ANALYTICAL SOLUTIONS TO TRANSPORTATION DECISION-MAKING

By

PEDRO CESAR LOPES GERUM

A dissertation submitted to the
School of Graduate Studies
Rutgers, The State University of New Jersey
In partial fulfillment of the requirements
For the degree of
Doctor of Philosophy
Graduate Program in Industrial and Systems Engineering
Written under the direction of
Melike Baykal-Gürsoy
And approved by

_____

_____

_____

_____

_____

New Brunswick, New Jersey
October, 2020

ABSTRACT OF THE DISSERTATION

Analytical Solutions to Transportation Decision-Making

by PEDRO CESAR LOPES GERUM

Dissertation Director:

Melike Baykal-Gürsoy

This dissertation introduces original analytical methodologies for decision-making in transportation systems. Moving away from the conventional yet burdensome simulation approaches, we advance closed-form solutions that describe transportation-related processes. It contains two parts. Part 1 concentrates on the problem of predicting congestion on roadways, and part 2 focuses on the problem of scheduling inspections for railway track maintenance.

Part 1 provides a faster and more efficient method to determine traffic density behavior for long-term congestion management using minimal statistical information. Applications include road work, road improvements, and route choice. The research adapts and generalizes two models (off-peak and peak hours) for the probability mass function of traffic density in a major highway. It then validates them against real data. The studied corridor experiences randomly occurring service deterioration caused by accidents and inclement weather, such as snow and thunderstorms. We base the models on queuing theory, and we compare the fundamental diagram with the data.

This research supports the validity of the models for each traffic condition under certain assumptions on the distributional properties of the associated random parameters. Different scenarios demonstrate traffic congestion and traffic breakdown behavior under various deterioration levels. Last, we include a direct expansion of the model for non-space-homogeneous segments. These models, which account for non-recurrent congestion, can improve decision-making with no extensive datasets or

time-consuming simulations.

Part 2 considers inspection and maintenance activities in railways. They are essential to preserving railways' safety and cost-effectiveness. Still, one of the leading causes of derailments, the stochastic nature of railway defect occurrence, is rarely present in the related literature. Defect occurrence has been investigated as a stand-alone problem by other authors. Even then, models concentrating on defect prediction demand large datasets of obscure parameters that can be costly or infeasible to gather.

We propose a new method that relies on customary data for predicting track and geometry defects. We then develop a holistic approach to scheduling inspection and maintenance activities that integrates the prediction of railway defects into the problem. This integration is robust and allows for different constraints, such as crew limitations via a Multi-Armed-Bandit framework. Results indicate a high accuracy rate in prediction and effective scheduling policies that are adaptable to varying levels of risk tolerance. Finally, we theorize that search games can solve the final decision of where to inspect within the pre-defined segment.

Dedication

To Jader, whose card games inspired my passion for probability.

Grandpa, I wish you were here with me.

Acknowledgments

The past five years have been a turning point in my life. This dissertation represents the end of this era. Many crossed my path during my doctorate, and I express my sincere gratitude to all of you.

I am thankful to my advisor, Dr. Melike Baykal-Gursoy, for guiding me throughout this journey. Who would have guessed that a meeting at OQ would lead to 5 years of fruitful collaborations and growth? I am indebted to you because you had the confidence in my skills and believed in my potential from the start. With you, I learned not to give up. I fondly remember the times I showed up at your office quite disheartened with my research, just to leave with renewed energy and restored motivation. I admire your positive attitude and your ability to motivate people around you.

Mom and dad, I appreciate your support and encouragement along the way. From the very beginning, you supported me, comforted me, and listened to me. Dad, thanks for challenging me from an early age (I still cannot believe you used to read Malba Tahan for me to sleep). Mom, thanks for being my greatest supporter. I know my passion for numbers and my critical thinking come from you.

Ana, you helped me keep it real (God knows how stuck up I would be if it were not for you). I learned invaluable lessons from you. Aunt Lucia, thank you for the long conversations about life and for helping me have a perspective on things. Grandma Ada, thanks for making me feel like a hero everytime we talk.

This project would also not have come to fruition if it were not for the friends I made along the way.

Special thanks to Mike, who co-led my (our?) first camping trip as a leader, and to Noah, who helped ease my transition to New Jersey. You were the first who made

me feel welcome to this place. Thanks to my fantastic roommates, who lifted the burden when work burned me out. Noah, Kevin, Greg, Andrew, Christian, John, Burcu, and Luis: you rock!

To my outdoorsy friends Mike, Kevin, Greg, Erin, Ethan, Bobby, Jake, Julia, and Emily, thanks for showing me a new world I did not know existed. Lisia, Pupim, Luiz, Karin, Erik, Silvio, Allison, Valdir, and Vito, thank you for sharing the homesick days with me. I wonder if I would have forgotten how a pão de queijo tastes like had I not met you.

To Laurent, Sabrina, Daniel, and Miguel, thanks for making my time in California full of adventures. Finally, I am thankful for my Rutgers fellow peers, Andrew, Burcu, Nooshin, Jingbo, Stam, Vidita, and Clara. You endured my highs and lows, and we bonded over our failures and successes. Ayca and Andrew, thank you for co-authoring my first publications.

Last, though not least, I express my appreciation to the professors who helped, advised, and guided me along the way. Drs. Coit, Albin, Jafari, Lidbetter, Patel, Luxhoj, Valizadegan, Defimov, and Katehakis, I appreciate your patience and your the eagerness to share knowledge.

A final thank you goes to those who supported me financially throughout my journey. I am thankful to the National Council for Scientific and Technological Development of Brazil, Rutgers University, and American Express for facilitating my doctoral studies.

Contents

List of Tables

List of Figures

Introduction

In recent times, population concentration, economic growth, and lifestyle changes have increased the demand for traffic infrastructure at an unprecedented rate, outpacing improvements to infrastructure. There is a constant emergence of innovative technologies for all modals of transportation, such as autonomous cars and enhanced track materials to overcome this increase in demand. This dissertation proposes solutions for everyday problems that decision-makers still face both on roadways and railways, despite the currently available technology. We apply and adapt novel developments in the operations research field to find optimal and near-optimal solutions for decisions, often in closed-form. We hope that these solutions may foster elegant innovations that are less prone to bias in the field, reducing its current reliance on data and simulation.

In this dissertation, problems in roadways and railways serve as the background for our theoretical development. The models extend to other fields and, with little work, should be readily applicable to them. This section presents an overview of the problems assessed throughout the dissertation. We also brief the reader on the approaches proposed in the literature and describe the flaws that we intend to overcome. The following chapters further discuss previous studies in more detail.

Predicting Roadway Congestions

The rate of congestions and delays is increasing. According to the Urban Mobility Scorecard (Schrank et al., 2015), the average extra time spent in traffic per commuter due to congestion in the USA was 42 hours in 2014 compared to 18 hours in 1982. Nationwide, this means 6.8 billion hours or $160 billion in extra fuel consumption per year. Even the introduction of autonomous vehicles may not curb congestion due to high demand unless policies are developed to encourage ridesharing or improve flow management.

This problem becomes increasingly ubiquitous as the worldwide population grows and concentrates (Thakur et al., 2012). Unfortunately, transportation infrastructure investments are still lagging, putting pressure on those responsible for planning transportation systems (Schrank et al., 2015). Severe travel delays are frequently the result of the inappropriate planning of transportation systems (Chiou, 2016). Poor planning may cause insufficient provision of link capacity, particularly under uncertain travel demand and with the arrivals of non-recurrent incidents. Such issues need to be accounted for in roadway designs, even when it is infeasible to gather the appropriate data. Therefore, transportation investments must be carefully considered to help alleviate congestion and prevent future traffic problems.

While traffic congestion and breakdown caused by excess cars during peak hours are common, other types of congestion have started to become more pervasive. Non-recurrent congestions are a significant contributor to the total delay of vehicle travel time (Skabardonis et al., 2003; Kwon et al., 2006). The two main factors that cause non-recurrent delays are weather deterioration and accidents that affect the capacity of the road. They account for well over half of the non-recurrent delays in urban areas, and nearly all non-recurrent delays in rural areas (Skabardonis et al., 1998).

Next, in chapter 2, we present current approaches to describe traffic density distributions from literature, explain the data used in the validation of our traffic density distribution models, and introduce our models. Based on the previous research (Baykal-Gürsoy and Xiao, 2004; Baykal-Gürsoy et al., 2009a,b), the models described in this dissertation extend, simplify, and validate their approach against a real dataset.

Baykal-Gürsoy and Xiao (2004); Baykal-Gürsoy et al. (2009a,b) derive the distribution of density in closed-form, but its computation is nontrivial. Few properties of the distribution can be directly derived from its complex closed-form, limiting its application. However, in realistic scenarios, we can place reasonable bounds on the parameters, leading to a simpler closed-form distribution. The properties of the resulting distribution can then be directly computed and have an intuitive meaning, allowing decision-makers to attain a better understanding of the behavior of the system. Moreover, Baykal-Gürsoy and Xiao (2004), Baykal-Gürsoy et al. (2009a), and Baykal-Gürsoy et al. (2009b) provide no validations against real data. Chapter 2 also addresses this issue by validating the model against data obtained in Wisconsin.

To the best of our knowledge, there is not an agreed-upon steady-state probability distribution for traffic density in roadways that experience non-recurrent congestion. Most recent research that incorporates non-recurrent congestion focus on detecting incidents rather than computing the steady-state distribution (Anbaroglu et al., 2014; Chen et al., 2016; Laharotte et al., 2017). Mathematically, they work on the transient portion of congestion behavior, thus focusing on real-time applications instead of planning and long-term decision-making. Hence, there is a gap in the literature for direct approaches to computing the steady-state distribution of traffic density when accounting for non-recurrent congestion.

This study addresses the critical need for analytical congestion assessment models that are neither overly simplistic nor excessively complex. An outcome is a systematic

method that uses simple, usually in hand parameters, to estimate the probability distribution of traffic density. They are:

1. mean traffic flows under normal and deteriorated conditions;
2. mean speeds under normal and deteriorated conditions;
3. incident frequency;
4. repair rate;
5. deterioration rate.

Rail Defect Prediction and Rail Inspection and Maintenance Scheduling

The demand for rail transport is experiencing a boost at a global level, subsequently straining railway companies' ability to maintain service quality (Sharma et al., 2018). Simson et al. (2000) and Budai-Balke et al. (2009) argue that track maintenance and renewal costs are one of the highest expenses for railway companies. The literature divides maintenance and inspections into two kinds: a) corrective ones, which are inspections to repair known existing defects, and b) preventive ones, which are inspections and adjustments on defect-prone tracks (Sharma et al., 2018) before defects are detected. Notwithstanding the need for efficient inspection and maintenance to ensure the safety and security of transported public and goods, these activities may become costly if done excessively. This dissertation proposes algorithms to search for adequate policies that improve railway companies' processes for the inspection and maintenance of tracks.

One of the main factors that impact the decision to inspect and maintain a track is the number of defects it is expected to have. Hence, accurate defect predictions

are an essential step in this process. The literature mostly treats the prediction of defects and the scheduling of maintenance activities as different problems due to their complex nature involving numerous constraints. Besides, literature avoids treating all possible defect types together because of the increase in complexity. A few studies (Sharma et al., 2018; Merrick and Soyer, 2015) integrate prediction and maintenance. Sharma et al. (2018) introduce a Markovian model to predict whether defects happen or not, and use Markov decision theory to determine the optimal inspection and repair policy. Merrick and Soyer (2015) employ a nonhomogeneous Poisson process to model the stochasticity of track failures when planning for their replacement.

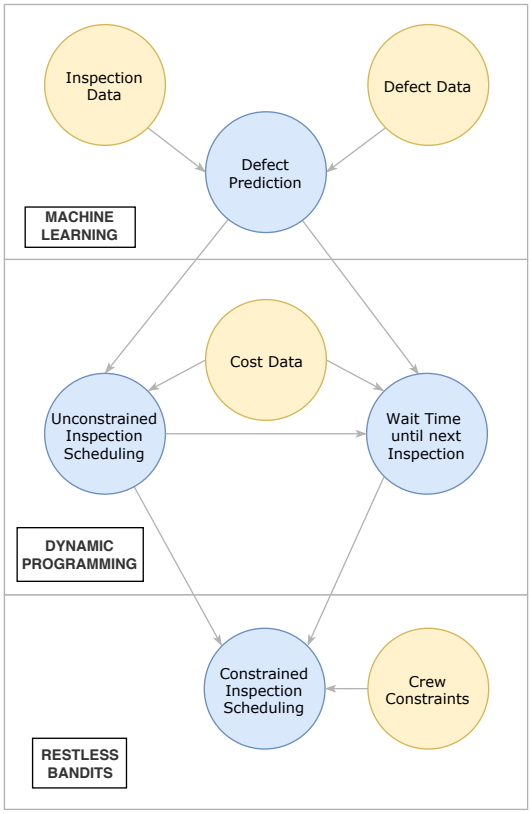


Figure 1: Flowchart of the integration between defect prediction and inspection.

In this dissertation, we propose an innovative integration of defect prediction and optimal scheduling, as well as solutions for issues both in defect predictions, and inspection and maintenance scheduling problem. First, we account for the stochastic

nature of defect prediction. Then, using the results obtained, we develop an iterative approach to schedule inspections with or without crew limitations optimally. Figure 1 visually describes the integration process proposed in chapters 5 and 6.

Lastly, inspection and maintenance crews are often unable to run throughout the whole segment, adding a constraint that limits the length of the inspection on a segment. Therefore, decision-makers may need to advise which sections to be prioritized. Current studies assume an equal likelihood across the segment for defect placement. However, in many cases, some areas within the segments are more defect-prone than others. Given the characteristic risk-aversion of this setup, this research proposes game-theoretical approaches to determine an optimal strategy for the inspection crews on where and when to inspect in chapter 7. This approach assumes Nature to be a hider that could place defects optimally. The decision-maker should minimize the expected cost of risk and the cost of inspection under the worst-case scenario. Search games have been commonly used in patrolling and security studies (Yolmeh and Baykal-Gürsoy, 2018; Lin and Singham, 2015; Roberts and Gittins, 1978), but adaptations for other applications are still incipient.

Part I

Roadways Recurrent and Non-recurrent Congestions — A queuing theory approach

Chapter 1

Literature Review

Several authors have derived solutions for the steady-state and the transient recurrent and non-recurrent congestion problem. However, proposed approaches for these problems face at least one of several shortcomings:

1. they require large datasets — and therefore are expensive or infeasible in most situations;
2. they lack a closed-form solution — and therefore require extensive computation;
3. they provide expected results, instead of full distributions — and therefore do not provide the full picture of the density behavior;
4. they do not account for non-recurrent congestion — and therefore lack robustness;
5. they focus on real-time traffic state estimation — and therefore lack predictive power;
6. they assume homogeneity of the auxiliary distribution parameters — and therefore limit the application to short segments.

This chapter reviews various approaches to congestion and density estimation proposed by the literature, and explains their flaws and opportunities for improvement.

1.1 Long-term congestion forecasting

The traditional methodology for traffic modeling was introduced by Lighthill and Whitham (1955) and Richards (1956). It approximates traffic as a deterministic fluid governed by a conservation equation relating the flow, speed, and occupancy, the so-called kinematic wave equation. While the initial models were powerful for modeling the emergent behavior seen in real traffic, they were mathematically cumbersome. Later models, like Newell (1993), made modifications that can accurately model traffic density with fewer technical complexities. However, these models are only capable of providing the expected flow parameters, not their distributions. Daganzo (1994) introduces the Cell Transmission Model (CTM), a numerical method to solve the kinematic wave equation. He also shows that the CTM can analyze non-recurrent incidents behavior on a scenario-by-scenario basis by temporarily altering initial conditions or parameters. The Switching-Mode (SMM) Model, proposed by Muñoz et al. (2003), provides an example of a CTM-based method with varying and parameters. Her approach consists of a combination of multiple CTMs, and it switches among them according to the current congestion level. The usefulness of this multiple-scenario approach is well demonstrated by the author, but it fails to include the stochasticity of real-life scenarios. It also does not provide the distributional information of the traffic density. Later, Morbidi et al. (2014) implement speed randomness into the original SMM model, assuming other parameters still to be deterministic.

An alternative approach divides traffic into much smaller sub-units, usually at the scale of a single-vehicle. Car-following models, like the Intelligent Driver Model proposed by Treiber et al. (2000), require extensive knowledge of driver characteristics.

These data are often costly or infeasible to gather, preventing the modeling of large or multiple studies. This shortcoming is particularly relevant in the planning of new areas where infrastructure does not exist. Cellular Automata models like Nagel and Schreckenberg (1992) do not have the same data requirements, and Daganzo (2006) shows that they converge to the solutions of the original formulation by Lighthill and Whitham (1955). The Cellular Automata models, as well as the CTM proposed by Daganzo (1994), can be stochastic, allowing traffic engineers to characterize the full distribution. Doing so, however, would require extensive simulation.

These simulation-based models are now standard in the industry, and few recent discoveries have occurred for long-term congestion forecast. Finally, even fewer papers have addressed non-recurrent incidents in their study of long-term behavior, although they are a significant contributor to the total delay of vehicle travel time and traffic breakdown (Skabardonis et al., 2003; Kwon et al., 2006). Baykal-Gürsoy and Xiao (2004) and Baykal-Gürsoy et al. (2009a,b) propose a model using queuing theory that accounts for non-recurrent congestion. They depict a segment of a roadway as two-state finite or infinite queues.

1.1.1 Probability of Traffic Breakdown

Part of the literature focuses on finding the probability of traffic breakdown. Traffic breakdown is triggered when a substantial speed decrease from the free flow speed is detected between two consecutive time intervals. This speed decrease drastically increases density, hence causing a sudden plunge in capacity. Kerner et al. (2002) adapt the original Cellular Automaton model to derive a theoretical probability for a spontaneous breakdown. With the further popularization of Cellular Automata models by Maerivoet and De Moor (2005), the concept of using simulation became a constant for the problem of finding the probability of traffic breakdown.

With time, other variations have surfaced, such as the Monte-Carlo simulation model proposed by Dong and Mahmassani (2012). Its novelty was the combination of a stochastic approach to macroscopic flow breakdown with a microscopic model of driver behaviors. However, these models face similar shortcomings – they require extensive computational resources and complex detailed data. Hence, probabilistic closed-form solutions for this problem have recently resurfaced in the literature (Arnesen and Hjelkrem, 2018; Han and Ahn, 2018). Even then, the solutions proposed are limited for planning purposes, as they provide little direction for the decision-maker and still require complex data information.

The problem of finding the probability of breakdown is contained in the more general problem of determining the full density distribution. The knowledge of the complete density distribution allows for direct computation of the probability of traffic breakdown. It also equips decision-makers with additional valuable information on the significance of different traffic parameters for breakdown and congestion.

1.2 Short-term congestion forecasting

Another problem extensively studied in the literature is the transient prediction of congestion, i.e., the forecast of congestion for a short period when the initial condition is known. Traditionally, CTM models were used to compute these behaviors, although still facing similar issues such as the need for extensive simulations. Kurzhanskiy (2009) created a variation of the original CTM designed explicitly for the transient prediction of traffic density. His contribution is a model that does now assume that cell capacities, arrivals, and measurement noises are known.

As a possible alternative for simulation-based models, Chrobok et al. (2004) also describe two simple prediction models for short-term forecasting using historical data. The constant model forecasts the same value for all horizons, while the linear model

fits a linear curve from the last N measured values. Zhang et al. (2016) propose the prediction of the congestion following an accident via a model that assumes the accident scene to be preserved. In both papers, the parameters used are deterministic. Conversely, Zhang et al. (2014) include the probabilistic nature of traffic in their genetic-algorithm-based approach for congestion prediction. The output, in this case, is binary (congestion/no-congestion).

The recent influx of data from new sources has been a boon for contemporary research in traffic congestion for dynamic systems, enabling machine learning approaches to progress. Several of these studies do not directly focus on density as a measure for congestion assessment. However, we conjecture that their implementation could be adapted to include density without a significant change in complexity. Initially proposed by Dougherty et al. (1993), these seminal models divide roadways into segments observed in discrete time units. Succeeding work by Dia (2001) suggests that dynamic architectures, such as Recurrent Neural Networks (RNNs), can perform better than simple fully-connected networks (MLPs). He uses the metrics of speed and flow as inputs for his model, and defines congestion as the combination of speed and flow parameters. More recently, Polson and Sokolov (2017) take advantage of the advancements in computational power and describe how a deep fully-connected neural network can accurately predict changes in traffic behavior caused by external events such as sports games or accidents. They also use metrics other than density in their model. Zhao et al. (2017) and Zhong et al. (2018) follow on the work of Dia (2001) and show that deep Long-Short Term Memory networks (LSTMs) are well suited to account for the time-dependence in traffic congestion. Lastly, Chen et al. (2018) develop a CNN-based approach to the same problem, yet their results are weaker than the ones obtained with other proposed architectures. These models also differ in the input used. Although most use common variations of traditional traffic

metrics (flow and speed), there is still no consensus in the literature regarding the most appropriate input for traffic density ML-based classifiers. To the best of our knowledge, no literature has implemented machine-learning-based predictive models that directly use density as the output parameter.

While these methods are promising for the future of traffic modeling, they are not well suited for planning applications. Typical solutions provide only expected values, and their accuracy decreases as the forecast horizon increases. These advantages are compelling for real-time traffic monitoring, but they do not transfer over to planning applications where these datasets are unavailable, and long-term distributional predictions are needed.

1.2.1 Traffic State Estimation and Incident Detection

Despite the predictive portion of the research, a considerable part of the literature centers on the detection of traffic states by tracking changes in available data. Their idea is that predictive analyses may be unnecessary in the short-term if real-time information is precise. Short-term predictive models mostly concentrate on the 10 to 40-minute window following the known instant. Therefore, the difference in response time may be sufficiently minor that a robust real-time estimator may be more useful than an ill-fit short-term predictor. Notwithstanding, short-term and long-term predictive models still benefit from this research as more accurately gathered data lead to more reliable predictions.

