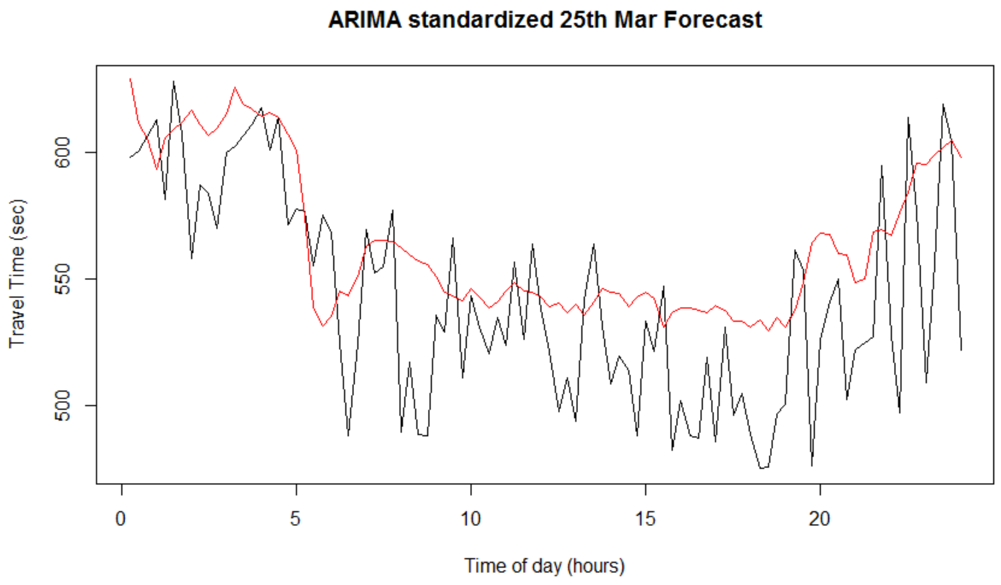
ARIMA model. The ARIMA model is an ARIMA(2,0,0) which doesn't have any differencing or moving average terms. The coefficients are 1.09 and -0.20 for the AR terms and 0.37 for the mean. The two biggest spikes on the partial autocorrelation plot are the first and second time lags, agreeing with the two autoregressive terms. Figure 2.4.9 shows an the forecast for the 25th of March. The model forecasts the general shape correctly however it is higher for the majority of the day than was observed.

ARIMA with selected Fourier terms. Figure 2.4.10 is a worse prediction than the ARIMA model. This is the ARIMA Fourier select model which is an ARIMA(2,0,0) model, like the standardized ARIMA model. The forecast has a part in the middle, between 8am and 12am, where it is much higher than the observed data, and the overall forecast is too high. The AR and intercept coefficients are very similar to the ARIMA model as 1.08, -0.20 and 0.37. The Fourier terms are -0.16, -0.35, 0.37, -0.61, -0.64, 0.01, 0.04 and 0.00.

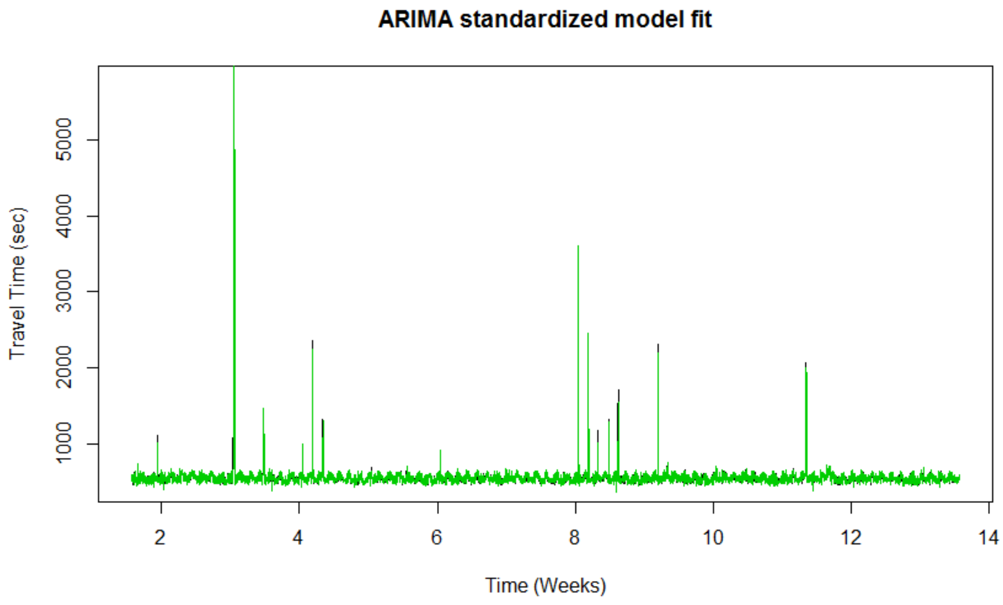
ARIMA with Fourier terms. The ARIMA model with Fourier terms provides the best forecast for the 25th March. The prediction is still slightly high in parts and the travel times do fluctuate a lot from one 15-minute to the next and the forecast follows the underlying pattern. Figure 2.4.11 shows that using the K=1 for the Fourier series provides a good estimate. This is an ARIMA(2,0,0) with coefficents for the AR of 1.08 and -0.20 and the intercept of 0.37 which is the same as the ARIMA with selected Fourier terms. The Fourier coefficents are -0.16 and -0.35.

Exponential smoothing. The forecast for exponential smoothing can be seen in Figure 2.4.12. The forecast slightly overestimates the day. The model fit slightly overestimates the highest peak. As with the model for January and February $\alpha = 1$.

Previous methods. Using the previous values only works if the whole of the previous day had no missing data. If there are any data missing it is unclear what should be used instead. One option would be to use the last value that had been observed



(a) Arima standardized model forecast. The observed data are black and the predictions are red.

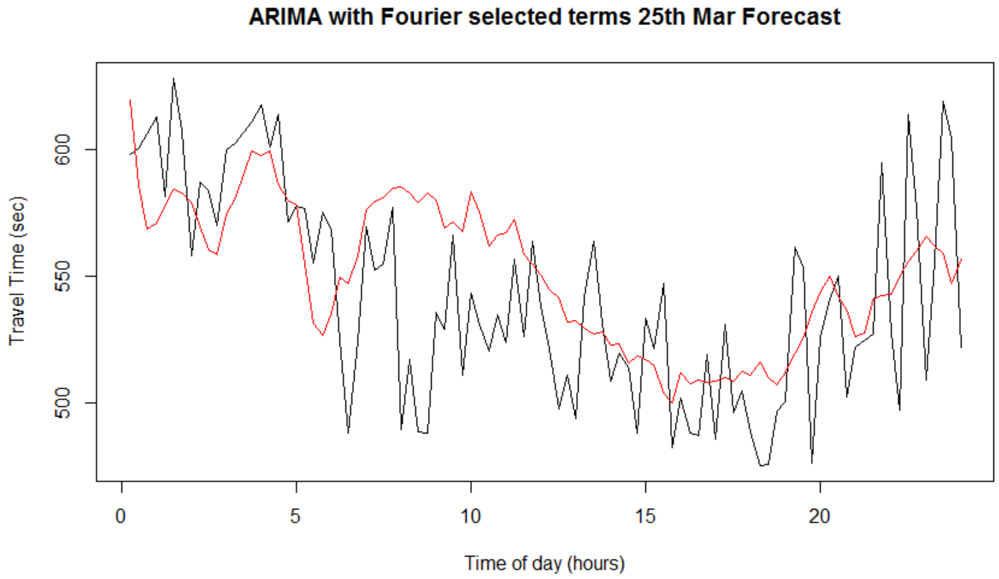


(b) Arima standardized model fit. The fitted values are in green while the observed values are in black.

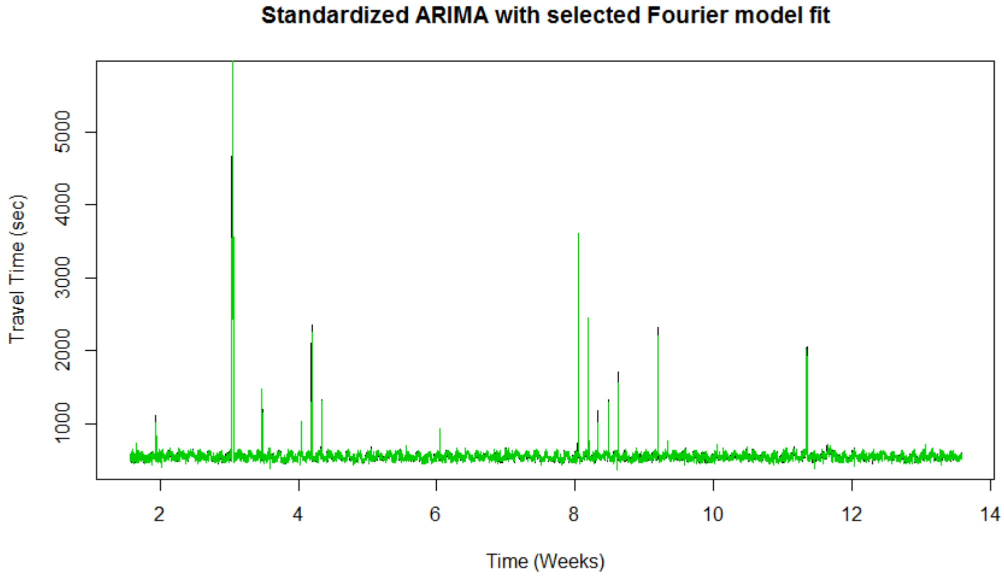
Figure 2.4.9: The fit and forecast for the standardized Arima model.

for that time index before the missing one.

Figure 2.4.13 shows the previous day forecasts which follow the general pattern



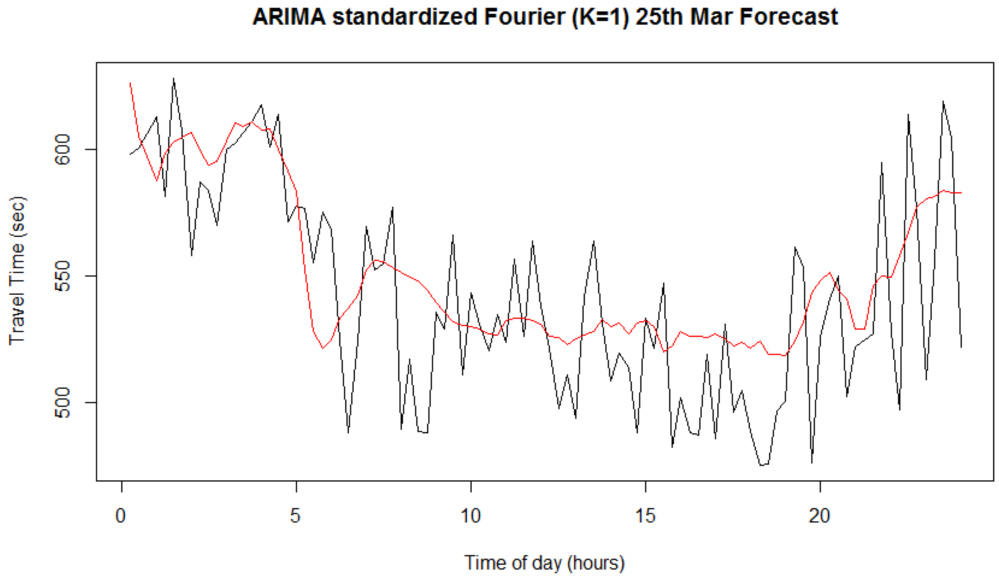
(a) Arima standardized forecast with selected Fourier terms. The observed data are black and the predictions are red.



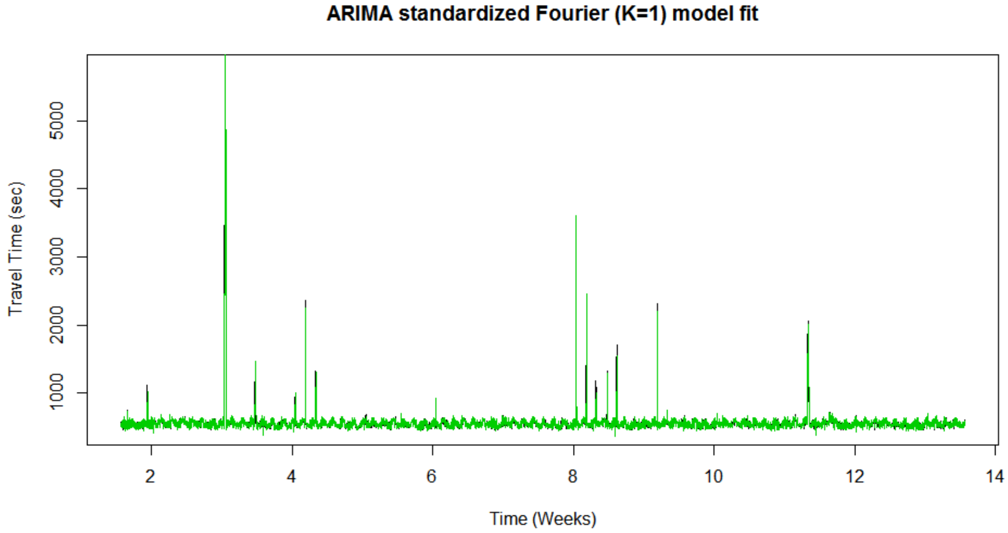
(b) Arima standardized fit with selected Fourier terms. The fitted values are in green while the observed values are in black.

Figure 2.4.10: The fit and forecast for the standardized Arima model with selected Fourier terms.

but are very spiky. This would prove a problem when inputting into the VRP problem as it could mean that if a vehicle enters the link a minute later it would emerge before.



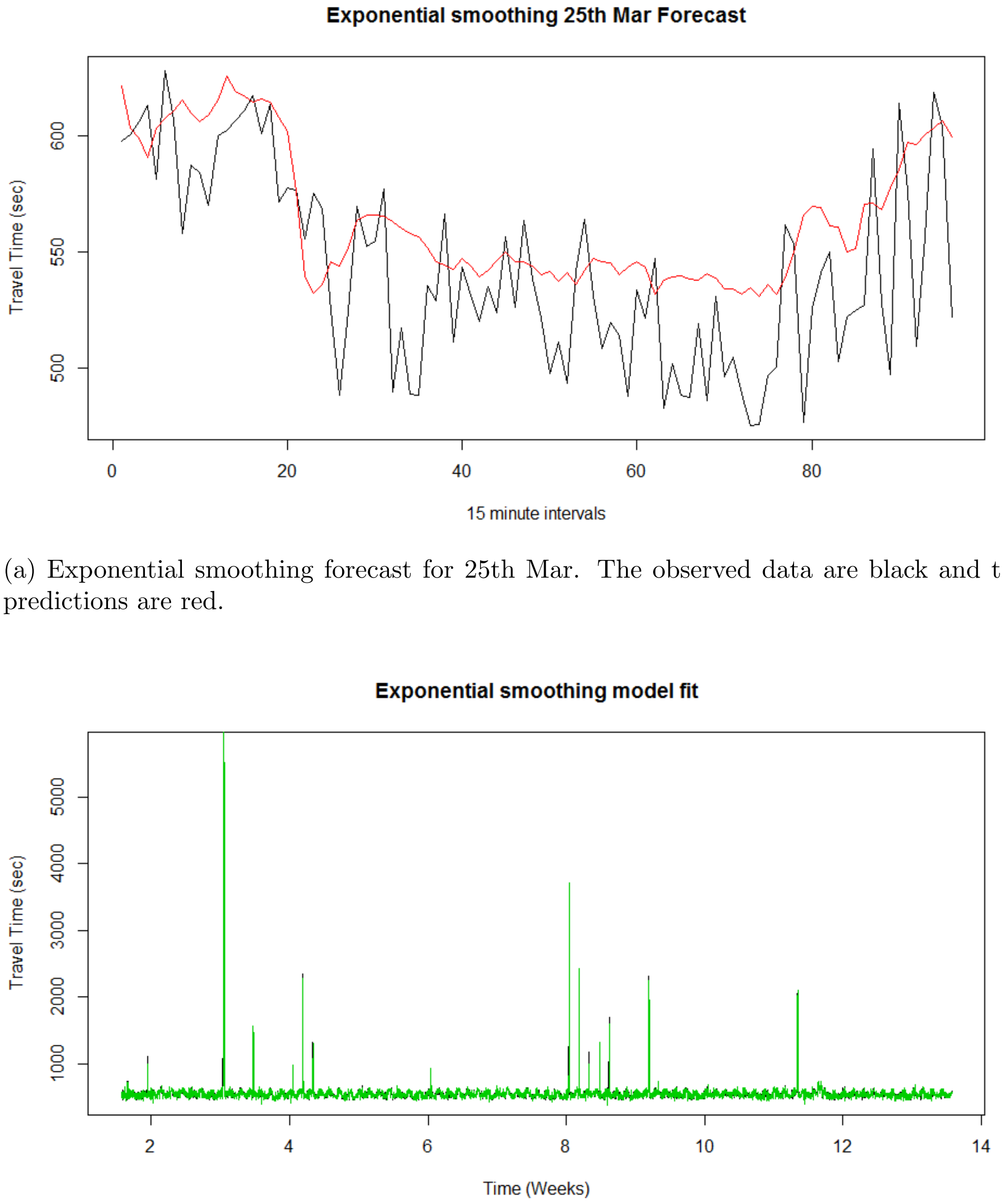
(a) Arima standardized Fourier 1 forecast. The observed data are black and the predictions are red.



(b) Arima standardized Fourier 1 fit. The fitted values are in green while the observed values are in black.

Figure 2.4.11: The fit and forecast for the standardized Arima Fourier 1 model.

This is called the FIFO or first in first out property as defined in Section 2.2.1.



(a) Exponential smoothing forecast for 25th Mar. The observed data are black and the predictions are red.

(b) Exponential smoothing fit. The fitted values are in green while the observed values are in black.

Figure 2.4.12: The fit and forecast for the exponential smoothing model.

Summary of methods. The forecasts for the 25th March are much closer to the observed travel time values. Figure 2.4.14 shows them all plotted on one graph. All of the ARIMA parts of the models are similar so they vary with the regression terms. The previous day values are close to the observed travel times.

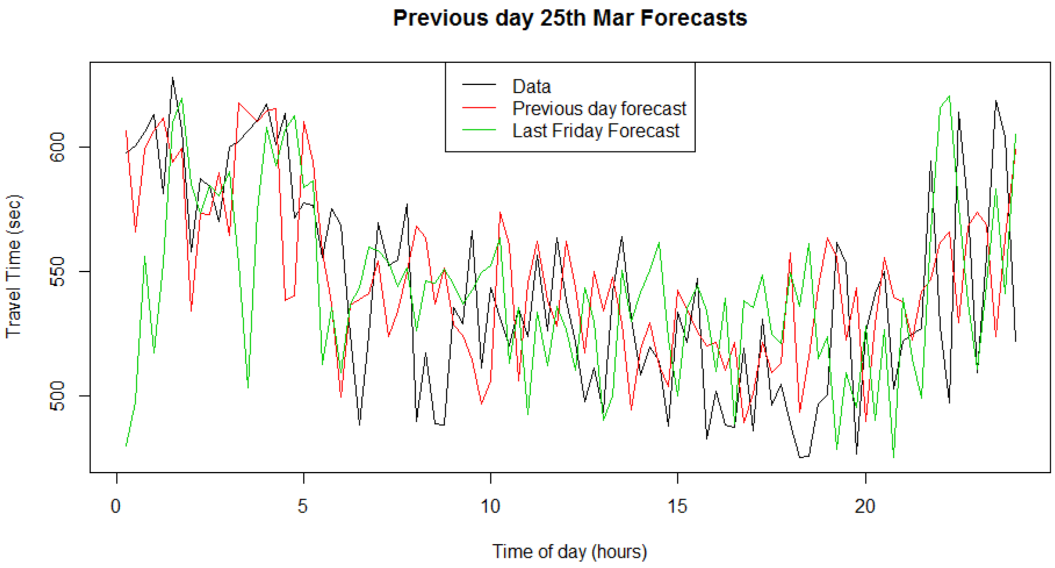


Figure 2.4.13: Previous value forecasts for 25th March.

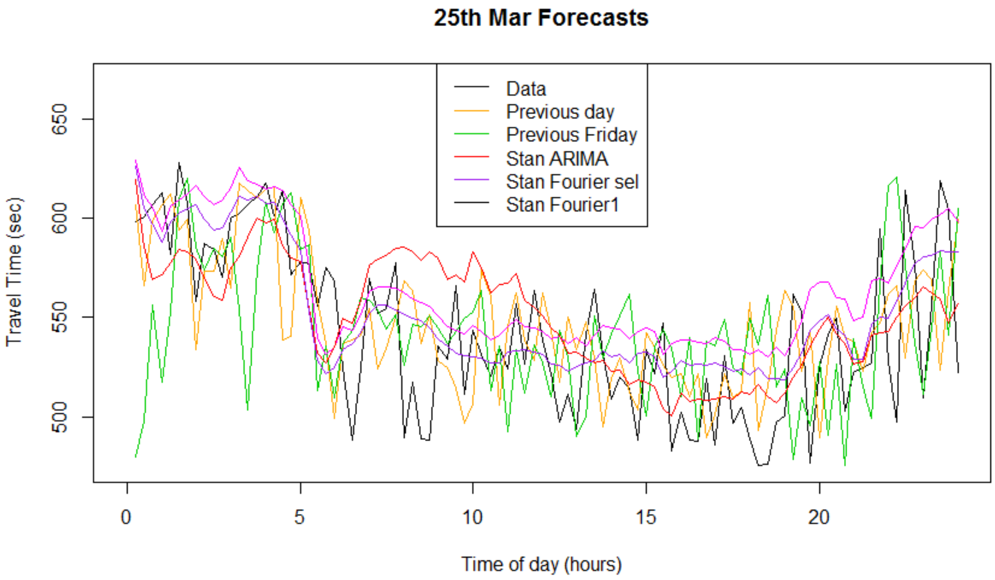


Figure 2.4.14: Forecasts of travel times for 25th March.

To check that the January/February/March model wasn't distorted by 25th March being a bank holiday the Fourier with K=1 was run again for Thursday 24th March. As can be seen clearly in Figure 2.4.15, using the K=1 for the Fourier series provides a good estimate with slight overestimation in between 7am and 10am.

This initial analysis will imply that the best method depends upon the day. A method is unlikely to be the best on one day but may be over multiple days. Classifying days into extra categories and using different models for different days is a potential avenue for further work. The Fourier with K=1 and ARIMA methods were the best on different days. Some of the methods greatly overestimate travel time peaks which would be an issue in terms of accuracy, while others miss rush hour peaks entirely. This was only over three days and hence we now look over a much longer period and over more days, necessitating the use of metrics rather than visual inspection to select the optimal method for all links.

2.5 Model diagnostics

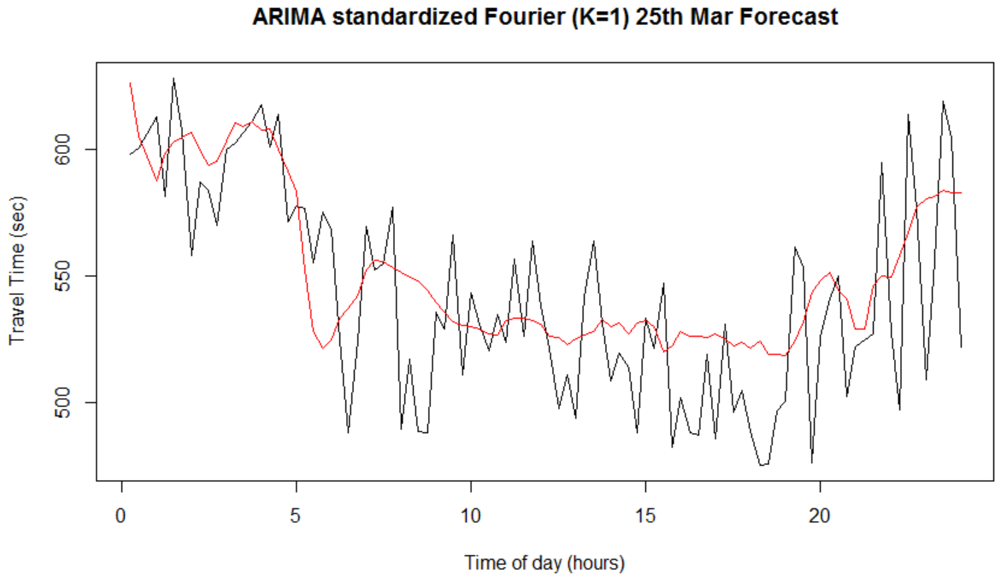
There are many different ways to measure how good the model is. We are interested in how well the forecast performs, as well as the overall model fit. The model needs to provide an accurate forecast for the entire 96 steps ahead, not just one so the model that fits the data the best may produce a terrible forecast.

Most measurements of the goodness of fit are based around the residual values. The residual value is the observed value, y_t, minus the fitted (or forecasted) value, $\hat{y}_t$. The simplest way to find a measure for the entire day is to add together all the residuals. This method has the disadvantage of disguising large errors if they are both positive and negative. This is

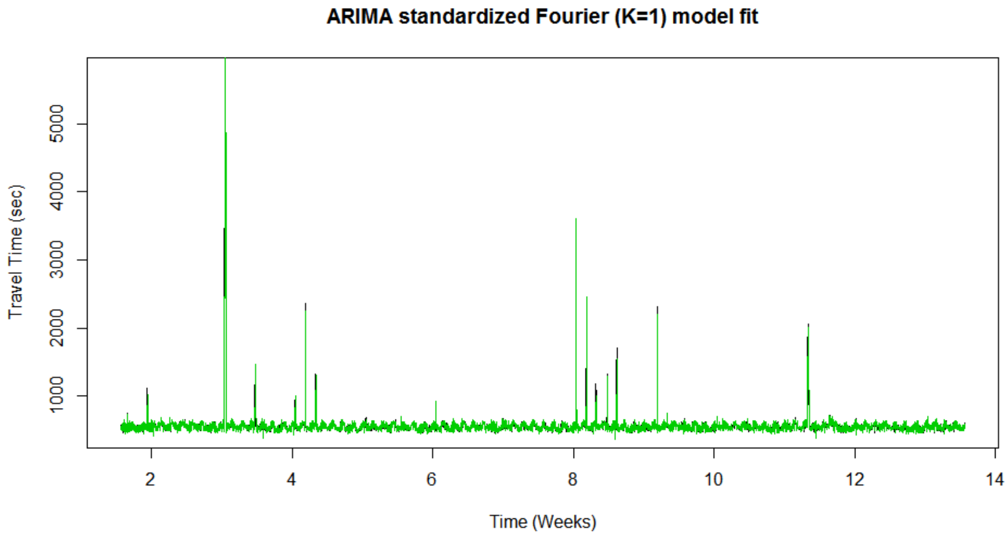
$$\sum_{t=1}^{T}(y_t - \hat{y}_t). \tag{2.5.1}$$

To ensure large errors cannot be cancelled out two main approaches have been suggested. One is to square the values and the other to take absolute values. Both of these ensure that all values are positive.

The three following error measures can be used to evaluate the forecasts (Ord and



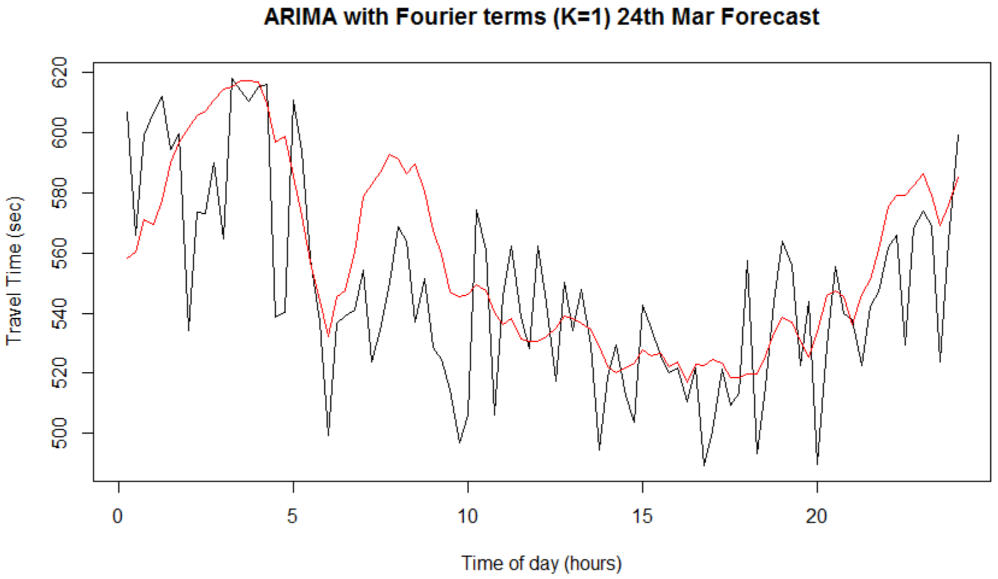
(a) Arima standardized forecast with selected Fourier terms. The observed data are black and the predictions are red.



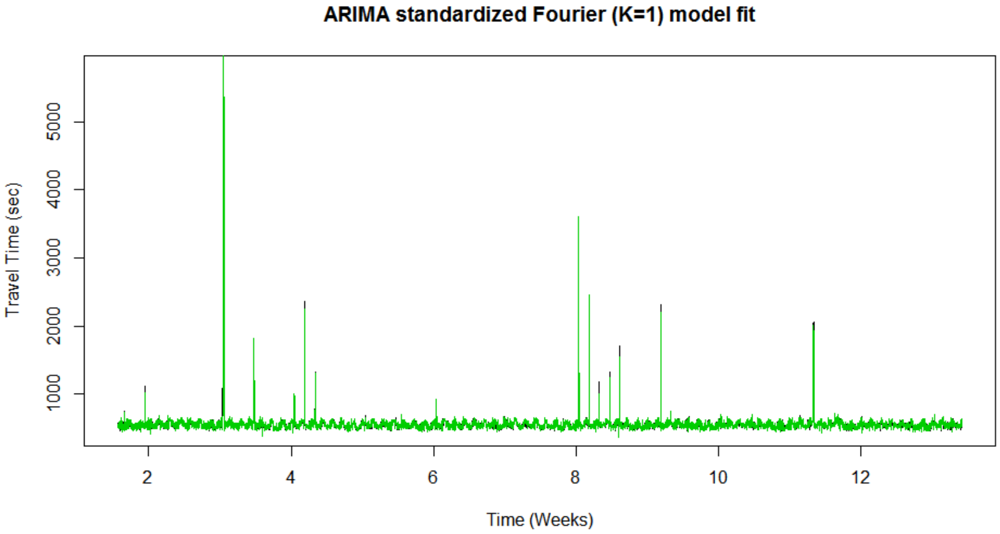
(b) Arima standardized fit with selected Fourier terms. The fitted values are in green while the observed values are in black.

Figure 2.4.15: The fit and forecast for the standardized Arima Fourier 1 model.

Fildes, 2013).



(a) Arima standardized forecast with selected Fourier and day terms. The observed data are black and the predictions are red.



(b) Arima standardized fit with selected Fourier and day terms. The fitted values are in green while the observed values are in black.

Figure 2.4.16: The fit and forecast for the standardized Arima Fourier day select model.

Root mean squared error. The RMSE or root mean squared error considers the mean of the squared standard errors. This will be affected by extremely large errors but presents a better overall picture of how far out the model will be from the true value, than simply considering the sum. This also enables the comparison of values from sequences of different lengths. The root is taken so the value is the same scale:

$$RMSE = \sqrt{\frac{\sum(y_t - \hat{y}_t)^2}{n}}. \tag{2.5.2}$$

Mean absolute error. The equivalent when considering the absolute values is the MAE or mean absolute error:

$$MAE = \frac{\sum |y_t - \hat{y}_t|}{n}. \tag{2.5.3}$$

Mean absolute percentage error. A further way to consider the error is to consider percentage errors. An error of 2 on a value of 3 is much worse than an error of 2 on a value of 100. The MAPE or mean absolute percentage error considers the absolute percentage error.

However these three measures depend on the size of the errors which will vary link by link. Ideally we want one type of model that can be implemented across the network. Hence we need to be able to compare the error metrics of multiple links. A link which has higher values overall is more likely to have bigger errors and therefore if a model fits worse on that link it will contribute more to the error metric than for another link. We therefore wish to use error metrics that are scale invariant, so we can choose the best method across all links.

To make the methods scale invariant we compare them to a benchmark method. This is normally the naive method, for which the previous day method is the most appropriate.

Average Relative Mean Absolute Error. Davydenko and Fildes (2013) suggest the use of the Average Relative Mean Absolute Error or AvgRelMAE. This is based upon the MAE from equation (2.5.3). We find the relative MAE between the current method and the naive method before by dividing each of the daily MAE calculations for each link by the corresponding ARIMA MAE estimate. The geometric mean of these values gives the AvgRelMAE. If the value is one then the method is equivalent to the previous day model, if it is less than one it is better and greater than one worse. Let MAE_i^s be the ith daily mean absolute error of the current method and MAE_i^b be the naive MAE. There are 96 values that make up each MAE, hence we calculate the AvgRelMAE as follows:

$$AvgRelMAE = \Big(\prod_{i=1}^{m}\Big(\frac{MAE_i^f}{MAE_i^b}\Big)^{96}\Big)^{(1/(96m))}. \tag{2.5.4}$$

Average Relative Median Absolute Error. We also include two other measures based upon the median as the data includes very high outliers. The first uses the Median Absolute Error rather than the MAE in the same manner as the AvgRelMAE. This considers the median of the errors rather than the mean.

Median Relative Mean Absolute Error The MedianRelMAE considers the median of the Relative errors, which in conjunction with the AvgRelMAE can identify if there are any particularly large relative errors. The MedianRelMAE is

$$MedianRelMAE = \text{Median}\Big(\Big(\frac{MAE_i^f}{MAE_i^b}\Big)^{96}\Big). \tag{2.5.5}$$

Comparison of measures. The measures all give numbers which are close to one, with the previous day being one for each measure. If the value is less than one then the measure considers the method better than the previous day model, if it is higher then it is worse.

All of these measures produce results based upon different things and thus it is impossible to compare them together directly. In most cases all three of the measures will suggest one model has the best forecasts. However there may also be cases whereby different measures suggest different models. In this case it is unclear which model to pick.

If all but one measure select one model then it would be reasonable to select that one. If all the measures suggest different models it is unclear which would be best. Which measure is likely to be favoured over the others depends upon the how the forecasts are used.

In our case we are using the model to provide one days forecast that will be input into a vehicle routing problem. This means that we want forecasts that are accurate across the entire day. We will pick the AvgRelMAE as it gives slightly higher weight to larger errors.

2.5.1 Results

Judging which method is best using only the 25th March as a forecast day is inappropriate, especially considering that the 25th was a bank holiday and may therefore have different travel patterns. As a further comparison between the methods we look first at the 23rd and the 24th of March, before considering more days and more links. These models are run using the data up to that day to tune the parameters and the standardization of the series. Hence the Auto Arima model for the 24th of March will have different coefficients and may use slightly different standardization terms to the 25th of March.

We use seven models to compare between - Exponential Smoothing, ARIMA, Fourier 1, Fourier Select, Fourier Select and Day effect, last same day and previous day.

Table 2.5.1 shows the two appropriate goodness of fit measures. We neglect the Median Relative MAE as there are only one value for each day.

With the exception of the 23rd of March the best for all measures is the Fourier with $K = 1$. The ARIMA forecast is better on the 23rd. Compared to the previous day estimate the previous day of the week is worse for all days, the Fourier 1 is always better over all three days while ARIMA and ES depend upon the day.

These results are generated for each day independently, whereas all three measures are designed to consider multiple days of results. We therefore combine them together to get Table 2.5.2. The best method is still the Fourier 1 and only the ARIMA and exponential smoothing are better than the previous day over most of the error measures. Exponential smoothing is better than the ARIMA over the three days.

		Error Measure	
	Method	AvgRelMAE	RelMedianAE
25 Mar	ARIMA	1.01	1.09
	Fourier1	**0.81**	**0.95**
	Fourier sel	1.37	1.69
	ARIMA Day & Fourier sel	0.98	1.15
	ES	1.02	1.06
	Previous Day	1.00	1.00
	Previous Weekday	1.14	1.26
24 Mar	ARIMA	0.99	1.22
	Fourier1	**0.83**	**0.77**
	Fourier sel	1.13	1.05
	ARIMA Day & Fourier sel	1.12	1.14
	ES	0.93	1.02
	Previous Day	1.00	1.00
	Previous Weekday.1	1.02	1.02
23 Mar	ARIMA	**0.79**	**0.82**
	Fourier1	0.86	0.97
	Fourier sel	0.97	1.15
	ARIMA Day & Fourier sel	1.10	1.35
	ES	0.80	0.90
	Previous Day	1.00	1.00
	Previous Weekday	1.09	1.31

Table 2.5.1: Table of error measure for forecasting link 7.8 by day.

Model recommendation for link 7.8. We therefore recommend the use of the ARIMA model with one Fourier coefficient for link 7.8 based upon the three days of data.

	Error Measure		
Methods	AvgRelMAE	RelMedianAE	MedianRelMAE
ARIMA	0.93	1.03	0.99
Fourier1	**0.83**	**0.89**	**0.83**
Fourier sel	1.15	1.27	1.13
ARIMA Day & Fourier sel	1.06	1.21	1.10
ES	0.91	0.99	0.93
Previous Day	1.00	1.00	1.00
Previous Weekday	1.08	1.19	1.09

Table 2.5.2: Table of error measures for forecasting the link 7.8 across the three days.

2.5.2 Overall Goodness of fit

We also wish to compare the method across multiple links to check that one method performs well on all the links in the network. Using different methods for different links would be infeasible when considering the size of the network that is used in the VRP of the petrol routing problem.

Because it takes a very long time to generate the Fourier model, over multiple links for multiple days, this has been ignored. Instead the Fourier 1 method has been selected. The ARIMA Day and Fourier select also takes longer to run than the other methods without obvious improvement, hence it has also been dropped. The results are compiled by calculating the next day as a forecast day, using only the information up to that point for January 8th to 25th March.

The links we used for this are 10.11, 9.10 6.10, 11.24 and 11.16, as described in Section 2.3.3. These are the links that will be used in the smallest network in Chapter 4.

Table 2.5.3 summarises the error metrics for each of the methods. The three ARIMA based methods perform much better than the naive previous day and previous weekday. Exponential Smoothing is slightly worse than the ARIMA methods but much better than the naive ones. The best method, although only just, is the ARIMA only method and as this is simpler and easier to run then it makes sense to use this.

	Error Measure		
Methods	AvgRelMAE	RelMedianAE	MedianRelMAE
ARIMA	**0.73**	**0.76**	**0.77**
Fourier1	0.74	0.76	0.78
Fourier sel	0.74	0.77	0.77
ES	0.80	0.83	0.81
Previous Day	1.00	1.00	1.00
Previous Weekday	1.03	1.03	1.04

Table 2.5.3: Table of the sum of error measures for forecasting each day between the 8th January and the 25th March for five links.

Consideration of different types of day. The travel time has two states - one where the travel time is approximately what we would expected over the day, and a second whereby a delay has occurred for some reason. The delays can be very large and may influence the models a lot. Hence we wish to consider how the model behaves on a standard day and on a delay day.

In a VRP we don't know if a delay will occur in the next day or not. However one model may be better at predicting when there is a delay due to recurrent congestion and produce a forecast that is closer to the observed travel time. Alternatively a model could be over influenced by an accident the previous day which is unlikely to occur the next day.

We must classify what a delay is and then locate any days which have a delay in. In Chapter 3 we consider a delay to be greater than two on the standardized scale but this means that a lot of days still include outliers. For link 9.10 using two means that only nine days have no outliers.

This issue arises because the AvgRelMAE is calculated using the daily MAE measures. Over the day most time periods have no delays, but if only one is above two, the whole day is classed as an outlier. One possible recourse to this is to calculate the MAE for each time point but this means that there is only one data point and the overall scale is different to the previous error metric tables.

The modelling of outliers requires a minimum number to be able to specify the distribution and hence may be overly cautious for our purpose. As an alternative we consider a day to be an outlier if it contains one or more travel times that are above five. This results in no outliers for link 10.11, which fits with the general pattern of the data for the link. For link 9.10 there are only 10 days with outliers, which is more indicative of extreme delays.

For each link we first classify a day as either an outlier or not and then calculate the error metrics for all days that are an outlier. Table 2.5.4 shows the error metrics if the day is an outlier and shows that the ARIMA is the best method and Previous Day the worst.

	Error Measure		
Methods	AvgRelMAE	RelMedianAE	MedianRelMAE
ARIMA	**0.80**	**0.74**	**0.86**
Fourier1	0.83	0.78	0.87
Fourier sel	0.83	0.78	0.86
ES	0.92	0.90	0.92
Previous Day	1.00	1.00	1.00
Previous Weekday	0.97	0.96	0.99

Table 2.5.4: Table of the sum of error measures for forecasting each day if the day is an outlier for five links.

Table 2.5.5 shows the error measures for the days that aren't outliers. This also agrees that the ARIMA method is best. The previous weekday is worse that the previous day whereas when there is an outlier it was slightly better than the previous day. Therefore the best method regardless of outliers is the ARIMA method.

Overall model recommendation. The overall model recommendation is to use the ARIMA model for all links of the network. The model performs best overall for the five links with all days considered. In addition the ARIMA is best when considering both days with delays in and days without, suggesting that even though it can't predict delays it doesn't perform any worse than the other models.

As expected the ARIMA isn't the best for all days and all links, as can be seen in

	Error Measure		
Methods	AvgRelMAE	RelMedianAE	MedianRelMAE
ARIMA	**0.72**	**0.76**	**0.75**
Fourier1	0.73	0.76	0.76
Fourier sel	0.73	0.77	0.77
ES	0.78	0.82	0.79
Previous Day	1.00	1.00	1.00
Previous Weekday	1.04	1.04	1.04

Table 2.5.5: Table of the sum of error measures for forecasting each day if the day is not an outlier for five links.

the three days for link 7.8. However the predictions aren't that far from the Fourier 1 model which was best for these and as the ARIMA model is simpler and therefore quicker to run the slight loss of accuracy on some days is acceptable.

2.6 Initial conclusions

Standardization is essential to ensure that the forecasts for the entire day retain the seasonal structure. For January/February the best method was using just a normal ARIMA, whereas for January/February/March the best was using a Fourier series with K=1 for link 7.8. Over the five links and over the entire time period the ARIMA model performs the best. The naive method of taking previous values is especially poor when a day has delays but is a better indication then using the same day the previous week in any other circumstance.

The degree of terms in the ARIMA models are always low, with the highest being 5. Most of the models only require 2 or 3 terms for the autoregressive and moving average parts. None of the models used differencing. However it should be noted that the AIC method chosen to select ARIMA models uses one step ahead forecasts rather than multiple step forecasts.

We considered days that had extreme outliers separately and the ARIMA model was still the best on these. However by making the errors scale invariant to the previous day error we have removed the consideration that the true error for an outlier day is larger than a standard day. As a result we will explore in more detail the outliers with the aim of being able to incorporate this into an improved model. A better understanding of when outliers occur and how severe they are will improve the prediction of travel time within the vehicle routing problem. Chapter 3 studies the behaviour of outliers, leading to models for the maximum size of the delay and the length of the delay as well as the probability of a delay happening. Chapter 5 discusses how we would approach combining these outlier models with the travel time forecasts generated either from this chapter or Chapter 4 within a VRP.

An alternative next step is to try and incorporate information from surrounding links into the forecasts. This may be of particular use when considering missing data, as the surrounding links can provide information as to the likely state of the link. The VRP uses the network and will require predictions from all the links rather than just one in isolation. Hence the most basic network model is to use the single link model on all the links. We look at models which use network information in Chapter 4.

Chapter 3

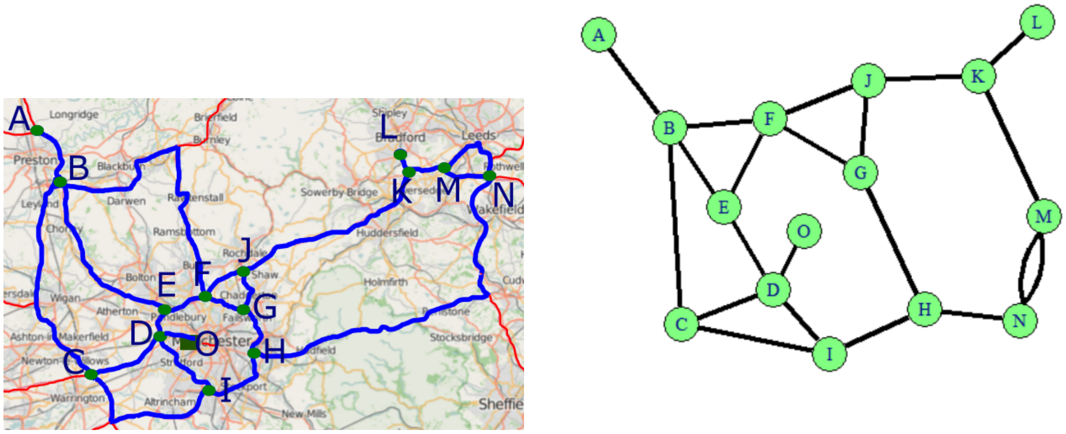
Analysis of extreme travel time delays

3.1 Introduction

As has been previously observed in Section 2.3.4, travel time data for traffic along roads usually follows an approximate weekly pattern with slight variation. However there are also large delays, which can be many times greater than the standard variation. There is no apparent pattern to when the delays occur. All of these factors suggest that these large delays behave differently to the weekly pattern and yet, if a vehicle gets caught in one of the delays it will have a great impact upon its journey. We therefore wish to model the extreme delays, with the aim of better understanding travel time across links and improving our current travel time model.

To examine the behaviour of delays on links we focus upon four links from a different network to the one introduced in Section 2.3.3. These links are from around Manchester and were chosen to represent different features of the travel time data, such rush hour occurring at different times. The Manchester network was originally selected to work upon throughout this thesis however due to a lack of A-roads the network from Section 2.3.3 was instead used to allow comparison between A-roads and motorways. The majority of work was complete upon this Chapter and hence uses the original network. The conclusions drawn from this chapter should be directly applicable to roads elsewhere in the larger road network, including the network from Section 2.3.3.

We first introduce the larger network, before defining the four links used in this section. The network is shown in Figure 3.1.1 which shows the network on the geographical road map and independently. The network is centred around the oil refinery at node O and stretches from Preston to Leeds.



(a) Road map.

(b) Grid layout.

Figure 3.1.1: Plots of the road layout of the Manchester network and the grid layout. Links of the network are identified in blue, with green circles as nodes. The oil refinery is a green square which is also a node. Each line represents two oppositely directed links. Map background OpenStreetMap contributors (2017).

The four links we look at are A.B (the M6 junctions 32 to 30), B.A (A.B in the opposite direction), F.E (M60 junctions 18 to 15) and C.D (M62 junctions 10 to 12). These represent two directions of the same road, a section of Manchester's ring road and a motorway section leading towards Manchester.

Chapter outline. We begin this chapter by defining a delay as a collection of successive time periods where the travel time is above a threshold and looking at the rate of occurrence of delays. We then study behaviour of the maximum size of a delay at a single time period within the delay, resulting in a generalized Pareto distribution. The duration of a delay is modelled by a discrete generalized Pareto distribution. Finally we combine the models of the rate, size and duration of a delay together to find the probability of a delay in an interval.

3.2 Delay event identification method

3.2.1 Justification for clusters

Large delays are characterized by the vehicle taking much longer to travel along the road than expected. This results in vehicles arriving at their destinations much later than expected, causing disruption to customers and potentially requiring routes to be cut short if drivers reach their regulated maximum driving hours. If some links suffer severe delays more frequently, or to a higher degree, then the delivery company may wish to take a slightly slower but more reliable route. Any unexpectedly short travel times are of little interest as the vehicles can usually unload early or wait until the start of the delivery slot.

We are therefore interested in any exceedances in travel times above a set threshold. Below this threshold we assume that all the variation in the travel time can be modelled by the single link model that has been selected using the analysis in Chapter 2. Anything above this threshold is an abnormal extreme delay which cannot be captured well by the single link model. Any exceedance of this threshold will be considered to have a larger delay from the journey than is predicted by the single link model.

The exceedances above the threshold cannot be considered to be independent. Travel time data are usually recorded at set intervals and any disruption to the travel time is often longer than a single interval. Hence a second exceedance depends upon the first. There are many different reasons for severe delays including a large volume of traffic and accidents. Large volumes of traffic take time to dissipate as a vehicle's speed is limited by those in front of them. Once the cause of the delay has been cleared the vehicles at the front of the queue can move freely, but the tail is still restricted and increasing in size. The collection of exceedances that result from a single reason are called an event.

If the events are independent of each other we can model the occurrence of the events and the behaviour of the travel time within events separately and then combine to make a model of the behaviour of events over time. If we assume that after every event the road conditions return to normal, then events on a stretch of road can be considered to be independent of each other. Events that result from extreme volumes of traffic, such as roadworks, may be more connected, however these usually have advance warning and could be accounted for separately. Hence we will assume that all events are independent.

In some instances there may be a chain of consequences from the original reason for the severe delay. For example a second accident could occur due to the queue from a first one. Any further delay from the second accident would be a direct result of the first, thus both accidents are considered to belong to a one event.

To model the events we need to identify them from the travel time data which is a time series. We have previously identified that an event can cause more than one exceedance. However by looking at the time series it is impossible to tell if two exceedances are from one or two events. Fortunately events only occur rarely and are very unlikely to occur successively. It is in fact impossible to locate the true start and end time of events, which are the physical delays, from the time series alone.

We therefore wish to infer where the events lie from the travel time data. From the point of view of travel time modelling we don't actually need to model the events themselves, only their effect upon the travel times. This effect can be simplified to the frequency of occurrence of events and the size of each event.

Thus we find and model the exceedances above the threshold. To do this we define a statistical object called a cluster, which is a set of exceedances. Ideally this cluster definition would be such that every cluster corresponds to all the exceedances that are due to a single physical event. If the definition is too broad a cluster can contain multiple events, but if it is too small then events can be split between clusters. By modelling the clusters we can predict the effect on the travel time. Each cluster is considered to be independent.

3.2.2 Cluster definition

In order to proceed we need a formal definition of a cluster. Smith and Weissman (1994) characterize clusters using two parameters, a threshold u, and a run length m, where a cluster is a set of neighbouring exceedances above the given threshold for a stationary sequence. Their definition is based upon observing the wait times between continuous sets of exceedances. These sets are infrequent and hence usually have large wait times, however there are also some very small wait times. Hence we consider joining two or more continuous sets into one cluster if the gap, or run length, isn't too large. In terms of travel time this means that the delay can recover before increasing again. This is particularly appropriate for values near the threshold where the absolute relative distance between two values is very small but only one is classed as an exceedance due to the in or out nature of a threshold model. Both of these values are large and the increase above the expected model is likely due to the same event.

To remove the weekly pattern we first standardize the data to ensure that the series is marginally identically distributed over time. This ensures that any outliers we observe belong to unusual delays for the time period. The standardization process is described in Section 2.3.5. The time series used in the rest of this chapter is assumed to be the standardized series. The use of the standardized series ties in with the single link model, of Chapter 2 which also uses the standardized series.

The mathematical definition of a cluster is as follows. Let y_t be the value of the series at time t. Let u be the threshold at which any value above the threshold is considered to be a long delay and hence belongs to the cluster. Thus if $y_t > u$ a severe delay is occurring. Let m be the run length or the minimum gap between two clusters.

We first define the start of a cluster. Let the cluster start at time t when both conditions in property (3.2.1) are satisfied,

$$y_t > u \qquad \text{and} \qquad \max(y_{t-m}, \dots, y_{t-1}) < u. \tag{3.2.1}$$

The travel time, at time t, is above the threshold but the previous m values are below so this is a new cluster which starts at time t.

A cluster ends at time τ, where $\tau > t$, when property (3.2.2) is satisfied,

$$y_\tau > u \qquad \text{and} \qquad \max(y_{\tau+1}, \dots, y_{\tau+m}) < u. \tag{3.2.2}$$

This identifies the last value of the cluster that is greater than the threshold. As the next m values are below the threshold the next time the series is above the threshold it will be the start of a new event.

However this definition of the end of a cluster doesn't guarantee that t and τ are the start and end times of the same cluster. We therefore require that property (3.2.3) is satisfied between the start and the end of the cluster,

$$\max(y_j, \dots, y_{j+m-1}) > u, \qquad t + 1 < j < \tau - m. \tag{3.2.3}$$

All continuous sequences, within the cluster, of values below the threshold have a length of less than m.

Properties (3.2.1) - (3.2.3) uniquely define one single cluster and for any time j between the start time, t, and end time, τ, y_j belongs to the cluster. Also all exceedances belong to one and only one cluster.

Figure 3.2.1 shows a collection of exceedances from the standardized travel time series for link A.B which we will use to explain how we can define the cluster while also considering events. The blue points show the exceedances as well the points in between which may be in the cluster. The inclusion of the two points below the threshold depends upon if there are one or two clusters. This suggests there may be one or two events that lead to the observed exceedances.

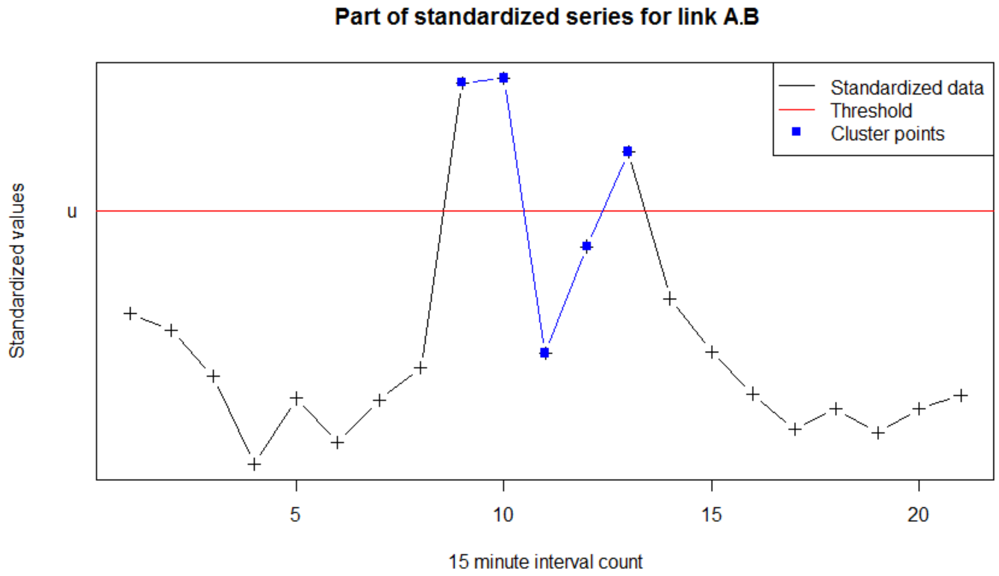


Figure 3.2.1: Plot of a cluster on the link A.B. The cluster is identified by the blue points, with the threshold line in red.

If m is three then the five blue points are one cluster as the longest consecutive series where the points are below the threshold line u has a length of two. This cluster starts at time $t = 9$ and ends at time $\tau = 13$. There are also at least three points below the threshold on either side of the cluster. If m was chosen to be one or two then there would be two clusters, one with two points, at times $t = 9$ and $\tau = 10$, and one with only one point, at time $t = \tau = 13$.

The number of clusters for the threshold u therefore depends upon the choice of m. We next discuss how to appropriately select both u and m.

3.2.3 Parameter choice

As identified in Section 3.2.2, events are defined by two parameters, the threshold u and the run length m. Both of these parameters must be appropriately selected, and it may be that the choice of one effects the other (Smith and Weissman, 1994). If the behaviour of the clusters of outliers may change as the threshold gets higher then the run length will be dependent upon the choice of the threshold and is denoted m_u.

In some instances there may be an obvious choice of the threshold, motivated by the situation, however in the case of travel time there is no clear choice. There are also modelling considerations for modelling both within clusters and between clusters which directly linked to the threshold level. If the threshold is too high then there will be too few outliers to construct a good model. However if the threshold is too low then some of the outliers won't be severe delays, and the clusters will be too large and too frequent.

The choice of the run length estimator, once the threshold has been selected, is even more difficult. If it is too small then clusters are separated into multiple parts, however if it is too high then clusters are incorrectly grouped together. Grouping multiple clusters together makes modelling them as a cluster much more difficult.

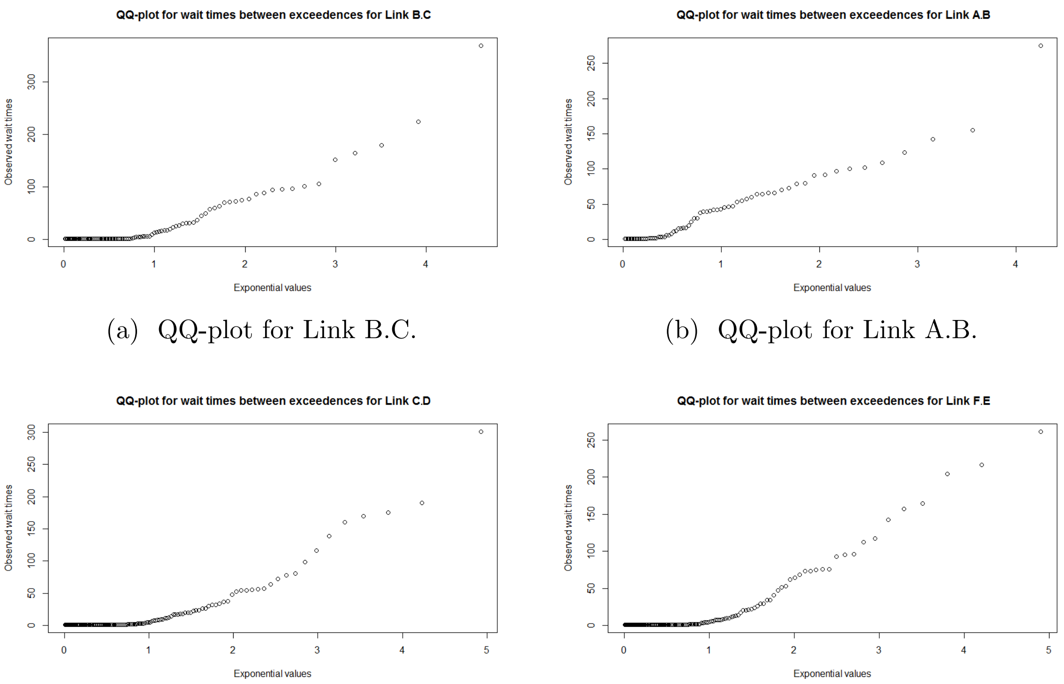
Ferro and Segers (2003) suggest a method of selecting m for given threshold parameter u which is based upon observing the wait times between successive exceedances. They assume that the series is strictly stationary with known marginal distribution function and first consider the case whereby all exceedances in a cluster happen at once. As the clusters occur simultaneously the times between exceedances in a cluster is zero and the run length m is zero. The times between clusters are exponentially distributed.

They then expand this to cases whereby the simultaneous occurrence of clusters is true in the limit. Instead of all the wait times being zero within clusters they will be zero or close to zero. Then by identifying at which point the distribution of these times changes from within clusters to between clusters an estimate for m can be found. An exponential qq-plot will show the two different distributions with a near horizontal line at the beginning where the inter-exceedance times are zero, followed by a straight line where the times are between clusters and follow an exponential distribution.

Because our data is recorded every 15 minutes, for each time point there can exist at most one exceedance. The between exceedance times are integer counts and the least they can be is 1. The times between exceedances within clusters will therefore be small, although not 0. However we can still use the idea of having the between clusters times having one exponential distribution and the within clusters having a different distribution to identify a possible m.

Due to the way properties (3.2.1) - (3.2.3) have been defined, m must be a positive integer. On the other hand u can be anything, although data analysis suggests that anything over five results in very few exceedances. We initially choose $u = 2$ as this ensures that the majority of the travel times are not considered to be outliers but the number of outliers is high enough that they can be modelled.

Figure 3.2.2 shows the qq-plot for all the times between all exceedances for links A.B, B.C, C.D and F.E when $u = 2$. In all four of the plots there is a clear concentration of points with a wait time of close to zero. The rest of the graphs follows an approximate straight line suggesting an exponential model may be appropriate for the between cluster times. This implies that m is small. There is no reason to assume that m is the same for all the links. However this assumption reduces the number of parameters and therefore makes the modelling easier. The plots of the four links all behave slightly differently, and have different distributions. By comparing the plots this suggests a value of $m = 3$ as $m = 1$ results in there being values off the straight line for the exponential distribution.



(a) QQ-plot for Link B.C.

(b) QQ-plot for Link A.B.

(c) QQ-plot for Link C.D.

(d) QQ-plot for Link F.E.

Figure 3.2.2: QQ-plots for links B.C, A.B, C.D and F.E for the times between exceedances. This is against the exponential quantiles and shows a high concentration with time interval one for the observed values for all links.

3.3 Modelling cluster occurrence

3.3.1 Cluster simplification

In the previous section we identified what a cluster is given the threshold u and the run length m. We now wish to model the clusters to get a better understanding of their size and when they occur. The clusters, by definition, can contain multiple exceedances because severe delays often last longer than one time period. As a result the exceedances within the same cluster aren't independent of each other.

One simple way of modelling the clusters is to condense each cluster to a single point. This point has a single magnitude and occurs at one point in time. Then from the single point we can infer the effect upon the travel time of a future severe delay. This delay is termed an event in Section 3.2, describing a single reason for a set of exceedances. By only using points the delay effects are much simpler to model. However the single point must be a sufficient representation of the clusters as a whole. We therefore wish to model the cluster as a single point such that one point represents one set of severe delays. This definition ensures that we only consider each cluster once and the models aren't biased by the multiple cluster exceedances occurring in succession.

We therefore need to select a single point that represents the entire cluster. To select this we have only the values within each cluster. Let y_t be the value of the standardized series at time t. Using the same notation as Section 3.2 we define the start and end times of cluster j as t_j and τ_j respectively. This gives the points $y_{t_j}, \ldots, y_{\tau_j}$ from which to select a single point. This could be the single point if $t_j = \tau_j$ or a combination of the points if $t_j < \tau_j$.

We wish to model the severity of the effect on the travel time from the exceedances. The higher the exceedances in a cluster the greater the effect on the travel time. Thus we choose to use the cluster maximum as the single point that represents the cluster. The cluster maximum represents the worst case impact of an event over a single interval as any other single point will result in a lower value. This also has a clear physical meaning and is easy to interpret.

We represent the maximum of cluster j as Y_j^c. The cluster maximum Y_j^c is the highest point within the cluster j, i.e.,

$$Y_j^c = \max(Y_{t_j}, \ldots, Y_{\tau_j}). \tag{3.3.1}$$

Having identified the single value as the cluster maximum we now need to locate the occurrence of the single point that represents the cluster. We could locate the cluster at the beginning or end of the cluster, but as the cluster maximum is where the cluster has the most impact upon the travel times we choose to locate the cluster there. The location of the jth cluster maximum is defined as the time point in the jth cluster which has the highest value, which is indicted by I_j^c, i.e.,

$$I_j^c = \text{argmax}(Y_{t_j}, \ldots, Y_{\tau_j}). \tag{3.3.2}$$

We now have a set of points, $(I_j^c, Y_j^c)_{j=1,\ldots,n_c}$ where n_c is the number of clusters, such that each pair represents one cluster. By the definition of a cluster $Y_j^c > u$, so each of these points is an exceedance above the threshold u. We wish to model these so we can better understand how the travel time in the future is likely to be affected by an event.

The representation of the clusters as points that occur in time suggests that a point process may be appropriate as the points appear to occur randomly (Davison and Smith, 1990). A point process models randomly occurring events that happen in continuous time but is a good approximation when events are recorded in discrete time. Here our event is a severe delay which occurs at a certain time but the data are only recorded every 15 minutes.

By definition the clusters must occur separately and it is reasonable to assume that anything that leads to an event, such as an accident, is entirely independent of the last such event. Thus it is reasonable that the history has no bearing upon the present. The clusters also occur infrequently. This suggests that a Poisson process may be an appropriate model for the occurrence of clusters.

A Poisson process is defined for all time. For any two times $t_1 < t_2$, let $N(t_1, t_2)$ be the number of events that occur in the time interval (t_1, t_2) (Cox and Isham, 1980). The events have an occurrence rate $\gamma(t)$, at time t, with $\gamma(t) \geq 0$ for all t. Then, considering the history H_t of the points in time up to the current time t, and letting $\delta \to 0_+$ we have:

$$P(N(t, t+\delta) = 1 | H_t) = \gamma(t)\delta + o(\delta) \tag{3.3.3}$$

$$P(N(t, t+\delta) > 1 | H_t) = o(\delta). \tag{3.3.4}$$

This implies that :

$$P(N(t,t+\delta)=0|H_t)=1-\gamma(t)\delta+o(\delta). \tag{3.3.5}$$

As a result of equation (3.3.4) it is very unlikely that two or more events occur at the same time. Equation (3.3.3) shows that the larger the occurrence rate $\gamma(t)$, the larger the probability that there will be a single event near time t.

There are three main properties of a Poisson process. The first is that the number of events in an interval follow a Poisson distribution. Specifically the interval (t_1, t_2) has:

$$N(t_1,t_2)\sim \text{Poisson}\left(\int_{t_1}^{t_2}\gamma(t)dt\right). \tag{3.3.6}$$

Both equation (3.3.3) and equation (3.3.4) show that the occurrence of an event is independent of when the last event occurred, or how many there have been previously, because they have no dependence upon the history H_t. This implies that for any two intervals (t_1, t_2) and (t_3, t_4) such that $t_1 < t_2 < t_3 < t_4$, the count of $N(t_1, t_2)$ is independent from $N(t_3, t_4)$.

Finally the occurrence times of events in the interval (t_1, t_2) have density

$$\frac{\gamma(t)}{\int_{t_1}^{t_2}\gamma(s)ds}. \tag{3.3.7}$$

There are two types of Poisson process, homogeneous and non-homogeneous, depending upon whether $\gamma(t)$ varies with t. The homogeneous Poisson process is a special case whereby the occurrence rate is constant through time and is referred to as γ, i.e., $\gamma(t) = \gamma$, and

$$N(t_1,t_2)\sim \text{Poisson}((t_2-t_1)\gamma). \tag{3.3.8}$$

In a homogeneous Poisson process, because the rate of occurrence is constant through time, equation (3.3.3) implies that the history of events and current time are irrelevant to the future events once γ is known. This is also known as being memoryless. This leads to the waiting times between the events having an exponential distribution, with rate γ, and the distribution of occurrence times in the interval (t_1, t_2) being uniform(t_1, t_2). This last property can be deduced from equation (3.3.7) as the rate is constant.

Likelihood ratio tests for constant rate. We wish to check whether or not a Poisson process is an appropriate model for the cluster occurrences and in particular whether a homogeneous process is a reasonable approximation. First we check visually whether there is any pattern to when clusters occur. If it is a homogeneous Poisson process the rate should remain the same for all time and hence there should be no pattern.

Figure 3.3.1 shows the occurrence of clusters throughout the month of January for link A.B. The clusters were selected with a threshold level of 2 and a run length of 3, using the parameters selected in Section 3.2.3. The cluster is located at the point in time of the highest outlier within the cluster. The severe delays occur infrequently. Some days have no clusters and some days have more than one.

How $\gamma(t)$ varies with time is a key difference between a non homogeneous and a homogeneous Poisson process. If $\gamma(t)$ is constant for all time the process is homogeneous. Therefore we wish to test if $\gamma(t) = \gamma$ is an appropriate model as opposed to letting the rate vary through time. An estimate for a constant rate is easy to find using maximum likelihood estimation but estimates for $\gamma(t)$ require either the specification of a parametric model or non-parametric class of smoothness.

Ideally we would let $\gamma(t)$ be able to vary throughout all time but we need to be able to estimate $\gamma(t)$ and clusters occur infrequently. We subset the data, to ensure we have enough data to fit each model for $\gamma(t)$. Within each subset $\gamma(t)$ is assumed to

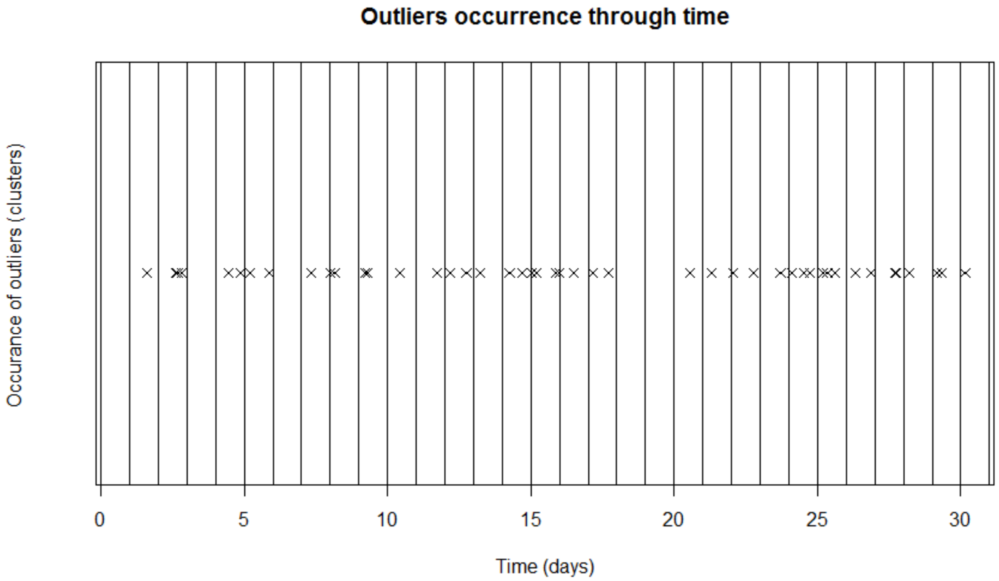


Figure 3.3.1: Outlier occurrence in time for link A.B. Each cross represents a cluster maxima in a fifteen minute interval.

be constant. This then represents an approximation of the non homogeneous Poisson process.

We first consider a subset to be of one day. Figure 3.3.2 shows the counts for each of the thirty one days and we wish to check if this is homogeneous. By equation (3.3.8) if this is true the counts come from a single Poisson distribution with mean $(t_e - t_1)\gamma$, where t_1 and t_e are the first and last time in the month. In the non homogeneous case each subset is homogeneous and hence has a Poisson distribution with mean $(t_{i+1} - t_i)\gamma$ such that t_{i+1} is the end of one day and t_i is the start of the same day, so the mean is 24γ if it is measured in hours.

The t_i causes the notation to get messy. If the subsets are all of equal length we can calculate $\tilde{\gamma}$ on the scale of one subset such $(t_2 - t_1) = 1$. To calculate the true rate we divide $\tilde{\gamma}$ by $(t_2 - t_1)$. For the ease of notation we henceforth refer to the rate as γ on the scale of the corresponding subsets.

To test whether all the days have the same rate γ we use a likelihood ratio test. This enables us to compare how likely it is that we observe the data if it is from a

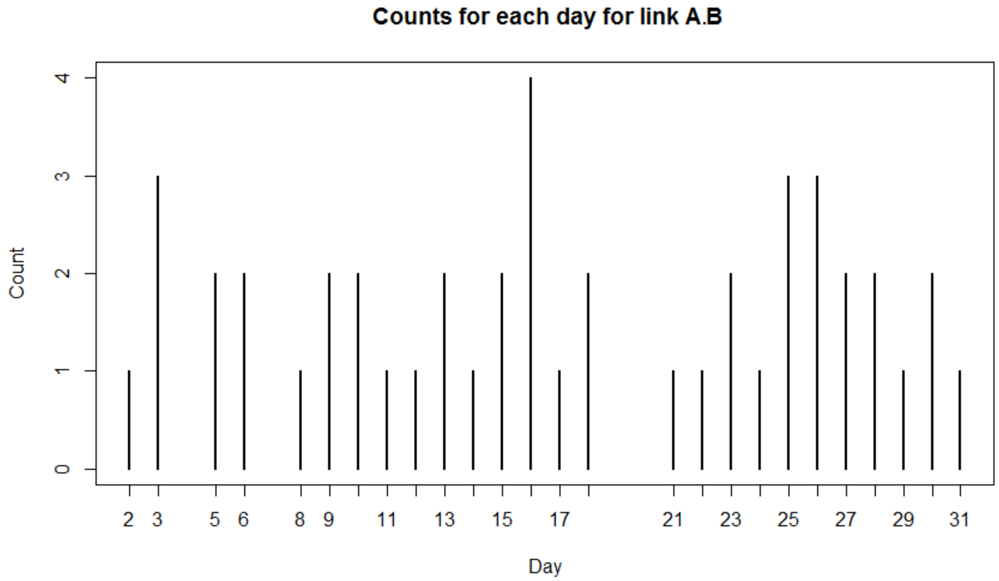


Figure 3.3.2: Cluster counts for each day for link A.B. There are no clusters on days without a line.

Poisson distribution with the same γ for all days, versus if we let γ_d be dependent upon the day d, in the month, where D is the set of all days in the month.

We have two hypothesis: H_0 is that the Poisson parameter is the same for all days and H_1 that the Poisson parameters vary for each day. Let D be the set of all the days in January. Then

$$\begin{aligned} H_0 :& \gamma_d = \gamma \in \mathbb{R}^+ && \forall d \in D \\ H_1 :& \gamma_d \in \mathbb{R}^+ && \forall d \in D. \end{aligned} \tag{3.3.9}$$

Let X_d be the number of occurrences of clusters on day d, then $X_d \sim \text{Poisson}(\gamma_d)$. The log-likelihood of observing X_d under the Poisson model is

$$l_d(\gamma_d, X_d) = -\gamma_d + x_d \ln(\gamma_d) - \ln(x_d!). \tag{3.3.10}$$

We want the log-likelihood of observing the counts in across the entire month. As the counts for each day are independent the full log-likelihood is therefore:

$$l((\gamma_d, X_d), d \in D) = \sum_{d \in D} l_d(\gamma_d, X_d). \tag{3.3.11}$$

The log-likelihood for the null hypothesis model, up to an arbitrary constant, then simplifies to:

$$l((\gamma, X_d), d \in D) = -n\gamma + \sum_{d=1}^{n} x_d ln(\gamma), \tag{3.3.12}$$

whereby n is the total number of days. The likelihood ratio test statistic is

$$\Gamma(X_d, d \in D) = -2\left(l((\hat{\gamma}, X_d), d \in D) - l((\hat{\gamma}_d, X_d), d \in D)\right). \tag{3.3.13}$$

When we calculate the test statistic with the thirty-one days in January in the set D we get a p-value of 0.72 and a test statistic of 25 with 30 degrees of freedom. This leads us to fail to reject the null hypothesis that the rate changes with time. This supports a homogeneous Poisson process. We conclude that there is no evidence to suggest that $\gamma(t)$ changes across time as t increases when compared against a hypothesis that it remains constant.

The rate, $\gamma(t)$, may also change dependent upon other factors such as the time of day or day of the week. We next look at cluster occurrence on a weekly scale and subset the data by the seven days of the week. This helps us to check that there are no differences between each day of the week.

Figure 3.3.3 shows that all days of the week have multiple clusters. The data are recorded in discrete time at 15-minute intervals. This means that when the clusters are overlaid to create a week map some clusters may occur in the same 15-minute interval. However they are very unlikely to occur at the same time within that interval so this is still appropriate for a Poisson process. To avoid losing the extra cluster from the data plot we record the clusters by the number rather than a cross. Over the month only twice does a cluster occur at the same weekly interval.

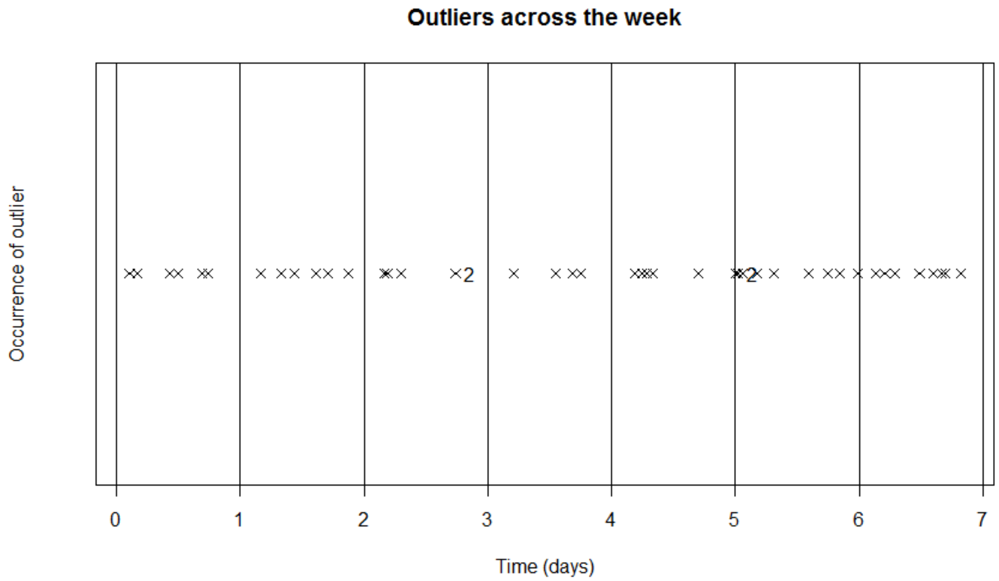


Figure 3.3.3: Outlier occurrence for a week, link A.B. A cross indicates a single cluster maxima and a number the count, if more than one cluster maxima shared the same fifteen minute interval of a week.

To test whether all the days in the week have the same rate γ we again use a likelihood ratio test where the alternative hypothesis has the rate be dependent upon the days of the week. The rate for day w is denoted by γ_w. Then

$$\begin{aligned} H_0 :& \gamma_w = \gamma \in \mathbb{R}^+ && \forall w \in W \\ H_1 :& \gamma_w \in \mathbb{R}^+ && \forall w \in W. \end{aligned} \tag{3.3.14}$$

The log-likelihoods for each day of the week can be found using equation (3.3.10), replacing the subscript d with w. Here x_w refers to the total number of clusters observed upon day of the week w.

The likelihood ratio test statistic (up to an arbitrary constant) is

$$\Gamma(x) = -2\left(l((\hat{\gamma}, X_w), w \in W) - l((\hat{\gamma}_w, X_w), w \in W)\right). \tag{3.3.15}$$

To evaluate this we need the maximum likelihood estimate $\hat{\gamma} = \sum_{i=1}^{7} \frac{x_w}{7}$. The degrees of freedom between the two models is six and the test statistic is 4.5 resulting in a p-value of 0.61. Hence we fail to reject H_0 and conclude that there is insufficient evidence that the days are different.

We finally look at the scale of one day to check if the clusters appear more frequently at particular times of the day. Figure 3.3.4 shows that the clusters appear throughout the day but over the thirty one days in January no more than three clusters ever have the same time in the day. A segregation of three hour intervals is shown on the picture.

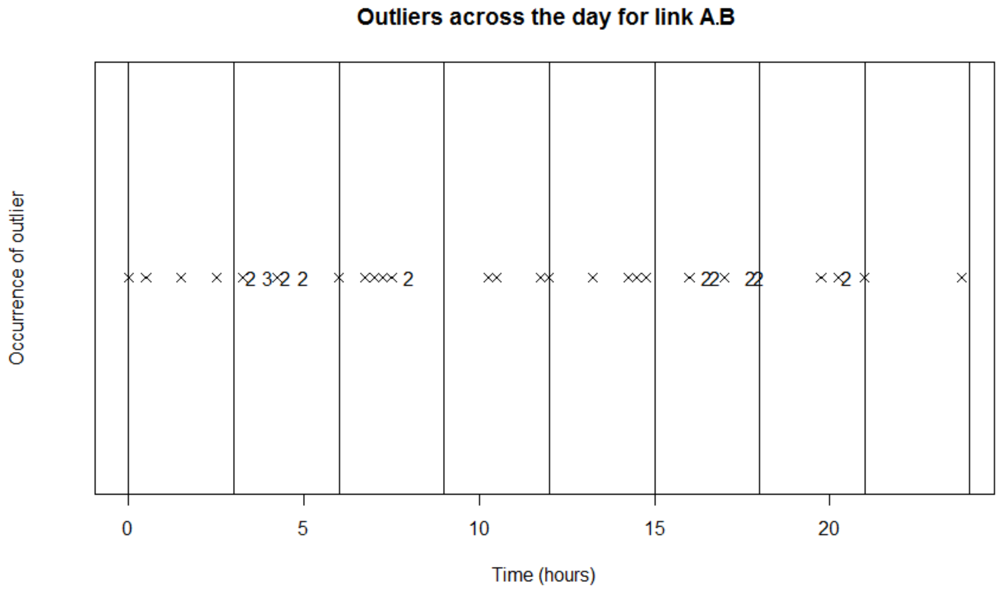


Figure 3.3.4: Outlier occurrence over the day for link A.B. The lines indicate the three hour segments.

We again use a likelihood ratio test to confirm whether the time of day affects the parameter γ. To reduce the number of different parameters we will split the day into 3 hour segments, each with 12 15-minute segments. As before our hypothesis are such that either γ is the same for all segments or that each segment s can have different

γ_s. The set S contains all 3 hour segments for one day. Then

$$\begin{aligned} H_0 :& \gamma_s = \gamma \in \mathbb{R}^+ && \forall s \in S \\ H_1 :& \gamma_s \in \mathbb{R}^+ && \forall s \in S. \end{aligned} \tag{3.3.16}$$

Using the previous method the likelihood ratio test statistic value is 10.4 with a p-value of 0.17. This suggests that at the 95% size of test there is insufficient evidence that the segments are different and we fail to reject H_0.