The concept of traffic state estimation was introduced by Wang and Papageorgiou (2005). It consists of determining all traffic variables (density, speed, and flow) at the current time instant based on real-time measurements. Initial developments proposed the usage of Kalman Filter based models for traffic state estimation (Sun et al., 2004; Wang and Papageorgiou, 2005). Wang et al. (2009) revised and upgraded the original

traffic state estimator model. Later on, Celikoglu (2014) and Celikoglu and Silgu (2016) propose the application of a Neural Networks (NNs) approach for traffic state estimation. They employ recent and current speed and flow data to determine in which region of the fundamental diagram traffic stands at the time. As a means to accurately represent the dynamically-changing Fundamental Diagram, they rely on simulation. The estimated Fundamental Diagram and NN classification model are then combined to compute the estimated density.

Current research to advance traffic state estimation uses microscopic simulation (Papadopoulou et al., 2018), a stochastic Lagrangian model with parametric uncertainties (Zheng et al., 2018), and relative flows from stationary and moving observers in a streaming-data-driven methodology (van Erp et al., 2018).

A separate approach also discussed in the literature is the advancement of the data generation process. Conventional detectors are inductive-loop traffic detectors that depend on a vehicle's ferrous body material to trigger its sensors. Qiu et al. (2010) propose the inclusion of probe data in the analysis, as subsequent detectors should have correlated outputs. Kwong et al. (2010) show that matching algorithms can be used to generate traffic data from wireless sensor networks efficiently. Lastly, other authors suggest innovative approaches, such as using smartphones as a source for data (Panichpapiboon and Leakkaw, 2017), and obtaining data from large-scale web camera pictures and videos (Zhang et al., 2017).

Indirectly, one of the main features these models aim to provide is incident detection. This particular problem has been studied separately with the emergence of new sources of data. Initially proposed by Ritchie and Cheu (1993), the research in artificial intelligence for incident detection only took off in more recent years. Lately, several authors have noted the potential of different AI approaches – Random Forests, Logistic Regression, and Neural Networks – for incident detection (Cheng et al., 2015;

Agarwal et al., 2016; Dogru and Subasi, 2018).

Despite the accurate results, these methods still struggle with the requirement of intense computational demands, and, in most cases, the dependency on simulation. They also do not produce a distributional output and have little long-term predictive application.

1.3 Stochastic Queuing Models

A theoretical framework adopted by some authors to predict congestion is queuing theory. Queuing analyses, together with deterministic (fluid-dynamics) models (May and Keller, 1967; Newell, 1971), are primarily used for performance evaluation and the synchronization of traffic lights (Newell, 1965). Early stochastic models assume individual vehicle arrivals to follow a Poisson process (Zheng and Liu, 2017; Wang and Ahmed, 2017; Gazis, 2006; J.N. Darroch and Morris, 1964; Tanner, 1953), or as platoons of vehicles (Daganzo, 1994; Alfa and Neuts, 1995; Dunne, 1967; Lehoczky, 1972) to represent the behavior of cars moving between traffic signals.

Cheah and Smith (1994) and Jain and Smith (1997) studied stochastic queues to explore the usefulness of finite server queuing models with state-dependent travel speed for modeling both pedestrian and vehicle traffic flows. In this model, vehicles arrive according to a Poisson process, and the total time to traverse the corridor is assumed to follow a general distribution. If the roadway is at capacity, new arrivals must take an alternative path. Consider vehicles traveling on a corridor, as depicted in Figure 1.1. The space occupied by one individual vehicle represents a moving server. The service cycle initiates when a car arrives at the corridor, and service (the act of traveling) occurs until the vehicle leaves the corridor. A server in this context is the moving vehicle-space, including the safe distance (space headway) to the car in front. The number of available servers is then the maximum number of vehicles that

the corridor can physically accommodate. Consequently, when there is no space left for a car to enter, i.e., all servers are full, the vehicle either is allowed to wait in a queue or is denied service (Cheah and Smith, 1994; Jain and Smith, 1997).

Other authors (Heidemann, 1996; Vandaele et al., 2000; Heidemann, 2001) studied a similar system but with a single server and infinite queuing capacity. Such a flow model is considered as a vertical queue that disregards the interdependence between vehicles within the same cell (Daganzo, 1994). Although some of these models focus on congestion, none of them include the occurrence of non-recurrent incidents as part of the model.

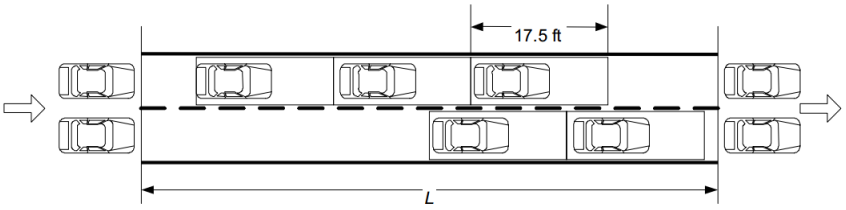


Figure 1.1: Graphical depiction of a two-lane roadway segment.

The queuing model proposed by Baykal-Gürsoy and Xiao (2004) and Baykal-Gürsoy et al. (2009a,b) is the only one that considers non-recurrent congestion.

Chapter 2

Traffic Density in Homogeneous Sections

This chapter proposes a mathematical framework for long-term congestion prediction that answers questions such as *1. How much would a change in traffic behavior impact traffic congestion? 2. How can we estimate the probability of traffic flow breakdown on roadways when minimal data is available?* Moreover, by defining traffic breakdown as the particular number of cars for a segment that causes speed to drastically decreased, this model explains how the probability of traffic breakdown changes according to changes in traffic parameters, and how improvements on the road clearance time impact congestion. A few examples of how to directly compute the probability of traffic breakdown with the model can be found in tables 2.4 and 2.5. This method simplifies and improves the assessment of non-recurring traffic density and its variability, leading to a reduction in congestion. Consequently, it leads to a safer, more efficient movement of people and goods, while also ensuring the timely delivery of critical resources for national security, emergency response and evacuation, and humanitarian relief efforts.

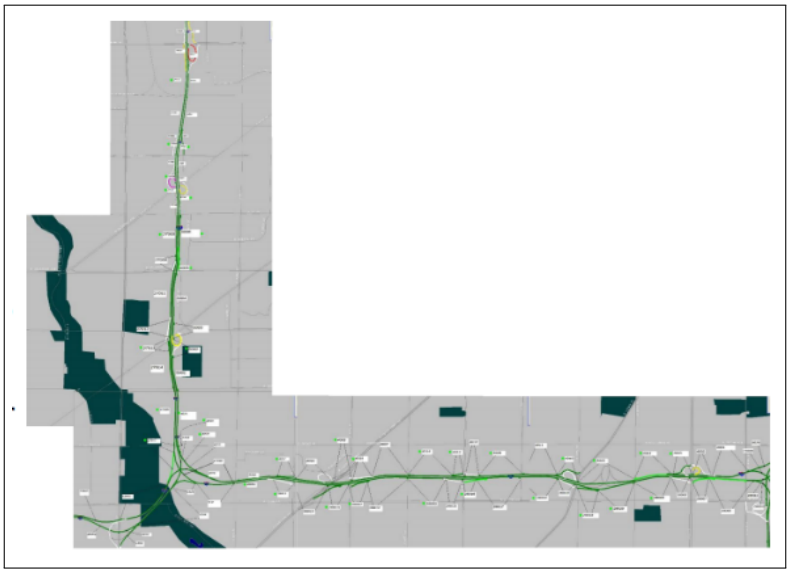


Figure 2.1: Map from where Data were Collected.

2.1 Measuring Traffic Data

In this section, we provide one example of how one may assess the congestion problem with real data. Data were collected from Wisconsin via 36 sensors in a 9-mile stretch, 18 on each side of the road (South-East and West-North Directions). Those sensors are inductive loop detectors embedded in the roadway depicted in Figure 2.1.

The sensors record three common indicators used in traffic models: speed, flow, and density. Density represents traffic congestion, counting the number of vehicles occupying a particular space unit at a specific time. Flow counts the number of cars passing through a certain point per unit time. Density, speed, and flow are closely related, as increasing flow also increases density initially. However, when more cars arrive than the highway can hold (surpassing the maximum flow), density keeps growing while flow decreases. The limit of the relationship occurs when traffic is down to a complete stop (flow is zero), hence representing the breakdown of traffic (Daganzo, 1994). Figure 2.2 represents a partial view of these relationships and showcases Milwaukee's data that this chapter uses for validation. Most current data-gathering

techniques cannot measure density directly but obtain speed, occupancy, and volume from which traffic density can be calculated (Kurzhanskiy, 2009). Occupancy is a proxy measure for density. It gives the percent coverage of the sensor per unit time, while volume is a proxy measure for flow, when the period analyzed is divided into equally spaced time intervals. In this section, we describe the methods used to compute density given speed, occupancy, and volume.

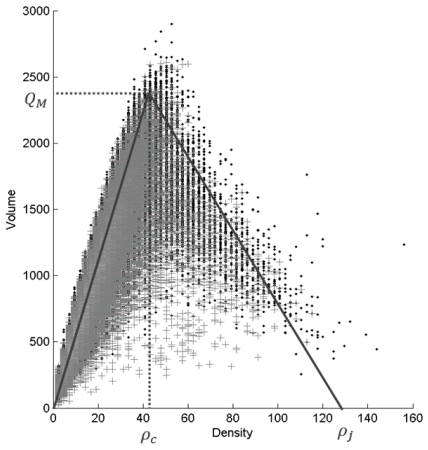


Figure 2.2: Fundamental Diagram.

The sensors provide occupancy measures for half-mile segments. In this chapter, we consider an average car length plus headway of 22 ft, obtained through the procedure described in Dailey (1999). We use ordinary least squares to find the best fit for the average car length plus headway. The number of cars, namely density, is then given by a simple ratio of occupancy times the length of the segment by the average length and headway of vehicles. However, the sensor's limitations cause occupancy data to be recorded in increments of 0.002. As a result, density generated from the data is rounded to specific values. Finally, although sensors also recorded speed, the data was truncated at the speed limit (at 60 mph for some sensors, 65 mph for others). A 'reconstructed speed' data set is then constructed from the volume and occupancy data using a method proposed by Dailey (1999).

As seen in Figure 2.1, there are numerous merges and forks through which vehicles may leave and arrive. This introduces further error to volume and occupancy, therefore making the data harder to analyze. However, for most scenarios, the additional noise does not undermine the validity of the model.

The most relevant traffic properties utilized in the validation are the frequency of incident occurrence, f, the duration until clearance and recovery – so-called incident clearance time, $\frac{1}{r}$, and the severity of capacity reduction caused by an incident, α. We gather data from accidents and weather conditions to better understand incidents. We then compare the sensors' data with data from weather conditions to find relations between incidents or traffic deterioration and extreme weather events such as fog, snowstorms, thunderstorms, rain, or normal conditions. Precipitation data were obtained from the Climate Data Online system of the National Climatic Data Center of the National Oceanic and Atmospheric Administration (NOAA, 2016). These data depict the hourly amount of precipitation in hundredths of inches recorded at the Milwaukee Mitchell International Airport weather station, for the same period as the traffic data. The data also include information on snow days and days with fog and thunderstorms. Our findings show that snowstorms and fog cause the highest impact on travel time.

After comparing the sensor's data with weather data and accident reports, we can split the speed, occupancy, and volume information into two situations: normal conditions (uptime), and adverse conditions (downtime). Downtime represents periods in which an accident or inclement weather condition deteriorates traffic. Furthermore, data are separated into peak hour and non-peak hour and had weekends and late nights removed for accuracy. This separation allows us to analyze the behavior of each group separately. Lastly, winter month's data is initially chosen to be the validation scenario during non-peak hours, because extra congestion due to weather

during wintertime tends to be predominant and extensive. Mean clearance time and the mean time to an incident are calculated using only the times considered.

Since the previous assumption could be biased towards long stretches of down periods, i.e., snowfalls that last several days, we also considered the case in which the system may only be affected by incidents, with the occasional heavy rain. In this situation, we analyze the frequency of incident occurrence, as well as the duration until clearance for the summer months during non-peak hours. We show that the model works in both situations.

2.2 Analytical Model

The main intention of this chapter is to analytically describe the probability mass function of the number of vehicles on a road segment (traffic density) while accounting for non-recurrent incidents. We also show that simulation and extensive datasets may be unnecessary for long-term planning. In this section, we expand on earlier research by creating a closed-form analytical model applicable to any roadway. A few of the assumptions used in the model are discussed and compared to a real-world dataset. The results include a simple approximation for the model proposed by Baykal-Gürsoy and Xiao (2004) and Baykal-Gürsoy et al. (2009b) during non-peak hours and a generalization for peak hours by Baykal-Gürsoy et al. (2009a). Their models consider a segment of a road operating in a two-state environment process as a Markovian queuing system. The two environment states represent the situation of the roadway, which could be under normal or adverse conditions. The latter refers to incidents such as snowstorms or accidents. They cause a reduction in the road capacity because of closures or blockages, or because drivers tend to slow down and increase the distance to the car in front for increased safety. The time in each environment state is assumed to be random.

2.2.1 Model Assumptions

Our model relies on the following assumptions:

1. Times to incidents and clearance times follow an exponential distribution.
2. Distributions for each segment can be generated independently.
3. Travel times are exponentially distributed, under normal and deteriorated conditions.
4. Arrivals follow a Poisson process, under normal and deteriorated conditions.

Assumption (1) is met. The histograms for the time to incident and clearance times are plotted together with a fitted exponential distribution in Figures 2.3 and 2.4 respectively. The distributions are fit with their maximum likelihood estimators. r^2 values, the proportion of total variation in the outcomes explained by the model, are reported for each fit. They indicate that both times to incident and clearance times can be modeled as exponential random variables. Assumption (1) allows us to model the environment with a Markovian model, for which servers change state according to a continuous-time Markov chain (CTMC).

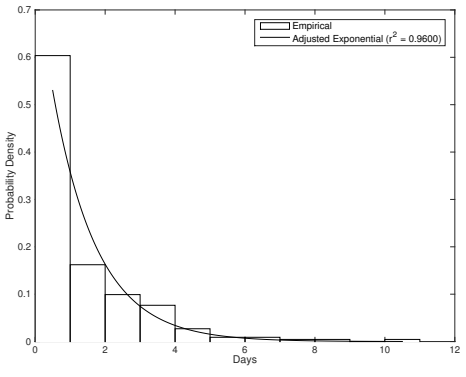


Figure 2.3: Time to Incident pdf.

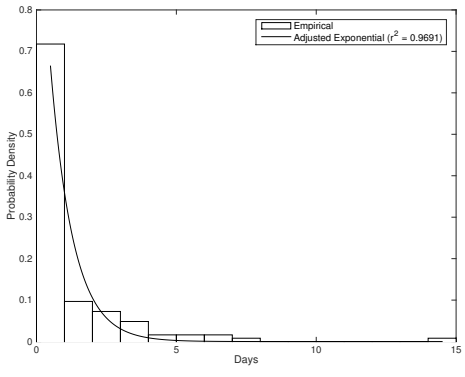


Figure 2.4: Clearance Time pdf.

Assumption (2) is also met. Although propagation effects will likely impact the parameters for upcoming segments, we expect that traffic engineers will use this model for individual segments in which they know or can estimate all the parameters. Section 2.2.2 lists these parameters, as well as their interpretation and methods of estimation. Capturing the correlation between sensors would only be necessary if the parameters were partially known, as depicted in figure 2.5. The modeling approach proposed in this chapter's modeling approach generates parameters from independent datasets for each segment, thus accounting for possible propagation effects. Therefore, distributions can be generated independently despite a possible correlation between the parameters of different segments.

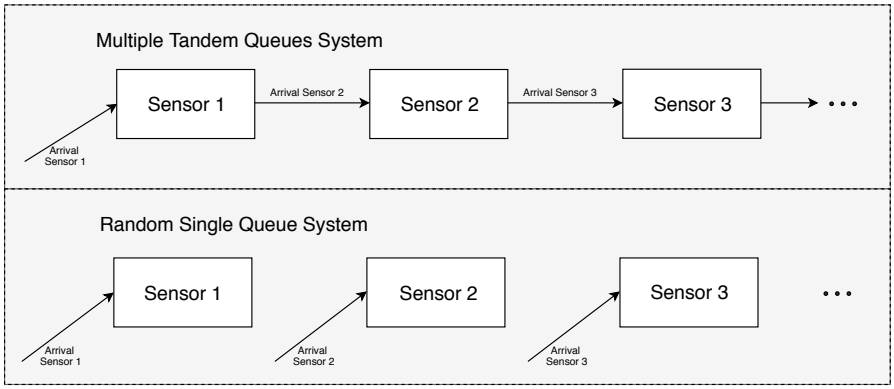


Figure 2.5: Representation of Arrivals for the System Proposed in this Model.

Furthermore, the correlations between sensors are obtained indirectly by the model. The time of breakdowns will likely be correlated between segments, as weather events and car accidents will frequently impact adjacent segments. This correlation is captured by the change in arrival and service rates during a breakdown.

Assumption (4) is also not met by our dataset. However, this assumption is commonly used in literature (J.N. Darroch and Morris, 1964; Gazis, 2006; Zheng and Liu, 2017; Wang and Ahmed, 2017). Although some methods in literature do not assume this, they introduce further complexity in the model, which then requires simulation

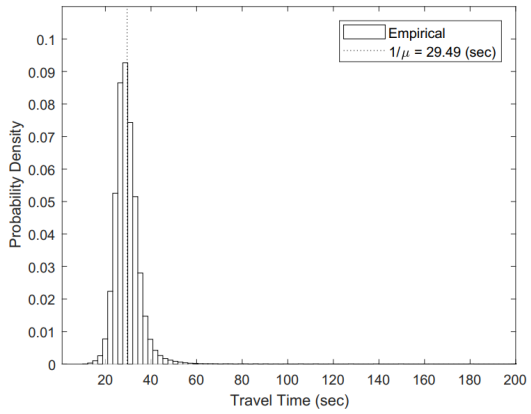


Figure 2.6: Travel Time pdf under Normal Conditions.

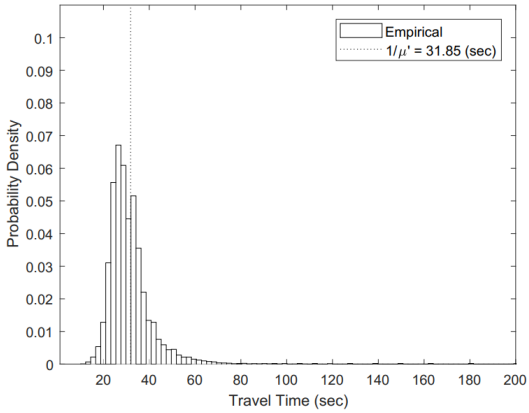


Figure 2.7: Travel Time pdf under Adverse Conditions.

for solutions. Numerical experiments indicate that many of these traffic simulations produce nonhomogenous Poisson arrival processes. An example can be found in figure 2.8, which contains the histogram of interarrival times at an arbitrary cell in a Cellular Automata model. Similar results can be found for other microscopic traffic models, suggesting that even models which do not explicitly assume that arrivals are Poisson ultimately reproduce a (possibly nonhomogenous) Poisson process.

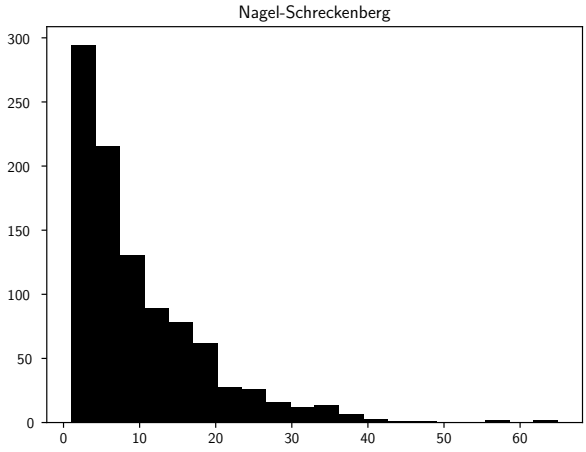


Figure 2.8: Histogram of Vehicle Interarrival Times from Nagel-Schreckenberg Simulation.

2.2.2 Model Structure and Parameters

Non-peak hours

In the queue shown in Figure 2.9, each state of the Markov chain represents both the number of cars in the system and the condition, normal or adverse, of the system. Figure 2.9 also depicts the definitions of parameters.

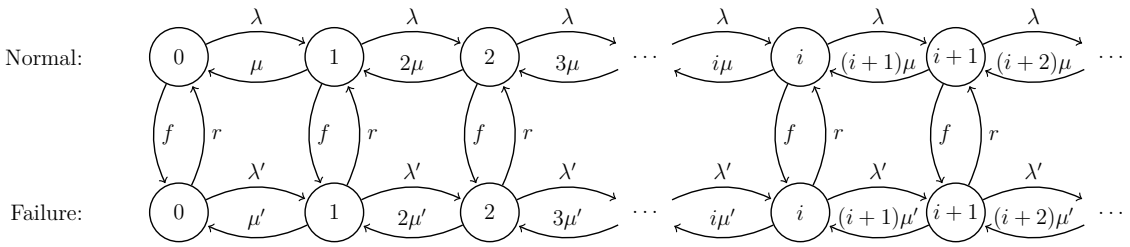


Figure 2.9: State Transition Diagram for a Markov-modulated $M/M/\infty$ Queue.