From the three likelihood ratio tests we conclude that there is insufficient evidence that $\gamma(t)$ varies with time or has different values for the time of day or day of the week and hence a homogeneous Poisson process with rate γ seems to be the best Poisson process model for this data. However these three tests aren't exhaustive of the things that could potential cause the rate to change and we may have missed an underlying feature.

For example, when considering the travel times, rather than the standardized series the greatest difference in times is during rush hour which is typically during the morning and the afternoon. During this time there are more cars on the road and therefore the rate of accidents may be higher.

Kernel density estimation. If $\gamma(t)$ is constant for all time and the process is a Poisson process then the occurrence times should be distributed in time according to a uniform distribution. Any large deviations from the uniform distribution will indicate features that we have failed to consider in the likelihood ratio tests.

One way to check this is to use kernel density estimation. This estimates the true density by smoothing the density from being concentrated at the observed points. If the resulting density plot is approximately constant the distribution is uniform and hence $\gamma(t)$ is constant for the range considered. Else the location of peaks and troughs suggest where the previously unconsidered features lie.

The smoothing effect depends upon the kernel used. There are different shapes of kernel and the width to which any point is smoothed over can also be adjusted. Common shapes include a Gaussian kernel which we will use, as well as a box or triangular shape.

A problem with using kernel density estimation on timed data is that the estimate performs poorly around the start and end of the period. Some of the density is lost because it is smoothed outside the region and clusters are only recorded within the period (Silverman, 1986). One suggested method to avoid losing the density is to reflect the missing density in the two edges. This ensures that no density is lost and that it is distributed close to where the points it came from were.

The resulting kernel density plot across time is shown in Figure 3.3.5. This is approximately uniform but suggests a possible slight increase in rate throughout the month. The red lines are 95% tolerance bounds calculated by sampling from the uniform distribution. The density estimation for each of these samples is calculated and the lines are the 95% tolerance bounds for each of the points that the density is estimated at. The error bounds are much larger at the start and end of the month because there is more information from neighbouring data there.

The kernel estimate line is comfortably within the tolerance bounds. To test this we will use another likelihood ratio test with the alternative hypothesis having two parameters - one for the first half and one for the second half. With a test statistic of 0.089 and a p-value of 0.79 we fail to reject the alternative hypothesis and conclude that $\gamma(t)$ is constant across time.

We also wish to check that the points are uniformly distributed throughout the week. The kernel density estimate should be independent of where the edge lies as Monday follows Sunday. Unlike before where we had no data outside the region a week is cyclical and hence the two edges of the region are connected. To avoid losing density from the flat representation of a week we can place the missing density onto the opposite edge which results in a continuous loop over the week.

The kernel density plot over the period of a week is shown in Figure 3.3.6. The

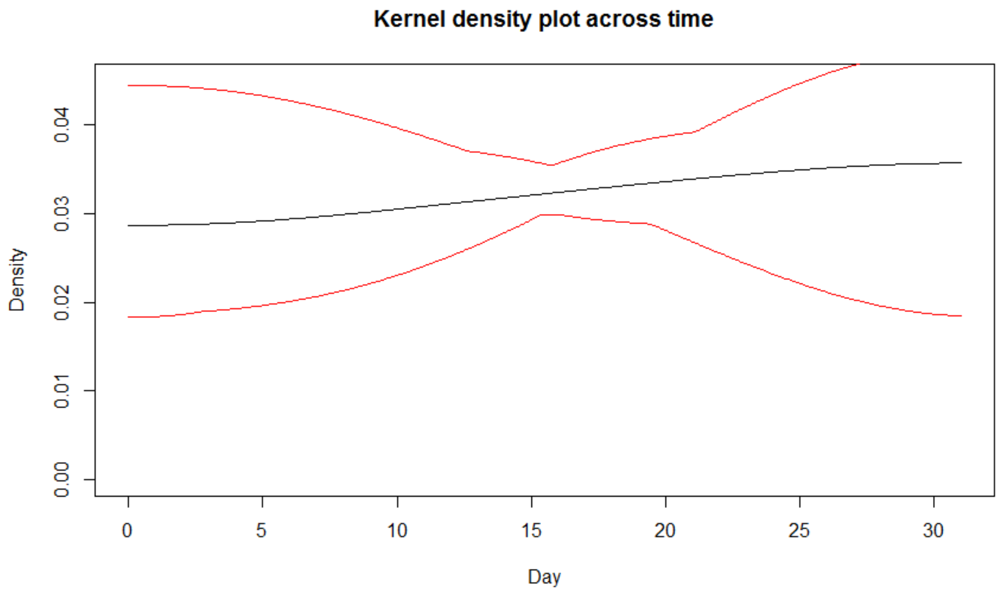


Figure 3.3.5: Kernel density plot of link A.B over month of January. The black line is the density plot, while the red are empirical tolerance bounds from an uniform distribution

plot is from Monday to Sunday. The cyclic method of calculating the density means that the tolerance bounds are approximately level through the week. The observed density lies within the tolerance bounds.

As with the plot over the month the plot is approximately uniform, with a potential slight dip in the middle of the week. This suggests a potential further test of week days and weekends being different.

The corresponding likelihood ratio test is where the alternative hypothesis is that γ is different for the weekends and the weekdays. This results in a test statistic of 3.4 and a p-value of 0.065 with one degree of freedom. Hence we conclude that there is insufficient evidence at the 5% confidence level that γ varies between the weekend and the weekdays and conclude that the occurrence rate is constant across the week.

The last uniformity check is for the time of day. A day is also cyclical so we can use the same method we used when checking the distribution of points over a week. Figure 3.3.7 shows the kernel density estimate for one day. The density line is nearly

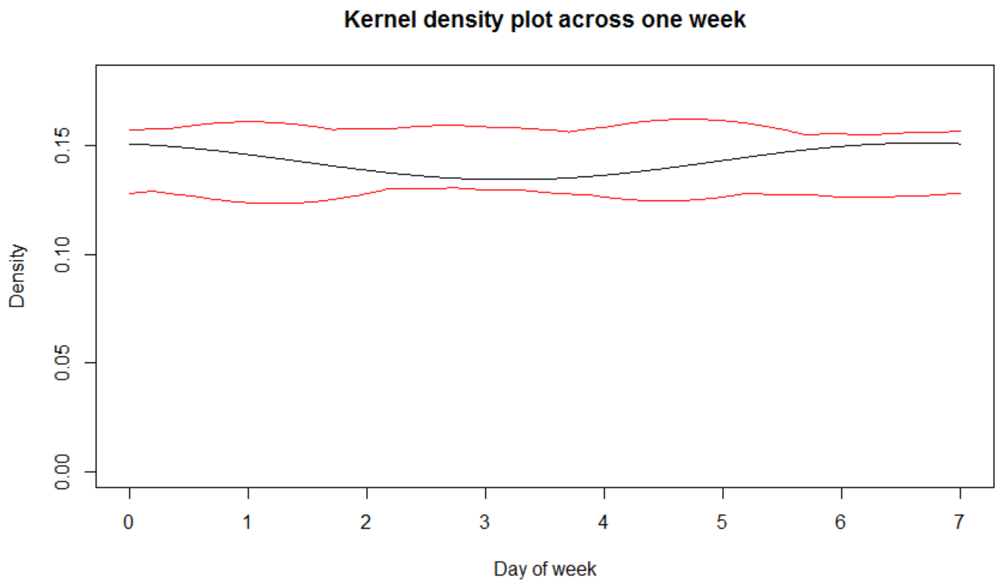


Figure 3.3.6: Kernel density plot of link A.B over a week with red uniform tolerance bounds.

horizontal with a very slight increase at about six and a slight decrease at about 18. As the density is clearly within the error bands of a uniform distribution we will conclude that we have no further need to test for differences across the time of day.

Between cluster time tests. If the cluster occurrences are a homogeneous Poisson process the times between clusters must have exponential distribution, with parameter γ, such that parameter remains the same for all time. Hence this must hold then for a homogeneous Poisson process. Our final test to check that γ is constant for all time within a homogeneous Poisson process is based upon checking this is true.

To test if they have an exponential distribution we must first estimate γ. Let n be the total number of clusters and I^c_j be the time of cluster j as before with I^c_0 being zero. Then we estimate γ by

$$\hat{\gamma} = \frac{n}{\sum_{j=1}^{n} I^c_j - I^c_{j-1}}. \tag{3.3.17}$$

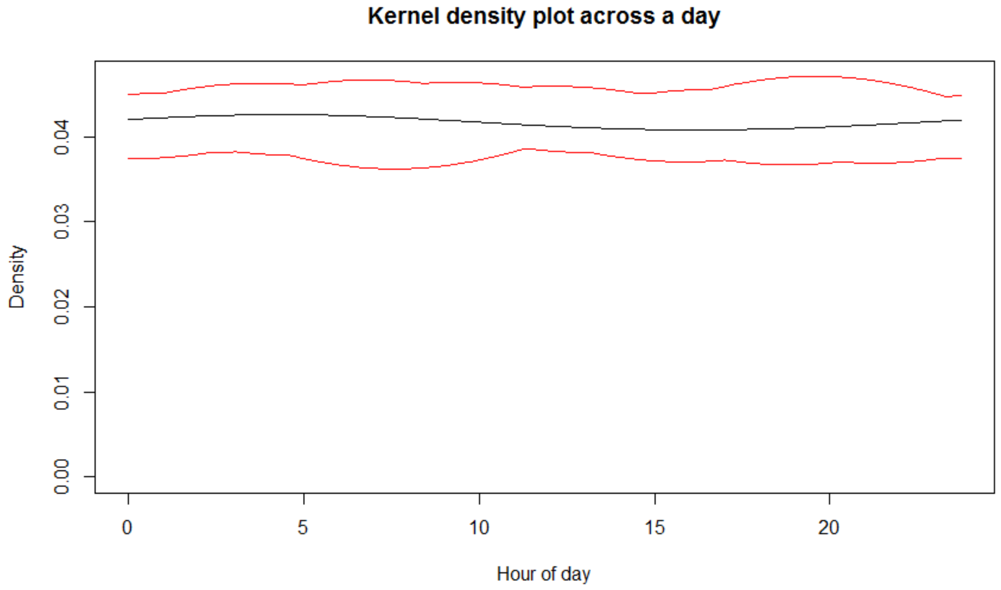


Figure 3.3.7: Kernel density plot of link A.B for a day with tolerance bounds for the uniform distribution in red.

This is the maximum likelihood estimator for γ and will give the same value as $\hat{\gamma}$ from the likelihood tests when measured upon the scale of fifteen minutes.

One way of checking if a set of points follows a distribution is to use a qq-plot. A qq-plot plots the observed quantile values in the series against the theoretical quantiles of an exponential distribution, with rate $\hat{\gamma}$. If the observed values come from the exponential distribution then they should form a straight line with gradient one. Some deviation from the straight line is to be expected, so to check that the deviation is acceptable tolerance bounds are needed.

The envelopes are generated by finding the highest and lowest values of the $i-th$ highest point from 200 samples, each of size n from Exponential($\hat{\gamma}$). This creates an empirical upper and lower bound which includes natural variation. If any point lies outside these bounds then there is evidence to suggest that the times between clusters are not exponentially distributed.

The plot in Figure 3.3.8 show the qq-plots of the wait times between clusters using the estimator $\hat{\gamma}$. The points are all within the envelope although there is some variation from the line.

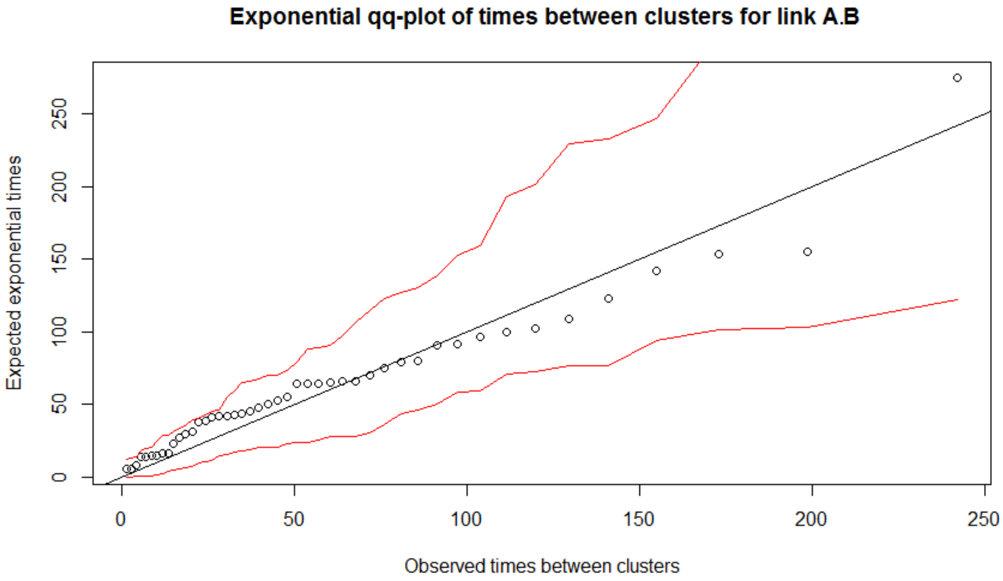


Figure 3.3.8: QQ-plot of wait times between clusters for link A.B with empirical confidence bounds in red.

Our data is only recorded discretely and the exponential distribution requires continuous time between the clusters. Instead an geometric distribution can be used. A geometric distribution is normally used to model the counts until a success or failure. Here we let a success be the occurrence of a cluster in the fifteen minute interval. Let W be the waiting time between two clusters and p be the probability of a cluster occurring, then

$$P(W = w) = (1 - p)^{w-1}p. \tag{3.3.18}$$

The geometric distribution has the same maximum likelihood estimator as the exponential distribution and the qq-plot is very similar as can be seen in Figure 3.3.9. All points lie within the exponential tolerance bounds. Given that the two plots both show that an exponential or geometric distribution is appropriate we can conclude that a homogeneous Poisson process is appropriate for link A.B.

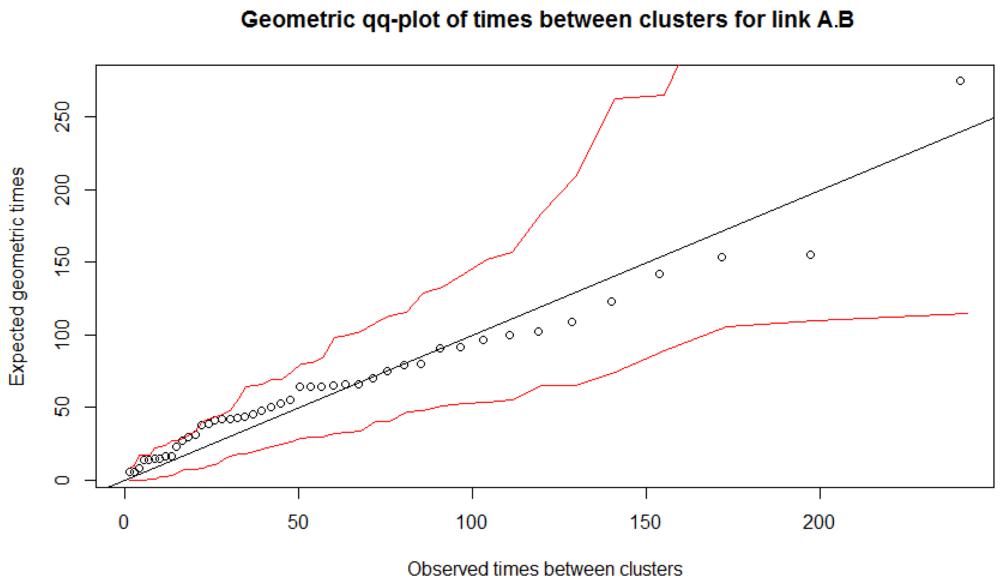


Figure 3.3.9: Geometric QQ-plot of wait times between clusters for link A.B with empirical bounds for the exponential quantiles shown in red.

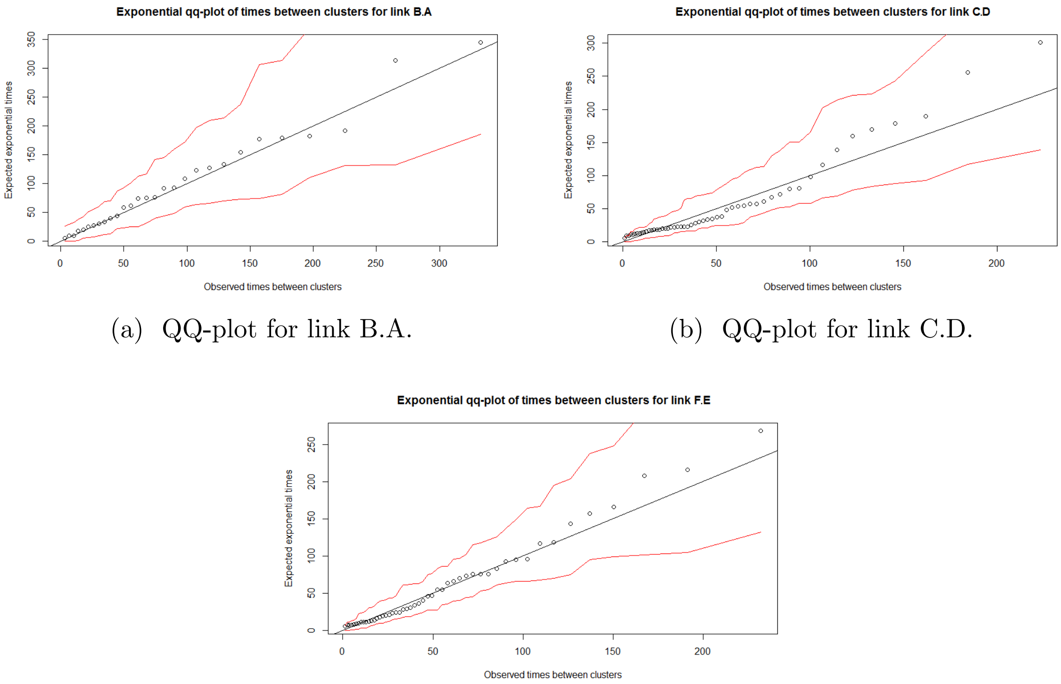
Conclusion of homogeneity. By combining the likelihood ratio tests of $\gamma(t)$, the kernel density plots and the waiting time distribution plots we conclude that, for link A.B, a homogeneous Poisson process is appropriate for the occurrence of clusters.

Other links

We now need to check if the other links also follow a homogeneous Poisson process. Given the links have different lengths and volumes it is likely that $\gamma(t)$ will change for the different links. If γ is constant for every link then the cluster occurrence follows a homogeneous Poisson process with rate γ_{ij} for link ij. The modelling of the effect of the delays upon the travel time then use similar models for all links which is much easier to implement across a large network.

We begin by looking at the qq-plots for the links B.A, C.D and F.E. These are shown in Figure 3.3.10. All the links appear to follow approximately a straight line although both C.D and F.E are close to the tolerance bounds when the time is small.

We next test the other links using the various likelihood ratio tests. When consid-



(a) QQ-plot for link B.A.

(b) QQ-plot for link C.D.

(c) QQ-plot for link F.E.

Figure 3.3.10: QQ-plot of wait times between clusters for links B.A, C.D and F.E with red empirical confidence bounds.

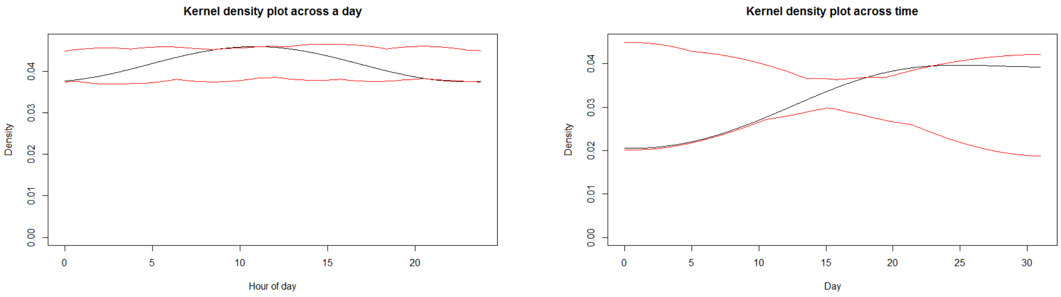
ering different γ_w for the days of the week the links C.D, F.E and A.B all fail to reject the null hypothesis that γ is constant. The corresponding p-values are 0.45, 0.74 and 0.28.

Link C.D rejects the null hypothesis that the rate is constant across time with a p-value of 0.022. Links B.A and F.E have a p-values of 0.51 and 0.32 which is evidence at the 5% level to fail to reject the null hypothesis and conclude that the rate doesn't change throughout the day.

The likelihood ratio tests of the change in $\gamma(t)$ over the hours in the day also result in different conclusions for the different links. Link B.A rejects the null hypothesis that $\gamma(t)$ is constant over the day at a 5% level, with a p-value of 0.0097 this is clearly rejected. Both links C.D and F.E fail to reject the null hypothesis at a 5% level. The p-value of link C.D is 0.87 and F.E is 0.43.

This implies a need to look in more detail at link B.A through the day and also link C.D through time. To do so we plot the corresponding kernel density plots. Figure 3.3.11a shows that the kernel density of link B.A over one day is much higher in the middle, between about 9 and 12. This is outside the tolerance bounds at the very peak. Conversely the density is lower than expected in the late evening/early morning from about 8pm to 1am. Similarly the plot for C.D across time in Figure 3.3.11b is also outside the tolerance bounds at around days 18-23 and is very close to the tolerance bounds throughout.

However when considering multiple links we would expect some variation from $\gamma(t)$ being constant for variations of t. Our overall aim is to be able to model the travel time over multiple links which requires us to model the extreme delays. Having different models for different links makes fitting the extreme delay models much more complicated and hence we accept the compromise that $\gamma(t)$ is constant for all time for all the links.



(a) Kernel density plot across day for link B.A.

(b) Kernel density plot across time for link C.D.

Figure 3.3.11: Kernel density plots where $\gamma(t)$ isn't constant with uniform tolerance bounds.

3.3.2 GPD model

Our clusters are defined by both the location and the cluster maxima. We have modelled the cluster occurrence in Section 3.3.1, however it is the values of the cluster maxima that determine the influence on the travel time. The cluster maxima can all be considered to be extreme values by their definition. There are a variety of different distributions that have been used for extreme values. One of these is the generalized Pareto distribution (GPD). The choice of the GPD as a model for extreme values will now be justified by considering how the events occur.

The following derivation applies for any arbitrary exceedances or cluster maximums, but we will discuss this with respect to the cluster maxima because that is what we are modelling. The values associated with the cluster are the cluster maxima, $Y_1^c, \ldots, Y_n^c$, where n is the total number of observed clusters. We assume that all the cluster maxima are independent and identically distributed. Because the travel time series has been standardized any variation due to the time of day should have been removed so they should be identically distributed. The cluster maxima should be independent because we have declustered the exceedances. Specifically, accidents shouldn't repeat and we should be able to anticipate congestion due to roadworks so each cluster is independent.

Let Y be a cluster maxima value. Then the occurrence times of Y are a homogeneous Poisson point process as justified in Section 3.3.1. We consider Y given it is above a threshold v. It should be noted that v is unrelated to u, the threshold for the clusters themselves which we have previously fixed in Section 3.2.3.

As Y are the cluster maxima then by definition they are bounded below by u, the cluster threshold level, and hence $Y > u$. If we choose $v < u$ then there may be isolated points in the standardized series that would be exceedances of the GPD threshold v but are not values of Y because they aren't in a clusters and hence can't be a cluster maximum. Hence because we are modelling the cluster maxima we require that $v \geq u$. We wish to find the probability that Y is above a certain value given it is above the threshold v. We consider this first asymptotically.

As we have assumed that the cluster maxima are i.i.d they must have some distribution $Y \sim F$ where F is an unknown distribution. Let the upper endpoint of F be denoted y^F. We take the limit as the threshold v tends to the upper endpoint. All values of Y will be positive as u is positive and $Y > u$. This gives the following result from Pickands (1971) which we can use to model Y. If we can find $c_v > 0$, a sequence that depends upon the threshold v such that, if, for $y > 0$,

$$\lim_{v \to y^F} P(Y > v + c_v y | Y > v) \to \bar{G}(y), \tag{3.3.19}$$

where the limit $\bar{G}(y) = 1 - G(y)$, with $G(y)$ a non- degenerate distribution function, then G must be of the form

$$G(y) = 1 - \left(1 + \xi \frac{y}{\sigma}\right)_+^{-\frac{1}{\xi}}, \tag{3.3.20}$$

with parameters $\sigma > 0$ and $\xi \in \mathbb{R}$. This distribution is called the generalized Pareto distribution or GPD, denoted GPD(σ, ξ). The notation $(y)_+ = \max(y, 0)$ defines the function as the inside of the bracket to be either zero or positive and hence the function (3.3.20) exists for all possible combinations of $\xi \in \mathbb{R}$, v, $\sigma > 0$ and y. In practice this defines the range of the distribution.

We assume that this limit (3.3.19) holds outside the limiting case where the threshold tends to the upper endpoint of F. So we assume that there exists a large enough $v < y^F$ such that probability that Y is greater than the level $v + c_v y$ given it is above the threshold is $G(y)$. This leads to the model for $y > 0$ of

$$P(Y > v + c_v y | Y > v) = \left(1 + \xi \frac{y}{\sigma}\right)_+^{-\frac{1}{\xi}}. \tag{3.3.21}$$

Replacing $c_v y$ with x implies $x > 0$ because c_v and y must be positive. This gives

$$P(Y > v + x | Y > v) = \left(1 + \xi \frac{x}{c_v \sigma}\right)_+^{-\frac{1}{\xi}}, \tag{3.3.22}$$

for $x > 0$. We can then replace $c_v \sigma$ with σ_v as c_v is just a scaling constant that depends upon the threshold choice and once v is fixed $c_v \sigma$ is a constant and thus σ_v

is a parameter of the distribution. Thus, for a large enough threshold v

$$Y - v|Y > v \sim GPD(\sigma_v, \xi). \tag{3.3.23}$$

The generalized Pareto distribution has two parameters; the shape, ξ, and the scale, $\sigma_v > 0$, as well as the threshold level v. The distribution is defined such that for the threshold v, for any values of Y that are above the threshold v then $Y - v$ has a GPD distribution.

The probability density function of the GPD is

$$g(y; \xi, v, \sigma_v) = \begin{cases} \frac{1}{\sigma_v}\left(1 + \frac{\xi(y-v)}{\sigma_v}\right)_+^{(-\frac{1}{\xi}-1)}, & \text{if } \xi \neq 0 \\ \frac{1}{\sigma_v} e^{-\frac{y}{\sigma_v}}, & \text{if } \xi = 0 \end{cases} \tag{3.3.24}$$

where $\sigma > 0$ and $\xi \in \mathbb{R}$.

The mean of the GPD is,

$$v + \frac{\sigma_v}{1 - \xi} \quad \text{if } \xi < 1. \tag{3.3.25}$$

The mean only exists if $\xi < 1$ and as we are modelling the travel times along a road it is reasonable to assume the mean exists. This implies that we should choose a value of ξ that has a finite mean and hence $\xi < 1$.

The second moment is

$$v^2 + 2v\frac{\sigma_v}{1 - \xi} + \frac{2\sigma_v^2}{(1 - \xi)(1 - 2\xi)} \quad \text{if } \xi < 1/2. \tag{3.3.26}$$

This leads to a variance of

$$\frac{\sigma_v^2}{(1 - \xi)^2(1 - 2\xi)} \quad \text{if } \xi < 1/2. \tag{3.3.27}$$

The distribution has several desirable properties. The tail is such that small values are much more likely to occur than larger values, while the likelihood of the higher values depends upon the parameters ξ and σ. When the shape parameter is zero this condenses to an exponential distribution, by taking the limit $\xi \to 0$. The value of the ξ parameter affects how the distribution behaves. The larger ξ is, the more weight there is in tail of the distribution.

When $\xi \geq 0$ then there is no upper limit and Y is unbounded above ($Y \geq v$). However, if ξ is negative, then the limit in terms of the standardized series is $v - \frac{\sigma_v}{\xi}$ such that $v \leq Y \leq v - \frac{\sigma_v}{\xi}$. Travel time delays can theoretically be infinite, or at least there is no obvious upper limit, which suggests that a choice of $\xi \geq 0$ is appropriate. Combining this with having a finite mean estimate leads to $0 \leq \xi < 1$ being most appropriate.

One very useful property of the GPD is the threshold stability property (Eastoe and Tawn, 2009). Suppose that for a fixed threshold level of $\tilde{v}$, that $Y - \hat{v}|Y > \tilde{v}$ has a GPD distribution with scale parameter $\tilde{\sigma}_{\tilde{v}} > 0$ and shape parameter $\tilde{\xi}$, i.e.,

$$Y - \tilde{v}|Y > \tilde{v} \sim GPD(\tilde{\sigma}_{\tilde{v}}, \tilde{\xi}). \tag{3.3.28}$$

Then the threshold stability property states that for any other threshold level v, such that $v > \tilde{v}$, the random variable $Y - v|Y > v$ also has a generalized Pareto distribution with scale parameter $\sigma_v = \tilde{\sigma}_{\tilde{v}} + \tilde{\xi}(v - \tilde{v})$ and shape parameter $\xi = \tilde{\xi}$, i.e.,

$$Y - v|Y > v \sim GPD(\sigma_v, \xi). \tag{3.3.29}$$

However this only holds for any possible threshold values $v > \tilde{v}$ above the current threshold $\tilde{v}$ and makes no guarantees about what the distribution is for any threshold values that would be lower. There exists one threshold $\hat{v}$ such that $Y - \hat{v}|Y > \hat{v} \sim GPD$ but for any $v < \hat{v}$ the distribution of $Y - v|Y > v$ is no longer GPD. The threshold stability property also holds here such that $v > \hat{v}$ is GPD and this $\hat{v}$ is a lower limit.

To fit the distribution we need to select one threshold level v, from which we can then find the best choices for ξ and σ_v. The threshold stability property means that there are many different appropriate choices available for the threshold level v. We wish to pick v as small as possible, while still ensuring the distribution is GPD as this ensures more data to fit the model. The more data to fit the distribution the closer the parameter estimates will be to the true values.

This gives the preferred choice for the threshold v to be $\hat{v}$ and we therefore wish to find $\hat{v}$. If $\hat{v}$ is small enough such that $Y - \hat{v}|Y > \hat{v} \sim GPD$ can be fitted well to the data, then the GPD is an appropriate model for the data.

3.3.3 Checking suitability of GPD model for link A.B

As we have just discussed, fitting a GPD model requires the selection of an appropriate threshold, which is itself dependent upon the threshold level to select the cluster maxima. The other parameters can be fitted by maximum likelihood once the threshold for the GPD model has been chosen. We first perform this analysis on link A.B.

The threshold stability plot will give a lower value for the threshold from the point of view of the GPD and the plot of the standardized series across the week will enable us to see how this compares to u and the outliers across the travel time.

Threshold stability plot. To identify $\hat{v}$ we can use a threshold stability plot. A threshold stability plot is a plot of the differing thresholds against the corresponding mean excess over the threshold. As the threshold stabilises the relationship between the threshold and the mean exceedance above the threshold of the cluster maxima should become linear. Hence the lowest v where the relationship is is linear is our choice for $\hat{v}$.

If the GPD is an appropriate model the mean is given by equation (3.3.25). The mean exceedance above the threshold is therefore the mean take away the threshold. The relationship between the scale parameters σ_v and $\sigma_{\tilde{v}}$ for two different thresholds v

and $\tilde{v}$ is $\sigma_v = \tilde{\sigma}_{\tilde{v}} + \xi(v - \tilde{v})$. The mean exceedance above the threshold, v, is therefore

$$\frac{\tilde{\sigma}_{\tilde{v}} + \xi(v - \tilde{v})}{1 - \xi}. \tag{3.3.30}$$

The mean exceedance above the threshold therefore varies linearly in v with gradient $\frac{\xi}{1-\xi}$.

Our previous analysis in Section 3.2.3 of the cluster maxima threshold led us to chose $u = 2$. The lowest possible value of $\hat{v}$ is therefore 2, as $v \geq u$. Hence we start the threshold stability plot at 2. The 95% confidence bounds are obtained empirically by bootstrapping with replacement all cluster maxima above the proposed threshold v. The plot ends just before there is only one exceedance as the confidence bounds would constrict to width zero.

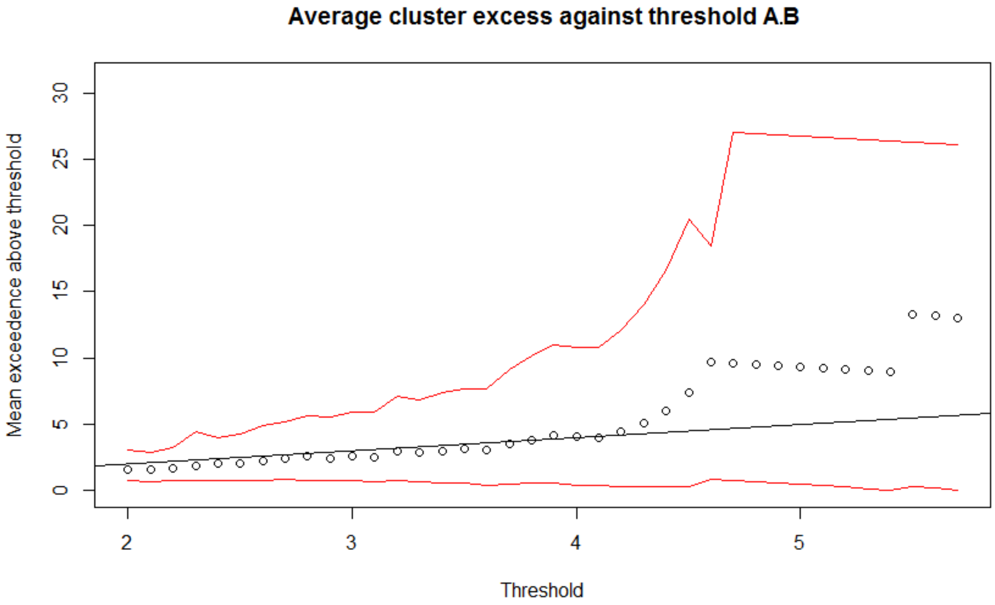


Figure 3.3.12: Threshold stability plot for link A.B. Bootstrapped 95% confidence bounds are shown in red which clearly enclose the illustrative line with gradient one, to indicate linearity of the points.

The threshold stability plot in Figure 3.3.12 shows that the mean exceedance above the threshold is gradually increasing. This is at a fairly steady rate from

2 until 4.3. As we don't yet have an estimate for ξ, we can't plot a line with the expected relationship between the threshold and the mean excess above the threshold. Hence, for illustrative purposes, a line with gradient 1 has been added to allow a comparison to a straight line. By eye a gradient of one is a reasonable approximation in the region from 2 to 4.3. Past 4.3 the threshold increases non linearly then increases and decreases in jumps due to the number of clusters maxima that are still exceedances being much lower. The steady deceases are because the values of the cluster maxima no longer occur between every point that the threshold is considered at. When there are fewer $Y - v$ then each cluster maxima is proportionally more of the mean exceedance above the threshold. Hence the increase is more pronounced, when the remaining cluster maxima values are much higher. As the line is within the 95% confidence bounds for the entire region this suggests that any value of v from 2 to 4.3 would be an appropriate choice.

Threshold choice. The threshold stability plot leads us to a range of $2 \geq v \geq 4.3$ for the GPD. As $u = 2$ then all the thresholds are cluster maxima values and hence can be considered as v as discussed in Section 3.3.2. However as $\hat{v}$ is the smallest value for which a GPD distribution is appropriate we select $\hat{v} = 2$. For simplicity, as $\hat{v} = u$, we now refer to the common threshold for both the cluster maxima and the GPD as the parameter $\hat{v}$.

We now confirm that the threshold of $\hat{v} = 2$ is appropriate given the standardized series. Figure 3.3.13 shows the standardized series for link A.B by day. It is from this series that the clusters are selected. The horizontal lines at 2 and -2 show that the majority of points are between 2 and -2, however we are only interested in positive exceedances. The maximum outlier is 31.7 and there are very few outliers above 5. There are enough cluster maxima from the corresponding clusters to fit a GPD where $v = 2$.

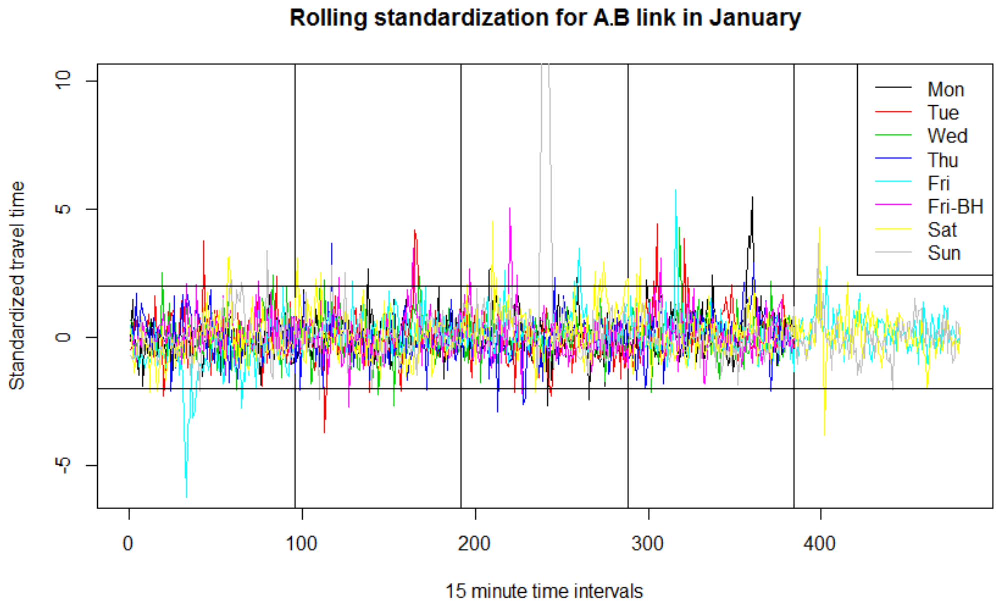


Figure 3.3.13: Standardization for link A.B.

GPD model. Having selected a minimum threshold of $\hat{v} = 2$ we now need to check if the GPD model is appropriate. If it is not then we can increase the threshold for the GPD, while u remains the same and refit the model.

The simplest GPD model to fit is one where both the shape and the scale parameter are the same for all time. However, as with the cluster occurrence models, the parameters may also change with time or other factors. The greater the number of parameters the more difficult it is to fit a model well because there is greater uncertainty around the parameters. Thus fewer models can be fitted from the same number of cluster maxima compared to the Poisson process models which have only one main parameter, $\gamma(t)$.

To reduce the number of parameters we consider using a common shape parameter and then allow the scale parameter to vary with time or the other factors (Heffernan and Tawn, 2001). This assumes that the different factors only affect the scale of the extreme delays rather than the underlying shape of the distribution. This has been found to be the case when modelling other extreme events such as extreme rainfall and wave heights as long as they where similar (Nadarajah et al., 1998). We therefore initially assume that a common shape parameter is appropriate and if the models perform poorly we will then allow the shape parameter to change.

Our previous tests over model suitability were based upon ensuring the properties of the model held true for the data or comparing two alternative models to each other with likelihood ratio tests. Due to the additional parameter in the GPD we will begin by considering simpler models and then checking if additional parameters are required.

In the Poisson process models in Section 3.3.1 we allowed the parameters to vary with time in different ways. These were through time, characterised by the individual days of the month having a unique mean, having the days of the week each with a different parameter and finally varying throughout the day so the time of the day effected the parameter value. Here there are too little data to fit parameters for each day of the month so we instead focus upon the remaining two ways of allowing the parameters to change.

The distribution is written as $GPD(\sigma_{\hat{v}}, \xi)$, however because we have chosen $\hat{v} = 2$ as the threshold we will drop the $\hat{v}$ from the $\sigma_{\hat{v}}$ to simplify the notation. As in Section 3.3.1 we have W as the set of the days of the week and S as all three hour segments in a day. Then we can characterise our three potential models as:

$$\text{Model } 0: \qquad (\sigma_t, \xi_t) = (\sigma, \xi) \qquad \forall t \tag{3.3.31}$$

$$\text{Model } 1: \qquad (\sigma_t, \xi_t) = (\sigma_w, \xi) \qquad w \in W \tag{3.3.32}$$

$$\text{Model } 2: \qquad (\sigma_t, \xi_t) = (\sigma_s, \xi) \qquad s \in S, \tag{3.3.33}$$

whereby (σ_t, ξ_t) are the parameters of the GPD at time t.

Model 0 is therefore the one where the shape and the scale are the same, it is the null model and the simplest. The first alternative model has a different scale parameter for each day of the week. The shape parameter remains the same irrespective of the day of the week. Similarly for the second alternative model, Model 2, the shape parameter is also constant for all time. The scale parameter changes for every three hour segment.