Parameters λ and λ' represent the arrival rate of the system in normal and adverse conditions, respectively. It is typically true that $\lambda \geq \lambda'$, although this is not necessary. Under the assumptions of section 2.2.1, the expected time between arrivals is given by $\frac{1}{\lambda}$ and $\frac{1}{\lambda'}$. This gives a method for estimating λ and λ' directly from traffic flow data:

$$\begin{aligned}\lambda &= \frac{1}{\mathbb{E}[\text{time between arrivals during normal conditions}]} \\ &= \mathbb{E}[\text{flow during normal conditions}], \\ \lambda' &= \frac{1}{\mathbb{E}[\text{time between arrivals during adverse conditions}]} \\ &= \mathbb{E}[\text{flow during adverse conditions}].\end{aligned}$$

Parameters μ and μ' represent the service rate of the system in normal and adverse conditions, respectively. The service rate is the instantaneous probability of a car leaving the segment of the road, and therefore, leaving the system. Because the

model allocates one server per vehicle, there are infinitely many servers (as long as full capacity is not reached). This assumption applies to non-peak hours, given the unlikelihood of traffic breakdown during non-peak hours. We remove this assumption when discussing peak hours. The more cars there are in the system, the more likely one of them will leave. Thus, if there are n cars on the segment, then the total service rate becomes $n\mu$ during normal conditions, and similarly $n\mu'$ during adverse conditions. Additionally, $\mu > \mu'$ because the service rate must be higher during normal conditions.

The service rate parameters come from the relationship between distance and speed. μ represents the rate of cars crossing a segment when the system is under normal conditions and is given by the ratio of the average speed and the segment length under normal conditions. Analogously, μ' represents the rate of cars crossing a segment when the system is under adverse conditions and is given by the ratio of the average speed and distance under adverse conditions. This interpretation allows us to estimate μ and μ' given the distance between sensors and the reconstructed speed (described in section 2.1):

$$\begin{aligned}\frac{1}{\mu} &= \frac{\text{segment length}}{\mathbb{E}[\text{speed reconstructed during normal conditions}]} \\ &= \mathbb{E}[\text{travel time during normal conditions}], \\ \frac{1}{\mu'} &= \frac{\text{segment length}}{\mathbb{E}[\text{speed reconstructed during adverse conditions}]} \\ &= \mathbb{E}[\text{travel time during adverse conditions}].\end{aligned}$$

Parameters f and r represent the incident rate when the system is in normal condition, and the repair rate when the system is in an adverse condition. They are determined from the incident and weather reports by averaging the time until an incident (accident or adverse weather condition) occurs and the clearance time of each incident. The failure rate (the number of times the system goes from normal

condition to adverse condition per unit time) is then given by one over the mean time to an incident, or the expected time that the system will remain in normal condition. Analogously, the clearance rate (the number of times the system goes from adverse condition to normal condition per unit time) is then given by one over the mean clearance time. For our dataset, the expected clearance time during summer is around 1 hour. The expected clearance time during winter is around 4.5 hours because snowfalls have a long lingering effect that takes longer to be cleared.

$$f = \frac{1}{\mathbb{E}[\text{time to incident}]} = \frac{1}{\mathbb{E}[\text{duration of normal condition}]},$$
$$r = \frac{1}{\mathbb{E}[\text{clearance time}]} = \frac{1}{\mathbb{E}[\text{duration of adverse condition}]}.$$

Keilson and Servi (1993) were the first to study such queues in a random environment. They derived the generating function of the stationary number of customers in the system in terms of Kummer functions. Baykal-Gürsoy and Xiao (2004) and Baykal-Gürsoy et al. (2009a,b) showed that the generating function reveals that the steady-state number of vehicles in the system is composed of two independent random variables. One represents the number of customers in an uninterrupted queue, and the other represents the customers accumulated during interruptions. In other words, the random number of cars on a corridor, X, is equal to $X = X_\phi + Y$ in steady-state, with X_ϕ representing the random number of vehicles accumulated during the normal condition, and Y representing the additional cars accrued due to incidents. Moreover, X_ϕ and Y are independent of each other.

Furthermore, the complete distributions of X_ϕ and Y are derived. In equilibrium, X_ϕ follows a Poisson, while $f_Y = p \cdot f_{Y_1} + (1-p) \cdot f_{Y_2}$ follows a mixture of two Poisson random variables Y_1 and Y_2 with random parameters coming from two different

truncated Beta distributions, as detailed below:

$$X_\phi \sim \text{Poisson}\left(\frac{\lambda}{\mu}\right), \tag{2.1}$$

$$Y_1 \sim \text{Poisson}\Big(B(a, b, -2\rho^*)\Big), \tag{2.2}$$

$$Y_2 \sim \text{Poisson}\Big(B(a+1, b+1, -2\rho^*)\Big), \tag{2.3}$$

$$\text{with: } a = \frac{f}{\mu},\ b = \Big(\frac{f}{\mu} + \frac{r}{\mu'}\Big),\ \rho^* = \frac{1}{2} \cdot \Big(\frac{\lambda}{\mu} - \frac{\lambda'}{\mu'}\Big),\ p = \frac{\Big(r + \frac{f\mu'}{\mu}\Big)}{(r+f)},\ c = -2\rho^*.$$

Once these parameters are determined, the model yields the probability mass function for the traffic density. We compare it with the empirical traffic density obtained from occupancy data to assess how good the fit is per sensor.

As mentioned before, each sensor is treated individually for validation purposes. Nevertheless, the outputs of the analytical models generated for each sensor are correlated outputs because of the inherent correlation in the input data.

Model improvements

The sensor's data yield parameters a and b that are on the order of 10^{-5} for winter months, and a in the order of 10^{-5} and b in the order of 10^{-3} for summer months, meaning that incident and clearance rates, f and r, are considerably lower than μ and μ'. This relationship among these parameters depicts an ordinary situation since the rate of accidents and clearances tend to be notably lower than the rate of cars crossing the segment. This section proposes a simple approximation for this case. The simplified model has the accuracy of the full model but is simpler to calculate.

Another underlying assumption used in this model is that the addition of new cars will not affect the travel time – i.e., that the distribution for travel time is independent of the number of vehicles in the system. For non-peak hours, the primary source of congestion is not the arrival rate, but rather system deterioration caused by non-recurrent events, such as accidents or weather conditions. The second half of this section considers peak hours, and it also disregards this assumption.

Applying these reasonable assumptions results in a single equation for the probability mass function of traffic density, which we later validate against the data. Moreover, we prove this equation to be just a mixture of two Poisson distributions when times to incidents and clearance times are considerably longer than travel times under normal and adverse conditions.

As proposed by Baykal-Gürsoy and Xiao (2004), the probability mass function of traffic density is a convolution sum between a Poisson random variable and a mixture of two Poisson random variables with random parameters coming from truncated Beta distributions. It represents the probability mass function for the state of an infinite queue subject to two-server states.

$$X = X_\phi + Y, \tag{2.4}$$

$$f_Y = p \cdot f_{Y_1} + (1-p) \cdot f_{Y_2}. \tag{2.5}$$

The probability mass function of Y_1 given in equation 2.2 can be explicitly written as:

$$P\{Y_1 = k\} = \int_0^c e^{-\gamma} \cdot \frac{\gamma^k}{k!} \cdot \frac{\Gamma(b)}{\Gamma(a)\Gamma(b-a)} \cdot \frac{\left(\frac{\gamma}{c}\right)^{a-1}\left(1-\frac{\gamma}{c}\right)^{b-a-1}}{c} d\gamma. \tag{2.6}$$

$P\{Y_2 = k\}$ can be computed through the same integral, but via substituting (a) to $(a+1)$ and (b) to $(b+1)$. Hence, derivations below can also be carried out for $P\{Y_2 = k\}$ following the same operations.

This integral is not robust for all ranges of a and b. Additionally, its lack of a closed-form prevents it from being readily applied. There are several effective methods for computing this, the most simple being solving the balance equations numerically. However, we will retain this closed-form as it allows for further simplifications.

Note that the exponential factor in equation 2.6 could be expanded by using Taylor series to rewrite it as a sum multiplied by the truncated beta function that is equal

to $\frac{\Gamma(k+n+a)\Gamma(b-a)}{\Gamma(k+n+b)}$ (see Olver (2010)).

$$P\{Y_1 = k\} = \sum_{n=0}^{\infty} \frac{(-1)^n}{n!} \int_0^c \gamma^n \cdot \frac{\gamma^k}{k!} \cdot \frac{\Gamma(b)}{\Gamma(a)\Gamma(b-a)} \cdot \frac{\left(\frac{\gamma}{c}\right)^{a-1}\left(1-\frac{\gamma}{c}\right)^{b-a-1}}{c} d\gamma$$

$$= \sum_{n=0}^{\infty} \frac{(-1)^n}{n!} \cdot \frac{c^{k+n}}{k!} \cdot \frac{\Gamma(b)}{\Gamma(a)\Gamma(b-a)} \cdot \int_0^c \left(\frac{\gamma}{c}\right)^{k+n+a-1} \left(1-\frac{\gamma}{c}\right)^{b-a-1} d\frac{\gamma}{c}$$

$$= \sum_{n=0}^{\infty} \frac{(-1)^n}{n!} \cdot \frac{c^{k+n}}{k!} \cdot \frac{\Gamma(b)}{\Gamma(a)\Gamma(b-a)} \cdot \int_0^1 x^{k+n+a-1}(1-x)^{b-a-1} dx$$

$$= \sum_{n=0}^{\infty} \frac{(-1)^n}{n!} \cdot \frac{c^{k+n}}{k!} \cdot \frac{\Gamma(b)\Gamma(k+a+n)}{\Gamma(a)\Gamma(k+b+n)}.$$

The full equation can then be simplified by expanding the convolution sum to:

$$\begin{aligned} X =& X_\phi + Y, \\ P(X=k) =& \sum_{q=0}^{k} P\{X_\phi = k-q\} P\{Y = q\} \\ P(X=k) =& \sum_{q=0}^{k} P\{X_\phi = k-q\}(p \cdot P\{Y_1 = q\} + (1-p) \cdot P\{Y_2 = q\}) \end{aligned}$$

$$\begin{aligned} &= \sum_{q=0}^{k} e^{-\lambda/\mu} \frac{(\lambda/\mu)^{k-q}}{(k-q)!} \Bigg[p \cdot \sum_{n=0}^{\infty} \frac{(-1)^n}{n!} \cdot \frac{c^{q+n}}{q!} \cdot \frac{\Gamma(b)\Gamma(q+a+n)}{\Gamma(a)\Gamma(q+b+n)} \\ &+ (1-p) \cdot \sum_{n=0}^{\infty} \frac{(-1)^n}{n!} \cdot \frac{c^{q+n}}{q!} \cdot \frac{\Gamma(b+1)\Gamma(q+a+1+n)}{\Gamma(a+1)\Gamma(q+b+1+n)} \Bigg]. \end{aligned} \tag{2.7}$$

This new equation solves the issues the integral had for the extreme points.

In practice it is typically true that $f \ll \mu$ and $r \ll \mu'$. This has the interpretation that the time a vehicle spends traveling a segment is much shorter than the time it takes for traffic to accumulate or disperse from an incident or clearance. Under this assumption, further simplifications are possible. Tricomi and Erdélyi (1951) prove

the following asymptotic approximation for the quotient of gamma functions:

$$\frac{\Gamma(z+a)}{\Gamma(z+b)} = z^{a-b}\left[1 + \frac{(a-b)(a+b-1)}{2z} + O(|z|)^{-2}\right]. \tag{2.8}$$

This is approximately 1 when $|a-b| \ll z$, which is implied by $f \ll \mu$ and $r \ll \mu'$. For $k = 0$:

$$P\{Y_1 = k\} = \sum_{n=0}^{\infty} \frac{(-1)^n}{n!} \cdot c^n \cdot \frac{\Gamma(b)\Gamma(a+n)}{\Gamma(a)\Gamma(b+n)}$$

$$= \frac{\Gamma(b)\Gamma(a)}{\Gamma(a)\Gamma(b)} + \sum_{n=1}^{\infty} \frac{(-1)^n}{n!} \cdot c^n \cdot \frac{\Gamma(b)\Gamma(a+n)}{\Gamma(a)\Gamma(b+n)}$$

$$\approx 1 + \frac{\Gamma(b)}{\Gamma(a)}(e^{-c} - 1). \tag{2.9}$$

For $k \geq 1$:

$$P\{Y_1 = k\} = \sum_{n=0}^{\infty} \frac{(-1)^n}{n!} \cdot \frac{c^{k+n}}{k!} \cdot \frac{\Gamma(b)\Gamma(k+a+n)}{\Gamma(a)\Gamma(k+b+n)}$$

$$\approx \frac{\Gamma(b)}{\Gamma(a)} \frac{e^{-c} \cdot c^k}{k!}, \tag{2.10}$$

The final equation is the weighted Poisson distribution.

This approximation is imprecise, since the ratio $\frac{\Gamma(k+a+n)}{\Gamma(k+b+n)}$ slowly diverges from 1 as $k+n$ grows bigger. However, the increase in the factorial terms in the equations grows faster, thus offsetting such divergence. Numerical experiments indicate that this approximation has no apparent adverse effect on the final density function.

The derivations to determine the equations to solve for $P\{Y_1 = k\}$ can be used in a similar way for $P\{Y_2 = k\}$, under the same assumptions (small a's and b's). Since the only parameters changing are $a = a + 1$ and $b = b + 1$, it is easy to see that again $\frac{\Gamma(k+a+1+n)}{\Gamma(k+b+1+n)}$ can be approximated as 1 for all k's; besides, $\frac{\Gamma(b+1)}{\Gamma(a+1)}$ can also be approximated to 1, allowing for even further simplification.

Thus, the probability mass function of Y_2, for all k's is:

$$P\{Y_2 = k\} = \sum_{n=0}^{\infty} \frac{(-1)^n}{n!} \cdot \frac{c^{k+n}}{k!} \cdot \frac{\Gamma(b+1)\Gamma(k+a+1+n)}{\Gamma(a+1)\Gamma(k+b+1+n)}$$

$$\approx \frac{\Gamma(b+1)}{\Gamma(a+1)} \frac{e^{-c} \cdot c^k}{k!}$$

$$\approx \frac{e^{-c} \cdot c^k}{k!}, \tag{2.11}$$

which is a Poisson distribution.

We can further derive the probability of traffic density as

$$f_X = \frac{r}{r+f} f_{X_\phi} + \frac{f}{r+f} f_{X_b},$$

where X_ϕ is a random variable that follows a Poisson with parameter (λ/μ), and X_b is a random variable that follows a Poisson with parameter (λ'/μ'). Proof is postponed to the Appendix. With a and b being small, the weight parameter of the mixture tends to $\frac{r}{r+f}$ for X_ϕ and $\frac{f}{r+f}$ for X_b. This is expected, because when incident and clearance rates are low, each car will likely spend the whole time in the same state of the queue.

Proposition 1. *The probability mass function of traffic density is the mixture of two Poisson random variables with rates* $\frac{\lambda}{\mu}$ *and* $\frac{\lambda'}{\mu'}$ *when the incident and clearance rate are considerably lower than* μ *and* μ'*. The mixing weight for the Poisson random variable with rate* $\frac{\lambda}{\mu}$ *is given by* $\frac{r}{r+f}$*.*

We will derive the probability mass function of the random number of cars on a segment from equations 2.1, 2.4, 2.5, 2.9, 2.10, and 2.11.

Proof for Proposition 1.

$$
\begin{aligned}
P\{X=k\} =& P\{X_\phi + Y = k\} = \sum_{q=0}^{k} P\{X_\phi = k-q\} \cdot P\{Y=q\} \\
=& P\{X_\phi = k\} \cdot P\{Y=0\} \\
&+ \sum_{q=1}^{k} P\{X_\phi = k-q\} \cdot (p \cdot P\{Y_1 = q\} + (1-p) \cdot P\{Y_2 = q\}) \\
=& \frac{e^{-\lambda/\mu} \cdot (\lambda/\mu)^k}{k!} \cdot \left[p \cdot \left(1 + \frac{\Gamma(b)}{\Gamma(a)} (e^{-c} - 1) \right) + (1-p) e^{-c} \right] \\
&+ \sum_{q=1}^{k} \frac{e^{-\lambda/\mu} \cdot (\lambda/\mu)^{k-q}}{(k-q)!} \left[p \cdot \left(\frac{\Gamma(b)}{\Gamma(a)} \frac{(e^{-c} c^q)}{q!} \right) + (1-p) \frac{(e^{-c} c^q)}{q!} \right] \\
=& \frac{e^{-\lambda/\mu} \cdot (\lambda/\mu)^k}{k!} \cdot \left[e^{-c} \left(1 + p \left(\frac{\Gamma(b)}{\Gamma(a)} - 1 \right) \right) + p \left(1 - \frac{\Gamma(b)}{\Gamma(a)} \right) \right] \\
&+ \left(1 + p \left(\frac{\Gamma(b)}{\Gamma(a)} - 1 \right) \right) \sum_{q=1}^{k} \frac{e^{-\lambda/\mu} \cdot (\lambda/\mu)^{k-q}}{(k-q)!} \cdot \frac{e^{-c} c^q}{q!} \\
=& \frac{e^{-\lambda/\mu} \cdot (\lambda/\mu)^k}{k!} \cdot \left[e^{-c} \left(1 + p \left(\frac{\Gamma(b)}{\Gamma(a)} - 1 \right) \right) + p \left(1 - \frac{\Gamma(b)}{\Gamma(a)} \right) \right] \\
&+ \left(1 + p \left(\frac{\Gamma(b)}{\Gamma(a)} - 1 \right) \right) \sum_{q=1}^{k} \frac{e^{-\lambda/\mu} \cdot (\lambda/\mu)^{k-q}}{(k-q)!} \cdot \frac{e^{-c} c^q}{q!}.
\end{aligned}
$$

Substituting $\left(1 - p\left(1 - \frac{\Gamma(b)}{\Gamma(a)}\right)\right) = m$, we have

$$P\{X = k\} = \frac{e^{-\lambda/\mu} \cdot (\lambda/\mu)^k}{k!} \left[e^{-c}m + (1-m)\right] + m \cdot \frac{e^{(\lambda/\mu+c)}[(\lambda/\mu + c)^k - (\lambda/\mu)^k]}{k!}$$

$$= (1-m) \cdot \frac{e^{-\lambda/\mu} \cdot (\lambda/\mu)^k}{k!} + m\left(\frac{e^{(\lambda/\mu+c)}[(\lambda/\mu + c)^k - (\lambda/\mu)^k]}{k!} + \frac{e^{-\lambda/\mu+c} \cdot (\lambda/\mu)^k}{k!}\right)$$

$$= (1-m) \cdot \frac{e^{-\lambda/\mu} \cdot (\lambda/\mu)^k}{k!} + m\left(\frac{e^{(\lambda/\mu+c)}(\lambda/\mu + c)^k}{k!}\right).$$

Therefore

$$P\{X = k\} = p\left(1 - \frac{\Gamma(b)}{\Gamma(a)}\right) X_g + \left(1 - p\left(1 - \frac{\Gamma(b)}{\Gamma(a)}\right)\right) X_b.$$

Now we will show that $p\left(1 - \frac{\Gamma(b)}{\Gamma(a)}\right)$ is approximately equal to $\frac{r}{r+f}$ for small a and b.

Claim: $p\left(1 - \frac{\Gamma(b)}{\Gamma(a)}\right) \approx \frac{r}{r+f}$ for small a and b.

For $x > 0$,

$$\Gamma(x) = \frac{\Gamma(x+1)}{x}.$$

The above relation for $x = a = \frac{f}{\mu}$ and $x = b = \frac{f}{\mu} + \frac{r}{\mu'}$, implies

$$\frac{\Gamma(b)}{\Gamma(a)} = \frac{\frac{f}{\mu}}{\frac{f}{\mu} + \frac{r}{\mu'}} \cdot \frac{\Gamma(b+1)}{\Gamma(a+1)}.$$

Using the approximation (2.8) for $z = 1$, since $a \ll 1$, $b \ll 1$, and $b - a \ll 1$, one can deduce that $\frac{\Gamma(b+1)}{\Gamma(a+1)}$ is approximately equal to 1. Thus,

$$\frac{\Gamma(b)}{\Gamma(a)} = \frac{\frac{f}{\mu}}{\frac{f}{\mu} + \frac{r}{\mu'}}.$$

Because $p = \frac{r+f\frac{\mu'}{\mu}}{r+f}$, the result follows

$$p\left(1-\frac{\Gamma(b)}{\Gamma(a)}\right) = \left(\frac{r+f\frac{\mu'}{\mu}}{r+f}\right)\left(1-\frac{\frac{f}{\mu}}{\frac{f}{\mu}+\frac{r}{\mu'}}\right)$$
$$= \left(\frac{r\mu+f\mu'}{\mu(r+f)}\right)\left(\frac{\frac{r}{\mu'}}{\frac{f}{\mu}+\frac{r}{\mu'}}\right) = \frac{r}{r+f}.$$

As a result, the probability mass function of density tends to the mixture $P\{X=k\} = \frac{r}{r+f}X_g + \frac{f}{r+f}X_b$, where X_g follow a Poisson with parameter (λ/μ) and X_b follow a Poisson with parameter (λ'/μ'). □

The error of the approximation (2.8) can be derived similarly to the previous proof:

$$\begin{aligned} Error &= p\left(1-\frac{\Gamma(b)}{\Gamma(a)}\right) - \frac{r}{r+f} \\ &= \left(\frac{r\mu+f\mu'}{\mu(r+f)}\right)\left(1-\frac{\frac{f}{\mu}}{\frac{f}{\mu}+\frac{r}{\mu'}}\cdot\frac{\Gamma(b+1)}{\Gamma(a+1)}\right) - \frac{r}{r+f} \\ &= \left(\frac{r\mu+f\mu'}{\mu(r+f)}\right) - \left(\frac{f\mu'}{\mu(r+f)}\cdot\frac{\Gamma(b+1)}{\Gamma(a+1)}\right) - \frac{\mu r}{\mu(r+f)} \\ &= \frac{f}{r+f}\cdot\frac{\mu'}{\mu}\cdot\left(1-\frac{\Gamma(b+1)}{\Gamma(a+1)}\right). \end{aligned}$$

Again, as a and b become smaller, the error converges to zero. Figure 2.10 shows the error for different ratios between f and μ and r and μ'. We can see that when they are around 1% of the μ and μ', the error is on the order of 10^{-3} and it decreases to lower than 10^{-5} when the ratios drop to 0.01%.

Although one may argue this result could follow from intuition, this proof formally demonstrates the result to be true for segments where the incident and clearance inter-times are much longer than the average travel time to cross the segment. The proof also shows that a more general formulation (equation 2.7) must be used for segments that do not meet these criteria. Note that this suggests the probability mass function for the density depends on the clearance and incident rate, thus supporting the impor-

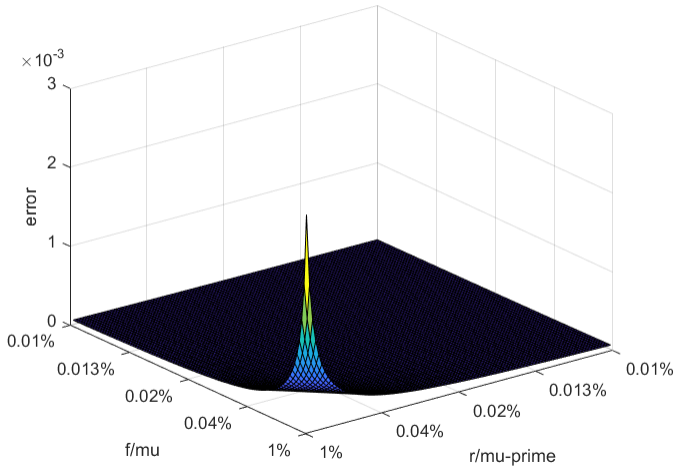


Figure 2.10: Error between the algebraically found weight and $\frac{r}{r+f}$ for different ratios of f and μ, and r and μ'.

tance of considering non-recurrent incidents in the model. This closed-form solution also allows traffic engineers to compute higher moments since they can be derived by weighing the moments from a Poisson distribution. The moment generating function (MGF) of a Poisson random variable Z with parameter ϕ is $M_Z(t) = e^{\phi(e^t-1)}$. The n^{th} factorial moment of the distribution can be computed by taking the n^{th} derivative of the MGF, then setting $t = 0$ (Ross, 1996). The central moments of this Poisson random variable Z are $E[Z] = \phi$, $V[Z] = \phi$, Skewness$[Z] = \sqrt{\phi^{-1}}$, and Kurt$[Z] = \phi^{-1}$. Since the traffic density random variable X is approximated as a mixture of two independent Poisson random variables, we can immediately write its central moments as described in Table 2.1.

$E[X] = \frac{r}{r+f} \cdot \frac{\lambda}{\mu} + \frac{f}{r+f} \cdot \frac{\lambda'}{\mu'}$	$V[X] = \frac{r}{r+f} \cdot \frac{\lambda}{\mu} + \frac{f}{r+f} \cdot \frac{\lambda'}{\mu'}$
Skewness$[X] = \frac{r}{r+f} \cdot \sqrt{\frac{\mu}{\lambda}} + \frac{f}{r+f} \cdot \sqrt{\frac{\mu'}{\lambda'}}$	Kurt$[X] = \frac{r}{r+f} \cdot \frac{\mu}{\lambda} + \frac{f}{r+f} \cdot \frac{\mu'}{\lambda'}$

Table 2.1: Central Moments for the traffic density distribution under non-peak hours.

Peak hours

When the system operates in peak hours, the higher arrival rate of cars is also a cause of congestion. In this case, the assumption that the travel time distribution is independent of the number of vehicles is not as consistent with real-life scenarios. Therefore, we need to account for another source of travel time deterioration: the current number of cars in the system.

For this case, we combine our deterioration model with the congested traffic model M/G/C/C discussed in Jain and Smith (1997). Let us assume $\mathbf{a} = [\mathrm{a}_1, \mathrm{a}_2, \mathrm{a}_3, \ldots]$ is a vector where component a_n represents the deterioration coefficient caused by congestion when n cars are present. $\mathrm{a}_1 = 1$ because a single car can travel at free-flow speed. Moreover, $0 \leq \mathrm{a}_n \leq 1$, and a_n is monotonically decreasing as n grows, meaning that cars arriving can only maintain or worsen system conditions. As an initial suggestion, Jain and Smith (1997) propose function a_n to be linearly or exponentially decreasing in n. In this chapter, we assume a linearly decreasing function for a_n, which avoids an overly fast deterioration caused by the additional cars.

Furthermore, since the probability of breakdown is not negligible in this scenario, the system capacity is truncated at a certain point C, i.e., we assume that no more cars arrive after C cars are in the system. The presence of C cars in the system represents a complete breakdown, where there is no space left for another car to arrive. The modeling thus follows an M/M/C/C queue in a random environment, represented in Figure 2.11. Note the inclusion of the parameters representing the extra congestion due to the accumulation of cars, and the limited capacity C of the system.

The solution for such a queue is described below, and more details can be found in Baykal-Gürsoy et al. (2009a).

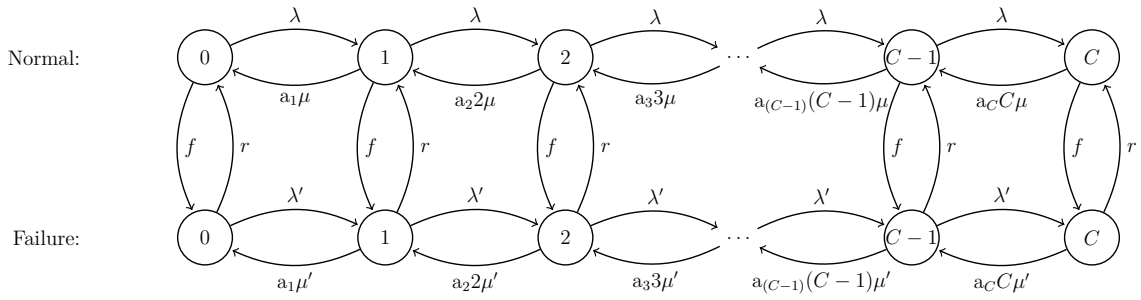


Figure 2.11: State Transition Diagram for a Markov-modulated $M/M/C/C$ Queue.

The balance equations are given by:

$$p_{iN}\left(\lambda + f + i\mu \mathrm{a}_i\right) = r p_{iF} + \left((i+1)\,\mu \mathrm{a}_{i+1}\right) p_{i+1,N} + \lambda p_{i-1,N}, \quad \text{for } i = 1, 2, \ldots, C-1,$$

$$p_{iF}\left(\lambda' + r + i\mu' \mathrm{a}_i\right) = f p_{iN} + \left((i+1)\,\mu' \mathrm{a}_{i+1}\right) p_{i+1,F} + \lambda' p_{i-1,F}, \quad \text{for } i = 1, 2, \ldots, C-1.$$

and the boundary equations are,

$$p_{0N}(\lambda + f) = r p_{0F} + \mu \mathrm{a}_1 p_{1N},$$

$$p_{0F}(\lambda' + r) = f p_{1N} + \mu' \mathrm{a}_1 p_{1F},$$

$$p_{CN}\left(C\mu \mathrm{a}_C + f\right) = r p_{CF} + \lambda p_{C-1,N},$$

$$p_{CF}\left(C\mu' \mathrm{a}_C + r\right) = f p_{CN} + \lambda' p_{C-1,F},$$

and the normalization equation is $\sum_{i=0}^{C}(p_{iN} + p_{iF}) = 1$.

Given a fixed value of C, determined as the maximum capacity of the road (the number of cars for which the roadway is in a breakdown), we can solve for all p_{iN} and p_{iF}, as long as the values of the vector $\mathbf{a}$ are available.

Note that, unlike the non-peak hours' framework, the probability mass function for density during peak-hours does not have an intuitive closed-form. We provide a straightforward and efficient framework to compute it numerically via a relatively small linear system of equations.

2.3 Validation

This section shows how the proposed model compares to the dataset. Here, we use a dataset to compute the parameters for the proposed model and compare it against the lognormal and Weibull distributions. These models also provide a reasonable fit to the data, but they require parameters that can only be found using observed density data. By comparison, the parameters for the proposed model are often known by traffic engineers or readily available with little data collection, allowing an effortless implementation of the model for different road segments.

2.3.1 Discussion and Findings

Results are obtained from the comparisons between the curve generated from the model and the data for each sensor, as depicted in figures 2.12, 2.13, and 2.14. The curve represents the cumulative distributions generated through the analytical model and from the density data computed from the occupancy data set. The main goal is to determine whether the curve from the analytical model is a good fit for the histogram and how it compares to other distributions used in literature. However, the dataset used is censored due to limitations in the sensors' precision. As a result, some points in the density distribution are misrepresented in the dataset. This limitation causes distribution tests, such as the Kolmogorov-Smirnov test, to fail for both our model and other models used in literature. Therefore, we used the confidence interval obtained from the Kolmogorov-Smirnov test (Massey Jr, 1951) along with an added uncertainty level to account for sensors' lack of precision to determine the validity of the model.

As the figures show, the empirical cumulative distribution is within the KS-test upper and lower bounds, and therefore matches the model's cumulative distribution.

Lastly, we compare the analytical model to other distributions used in literature employing the Akaike Information Criterion.

Non-peak hours

The best approach when using this analytical model is to separate specific periods in which these parameters behave according to the assumptions. The model assumes that time to an incident and clearance times are exponentially distributed, as well as the interarrival times for normal and adverse conditions, and travel times during normal and adverse conditions. As an example, January and February, months that are known to be the snowiest, can have a model, and the remaining months separated in two groups, dry season and rainy season, can have two other models. Such definitions depend on local factors and should be determined individually.

Non-peak hours are defined as the times between 10 a.m. and 1 p.m., Tuesday to Thursday.

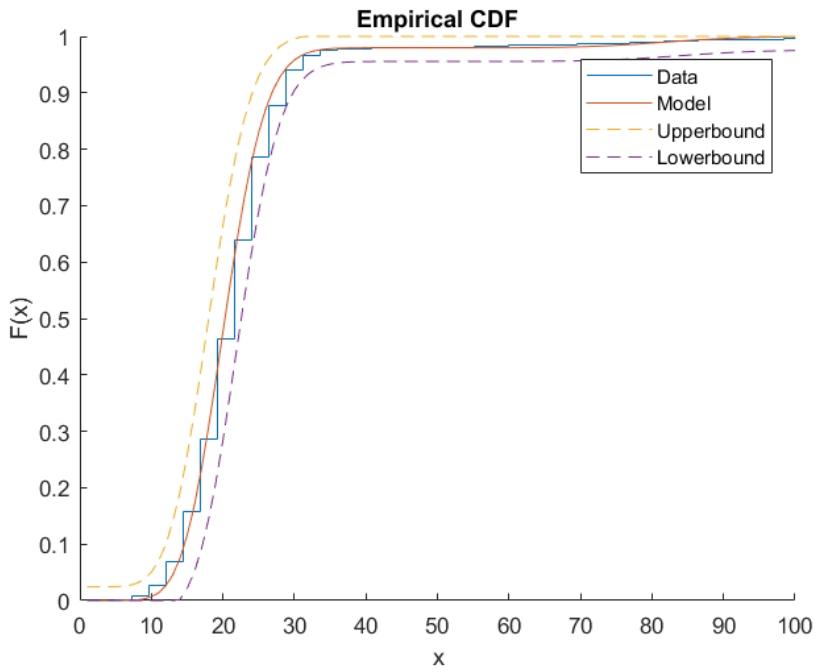


Figure 2.12: Model's analytical and data's empirical CDFs (non-peak hours/winter).

Figures 2.12 and 2.13 depict one of the 36 sensors comparison for non-peak hours

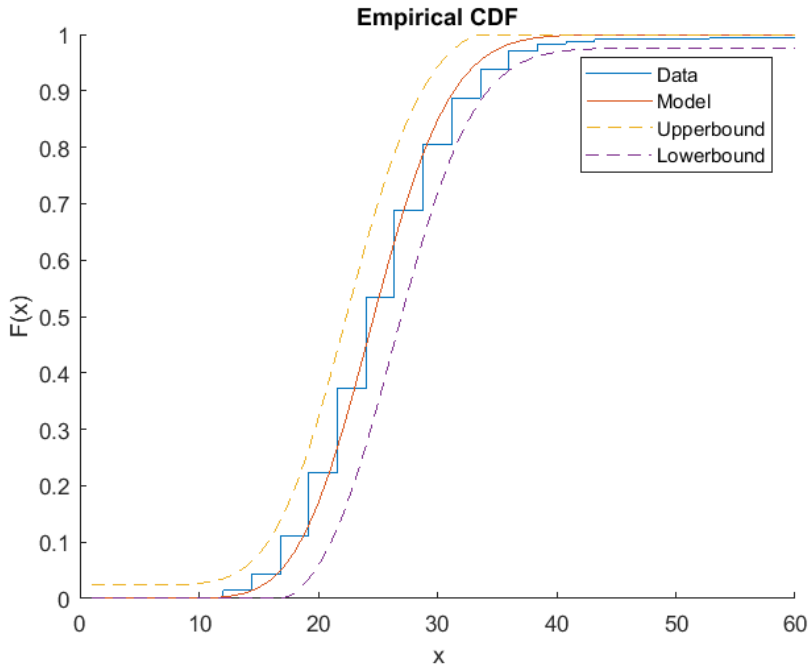


Figure 2.13: Model's analytical and data's empirical CDFs (non-peak hours/summer).

during winter and summer months. The upper and lower bounds represent the 5% significance level of the model distribution (Massey Jr, 1951) when accounting for data uncertainty due to the censored dataset. Since the empirical distributions reside within the bounds, the model is valid for both sets of months. Non-peak hours data were gathered from 11 a.m. to 1 p.m., Tuesday to Thursdays. The main difference between the two figures is the slightly thicker tail in figure 2.12, mainly caused by snowstorms that tend to have a more lasting impact than the heavy rains and accidents in the summer.

Table 2.2 indicates that the model closely matches the goodness of fit of lognormal and Weibull, two commonly used distributions (the complete AIC list for each sensor is given in Table 2.9, Appendix 2). Hence, our model is valid for roads in which incident and clearance rates (f and r) are much lower than μ and μ'. The values also suggest the validity of the model for both winter and summer months, although

	Model Winter	LNormal Winter	WBull Winter	Model Summer	LNormal Summer	WBull Summer
Mean	4931	4941	5042	1362	1173	1370
Best AIC (out of 33 sensors)	30	2	1	13	19	8

Table 2.2: Adapted AIC comparison between model and commonly used distributions during non-peak hours.

parameters may need adjustments for the analyzed period.

These results endorse the validity of this queuing theory approach, suggesting the simplified analytical model (a mixture of two Poisson distributions) is robust enough. It matches the performance of or performs better than other distributions used in literature. Furthermore, the results suggest that the assumptions made by the model are valid.

Moreover, a significant advantage of this analytical model is that it does not require detailed data, but rather aggregate parameters that can be easily estimated. Furthermore, minor errors in estimation are not overly harmful to the performance of the model. The derivatives of expected value and distribution of density with respect to the parameters $\lambda, \lambda', \mu, \mu', f, r$ allow for sensitivity analysis. Except for λ and λ', this model is robust to errors in estimation. For similar values to those seen on all sensors, a 10% error in μ or μ' results in a 1% error in expected density. The model is sensitive to changes in λ and λ', only when $\lambda \approx \lambda'$. This model is, therefore, appropriate as an initial gauge on how traffic density will behave in new roadways, in roadways for which data are scarce, and in roadways that may have endured some change in behavior.

Peak hours

Similar to non-peak hours, the results for peak hours also proved to be very robust. Peak hours are defined as the times between 8 a.m. and 10 a.m. for Southeast direction, and between 6 p.m. and 8 p.m. for Northwest direction, Tuesday to Thursday. These periods are chosen to match the commute from residential areas to work locations in the morning and back in the evening.

The values of **a** are calculated from a function of the number of cars in the system and the road segment capacity. They determine the deterioration in service level (travel time) caused by the presence of more cars, thus causing drivers to drive more carefully and slowly. Jain and Smith (1997) suggest that such a function could be linear and follow $a_n = \frac{C+1-n}{C}$, where C is the capacity of the segment. In our dataset, segments are half-mile long and contain three lanes. Thus, we can obtain C as

$$C = \frac{\text{length segment} \cdot \text{number of lanes}}{\text{average length of a car}} = \frac{0.5 \cdot 3}{22/5280} = 360,$$

giving $a_n = \frac{361-n}{360}$.

An alternative approach is to generate **a** via travel time computed from the speed data. By comparing average speed data for each density point in each sensor, we can directly generate **a**. For each density point, we can observe every car's speed, thus creating an array of speed data points. For density points with fewer than three speed data points, we assume speed remained the same as the previous density data point to prevent outliers.

Although the two approaches yield good fits, the one that generates **a** from a linear function covers a little more of the variance of the data. The average r^2 for the linearly generated **a** is 0.776, and for the **a** coming from speed data is 0.762. Hence, we choose to present the results obtained with the linear function for **a**.

Figure 2.14 presents the empirical CDF generated from the data along with the model's CDF. The upper and lower bounds represent the 5% confidence interval of the model distribution (Massey Jr, 1951) when accounting for data uncertainty due

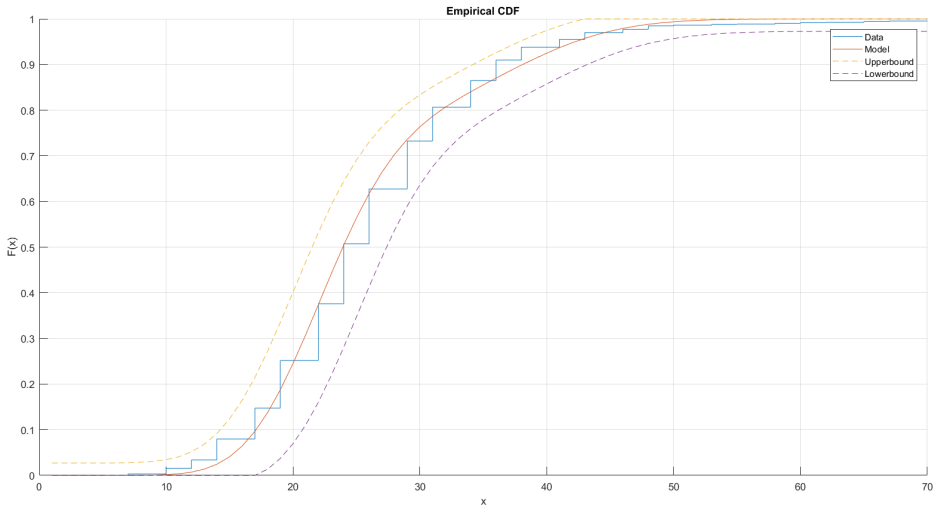


Figure 2.14: Model's analytical and and data's empirical CDFs (peak hours/summer).

to the gaps in the dataset. Since the empirical distribution resides within the bounds, the model is valid.

	Model Winter	LNormal Winter	WBull Winter	Model Summer	LNormal Summer	WBull Summer
Mean	5236	5377	5433	804	786	812
Best AIC	33	3	1	5	23	6
(out of 36 sensors)	(1 tie)	(1 tie)		(1 tie)	(1 tie)	

Table 2.3: Adapted AIC comparison between model and commonly used distributions during peak hours.

Table 2.3 compares the goodness of fit for our model with other commonly used distributions for traffic density (the complete AIC list for each sensor is described in Table 2.10, Appendix 2). Adapted AICs (following the same algorithm described for non-peak hours) are calculated using the maximum log-likelihood function from the data for the model, lognormal, and Weibull distributions. The MLEs calculated for the model are obtained using the Nelder-Mead method with five iterations (Nelder

and Mead, 1965), which implies that certain improvements can still be achieved in the model's AIC if more iterations are performed. Results indicate that the analytical model performs slightly better than the other two fits during winter, and is consistent with the other two fits during summer months. They reiterate the importance of our model because it does not depend on large sets of data that can be expensive or infeasible to gather, differently from the other distributions.

2.4 Applications

2.4.1 Sensitivity on the deterioration level α

We consider the effect of the ratio between μ' and μ, which we call α. The lower the α is, the worse the system becomes when it deteriorates. A value of $\alpha = 0.80$ means service frequency drops 20% when the system deteriorates due to an incident. This parameter is one of the most difficult to change – it mainly reflects the overall infrastructure's resilience to incidents. A change in this parameter could represent the aging of a system or a significant change in utilization. The model presented in this chapter can help traffic engineers understand the effect a variation in the deterioration level would have on congestion.

Tables 2.4 and 2.5 contain several points on the tail distribution for density during peak and non-peak hours. For non-peak hours, we set arrival rate is 15 cars per minute, expected travel time is 1 minute under normal conditions, expected time to incident is 41 hours and expected clearance time is 28 hours. For peak hours, we set the arrival rate is 25 cars per minute, expected travel time is 1 minute under normal conditions (before accounting for deterioration caused by congestion), and expected time to incident is 41 hours and expected clearance time is 28 hours. The expected travel time under adverse conditions vary with α as $\frac{1}{\alpha\mu}$, i.e., $\frac{1}{\alpha}$E[travel time under

normal conditions].

From the tables, we expect to see more than 11 cars with more than 30% probability when $\alpha = 0.8$ during non-peak hours and more than 18 with the same probability during peak hours. It is interesting to see that, as α decreases, the tail becomes thicker. The tables also show that the left side of the tail distribution is very similar for both non-peak and peak hours (as shown for $P\{X > x\} \geq 70\%$). In this part of the distribution, α has little impact. However, on the right side of the distribution (as shown for $P\{X > x\} \leq 30\%$) for peak hours, the addition of cars caused by a non-recurrent incident causes traffic to become worse when α is small, creating a cascading effect in traffic congestion.

$P\{X > x\}$	99%	70%	30%	1%	E[number of cars]
α=0.8	4	8	11	17	8.63
α=0.6	4	8	11	20	8.99
α=0.4	4	8	12	28	9.71
α=0.2	4	8	12	52	11.86

Table 2.4: Tail Probability for different levels of deterioration during non-peak hours.

$P\{X > x\}$	99%	70%	30%	1%	E[number of cars]
α=0.8	8	14	18	27	15.01
α=0.6	8	14	18	33	15.69
α=0.4	8	14	19	49	17.14
α=0.2	8	14	19	112	22.81

Table 2.5: Tail Probability for different levels of deterioration during peak hours.