The shape parameter effects the overall behaviour of the distribution, whereas the scale affects the intensity. Thus it makes more sense for the scale parameter to change, as one day of the week may have more vehicles and therefore delays are longer.

We first fit GPDs to the data for Model 0 and Model 1. The parameter values can be compared, as can the likelihood ratio statistic to see which is the better model. We will then do the same for Model 2 which can only be compared to Model 0 as the two sets aren't subsettable. If two sets are subsettable then all the parameters for one are contained in the set of parameters for the other. Days of the week and time of the day are such that the values on one day of the week will have different times of day and vice versa. Hence the two models aim to model different features, both of which may be important and comparing the two together will only show if one is much more important than the other. If they are both relevant then the difference in the log-likelihoods is likely to be small and hence a likelihood ratio test will fail to give the required result as to whether or not the extra parameters should be included. After this we look at the general fitting of the GPD models to the data.

To fit the parameters we use maximum likelihood estimation. Under Model 0 the log-likelihood of a single point in a GPD distribution is,

$$l_0(\sigma, \xi) = -\log(\sigma) - \left(1 + \frac{1}{\xi}\right) \log\left(1 + \frac{\xi(X - v)}{\sigma}\right). \qquad (3.3.34)$$

Let n be the total number of clusters for the threshold u. Hence the log-likelihood of observing the cluster maxima, $(X_1, \ldots, X_n)$ is

$$l_0(\sigma, \xi) = -n\log(\sigma) - \left(1 + \frac{1}{\xi}\right) \sum_{j=1}^{n} \log\left(1 + \frac{\xi(X_j - v)}{\sigma}\right). \qquad (3.3.35)$$

Model 0 generates values for the parameters estimates of $\hat{\xi} = 0.432$, $\hat{\sigma} = 0.788$. The value of ξ, the shape parameter, is between the bounds of zero and one as identified in the previous section as being necessary for a GPD model that has physical sense.

We compare this with Model 1 which has different scale parameters for each day. We now subset the cluster maxima X by their day of the week such that $X_{w,j}$ is the jth cluster maxima on day of the week w.Thus the log-likelihood is now

$$l_1(\sigma_w, \xi, w \in W) = \sum_{w=1}^{7} \left(-n \log(\sigma_w) - \left(1 + \frac{1}{\xi}\right) \sum_{j=1}^{n_w} \log\left(1 + \frac{\xi(X_{w,j} - v)}{\sigma_w}\right)\right), \tag{3.3.36}$$

whereby n_i is the number of clusters on day of the week i.

Figure 3.3.14 shows the different parameter estimates with the corresponding parameter estimates from Model 0. The seven days of the week are indicated by their respective start letter as a subscript. Each parameter estimate is surrounded by the delta method based 95% confidence intervals.

The scale parameter estimate from Model 0 is inside all the confidence bounds for the differing day of the week scale parameter estimates and all estimates are close to each other. The estimate for ξ from Model 0 is very close to the shape parameter estimate suggested by Model 1. This suggests that the two models are similar but we will perform a likelihood ratio test to check if this is true.

Our two hypothesis are

$$\begin{aligned} H_0 :&(\sigma_t, \xi_t) = (\sigma, \xi) && \forall t \\ H_1 :&(\sigma_t, \xi_t) = (\sigma_w, \xi) && w \in W. \end{aligned} \tag{3.3.37}$$

Let T be the set of all times where a cluster maxima is observed. There are six degrees of freedom between the two models. The likelihood ratio test statistic is therefore

$$\Gamma = -2\left(l_0((\hat{\sigma}, \hat{\xi}, X_t), t \in T) - l_1((\hat{\sigma}_w, \hat{\xi}, X_t), w \in W, t \in T)\right). \tag{3.3.38}$$

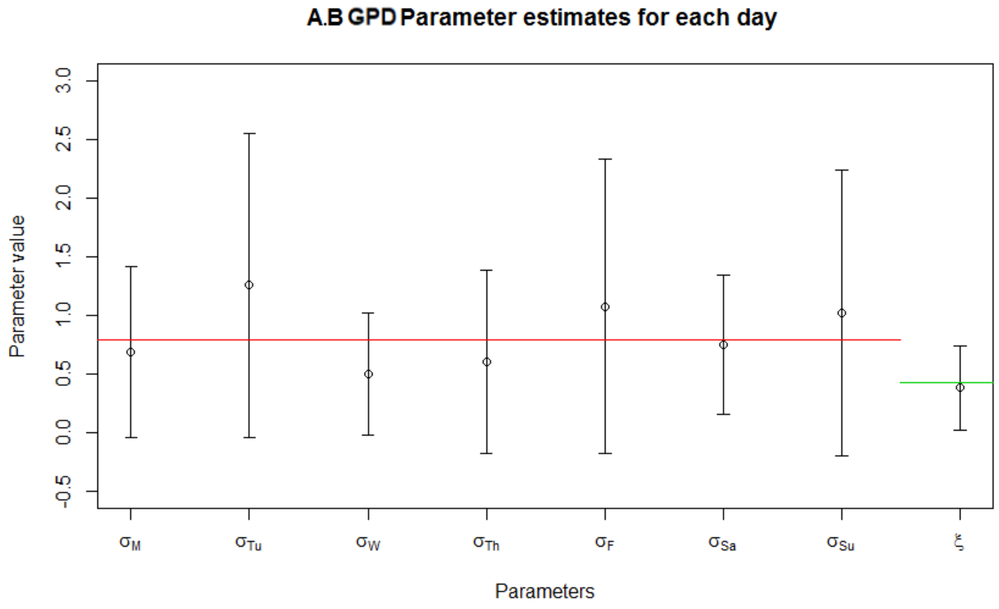


Figure 3.3.14: All parameter estimates for Model 1, with σ_w and ξ. $\hat{\sigma}$ (Model 0) = red line, $\hat{\xi}$ (Model 0) = green line estimate. The $\hat{\sigma}$ (Model 0) estimate lies wholly within the respective confidence intervals for the Model 1 parameter estimates, as does $\hat{\xi}$ (Model 0) within the 95% intervals for $\hat{\xi}$ (Model 1).

This gives a test statistic of 2.22 which has a p-value of 0.899. Therefore we would fail to reject the null hypothesis at the 5% significance level and conclude that there is insufficient evidence that the scale parameters vary by day of the week.

We next check whether the scale parameter varies throughout the day. This is Model 2 and we compare it directly to Model 0. Model 2 divides the day into eight segments, s, each of length three hours. Let $X_{s,j}$ be the jth cluster maxima of segment s with there being n_s cluster maxima in segment s. The log-likelihood for Model 2 is

$$l_2(\sigma_s, \xi, s \in S) = \sum_{s=1}^{8} \left(-n \log(\sigma_s) - \left(1 + \frac{1}{\xi}\right) \sum_{j=1}^{n_s} \log\left(1 + \frac{\xi(X_{s,j} - u)}{\sigma_s}\right)\right). \quad (3.3.39)$$

The null hypothesis, as before, is that Model 0 is the most appropriate and the alternative is that Model 2 fits the data better. Our two hypothesis are

$$\begin{aligned} H_0 :&(\sigma_t, \xi_t) = (\sigma, \xi) && \forall t \\ H_1 :&(\sigma_t, \xi_t) = (\sigma_s, \xi) && s \in S. \end{aligned} \tag{3.3.40}$$

The likelihood ratio test statistic is therefore

$$\Gamma = -2 \left(l_0((\hat{\sigma}, \hat{\xi}, X_t), t \in T) - l_2((\hat{\sigma}_s, \hat{\xi}, X_t), s \in S, t \in T) \right). \tag{3.3.41}$$

There are seven degrees of freedom between the two models. This results in a test statistic of 10.6 and a p-value of 0.155. As with having different scale parameters for different days we fail to reject the null hypothesis at the 5% significance level and conclude there isn't enough evidence that the scale parameter changes with the time of day.

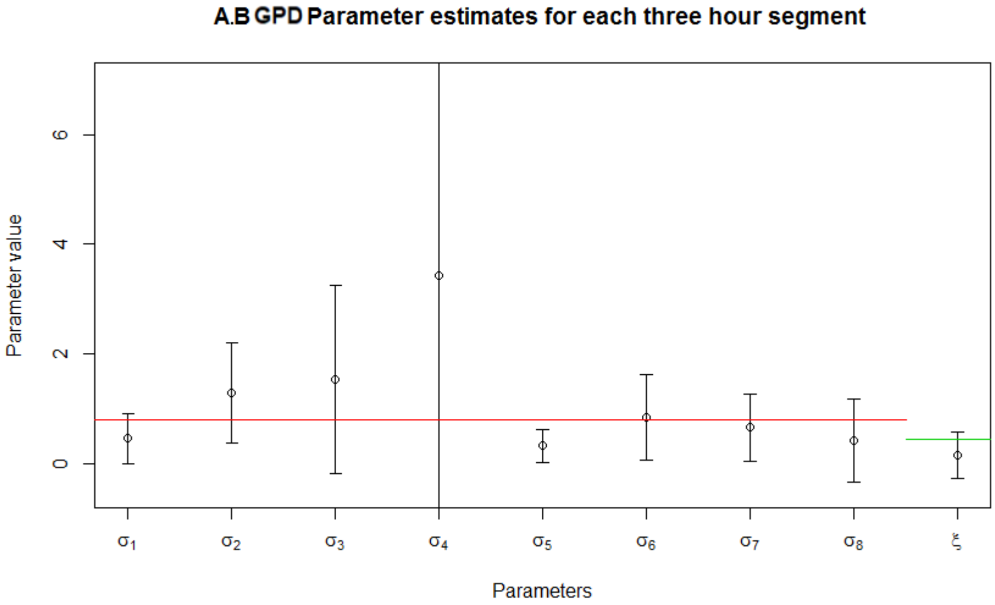
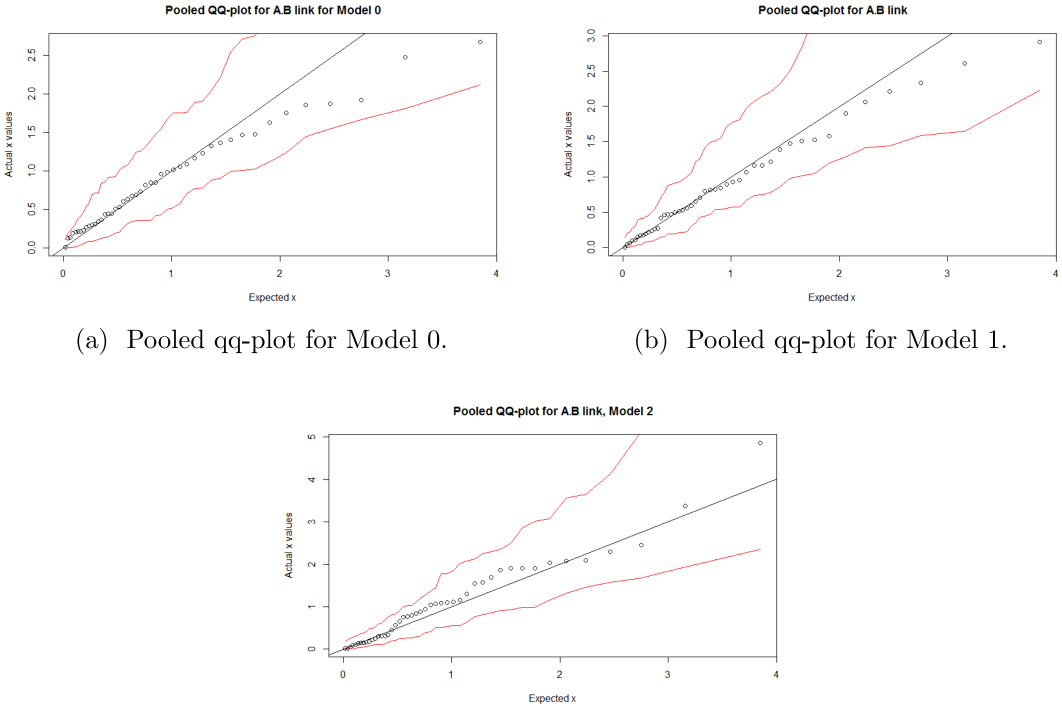


Figure 3.3.15: Scale and shape parameter estimates for Model 2 with 95% delta method confidence intervals. The ξ parameter estimate from Model 0 is shown in green. The red line shows the scale parameter estimate from Model 1.

As before we can examine the differing parameter estimates. The ξ estimate and its confidence interval are both smaller than for Model 1. However the ξ estimate from Model 0 is still within the confidence interval. There is much more variation in the scale estimates for Model 2 compared to Model 1. The estimate for σ_5 is such that the confidence interval doesn't contain the σ estimate from Model 0.

We can check that the GPD is a suitable model for the cluster maxima by using qq-plots, as in the previous subsection. We generate plots for each of the three models which are shown in Figure 3.3.16. These are plotted on an exponential scale by using a probability integral transform.



(a) Pooled qq-plot for Model 0.

(b) Pooled qq-plot for Model 1.

(c) Pooled qq-plot for Model 2.

Figure 3.3.16: Pooled qq-plots for the three different GPD models for link A.B. Model 0 has only one scale parameter, Model 1 has different scale parameters for each day of the week and Model 2 for different times of the day. The delta method confidence intervals are shown in red in addition to the line if the two quantiles were the same.

The model which is closest to the straight line is Model 2 as can be seen in Figure 3.3.16c. However both of the qq-plots for Model 0 and Model 1 are within the 95%

tolerance bounds. Figure 3.3.16a shows that Model 0 deviates the most, especially at the start and end of the range.

3.3.4 Other roads

The link A.B is only one road and we wish to check that a GPD model is appropriate for all roads across the network. We therefore repeat the checks in Section 3.3.3 for link B.A, F.E and C.D.

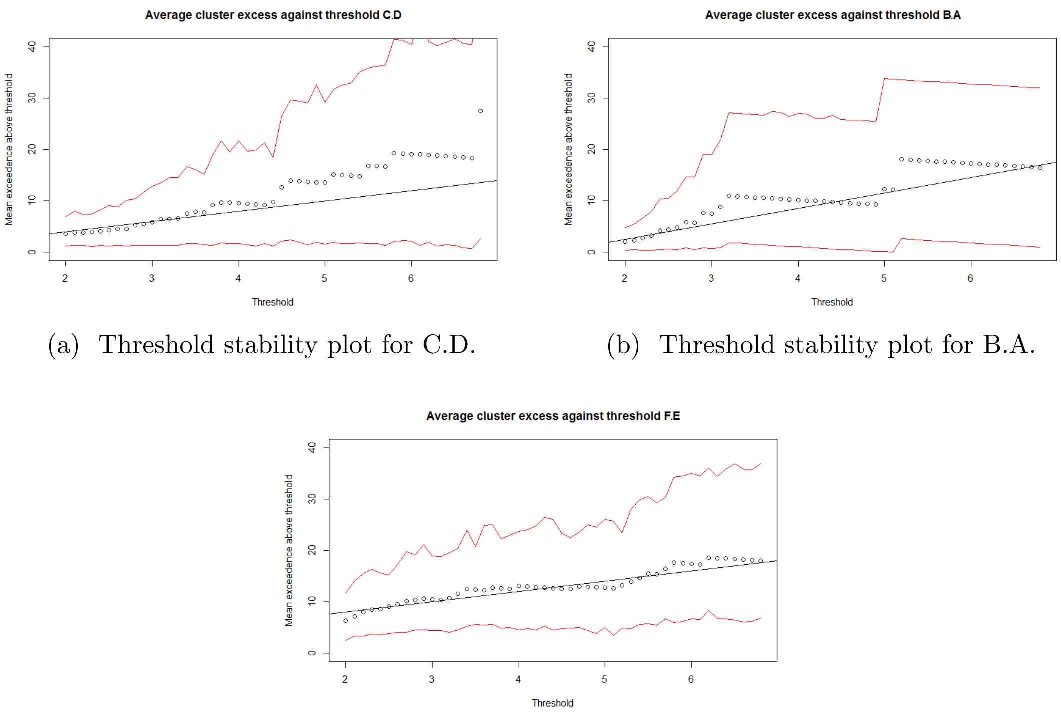
We first check that $\hat{v} = 2$ is appropriate for all three links with the threshold stability plot. The three links are shown in Figure 3.3.17 with illustrative lines to check the linearity. These lines have different gradients to the link A.B and each other. Links C.D and F.E both have gradient 2 and link B.A has gradient 3. This suggests that the shape parameter may be different for the different links. All three lines fit within the 95% confidence bounds. All three are approximately linear at $v = 2$, although link F.E is slightly under the illustrative line suggesting that $\hat{v}$ may be slightly above 2. However considering this line is only illustrative and the difference is small then this is probably natural variation.

The threshold for B.A remains stable for a much shorter period of time. There are only 23 total clusters observed compared to about 40 on the other three links. There are fewer larger values and hence the plot becomes unstable sooner.

We therefore proceed with the likelihood tests assuming that $\hat{v} = 2$ is true. As with link A.B we compare both Model 1 and Model 2 with Model 0, the GPD with a single scale and shape parameter.

To begin we fit GPDs for each link with only one scale and one shape parameter. This is Model 0. Table 3.3.1 shows the differences between the GPD parameters for the three links, with link A.B included for comparison. The shape parameters vary for each link and for F.E is above the limit of 1 that we identified for ξ in Section 3.3.2. The scale parameters also vary.

Given these differences in values for Model 0 we now need to check whether Models



(a) Threshold stability plot for C.D.

(b) Threshold stability plot for B.A.

(c) Threshold stability plot for F.E.

Figure 3.3.17: Threshold stability plots for C.D, B.A and F.E with possible gradient lines to allow stability checks. The delta method confidence intervals are shown in red.

Link	σ	ξ
A.B	0.79	0.43
B.A	0.39	0.76
C.D	1.1	0.66
F.E	0.93	1.3

Table 3.3.1: The shape and scale parameters from Model 0 are presented for each of the four links.

1 and 2 for the other links are significant. To do this we use a likelihood ratio test.

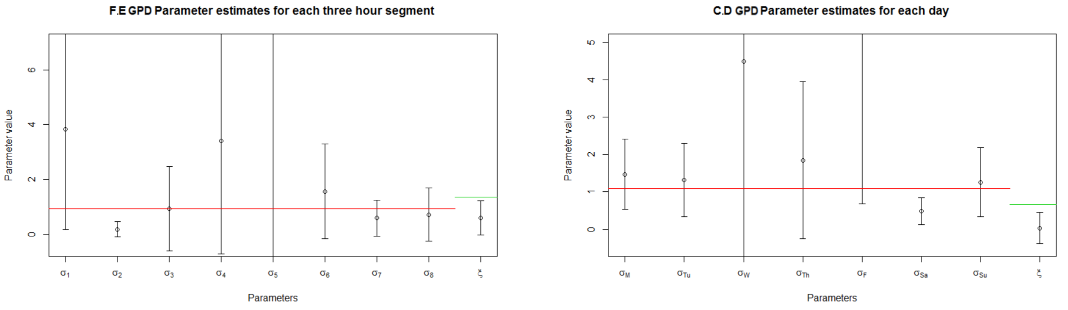
For link B.A we use a likelihood ratio test to compare Models 0 and 1. The test statistic is 2.58 with a p-value 0.859. The comparison between Models 0 and 2 has a slightly lower test statistic of 6.65 and a p-value of 0.466. Hence we conclude that neither Model 1 nor Model 2 are significant at the 5% significance level to Model 0.

For link C.D the test between Model 0 and Model 1 is significant at the 5% level with a p-value of 0.00260 and a test statistic of 20.2. The ξ value is very low at 0.032

to 3dp. The test for the time of day however is not significant at the 5% significant level. The p-value is 0.0737 and the test statistic 12.9.

When comparing Models 0 and 1 for link F.E with a likelihood ratio test the test statistic is 7.96 and the p-value is 0.241. For Models 0 and 2 the test statistic is 19.2 and the p-value is 0.00768. Hence we would reject a difference between Models 0 and 1 but not between 0 and 2 at the 5% significance level.

We therefore wish to look at the parameter estimates for Model 1 for link C.D and Model 2 for link F.E. These are shown in Figure 3.3.18. For both the shape parameter is outside the confidence intervals. For link F.E this means that ξ is now between 0 and 1 however for link C.D the estimate is only just above 0. Both models have one scale estimate which is very large. In link C.D σ_F is 17.3 and for link F.E σ_5 is 12.9.



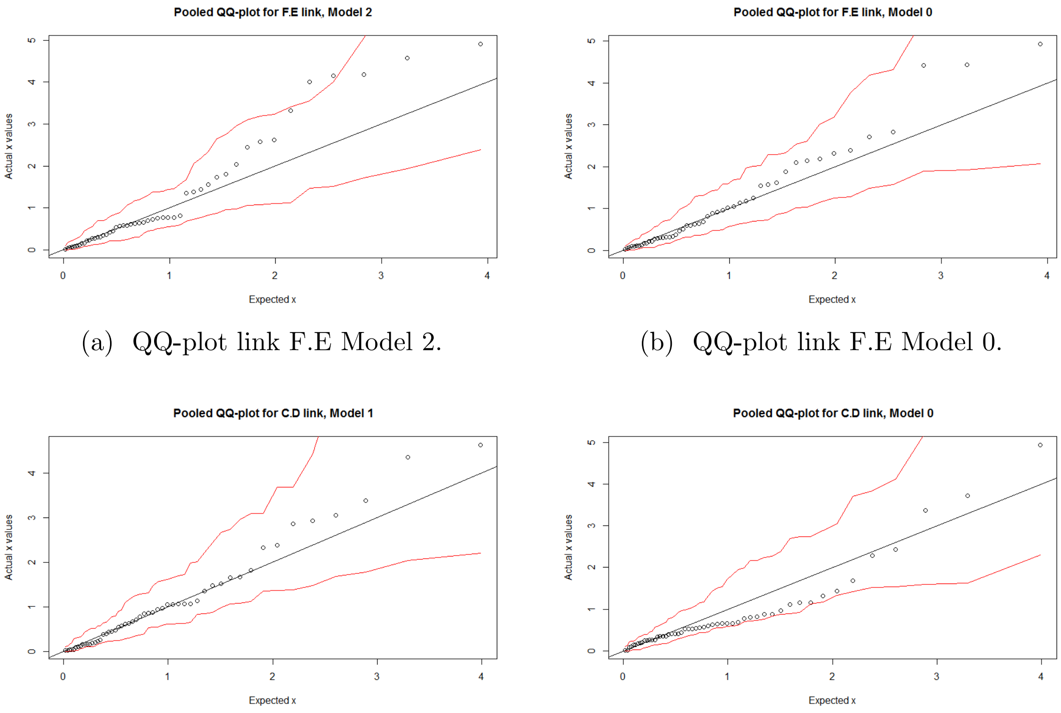
(a) Parameter estimates for link F.E Model 2.

(b) Parameter estimates for link C.D Model 1.

Figure 3.3.18: Parameter estimates with delta method confidence intervals for the two significant models from the likelihood ratio tests. Corresponding shape and scale parameter estimates for Model 0 from the respective link are shown in red and green.

We can also look at the qq-plots for the links. These are shown in Model 3.3.19. By eye for link F.E, Model 0 appears to fit better. For link C.D the Model 0 qq-plot has a lot of deviation for the link, particularly in the early to middle section. This suggests that Model 1 provides a better fit.

Maximum delay model conclusion. For two of the links, A.B and B.A the best model of the cluster maxima is the GPD with one shape and one scale parameter.



(a) QQ-plot link F.E Model 2.

(b) QQ-plot link F.E Model 0.

(c) QQ-plot link C.D Model 1.

(d) QQ-plot link C.D Model 0.

Figure 3.3.19: QQ-plots for the two significant models from the likelihood ratio tests. The Model 0 qq-plots are shown for comparison.

The other two links both select models with scale parameters based upon day of the week (C.D) and time of day (F.E). However given the large number of links in the network it is unrealistic to model each link in such detail. As we would expect some deviation from the basic models then we conclude that the best overall model for all links is the GPD with one shape and one scale parameter.

3.4 Modelling the cluster length

We wish to find the overall impact of delay due to a cluster. We have modelled the impact by condensing every delay into a single point and then modelling the maximum height of a cluster in Section 3.3.2. However these clusters exist at multiple points in time. We therefore wish to model the length of the clusters.

We will consider two ways of considering the length of the cluster. The cluster duration isn't considered in the extremes literature. We define the cluster duration as the time from the start of the cluster to the end of the cluster. From Section 3.3.1 we have the defined that the jth cluster begins at time t_j and ends at time τ_j. Hence the cluster duration, D_j is:

$$D_j = \tau_j - t_j. \tag{3.4.1}$$

The definition of cluster length from the extremes literature is the number of time periods within a cluster that the series is above the threshold (Smith and Weissman, 1994). Hence we now defined the cluster length L_j as:

$$L_j = \sum_k I_{\{y_k > u \,\wedge\, t_j \leq k \leq \tau_j\}}. \tag{3.4.2}$$

When modelling the roads in real time the duration of the cluster is particularly important as we wish to avoid sending vehicles down a road before the delay from the cluster is over. When scheduling a day in advance the occurrence of any clusters are unknown and hence their length will give a better indication of the size of the overall delay due to an incident on the link across time. The duration gives how long the effect is spread over, even though it's not even. We therefore look at both the duration and the cluster length.

By definition both L_j and D_j must be integers and to enable the two to be directly compared the durations are calculated on the scale of 15 minutes, such that there is a gap of one between each time point. The cluster length is at most equal to the duration and by the definition of a cluster if the cluster length is one the duration is also one.

As with the cluster maxima the cluster lengths may be entirely random, or may depend upon different factors. To begin we calculate the mean cluster length and the mean duration of all clusters for one link. The mean cluster length for link A.B is 1.52 while the duration is 1.80. In addition we look at the proportions of each type of cluster size. This is given in Table 3.4.1. Most clusters are of length/duration one. For values higher than one the larger the cluster size the smaller the proportion. The maximum cluster length is seven which is also the maximum duration. However most of the proportion for the length is with a smaller cluster size. This shows that several clusters contain a dip below the threshold.

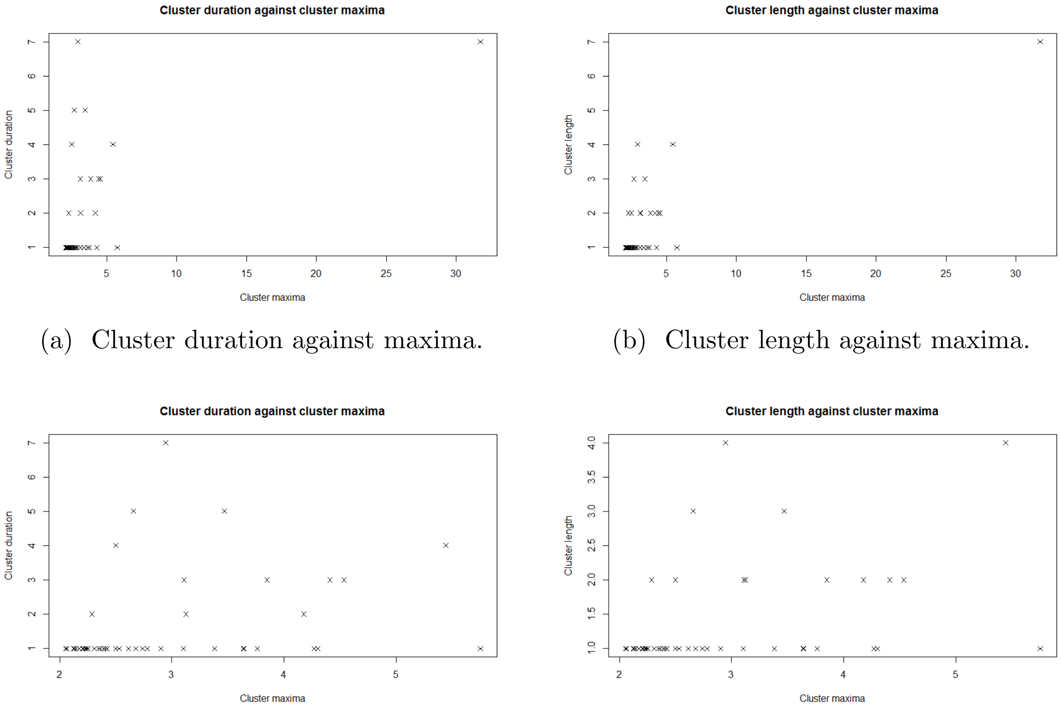
Cluster size	1	2	3	4	5	6	7
Proportion (duration)	0.72	0.07	0.09	0.04	0.04	0	0.04
Proportion (length)	0.72	0.17	0.04	0.04	0	0	0.02

Table 3.4.1: Table of the proportions of each cluster length and duration for link A.B.

One factor which may affect the cluster length is the size of the cluster maxima. Figure 3.4.1b shows the plot of cluster length against the cluster maxima while Figure 3.4.1a shows the duration against the cluster maxima. There is one clear outlier which is also an outlier with respect to the cluster maxima. This obscures the behaviour of the rest of the points so we look at the plot again without the outlier. From Figure 3.4.1d we can see that the majority of clusters are of length one and these are spread over the range of cluster maxima. When considering the cluster duration as shown in Figure 3.4.1c the distribution of points is similar but with higher numbers. Of the clusters that are longer than one there is a slight positive correlation with the cluster maxima but this isn't too strong.

The cluster length could change throughout time, by day or with the time of day. Figure 3.4.2 shows the plots of the cluster length and durations against the time of day, the day of the week and throughout time. There is little difference between the plots for length or duration, except that the length are lower. The spread of clusters of length one or duration one on all three scales appears random. Figure 3.4.2a shows that when considering clusters of length greater than one over a day the spread is also relatively random.

Over the scale of a week both Wednesday and Thursday have no clusters greater than length one as can be seen in Figure 3.4.2c. Over the whole period of January



(a) Cluster duration against maxima.

(b) Cluster length against maxima.

(c) Zoomed in plot of duration.

(d) Zoomed in plot of length.

Figure 3.4.1: Plots of cluster length and duration against cluster maxima for Link A.B.

Figure 3.4.2e has a possible slight increase in cluster length in middle of the month.

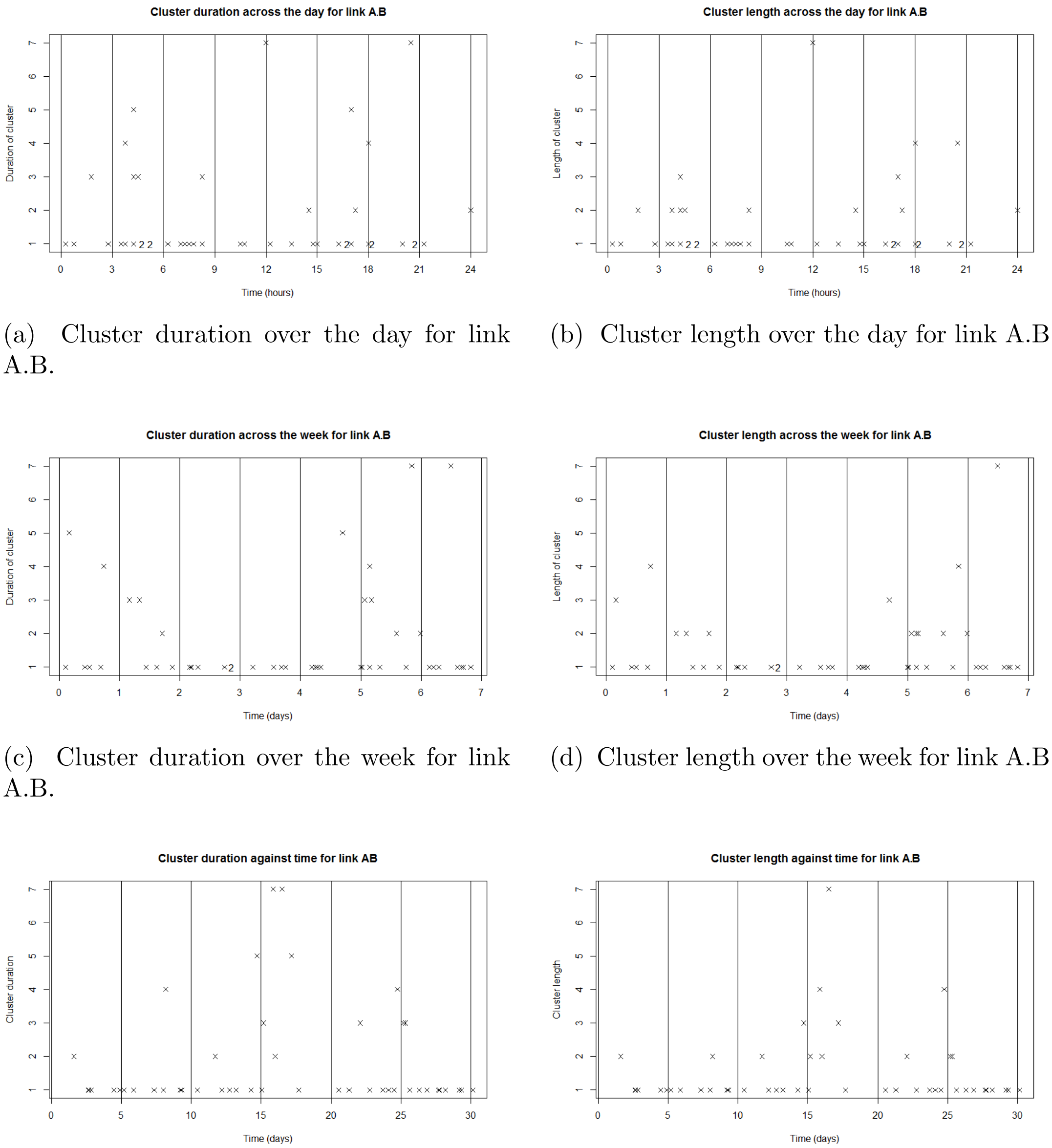
3.4.1 Cluster lengths in other links

To analyse the other links and compare them with each other we first calculate the mean cluster length for each link. Table 3.4.2 shows the average of each link. A.B and B.A are similar to each other but are both smaller than C.D and F.E which are also close to each other.

Link	A.B	B.A	C.D	F.E
Mean duration	1.80	1.70	2.98	3.1
Mean length	1.52	1.55	2.60	2.68

Table 3.4.2: Table of the average cluster length for links A.B, B.A, C.D and F.E.

To examine this in greater detail we look at the distributions of the cluster lengths



(a) Cluster duration over the day for link A.B.

(b) Cluster length over the day for link A.B.

(c) Cluster duration over the week for link A.B.

(d) Cluster length over the week for link A.B.

(e) Cluster duration over time for link A.B.

(f) Cluster length over the time for link A.B.

Figure 3.4.2: Plots of cluster length against different time indexes for Link A.B. Repeated cluster lengths for the same time point are shown by the number rather than a cross.

and durations. Table 3.4.3 shows that for all four links the vast majority of clusters are of duration one. Links F.E and C.D have a lower proportion of clusters being one. The highest cluster length for Link A.B is seven which is much lower than seventeen and eighteen for links F.E and C.D respectively. All four links have some of all cluster lengths up to four.

Duration	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Link A.B	0.72	0.07	0.09	0.04	0.04		0.04											
Link B.A	0.84	0.04	0.04	0.02		0.02					0.04							
Link C.D	0.57	0.04	0.11	0.11	0.06	0.04					0.02	0.02		0.02				0.02
Link F.E	0.60	0.08	0.04	0.06	0.08			0.02		0.06	0.02		0.02				0.02	

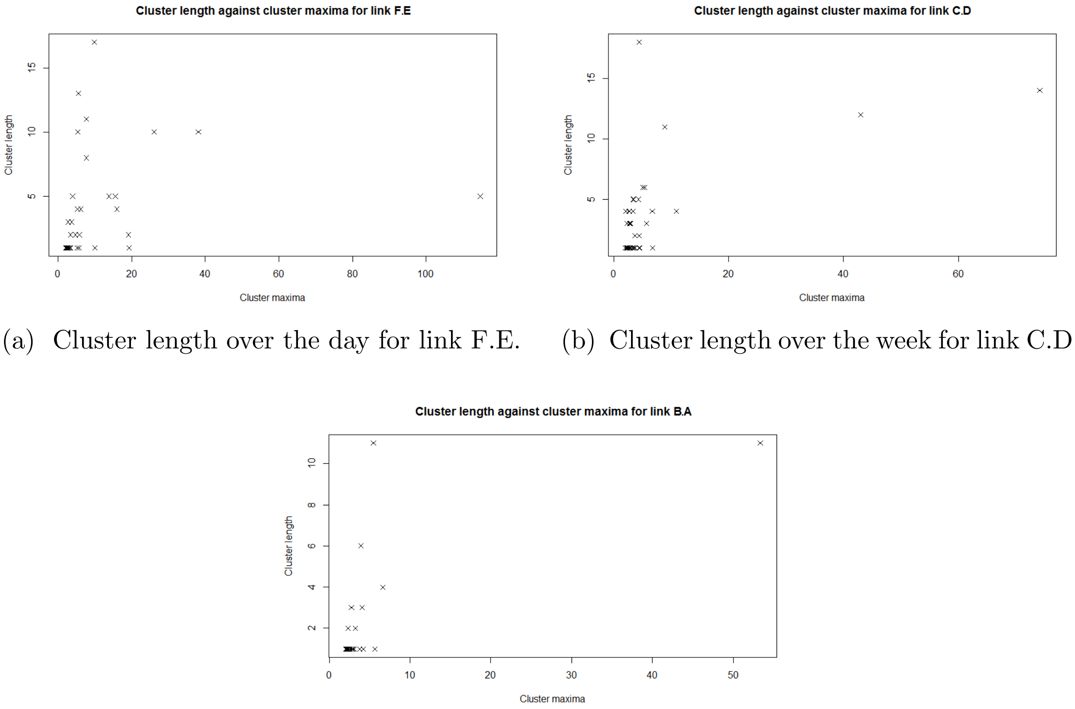
Table 3.4.3: Table of cluster duration proportions for links A.B, B.A, C.D and F.E. Any zeros have been left blank to highlight where the proportions are.

Table 3.4.4 shows the proportions of the cluster lengths for the different links. Unlike with the duration the second highest proportion for all links is length two. Links C.D and F.E still have some of the proportion that is higher than links B.A and A.B but this is lower than the duration.

Length	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Link A.B	0.72	0.17	0.04	0.04			0.02											
Link B.A	0.84	0.06	0.02	0.02	0.02	0.02					0.02							
Link C.D	0.57	0.17	0.08	0.04	0.04	0.04			0.02	0.02				0.02	0.02			
Link F.E	0.60	0.10	0.08	0.02	0.06	0.02	0.06			0.02			0.04					

Table 3.4.4: Table of cluster length proportions for links A.B, B.A, C.D and F.E. Any zeros have been left blank to highlight where the proportions are.

We next look at the cluster maxima in comparison with the cluster length. As with the link A.B there is a slight positive correlation between the cluster maxima and cluster length. The exceedingly large cluster maxima all have lengths of at least five. For link B.A the clusters of size one are spread throughout the range of most cluster maxima (below seven) but with a higher concentration lower down. For both links F.E and C.D this concentration of length one is much more pronounced, with there being more clusters with length above one than length below one as the cluster maxima gets higher.

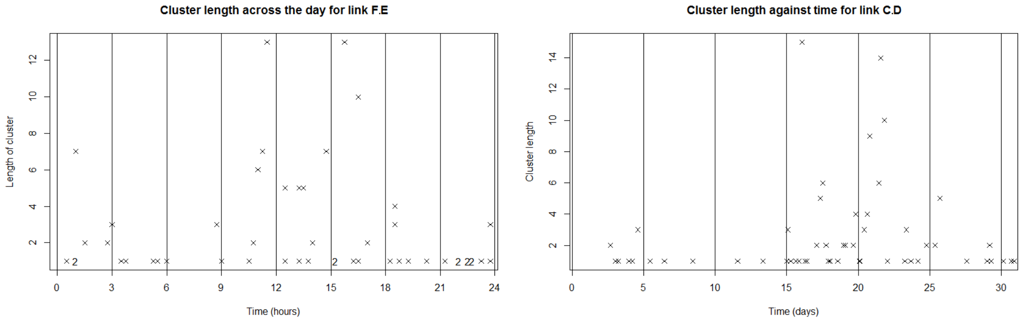


(a) Cluster length over the day for link F.E. (b) Cluster length over the week for link C.D.

(c) Cluster length over time for link B.A.

Figure 3.4.3: Plots of cluster length against cluster maxima for links B.A, C.D and F.E.

Figure 3.4.4a shows an increase in cluster length in the middle part of the day from about 10am to 5pm. The GPD model for the cluster maxima with different scale parameters for the time of day fitted better than a single scale parameter. Similarly Figure 3.4.4b shows an increase between days 15 and 25 for link C.D. This however does not correspond with the preferred model for link C.D which allowed the scale parameter to change dependent upon the day of the week. The plots for the links B.A, C.D and F.E for the other time factors all appear to be random.