More graphically, on Figures 2.15 and 2.16, we can see how the probability mass function behaves as α decreases for both non-peak and peak hours. Parameters used

are:

1. arrival rate of 15 cars per minute for non-peak hours and of 15 cars for peak hour
2. expected travel time of 1 minute under normal conditions (before accounting for deterioration caused by congestion during peak hours)
3. expected time to incident of 41 hours
4. expected clearance time of 28 hours

At first, the tail starts to thicken. However, after some threshold, the system becomes bimodal, with an evident separation between the normal condition and the adverse condition distributions. This bimodality is more prominent during peak hours than non-peak hours.

The probability of traffic breakdown directly follows from the resulting distribution. Traffic engineers may determine the number of cars that causes a traffic breakdown on this particular road, and then have the probability of breakdown determined by the model. This model is flexible enough to provide decision-makers with such a measure for different definitions of traffic breakdown since the literature has yet to agree on a specific definition.

2.4.2 Example on the usage of the model for planning

In this section, we provide an example of how traffic engineers can apply this new model in their decision-making process. Suppose traffic engineers plan to build a highway with different 0.5-mile sections. They consider various investments, which alter multiple parameters in the model. In particular, they will look at the number of lanes to be built and the budget for nearby service vehicles and first responders. They

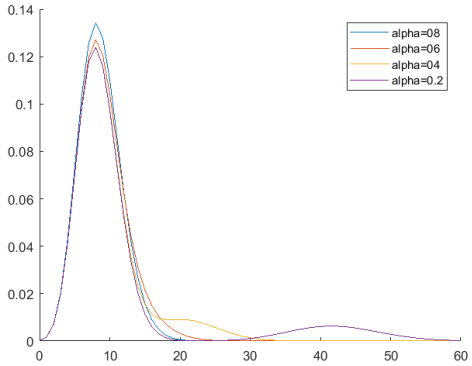


Figure 2.15: PMF's for non-peak hours

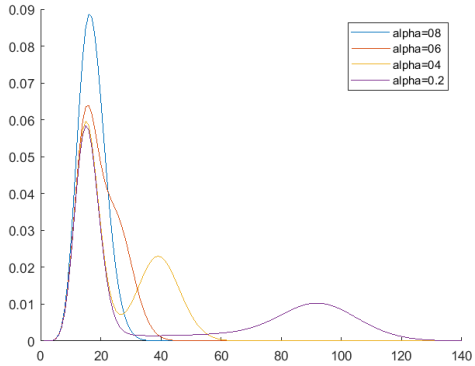


Figure 2.16: PMF's for peak hours

assume all 0.5-mile sections within a region behave similarly. Usually, such a study would entail collecting data for similar existing highways and simulating changes to them. However, the proposed model enables these analyses without such data or simulations.

In order to understand how the highway would behave throughout different periods, they separate the analysis into three period groups: 1. low usage, which includes late nights, early mornings, weekends, and holidays; 2. medium usage, which includes weekdays, but during times that avoid the main commute rush; 3. high usage, which includes weekdays peak-hours. By comparing the region with other similar locations, the traffic engineers were able to estimate the following parameters:

From the results, we can assume that the system is unlikely to reach full capacity under Medium Usage, and we, therefore, use the non-peak hours' method to calculate the density distribution. For the High Usage time frame, we use the peak-hours model instead.

We first explore the effect of lane count, assuming an average clearance time of 30 minutes. Increasing the lane count is equivalent to increasing the capacity of the system. Table 2.7 shows the probabilities of seeing more than 10% of the road occupied and the probability of having less than 90% of the roadway occupied under

	Low Usage	Medium Usage	High Usage
λ	160	650	1100
λ'	150	630	790
μ	60	21	18
μ'	40	14	12
f	0.00002	0.005	0.02
r	2	2	2

Table 2.6: Parameters estimated by the traffic engineers for the location in which the new highway will be built.

Medium and High Usage given the number (1, 2 or 3) of lanes. C represents the capacity of the system, which is a function of the number of lanes. We immediately see the trade-off between under-utilization in medium usage periods and over-utilization in high-usage periods. The correct selection of lanes depends on the constraints faced by the decision-maker. For the sake of discussion, we select two lanes, which will be over-utilized 25% of the time during peak hours.

	Medium Usage		High Usage	
Lanes	$P\{X > 0.1 \cdot C\}$	$P\{X < 0.9 \cdot C\}$	$P\{X > 0.1 \cdot C\}$	$P\{X < 0.9 \cdot C\}$
1	0.9999	1	1	0.2306
2	0.8799	1	1	0.7583
3	0.1608	1	1	0.9998

Table 2.7: Probability of under-utilization and over-utilization for different lane counts.

We now consider changes to the clearance time, which can be affected by other investments, such as the response time of first responders and service vehicles. Again

we see a trade-off that must be reconciled by the decision-maker – over-utilization can be reduced, but only by making significant reductions in clearance time. Another consideration is that increasing the clearance rate (i.e., decreasing expected clearance time) has a stronger impact on lowering over-utilization than under-utilization.

		Med. Load		High Load	
E[Clearing Time]	r	$P\{X > 24\}$	$P\{X < 216\}$	$P\{> 24\}$	$P\{< 216\}$
8.5 min	7	0.8796	1	1	0.9092
15 min	4	0.8797	1	1	0.8587
30 min	2	0.8799	1	1	0.7583
5 h	0.2	0.8825	1	1	0.2146
50 h	0.02	0.9036	1	1	0.0002

Table 2.8: Probability of under-utilization and over-utilization for different clearance times, with 2 lanes.

APPENDIX – Full list of AICs for each sensor

Non-peak hours

Sensor SE then WN	Model Winter	Lognormal Winter	Weibull Winter	Model Summer	Lognormal Summer	Weibull Summer
1	6057	5763	5607	783	784	804
2	4775	4795	5058	747	749	814
3	6427	5867	5732	817	814	840
4	4965	4985	5453	767	791	882
5	4555	4794	5307	738	769	862

Continued on next page

Sensor SE then WN	Model Winter	Lognormal Winter	Weibull Winter	Model Summer	Lognormal Summer	Weibull Summer
6	4740	4849	5250	789	787	800
7	4948	4970	5405	776	779	778
8	4752	4819	4792	1303	1233	1344
9	4636	4708	4664	787	792	786
10	4865	4866	4938	771	776	774
11	4956	5007	4995	822	830	820
12	4689	4754	4721	820	832	831
13	4078	4157	4263	700	700	705
14	4762	4814	4874	800	803	802
15	4762	4838	4791	816	816	818
16	4670	4740	4706	802	803	805
17	5009	5070	5021	804	804	804
18	5983	5337	5361	830	826	831
1	4822	4862	4869	833	829	810
2	5334	5319	5507	891	893	882
3	4724	4818	5251	806	807	802
4	4724	4818	5251	806	807	802
5	4867	4889	4899	988	887	1070
6	5003	5026	5018	1034	907	1073
7	4795	4832	4825	993	892	1102
8	3990	4125	4007	1210	1062	1364
9	4740	4802	4770	1058	929	957
10	4904	4942	4937	1328	1111	1147

Continued on next page

Sensor SE then WN	Model Winter	Lognormal Winter	Weibull Winter	Model Summer	Lognormal Summer	Weibull Summer
11	5106	5131	5127	1633	1271	1280
12	5319	5317	5287	1727	1632	1589
13	5642	5567	5549	1717	1588	1623
14	4900	4959	4914	1872	1972	2099
15	4722	4757	4947	4854	3187	N/A
16	4480	4620	4884	6314	4013	N/A
17	4890	4956	5196	3606	2302	9400
18	4941	5014	5328	2699	2437	5473

Table 2.9: Full list of the adapted AIC comparison between model and commonly used distributions during non-peak hours.

Peak hours

Sensor SE then WN	Model Winter	Lognormal Winter	Weibull Winter	Model Summer	Lognormal Summer	Weibull Summer
1	5505	6026	5767	774	772	780
2	4938	5132	5149	746	743	746
3	5663	6055	5775	800	796	798
4	5112	5302	5358	775	768	775
5	4777	5182	5341	736	728	734
6	5020	5147	5366	762	747	746
7	5255	5357	5626	771	776	767
8	N/A	5830	5990	874	831	883

Continued on next page

Sensor SE then WN	Model Winter	Lognormal Winter	Weibull Winter	Model Summer	Lognormal Summer	Weibull Summer
9	5000	5257	4999	790	774	795
10	5230	5426	5262	777	786	831
11	5245	5494	5353	867	812	868
12	5163	5351	5247	829	809	854
13	4577	4982	5141	789	687	721
14	5343	5468	5626	822	798	807
15	5385	5678	5922	825	798	863
16	5643	5683	5943	747	797	867
17	6003	6050	6263	886	853	886
18	5955	6062	6262	926	846	907
1	5391	5395	5510	823	809	852
2	5765	5770	5847	882	876	920
3	5221	5245	5258	797	788	796
4	5221	5245	5258	797	788	796
5	5286	5310	5323	784	778	780
6	5345	5375	5374	806	801	795
7	5212	5241	5249	827	761	767
8	4505	4533	4507	747	747	753
9	4929	4961	4952	781	775	772
10	5169	5195	5198	792	787	783
11	5138	5152	5199	886	827	817
12	5351	5506	5476	809	772	788
13	5554	5734	5689	828	809	818

Continued on next page

Sensor SE then WN	Model Winter	Lognormal Winter	Weibull Winter	Model Summer	Lognormal Summer	Weibull Summer
14	5039	5121	5180	779	758	806
15	4851	4864	4945	756	763	828
16	4814	4818	4972	738	746	821
17	5236	5228	5482	816	787	837
18	5402	5402	5784	806	803	871

Table 2.10: Full list of the adapted AIC comparison between model and commonly used distributions during peak hours.

Chapter 3

Traffic Density in Non-homogenous Sections

This section focuses on expanding the results of chapter 2 to long segments in which parameters may not remain homogeneous. We initially divide this long segment into smaller sections in which the parameters remain homogeneous. Then, we study techniques to combine the information from each short segment to obtain the distribution for the long stretch. We also validate the results against the Wisconsin data described in chapter 2.

We concentrate on results for non-peak hours because of mathematical convenience. Winter data are used for validation.

3.1 Analytical Model

The model proposed in this chapter follows the approximated distribution derived in chapter 2. Again, the approach models traffic as the Markov-modulated queue depicted in Figure 2.9.

The set up of the problem proposed assumes parameters to be homogenous across

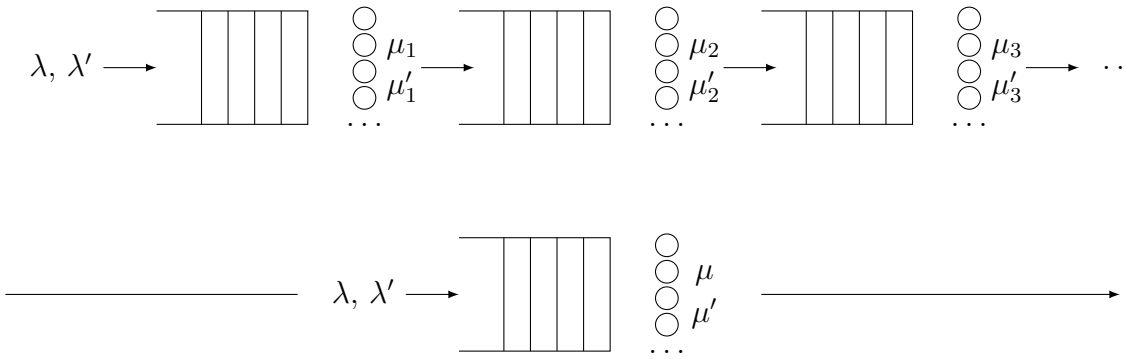


Figure 3.1: Representation of the tandem-queue and its single-queue alternative.

space. While this may be true for sufficiently small segments, curves, ramps, and other road features may change the rate parameters. In this chapter, we propose a tandem-queue approach to account for spatial changes in rate parameters. We maintain the assumptions that failure rate and repair rate being considerably smaller than the service rates. For comparison purposes, we consider two alternatives, depicted in Figure 3.1. The first breaks the segment into 0.5-mile subsections. It considers subsections as queue systems that form a tandem-network when combined. The latter considers the whole segment as a single queue, as discussed in chapter 2. This comparison will determine whether a tandem-queue approach is necessary.

This study proposes that a sequence of the Markov-modulated queue depicted in Figure 2.9 can be approximated as a tandem-queue system in which the distribution of customers in the system has the product-form:

$$\lim_{t\to\infty} P\left\{X_i(t) = x_i, 1 \leq i \leq N\right\} = \prod_{i=1}^{N} P\{X_i = i\}$$

For this, we argue the following needed assumptions (Kulkarni, 2016):

1. It has N service stations (nodes) — where N represents the number of segments considered;

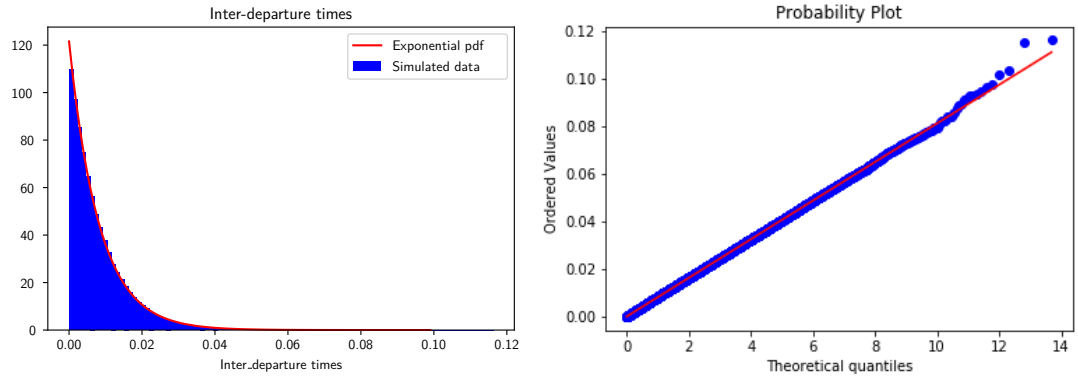
2. The departure process follows the same distribution of the arrival process — as discussed in this section;

3. There is infinite waiting room in each node — cars can remain stopped for as long as needed;

4. There are no external arrivals other than in the first node;

5. There are no middle departures from the system — cars must join the next segment after completing one.

Assumptions (3), (4), and (5) are met from the way the problem is set, as depicted in Figure 2.9. Assumption (1) is met because we assume a finitely long segment. Assumption (2) is not met because service times are not independent and identically distributed. However, we make this assumption for mathematical convenience and because this assumption closely matches what we see in simulation. Figure 3.2 depicts the inter-departure times, and the corresponding exponential Q-Q plot from a 10-thousand hour simulation. Notice that we assume the departure interarrival times to follow a mixture of exponentials, a more general distribution than the one pictured in Figure 3.2.



(a) (b)

Figure 3.2: Simulated queue inter-departure times.

The following proof for assumption (2) is an adaptation of the work developed by Mirasol (1963).

Proposition 2 (Departure Process)**.** *The departure process of a MAP/G/∞ queue follows the same distribution of the arrival process.*

Proof. **Departure process for a MAP/G/∞ queue**

Let us take a look at the queue when starting at time zero with no events, and analyze it at time t and then at $t+T$. By finding solutions in this structure, we can then push t to the limit so that we can see the behavior in steady-state. Mirasol (1963) has used this approach to prove the departure distribution for M/G/∞ queues. We will follow similar steps to prove the departure distribution for a MAP/MSP queue in a random environment.

The process starts at time 0 with no customers. The arrival period is denoted as subscripts and departures as superscripts. A thorough definition of each notation follows.

- Departures:
  - $\Psi^{(t,t+T)}$ = number departing the system in (t,t+T),
  - $\Psi_1^{(t,t+T)}$ = portion of $\Psi(t,t+T)$ that arrived in (0,t),
  - $\Psi_2^{(t,t+T)}$ = portion of $\Psi(t,t+T)$ that arrived in (t,t+T).

Note that $\Psi^{(t,t+T)} = \Psi_1^{(t,t+T)} + \Psi_2^{(t,t+T)}$.

- Survivals:
  - $\gamma(t+T)$ = number surviving in the system at t+T,
  - $\gamma_1(t+T)$ = portion of $\gamma(t,t+T)$ that arrived in (0,t),
  - $\gamma_2(t+T)$ = portion of $\gamma(t,t+T)$ that arrived in (t,t+T).

Note that $\gamma(t+T) = \gamma_1(t+T) + \gamma_2(t+T)$.

- Leaving Probabilities (arrival state i can be):
  - u = Pr{an arrival in (0,t) leaves in (t,t+T)},
  - v = Pr{an arrival in (0,t) leaves in (t+T, ∞)},
  - w = Pr{an arrival in (t,t+T) leaves in (t,t+T)},
  - z = Pr{an arrival in (t,t+T) leaves in (t+T, ∞)},
  - $H(x)$ = Pr{service time $\leq x$},

In a nonhomogeneous Poisson process having n events occuring in (0,t), the instances in time when these events occured is the joint distribution of $U(1), \ldots, U(n)$, the order statistics of n i.i.d. Unif$(0,t)$ random variables (Campbell's theorem).

Therefore:

$$u = \int_0^t \frac{H(T+x) - H(x)}{t} dx \qquad w = \int_0^T \frac{H(x)}{T} dx$$
$$v = \int_0^t \frac{1 - H(T+x)}{t} dx \qquad z = \int_0^T \frac{1 - H(x)}{T} dx$$

Let $F(m,n;t,T) = Pr\{\gamma(t+T) = m, \Psi^{(t+T)} = n\} = \sum_{m_b=0}^{m} \sum_{n_b=0}^{n} Pr\{\gamma_g = m - m_b, \Psi_g^{(t+T)} = n - n_b \mid \gamma_b = m_b, \Psi_b^{(t+T)} = n_b\} \cdot Pr\{\gamma_b = m_b, \Psi_b^{(t+T)} = n_b\}$. We want to consider this joint probability, because it allows us to condition on the number of arrivals, which has a known distribution. Let the number of arrivals in a period x be denoted as $N(x)$. We can condition F on the number of arrivals in $(0,t)$ – say $N(t) = k$, and the number of arrivals in $(t, t+T)$ – say $N(T) = j$. Note that $\Psi_2^{t,t+T}$ is totally dependedent on the j and $\gamma_2(t+T)$,

$\Psi_2^{t,t+T} = j - \gamma_2(t+T)$.

$$F(m,n;t,T) = \sum_{i=0}^{m}\sum_{j=i}^{n+i}\sum_{k=m+n-j}^{\infty} P\{N(t)=k, N(T)=j\}\cdot$$

$$P\{\gamma_2 = i, \gamma_1 = m-1, \Psi_1 = n-(j-i) \mid N(t)=k, N(T)=j\}$$

The limits of the summation are defined from the relationships among the parameters. As i represents the number of survivals at time $t+T$ that arrived between t and $t+T$, it has to be less than the total number of survivals, which implies $i < m$. Also, there are j arrivals in the period $(t, t+T)$. Therefore, the number of arrivals j must be greater than the survivals i, $j > i$. We also know that n departures happened in $(t, t+T)$. Since we know that $\Psi_2^{t,t+T} = j-\gamma_2(t+T) = j-i$ has be less than n, $j < n+i$. Finally, the number of arrivals in $(0,t)$ can be no less than the total survivals at $t+T$ plus total departures minus the the arrivals in $(t, t+T)$, implying $k > m+n-j$.

We can now show that this joint distribution is just the product of the departure

events and the survival events, proving that they are independent.

$$F(m,n;t,T) = \sum_{i=0}^{m}\left[\sum_{j=i}^{n+i}\Big(P\{\gamma_2 = i \mid N(T)=j\}P\{N(T)=j\}\right.$$
$$\left.\sum_{k=m+n-j}^{\infty} p\{\gamma_1 = m-1, \Psi_1 = n-(j-i) \mid N(t)=k\}P\{N(t)=k\}\Big)\right]$$

$$= \sum_{i=0}^{m}\left[\sum_{j=i}^{n+i}\binom{j}{i} z^i w^{j-i}\left(\frac{r}{r+f}e^{-\lambda T}(\lambda T)^j/j! + \frac{f}{r+f}e^{-\lambda' T}(\lambda' T)^j/j!\right)\cdot\right.$$
$$\sum_{k=m+n-j}^{\infty}\frac{k!\; u^{m-i}v^{n-j+i}(1-u-v)^{k-m-n+j}}{(m-i)!(n-j+i)!(k-m-n+j)!}$$
$$\left.\left(\frac{r}{r+f}e^{-\lambda t}(\lambda t)^j/j! + \frac{f}{r+f}e^{-\lambda' t}(\lambda' t)^j/j!\right)\right]$$

$$= \left(\frac{r}{r+f}\right)^2\left\{\frac{e^{-\lambda(zT+vt)}[\lambda(zT+vt)]^m}{m!}\right\}\left\{\frac{e^{-\lambda(wT+ut)}[\lambda(wT+ut)]^n}{n!}\right\} +$$
$$\left(\frac{f}{r+f}\right)^2\left\{\frac{e^{-\lambda'(zT+vt)}[\lambda'(zT+vt)]^m}{m!}\right\}\left\{\frac{e^{-\lambda'(wT+ut)}[\lambda'(wT+ut)]^n}{n!}\right\} +$$
$$\left(\frac{r}{r+f}\right)\left(\frac{f}{r+f}\right)\left\{\frac{e^{-\lambda(zT+vt)}[\lambda(zT+vt)]^m}{m!}\right\}\left\{\frac{e^{-\lambda'(wT+ut)}[\lambda'(wT+ut)]^n}{n!}\right\} +$$
$$\left(\frac{f}{r+f}\right)\left(\frac{r}{r+f}\right)\left\{\frac{e^{-\lambda'(zT+vt)}[\lambda'(zT+vt)]^m}{m!}\right\}\left\{\frac{e^{-\lambda(wT+ut)}[\lambda(wT+ut)]^n}{n!}\right\}$$

$$= \left\{\frac{r}{r+f}\cdot\frac{e^{-\lambda(zT+vt)}[\lambda(zT+vt)]^m}{m!} + \frac{f}{r+f}\cdot\frac{e^{-\lambda'(zT+vt)}[\lambda'(zT+vt)]^m}{m!}\right\}\cdot$$
$$\left\{\frac{r}{r+f}\cdot\frac{e^{-\lambda(wT+ut)}[\lambda(wT+ut)]^n}{n!} + \frac{f}{r+f}\cdot\frac{e^{-\lambda'(wT+ut)}[\lambda'(wT+ut)]^n}{n!}\right\}.$$

Therefore, they are independent. Note that $wT+ut$ can be rewritten as $\int_t^{t+T} H(x)dx$. Also note that $\lim_{t\to\infty}\int_t^{t+T} H(x)dx = T$, as it represents the area of a rectangle with height 1, and width $(t+T)-t = T$. Hence:

$$P\left\{\Psi^{(t,t+T)}=n\right\}=\frac{r}{r+f}\frac{e^{-\lambda(wT+ut)}[\lambda(wT+ut)]^n}{n!}+\frac{f}{r+f}\frac{e^{-\lambda'(wT+ut)}[\lambda'(wT+ut)]^n}{n!}$$

$$\lim_{t\to\infty}P\left\{\Psi^{(t,t+T)}=n\right\}=\frac{r}{r+f}\cdot\frac{e^{-\lambda T}(\lambda T)^n}{n!}+\frac{f}{r+f}\cdot\frac{e^{-\lambda' T}(\lambda' T)^n}{n!}$$

□

As such, we can consider the long stretch of the road consisting of the sequence of N segments to be an Open-Jackson Network Tandem Queue.

$$\lim_{t\to\infty}P\left\{X_i(t)=x_i,1\le i\le k\right\}=\prod_{i=1}^{k}\left(\frac{r_i}{r_i+f_i}\left(\frac{\lambda}{\mu_i}\right)^{x_i}\frac{e^{(\lambda/\mu_i)}}{x_i!}+\frac{f_i}{r_i+f_i}\left(\frac{\lambda'}{\mu_i'}\right)^{x_i}\frac{e^{(\lambda'/\mu_i')}}{x_i!}\right). \tag{3.1}$$

3.1.1 Computing Density from The Product Form

As a consequence of equation 3.1, the total number of customers in the network follows a distribution that is the sum of each segment's distribution separately. Therefore, we can obtain the overall distribution via convolution. Let A be the distribution of the number of cars in the steady-state.

$$A=\lim_{t\to\infty}\sum_{i=1}^{k}X_i(t)$$

$$P\{X_i=x_i\}=\frac{r_i}{r_i+f_i}\left(\frac{\lambda}{\mu_i}\right)^{x_i}\frac{e^{(\lambda/\mu_i)}}{x_i!}+\frac{f_i}{r_i+f_i}\left(\frac{\lambda'}{\mu_i'}\right)^{x_i}\frac{e^{(\lambda'/\mu_i')}}{x_i!}$$

Theorem 1. *Let A be a sum of $k>1$ mixtures of two Poisson with weights $p_1^j, p_2^j = 1-p_1^i$, and with parameters λ_s^j, where j represents one particular mixture in the sum ($j=1,2,\dots k$), and s represents the inner random variables of each mixture ($s=1,2$).*

Then A is a mixture of 2^k Poissons named A_ω, where $\omega=\{s_1,s_2,...,s_k\}$ is a

string containing a particular combination of s's for each j. The weights is defined as the product $\prod_{j=1}^{k} p_{s_j}^{j}$. *The parameter for each* A_ω *is defined as* $\sum_{j=1}^{k} \lambda_{s_j}^{j}$.

We will prove this using two alternative approaches.

1.) By using direct convolution

Proof. We will prove by induction.

Base case

Let X_1 be a Poisson RV with parameter λ_1 and X_2 be a Poisson RV with parameter λ_2. Let Y_1 be a Poisson RV with parameter μ_1 and Y_2 be a Poisson RV with parameter μ_2.

Then, let X and Y be two mixtures:

$$f_X = p_1 f_{X_1} + p_2 \cdot f_{X_2}$$

$$f_Y = q_1 f_{Y_1} + q_2 \cdot f_{Y_2},$$

where $p_1 + p_2 = 1$ and $q_1 + q_2 = 1$. We want to derive the probability mass function for the RV A, where A = X+Y.

$$\begin{aligned}
f_A(n) &= \sum_{i=0}^{n} f_X(i) f_Y(n-i) \\
&= \sum_{i=0}^{n} \left(p_1 \cdot f_{X_1}(i) + p_2 \cdot f_{X_2}(i)\right) \cdot \left(q_1 \cdot f_{Y_1}(n-i) + q_2 \cdot f_{Y_2}(n-i)\right) \\
&= p_1 \cdot q_1 \sum_{i=0}^{n} f_{X_1}(i)\, f_{Y_1}(n-i) + p_1 q_2 \sum_{i=0}^{n} f_{X_1}(i)\, f_{Y_2}(n-i) \\
&\quad + p_2 q_1 \sum_{i=0}^{n} f_{X_2}(i)\, f_{Y_1}(n-i) + p_2 q_2 \sum_{i=0}^{n} f_{X_2}(i)\, f_{Y_2}(n-i) \\
f_A(n) &= p_1 q_1 \cdot f_{A_{11}}(n) + p_1 q_2 \cdot f_{A_{12}}(n) + p_2 q_1 \cdot f_{A_{21}}(n) + p_2 q_2 \cdot f_{A_{22}}(n),
\end{aligned}$$

where A_{ij} is a Poisson RV with parameter $\lambda_i + \mu_j$. This finishes the proof for $k = 2$.

Induction Step

Assume that, for an arbitrary k, the claim is valid. We will show that it remains valid for $k+1$.

Let X be mixture of 2^k Poisson RV $(X_1, \ldots, X_{2^k})$ with weights w_j and parameters $\rho_j, j = 1, 2, \ldots 2^k$. Let Y be a mixture of 2 Poisson RV (Y_1, Y_2) with weights p_1 and $p_2 = 1 - p_1$ and parameters λ_1 and λ_2.

$$\begin{aligned}
f_A(n) &= \sum_{i=0}^{n} f_X(i) f_Y(n-i) \\
&= \sum_{i=0}^{n} \left(\sum_{j=1}^{2^k} w_j \cdot f_{X_j}(i) \right) \cdot \left(p_1 \cdot f_{Y_1}(n-i) + p_2 \cdot f_{Y_2}(n-i) \right) \\
&= \sum_{i=0}^{n} \sum_{j=1}^{2^k} \left(p_1 w_j \cdot f_{X_j}(i) f_{Y_1}(n-i) + p_2 w_j \cdot f_{X_j}(i) f_{Y_2}(n-i) \right) \\
&= \sum_{j=1}^{2^k} \sum_{i=0}^{n} \left(p_1 w_j \cdot f_{X_j}(i) f_{Y_1}(n-i) + p_2 w_j \cdot f_{X_j}(i) f_{Y_2}(n-i) \right) \\
&= \sum_{j=1}^{2^k} \left(p_1 w_j \sum_{i=0}^{n} \left(f_{X_j}(i) f_{Y_1}(n-i) \right) + p_2 w_j \sum_{i=0}^{n} \left(f_{X_j}(i) f_{Y_2}(n-i) \right) \right) \\
&= \sum_{j=1}^{2^k} p_1 w_j f_{A_{j1}}(n) + \sum_{j=0}^{2^k} p_2 w_j f_{A_{j2}}(n),
\end{aligned}$$

where A_{js} is a Poisson random variable with parameter $\rho_j + \lambda_s$, $j = 1, 2, \ldots, k$, and $s = 1, 2$. Therefore, A is a mixture of 2^{k+1} Poisson RV with corresponding weights for each A_{js} defined as $w_j p_s$, and parameter $\rho_j + \lambda_s$, $j = 1, 2, \ldots, k$, and $s = 1, 2$. □

2) By using generating functions

Proof. Also by induction.

Let $G_X(z)$ be the probability generating function for the RV X, i.e., $G(z) = E\left[z^X\right]$. Then, if we have a random variable $A = X + Y$, where X and Y are independent, $G_A(z) = E\left[z^{X+Y}\right] = E\left[z^X z^Y\right] = E\left[z^X\right] E\left[z^Y\right] = G_X(z) G_Y(z)$.

For a Poisson RV with parameter λ, G(z) = $e^{\lambda(z-1)}$. Given the properties of expectation, we also know that the probability generating function of a mixture of two RV is a mixture of the generating functions for each of the RV.

Base case

Let $f_X = p_1 f_{X_1} + p_2 f_{X_2}$ and $f_Y = q_1 f_{Y_1} + q_2 f_{Y_2}$, where $p_1 + p_2 = 1$, $q_1 + q_2 = 1$, and X_1, X_2, Y_1, and Y_2 are Poisson distributed with parameters λ_1, λ_2, μ_1, and μ_2, respectively. Then

$$
\begin{aligned}
&G_A(z) \\
&= \left(p_1 e^{\lambda_1(z-1)} + p_2 e^{\lambda_2(z-1)}\right)\left(q_1 e^{\mu_1(z-1)} + q_2 e^{\mu_2(z-1)}\right) \\
&= p_1 q_1 e^{(\lambda_1+\mu_1)(z-1)} + p_2 q_1 e^{(\lambda_2+\mu_1)(z-1)} + p_1 q_2 e^{(\lambda_1+\mu_2)(z-1)} + p_2 q_2 e^{(\lambda_1+\mu_2)(z-1)}
\end{aligned}
$$

This finishes the proof for k = 2.

Induction Step

Assume that, for an arbitrary k, the claim is valid. We will show that it remains valid for $k+1$.

Let X be mixture of 2^k Poisson RV $(X_1, \ldots, X_{2^k})$ with weights w_i and parameters $\rho_j, j = 1, 2, \ldots 2^k$, as defined in the setup. Let Y be a mixture of 2 Poisson RV (Y_1, Y_2) with weights p_1 and $p_2 = 1 - p_1$, and parameters λ_1 and λ_2. Then $G_X(z) = \sum_{j=0}^{2^k} w_j e^{\rho_j(z-1)}$, and $G_Y(z) = p_1 e^{\lambda_1(z-1)} + (1-p_1)\, e^{\lambda_2(z-1)}$.

$$
\begin{aligned}
G_A(z) &= G_X(z) G_Y(z) \\
&= \left(\sum_{j=1}^{2^k} w_j e^{\rho_j(z-1)}\right)\left(p_1 e^{\lambda_1(z-1)} + p_2 e^{\lambda_2(z-1)}\right) \\
&= \sum_{j=1}^{2^k} w_j e^{\rho_j(z-1)} p_1 e^{\lambda_1(z-1)} + \sum_{j=1}^{2^k} w_j e^{\rho_j(z-1)} p_2 e^{\lambda_2(z-1)} \\
&= \sum_{j=1}^{2^k} p_1 w_j e^{(\rho_j+\lambda_1)(z-1)} + \sum_{j=1}^{2^k} p_2 w_j e^{(\rho_j+\lambda_2)(z-1)}.
\end{aligned}
$$

Therefore, A is a mixture of 2^{k+1} Poisson random variables with parameters $\rho_j + \lambda_s$, and corresponding weights defined as $w_j p_s$, where $j = 1, \ldots, 2^k$, and $s = 1, 2$. □

Applying Theorem 1 for Traffic Density

From chapter 2 and Lopes Gerum et al. (2019b), we know that during non-peak hours the probability mass function for traffic density in one 0.5-mile segment can be represented as:

$$f_X(n) = \frac{r}{r+f} f_{X_\phi}(n) + \frac{f}{r+f} f_Y(n),$$

where X_ϕ follows a Poisson with parameter $\frac{\lambda}{\mu}$, and Y follows a Poisson with parameter $\frac{\lambda'}{\mu'}$. Looking at a longer stretch, composed of k 0.5-mile segments, we have the distribution for the overall stretch determined as:

$$f_{X_{\text{all}}}(n) = \sum_{j=0}^{2^k} p_j f_{X_j}(n),$$

where j corresponds to each possible combination of systems behavior for each segment. For each j, the corresponding weight is, therefore, the product of all weights corresponding to each segment's behavior, and the parameter for the corresponding Poisson RV is the sum of the parameters for each segment, given the behavior.

As an example, let us look at two 0.5-mile segments, which form 1-mile segment. The parameters for each segment i is given as λ, λ', μ_i, μ_i', f_i, r_i, $i = 1, 2$. In this case,

$$f_{X_{\text{all}}} = \frac{r_1}{r_1+f_1}\frac{r_2}{r_2+f_2} f_{X_1} + \frac{r_1}{r_1+f_1}\frac{f_2}{r_2+f_2} f_{X_2} + \frac{f_1}{r_1+f_1}\frac{r_2}{r_2+f_2} f_{X_3} + \frac{f_1}{r_1+f_1}\frac{f_2}{r_2+f_2} f_{X_4},$$

and the parameters for X_1, X_2, X_3, and X_4 are $\left(\frac{\lambda}{\mu_1} + \frac{\lambda}{\mu_2}\right)$, $\left(\frac{\lambda}{\mu_1} + \frac{\lambda'}{\mu_2'}\right)$, $\left(\frac{\lambda'}{\mu_1'} + \frac{\lambda}{\mu_2}\right)$, and $\left(\frac{\lambda'}{\mu_1'} + \frac{\lambda'}{\mu_2'}\right)$, respectively.

3.2 Validation

We use the data described in section 2.1 to validate the proposed model. After filtering, we split the data in two groups, training and testing. The first group is used to compute the parameters and to estimate the best fit for the distributions. The testing group serves as an out of sample dataset to compute performance metrics, such as the Akaike Information Criterion (AIC). 80% of the data is used as training, and 20% as testing.

This validation determines whether the separation in smaller segments is necessary. It compares the fit to the data of the tandem-queue-based distribution against the fit of combining the segments in a single long stretch with a single arrival rate and single service rate – i.e., disregarding spatial variation in the rate parameters. Then it verifies how the proposed model compares with the standard lognormal distribution used in literature.

3.2.1 Aggregate Parameters for the Single-Queue Approach

Let us combine k sequential segments into a long stretch. The segment index is used to separate each parameter set from each other. One assumption made is that, if one segment fails, the whole stretch fails. This assumption implies that the travel time deteriorates in the whole stretch.

This assumption is valid for our dataset during winter times. When looking at all two subsequent sensors at a time, we see the periods in which one sensor failed while the other did not are, on average, 0.63% of the failed time, with a maximum of 2.56% of the failed time for a particular pair during the whole analyzed period. Since snow might skew this, we also analyzed the proportion for summer months. In this case, we see the periods in which one sensor failed while the other did not are, on average,

99.2% of the failed time, with a maximum of 100% of the failed time for a particular pair. Therefore, although the sensors are fully correlated during the winter because of snow, they are independent over the summer.

We use winter data for validation, and we obtain the parameters from the equations in Section 3.1.

3.2.2 Parameter Fitting for a Mixture of Poissons

Because we try to compare the performance of the convolutional distribution to one used in literature, we choose the parameters best fit for the data, rather than using the parameter estimation methods discussed. This choice assures fairness in the comparison, since finding the parameters for lognormal demand usage of the whole dataset. We remind the reader that this process is not needed when estimating the distribution with the aggregate parameters, a considerable advantage of our formulation.

The best-suited parameters $\underset{\sim}{\theta}$ for a given distribution f can be estimated from N data points by maximizing the complete likelihood function,

$$\mathcal{L}(\underset{\sim}{x} : \underset{\sim}{\theta}) = \prod_{i=1}^{N} f_{\underset{\sim}{\theta}}(x_i).$$

Typically, the optimization is performed in the log realm, to avoid numerical issues. Therefore:

$$\underset{\sim}{\theta}^* = \arg\max_{\underset{\sim}{\theta}} \log \mathcal{L}(\underset{\sim}{x} : \underset{\sim}{\theta}) = \arg\max_{\underset{\sim}{\theta}} \sum_{i=1}^{N} \log f_{\underset{\sim}{\theta}}(x_i).$$

In the case of large mixtures, this maximization problem can be hard to solve. The constraints on the weight parameters and the maximization of the logarithm of a sum often cause numerical issues that impact the instability of non-linear optimization solvers. An alternative approach for such problems is an application of the Expectation-Maximization algorithm described in Bilmes et al. (1998) and Tomasi

(2004). It separates the global optimization problem into two smaller ones. Starting from a randomly assigned set of parameters, it then iterates until convergence. We avoid the problem of maximizing the logarithm of a sum by bounding the log-likelihood function using Jensen's inequality. The derivation follows:

Derivation for the algorithm

Adapted from Bilmes et al. (1998) and Tomasi (2004).

Suppose we want to fit parameters for a mixture of K Poisson random variables. We denote each pmf as $g_j(x) = \frac{e^{-\lambda_j}\lambda_j^x}{x!}$. The objective is to find the best parameters λ_ks and weigths p_ks for a dataset containing N points. From the initial parameters, we determine the *membership probabilities* of each data point, $p(k \mid x_i)$ for each $i = 1, ..., N$. These represent the probabilities that datapoint x_i belongs to random variable k. From Bayes' rule:

$$p(j \mid x_i) = \frac{p(x_j \mid k)p_j}{p(x_i)} = \frac{g_j(x_i)p_j}{\sum_{m=1}^{K} g_m(x_i)p_m}.$$

In practice, if a datapoint x_i, a weight p_j, or a parameter λ_j are large, numerical issues may arise. This becomes increasingly problematic the longer the stretch is, because more cars are expected to be present at any given time. To overcome this numerical instability, we alter this equation to

$$p(j \mid x_i) = \frac{p_j \frac{e^{-\lambda_j}\lambda_j^{x_i}}{x_i!}}{\sum_{m=1}^{K} \frac{e^{-\lambda_m}\lambda_m^{x_i}}{x_i!} p_m} = \frac{\exp\left(\log\left(p_j\right) - \lambda_j + x_i \log\left(\lambda_j\right)\right)}{\sum_{m=1}^{K} \exp\left(\log\left(p_m\right) - \lambda_m + x_i \log\left(\lambda_m\right)\right)},$$

and we set a lower bound to the weight parameters. If these algebraic changes are not sufficient to prevent numerical issues caused by the magnitude of the datapoint, a simple scaling of the data solves the problem.

We now determine the best parameters while keeping the *membership probabilities* fixed. From Jensen's inequality, we have that,

$$\log \sum_{j=1}^{K} \pi_k \alpha_k \geq \sum_{j=1}^{K} \pi_j \log \alpha_j,$$

because of the convexity of the logarithm. By replacing π_j with $p_j g_j(x_i)$ and α_j with $\frac{p_j g_j(x_i)}{p(j \mid x_i)}$, we have:

$$\log \mathcal{L}(\underset{\sim}{x} : \underset{\sim}{\theta}) = \sum_{i=1}^{N} \log \sum_{j=1}^{K} p_j g_j(x_i) \geq \sum_{i=1}^{N} \sum_{j=1}^{K} p(j \mid x_i) \log \frac{p_j g_j(x_i)}{p(j \mid x_i)}.$$

We focus on maximizing the bound determined in the inequality. Note that the numerator in the logarithm is fixed at this point and can, therefore, be disregarded by separating it out of the logarithm. We call the new objective function to maximize $b(\underset{\sim}{\theta})$:

$$\begin{aligned}
b(\underset{\sim}{\theta}) &= \sum_{i=1}^{N} \sum_{j=1}^{K} p(j \mid x_i) \log p_j g_j(x_i) \\
\frac{\partial b}{\partial \lambda_j} &= \frac{\partial}{\partial \lambda_j} \sum_{i=1}^{N} \sum_{j=1}^{K} p(j \mid x_i) \log p_j g_j(x_i) \\
&= \sum_{i=1}^{N} \sum_{j=1}^{K} p(j \mid x_i) \frac{\partial}{\partial \lambda_j} \log p_j g_j(x_i) \\
&= \sum_{i=1}^{N} p(j \mid x_i) \left(\frac{x_i}{\lambda_j} - 1 \right).
\end{aligned}$$

By equating the derivative to zero, we can estimate λ_j for each j as:

$$\lambda_j = \frac{\sum_{i=1}^{N} p(j \mid x_i) x_i}{\sum_{i=1}^{N} p(j \mid x_i)}.$$

Once the λ_js are determined, we can find the best p_j for these new parameters. Given the constraint that the weights must add up to 1, we add a Lagrange multiplier.

$$\begin{aligned}
b(\underset{\sim}{\theta}) &= \sum_{i=1}^{N} \sum_{j=1}^{K} p(j \mid x_i) \log p_j g_j(x_i) + \mu \left(\sum_{j=1}^{K} (p_j) - 1 \right) \\
\frac{\partial b}{\partial p_j} &= \frac{\partial}{\partial p_j} \sum_{i=1}^{N} \sum_{j=1}^{K} p(j \mid x_i) \log p_j g_j(x_i) + \mu \left(\sum_{j=1}^{K} (p_j) - 1 \right) \\
&= \sum_{i=1}^{N} \frac{\partial}{\partial p_j} p(k \mid x_i) \log p_j g_j(x_i) + \frac{\partial}{\partial p_j} \mu \left(p_j - 1 \right) \\
&= \sum_{i=1}^{N} \frac{p(j \mid x_i)}{p_j} + \mu.
\end{aligned}$$

By summing of j and equating it to zero, we note that $\mu = -N$.

$$0 = \sum_{i=1}^{N} \frac{p(j \mid x_i)}{p_j} + \mu$$

$$\sum_{j=1}^{K} \mu p_j = -\sum_{i=1}^{N} \sum_{j=1}^{K} p(j \mid x_i)$$

$$\mu = -N.$$

Therefore, we can estimate p_j for each j as:

$$p_j = \frac{\sum_{i=1}^{N} p(j \mid x_i)}{N}.$$

With this algorithm, one can then estimate both the parameters and the weights by iterating until convergence of parameters and weights. Algorithm 1 summarizes the steps. We use this method to determine the best parameters and weights for the mixtures and the standard MLE formulatio for the lognormal.