(a) Cluster length over the day for link F.E.(b) Cluster length over the week for link C.D.

Figure 3.4.4: Plots of cluster length against different time scales.

3.4.2 Pooling across roads

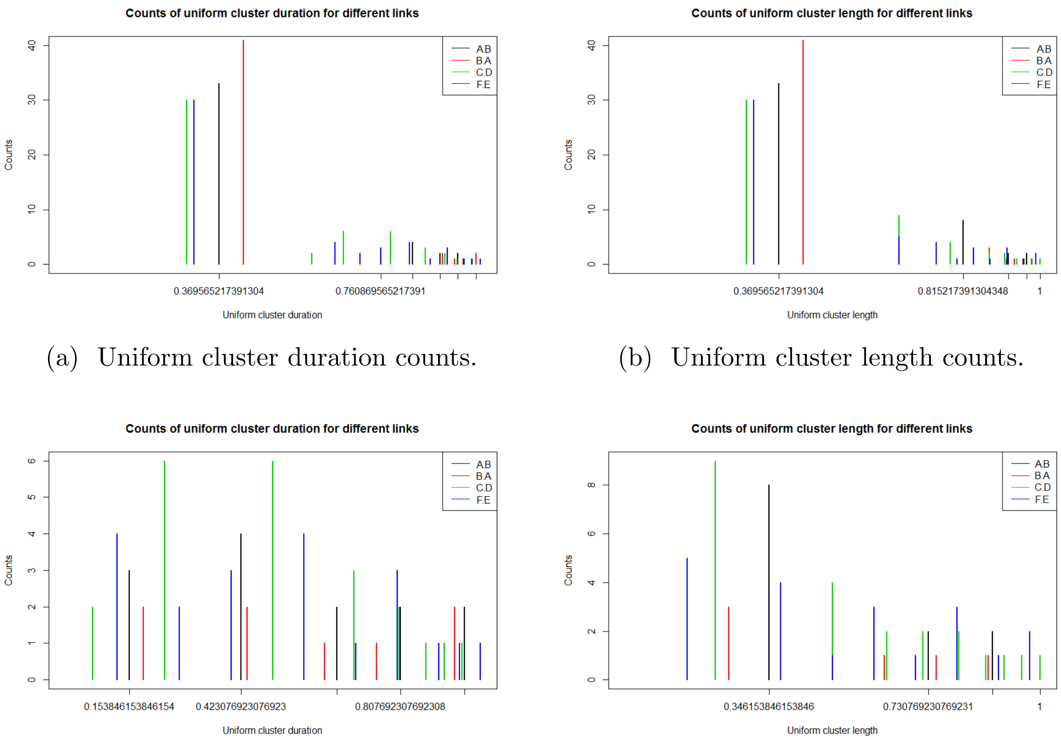
Due to the different scales of the cluster lengths and durations it is difficult to tell if the links behave in a similar manner. To enable us to compare them we will transform each link to a Uniform(0,1) distribution using probability interval transform. This transforms the data from one distribution to another that is all on the same scale. The empirical distribution function allows us to rank the data without making any additional assumptions on the distribution.

The empirical distribution function is:

$$\frac{\text{Rank}(x)}{\text{length}(x)}. \tag{3.4.3}$$

As our data is discrete with multiple clusters having the same length this leads to blocks for for the length or duration. We transform the cluster lengths and durations for each length. The resulting duration mean for each link to 2dp is 0.51 which is the same as the cluster length mean.

We then look at the counts of each uniform cluster duration for each link. This is shown in Figure 3.4.5. As before all links are dominated by the equivalent uniform size of what were cluster size one. We therefore look at these without the ones. On the uniform scale there is a slight decrease in number as the duration and length increases. The effect is more pronounced for the cluster length.



(a) Uniform cluster duration counts.

(b) Uniform cluster length counts.

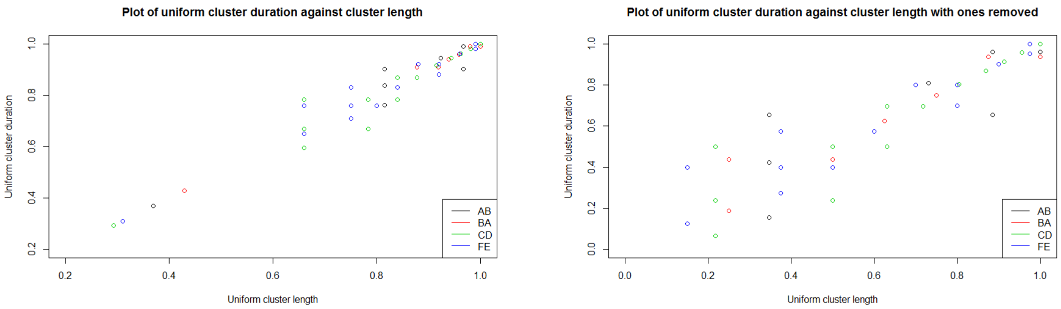
(c) Uniform cluster duration counts without duration one.

(d) Uniform cluster length counts without length one.

Figure 3.4.5: Plots of cluster length and duration counts for each link, both including, and excluding the cluster of size one.

We also check whether the relationship between cluster length and cluster duration is the same for all links. Figure 3.4.6 shows plots of the uniform cluster duration against uniform cluster length both including the clusters of length one and excluding them. There is a clear positive correlation although this is more visible for the one including the clusters of size one. Although not apparent from the plot the majority of points are located at the equivalent uniform scale to a cluster size of one.

We also wish to check the relationship of the cluster maxima against the cluster durations and lengths for all links. However the cluster maxima size depends upon the link and hence we need to transform them to one scale. Because the cluster maxima are distributed with a generalized Pareto distribution we can transform each



(a) Uniform cluster duration against length.

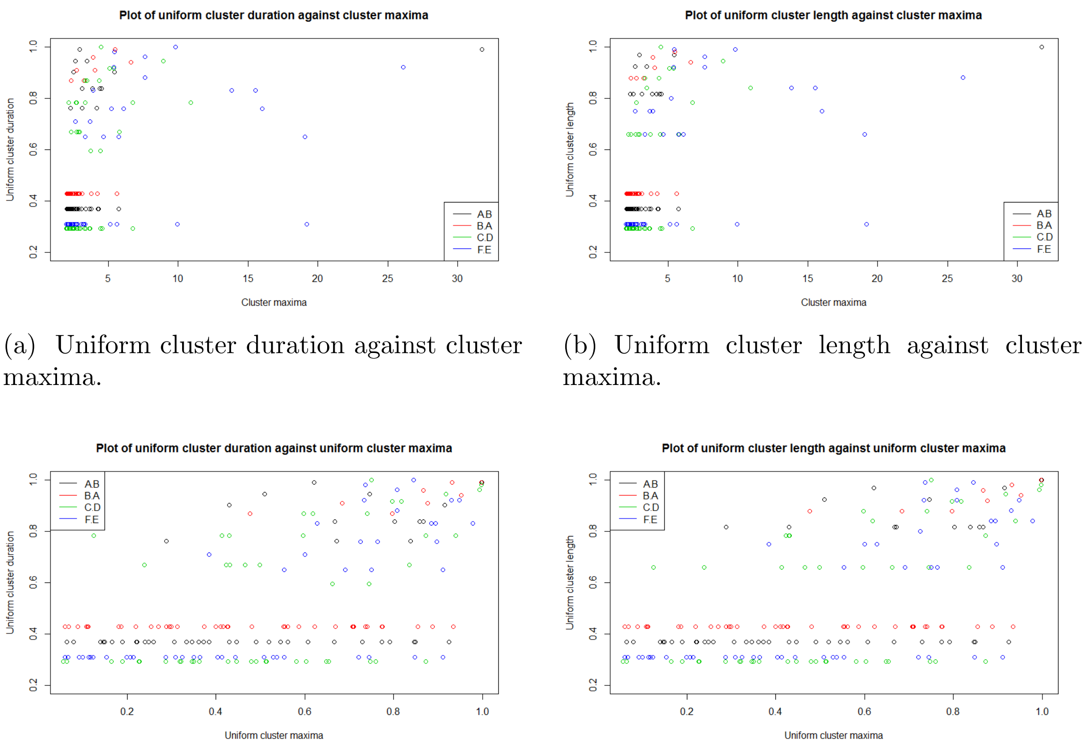
(b) Uniform cluster duration against length without ones.

Figure 3.4.6: Uniform cluster duration against uniform cluster length for different links. One plot includes the ones, while the other has them removed to see the effect they have.

link to uniform scale. The transformed and untransformed plots for cluster maxima against cluster length/duration are shown in Figure 3.4.7. These show a possible slight positive relationship between cluster maxima size and the cluster duration/length.

To check whether the relationship between the cluster maxima and duration or length is independent we calculate the Kendall Tau for each set of points. The Kendall Tau value measures if the ranks for both the uniform cluster maxima and the uniform cluster length or duration are the same. The Kendall Tau values are 0.260 for the uniform cluster duration and uniform cluster maxima and 0.266 for the uniform cluster length and uniform cluster maxima.

However the calculation of the Kendall Tau value assumes that all the values of the points are unique whereas our data contains a large number of ties. To account for this we randomly assign the uniform cluster maxima and uniform cluster lengths into new pairs and calculate the Kendall Tau values for these to give us a 95% interval. The interval for the uniform duration and uniform maxima is {-0.223,0.215} and for the uniform length and uniform maxima is {-0.215,0.212}. Neither of these intervals contain the observed Kendall Tau values and hence we conclude that there is some relationship between the uniform cluster maxima and the uniform cluster duration or length.



(a) Uniform cluster duration against cluster maxima.

(b) Uniform cluster length against cluster maxima.

(c) Uniform cluster duration against uniform cluster maxima.

(d) Uniform cluster length against uniform cluster maxima.

Figure 3.4.7: Uniform cluster size against cluster maxima for different links.

We now wish to try and fit a distribution to the lengths and durations. From Figure 3.4.6 it appears that when the links are transformed to a uniform scale the uniform cluster length and uniform cluster duration are approximately equal. This suggests that if one distribution models the cluster lengths, when allowing it to vary for each link, then a similar distribution will be appropriate for the cluster durations.

One possible distribution is the geometric distribution. The geometric distribution models the number of trials, k, needed to get one success, given we start with a failure. The Geometric distribution is given as:

$$P(X = k) = (1 - p)^{k-1}p, \tag{3.4.4}$$

whereby p is the probability of success. Depending upon if we are modelling the duration or the cluster length X will be either D or L. In our case a failure is a delay and hence each fifteen minute interval is a trial in which the delay will either remain or clear up for the next fifteen minutes. This follows more clearly for the duration whereby this is the overall end of the period whereby the delay has any effect.

To check whether the geometric distribution is appropriate we produce qq-plots. These are shown in Figures 3.4.8 and 3.4.9. To generate these we first have to find the maximum likelihood estimator for the geometric distributions given the observed cluster lengths and durations. The maximum likelihood estimator for the length is:

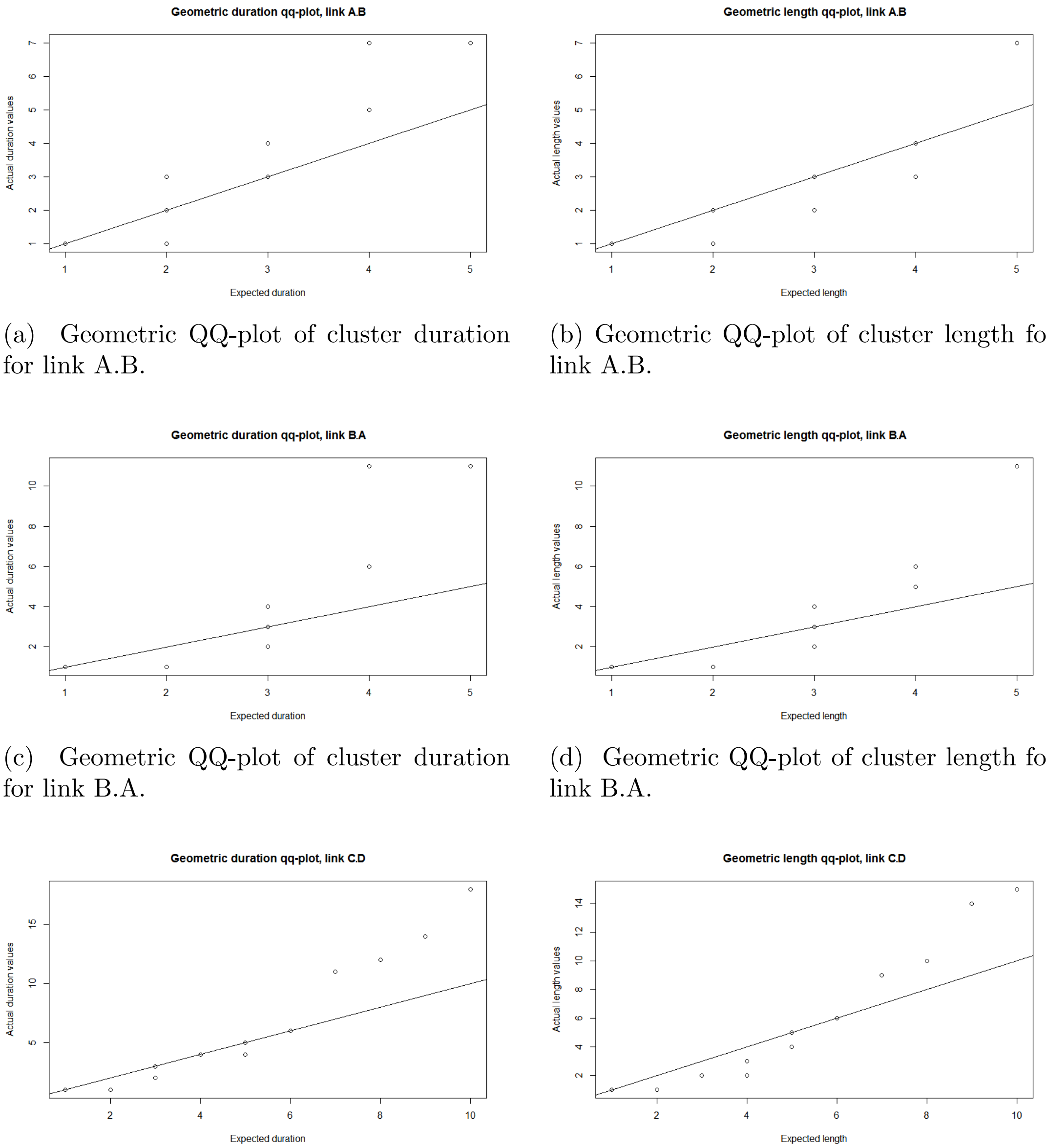
$$\hat{p} = \frac{n}{\sum_j L_j}, \tag{3.4.5}$$

whereby the number of clusters is n.

The qq-plots are approximately close to the line but do have outliers, particularly in the tails. There are a lot fewer points than clusters because all values must be integer and therefore there are a lot that occur at the same point. We therefore consider a distribution with heavier tails as an alternative.

When considering the cluster maxima we used the GPD as a heavier tailed alternative. Cluster maxima are continuous whereas the cluster lengths and durations are discrete integer values. We hence need to adapt the GPD to this discrete setting.

The simplest way to make the GPD discrete is to round the continuous values to the integer lengths. However the cluster length and duration are at a minimum one, while the GPD starts at zero. If we shift the GPD such that the threshold is 0.5 we can round any value that is between $j - 0.5$ and $j + 0.5$ to j, where j is an integer of at least one. To calculate the probability for each j this we use the survival function,



(a) Geometric QQ-plot of cluster duration for link A.B.

(b) Geometric QQ-plot of cluster length for link A.B.

(c) Geometric QQ-plot of cluster duration for link B.A.

(d) Geometric QQ-plot of cluster length for link B.A.

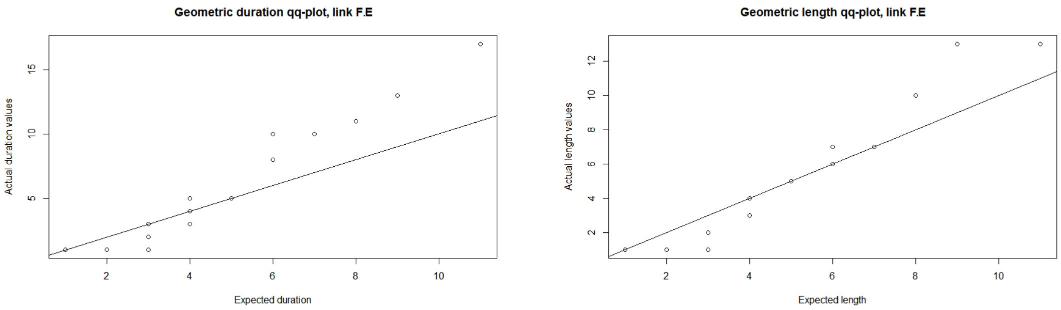
(e) Geometric QQ-plot of cluster duration for link C.D.

(f) Geometric QQ-plot of cluster length for link C.D.

Figure 3.4.8: Geometric QQ-plot of cluster duration and lengths for different links.

S, of the GPD. The probability of observing cluster length j is:

$$
\begin{aligned}
P(L=j) =& S(j-0.5) - S(j+0.5) && (3.4.6)\\
=& (1+\xi\frac{j-0.5-0.5}{\sigma})^{-1/\xi} - (1+\xi\frac{j+0.5-0.5}{\sigma})^{-1/\xi} && (3.4.7)\\
=& (1+\xi\frac{j-1}{\sigma})^{-1/\xi} - (1+\xi\frac{j}{\sigma})^{-1/\xi}. && (3.4.8)
\end{aligned}
$$



(a) Geometric QQ-plot of cluster duration for link F.E.

(b) Geometric QQ-plot of cluster length for link F.E.

Figure 3.4.9: Geometric QQ-plot of cluster duration and lengths for link F.E.

This is valid for any $j + 0.5 \leq 0.5 - \sigma/\xi$ if $\xi < 0$. As before with the GDP $\sigma > 0$. The survival function must be positive but it may reach the maximum L within the interval. To counteract this we set $S(j + 0.5) = 0$ if it would otherwise be negative for the highest observed j. The total likelihood is therefore:

$$P(L = \{l_1, \ldots, l_n\}) = \prod_{i=1}^{n} P(L = l_i). \tag{3.4.9}$$

For the duration we substitute D for L. Hence we need to find the optimal parameters given the sample for each link which we can do by maximizing the log-likelihood.

Table 3.4.5 shows the different ξ and σ values for the different links. The subscripts d and l indicate if they are for the length or the duration. The parameter values vary from link to link and the duration values are higher than the lengths as would be expected.

Link	A.B	B.A	C.D	F.E
ξ_d	0.589	1.12	0.634	0.940
σ_d	0.592	0.175	1.14	0.800
ξ_l	0.265	0.916	0.599	0.732
σ_l	0.672	0.220	0.970	0.875

Table 3.4.5: Table of the GPD parameter sizes for length and duration for links A.B, B.A, C.D and F.E.

For further comparison Figure 3.4.10 shows these values along with the standard errors for the ξ parameters. We can clearly see that a single ξ value would fit within the standard error intervals of both the duration and length which would fit with the common shape parameter literature discussed in Section 3.3.3. However given that the optimal shape values for all the duration GPDs are higher than the length ones then we will test these separately. We then test if a single shape parameter is appropriate by conducting a set of likelihood ratio tests to see if a single shape parameter with either a single scale parameter or different scale parameters are preferable. Unlike in the previous likelihood ratio tests for the cluster maxima the three models are nested and hence can be compared directly to each other.

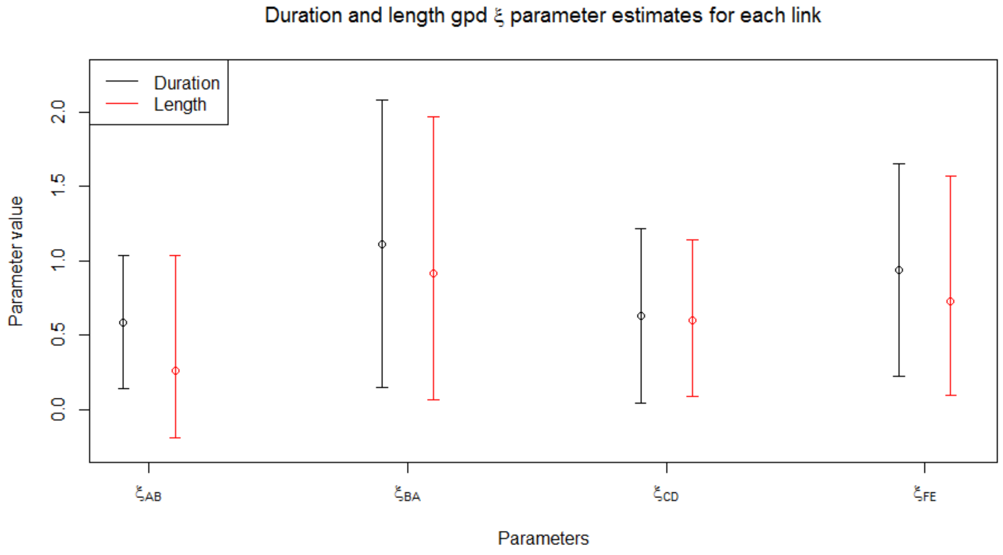


Figure 3.4.10: Gpd parameter estimates for ξ for the cluster duration and lengths for different links.

We have three models, M_0 whereby all links have the same ξ and σ. Then we have M_1 whereby all links have the same shape parameter ξ but different scale parameters σ_{ij} for each link. The last model is M_2 which has unique ξ_{ij} and σ_{ij} values for each link. To compare the three models we use likelihood ratio tests. We first compare models M_0 and M_2.

$$\begin{aligned} H_0 :&\xi_{ij} = \xi, \sigma_{ij} = \sigma \in \mathbb{R}^+ && \forall ij \in E \\ H_1 :&\xi_{ij}, \sigma_{ij} \in \mathbb{R}^+ && \forall ij \in E. \end{aligned} \tag{3.4.10}$$

The degrees of freedom between M_0 and M_2 is 6, and when looking at the cluster lengths has a test statistic of 15.65 and a p-value of 0.016. Hence we reject the null hypothesis that the shape and scale parameters are the same and instead conclude that there is evidence that the shape and scale parameters vary link by link.

We next compare whether or not having a common shape parameter across links is preferable. Our two hypothesis are

$$\begin{aligned} H_0 :&\xi_{ij} = \xi, \sigma_{ij} \in \mathbb{R}^+ && \forall ij \in E \\ H_1 :&\xi_{ij}, \sigma_{ij} \in \mathbb{R}^+ && \forall ij \in E. \end{aligned} \tag{3.4.11}$$

There are three degrees of freedom between M_1 and M_2. The test statistic is 2.37 resulting in a p-value of 0.50. Hence we conclude that there is insufficient evidence that the shape parameter varies link by link and instead choose the model that has a common shape parameter and varying scale parameters.

For completeness we then compare M_1 to M_0. This also has three degrees of freedom and with a p-value of 0.0041 we would reject the model with the same shape and scale parameters and instead conclude that there is evidence to suggest that M_1 is the preferred model.

We also look at the three comparative models for the cluster durations. Between M_0 and M_2 the p-value is 0.018, between M_0 and M_1 it is 0.0032 and between M_1 and M_2 0.70. This leads to the same conclusion as for the cluster lengths, whereby the best model is M_1.

Hence we conclude that M_1 is the best model. The cluster lengths have a shape parameter of 0.624 which is within the confidence intervals of the shape parameters for each link. The cluster durations have a shape parameter of 0.813. The scale parameters are shown in Table 3.4.6a.

The exponential model has shape parameter 0. Hence as a final comparison we check whether this lies within the confidence intervals of the discrete GPD shape parameters. The cluster length shape parameter confidence interval is $(0.328, 0.920)$ and the cluster duration shape parameter is $(0.462, 1.165)$. Hence the zero is outside both intervals, which confirms that the cluster lengths and durations have a heavier tail than the exponential distribution as we expected.

Link	A.B	B.A	C.D	F.E
σ_d	0.479	0.257	0.988	0.890
σ_l	0.486	0.318	0.951	0.960

(a) Table of the GPD scale parameter sizes for length and duration for links A.B, B.A, C.D and F.E.

Parameter	Value
ξ_d	0.813
ξ_l	0.624

(b) Shape parameters for the duration and length.

Table 3.4.6: Parameters for model M_1 for the cluster durations and cluster lengths.

Concluding remarks on the cluster length. In this section we have considered both the cluster length, which is the number of intervals a delay is extreme and the duration which is the number of intervals from the start of the delay to the end. The cluster duration is at least as long as the cluster length but both are of interest when modelling the extreme delays. We have concluded that the best model for modelling the cluster lengths for all the links is the discrete GPD with a common shape parameter and a scale parameter unique to each link. This is the same as for the cluster duration, except the parameter values differ.

3.5 Overall delay model

In Section 3.2 we modelled the rate of occurrence of delays by grouping together outliers from the same delay in one cluster. We then modelled the cluster maxima in

Sections 3.3.3 and 3.3.4 and the cluster lengths in Section 3.4. We now restate these models for clarity.

The delays occur following a Poisson process with a rate that is constant and unique to each link. The maximum delay size is modelled by a generalized Pareto distribution with shape and scale parameters that are unique to each link. The cluster length is modelled by a discrete generalized Pareto distribution with a common shape parameter for all links and unique scale parameters.

By combining the models for the cluster maxima and the cluster length we can model the impact of extreme delays upon the travel time. However, this model gives us only the worst possible delay and the overall length, rather than the behaviour across the entire cluster. Modelling the behaviour of the delay over time requires a model with a lot more complexity. We discuss combining the delay models together in Section 5.2.2.

Combining the two models gives a model for a delay conditional on the delay happening. This means that we would need to know the start time of a delay in order to use the model for a future delay. However, in the future we don't know when a delay is going to occur. As a result we wish to calculate the probability of a delay occurring in the future, so that we can incorporate expected delay occurrences in our decision making.

3.5.1 Probability of a delay

We begin by looking at the probability that any of the 15 minute intervals of the 96 in a day are in a state of delay. Figure 3.5.1 shows a section of a day with delays as they occur in continuous time. A cross indicates the delay starting and a small vertical line the end of the delay. We only observe at 15 minute intervals, irrespective of when in the interval the start and end of the delay occur, and we classify an interval as being in a state of delay if any part of a delay occurs in it. The 15 minute intervals that are in a state of delay are therefore 6, 7 and 8 from the first delay and 14 and 15 for the second.

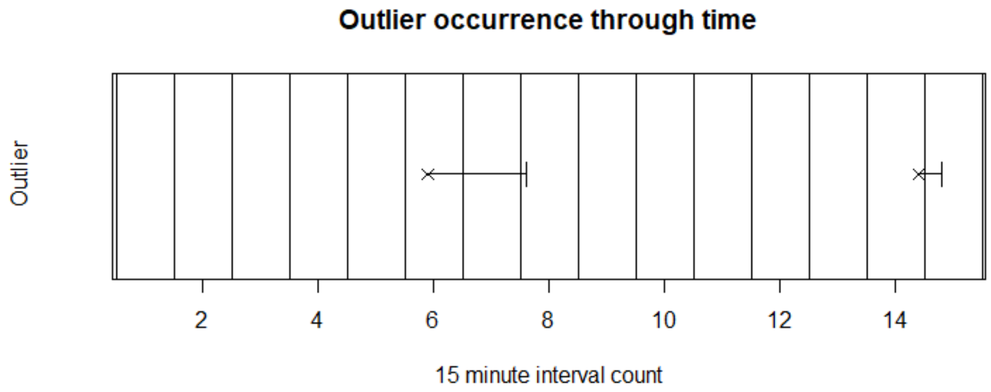


Figure 3.5.1: Cluster counts for each day for link A.B.

In Section 3.3.1, after standardization, we concluded that we would model the rate, $\gamma(t)$, of cluster occurrences over time, as constant over the day to reduce the number of parameters to estimate. As the rate is constant throughout the day the probability that a delay occurs in any interval is the same and the rate is unique to each link. We don't need to make any adjustments at the start of the day for delays that started at the end of the previous night.

The probability that a delay occurs in a 15 minute interval depends upon the probability of a delay occurring and the length of the delays. Let P_d be the probability a single interval being in a state of delay and R be the total number intervals that are in a delay state over the day. Then

$$P_d = \frac{E(R)}{96}. \tag{3.5.1}$$

Ideally we would be able to calculate $E(R)$ directly but this expression is very difficult to evaluate analytically. As clusters last for multiple intervals and their occurrence follows a Poisson process it is possible, but very unlikely, that the delays may overlap. Thus R is the sum of all the non-overlapping delays and hence the number of delay intervals that the ith delay contributes to R depends upon all the delays before that. The expectation of R is a large expression of multiple sums and integrals that cannot be simplified due to the dependence between them.

If two delays cannot occur at the same time then the number of delay intervals, R is the sum of the length of each delay that is inside that day. We can therefore approximate the expectation of the number of intervals of delay by the expectation of the length, L multiplied the expectation of the number of delays in an day. We define N to be the number of delays that start in that day. Hence:

$$E(R) \approx E(N)E(L). \tag{3.5.2}$$

This approximation works because a day is cyclic. Some lengths will extend beyond the end the day but these correspond to the delay intervals of continuing delays that started before the day started. This is applicable because the rate of occurrence is constant.

We can use the discrete GPD distribution from equation (3.4.8) to find the probability that a cluster is a of a certain length. Using the notation of Section 3.4 L is the length of the cluster. Let c_L be large enough that $P(L \geq c_L + 1) \approx 0$. Then

$$E(L) = \sum_{l=1}^{c_L} \left((1 + \xi \frac{l-1}{\sigma})^{-1/\xi} - (1 + \xi \frac{l}{\sigma})^{-1/\xi} \right) l. \tag{3.5.3}$$

From Section 3.2.3 we know that the standardized delays occur according to a Poisson Process with constant rate across the day such that each interval has the same probability of a delay starting within it. Let N be the number of clusters in a day which is Poisson distributed with rate γ in a day. Then the probability of observing n clusters in a day is

$$P(N = n) = \frac{\gamma^n e^{-\gamma}}{n!}. \tag{3.5.4}$$

We use maximum likelihood to estimate γ. Using the same notation as Section 3.3.1, let x_d be the number of clusters observed on day d. Then

$$\hat{\gamma} = \sum_{d=1}^{31} \frac{x_d}{31}. \tag{3.5.5}$$

The estimated expectation is therefore

$$E(N) = \hat{\gamma}. \tag{3.5.6}$$

We can then combine the two expectations to get an approximation for $E(R)$. Hence the probability an interval is in a state of delay is approximately

$$P_d \approx \frac{\hat{\gamma}E(L)}{96}. \tag{3.5.7}$$

3.5.2 Evaluation of P_d by Monte Carlo methods

As noted in Section 3.5.1, equation (3.5.1) is too complicated to evaluate analytically. We instead find the probability that an interval is in a delay state using Monte Carlo methods. Using this we can check whether our simpler probability, the expectations estimation of equation (3.5.7), that includes overlaps is an acceptable approximation or not.

To set up the simulation we have to first clarify what we are modelling. From Section 3.2 we have the cluster inter-arrival times being exponentially distributed. Then from Section 3.4 the cluster length is a discrete generalized Pareto distribution.

We can therefore generate a sequence of when the delays occur in time, with their accompanying lengths. From this we can use an indicator function to show whether or not there is a delay happening in interval t. Let I_{ti} be the indicator function for the delay in interval t of the ith simulation. If we then find the sum of the indicator function over one day and divide it by 96 we get the estimated probability from the

ith simulation that an interval has a delay as

$$\hat{p}_i = \frac{\sum_{t=1}^{96} I_{ti}}{96}. \tag{3.5.8}$$

This only needs the indicator function between the times 1 and 96. However we need to generate the arrival times and cluster lengths for a large period of time beforehand because the arrival times have no fixed start time and the simulation requires a set start time. If the simulation starts at time 0 and only runs for 96 periods this requires the assumption that there is no event occurring at time 0. The time beforehand that we use as burn in time to eliminate the fact that the simulation is started at a particular time by letting it settle into its natural state.

For the purpose of this simulation we have chosen to select 10^4 as the start time. This is just over 100 days beforehand. The simulation is then run n times to ensure that the average is close to what we would expect. The returned probability of the simulation is then:

$$\hat{p} = \frac{\sum_{i=1}^{n} \hat{p}_i}{n}. \tag{3.5.9}$$

Here $\hat{p}$ is an estimate of P_d, the probability that an interval is in a state of delay. We find an uncertainty interval for the Monte Carlo method by considering the variance of our estimator $\hat{p}$. Then

$$\text{Var}(\hat{p}) = \frac{\text{Var}(\sum_{i=1}^{n} \hat{p}_i)}{n^2}. \tag{3.5.10}$$

The $\hat{p}_i$ are independent of each other as each run of the simulation is independent of the others. Thus

$$\text{Var}(\hat{p}) = \frac{\sum_{i=1}^{n} \text{Var}(\hat{p}_i)}{n^2}. \tag{3.5.11}$$

We then define $m_i = \sum_{t=1}^{96} I_{ti}$ and substitute this into equation (3.5.8). Using that the m_i are identically distributed we then get that:

$$\text{Var}(\hat{p}) = \frac{\text{Var}(m_i)}{96^2 n}. \tag{3.5.12}$$

Using this we can construct confidence bounds of plus and minus two standard errors of $\hat{p}$ to give a range of values which P_d is likely to fall in. The more simulations we run the smaller the variance, and thus the smaller the Monte Carlo uncertainty interval. We wish to pick n large enough so that we can identify if the expectations estimation, of equation (3.5.7), lies within these bounds.

The simulation is run for each of the four links using the optimal parameters identified in the models in the previous sections. We use the values from Table 3.4.6 for the length to give the optimal ξ and σ values for the discrete GPD for each link and the rate of occurrence values from Section 3.4. This gives the probability of delay for one interval for each of the four links. We choose $n = 10^6$ to ensure that the standard error for $\hat{p}$ is small enough. We then calculate equation (3.5.7) using the same parameters to give the probability of a delay when ignoring overlaps. The probabilities are presented in Table 3.5.1.

	Link A.B	Link B.A	Link C.D	Link F.E
Monte Carlo Approximation	0.0399	0.0295	0.0763	0.0667
(3.5.7)	0.0408	0.0301	0.0798	0.0694
Lower Monte Carlo interval	0.0397	0.0282	0.0734	0.0640
Upper Monte Carlo interval	0.0401	0.0308	0.0792	0.0693

Table 3.5.1: Probabilities of delay for one interval (P_d) for each of the simulation and the estimation from (3.5.7).

We can compare the estimation from equation (3.5.7) to the values obtained from the simulation. If the value is inside the Monte Carlo interval we conclude that the estimation is close enough to the simulation that it is an acceptable approximation of P_d. Only the value for Link B.A is inside the Monte Carlo interval and this is only just inside. This would lead us to conclude that the estimations, other than for link B.A, are different from the true value of P_d.

All of the estimations from (3.5.7) are slightly higher than the Monte Carlo Approximation which is to be expected as intervals with overlapping delays will be counted more than once. However the Monte Carlo Approximation and the estimation both agree to 2dp and the link B.A is the same to 3dp. The fact that they are as close as they are suggests that very few delays overlap each other. Ideally we would therefore use the simulation to estimate the probabilities for each link and as these probabilities remain the same over time then they would only need to be calculated once. However, this has a high computational cost so if the network is very large we need to use an approximation. The expectations estimation of equation (3.5.7), of the probability of a delay in an interval is suitably accurate for practical usage.

3.6 Conclusion

In this chapter we have created a model for the extreme delays that occur in travel time upon sections of road. We group together delays that occur for the same reason into clusters, as the clusters can be considered to be independent of each other. These clusters occur according to a Poisson process and can be described by their duration, size and probability of occurrence. The behaviour of these delays was concluded to be the same over time, but vary link by link. The size of the maximum delay is modelled by a generalized Pareto distribution, with parameters unique to each length. The length meanwhile can be modelled by a discrete generalized Pareto distribution, with a common shape parameter but different scale parameters for each link.

As extreme delays occur at random we cannot predict their occurrence in advance. We can find the probability of a cluster happening by using the link occurrence rate. We discuss ways in which our delay model can be combined with the single link model (Chapter 2) or the network model (Chapter 4) in Chapter 5.

Chapter 4

Travel time models over a network

4.1 Introduction

Chapter 2 models the travel time of a single link in isolation, selecting a method that uses only information from one link to generate a forecast. The petrol routing problem, as outlined in Section 1.1.3, requires travel time forecasts for the next day for a network. The network is a set of the links and when using the single link model these are all calculated separately.

The set of links are a representation of the road network that is relevant to the specific petrol routing problem. Delays can propagate backwards across multiple links or spread onto the surrounding roads. We therefore wish to consider using multiple links to model the travel time on each link, rather than treating each link separately.

To this end we assume that we can only use information within the specified network to produce forecasts. Our network is a subset of the full road network and hence there will be other roads that connect to the network. To simplify the problem we assume that the surrounding roads have no effect on the roads in our subset which remains the same so the network is a fixed network.

There are two types of road that connect to the network - those that are too small to be managed by Highways England, and the roads that are outside the network boundaries but have travel time data in the same dataset used for the other roads. The roads for which we have no data are smaller with less traffic so the assumption that they would make no contribution to the model is more reasonable. One possibility for the roads outside the network would be to include them as inputs to the model but not forecast using them. However we first build our model without considering these additional roads.

At the very simplest level a network model consists of a function G that takes the previous travel times for the all the links and outputs the predictions for those same links. The mathematical notation when we consider only the travel times that are one lag away $(t-1)$ is:

$$(\hat{y}_{1,t}, \ldots, \hat{y}_{N,t}) = G(y_{1,t-1}, \ldots, y_{N,t-1}). \tag{4.1.1}$$

The number of lags can be extended to try and improve the prediction. In this chapter we will look at a variety of methods that use multiple links to forecast the travel time on individual links. The method G could use all the links in the forecast of each link or it could be restricted. Using more links is likely to have a high computational cost but could potentially capture more information about the travel times in the future.

A large number of variables often need to be estimated in a network model. Let N be the total number of links in the network. In the single link model each link required a limited number of parameters and these are estimated separately. This is a complexity of N times the number of parameters in each link model.

Network level models have parameters which depend upon either the whole network or a subset of nodes. Calculating one parameter across the network normally has a higher computational cost than a parameter for one link. However calculating one parameter may be quicker than many individual ones and having additional parameters can lead to overfitting. Thus there is a trade off between fit and computational time which increases as the network size increases. The full network for the petrol routing problem, as shown in Figure 1.1.4 in Section 2.3.3, contains over fifty links, hence we wish to reduce the complexity of the model. Simpler models are also easier to interpret, both from the point of view of why they work and identifying if something is wrong.

This consideration motivates restricting the number of links to a subset of links from which we calculate parameters by considering the network structure. Links that are closer together and connected by the network are more likely to contain useful information in forecasting each other.

4.1.1 Network structure

There are many different ways that the network structure can be incorporated into the forecasting model. Our road network is made up of the major A-roads and motorways in the West Midlands area, with links connected at nodes. Some of the links are directly connected together as they form the same roads, while others are connected by roundabouts or other junctions.

Primarily we aim to use the network structure to limit the number of parameters. Only those links that are close enough over the network will have associated parameters that need to be calculated when forecasting a link. Beyond these links the effect of traffic from surrounding links is considered to be negligible. In some network forecasting models the problem gives a clear indication of the extent of effects between any two points in the problem. However it is unclear how to create a connection matrix for travel time relationships over road networks.

Roads that are directly connected are more likely to have travel times that are affected in a similar way. Whether to also include links that aren't adjoining may depend upon additional factors like the length of the link or the road types. Including these in the model when they aren't needed will result in a much higher computational complexity. Hence we will need to carry out sensitivity analysis to determine the correct connections between links so they form inputs into the models. This will be discussed in more detail in Section 4.1.2.

4.1.2 Network connections

As identified in equation (4.1.1), a model on a network takes all N links and returns forecasts of each of these links. This usually results in a very complicated function G. There are two different ways we can use the network connections to improve forecasts from the function G. We can split G into N different functions such that G_a forecasts link a only. However if $G_a = G$ there is no advantage to splitting the model up.

Some models only need a few links within G_a to the forecast link a. If G_a only contains parameters that are unique to forecasting link a then this will be much simpler than G and each link can be forecast separately. Hence using G_a for each link is often quicker than using G to forecast the whole network as we can work with only a subset of the links at a time. In order to use the models we need to identify these subsets.

Ignoring the network structure, we are identifying a set, R_a, of all the links that are considered to be relevant to forecasting the travel time on link a. Any other links can be ignored when forecasting for link a because they don't contain enough useful information. Hence, using one time lag,

$$\hat{y}_{a,t} = G_a(\{y_{r,t-1} | r \in R_a\}). \tag{4.1.2}$$

As we require forecasts for all links in the network we must specify R_a for each of the links. In order to keep the notation appropriate for the full network level models we summarise this as a matrix R such that $R_{a,b}$ is one if link b is in the set R_a, else it is zero.