3.3 Discussion and Findings

For our validation, we analyze groups of sensors of varying lengths. The data for each group is the sum of the minute-stamped traffic density data of all sensors belonging to that group. We compare how combining the segments and treating them as a single queue perform versus treating the system as a tandem queue. Finally, we also compare the goodness of the fit of the tandem-queue approach against a lognormal fit. The results are presented in Table 3.1.

For each group of sensors, we generate the distributions of all three models. Part of the distributions are depicted in Figure 3.3. It displays stretches of varying lengths in the South-East direction during January and February.

Lognormal seems to better fit the data for long segments, but Figure 3.3 suggests that extreme tails are better captured by our model. We remind the reader that

Algorithm 1 Expectation Maximization to Estimate Parameters for a Mixture of Poissons

Require: $\underset{\sim}{x}$ with N datapoints, and K, the number of mixtures.

Initialize with a random set of K λ_js and weights p_js.
while convergence is not obtained **do**
    for i = 1 to N **do**
        for j = 1 to K **do**
            $a_{ij} = e^{-\lambda_j} \lambda_j^{x_i} p_k$
        end for
    end for
    for j = 1 to K **do**
        for i = 1 to N **do**
            $p(j \mid x_i) = \frac{a_{ij}}{\sum_{m=1}^{K} a_m}$
        end for
        $\lambda_j = \frac{\sum_{i=1}^{N} p(j \mid x_i) x_i}{\sum_{i=1}^{N} p(j \mid x_i)}$
        $p_j = \frac{\sum_{i=1}^{N} p(j \mid x_i)}{N}$
    end for
end while

	Mean AIC				
Stretch Length	Single Queue	Tandem	Lnormal	Diff.	Tandem Wins
0.5 mi	6833	-	6785	<1%	7/15
1 mi	8602	8205	7959	3%	3/14
1.5 mi	9752	9199	8672	6%	0/12
2 mi	11512	9660	9193	5%	1/10
2.5 mi	12448	10336	9601	7%	0/8

Table 3.1: Tues–Thurs; Jan and Feb; 10am-1pm; SE direction; Sensors 2-10 and 12-17

the proposed model could be generated with intuive aggregate parameters, while lognormal cannot. Table 3.1 also indicates that the proposed model has AICs that are close to lognormal.

Combining the segments and treating the stretch as a single queue indicates a bias towards overestimating the number of cars present. This bias likely occurs because of the presence of exit ramps in the stretch. Although this likely affects the tandem-queue approach as well, Figure 3.4 agrees with the results of Table 3.1 and supports that considering parameters separately dramatically improves performance.

The resulting AICs presented in table 3.1 demonstrate that the tandem-queue framework performs marginally worse than the lognormal when the stretch is sufficiently short. In the particular length of 0.5-mile, the queueing theory approach performed better than lognormal in almost half of the analized sensors. We hypothesize that the reason for this is because the locations in which the model performed better than lognormal are straight sections, as pictured in Figure 3.5 (the full list of AICs can be found in the Appendix).

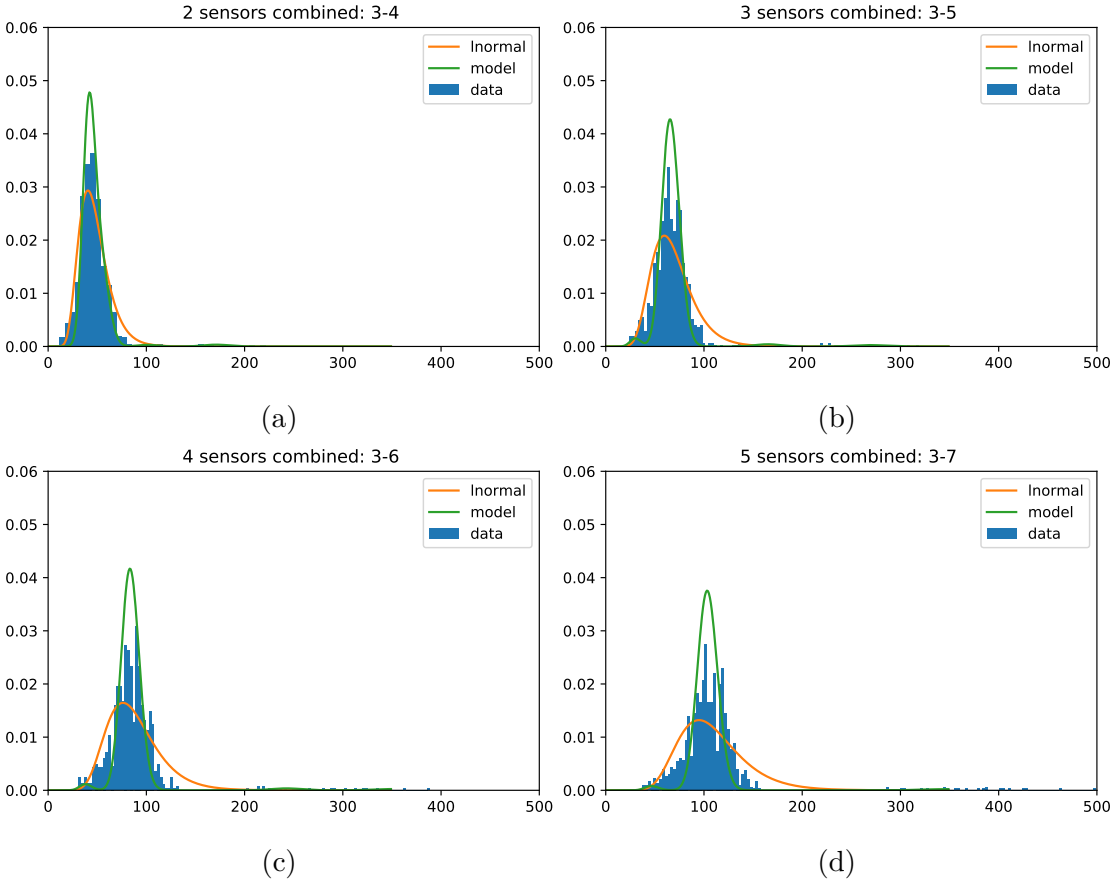


Figure 3.3: Best fit for model and lognormal in varying lengths.

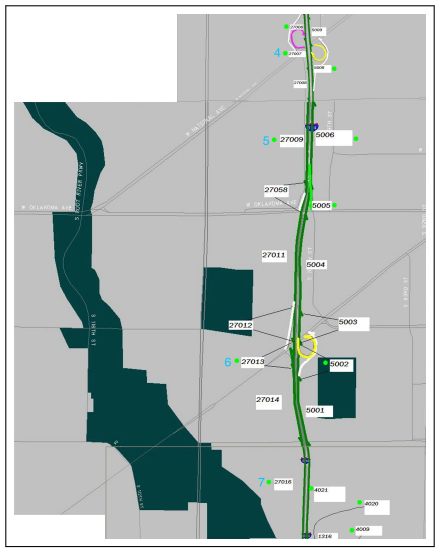


Figure 3.5: Zoomed-in map for sensors whose data were best fit by the proposed model.

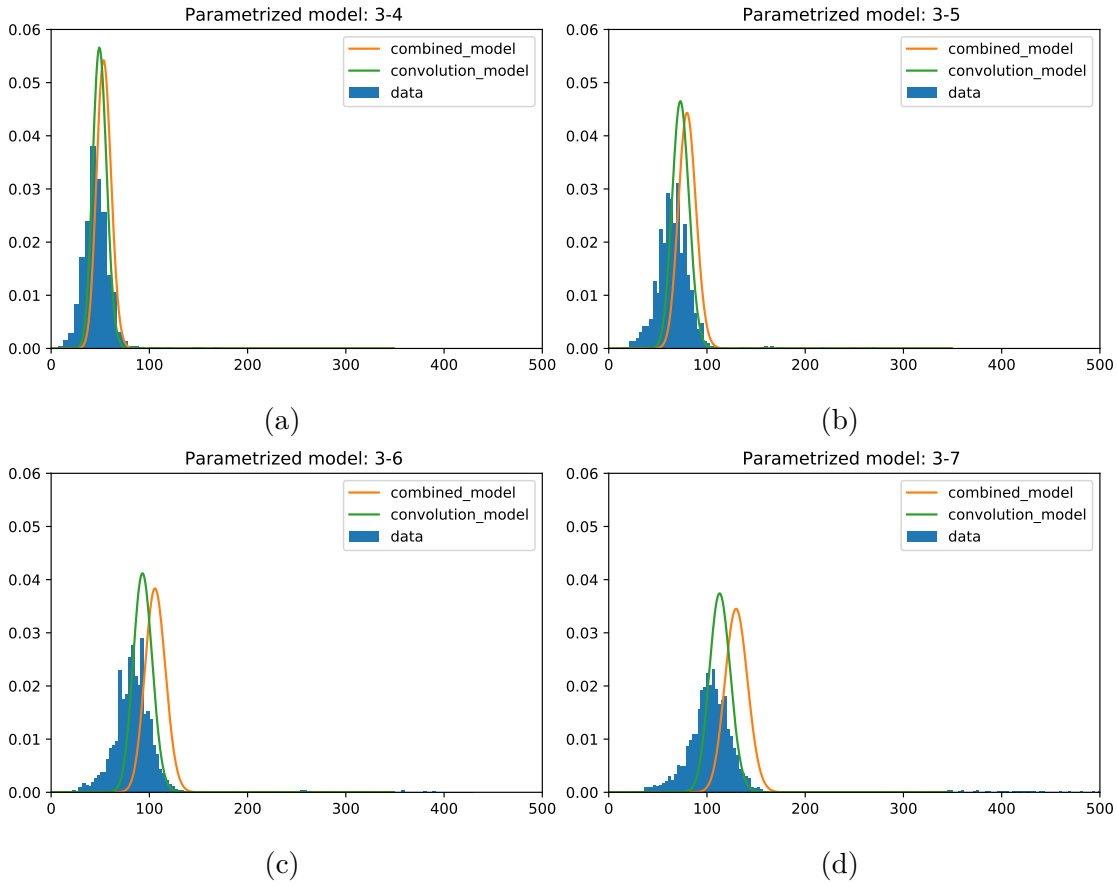


Figure 3.4: Traffic Density pmf's for varying stretch lengths

For longer stretches, the lognormal distribution seems to be a better fit, likely due to additional uncertainties caused by ramps and merges. Nonetheless, the tandem-queue model allows for direct computation without the need for simulation or massive amounts of data. Mere knowledge of aggregate parameters is sufficient for obtaining an accurate initial gauge of the distribution of traffic density.

APPENDIX – Full list of AICs for group of sensors (SE Direction)

Sensor Group	Lognormal	Model	Difference
(2)	6616.064329	6706.689514	90.625185
(3)	7241.579373	7420.351889	178.772516
(4)	6941.690609	7043.023467	101.332858
(5)	6976.848200	6846.907029	-129.941171
(6)	6923.414294	6913.936331	-9.477963
(7)	7176.321850	7230.446762	54.124913
(8)	6742.180870	6721.037478	-21.143391
(9)	6510.933888	6438.630752	-72.303136
(10)	6678.489760	6675.813118	-2.676643
(2, 3)	8132.120154	8312.090917	179.970763
(3, 4)	8266.963759	8939.652260	672.688501
(4, 5)	8209.168325	8114.226066	-94.942259
(5, 6)	8127.608853	8056.259081	-71.349772
(6, 7)	8303.057533	8109.814437	-193.243096
(7, 8)	8082.074626	8412.461911	330.387285
(8, 9)	7612.641413	7925.161553	312.520140
(9, 10)	7729.546542	8075.208779	345.662237
(2, 3, 4)	8915.556127	9397.999867	482.443741
(3, 4, 5)	8968.186073	9769.043441	800.857367

Continued on next page

Sensor Group	Lognormal	Model	Difference
(4, 5, 6)	8888.413364	9339.226056	450.812692
(5, 6, 7)	8979.481643	9021.974507	42.492864
(6, 7, 8)	8843.769424	9229.792986	386.023562
(7, 8, 9)	8519.538627	9153.381768	633.843141
(8, 9, 10)	8328.177821	8439.595207	111.417386
(2, 3, 4, 5)	9440.478266	9977.768101	537.289835
(3, 4, 5, 6)	9449.249861	9310.562550	-138.687311
(4, 5, 6, 7)	9494.534061	9776.934586	282.400525
(5, 6, 7, 8)	9378.801319	9582.570731	203.769411
(6, 7, 8, 9)	9153.699375	9371.976896	218.277521
(7, 8, 9, 10)	8960.664078	9264.044350	303.380272
(2, 3, 4, 5, 6)	9824.970142	10731.551441	906.581298
(3, 4, 5, 6, 7)	9913.180390	10469.584261	556.403871
(4, 5, 6, 7, 8)	9802.216136	10278.987066	476.770930
(5, 6, 7, 8, 9)	9624.691801	10007.094243	382.402442
(6, 7, 8, 9, 10)	9476.898664	9953.910466	477.011802
(12)	6722.019504	6589.541613	-132.477891
(13)	5876.105482	5782.412761	-93.692721
(14)	6644.163897	6713.277376	69.113480
(15)	6726.669374	6879.733407	153.064033
(16)	6631.546961	6827.397901	195.850940
(17)	7188.894622	7781.590979	592.696356
(18)	6971.432413	NaN	NaN

Continued on next page

Sensor Group	Lognormal	Model	Difference
(12, 13)	7279.070223	7384.582621	105.512398
(13, 14)	7667.674507	7936.763552	269.089044
(14, 15)	7687.887747	8094.220382	406.332635
(15, 16)	8093.113335	8324.562822	231.449488
(16, 17)	8082.186279	8430.295575	348.109295
(17, 18)	8157.556223	8755.015071	597.458848
(12, 13, 14)	8201.288497	8552.477523	351.189026
(13, 14, 15)	8294.558679	9032.536325	737.977647
(14, 15, 16)	8562.843790	9433.897149	871.053359
(15, 16, 17)	8849.409436	9512.502706	663.093270
(16, 17, 18)	8722.291724	9508.861537	786.569812
(12, 13, 14, 15)	8701.000442	9836.005799	1135.005357
(13, 14, 15, 16)	8927.582001	9757.260249	829.678249
(14, 15, 16, 17)	9174.479403	9756.748822	582.269419
(15, 16, 17, 18)	9251.391537	9973.229226	721.837689
(12, 13, 14, 15, 16)	9219.588584	10098.345775	878.757191
(13, 14, 15, 16, 17)	9428.916595	10582.607773	1153.691178
(14, 15, 16, 17, 18)	9522.924462	10567.748367	1044.823905

Part II

Predictive Inspection and Maintenance Scheduling for Railway Tracks

Chapter 4

Literature Review

Literature has introduced several methods for the problems of determining the number of defects on a track, and scheduling railway inspections. However, these methods face at least one of several shortcomings listed below.

Defect Prediction

1. they use data that is often unavailable or costly to gather — and therefore are infeasible for most practical applications;

2. they are too simplistic — and therefore neglect important features that are easily accessible.

Maintenance and Inspection scheduling

1. they assume the number of defects to be deterministic — and therefore fail to account for the stochasticity of defect generation;

2. they include too many constraints in their model — and therefore can be insolvable with the currently available computers.

A noteworthy additional flaw not yet extensively acknowledged in the literature is the absence of integration methods between solutions for each problem. This integration is particularly relevant because track defects are one of the major causes of derailments. This chapter discusses various approaches proposed in the literature.

4.1 Railway Defect Prediction

Earlier studies use summary statistics, particularly standard deviation, to evaluate the risk of defects or derailments (Hamid and Gross, 1981). However, these summary statistics are not constant over time and depend on conditions such as season, and load. New studies implement complex models for more accurate reflections of real-world conditions that benefit from probabilistic and stochastic approaches.

A portion of the more recent literature uses classification and regression-based data analysis to predict defects. Sharma et al. (2018) develop a Track Quality Index (TQI) based Markov Model. They divide tracks into 0.1-mile segments and compare Random Forests (RF), Logistic Regression (LR) and Support Vector Machines (SVM) models that predict a binary defect/no defect outcome. Martey et al. (2017) attempt to predict geometric defects in a big data environment. However, since analyzing big data has execution time constraints, they combine tracks so that similar tracks can be studied together. They also employ Principle Component Analysis (PCA) at every cluster to observe the determinants that cause defects and then predict the number of defects via Linear Regression, Random Forests, and Support Vector Regression techniques. In both studies (Sharma et al., 2018; Martey et al., 2017), Random Forests outperforms other methods.

Moridpour et al. (2017) develop a regression-based model that implements the degradation level of light rail tracks using Artificial Neural Networks (ANNs). A similar study conducted by Güler (2014) predicts degradation due to geometric defects

using ANN. While Güler (2014) prioritize geographic and train-related inputs, Moridpour et al. (2017) include urban and traffic-related inputs. Both authors report that Artificial Neural Networks provide a success rate greater than 70%. One limitation of such approaches is that they require detailed data, without which the performances decrease drastically.

4.2 Railway Maintenance and Inspections

The scheduling of inspections and maintenance has been described as a complex, multi-variate problem of significant importance (Fan et al., 2011; Chen et al., 2014; Peralta et al., 2018; D'Ariano et al., 2019; Ghofrani et al., 2018). Previous studies attempt to solve this problem by various adaptations of well-known optimization problems such as the Vehicle Routing Problem (VRP) or the Traveling Salesman Problem (TSP). However, most studies assume that defects are deterministically known. They fail to account for the stochasticity of where and when defects occur and thus disregard the prediction problem.

Heinicke et al. (2015) characterize maintenance scheduling as a multi-depot VRP with time windows. Fan et al. (2011) also consider a VRP, but only include transportation (travel) costs. Despite the effort in modeling, Fan et al. (2011) and Heinicke et al. (2015) do not provide algorithms to solve the model. They recognize that the models are NP-hard, thus making a polynomial run time to reach the optimal solution difficult to achieve.

Following the adaptations of well-known optimization problems, Camci (2014) coins the term Traveling Maintainer Problem that represents a scheduling TSP with repair, and inspections costs added to travel costs. Pour et al. (2018) construct a Mixed Integer Programming model to assign teams according to their capabilities near-optimally. A Lagrangian relaxation-based solution for this model is developed

by Luan et al. (2017). A more extensive study is then conducted by Lidén and Joborn (2017), who further incorporate the railway traffic into inspection and maintenance scheduling. Nonetheless, all of these studies mention that the large number of variables results in considerable execution time due to the NP-hardness of the mathematical programs. Hence, heuristics and metaheuristics are frequently implemented to obtain a feasible schedule (Camci, 2014; Andrade and Teixeira, 2016; Khalouli et al., 2016), but that is not guaranteed to be optimal. For some modeling approaches, column generation methods have proven successful in finding exact optimal solutions for large-sized problems (Nishi et al., 2011; Lannez et al., 2015).

The incorporation of risk-based qualitative data into the maintenance scheduling problem has also been considered in the literature. Jiang et al. (2003) employs multiple degradation states with associated probabilities of failure to decide on the maintenance schedule. This work is further developed by Consilvio et al. (2016), who helps determine threshold levels to achieve a tolerable degradation range. Recently, Wang et al. (2018) break down failure consequences into fuzzy linguistic classes from negligible to catastrophic, and qualify failure likelihood from very low to very high.

In a recent review of railway transportation, Ghofrani et al. (2018) classify railway maintenance as: a) condition-based, b) preventive, and c) corrective maintenance. Predictive maintenance is embedded within condition-based maintenance. According to Ghofrani et al. (2018), literature mostly deals with track defects using corrective maintenance, and scheduling is mostly planned in cases when defects are already known. Moreover, condition-based maintenance is mostly exploited for vehicle maintenance but rarely applied to tracks. Turner et al. (2016) provide an extensive literature review on planning and scheduling of railway traffic in Europe, also discussing a few studies that incorporate maintenance activities into transportation planning and scheduling.

Chapter 5

Railways Track and Geometry Defect Predictions

Defects in a railway system are a measure of track and railway quality. Hence, various studies implemented applications focused on qualifying track quality based on the number and size of defects present. This chapter presents a new approach to defect prediction that builds from the literature but allows for a more direct application. These advancements are also used in later chapters to help decision-makers better schedule inspections and maintenance in tracks.

In rail terminology, geometry defects are horizontal and vertical misalignments in the track (Sharma et al., 2018). In contrast, rail defects include track wear such as corrosion or impairments such as broken rails or cracks (Clark, 2004). Some prior work classifies defects as yellow or red, depending on their severity (He et al., 2015; Cárdenas-Gallo et al., 2017). Yellow defects are minor defects, such as superficial cracks or buckling, and satisfy the Federal Railroad Administration (FRA) standards. On the other hand, red defects are significant defects (e.g., broken rails) that do not meet FRA standards and need immediate repair. In this dissertation, we classify defects as yellow or red, following the literature.

Next, we will describe the data that will be later used in the examples.

5.1 Preprocessing Rail Data

Most rail companies gather data to understand the behavior of their processes. This data includes information on defects and inspections. However, many companies may struggle to maintain the accuracy of such data due to human interaction in data registration, and the lack of enforced standards. Furthermore, some companies also struggle with the availability of detailed data for prediction and scheduling. Therefore, there is a trade-off between the benefits of process optimization through data analysis and the cost of obtaining accurate and detailed data. In this section, we explain faulty data treatment and noise filtering while maintaining an acceptable level of data for modeling purposes.

We use rail data gathered over two years (2016 and 2017). The defect database includes a list of all rail and geometry defects found in each segment of the network. It details the time and the location in which the defect was found, and the type and severity (yellow or red) of the defect found. The inspection data include the type of inspection, the segment inspected, and the time of inspection. However, the inspection data do not provide the exact beginning and end of inspection locations. Moreover, segments vary from 300m to 60km in length, with an average of 17 km. Figure 2 displays spatio-temporal defect observations on a particular segment. The horizontal axis represents the whole length of the rail segment, while the vertical axis represents time. The blue dots on the vertical axis correspond to the inspections. Finally, a third database contains the daily load information for each segment. Load information encompasses the total gross tonnage endured by the segment, including the weight of wagons. The number of wagons transported or wheels in contact with the track were not available.