When G cannot be simplified into a structure similar to G_a we consider a second way which allows us to generate any non global parameters using only a subset of the links. Then R becomes an input into the model, where each row represents a different parameter. Hence we can rewrite equation (4.1.1) as:

$$(\hat{y}_{1,t}, \ldots, \hat{y}_{N,t}) = G(y_{1,t-1}, \ldots, y_{N,t-1}, R). \tag{4.1.3}$$

When $R = I$, the matrix represents the situation where there are no sets of links that need to be considered together. If $G_a(y_a)$ exists for each link a then the model is a single link model of the form from Chapter 2, rather than a network model as all links are independent of each other.

We wish to use the network structure to specify R. A logical starting set for R_a is those that are directly connected to the link a, including the link a itself. If the set R_a doesn't contain the link a, the method ignores any prior knowledge of the link it is forecasting. This provides a small subset of links that can be considered as potentially important for the prediction of link a. However this provides only a starting point. It could be that links of up to two away from a, or even further, provide useful information. The more links we consider the more complex the model is, which could either lead to over fitting, or for the model to take much longer to run.

In addition there may be other factors that are important. Links are of differing lengths and two links combined may be shorter than another single link. In this case it may be significant to consider a link that isn't directly connected, alongside the other links that are one degree closer. Some links that connect together are actually the same stretch of road, with junctions at the connecting nodes which means that they may be more likely to provide information if they are commonly travelled in succession.

We therefore have several different network connections that may be relevant. The first is $R_a = a$, and the second type are $R_a = R_a^c$, where R_a^c are the set of all links that are at most c away from a. The first case is a subset of the second, where $c = 0$. We refer to the links that are exactly d away from link a as d-degree neighbours and thus R_a^c contains all the d-degree neighbours where $d = (0, \ldots, c)$.

Then we have the case whereby some additional roads that are further away are also in the subset. This can be considered to be related to the second case, R_a^c as long as c is large enough, except we have removed more links. To identify any sets that are like this, we can calculate the model parameters using the set R_a^c on a test set, and if all the parameters relating to $R_{a,b}^c$ are zero (or close to zero) then we can remove link b from the set R_a. This is equivalent to setting them to be zero in the matrix R, which reduces the number of parameters that the forecasting model has to calculate.

Identifying the network connections by using R results in links either being included or not included with no reference to the potential size of the connection between links. In spatial problems observations are made at single points in space. Then a distance metric can be used to judge how far one point is from another, which in turn can be incorporated into the matrix R as a weighted value. Weightings are included as it reduces the number of parameters that need to be calculated by a priori estimating the contributions up to a constant. A single parameter can then normalise the contributions of all the weightings for one link or network. As the majority of the methods that follow use only the binary version of R we will leave any adaption of the matrix by weighting to be incorporated into the method itself.

The matrix R has many different names and forms depending upon where and how it is used. If weights are included then it is known as the weighting matrix. If it identifies links that are first degree neighbours then it is also known as the adjacency matrix. However as the set R_a also includes itself it is not a true adjacency matrix in the sense of the definition used in network analysis. We look at selecting the set R_a is Section 4.4

Chapter outline. The format of this chapter is similar to that of Chapter 2. First we describe the different network level models. These models were first applied first to a five link sub-network, and those that took over half a day to return predict the forecasts for a single day and their more complicated extensions were removed from consideration. We then applied the remaining models to a larger twenty link network. The methods are then analysed by comparing them with the best method from the single link chapter. Finally we select the overall best method as measured by predictive ability.

4.2 Network level methods

In this section we discuss the network level models that can be used for forecasting travel time upon road links, starting from the simplest model, which is adapted to form more complicated models. The simplest model is to use the single link model from Chapter 2 for each link. The forecasts from the set of single link models can then be expanded by using linear combinations to create the next level of model complexity. This is the second set of models.

Finally we consider models which use either all the links, or a subset of the links in a full network level model. These generate the forecasts for all the links at once and hence are generally more computationally intensive. Some of the more complicated models are too computationally intensive to be calculated for the petrol routing problem but are nevertheless included as a comparison.

The properties of the models that are discussed in this section are summarised in the following table, Table 4.2.1. The models are grouped together into the three types of model complexity. The properties are whether or not the models use linear combinations of forecasts from single link models, if the model considers the network connections and if all the links are forecasted at one time.

The majority of these methods have not been used on travel time data in the format that will then be used inside a VRP and none have considered the important considerations that we associated with generating predictions in Chapter 2. We begin by looking at the methods that use linear combinations to create travel time predictions.

Method	Linear combination	Network connections	All links
single link ARIMA	NA	NA	N
bubble	Y	Y	N
link beta	Y	Y	N
time dpt beta	Y	Y	N
hierarchical	Y	Y	N
sparse VAR	N	N	Y
constrained VAR	N	Y	Y
VAR	N	N	Y
VARMA	N	N	Y
STARIMA	N	Y	Y

Table 4.2.1: Table summarising the network level methods by whether they use the road network connections and whether they use linear combinations of the single link method forecasts.

4.2.1 Linear combination methods

The first type of models we consider are those which use a single link model to generate initial forecasts for each link and then try and improve them by using linear combinations of the results. The network characteristics, as discussed in Section 4.1.2, are used to select the set of relevant links, R_a, for each link and then the forecast for link a is found by finding the best linear combination of this subset. In the case where $R = I$ the model simplifies to a single link model and we therefore include our summary of that model in this section.

These models are the simplest type of network level model that can incorporate additional links into the forecast of one link. As discussed in Chapter 2 there are a wide variety of methods that use only a single link at a time. However, we will select the ARIMA model to use as the base model, motivated by our findings in Section 2.5.2 that it was the optimal single link model over the six links. Fitting N separate ARIMA models is less computationally intensive than fitting the N ARIMA models simultaneously across the links. We require forecasts for all N links independently of the R matrix.

In a slightly more formal notation we generate the forecast for link a at time $t+1$, $\hat{Y}^a_{t+1}$, by taking a linear combination of the ARIMA forecasts for time $t+1$ for the links that are in R_a. We denote the ARIMA forecast for link b as $\tilde{Y}^b_{t+1}$, where $b \in R_a$ is the set of relevant links for a. This notation is used to keep the overall forecast the same as in Chapter 2. The parameters we estimate, β_{ab}, are unique for link a. Then the forecast for link a is:

$$\hat{Y}^a_{t+1} = \sum_{b \in R_a} \beta_{ab} \tilde{Y}^b_{t+1}. \tag{4.2.1}$$

The above equation gives the formula required for G_a in equation (4.1.2) for each link a.

Alternatively we can write the forecasts for the entire network using matrix notation. Let W be the matrix with the corresponding β coefficients. If $R_{ab} = 1$ then $W_{ab} = \beta_{ab}$, else both are zero. This summarises all the parameter values for one method. Let $\hat{\mathbf{Y}}_{\mathbf{t+1}}$ and $\tilde{\mathbf{Y}}_{\mathbf{t+1}}$ respectively be the vectors of the forecasts for the links using the method and the ARIMA model. Then the matrix equivalent is:

$$\hat{\mathbf{Y}}_{\mathbf{t+1}} = W \tilde{\mathbf{Y}}_{\mathbf{t+1}}. \tag{4.2.2}$$

In the single link model case no extra parameters are required. For the other methods the complexity increase depends upon two parts - the R matrix and if the parameters also vary with time. All the methods use the same R matrix and this complexity is therefore defined by the decisions made about the network structure. If we select all the links in the network as relevant for all links this results in $O(N^2)$ more parameters. The number of non-zero parameters in any other R matrix will be between these two extremes. Let n_R be the number of non zero coefficients in the matrix R. Then any method using R is of order n_R.

The methods that use the linear combinations differ in how they identify the βs. We explain the differences in calculating W here before we discuss each method in detail later in this section. *hierarchical* forecasts are generated using slightly different matrices than the matrix form of equation (4.2.2). However they can be converted to give a comparable W matrix. Both the hierarchical and the *bubble* methods select the relationships between the links before the forecast values are seen. This can be seen as equivalent generating the W matrix from predetermined weights. As they are predetermined the computational cost is less than other methods. The hierarchical method predetermines one parameter for each possible hierarchy option which remains the same for all time. The bubble method has a W matrix that is unique for each of the 672 time steps in the week as it depends upon the median values. However these are determined by $2N + 1$ parameters. The number of non zero beta values is $672n_R$ across the week or $96n_R$ for one day.

The *link beta* and the *time dependent beta* methods select the βs dependent upon the forecast values of the fitted models. The link beta fits one parameter for each non-zero entry in the W matrix and therefore has n_R parameters to fit. The time dependent beta fits n_R parameters for each time interval, up to one week. Therefore the most parameters to generate for this method is the same as the bubble method of $672n_R$.

Single link model. The first model we consider is the case where $R_a = a$ and $\beta_a = 1$, hence $W = I$. This corresponds to the single link model from Chapter 2 and is the benchmark against which we test all network models.

The ARIMA model will be different for each link, both in the number of parameters and the parameter coefficients. As this is a network problem we refer to the specific link we are forecasting on as link a. For clarity we now restate the ARIMA forecast equation with reference to the link specific parameters. The forecast for link a is

$$(1-B)^d\Big(1-\sum_{i=1}^{p_a}\phi_{ai}B^i\Big)\hat{y}_{at} = c_a + \Big(1+\sum_{i=1}^{q_a}\theta_{ai}B^i\Big)e_t. \tag{4.2.3}$$

The model is run for each link and as $\beta_a = 1$, the forecast for link a from the ARIMA model will give the overall forecast for link a, hence we use $\hat{y}$ rather than $\tilde{y}$ for the ARIMA model predictions for the rest of the linear equation methods.

Bubble model. The bubble method preselects the weight of each β given the distance between links within the network. We developed this method with the intention of reducing the computational complexity of the network model by generating the weights before the model is run. In the spatial time series literature the spatial distance between two points is often used as a measure and then all distances are scaled by an appropriate factor. However in our case the travel time is predicted for the entire link and as the vehicle can be at any point along the link, there is no single point from which to measure the distance.

The VRP requires a forecast for the whole link as if it were a single point. Condensing each link to a single point will lose some dependence structure even if additional considerations are taken. Consider a vehicle travelling along a link. Vehicles come from roads that connect to the link at the start and traffic can propagate backwards from roads the link feeds into. Hence the current influence of surrounding links at one time depends upon where a vehicle is on the link.

To estimate the effect of surrounding roads we consider a single vehicle at one point on the link that normally takes 30 minutes to traverse. The vehicle is in the middle of a link and we wish to forecast the travel time to the end of the link of another vehicle at the same location in 5 minutes time. The second vehicle is already some way into the link and hence vehicles on the feeder links before are unlikely to catch it up. If we assume constant travel time then they will always be behind. Similarly even if the links ahead stop completely the effect is unlikely to reach back to the middle within the 5 minutes.

Instead consider that the first vehicle is at the beginning of the link. The end links won't affect it but the second vehicle (and those around it) were traversing the feeder links 5 minutes ago. These feeder links give an indication of how busy it will be for a vehicle at the beginning of the link in the future and explain why we also consider the vehicles behind to affect the future travel time. At the end of the link a similar situation occurs whereby it is the end links that affect the travel time and the feeder links can be ignored. We define these influences such that that the feeder links have an influence up to δ_1 minutes in while the end links have influence from $Y_t - \delta_2$ minutes, where Y_t is the travel time for arrival time t. Beyond this point their influence is negligible.

From the point of view of the vehicle anything within a *bubble* of radius s can influence that vehicle's travel time. We then consider how much of the bubble is on each of the links as the initial values to calculate the influence of each link. Figure 4.2.1 shows the a bubble (in orange) around a blue vehicle on link a which is also spread onto links b and c.

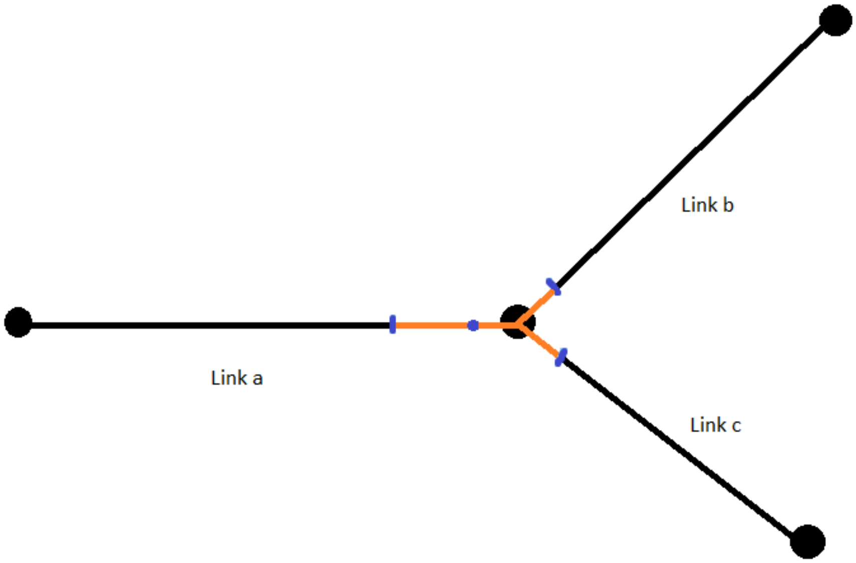


Figure 4.2.1: Bubble (shown in orange) around a vehicle located at the blue dot on link a.

However we wish to consider the initial values at every point on the link simultaneously. These spatial points are observed in time, starting at time t and ending at time $t + Y_t$ which is unknown. We estimate Y_t by $l_{\tau a}$, which we take to be the median travel time for link a at time index $\tau(t)$. However, since all the β calculated for one time point have the same τ, we drop this subscript to improve the readability.

The values of β depend upon s, the median travel times of each of the links l and the weighting α_i for the in-links and the out-links. This enables a correction to be added if feeder links have more influence that the out links, or if one link has different properties. We note that if the sum of α values for the in or the out links is greater than one then overall size of the bubble when it reaches them will be larger than one. The longer the link is the less effect the surrounding links will have as vehicles spend more time in the middle of the link where their influence is negligible is larger.

Changing s has much more effect on β than changing α_i but this affects all the β values, whereas altering α_i allows adjustments to a single β_{ai}. All the β values will be positive because they are a measure of how close a vehicle is to a link.

We generate the travel time predictions by combining the β values and their corresponding ARIMA predictions, according to equation (4.2.1). Thus if the β values are too large then the predictions for each link will be too large. Let $\tilde{\beta}_{ai}$ be the un-scaled coefficient for link i, when considering link a which we need to scale. As everything is calculated on a standardised scale an initial scaling would be for the βs to sum to one. This occurs if we normalise as follows:

$$\beta_{ai} = \frac{\tilde{\beta}_{ai}}{\sum_{b \in R_a} \tilde{\beta}_{ab}}. \tag{4.2.4}$$

The scaled values are between zero and one. The β coefficients depend upon how large the median travel times (l_i) are compared to s. For each link i in R_a we can find the formula of the amount of the bubble that is on link i at time t. By integrating this with respect to time we can find $\tilde{\beta}_{ai}$. Table 4.2.2 shows the $\tilde{\beta}$ coefficients for each possible combination before they are normalised. Thus $\tilde{\beta}_{ai}$ is the coefficient required for equation (4.2.4).

We initially use $s = 5$ minutes, or 300 seconds and choose $\alpha_i = \frac{1}{|R_a^+|}$ where R_a^+

	$l_a > s$		$l_a < s$		
$\tilde{\beta}_{aa}$	$l_a - \frac{s}{2}$		$\frac{l_a^2}{2s}$		
	$l_i > \alpha_i s$	$l_i < \alpha_i s$	$l_i > \alpha_i s$	$\alpha_i s > l_i > (s - l_a)\alpha_i$	$l_i < (s - l_a)\alpha_i$
$\tilde{\beta}_{ai}$	$\frac{\alpha_i s}{4}$	$(1 - \frac{l_i}{2s})\frac{l_i}{2}$	$\frac{l_a \alpha_i}{2}(1 - \frac{l_a}{2s})$	$\frac{l_i}{2s}(l_a - \frac{1}{2}(l_i - \alpha(s - l_a))$	$\frac{l_i l_a}{2s}$

Table 4.2.2: Table of the β coefficents for each link before normalizing.

is all the links into link a if i is a feeder link, with the same for the out links. We increase s in Section 4.4 to extend onto more links. All of the β parameters can be found in advance by using the median travel times for each link and time. This gives the W matrix for each time period.

Hierarchical model. Hierarchical forecasts are based upon the premise that the data can be categorised into multiple hierarchies of ordered importance (Athanasopoulos et al., 2017). This method hasn't previously been applied to travel time data. Each time series for the travel is categorised by multiple features such as the road type and speed limit. One feature such as an road type is a hierarchy, which has categories motorway and A-road. By forecasting each level of the hierarchy simultaneously we may be able to generate a better forecast, as we share information about one type of hierarchy across the all occurrences, rather than forecasting each combination individually. The order of these levels affects the way the forecasts are calculated. The highest hierarchy usually includes all the time series, which in our case are the individual links.

In a road network there are no obvious additional levels for the hierarchies so we have to impose artificial ones. These artificial hierarchies are an alternative way of selecting links that can provide influences on each other. Before we selected a set R_a for each link. In a hierarchy all links in a set are considered relevant to each other and links can be in multiple hierarchies. The number of hierarchies is generally much smaller than the number of links. In what follows we start with three levels - the base level whereby each link is separate, the top level which has all the roads in and for the five link network there is an addition one which groups the roads by if they go into or out of the main link.

The hierarchies are represented by a summing matrix S. Each column represents one link and each row represents one possible categorisation. Categorizations are worked from the top of the hierarchy down and represent all possible options. If link i is in categorisation k then S_{ki} is one, else it is zero. The top row will be all the links, then the next rows represent each categorisation c_2 for the second level of the hierarchy, where c_2 is an option from hierarchy 2. Each one of the next set of rows are categorised as (c_2, c_3). This is repeated until the last set of rows which are the individual links themselves.

Hierarchical forecasts generate forecasts for each of the rows of the summing matrix. If all the hierarchy level i forecasts with c_{i-1} in are added together it should give the forecast for c_{i-1}. Thus the lower estimates need to be reconciled with the upper level hierarchies.

Let $\mathbf{\tilde{Y}_t}$ be the ARIMA forecast vector for the links. The prediction for all the different Hierarchy options, $\mathbf{\hat{h}_t}$ is:

$$\mathbf{\hat{h}_t} = S(S'VS)^{-1}S'V'\mathbf{\tilde{y}_t}. \tag{4.2.5}$$

The last N terms of the vector $\mathbf{\hat{h}_t}$ give the predictions for each of the single link hierarchies, due to the ordering of S. Before we used the W matrix to show the parameter values. However the Hierarchical method has parameter values for all the different rows of the summing matrix. The corresponding W is therefore the last N rows of $S(S'VS)^{-1}S'V'$ as it also includes the coefficients relating to the higher levels of the hierarchies. We estimate the covariance matrix V using the fitted mean squared error for each series for the diagonal values and set all other terms to zero. As these are estimated using the true values rather than the forecasts from the ARIMA model, then it is considered to be preselected, as they are selected independently of the ARIMA forecasts.

Link beta. We now consider methods that use the fitted values from the ARIMA models to select the β values. The first of these is the link beta method which calculates a β_{ab} value for link b, if it is in the set R_a, as defined in Section 4.1.2. The total number of parameters for the method is therefore n_R.

The β_{ab} are selected by minimising the sum of squared errors. This uses the fitted ARIMA values for each of the links and observed values for link a as the correct values. Thus for the forecasts for the day that starts at time index $t+1$ we select the β values that minimise

$$\sum_{j=1}^{t}(y_j^a - \sum_{b \in R_a} \beta_{ab}\tilde{y}_j^b)^2. \tag{4.2.6}$$

The forecast for the link is found using equation (4.2.1) with the linear combination of the betas and the ARIMA forecasts from the optimal ARIMA models for each link.

Time dependent beta. The time dependent beta method is similar to the link beta method, except that we allow the β values vary with time. As our unstandardised data includes a weekly pattern we choose to include one value for each of the time intervals in a week. This is 672 parameters in total but forecasting one day in advance requires only 96. We therefore refer to the coefficient for link b when predicting link a as $\beta_{ab\tau}$, whereby τ is the weekly index. The other 576 parameters are used to check for the method accuracy. The sum of squares equation for all the βs is

$$\sum_{j=1}^{t}(y_i^a - \sum_{b \in R_a} \beta_{ab\tau(t)}\tilde{y}_j^b)^2. \tag{4.2.7}$$

This equation has $672n_r$ different parameters which is a lot of parameters to optimise at once. However because the $\beta_{ab\tau}$ values are independent for each τ we can calculate them separately, using only the values of t that have value τ. This reduces the number of parameters to n_r for each of the weekly index points. However it also reduces the number of points we have to calculate the parameters as compared to the link beta method. Whereas for that method we could use all the data points to fit the β values, now we have to divide that by 672 as we can only use data points with the correct time index. This means that the method can potentially result in over fitting.

The required number of values to calculate increases by a factor of 96 from the link beta method. Thus the total number of parameters is $96n_R$.

Summary of linear combination methods. We summarise the above methods in Table 4.2.3 by the number of additional parameters and whether they are preselected. The bubble and the time dependent beta methods have the most parameters but because the bubble is preselected it should be quicker. The more parameters there are, the more likely a model is to be over fitted due to picking up trends in the data that don't exist. The bubble parameter values are limited to be between zero and one, whereas the others can take any value.

Method	Additional parameters	Pre selected
single link ARIMA	0	NA
bubble	96 n_R	Y (median travel time)
hierarchical	N^2	N (SoS)
link beta	n_R	Y (covariance)
time dpt beta	96 n_R	N (SoS)

Table 4.2.3: Table of linear combination model characteristics. The additional parameters are the number of non zero β parameters in W.

4.2.2 Multivariate models

The second set of models we look at are multivariate models. These methods use the data for several links simultaneously to generate parameters. Some of these parameters use the data from all the links, while others subset the links to reduce the problem size. The previous methods use linear combinations of forecasts for models which are generated on single links.

As the ARIMA model is the best model in the single link setting we begin by adapting this to the multivariate setting. One multivariate form of the ARIMA model is called VARIMA, the *vector autoregressive integrated moving average.* The travel time predictions for each link are calculated in a vector with connections between the links created by a matrix. However due to the computational complexities involved in calculating this model we simplify it to a VAR model. This removes both the differencing within the model calculation and the MA terms.

In Section 2.4.1 the optimal ARIMA model has 2 AR and 1 MA term for the 25th March but considering averages of the ARIMA models used to forecast all 78 days and six links leads to in an average of 1.7 AR terms and 2.06 MA terms. This would suggest that we should include the MA terms but without running the ARIMA models again specifying to only include the MA or AR terms we cannot be sure which to use. We therefore begin by using only AR terms with the potential to change to a using the MA terms if the models perform poorly.

The VAR model uses autoregressive terms across multiple variables. In the case of the petrol routing problem each link is classed as a separate component of the vector. We represent the connections between links by considering their correlations.

As before we define $\mathbf{Y_t}$ as the observed vector of travel times for all links at time t, with $\hat{\mathbf{Y}}_\mathbf{t}$ as the predictions for the same time. In the single link ARIMA the coefficients for the AR and MA terms are dependent only upon the link a. However in the multivariate case some parameters also involve other links. In the previous section we defined W as the matrix that includes the β coefficients. In the ARIMA model ϕ parameters represent the AR part of the model and θ the MA part. These are modelled separately and both can include the other links. The ϕ terms depend upon lags so we therefore define B_i as the matrix of ϕ coefficients for lag i (Christiano, 2012). Then, if c is a constant, the VAR model is:

$$\mathbf{\hat{Y}_{t+1}} = c + \sum_{i=1}^{p} B_i \mathbf{Y_{t-i+1}} + \mathbf{e_t}. \tag{4.2.8}$$

The relationships between the links are governed by the matrix B_i which is unique for the lag. The combination of these matrices form the autoregressive terms and are generated up to P lags. The larger P is the longer the method takes to run because there are more variables.

This model has N^2p parameters that need to be calculated simultaneously across the network but some of these can be set to zero, similar to the linear combination models. We start with the simplest model and then increase the model complexity.

Constrained VAR. A simple way to reduce the number of parameters in the VAR model is to preselect the parameters we think are going to be significant. All other parameters are automatically set to zero. We call this method the constrained VAR method.

We use the same R matrix as the linear combination models in Section 4.2.1 that identifies which links are important to each other as the input. This applies to each of the p matrices. The number of parameters for this simplified method is therefore $|R|p$.

The input matrix predefines which coefficients will be zero. The coefficients of row a of the B_i matrices are selected by calculating

$$(X'_{r_a} X_{r_a})^{-1} X_{r_a} Y_a. \tag{4.2.9}$$

Here X_{r_a} is the matrix of the travel time values at the correct order for all links that are non-zero in the R matrix and Y_a is vector the observed travel times on link a.

Sparse VAR. Another way of reducing the number of parameters needed for the full VAR models is the sparse VAR method. This selects a model with at least some of the parameters being zero.

The coefficients for each row of the matrices B_i are found separately. However the different lags must be considered at the same time. A generalised linear model is fitted for each link using all the data by using a penalised maximum likelihood. One penalised maximum likelihood is the LASSO which selects a model with some of the coefficients set to zero (Tibshirani, 1996). This model is selected by minimising:

$$\frac{1}{\sum_a K_a}\sum_{a=1}^{N}\sum_{j=1}^{K_a}(y_a(j) - \sum_{p=1}^{P}\sum_{i=1}^{K_a} y_a(j-p)\gamma_{ai})^2 + \lambda||\gamma||_1. \tag{4.2.10}$$

The value of λ is problem-specific and is selected in the model by cross-validation. The row B_i matrices are reduced to a vector γ and K_a is the total number of observations for link a. The ℓ_1 norm for the γ vector considers the value of the differences in each γ_{ai} value and thus favours models with a coefficient vector that has a high number of zeros.

VAR. For the full VAR model we calculate B_i without any assumptions on variables being zero. This is calculated using equation (4.2.9) except all links are used to calculate the parameters for each row a.

Hence there are N^2p parameters to calculate over the whole network at once. As N gets larger this becomes much more computationally prohibitive. We can use both the predictions from this and the corresponding B_i matrix to check whether the constrained model and the sparse model provide similar predictions and that the parameters that have been set to zero in these models are close to zero in the full model.

VARMA. The previous models three models in this section use VAR models. However we can also consider a VARMA model which includes moving average terms as well as autoregressive ones which is much closer to the ARIMA model selected for the single link case. Let E_i be the coefficient matrix for the moving average terms at lag i. Then the VARMA model is:

$$\hat{\mathbf{Y}}_{\mathbf{t+1}} = c + \sum_{i=1}^{p} B_i \mathbf{Y}_{\mathbf{t-i+1}} - \sum_{i=1}^{q} E_i \mathbf{e}_{\mathbf{t-i}} + \mathbf{e}_{\mathbf{t}}. \tag{4.2.11}$$

The added complexity of this model makes it impractical both to run in an everyday setting and to generate the forecasts required to compare this method to the others over the 78 days of forecasts.

There are of course other multivariate models such as the STARIMA model and its variants. The STARIMA model is a spatial and temporal version of the ARIMA model. Khan et al. (2012); Min et al. (2009) both study the travel flow using STARIMA models while Salamanis et al. (2016) use a variant to evaluate travel time prediction at different spatial lags. The spatial part of the STARIMA model comes from the weighting matrix, S, which is multiplied by the corresponding AR and MA parameters. The spatial order represents the how far two links are away from each other, so S^k_{ij} will be non-zero only if the links i and j are k away from each other. Each row sums to 1. There is a single temporal spatial AR or MA coefficient regardless of the spatial location. The STARIMA model, when ∇^d is the differencing can be written as:

$$\nabla^d \hat{\mathbf{x}}_{\mathbf{t}} = \sum_{i=1}^{p} \sum_{k=1}^{m_i} \phi_{ik} S^k \nabla^d \mathbf{x}_{\mathbf{t-i}} - \sum_{i=1}^{q} \sum_{k=1}^{n_i} \theta_{ik} S^k \mathbf{e}_{\mathbf{t-i}} \tag{4.2.12}$$

Other versions include the GSTARIMA model or Generalized STARIMA model in Min et al. (2010) replaces the MA and AR terms by matrices while the MSTARMA model in Min and Wynter (2011) is used to predict travel time by using considers the speed a vehicle is travelling to generate the spatial matrix. This model has no spatial lag and hence is a simpler version of the STARIMA model. However they aren't applicable in this case because even the simpler MSTARMA requires further simplification to be able to estimate the number of parameters. As a result, in a VRP setting the number of parameters to estimate and the work required to generate them is too high, and we don't consider them further in this chapter.

Summary of all the methods. The linear combination methods are all of the form:

$$\hat{\mathbf{Y}}_{\mathbf{t+1}} = W\tilde{\mathbf{Y}}_{\mathbf{t+1}}.$$

They first use an ARIMA model on each link to generate the forecasts $\tilde{\mathbf{Y}}$ which is then multiplied by the matrix W to calculate the overall forecasts. The methods differ in how W is calculated.

Method	W	Preselected values
ARIMA	I	NA
bubble	$f(\alpha, s, l^{t+1})$	l-travel time, s, α vector
hierarchical	$S(S'VS)^{-1}S'V'$ (last N rows)	S-summing matrix V-covariance matrix
link beta	$\min \sum_{j=1}^{t}(y_i^a - \sum_{b \in R_a} \beta_{ab}\tilde{y}_j^b)^2$	
time dpt beta	$\min \sum_{j=1}^{t}(y_i^a - \sum_{b \in R_a} \beta_{ab\tau(t)}\tilde{y}_j^b)^2$	

Table 4.2.4: Table of linear combination model characteristics.

We then look at the VAR based methods which are based upon the equation

$$\hat{\mathbf{Y}}_{\mathbf{t+1}} = c + \sum_{k=1}^{p} W_{\phi k}\mathbf{Y}_{\mathbf{t-i+1}} + \mathbf{e}_{\mathbf{t}}. \tag{4.2.13}$$

To directly compare models of theses to the linear combination methods we need to rewrite the linear combination equation to be in terms of $\mathbf{Y}$ rather than $\tilde{\mathbf{Y}}$. However ARIMA models of the form of equation (4.2.3) are very difficult to rearrange into a form that isolates $\tilde{Y}$ because they include differencing. Thus direct comparison is very difficult.

In theory the VAR methods use all of the data at once to generate the B_i matrix. However the methods used to select the values of the B_i matrices mean that each link can be calculated separately. The R matrix is used as an input into the constrained VAR method only.

Method	row a of B_i for link a
constrained VAR	$(X'_{r_a} X_{r_a})^{-1} X_{r_a} Y_a$, r_a from R
sparse VAR	$\min \frac{1}{N} \sum_{j=1}^{N} (y_j - \sum_{p=1}^{P} \sum_{i=1}^{N} y_{j-p} \beta_{ai})^2 + \lambda \|\|\beta\|\|_1$
VAR	$(X'X)^{-1} X Y_a,$

Table 4.2.5: Summary of VAR methods.

4.3 Applying network travel time forecasting methods to real data

To illustrate the differences between the methods we initially test the methods on a small subset of the test network with only five links. This enables us to compare the methods using real data that could be used in a full scale petrol routing problem. A link is a directional section of road that the vehicle travels along, with nodes at intersections between links. The network consists of links 10.11, 11.24, 9.10, 6.10 and 11.16 which were introduced in Section 2.3.3. These are the connecting links of link 10.11. Figure 4.3.1 shows how the links are connected, with two in links and two out links. The four connecting links are also connected to other links outside the test network.

In order to use the network methods we need to define R. We put the first link in R as Link 10.11. As there are only five links all links are either first degree neighbours, second degree neighbours or cannot connect due to being directional. Hence we select R as the matrix of all links that are first degree neighbours. Therefore:

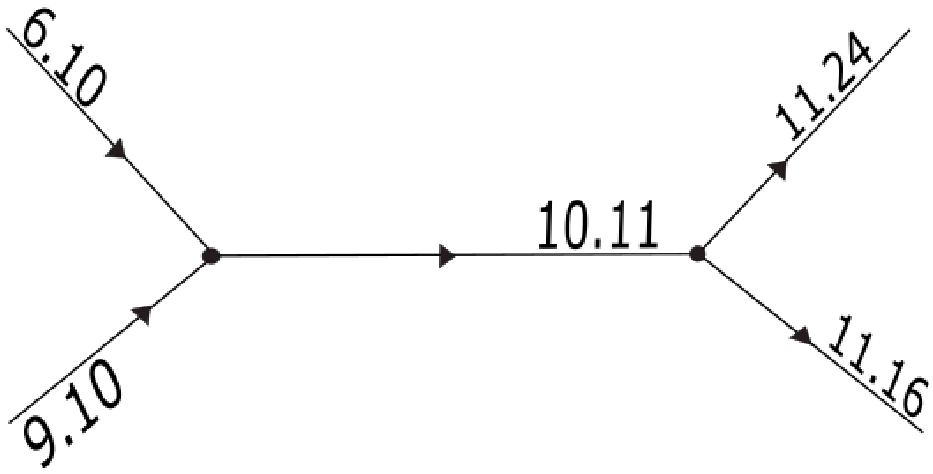


Figure 4.3.1: The five link network.

$$R = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \end{bmatrix}. \tag{4.3.1}$$

Link 10.11 is the first link and is connected to all others and itself. The other links are only connected to link 10.11 and themselves. For example $R_{9.10} = \{9.10, 10.11\}$.

This section is organised as follows. We first discuss the data formatting that we used to be able to apply the methods. Then we apply the forecasting methods to the five link network and analyse the results. Initially we visually inspect the forecasts for the different methods on link 10.11 for the 25th March 2016, which was Good Friday, and then the other four links for the same day. We also compare the β values for the methods and the computational time taken to generate these results before looking at the error metrics for all the days we can forecast.

4.3.1 Data formatting

Some of the network methods require that there is no missing data in the travel time inputs. Thus once a missing value appears in the dataset, it is impossible to generate any estimates using that method from that point on. If a model uses all the links at once then once one link has a missing value we cannot predict any of the links. In order to be able to continue to predict using these methods after a value is missing we replace the missing values with an estimate.

To do this we need to provide a value that is sensible given the rest of the travel time values and is in keeping with the magnitude of the surrounding values. One possibility is to set all the missing values to zero, however this may bias the dataset in a particular direction. Given that the ARIMA model provides a good model for the single link model we propose using estimates from the ARIMA model.

To ensure that the data values remain the same from one day to the next we will use the ARIMA model from the last day of the dataset to fill in the missing values. This generates complete time series for each link for the entire period. These will be considered as the observed travel time values for the remainder of this chapter.

4.3.2 Forecasts for 25th March

We have applied all of the methods discussed in Table 4.2 to a small network of five links. Due to the complexity of the method resulting in prohibitively long computational times the VARMA, the STARIMA and associated methods haven't been predicted.

We initially focus upon forecasting the 25th March for link 10.11 because this is the only link in the network that has all its connecting links to use to forecast. This is also the same link that we focused upon in Chapter 2. As discussed in Section 2.4.1 the 25th March is a bank holiday and hence we would expect this not to be a typical day. However it is the last day in the three month period and hence has the most data to forecast with. To keep consistency with Chapter 2 we use this day as an illustrative example, but make conclusions across many days and multiple links.

To calculate the coefficients for the three VAR versions we use the MTS package in R (Tsay, 2015). The sparse VAR and constrained VAR also use the fastVAR package in R (Wong, 2012). The ARIMA coefficients are calculated as in Section 2.4.

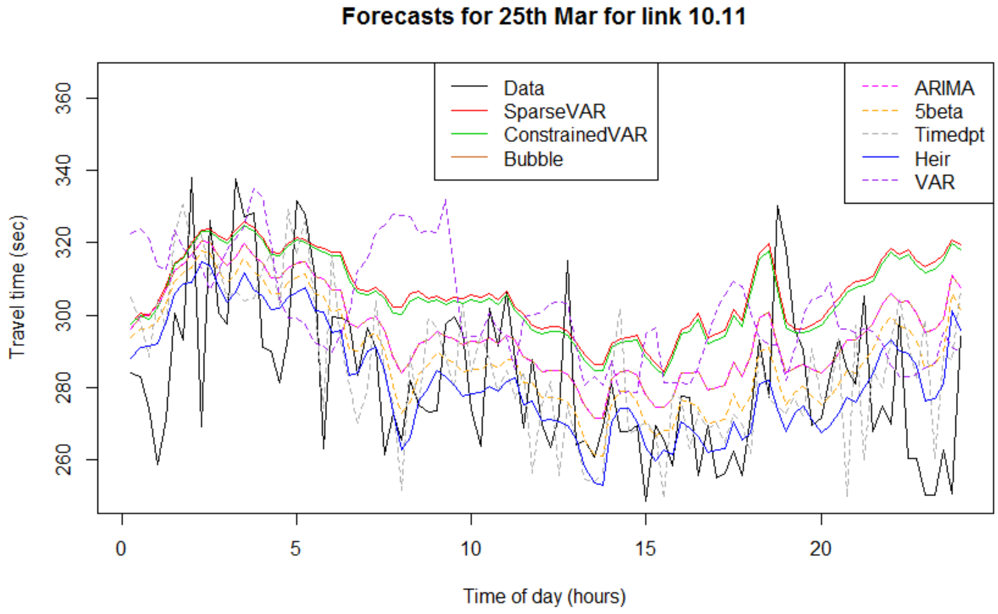


Figure 4.3.2: Forecasts for all the methods of link 10.11 for 25th March.

Figure 4.3.2 shows the predictions for all the methods for link 10.11 for the 25th of March 2016. All the methods overestimate the travel times. This could be because 25th March is a bank holiday and the traffic patterns may be different. Many of the methods predict a similar structure for the day, with the difference being how high the travel time predictions are. The VAR has a unique structure which overestimates between about 7am and 9am, when rush hour would be. The time dependent beta is a lot less smooth than the other methods. For some time points it is a lot closer to the observed travel time, while in others it is further away. A visual inspection would imply that either the Hierarchical or time dependent methods are best for link 10.11 for the 25th March.

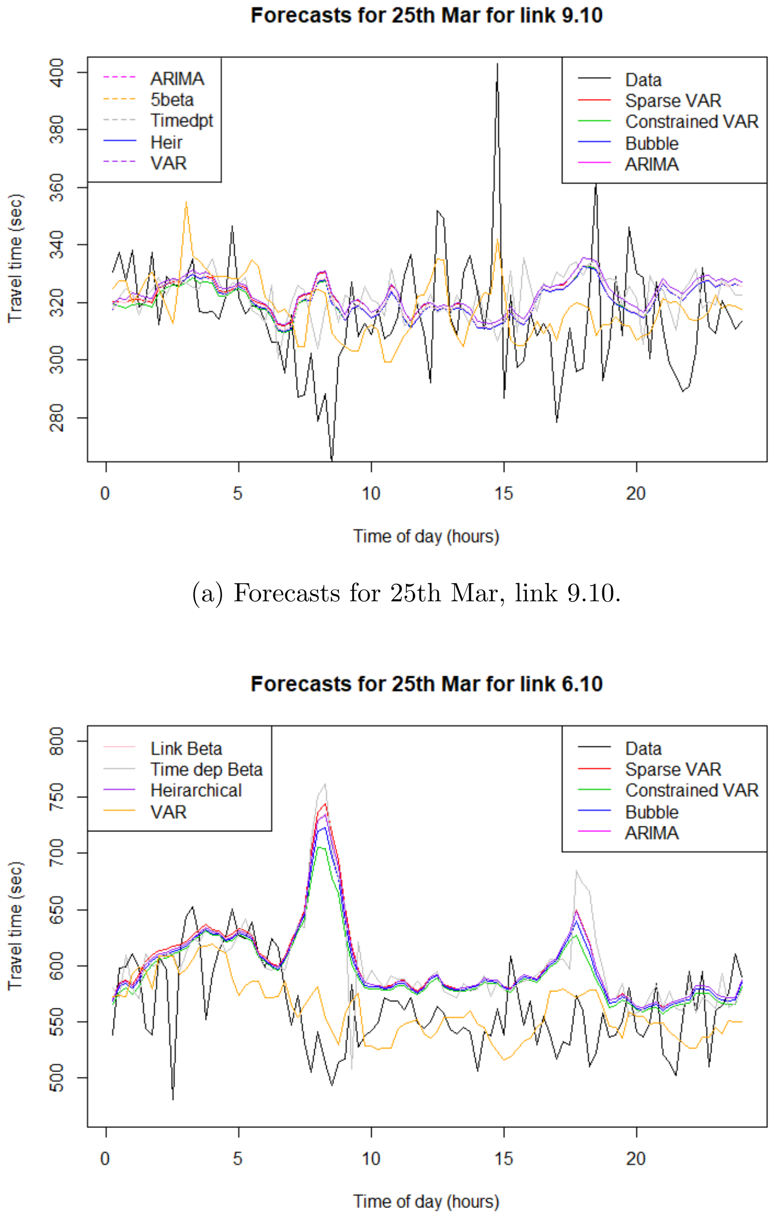
We now look at the other four links as shown in Figures 4.3.3 and 4.3.4. For three of the links the VAR model is better as all of the other methods predict the travel time to be much longer around the periods when rush hour would be expected. All the methods are fairly similar for link 9.10, where the travel time predicts remain within a 50 second band for the entire day. None of the methods predict the much lower travel time at around 8am or the sudden jump just before 3pm. The second is to be expected as it is much higher than the surrounding values, and may be high enough to be an outlier. We have previously discussed identifying and modelling outliers in Chapter 3.

Beta variations. We now examine how W, which contains the β values, vary for each linear combination method. Every day the model is run we get a different W matrix. Let W_d matrix is comprised of the estimates of the β parameters for forecasting day d. We start with the ARIMA method, which only uses the link it is forecasting and is the same for each day so

$$W_d = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}, \quad \forall d \in \{1, \ldots, 78\}. \tag{4.3.2}$$

As mentioned in Section 4.2.1 for the Link Beta and Hierarchical the W_d matrices, and hence the β parameters, remain the same over time within the day, whereas the values for the bubble method and time dependent beta methods have different β values for each time point within the week. When we consider the daily variation as well there are 96×78 different $W_{d\tau}$ matrices in total, where τ is the 15-minute index over the day.

We choose to summarise all the W matrices, of all the values over time, by two matrices, one for the mean, which we denote $\overline{W}$ and the other the standard deviation,

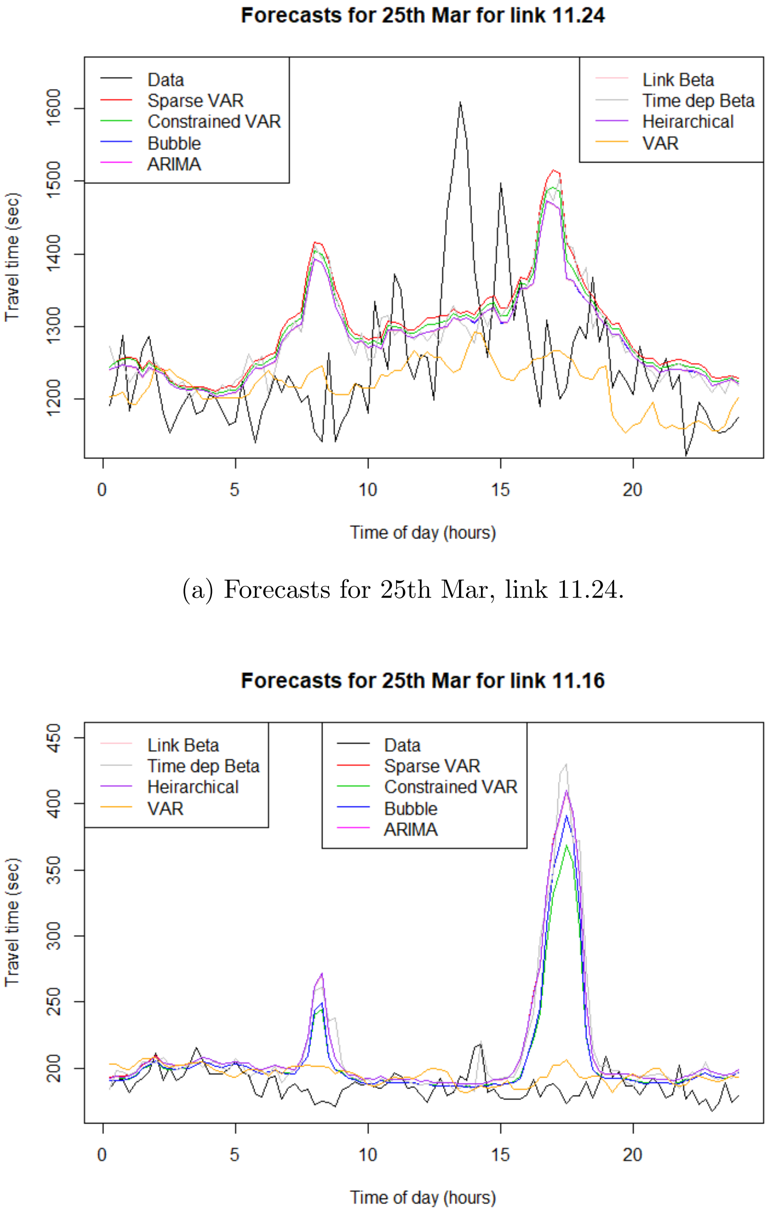


(a) Forecasts for 25th Mar, link 9.10.

(b) Forecasts for 25th Mar, link 6.10.

Figure 4.3.3: The forecasts for the 25th of March for the links 9.10 and 6.10.

which we denote W_σ. These have been chosen to give an idea of where we expect the β values to be and how much they vary by. By removing the τ dimension from the $W_{d\tau}$ matrices we lose the ability to summarise the β values at each of the 96 time points but this allows an easier comparison between the five methods in terms of forecasting performance.



(a) Forecasts for 25th Mar, link 11.24.

(b) Forecasts for 25th Mar, link 11.16.

Figure 4.3.4: The forecasts for the 25th of March for the links 11.24 and 11.16.

We use the mean of W to get an idea of what the true values of W are. Each W_d matrix is itself an estimate of the β values for time t. For each time there exist true parameters which we will never observe. We calculate $\overline{W}$ to estimate the mean of the true β parameters. If we assume that W_d is constant through time then the mean is an asymptotically unbiased and consistent estimator of W_d and hence we can use delta method confidence intervals of where the mean of each parameter should lie. This gives us a 95% significance level that if two methods, (1) and (2) are significantly different if they aren't within two standard deviations of each other. While the assumption that W_d is constant may not be true, particularity when considering $W_{d\tau}$ this still provides a clear significance level of whether or not two parameters are the same.

We compare all methods to the single link W_d as this is the base model. Any parameters that have the corresponding parameter from matrix (4.3.2) outside their significance level are indicated in bold.

Equation (4.3.3) shows the mean 5 link beta W matrix, and the corresponding standard deviation of W. The majority of the weight is on the β_{aa} value for each link but the means for all the β_{ab} are non zero. However all of these parameters are within two standard deviations of zero. The first W value is 0.75 and one is just inside two standard deviations of this. Therefore the method suggests that while the contribution from link 10.11 is less than itself it is not significantly so. None of the other values are significantly different from matrix (4.3.2) at the 95% significance level. When combined with the sum of the column for 10.11 being 1 this suggests that R_a may be the correct size.

$$\overline{W} = \begin{bmatrix} 0.75 & 0.16 & 0.05 & 0.05 & 0.15 \\ 0.16 & 0.74 & 0 & 0 & 0 \\ 0.12 & 0 & 0.94 & 0 & 0 \\ -0.02 & 0 & 0 & 0.92 & 0 \\ -0.01 & 0 & 0 & 0 & 0.92 \end{bmatrix}, \quad W_\sigma = \begin{bmatrix} 0.15 & 0.14 & 0.13 & 0.11 & 0.32 \\ 0.14 & 0.32 & 0 & 0 & 0 \\ 0.21 & 0 & 0.11 & 0 & 0 \\ 0.05 & 0 & 0 & 0.05 & 0 \\ 0.11 & 0 & 0 & 0 & 0.15 \end{bmatrix}. \tag{4.3.3}$$

By construction all the columns of the mean bubble W matrix in equation (4.3.4) sum to one. The standard deviations are also very small which is to be expected because the median travel times vary only slightly over the days and hence the β values can vary only slightly. Therefore all parameters are significantly different to the ARIMA parameters in matrix (4.3.2). This method, when $s = 300$ and $\alpha = 0.5$, puts a lot of weight onto the other β_{ab} values, as opposed to β_{aa}.

$$\overline{W} = \begin{bmatrix} \mathbf{0.51} & \mathbf{0.18} & \mathbf{0.08} & \mathbf{0.03} & \mathbf{0.33} \\ \mathbf{0.12} & \mathbf{0.82} & 0 & 0 & 0 \\ \mathbf{0.12} & 0 & \mathbf{0.92} & 0 & 0 \\ \mathbf{0.12} & 0 & 0 & \mathbf{0.97} & 0 \\ \mathbf{0.12} & 0 & 0 & 0 & \mathbf{0.67} \end{bmatrix}, \quad W_\sigma = \begin{bmatrix} 0.02 & 0.01 & 0.01 & 0.00 & 0.05 \\ 0.01 & 0.01 & 0 & 0 & 0 \\ 0.01 & 0 & 0.01 & 0 & 0 \\ 0.01 & 0 & 0 & 0 & 0 \\ 0.01 & 0 & 0 & 0 & 0.05 \end{bmatrix}. \tag{4.3.4}$$

Equation (4.3.6) shows that the time dependent standard deviation matrix has the highest values. This is because it has a lot of parameters to fit and the best values for these were of much higher absolute magnitude when there was less data to fit them. The sum of the columns are very different than the 5 link Beta and the bubble methods. For link 10.11 it is 2.5 and for the last link nearly 6. If all the links have initial predictions are about the same size and magnitude the overall travel times will be much larger, which is a very large change to the original ARIMA predictions. Both zero and one are within two standard deviations of all of the β parameters in matrix (4.3.6) and hence none of them are statistically significant at the 95% level from matrix (4.3.2).

$$\overline{W} = \begin{bmatrix} 0.79 & -0.35 & 0.52 & -1.03 & 0.13 \\ 0.92 & -0.09 & 0 & 0 & 0 \\ 0.49 & 0 & 0.10 & 0 & 0 \\ 0.48 & 0 & 0 & 1.27 & 0 \\ -0.22 & 0 & 0 & 0 & 5.69 \end{bmatrix}, \tag{4.3.5}$$

$$W_\sigma = \begin{bmatrix} 76.56 & 33.08 & 38.45 & 153.76 & 103.24 \\ 50.27 & 36.66 & 0 & 0 & 0 \\ 40.19 & 0 & 35.56 & 0 & 0 \\ 23.97 & 0 & 0 & 139.42 & 0 \\ 19.69 & 0 & 0 & 0 & 443.77 \end{bmatrix}. \qquad (4.3.6)$$

Because of the way that the Hierarchical method is calculated there are β values for each of the links in equation (4.3.7). However these values are determined by fewer parameters. We present the β values to enable comparison between the methods.

As a result of the different method to calculate the β values their magnitudes are very different when compared to the other methods. With the exception of the third and fourth links the biggest value are for the links themselves. Most of the other parameters are small and if the magnitude of β_{ab} is large then β_{ba} is also large. The method of calculation also ensures that the standard deviations are low, in some cases zero to 2dp. As a result of the small standard deviations in matrix (4.3.8) all but two of the parameters are statistically significant from the single link model of matrix (4.3.2) at the 95% level.

$$\overline{W} = \begin{bmatrix} \mathbf{0.30} & \mathbf{-0.13} & \mathbf{-0.13} & \mathbf{-0.11} & \mathbf{-0.11} \\ \mathbf{-0.07} & \mathbf{0.69} & \mathbf{-0.31} & \mathbf{0.01} & \mathbf{0.01} \\ \mathbf{-0.12} & \mathbf{-0.53} & \mathbf{0.47} & \mathbf{0.01} & \mathbf{0.01} \\ \mathbf{-0.11} & 0.02 & 0.02 & \mathbf{0.45} & \mathbf{-0.55} \\ \mathbf{-0.06} & \mathbf{0.01} & \mathbf{0.01} & \mathbf{-0.31} & \mathbf{0.69} \end{bmatrix}, \qquad (4.3.7)$$

$$W_\sigma = \begin{bmatrix} 0.06 & 0.02 & 0.02 & 0.02 & 0.02 \\ 0.02 & 0.05 & 0.05 & 0.00 & 0.00 \\ 0.03 & 0.07 & 0.07 & 0.00 & 0.00 \\ 0.02 & 0.01 & 0.01 & 0.06 & 0.06 \\ 0.02 & 0.00 & 0.00 & 0.07 & 0.07 \end{bmatrix}. \qquad (4.3.8)$$

The average β values which can be found in the $\overline{W}$ matrices of all the methods lie between -1.03 and 5.69. Both of these are from the time dependent method with also has very high standard deviations. The high variance is in part due to the over-fitting that occurs, resulting in a much wider variation of β parameters since for each of the 78 days there are 96 possible values.

With the exception of the Bubble method which is constricted to positive β values all the methods have at least one negative mean β value although for the 5 link beta method these are very close to zero. Only the Bubble method and the Hierarchical method are statistically significant from the single link model W_d at the 95% level. However some parameters in the link beta method are only just within this window and for all methods the mean $\overline{W}$ has none zero values for all parameters in R_a. It is also possible that W_d is not constant through time.

Computational time. In this section we consider the computational time each method takes to run. This is taken as a proxy for computational complexity and is presented in Table 4.3.1. Generally the longer the time series the longer a method takes to run, however the forecasts for some days will be shorter than others. For this reason we present the time for all the days in the time period January-March 2016. From this it is clear which methods are faster. All times are given in seconds. The following times are the total time for running on a HPC cluster.

The time dependent beta method takes the longest to run, followed by the VAR method. The ARIMA, link beta, Hierarchical and Bubble all take approximately the same time.

In order to see how the computational time varies with the size of the network we have run all the models for two additional sizes, one network with 10 links and another with 20 links. The 20 link network results are explored in more detail in Section 4.4. The results are shown in Figure 4.3.5. To enable easier comparison we have grouped the majority of the linear combination methods together in one colour,

	Time(seconds)	Time (minutes)
Sparse VAR	170.69	2.84
Constrained VAR	919.27	13.32
Bubble	1264.158	21.07
ARIMA	1233.769	20.56
Link beta	1136.094	18.93
Time dpt	2485.886	41.43
VAR	1937.026	32.28
Hierarchical	991.253	16.52

Table 4.3.1: Time taken to run each network method for 5 link network.

with the exception of the time dependent method which is clearly larger. As this method needs to calculate beta values for all 696 time periods for each link in the network it follows that it has a greater computational time than the other linear combination methods.

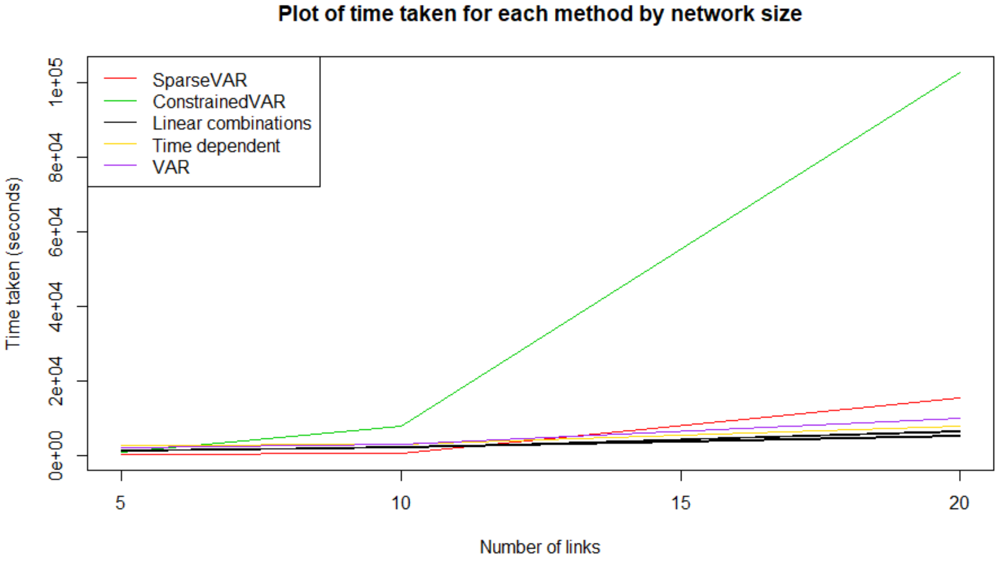


Figure 4.3.5: Comparison of computational time for each method over different sized networks.

All methods increase in computational time with the number of links. The increase for the VAR based methods is greater, with the Constrained VAR going from one of the quickest to the slowest by a considerable margin. The gradient of both lines between the 5 link and the 10 link times and the 10 link and the 20 link times is steepest for the constrained VAR. By comparison at 10 links the sparse VAR is still the quickest but at 20 links is the second slowest. The graph supports the view that while at small networks the VAR methods have quick computational times they cope very poorly when the network size is increased. As a VRP network will be larger than 20 links the Constrained VAR method is too computationally intensive to be used in predictions for the petrol routing problem.

4.3.3 Model diagnostics and performance

This section describes the forecasting performance in terms of the error metrics found in Section 2.5 where a detailed explanation of their selection is given. As before we look at the performance of both the fitted model over all 78 series that were used to fit the model and the forecasting performance for the 78 days.

Fitted summary. Table 4.3.2 shows the error measures for each of the methods for the fitted values of all the models. For all links the best fitted method is the time dependent beta. This is because when there is less data the method has a lot of parameters so overfits the data.

The VAR, Sparse VAR and Link Beta methods all perform better than the ARIMA method over all the measures. The worst fitted model is the Bubble method which is worse than the single link ARIMA on all but the Median Relative MAE.

	AvgRelMAE	RelMedianAE	MedianRelMAE
SparseVAR	0.96673	0.95722	0.99992
ConstrainedVAR	0.97668	0.95822	1.00186
Bubble	1.04841	1.01779	0.99912
ARIMA	1.00000	1.00000	1.00000
Link beta	0.97868	0.97339	0.99984
Timedpt	**0.00000**	**0.00000**	**0.68878**
VAR	0.87067	0.83152	0.99916
Heir	0.99998	1.00028	1.00000

Table 4.3.2: Fitted error measures table for all 5 links.

Forecast summary. While a good fit is important we want the forecasts to perform well. Table 4.3.3 shows the forecast error measures for the methods described in Sections 4.2.1 and 4.2.2. These are presented for all the links combined together into a single value for each error metric. The best method for forecasting each link varies from method to method, and the worst is the time dependent one. This is due to the time dependent model having too many parameters and therefore overfitting the data.

The only method that is better for all three error metrics is the Link Beta. The VAR, SparseVAR and Hierarchical methods are all worse or no better than the ARIMA model.

These remarks shows the importance of considering the metrics over all the days rather than the visual inspections of the 25th of March which implied that the VAR or time dependent methods were the best.

	AvgRelMAE	RelMedianAE	MedianRelMAE
SparseVAR	1.02186	1.02831	1.00150
ConstrainedVAR	0.99916	1.00418	**0.99787**
Bubble	1.00100	1.00510	0.99976
ARIMA	1.00000	1.00000	1.00000
Link beta	**0.99471**	**0.99396**	0.99976
Timedpt	1.47862	1.27560	1.14066
VAR	1.02176	1.02784	1.06082
Heir	1.00010	1.00017	1.00000

Table 4.3.3: Forecast error measures table for all 5 links.

Concluding remarks for the 5 link dataset. The forecast summary suggests that the best method overall is the Link beta method. This method is moderate in terms of the computational time (Table 4.3.1) and was also better than the ARIMA method on all three fitted error metrics (Table 4.3.3.)

4.4 Range of influence

In Section 4.1.2 we discussed the idea of limiting the number of links used to forecast link a to a set R_a. We now investigate how big R_a should be to be optimal in terms of the forecasting performance. We term the *range of influence* as how far away the links have influence on the travel time. We denote the number of links away that we are including by the parameter c. Once we have found the correct range of influence we have a starting set for R_a.

For the five link network we only considered $c = 1$ (and $c = 0$ for the single link model) so that only links that were first degree neighbours were included. Due to the arrangement of these five links any larger c would require us to consider links outside the network and hence it was necessary to limit c this way. However in a petrol routing problem the size of the network will be much larger than 5 links and hence c could be larger than one.

If we include more links, in the set R_a, the model will predict β parameters for them which are likely to be non-zero. If these parameters are very close to zero then we can assume that their influence is negligible and remove them from the model. The greater the number of parameters the more risk there is of over-fitting, as we found with the time dependent model in Chapter 2 and the model will have a much larger computational time. Hence if the predictions are worse, or the same, then we will automatically select the simpler model. Any predictions must be produced within a reaonable time and hence the selection of c for use within the model is a balance between what may be the true range of influence and the ability of the model to calculate accurate predictions.

Let R_a^c be the set of links that are at most c-degree neighbours from link a. If c is too small then our predictions will be less accurate. In this section we initially consider different sizes of c to find the range of influence by examining how changing c affects W and the overall prediction accuracy. The range of influence provide an easy way to identify a set of links to use.

However there may still one or two links outside the range of influence that are important when forecasting link a, for example if a link is very short. Conversely some of the parameters within the set may be very close to zero. We therefore discuss how we can identify and select alternative sets to consider. These alternative sets include the hierarchies that the Hierarchical forecasting model uses.

We also have to consider the bubble method slightly differently. A link is ignored if it is too far away in terms of s. When the model is run with the set R_a^1 all the links that are first degree neighbours are included and if the bubble extends beyond the end of a link, the extra part of the bubble is ignored. This extra part occurs if the length of a link is less than $\alpha_i s$ but the majority of links in the dataset are longer than this. Thus if we increase c most of the extra links will be too far away without either an increase of either s or α_i.

The size of α_i is a measure of how important link i is to link a without considering the bubble. At most half of the bubble can be on link i, whereas the full bubble can be on link a if it is long enough. If there are multiple input and output links then the bubble will cover all of them at the same time. Therefore it should be split between them. This means that the maximum value of α_i is one and hence to increase the neighbourhood we need to increase s.

For this section we introduce a 20 link network as we need to be able to compare links that have at least one link between them.

New network. To check how large the neighbourhood of influence is we need to use a larger network than our current 5 link network. We have selected a 20 link network which includes the original five link network.

Figure 4.4.1 shows a ten link network, each of which is bi-directional, representing the 20 link network. It is impossible to show 20 bi-directional links on the map as the two directional links occupy the same space on this scale. All of the links in the five link network have additional connections to Section 4.3.2. In the case of link 10.11

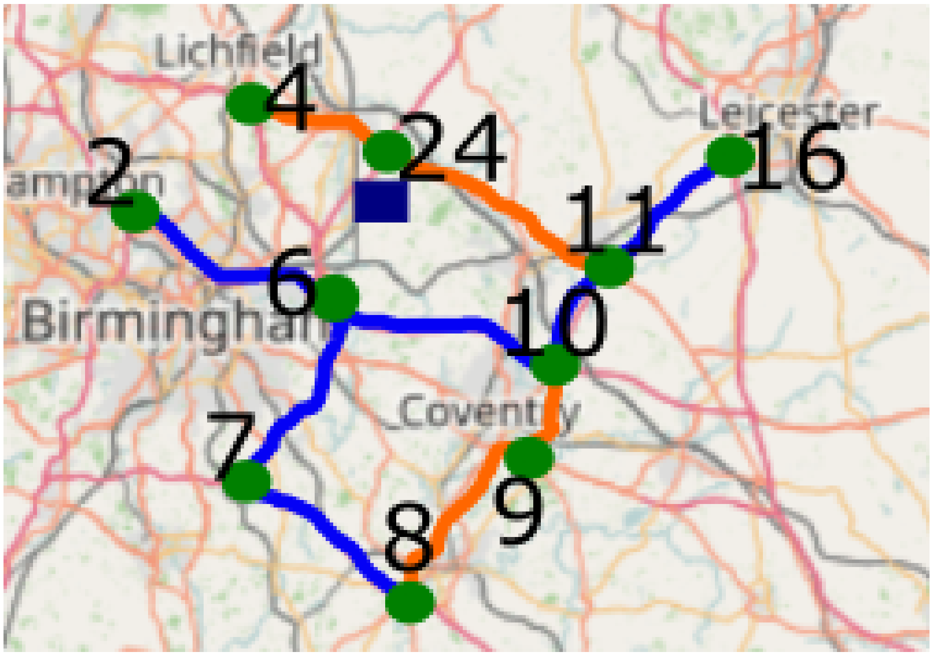


Figure 4.4.1: The 20 link network. Each line between a pair of nodes represents two directional links in opposing directions. A link from node i to node j is named $i.j$, whereas the reverse link is $j.i$. Map background OpenStreetMap contributors (2017).

this is because the 20 network includes the reverse link 11.10. Despite being a first degree neighbour link 11.10 was neglected from the 5 link network due to the fact that very few vehicles immediately travel back on themselves. The other four links now include the connections that were missing from the 5 link network due to them being further away from link 10.11.

To be more concise instead of reporting R^1 for when $c = 1$ and R^2 for when $c = 2$ we instead summarise them together in $\tilde{R}$. In $\tilde{R}$ any links that are in $c = 1$ are coded one while any links that are second-degree neighbours are coded 2. This allows us to clearly see the differences between the two as R^1 would replace any twos with a zero whereas R^2 would replace the twos with a one. As before we select the first link as link 10.11 so for the 20 link network the matrix is:

$$\tilde{R} = \begin{bmatrix}
1 & 1 & 1 & 1 & 1 & 1 & 2 & 0 & 0 & 2 & 0 & 0 & 2 & 0 & 2 & 2 & 2 & 0 & 2 & 2 \\
1 & 1 & 2 & 2 & 2 & 2 & 0 & 2 & 2 & 0 & 2 & 0 & 1 & 2 & 1 & 1 & 0 & 0 & 0 & 0 \\
1 & 2 & 1 & 2 & 2 & 2 & 1 & 2 & 2 & 1 & 0 & 2 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
1 & 2 & 2 & 1 & 2 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 2 & 1 & 1 \\
1 & 2 & 2 & 2 & 1 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 1 & 1 \\
1 & 2 & 2 & 2 & 2 & 1 & 0 & 2 & 2 & 0 & 0 & 0 & 0 & 2 & 1 & 1 & 0 & 2 & 1 & 1 \\
2 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 2 & 2 & 0 & 0 & 0 & 2 & 0 & 0 & 0 & 0 & 0 \\
0 & 2 & 2 & 0 & 0 & 2 & 1 & 1 & 2 & 1 & 0 & 2 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 2 & 2 & 0 & 0 & 2 & 1 & 2 & 1 & 1 & 1 & 2 & 2 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
2 & 0 & 1 & 0 & 0 & 0 & 2 & 1 & 1 & 1 & 2 & 1 & 0 & 2 & 2 & 0 & 0 & 0 & 0 & 0 \\
0 & 2 & 0 & 0 & 0 & 0 & 2 & 0 & 1 & 2 & 1 & 1 & 1 & 2 & 2 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 2 & 0 & 0 & 0 & 0 & 2 & 2 & 1 & 1 & 1 & 2 & 1 & 0 & 2 & 0 & 0 & 0 & 0 \\
2 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 0 & 1 & 2 & 1 & 1 & 2 & 2 & 0 & 0 & 0 & 0 \\
0 & 2 & 0 & 0 & 0 & 2 & 0 & 0 & 0 & 2 & 2 & 1 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\
2 & 1 & 1 & 0 & 0 & 1 & 2 & 1 & 1 & 2 & 2 & 0 & 2 & 0 & 1 & 2 & 0 & 0 & 2 & 2 \\
2 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 2 & 2 & 1 & 2 & 1 & 0 & 0 & 2 & 2 \\
2 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 2 & 2 \\
0 & 0 & 0 & 2 & 2 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 \\
2 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 2 & 2 & 0 & 1 & 2 \\
2 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 2 & 2 & 1 & 2 & 1
\end{bmatrix}. \qquad (4.4.1)$$

We first compare the fit and forecast error metrics using $c = 1$ and $c = 2$ for all links.

4.4.1 Comparison of forecasting performance to identify the range of influence

	AvgRelMAE	RelMedianAE	MedianRelMAE
Sparse VAR	0.63092	0.52860	0.98172
Constrained VAR	1.10106	0.98380	1.00194
Bubble	1.37977	0.65081	1.00050
ARIMA	1.00000	1.00000	1.00000
Link beta	0.82736	0.73937	0.99770
Timedpt	**0.00000**	**0.00000**	**0.39490**
VAR	1.07432	1.03889	0.99954
Hierarchical	0.99968	1.00050	1.00000
Constrained VAR	1.50823	1.22185	1.00966
Bubble	0.07108	0.95684	0.83496
Time dpt	0.00000	0.00000	0.00000
Link beta	0.84973	0.75710	0.99712
Hierarchical	1.00000	1.00000	1.00000

Table 4.4.1: Fitted error metrics for the methods applied to the 20 link data. The first set are the $c = 1$ results, the second the $c = 2$ results.

Error metrics. The fitted error metrics in Table 4.4.1 agree with the 5 link metrics that the time dependent method is best at fitting. The AvgRelMAE and RelMedianAE are zero for both the $c = 1$ and the $c = 2$ cases. When including the second degree neighbours the MedianRelMAE is also zero, implying that including the second degree neighbours provide the best fit overall. The next lowest AvgRelMAE is the bubble for the $c = 2$ case, which is considerably lower than the others. This method is also better than the ARIMA for the two other measures. The next best is the sparse VAR and the link beta for both cases is the only other method that is better than the ARIMA across all three metrics. The constrained VAR is worse at fitting when including the second degree neighbours and the hierarchical is better for the RelMedianAE but worse for the AvgRelMAE.

Table 4.4.2 shows the Forecast error metrics. Unlike with the 5 link network the constrained VAR method is much better than any of the other methods. The constrained VAR method is better without including the 2-degree neighbouring links. Only the hierarchical method is better when including more links and this is only slightly. The difference between when $c = 1$, and when $c = 2$, for the link beta method is small. The bubble method, for the $c = 1$, and link beta method, for

	AvgRelMAE	RelMedianAE	MedianRelMAE
Sparse VAR	0.98573	1.01031	0.99912
Constrained VAR	**0.79381**	**0.78270**	**0.99121**
Bubble	0.99979	0.98554	0.99730
ARIMA	1.00000	1.00000	1.00000
Link beta	0.96081	0.95478	0.99871
Time dpt	18.50947	5.13130	1.33136
VAR	1.03522	1.06106	1.00018
Hierarchical	1.00001	1.00008	1.00000
Constrained VAR	0.82102	0.82032	0.99522
Bubble	1258.56115	6542.59516	7.16781
Time dpt	522.17839	29.88332	4.20349
Link beta	0.96695	0.96493	0.99852
Hierarchical	1.00000	1.00000	1.00000

Table 4.4.2: Forecast error metrics for the methods applied to the 20 link data. The first set are the $c = 1$ results, the second the $c = 2$ results.

both cases, are both better than the ARIMA method across all three metrics. The bubble method for the $c = 2$ case is by far the worst method at forecasting. The time dependent method also performs poorly, as in the 5 link case, especially when including the extra links. This is likely to be because the time dependent method overfits, and hence, when including even more parameters when $c = 2$, it overfits even more. The potential issue of overfitting applies to all methods when $c = 2$, hence the fact that most of the methods are better with less parameters implies that the correct range of influence is $c = 1$.

When comparing the fit and forecast error metric values for each method and range of influence the majority of the forecasts are worst. In general we would expect the forecast values to be higher as the fitted values are calculated using the data the model parameters are calculated using. However these measures are different because we calculate them relative to the fitted ARIMA values and the forecast ARIMA MAE and MedianAE. Hence the two aren't directly comparable.

Computational time. When we scale the methods up to the 20 link network from the 5 link network which methods are quicker changes. The times (in seconds) are

	$c = 1$	$c = 2$
SparseVAR	15478.68	NA
ConstrainedVAR	102761.12	107962.14
Bubble	6225.13	4372.16
ARIMA	5010.43	NA
Link beta	5396.47	6113.94
Timedpt	7809.89	5483.86
VAR	10020.28	NA
Heir	6542.66	**4014.04**

Table 4.4.3: Computational time for $c = 1, 2$ for 20 link network. Methods which don't consider the range of influence are recorded in $c = 1$.

presented in Table 4.4.3. Before the sparse VAR and constrained VAR where quickest, now these, as well as the VAR, take much longer than the linear combination methods. The constrained VAR takes the longest, to the extent that it is impractical to have to run the model every day. We would expect the 2 away methods to take longer as they have more parameters to calculate but this isn't the case in the time dependent, the hierarchical or the bubble cases. The quickest method is the Hierarchical 2 away method which is unexpected and may be due to the cluster being less busy when it was run.

We can however see that the methods that use linear combinations and the single link model take approximately the same time when considered on a larger scale. The VAR related methods are orders of magnitude larger than the linear combination methods, with the constrained VAR taking at least 10 times as long. This difference will only increase as the size of the network increase, as we discussed in Section 4.3.2.

As a result of the lack of scalability we would have to consider the constrained VAR to be unusable for the petrol routing problem, despite its better forecasting performance. The second best method in terms of forecasting performance is the link beta method which is much more scaleable in terms of computational time when the network size is increased.

4.4.2 Beta comparison

We now look at how the β values change between the 2 away and 1 away methods. If the extra parameters in the two away matrix are zero or close to zero, with the rest of the parameters being the same, then we can conclude that the range of influence is one link only. However if there are large changes and the two away method is much better at forecasting then the range of influence may be even larger than 2 links away.

In Section 4.3.2 we compared the β values for the 5 link network. We now focus on the β values for link 10.11 as the W matrix is now 20 by 20. In the 1 away there are 5 connecting links, whereas when also considering the 2 away links there are 13 links.

Link beta comparison. Table 4.4.4 shows the mean beta parameters for both the 1 link away and 2 links away values for the link beta method. The values for the six links in both cases are similar. With the exception of the second to last mean β value they are all under 0.1 and hence small. The second to last one is 0.19. All of the mean standard deviations are within two standard deviations of zero.

There is greater variability in the six links that are one away, with the exception of the second to last one and the fifth to last one, which both have a standard deviation of nearly 0.2. This supports the idea that most of the 2 away links aren't necessary and when combined with the better forecasting performance result we conclude that for the link beta method $c = 1$.

Hierarchical Range. Due to the method of calculation of the β values there are 20 possible values for the Hierarchical method as we have had to convert the original parameter matrix to form W to allow comparison between the methods. In addition the two cases which we call $c = 1$ and $c = 2$ aren't the same as for the selection of R. Here we consider $c = 1$ to be the case which splits the links by whether they are north/south links or east/west links and $c = 2$ further grouped the roads by motorway

2 away mean	1 away mean	Sd 2 away
0.68	0.72	0.21
0.17	0.17	0.15
0.07	0.11	0.17
-0.08	-0.02	0.15
-0.03	-0.00	0.15
-0.04	0.03	0.21
0.00		0.06
-0.00		0.08
0.08		0.05
0.03		0.20
0.04		0.08
0.01		0.04
0.19		0.18
0.04		0.12

Table 4.4.4: Mean and standard deviation of beta parameters for link 10.11 for link beta method.

or A-roads.

The β values for link 10.11 can be seen in Table 4.4.5. The parameters in bold are those not in $R^2_{10.11}$. Several of the mean parameters are zero but these change from $c = 1$ to $c = 2$. The standard deviations for all the parameters are small and interestingly for $c = 2$ the link itself is classed as zero so of no importance. Considering that the Hierarchical method forecasting performance was better for $c = 2$ than $c = 1$ this is especially unexpected. The Hierarchical method by design picks up slightly different things in the forecast to the other methods that use R but this implies that there is no information from the link itself when forecasting it. Several of the parameters are significantly different supporting the idea the two cases are different.

Time dependent comparison. We present the mean and standard deviations for the β parameters for the $c = 1$ and $c = 2$ in Table 4.4.6. The standard deviations are very high, with the exception of the last three parameters. Unlike in the link beta case whereby the mean parameters that are only in the $c = 2$ case where clearly smaller and had less variance in this case they are still large and also have high variance.

2 away mean	1 away mean	Sd 2 away
-0.00	0.89	0.00
-0.00	-0.05	0.00
-0.00	-0.09	0.00
0.26	-0.00	0.01
-0.00	-0.06	0.00
-0.00	-0.11	0.00
-0.20	-0.00	0.04
-0.19	-0.00	0.05
-0.24	-0.00	0.07
-0.21	-0.00	0.09
-0.00	-0.04	0.00
-0.00	-0.06	0.00
-0.00	-0.13	0.00
-0.00	-0.06	0.00
-0.00	-0.05	0.00
-0.00	-0.11	0.00
0.19	-0.00	0.03
0.17	-0.00	0.04
-0.00	-0.12	0.00
0.24	-0.00	0.02

Table 4.4.5: Mean and standard deviation of beta parameters for link 10.11 for Hierarchical method.

The high variance means that it is impossible to conclude that any of the parameters associated with the second degree neighbours should be zero, because while zero lies within the confidence bounds then so do a lot of other values. However when we consider that the forecasting performance is much better when $c = 1$ we conclude that the region of influence is $c = 1$.

Bubble. As with the Hierarchical method the two cases are slightly different as we have to change the value of s to allow extra links to be considered. When $c = 1$ we use $s = 300$ and for $c = 2$ we use $s = 600$. The bubble values are quite different from the cases $c = 1$ and $c = 2$. This is due to the fact that the link 10.11 is only about 300 seconds long. When we increase the bubble size to 600 the bubble is on the connecting links for the entire period. The amount of the bubble on link 10.11 is limited by the travel time along the link rather than s. The 4th link, for which the

2 away mean	1 away mean	Sd 2 away
-0.91	-0.28	137.85
2.25	0.90	84.64
0.24	0.31	52.10
0.62	0.02	63.42
-0.41	0.09	45.22
-0.79	-0.35	42.89
0.82		41.97
-0.17		21.92
-0.83		56.34
-0.15		14.42
0.25		13.19
-0.00		0.72
0.00		0.00
0.00		0.00

Table 4.4.6: Mean and standard deviation of beta parameters for link 10.11 for time dependent method.

average travel time is above 1000 is limited by the size of the bubble s, and the α values. Thus the mean β value associated with this link is larger than the value for link 10.11. This is likely to explain why it is so poor at forecasting.

Most of the second degree neighbours are zero as the bubble is split between the first degree neighbours and if they are long then the bubble ends on them. The two that are non-zero are very small, however due to the very small variance zero isn't in their confidence interval. The small variance leads to us concluding the parameters for the methods are significantly different. When combining this with the forecasting performance we would conclude that $c = 1$ is the best neighbourhood, where $s = 300$. We could also consider changing the value of s and the α values but this is unlikely to give large improvements, and in the case of increasing s is likely to cause the predictions to get much worse.

Range of influence conclusions. The majority of the methods perform best with only 1 link away rather than 2. The mean β parameters for the links that are 2 away are very close to zero and most have little variation. When combined with the

2 away mean	1 away mean	Sd 2 away
0.19	0.46	0.01
0.09	0.11	0.00
0.14	0.11	0.00
0.28	0.11	0.00
0.11	0.11	0.01
0.16	0.11	0.00
0.00	0.00	0.00
0.00	0.00	0.00
0.00	0.00	0.00
0.00	0.00	0.00
0.00	0.00	0.00
0.00	0.00	0.00
0.01	0.00	0.00
0.01	0.00	0.00

Table 4.4.7: Mean and standard deviation of beta parameters for link 10.11 for the bubble method.

forecasting performance results also agreeing that $c = 1$ is better we conclude that using one link away is the best choice for R_a and hence $c = 1$.

4.5 Initial conclusions

The quickest in terms of absolute time for the 20 link method is the hierarchical method but the other linear combination methods also have very similar run times. They also scale appropriately for when the network is largest.

Without consideration of the practicality of implementing the method we would select the constrained VAR with $c = 1$ due to its forecasting accuracy being the best by a considerable way. However the time it takes to run is a major issue, even when running for a single day at a time. On the small 5 link network it is the second quickest but on the 20 link network it is by far the longest and the plot in Figure 4.3.5 suggests that the time will only increase compared to the other methods for larger networks. The test network suggested for a VRP in Section 2.3.3 has 54 links and this is a small network. As a result the computational complexity required for the forecasts will be too great.

We therefore need to compromise on the accuracy in order to get a usable result. The link beta method, for either range of influence, has better forecasting accuracy on all three error metrics than the ARIMA method for both the 5 link and 20 link network. In addition it only takes slightly longer to run for both sizes of network, as it is a linear combination method. The method is therefore scalable to larger networks and for a slightly better accuracy it is worth the minimal additional time to generate the forecasts.

The range of influence that should be considered for each link is small, only up to and including the first degree neighbours. This holds for all methods that we can consider this for. The inclusion of additional parameters, both from additional links and from more time steps, results in over-fitting, leading to poor forecasts. We therefore conclude that the most practical method to use is the link beta method with a range of influence of $c = 1$.

Chapter 5

Conclusions and Further Work

5.1 Summary of the thesis

This thesis introduced a version of the vehicle routing problem, which we called the petrol routing problem, which finds the optimal route for a vehicle delivering items to customers. Roads that the vehicles can travel on were represented by links in a network. Routes were created by combining the links together. Each link has attributes, such as the travel time along them, that are used to select the optimal route. These attribute values needed be accurate to ensure the optimal route was selected. The thesis focused on how to predict travel times that were time of day dependent, calculated for every link in the network and for all 96 15-minute time periods in the next 24 hours such that they could be input into the petrol routing problem to find the optimal route. The petrol routing problem used the travel time associated with the time period that the vehicle arrives at the start of the link as the travel time for the link.

Models for travel time along a single link in a network were developed. Previous models including exponential smoothing and variations of the ARIMA model were tested on a dataset to test their effectiveness at providing forecasts. Travel times of a set of links in England form the dataset upon which the models were tested. The forecasts generated from these links could be used as inputs into a real life petrol routing problem as the network formed from these links is of the major roads for one area. Investigation of the dataset found weekly patterns in the travel time. However most models assume stationarity throughout time because it is simpler to model, so we removed the non-stationary pattern from the dataset by using a time of day and day of the week standardization before adding the pattern back into the final predictions. The predictions for the travel times where judged based upon suitability for the 24 hours of predictions that are required in the petrol routing problem. The VRP requires accurate input predictions and hence we selected the model with the best prediction accuracy. We considered the prediction accuracy by calculating the average relative mean absolute error over multiple links and multiple days to ensure we only had one type of model to fit for the whole network. The preferred model was the ARIMA model for our dataset, with different AR and MA orders being optimal for different links. However the preferred model could be different if using alternative roads or a slightly different VRP problem. If applying the models to different travel time datasets, checks must be made to ensure stationarity can be achieved after standardization.

We identified that there were infrequent but large delays which can last for multiple consecutive time periods. We set a threshold above which the travel time on a link was a classified as an outlier and hence part of a large delay. These delays were found to occur randomly after standardization and as the single link models focus upon the general behaviour rather than the extreme they fail to capture this behaviour if delays occurs within the predicted 24 hours. We therefore investigated how to model the delays separately by considering the duration of a delay, the maximum additional travel time due to the delay and the probability of a delay occurring. A Poisson process was used to model the occurrence of the delays. The maximum travel time due to the delay was found to be well modelled by a generalized Pareto distribution and the duration by a discrete version of the generalized Pareto distribution. By combining the three models together we found the expected behaviour of delays along a link. This model is designed to be applied to all links of our network as similar behaviour of delays is expected. For alternative networks, resulting from a different dataset, the model could be applied if the delays behave in a similar way. However the threshold could change and different distributions may be more appropriate for the three parts of the model.

We investigated using information from neighbouring links to generate travel time predictions upon each link. These methods were split by using the single link model to generate initial predictions, which were improved upon by using linear combinations of the initial predictions and using all of the links at once to generate global parameters. The methods were evaluated by considering the prediction accuracy and computational effort, as well as which links should be considered in the neighbourhood when predicting a single link. Computational effort was important as the size of the network that is required for the petrol routing problem is likely to be large. The model which had the best prediction accuracy when discounting those with too high a computational effort was the link beta method with a range of influence of size one; i.e., the neighbourhood included the link of interest and all other links directly connected to the link itself. The link beta method is a linear combination method which adjusts the initial predictions for the single link model, which is an ARIMA model, by fitting a β parameter by the sum of least squares for each link in the range of influence. By comparing all the accuracy results to the ARIMA model as the baseline we saw that the link beta model was an improvement. As with the single link findings the optimal model can change dependent upon the network used. The models tested here were developed from the ARIMA model being the optimal single link model for our dataset and alternative models would need to be developed if the optimal model for a different dataset was something other than an ARIMA model. However the linear combination methods are versatile as they can use single link model travel time predictions from any model.

We now consider some avenues for further work that lead on from the work in this thesis.

5.2 Combining the delay model with the travel time model

The ultimate purpose of the petrol routing problem, or indeed any vehicle routing problem (VRP), is to select the optimal route for the vehicles to use when delivering their goods to the customers. As such both the travel time and potential delays are inputs into the VRP and we must consider the overall affect on the VRP when looking at how to combine the two to give an overall model of the travel time behaviour over the links.

From the point of view of the VRP we consider delays in addition to the travel time because customers wish to have their goods delivered on time. In addition, when transporting petrol, getting stuck in a traffic jam also increases the risk, as if there should be an accident there are many surrounding cars and the petrol is in the vehicle for longer. There are also driving time limits for drivers, meaning they must stop or return early. Hence a company would consider selecting a route that is slightly slower overall but has a much lower chance of a long delay, resulting in the route taking much longer overall.

The ways in which we can consider including both the travel time and the delay models are limited by the design of the VRP and how it accepts inputs. The petrol routing problem, as we defined it in Section 1.1.3, uses a single travel time value for each link as an input. This value varies through time to account for the daily travel changes as observed in Section 1.1.7. The travel time is therefore deterministic and the VRP assumes that the travel time that is input is the travel time that will be observed, hence the need for the travel times to be as close to what will be observed as possible.

In this formulation the travel time is used to ensure that routes aren't too long and that the vehicles arrive when customers are open. The VRP also outputs a schedule that enables the delivery company to predict when they will arrive at a customer and when to initially leave the depot. The travel time doesn't directly contribute to the objective function, as defined in Section 1.1.5, which decides which route is best. However in another version of the problem the shortest travel time could be either the objective or an additional objective.

We begin by summarising the main features of our models for the travel time and the delays, before discussing these in more detail. The models for the base travel time and the delays are all generated using the same time series which is recorded every 15 minutes and has been standardized from the original travel time values. The standardization removes the daily and weekly patterns and the initial outputs from both models are on the same scale, so the trend must be reintroduced to produce travel time predictions for a given time interval. The travel time predictions for the base travel time vary from interval to interval while the delay model is the same over time.

The travel time model which we developed in Chapters 2 and 4 produces 96 travel time predictions for each link. The 96 predictions cover a 15 minute interval and hence provide the travel time estimates for the entire day that we wish to run the VRP for. These models are unique to each link, but as with Chapter 2, we drop the link reference from the notation for ease of reading.

By contrast the delay model from Chapter 3 is more complicated. Every 15 minute interval is classified as either being in a state of delay for its entire length or having no delay at all. This is because the data are recorded only at 15 minute intervals and from this single value we cannot infer anything about a smaller time scale. This is on the same scale as the network level model for the travel time which enables the combination of the two. A delay can extend over multiple intervals. The model consists of three parts, one for the occurrence of delays, and two models for the behaviour of a delay conditional on a delay occurring. These two models are for the maximum size of the delay and its duration.

The exact models are summarised in Section 3.5. The rate of occurrence of delays is a Poisson process and the times between delays are exponentially distributed with a constant rate through time. The size of the maximum delay is modelled by a generalized Pareto distribution with parameters that are unique to each link. The duration of delays are measured in intervals so a discrete random variable model is required. The geometric distribution was not sufficiently flexible so instead we model duration by a discrete version of the generalized Pareto distribution, with a common shape parameter for all links and a scale parameter that is unique to each link.

How we use these three models depends upon how we incorporate the delay with the base travel time. One common feature of each of the approaches proposed below is that we model using the maximum delay sizes. Delays that occur over multiple intervals have different sizes of delay for each interval. However modelling this over all the links in the network would be too complicated. For simplicity we ignore this profile of the size of the delay and use the conservative approach of approximating all values by the maximum cluster delay size. This ensures we cover the worst case scenario.

If the variance between the potential travel times is very small then the observed travel time of any route will be very similar to those given from the VRP using the travel time predictions for each link. We know from Section 1.1.7 that delays can occur on a link and these cause the observed travel time to be much higher than the prediction. There is therefore high variation in the travel times that are observed, potentially leading to the VRP selecting a route that will be infeasible due to a customer being shut when they actually arrive.

As the proposed VRP of Section 1.1.3 is not online then most delays currently happening will dissipate before the vehicles are scheduled to leave the depot. An alternative branch of further work would be to incorporate the delay models into an online VRP to model the future of a delay that is currently happening.

We now consider five ways of incorporating both the delays and the travel time models in the VRP. These are: increasing the travel time by expected amount of delay in Section 5.2.1, delay insertion in Section 5.2.2, model switching in Section 5.2.3, including an additional objective for delay in Section 5.2.4 and making the travel time stochastic in Section 5.2.5. They have been ordered by how difficult it would be to alter the VRP to incorporate them, from the simplest to the most complex.

All five approaches in this section use standardized travel times, as the delays travel times are calculated using the same standardized scale. Once the standardized travel times and delays have been combined into new predictions these then need to be transformed into the travel times. We rearrange equation (2.3.2) from Section 2.3.5, with the corresponding median and IQR for the time t to calculate the travel time of the link for all the suggested models.

5.2.1 Increase travel time by expected amount of delay

The simplest way is to keep the same VRP, using the travel time as an input. However we wish to change the travel time values that are input to disadvantage those roads which have the potential for longer delays. Therefore the travel time input will be larger than is estimated by the network model.

An obvious first choice for this would be to adjust the base travel times to be the expected travel times which includes the potential for delays. The base travel times are a single value travel times which can be taken as the expected travel time for the model. Thus including the delays via the expectation is a logical extension. The VRP uses a single travel time for each interval as an input and hence we are interested in the probability of a delay happening in that single interval, rather than the duration of the delay or the rate of occurrence.

In Section 3.5 we confirmed that the estimation of p_d, the probability of the interval being in a state of delay, by equation 3.5.7 is acceptable. If the model is not in a state of delay then we observe the base travel time and hence we have the probability of observing the base travel time.

We can then use find the expected size of the maximum size of delay model which is a Generalised Pareto distribution. From equation (3.3.25) in Section 3.3.2 the expected maximum size of delay, $E(D)$, is:

$$E(D) = v + \frac{\sigma_v}{1-\xi} \quad \text{if } \xi < 1. \tag{5.2.1}$$

Let $\tilde{y}(t)$ be the standardized base travel time from the network model in Section 4.4.1, for interval t and p_d be the probability of a delay in an interval. We then find the maximum expected standardized travel time, which we denote $\hat{y}(t)$, for the same interval t, by:

$$\hat{y}(t) = \tilde{y}(t)(1 - p_d) + E(D)p_d. \tag{5.2.2}$$

This is a mixture model, which combines two models by giving their values different weights. A more general formulation of (5.2.2) may be required. One possible such formulation takes a function $f(D)$ that takes the possible delays and converts them to single value. In equation (5.2.2) this function is the expectation. An alternative function could be the standard deviation. This would advantage links with low variance and low travel time. Similarly we weight the values by a factor a, $0 \leq a \leq 1$, which before was p_d. Then the travel time prediction of a mixture model is:

$$\hat{y}(t) = \tilde{y}(t)a + f(D)(1 - a). \tag{5.2.3}$$

Using any other parameters than those used in equation (5.2.2) causes issues with estimating the arrival times at customers as $\hat{y}(t)$ is used within the VRP to calculate the arrival times. A weighting that results in a much smaller or much larger $\hat{y}(t)$ will lead to incorrect schedules because $\hat{y}(t)$ is the travel time along a link. As discussed in Section 5.2 we use the maximum delay size rather than considering the shape of the delay over time, thus using the parameters in equation (5.2.2) will be an increase on the true expectation. However this increase is likely to be small.

The use of alternative mixture models is more appropriate when one of the objectives is the travel time. Then the weightings can take into account how much the delivery company wishes to avoid potential delays.

5.2.2 Delay insertion

Ideally we would be able to predict exactly when the delays occur and then we can model the travel time during the delay using the duration and size of delay. However delays occur randomly, uniformly distributed over time after we have standardized and hence we cannot predict when they occur. One idea would be to randomly insert delays into the travel time profile of the 96 predictions for the day by sampling from the model of Chapter 3. This generates one travel time profile, out of many possible profiles, that could be observed. If we use this single travel time profile in the VRP it generates a distorted optimal route and hence we discuss using replicated simulations of profiles once we have introduced how to create a single travel time profile.

We have two models, one for the base travel time, $\tilde{y}(t)$ and the other for the delay. If a delay is predicted we sample from the maximum size distribution to find the maximum size and use this value instead of the network travel time. We use this maximum size for the length of the delay as sampled from the duration distribution. The travel time profile will consist of single values and hence we can use it in our proposed VRP from Section 1.1.3 and in any VRP that uses travel times that aren't stochastic.

Let i_j be the start time of delay j which is sampled from the exponential distribution. There are N total delays which depends upon the samples drawn for the time between delays. We then sample the maximum delay, m_j, from the generalised Pareto distribution and sample the duration of the delay, l_j, from the discrete generalised

Pareto distribution. Then, with $\hat{y}(t)$ and $\tilde{y}(t)$ as in Section 5.2.1, the standardized travel time at interval t is;

$$\hat{y}(t) = \begin{cases} m_j, & \text{if } i_j \leq t < i_j + l_j, \qquad \forall j = \{1, \ldots, N\} \\ \tilde{y}(t), & \text{otherwise.} \end{cases} \tag{5.2.4}$$

As in Section 5.2 we then transform the standardized values to create the travel time predictions.

When the VRP model for the petrol routing problem, as introduced in Section 1.1.3, calculates the optimal route it can only use one travel time profile for each link. This single profile is one of many that could be observed and different optimal routes are likely to be returned by the VRP if different profiles are used. We could instead use simulation to select the optimal route. By generating a large number of travel time profiles and then repeatedly solving the VRP to find the best route for each profile the possible routes can be ranked according to how well they perform over all the profiles. We would then select the best ranked route as the optimal route. As long as a large enough number of travel time profiles are generated the optimal route that is selected will be stable with repeat to the simulation process. However solving the VRP multiple times will have a high computational cost.

5.2.3 Model switching

A variation on the delay insertion in Section 5.2.2 is to use the base travel time to decide when to use the delay model rather than randomly insert the delays. If the travel time is above the threshold for a delay we use the delay model to predict the delay and then switch back to the travel time model when the delay has finished. We use equation (5.2.4) but the methods differ in how they select i_j and N. In the delay insertion method we sampled these values whereas when model switching they are determined by the travel time model. We define i_j as the start of the jth delay using equation (3.2.1) from Section 3.2.2. We have already identified, in Section 3.2.3, that for our dataset the threshold is $u = 2$ and that the maximum interval gap when the travel time can be less that the threshold within the same cluster is 3. Then we define i_j such that;

$$y_{i_j} > u \qquad \text{and} \qquad \max(y_{i_j-m}, \ldots, y_{i_j-1}) < u. \tag{5.2.5}$$

As with the model in Section 5.2.1, we can make the model more general by using $f(D)$ instead of the maximum size m_j. To ensure that $\hat{y}(t)$ can be input into the VRP $f(D)$ must produce a single value. One option for the function f would be to chose the expectation of D given that a delay is happening and this is constant across time. All values of $f(D)$ should be above the threshold 2 as a delay is occurring. We therefore adapt equation (5.2.4) to the more general form of

$$\hat{y}(t) = \begin{cases} f(D), & \text{if } i_j \leq t < i_j + l_j, \qquad \forall j = \{1, \ldots, N\} \\ \tilde{y}(t), & \text{otherwise.} \end{cases} \tag{5.2.6}$$

One of the main advantages of this method is that it is easy to understand and explain. However the inclusion of the delays isn't random as they are determined by the values of $\tilde{y}(t)$. The method also has the same problem with optimal route selection as the delay insertion method of Section 5.2.2 which we discussed overcoming with simulation. Hence it is preferable to use the delay insertion method, as we identified in Chapter 3 that the delays are randomly occurring after standardization and hence should be modelled as such.

5.2.4 Additional objective for delay

We briefly discussed multi-objective VRP problems in Section 1.2. Many of the hazardous material transportation problems are multi-objective as the main objective is reducing risk and the delivery company is also interested in other things such as reducing cost, for which the travel time is often used as a proxy.

Having multiple objectives makes it much more difficult to select the optimal route as the objectives are often conflicting and thus routes that are good for one objective will be poor for another. It is therefore necessary to report multiple solutions with their corresponding values and let the user decide on the optimal solution from these potential solutions. The multiple solutions are such that any other route is unable to improve upon one objective without making the other one worse. Solving a multi-objective VRP requires different methods to a single objective VRP as it needs to output multiple solutions. Wang and Lin (2013) use simulation to reach a solution rather than solving the VRP directly.

It is therefore possible to include an additional objective for the delay. This will give the user the ability to check how likely a long delay is on the proposed optimal route and if choosing the route with slightly more risk is much less likely to have an delay.

We then have to choose how to characterise this delay objective function. As the delay model is a stochastic model various summary statistics need to be extracted from the model to form an objective. One example is the expected delay, which we can calculate for each link individually and then add together to get the expected delay for the whole route.

We could also use the mixture model of equation (5.2.3) as the objective function. If we just consider the delay then this corresponds to the case when $a = 0$. The weightings can then be changed such that the objective favours lower travel times more, such as is suggested in Section 5.2.1, or gives somewhere in between the two cases.

5.2.5 Make travel time stochastic

As mentioned in the literature review of VRP models in Section 1.2, if we make the VRP too complicated then it becomes too computationally intensive to solve. Using a stochastic travel time is one way that increases the complexity greatly. This is because the introduction of uncertainty makes it a lot more difficult to select the optimal route. If the uncertainty contributes directly to the objective then the objective becomes a distribution for each route. One distribution is unlikely to be better than all other route distributions for all possible travel time values. In the case of the petrol routing problem, travel time doesn't directly affect the object but the constraints that involve travel time are now all distributions. These are also difficult to incorporate.

The delay model is already stochastic as it is comprised of the rate of occurrence, as well as distributions of the maximum size and the duration as summarised in Section 3.5. As noted in Section 5.2.1, when we focus on a single time period we can instead simplify the rate of occurance and duration into the single probability of the interval being in a state of delay. As delays have been found to occur uniformly over time, after standardization, the distribution of the maximum delays will be the same for all the time points. To obtain the probability distribution of the maximum delay size for any of the 96 time periods we multiply the probability of the interval being in a delay state, p_d, from Section 5.2.1, by the Generalised Pareto distribution for the maximum size of delay as identified in Section 3.3.2.

Equation (3.3.24), from Section 3.3.2, is the density function for a GPD with parameters ξ and σ. The density function, f, of the maximum size of delays, m is therefore:

$$f(m) = \begin{cases} \frac{1}{\sigma}\left(1 + \frac{\xi(m-v)}{\sigma}\right)_+^{(-\frac{1}{\xi}-1)}, & \text{if } \xi \neq 0, m > v \\ \frac{1}{\sigma}e^{-\frac{m}{\sigma}}, & \text{if } \xi = 0, m > v. \end{cases} \tag{5.2.7}$$

We now combine equation (5.2.7) with the base travel time model to generate the overall travel time density function including the delays. Let g be the probability density function for the overall standardized travel time. Then the overall standardized travel time density function at time t is:

$$g_t(y) = \begin{cases} \mathbb{1}_{y=\tilde{y}(t)}(1-p_d) + \mathbb{1}_{y>v} p_d \frac{1}{\sigma}\left(1+\frac{\xi(y-v)}{\sigma}\right)_+^{(-\frac{1}{\xi}-1)} & \text{if } \xi \neq 0, \\ \mathbb{1}_{y=\tilde{y}(t)}(1-p_d) + \mathbb{1}_{y>v} p_d \frac{1}{\sigma} e^{-\frac{y}{\sigma}}, & \text{if } \xi = 0. \end{cases} \tag{5.2.8}$$

To calculate the density function of a route requires the joint density function of all the links that make up that route. However, the distribution of each link depends upon the arrival time at the start of the link which depends upon the travel time of the previous link which is also a distribution. Even combining two independent generalised Pareto distributions is very difficult. One possible approach is to simplify the continuous density functions into discrete distributions with only C possible values for the delay size. Zhang et al. (2013) distribute these C travel times evenly over the possible travel times such that all the differences between the cumulative distribution of successive travel times are the same.

In addition, even the base travel time will have some degree of variability. If there is a lot of variability we would need to model the base travel time as stochastic, as discussed in Section 2.2.3, and then combine this with the delay size distribution. The resulting distribution will be of a non standard form, requiring simplification in order to combine the distributions across links. One simplification considered by the models in Section 2.2.3 is to make the stochastic distribution invariant of the time of day but this loses the weekly pattern. A compromise could be to increase the interval size or to group similar intervals together.

A further issue with converting the travel time to being stochastic is considering time windows within which the customers must be served. We discussed hard and soft time windows in Section 1.2. A vehicle must arrive within the time window if it is hard and when delays result in a high variability between travel times then hard time windows are impossible to include as the range of possible arrival times is very large as getting delayed on all links could happen, although it is very unlikely. With soft time windows vehicles can arrive outside the time window but there is a cost for doing so. Jula and Dessouky (2006) use a soft time window with stochastic travel time but have a very small network.

The travel time is also used to generate schedules for the drivers to see where they should be and at what time. Schedules created using stochastic travel times will be high in variability making them difficult to interpret. The driver doesn't need to know that there is a very small possibility that he will arrive 2 hours late. Schedules that are easy to interpret could be generated using the expected travel time as discussed in Section 5.2.1. This removes the stochastic nature of the travel time from the schedules that were presented to the drivers but retains it in selecting the optimal route.

5.3 Additional further work

As discussed in Section 1.1.6 the travel time model is an input to the VRP. Hence, in order to fully evaluate the travel time model, we need to analyse how it affects the optimal solution. This applies both to the network model from Chapter 4 and to the combined models for the delay and the travel time in Section 5.2.

A VRP model assumes that all of the inputs to it are correct and hence selects the optimal solution from this. Thus using travel time predictions means that the best route the next day may not be the one the VRP selects. We differentiate between these two solutions as the predicted optimal solution for the route selected when using the travel time model and the true optimal solution as the optimal solution from the observed travel times. The optimal solution consists of the route each vehicle must follow and the schedule which is when a vehicle arrives at each customer.

Ideally the true optimal solution and the predicted optimal solution would be the same. However due to the variations in travel times this is very unlikely. However if the routes are the same, with a similar departure time then using the predicted optimal route will result in the true optimal route and the travel time model is therefore appropriate. We can therefore compare the two routes over many days, by first using the predicted values and then using the observed ones. However finding a solution for the VRP is very computationally intensive (El-Sherbeny, 2010).

Another way we can test the travel time model in the VRP is to use sensitivity analysis to see how the predicted optimal solution changes when the input values vary. We can explore how changing the travel time slightly affects the predicted optimal solution. Janssens et al. (2009) suggest evaluating the sensitivity to change by using time Petri nets instead of simulating using the whole VRP. Fleischmann et al. (2004) compare the effect of constant and time varying travel times while also varying the density of time windows by using different heuristics. These methods could be adapted to compare the travel time predictions.

A further consideration of the travel time predictions that are whether the resulting travel time profiles, as introduced in Section 2.2.1, are appropriate for use within the VRP. Due to the format of our dataset the travel time profile of all our models are step functions. Step functions frequently violate the properties and assumptions of travel time upon roads that Carey et al. (2003) and Horn (2000) introduced. We would need to check our step function against this and if necessary consider adapting it to a slanted step function such as Malandraki and Daskin (1992) use.

We can then evaluate the VRP models which include the delays as outlined in Section 5.2. As we are focusing on the petrol routing problem of Section 1.1.3, we can ignore the model switching approach of Section 5.2.3 because the delays are at the wrong times. We therefore have three discrete models - the expected travel time with the delays included, an additional delay objective and the delay insertion method. We can compare the predicted optimal routes for all three and find out how the main objective value changes from model to model.

Evaluating a stochastic travel time model, as discussed in Section 5.2.5, within a

VRP is much more difficult. We can't easily compare it to the deterministic models because they are formulated in a completely different way. However, stochastic network design has considered using deterministic solutions to create stochastic solutions and evaluates the deterministic solution based upon the stochastic setting (Wang et al., 2019). These ideas may be able to be adapted to our VRP problem.

One of the difficulties with the stochastic travel time is the difficulty in verifying that the distribution is correct. As with the deterministic models we can conduct sensitivity analysis to see how the results change when the distribution is altered slightly. Noorizadegan and Chen (2018) conduct sensitivity analysis to check whether the selected route is in fact feasible.

Bibliography

Abkowitz, M. and Cheng, P. D.-M. (1988). Developing a risk/cost framework for routing truck movements of hazardous materials. *Accident Analysis & Prevention*, 20(1):39–51.

Akgün, V., Erkut, E., and Batta, R. (2000). On finding dissimilar paths. *European Journal of Operational Research*, 121(2):232–246.

Akgün, V., Parekh, A., Batta, R., and Rump, C. M. (2007). Routing of a hazmat truck in the presence of weather systems. *Computers and Operations Research*, 34(5):1351–1373.

Ando, N. and Taniguchi, E. (2006). Travel time reliability in vehicle routing and scheduling with time windows. *Networks and Spatial Economics*, 6(3-4):293–311.

Androutsopoulos, K. N. and Zografos, K. G. (2010). Solving the bicriterion routing and scheduling problem for hazardous materials distribution. *Transportation Research Part C: Emerging Technologies*, 18(5):713–726.

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

Aron, M., Bhouri, N., and Guessous, Y. (2014). Estimating travel time distribution for reliability analysis. In *Transport Research Arena 2014, Paris*.

Athanasopoulos, G., Hyndman, R. J., Kourentzes, N., and Petropoulos, F. (2017). Forecasting with temporal hierarchies. *European Journal of Operational Research*, 262(1):60 – 74.

Bertsimas, D. J. and van Ryzin, G. (1993). Stochastic and dynamic vehicle routing in the euclidean plane with multiple capacitated vehicles. *Operations Research*, 41(1):60–76.

Bowler, L. A., Aff, M. P., and Mahmassani, H. S. (1998). Routing of radioactive shipments in networks with time-varying costs and curfews. Technical Report ANRCP-1998-11, Amarillo National Resource Center for Plutonium.

Cao, P., Miwa, T., and Morikawa, T. (2014). Modeling distribution of travel time in signalized road section using truncated distribution. *Procedia - Social and Behavioral Sciences*, 138:137–147.

Carey, M., Ge, Y. E., and McCartney, M. (2003). A whole-link travel-time model with desirable properties. *Transportation Science*, 37(1):83–96.

Carotenuto, P., Giordani, S., and Ricciardelli, S. (2007a). Finding minimum and equitable risk routes for hazmat shipments. *Computers & Operations Research*, 34(5):1304–1327.

Carotenuto, P., Giordani, S., Ricciardelli, S., and Rismondo, S. (2007b). A tabu search approach for scheduling hazmat shipments. *Computers & Operations Research*, 34(5):1328–1350.

Chang, T.-S., Nozick, L. K., and Turnquist, M. A. (2005). Multiobjective path finding in stochastic dynamic networks, with application to routing hazardous materials shipments. *Transportation Science*, 39(3):383–399.

Christiano, L. J. (2012). Christopher A. Sims and Vector Autoregressions. *Scandinavian Journal of Economics*, 114(4):1082–1104.

Cox, D. and Isham, V. (1980). *Point Processes*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis.

Current, J. and Ratick, S. (1995). A model to assess risk, equity and efficiency in facility location and transportation of hazardous materials. *Location Science*, 3(3):187–201.

Davison, A. C. and Smith, R. L. (1990). Models for exceedances over high thresholds. *Journal of the Royal Statistical Society. Series B (Methodological)*, 52(3):393–442.

Davydenko, A. and Fildes, R. (2013). Measuring forecasting accuracy: The case of judgmental adjustments to sku-level demand forecasts. *International Journal of Forecasting*, 29(3):510–522.

Department for Transport (2018). Domestic road freight statistics, united kingdom 2017. Technical report, National Statistics.

Department for Transport (2019). Continuing survey of road goods transport (great britain). Technical report, National Statistics.

Desai, S. and Lim, G. J. (2013). Solution time reduction techniques of a stochastic dynamic programming approach for hazardous material route selection problem. *Computers & Industrial Engineering*, 65(4):634 – 645.

Dong, W., Vu, H. L., Nazarathy, Y., Vo, B. Q., Li, M., and Hoogendoorn, S. P. (2013). Shortest paths in stochastic time-dependent networks with link travel time correlation. *Transportation Research Record: Journal of the Transportation Research Board*, 2338:58–66.

Eastoe, E. F. and Tawn, J. A. (2009). Modelling non-stationary extremes with application to surface level ozone. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 58(1):25–45.

El-Basyony, K. and Sayed, T. (2014). A framework for an on-demand dangerous goods routing support system for the metro Vancouver area. *Journal of Engg. Research*, 2(3):19–39.

El-Sherbeny, N. A. (2010). Vehicle routing with time windows: An overview of exact, heuristic and metaheuristic methods. *Journal of King Saud University - Science*, 22(3):123 – 131.

Erkut, E. and Alp, O. (2007). Integrated routing and scheduling of hazmat trucks with stops en route. *Transportation Science*, 41(1):107–122.

Errico, F., Desaulniers, G., Gendreau, M., Rei, W., and Rousseau, L. (2013). The vehicle routing problem with hard time windows and stochastic service times. *Les Cahiers du GERAD, Quebec, G-2013-45*.

Ferro, C. A. T. and Segers, J. (2003). Inference for clusters of extreme values. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 65(2):545–556.

Fleischmann, B., Gietz, M., and Gnutzmann, S. (2004). Time-varying travel times in vehicle routing. *Transportation science*, 38(2):160–173.

Gajewski, B. and Rilett, L. (2004). Estimating link travel time correlation: an application of Bayesian smoothing splines. *Journal of Transportation and Statistics*, 7(2-3):53–70.

Gómez, A., Mariño, R., Akhavan-tabatabaei, R., Medaglia, A. L., and Mendoza, J. E. (2016). On modeling stochastic travel and service times in vehicle routing. *Transportation Science*, 50(2):627–641.

Guessous, Y., Aron, M., Bhouri, N., and Cohen, S. (2014). Estimating travel time distribution under different traffic conditions. *Transportation Research Procedia*, 3:339–348.

Guin, A. (2006). Travel time prediction using a seasonal autoregressive integrated moving average time series model. *Proceedings of the IEEE ITSC*, 285:493–498.

Han, J., Lee, C., and Park, S. (2014). A robust scenario approach for the vehicle routing problem with uncertain travel times. *Transportation Science*, 48(3):373–390.

Health and Safety Executive (2014). Unloading petrol from road tankers - dangerous substances and explosive atmospheres regulations 2002.

Heffernan, J. E. and Tawn, J. A. (2001). Extreme value analysis of a large designed experiment: A case study in bulk carrier safety. *Extremes*, 4(4):359–378.

Highways England (2016). Online travel time files, http://tris.highwaysengland.co.uk, accessed 25-09-2017.

Highways England (2017). Highways england strategic road network initial report, pr128/17. Technical report, National Statistics.

Horn, M. E. T. (2000). Efficient modeling of travel in networks with time-varying link speeds. *Networks*, 36(2):80–90.

Huang, B., Cheu, R. L., and Liew, Y. S. (2010). GIS and genetic algorithms for HAZMAT route planning with security considerations. *International Journal of Geographical Information Science,*, 18(8):769–787.

Huang, L. and Barth, M. (2008). A novel loglinear model for freeway travel time prediction. In *2008 11th International IEEE Conference on Intelligent Transportation Systems*, pages 210–215.

Hyndman, R. J. (2010). Hyndsight (epub). Blog posted at https://robjhyndman.com/hyndsight/longseasonality/.

Ichoua, S., Gendreau, M., and Potvin, J.-Y. (2003). Vehicle dispatching with time-dependent travel times. *European Journal of Operational Research*, 144(2):379–396.

Janecek, A., Valerio, D., Hummel, K. A., Ricciato, F., and Hlavacs, H. (2015). The cellular network as a sensor: From mobile phone data to real-time road traffic monitoring. *IEEE Transactions on Intelligent Transportation Systems*, 16(5):2551–2572.

Janssens, G. K., Caris, A., and Ramaekers, K. (2009). Time petri nets as an evaluation tool for handling travel time uncertainty in vehicle routing solutions. *Expert Systems with Applications*, 36(3, Part 2):5987 – 5991.

Jie, G. (2010). Model and algorithm of vehicle routing problem with time windows in stochastic traffic network. In *2010 International Conference on Logistics Systems and Intelligent Management (ICLSIM)*, volume 2, pages 848–851.

Jula, H. and Dessouky, M. (2006). Truck route planning in nonstationary stochastic networks with time windows at customer locations. *IEEE Transactions on Intelligent Transportation Systems*, 7(1):51–62.

Karkazis, J. and Boffey, T. B. (1995). Optimal location of routes for vehicles transporting hazardous materials. *European Journal of Operational Research*, 86(2):201–215.

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

Khan, R.-A.-I., Landfeldt, B., and Dhamdhere, A. (2012). Predicting travel times in dense and highly varying road traffic networks using STARIMA models. *Technical report (University of Sydney. School of Information Technologies)*, (February):1–24.

Kheirkhah, A., Navidi, H., and Messi Bidgoli, M. (2015). A bi-level network interdiction model for solving the hazmat routing problem. *International Journal of Production Research*, 7543(October):1–13.

Khosravi, A., Mazloumi, E., Nahavandi, S., Creighton, D., and Van Lint, J. W. C. (2011). A genetic algorithm-based method for improving quality of travel time prediction intervals. *Transportation Research Part C: Emerging Technologies*, 19(6):1364–1376.

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

Li, X., Tian, P., and Leung, S. C. (2010). Vehicle routing problems with time windows and stochastic travel and service times: Models and algorithm. *International Journal of Production Economics*, 125(1):137–145.

Lindner-Dutton, L., Batta, R., and Karwan, M. H. (1991). Equitable sequencing of hazardous materials shipments. *Transportation Science*, 25(2):124–137.

List, G. F. and Mirchandani, P. B. (1991). An integrated network/planar multiobjective model for routing and siting for hazardous materials and wastes. *Transportation Science*, 25(2):146–156.

Malandraki, C. and Daskin, M. S. (1992). Time dependent vehicle routing problems: Formulations, properties and heuristic algorithms. *Transportation Science*, 26(3):185.

Marianov, V. and Revelle, C. (1998). Linear , non-approximated models for optimal routing in hazardous environments. *The Journal of the Operational Research Society*, 49(2):157–164.

Miller-Hooks, E., Hani, and Mahmassani, S. (1998). Optimal routing of hazardous materials in stochastic, time-varying transportation networks. *Transportation Research Record*, page 151.

Miller-Hooks, E. D. and Mahmassani, H. S. (2000). Least expected time paths in stochastic, time-varying transportation networks. *Transportation Science*, 34(2):198–215.

Min, W. and Wynter, L. (2011). Real-time road traffic prediction with spatio-temporal correlations. *Transportation Research Part C: Emerging Technologies*, 19(4):606–616.

Min, X., Hu, J., Chen, Q., Zhang, T., and Zhang, Y. (2009). Short-term traffic flow forecasting of urban network based on dynamic starima model. In *2009 12th International IEEE Conference on Intelligent Transportation Systems*, pages 1–6.

Min, X., Hu, J., and Zhang, Z. (2010). Urban traffic network modeling and short-term traffic flow forecasting based on GSTARIMA model. In *13th International IEEE Conference on Intelligent Transportation Systems*, pages 1535–1540.

Miranda, D. M. and Conceicao, S. V. (2016). The vehicle routing problem with hard time windows and stochastic travel and service time. *Expert Systems with Applications*, 64:104–116.

Nadarajah, S., Anderson, C. W., and Tawn, J. A. (1998). Ordered multivariate extremes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(2):473.

Nikovski, D., Nishiuma, N., Goto, Y., and Kumazawa, H. (2005). Univariate short-term prediction of road travel times. In *Proceedings. 2005 IEEE Intelligent Transportation Systems Conference*, pages 1074–1079.

Noorizadegan, M. and Chen, B. (2018). Vehicle routing with probabilistic capacity constraints. *European Journal of Operational Research*, 270(2):544 – 555.

Nozick, L. K., List, G. F., and Turnquist, M. A. (1997). Integrated routing and scheduling in hazardous materials transportation. *Transportation Science*, 31(3):200–215.

OpenStreetMap contributors (2017). Planet dump retrieved from https://planet.osm.org . https://www.openstreetmap.org.

Ord, J. and Fildes, R. (2013). *Principles of Business Forecasting.* South-Western Cengage Learning.

Patel, M. H. and Horowitz, A. J. (1994). Optimal routing of hazardous materials considering risk of spill. *Transportation Research Part A: Policy and Practice*, 28(2):119 – 132.

Pickands, J. (1971). The two-dimensional poisson process and extremal processes. *Journal of Applied Probability*, 8(4):745–756.

Pradhananga, R., Taniguchi, E., and Yamada, T. (2010). Ant colony system based routing and scheduling for hazardous material transportation. *Procedia - Social and Behavioral Sciences*, 2(3):6097 – 6108.

Pradhananga, R., Taniguchi, E., Yamada, T., and Qureshi, A. G. (2014). Environmental analysis of Pareto optimal routes in hazardous material transportation. *Procedia - Social and Behavioral Sciences*, 125:506–517.

Qiang, M., Der-Horng., L., and Cheu, R. L. (2005). Multiobjective vehicle routing and scheduling problem with time window constraints in hazardous material transportation. *Journal of Transportation Engineering*, 131(9):699–707.

Rahmani, M., Jenelius, E., and Koutsopoulos, H. N. (2015). Non-parametric estimation of route travel time distributions from low-frequency floating car data. *Transportation Research Part C: Emerging Technologies*, 58:343–362.

Russell, R. A. and Urban, T. L. (2008). Vehicle routing with soft time windows and Erlang travel times. *Journal of the Operational Research Society*, 59(9):1220–1228.

Salamanis, A., Kehagias, D. D., Filelis-Papadopoulos, C. K., Tzovaras, D., and Gravvanis, G. A. (2016). Managing Spatial Graph Dependencies in Large Volumes of Traffic Data for Travel-Time Prediction. *IEEE Transactions on Intelligent Transportation Systems*, 17(6):1678–1687.

Setak, M., Habibi, M., Karimi, H., and Abedzadeh, M. (2015). A time-dependent vehicle routing problem in multigraph with FIFO property. *Journal of Manufacturing Systems*, 35:37–45.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis.* Monographs on statistics and applied probability. Chapman and Hall, London ; New York.

Smith, R. L. and Weissman, I. (1994). Estimating the extremal index. *Journal of the Royal Statistical Society, Series B (Methodological)*, 56(3):515–528.

Sumalee, A., Watling, D., and Nakayama, S. (2006). Reliable network design problem: Case with uncertain demand and total travel time reliability. *Transportation Research Record: Journal of the Transportation Research Board*, 1964:81–90.

Ta, D., Dellaert, N., van Woensel, T., and de Kok, T. (2013). Vehicle routing problem with stochastic travel times including soft time windows and service costs. *Computers & Operations Research*, 40(1):214–224.

Taniguchi, E. and Shimamoto, H. (2004). Intelligent transportation system based dynamic vehicle routing and scheduling with variable travel times. *Transportation Research Part C: Emerging Technologies*, 12(3):235–250.

Tarantilis, C. and Kiranoudis, C. (2001). Using the vehicle routing problem for the transportation of hazardous materials. *Operational Research*, 1(1):67–78.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288.

TomTom (2015). Online files, goo.gl/aGjQ3r.

Tsay, R. S. (2015). *MTS: All-Purpose Toolkit for Analyzing Multivariate Time Series (MTS) and Estimating Multivariate Volatility Models.* R package version 0.33.

Turner, S. M., Eisele, W. L., Benz, R. J., and Douglas, J. (1998). Travel time data collection handbook. Technical Report 5, Department of Transport.

US Department of Transportation (2012). 2012 commodity flow survey, hazardous materials. Technical report, U.S. Census.

Wang, X., Crainic, T. G., and Wallace, S. W. (2019). Stochastic network design for planning scheduled transportation services: The value of deterministic solutions. *INFORMS Journal on Computing*, 31(1):153–170.

Wang, Z. and Lin, L. (2013). A Simulation-based algorithm for the capacitated vehicle routing problem with stochastic travel times. *Journal of Applied Mathematics*, 2013(127156):1:10.

Westgate, B. S., Woodard, D. B., Matteson, D. S., and Henderson, S. G. (2014). Large-network travel time distribution estimation for ambulances. *European Journal of Operational Research*, 252(1):322 – 333.

Wijeratne, A. B., Turnquist, M. A., and Mirchandani, P. B. (1993). Multiobjective routing of hazardous materials in stochastic networks. *European Journal of Operational Research*, 65(1):33–43.

Wong, J. (2012). *fastVAR*. R package version 1.9.9.

Wu, C.-H., Wei, C.-C., Su, D.-C., Chang, M.-H., and Ho, J.-M. (2003). Travel time prediction with support vector regression. In *Proceedings of the 2003 IEEE International Conference on Intelligent Transportation Systems*, volume 2, pages 1438–1442.

Zhang, J., Lam, W. H. K., and Chen, B. Y. (2013). A stochastic vehicle routing problem with travel time uncertainty: Trade-off between cost and customer service. *Networks and Spatial Economics*, 13(4):471–496.

Zhang, Y., Haghani, A., and Zeng, X. (2015). Component GARCH models to account for seasonal patterns and uncertainties in travel-time prediction. *IEEE Transactions on Intelligent Transportation Systems*, 16(2):719–729.

Zografos, K. G. and Androutsopoulos, K. N. (2004). A heuristic algorithm for solving hazardous materials distribution problems. *European Journal of Operational Research*, 152(2):507–519.