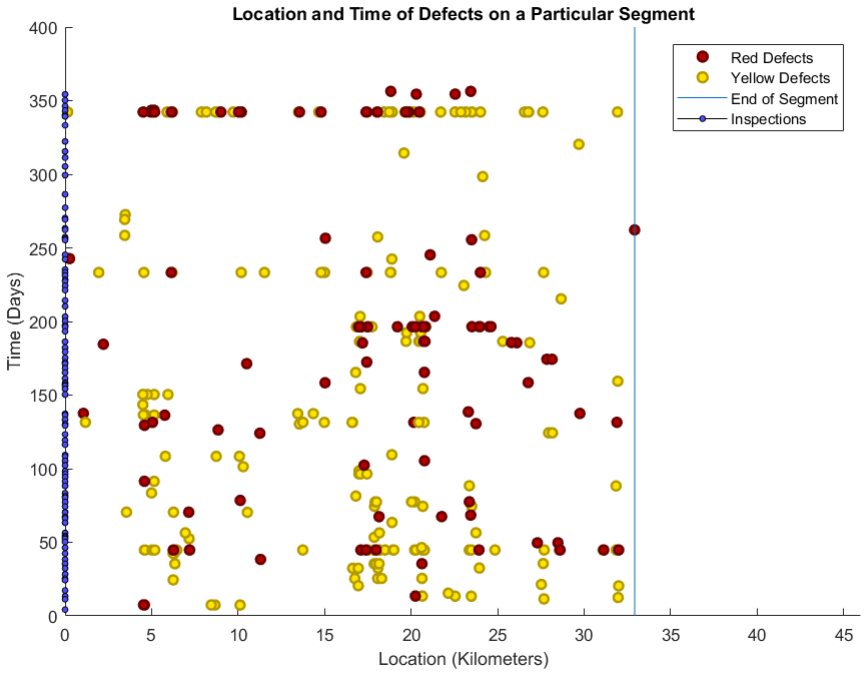


Figure 5.1: Spatiotemporal Defect Observations.

The defect data do not contain records for the corresponding inspection and vice versa. To better assess the distribution of the number of defects per inspection, each defect (rail or geometry) found in the defect dataset is matched to a corresponding inspection. This matching is one-to-many: a defect can be found during only one inspection, but an inspection can find multiple defects. Through this process, we achieve 91% match, losing only 9% of the defect data. For each inspection, we match the defects located in the inspected segment, found through the specific inspection type, and registered during the time of the inspection. We add a buffer of no more than one day to represent the delay between the time the inspection outcomes recording and the inspection execution time. Unmatched data may represent a human error, as well as data systems' mismatches. Faulty data, such as inspections with negative execution times, or with negative inter-arrival times, are also removed, along with the defects that match them.

The raw data contains more than 26 000 inspections and 82 000 defects. Broken rails are one of the most commonly encountered red defect types (Figure 5.2). Moreover, they are the defect most often rated as red: around 96% of broken rails are classified as red. Table 5.1 shows the defect types that are most commonly rated as red. Note that the visual and ultrasound cracks usually precede broken rails.

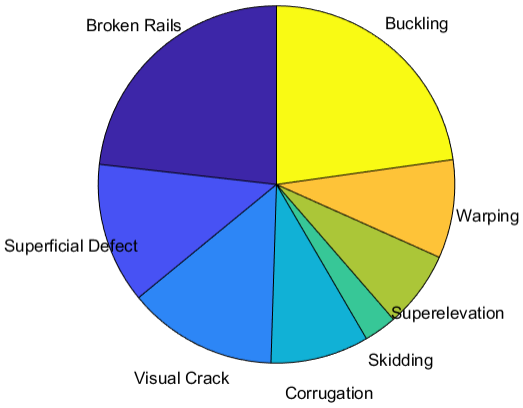


Figure 5.2: Distribution of Defects Classified as 'Red' per Type of Defect.

Defect Type	Percentage of Defects from each type classified as 'red'
Broken rails	95.97%
Visual Cracks	58.05%
Ultrasound cracks	52.52%
Warping	39.88%
Buckling	32.25%

Table 5.1: Defect types with highest rates of red defects.

5.2 Non-linear Regression Models for Defect Prediction

Literature agrees that predicting the number of defects in a track is mostly a non-linear problem. Most authors propose closed-form equations for prediction, but advanced 'black-box' techniques are succeeding them. Machine-learning approaches, such as RF and ANNs, have recently been employed for defect prediction (Güler, 2014; Moridpour et al., 2017; Sharma et al., 2018), but very little has been done on the integration of such prediction models with inspection and maintenance scheduling policies. Furthermore, many studies employ features that are often not available to rail companies. In this study, we focus on methods that use widely available data. In chapter 6, we integrate these approaches into the scheduling model to increase the efficiency of inspection and maintenance scheduling policies. In this section, we use a conventional methods to predict the number of yellow and red defects: Random Forests (Sharma et al., 2018).

The related data used to generate the input features are commonly available or easy to gather, such as the current month and season, and the gross load endured since the previous inspection. In both methods, the set of all features contains a) time in days since the last inspection, b) the gross load endured by the tracks since the last inspection, c) month, d) season, and e-f) the number of yellow and red defects found in the previous inspection.

Random Forest (RF) is an ensemble regression model that combines independent regressors, namely decision trees. Each decision tree is constructed by randomly selecting a subset of features from the set of all features, and by using the bootstrap aggregation technique. The output is determined by averaging the outputs of all decision trees. The use of multiple trees reduces the chance of over-fitting and decreases

the variance (Friedman et al., 2001). Each tree is generated by sampling the training data with replacement, and by using a randomly generated set of five features from the seven possible features. Then, trees are grown with the criterion of using the maximum impurity gain from all candidates to split branches. The impurity of each node is calculated with the Gini's diversity index to determine the impurity of each node (Breiman et al., 1984). We grow 300 trees in this modeling approach.

Yellow defects are more common than red defects, and the prediction model is run separately for both. The inputs of the model, as mentioned above, include the number of red and yellow defects the previous inspection has found; hence, the inputs are inspection related. The number of inspections is the same for both red and yellow defects; any inspection can find both types. Thus, red and yellow defect prediction models have the same number of inputs. The prediction model determines different structures for the relationship between the inputs and the outputs for red or yellow defects.

Finally, since the results of the predictions are real numbers, we use Multinomial Ordinal Regression (MRO) to assign each prediction to its corresponding number of defects.

5.2.1 Results

In this section, we provide the results of using the Random Forest algorithm to predict rail defects of different severity levels. We use the standard loss function that maximizes precision by analyzing mean absolute error and mean squared error. The % exact match represents the precision of the model, namely the percentage of testing data whose predictions perfectly matched the corresponding label. However, some risk-averse decision-makers may consider false negatives worse than false positives. For these cases, we propose a new formulation in Section 5.2.2.

Cluster	MAE	MSE	E[over]	E[under]
1	0.40	0.59	1.50	1.28
2	0.51	0.82	1.44	1.44
3	0.53	0.77	1.31	1.31
4	0.41	0.62	1.88	1.28

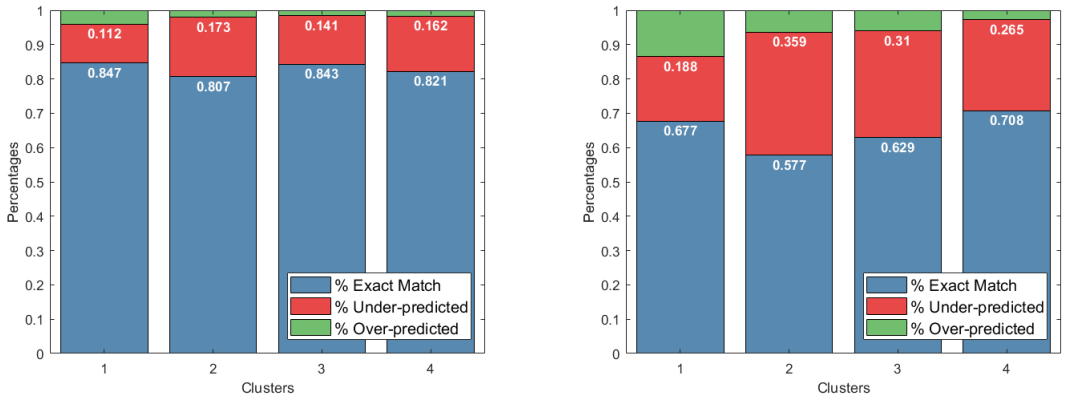
Table 5.2: Yellow Defects - Random Forest Error Results.

This method exhibits a small error rate that matches the one in literature, despite the fewer and more commonly found features used. Sharma et al. (2018) have recently predicted the existence of a geometry defect at an accuracy rate of 75-77%. Cárdenas-Gallo et al. (2017) apply the red/yellow distinction to railway defects, and their accuracy is also around 80%. Due to the high likelihood of inspections finding few to no defects, the models tend to undershoot more often than overshoot.

Table 5.2 displays the prediction of yellow defects during a walking inspection (0, 1, or 2 or more defects). The metrics Mean Absolute Error and Mean Squared Error are displayed in the number of defects per inspection, and the round numbers on the expected overestimation are caused by a few overestimates. Predicting a more granular number of defects demands more data points than are available in this study, but an increase in granularity is recommended as more data become available. 80% of the data is used to train the model, with the remaining 20% used for testing. The third and fourth columns describe the expected over or under-predicted number of defects. The average number of over-predicted defects is significantly skewed by extremely infrequent data points with no defect when it is predicted to have two defects. Note that over-prediction of defects occurs in less than 2% of all inspections, except in one cluster, where it occurs in around 4% of inspections.

Figures 5.3a and 5.3b depict the average accuracy for all clusters during one run.

The overall accuracy of perfect (spot-on) predictions for testing data is around 82% for red defects and 62% for yellow defects on average. Yellow defects are harder to predict due to their high variance in the data. Given the risk-averse characteristics of the problem, it may be preferable to over-predict rather than to under-predict if one considers over-predictions acceptable.



(a) Red Defects - RF Prediction Results. (b) Yellow Defects - RF Prediction Results.

Figure 5.3: Underestimation, Overestimation and Exact Prediction Rates.

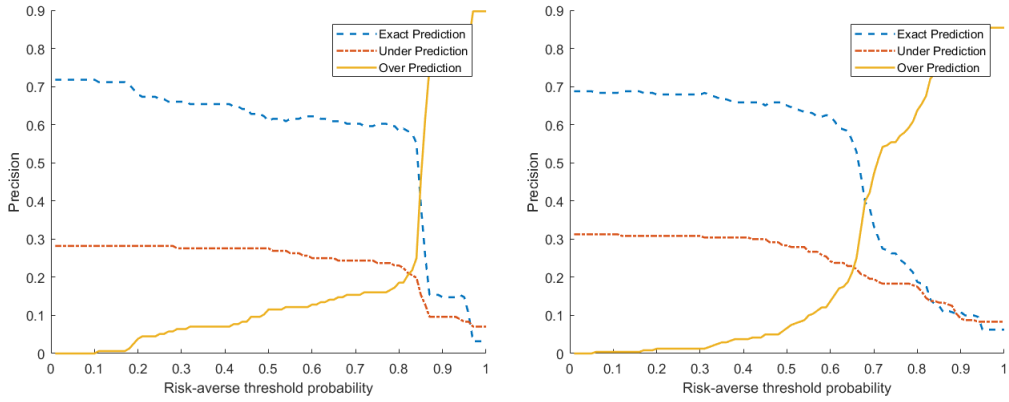
5.2.2 Risk-Averse Adaptation

In risk-averse screening systems, reducing missed observations is more critical than removing unnecessary screenings (Thomas et al., 2001). In railway systems, reducing missed observations refer to an overlooked rail or geometric defects on a rail segment, whereas an unnecessary screening refers to maintenance or inspection done in a defect-free rail segment. The risk of derailment and the possible consequences that stem from missing a defect may be more costly than the cost of ensuring that a segment remains defect-free. Hence, weights for underestimations become stricter than in overestimations in such cases.

Results obtained by RF tend to favor under-shooting due to the class imbalance of the data. However, decision-makers may choose to weight under-shooting more

heavily, since missing a defect may have worse consequences than over-shooting. The logistic regression results provide the confidence level of the model for each possible prediction option. The standard selection chooses the prediction with the highest confidence level, assuring a higher proportion of exact matches.

Risk-averse decision-makers may want to change the way the model chooses the prediction by going iteratively from the maximum possible number of defects predicted to the minimum, summing the confidence levels each time. Once a certain threshold is obtained, the number of defects is the prediction. Figure 5.4 depicts how proportions of exact matches, under-shooting, and over-shooting change for each threshold in one particular group of segments when using RF. It suggests that a steep increase in over prediction is the trade-off for decreasing under prediction.



(a) Different thresholds for red defects. (b) Different thresholds for yel. defects.

Figure 5.4: Risk-Averse Results when Changing Model Thresholds.

Chapter 6

Railways Inspection and Maintenance Scheduling

Most authors assume a known list of defects across the network before scheduling maintenance activities. However, when this information is not known with certainty, one has to resort to stochastic methods to find the optimum inspection policies. In this section, we propose a Markov decision process (Ross, 1992; Puterman, 1994; Bertsekas, 2007) model that integrates the stochastic nature of defect occurrence into scheduling. Through this model, one can determine the optimal inspection policy for a specific segment over an infinite time horizon. An assumption used is that all inspections are performed at the beginning of the day, and all defects observed during the inspection are repaired instantaneously. New defects might arise during the day after the inspection. Hence, one can model the segment deterioration state as a discrete time Markov process that depends on the previous state and the action taken in that state. Then, the Markovian decision problem to determine the sequence of actions that minimize the discounted cost incurred over an infinite horizon becomes

$$\nu(s(0)) = \min_{a(t)\in\mathcal{A}} E\left[\sum_{t=0}^{\infty} \alpha^t c(s(t), a(t))|\, s(0)\right], \tag{6.1}$$

where $\nu(\cdot)$ denotes the optimal total discounted cost, i.e., value function, starting in state $(\cdot)$, $s(t) \in \mathcal{S}$ denotes the segment state at time t with $\mathcal{S}$ representing the set of states, $a(t) \in \mathcal{A}$ denotes the binary variable representing if the segment is inspected or not at time t with $\mathcal{A}$ representing the set of actions, and $c(\cdot, \cdot)$ denotes the expected cost incurred when the state and action information are given. α represents the discount factor, and the notation $E[\cdot|s(0)]$ represents the conditional expectation of the discounted cost incurred over the infinite horizon given the initial deterioration state of the segment. Later on, we generalize our approach to the case of multiple segments under a constraint on the availability of inspection teams.

6.1 Dynamic Programming Formulation

A direct method to generate the optimal policy for equation 6.1 is to use the following dynamic programming equation (Ross, 1992; Puterman, 1994),

$$\nu(s) = \min_{a \in \mathcal{A}} \left\{ c(s,a) + \sum_{s' \in \mathcal{S}} \alpha \nu(s') \cdot p\{s'|s,a\} \right\}, \forall s \in \mathcal{S}, \tag{6.2}$$

with p denoting the transition kernel. $\nu(s)$ represents the expected discounted cost to go starting in state s from the current period onwards under the optimal policy. Let $c(s,a)$ correspond to the expected cost of inspection and repair if the action is 'inspect', and the expected cost of the risk of derailment otherwise. Since the state space, $\mathcal{S}$, and the action space, $\mathcal{A}$, are finite, $c(s,a)$ is uniformly bounded. $p\{s'|s,a\}$ denotes the probability of moving to a new state s', given the current state s and the action taken in the current period, a. These probabilities are estimated from the data. In this problem, we fix the discount rate as $\alpha = 0.95$.

Let us define the state of the segment as a composition of the current level of rail deterioration and the load during the next period. The rail deterioration level, or state, can be 1, when no red or yellow defects are present; 2, when no red defect

is present; and 3, when at least one red defect is present. Load is also separated into two states as high and low, where the cut-off is the median load endured by the whole network, obtained from data. There are, therefore, six possible states: $(1, L), (2, L), (3, L), (1, H), (2, H), (3, H)$, denoted as $1, 2, \ldots, 6$, respectively.

After filtering the inspection data by one day inter-inspection time within each cluster, the transition matrix for the action 'inspect' is obtained from the frequency of each state change. These normalized frequencies correspond to the maximum likelihood estimators of the transition probabilities (Bartlett, 1951; Anderson and Goodman, 1957).

State	1	2	3	4	5	6
1	0.5266	0.1339	0.1072	0.1518	0.0447	0.0358
2	0.1772	0.2910	0.1899	0.2278	0.1013	0.0128
3	0.2058	0.1912	0.2793	0.1912	0.0148	0.1177
4	0.0981	0.0785	0.0719	0.4705	0.2026	0.0785
5	0.0556	0.1805	0.0695	0.3471	0.2916	0.0556
6	0.2352	0.0590	0.1471	0.3233	0.0590	0.1765

Table 6.1: Transition Matrix for the action 'inspect'.

One methods is used to validate the transition matrix given in Table 6.1:

Correlation with the two-step matrix: A validation method used by Sharma et al. (2018) is based on the two-step state transition matrix, $P2$. If the one-step transition matrix is P, then the correlation between P^2 and $P2$ must be high. In our case, the correlation is found to be strong at level 0.8633.

This methods indicates the validity of the transition matrix for 'inspect', which is presented in Table 6.1. We use this transition matrix as a starting point to determine

State	1	2	3	4	5	6
1	0.5266	0.1339	0.1072	0.1518	0.0447	0.0358
2	0.0000	0.6605	0.1072	0.0000	0.1965	0.0358
3	0.0000	0.0000	0.7677	0.0000	0.0000	0.2323
4	0.0981	0.0785	0.0719	0.4705	0.2026	0.0785
5	0.0000	0.1765	0.0719	0.0000	0.6731	0.0785
6	0.0000	0.0000	0.2485	0.0000	0.0000	0.7515

Table 6.2: Transition Matrix for the action 'do not inspect'.

the transition matrix for the action 'do not inspect'.

Let us look at the transition matrix for the action 'inspect' separately for high and low load (only considering the 3 deterioration states). The first row of the transition matrix for the action 'inspect' represents the probability of no red and no yellow defects happening, given as $\bar{P}_{red}$ and $\bar{P}_{yellow}$, respectively, the probability of no red defects happening and some yellow defects happening (P_{yellow}), and the probability of at least one red defect happening (P_{red}), given as $(\bar{P}_{red} \cdot \bar{P}_{yellow}, \bar{P}_{red} \cdot P_{yellow}, P_{red})$. We use these parameters to generate the transition matrix for the action 'do not inspect'.

The transition probability for the action 'do not inspect' will be an upper triangular matrix, because no repairs are performed when no inspection happens. The first row is composed of $(\bar{P}_{red} \cdot \bar{P}_{yellow}, \bar{P}_{red} \cdot P_{yellow}, P_{red})$, which is the same as the transition matrix for the action 'inspect'. The second row is formed by $(0, \bar{P}_{red}, P_{red})$, because the system remains in state 2 if no red defects are generated, and moves to state 3 otherwise. Finally, the third row is formed as $(0, 0, 1)$. With the transition probability for the action 'do not inspect' determined for high and low load, we can rebuild the 6-state transition matrix through algebraic manipulations using the

3-state low load transition matrix, the 3-state high load transition matrix, and the 2-state load transition matrix. The matrices obtained are shown in Tables 6.1 and 6.2. These matrices are based on the data and may not represent the true tables. A structured approach on how to continuously update these matrices with the addition of incoming data, while still optimally choosing segments to inspect is presented in section 6.2.

When $a = 0$, $c(s, a)$ can be defined as the risk cost dependent on the number of defects existing at the beginning of the current period and the likelihood of new defects occurrence. This cost can be computed by multiplying the cost of a derailment by the probability that a derailment will occur given the number of defects present. Using Bayes' rule, the probability of a particular defect causing a derailment, $p\{\text{der.}|\text{def.}\}$, can be expressed as:

$$p\{\text{der.}|\text{def.}\} = \frac{p\{\text{def.}|\text{der.}\} \cdot p\{\text{der.}\}}{p\{\text{def.}\}}.$$

Given a derailment, the probability of its being caused by a certain defect, namely $p\{\text{def.}|\text{der.}\}$, is calculated by Liu et al. (2012). When comparing the data provided in Liu et al. (2012) with our data, we infer that the probability of a derailment being caused by a red defect is 22.6% and the probability of a derailment being caused by a yellow defect is 9.6%.

Furthermore, Anderson and Barkan (2005) derive the probability of a derailment, depending on the length of the train. On average, they report the probability of a train derailing as $p\{\text{der.}\} = 0.720 \cdot 10^{-3}$. Finally, the probability of defect occurrence $p\{\text{def.}\}$ is given by P_{red} and P_{yellow} that are obtained from the transition matrix for the action 'inspect'.

Therefore,

$$p\{\text{der.}|\text{red defect}\} = \frac{22.6\% \cdot 0.720 \cdot 10^{-3}}{P_{red}},$$

$$p\{\text{der.}|\text{yellow defect}\} = \frac{9.6\% \cdot 0.720 \cdot 10^{-3}}{P_{yellow}}.$$

The cost of taking action 'do not inspect', $c(s,0)$, is just the probability of at least one of the defects causing a derailment. It can be obtained from the equation below, where c_{der} represents the expected derailment cost, and R and Y are random variables denoting the number of red and yellow defects present, respectively:

$$c(s,0) = c_{der} \cdot \sum_{r,y} \left[1 - \Big((1 - p\{\text{der.}|\text{red def.}\})^r \cdot (1 - p\{\text{der.}|\text{yel. def.}\})^y\Big)\right] p\{R = r, Y = y|s\}.$$

In this study, the joint distribution conditioned on the current state, $p\{R = r, Y = y|s\} \ \forall \ r, y$, is obtained empirically from the data. Finally, we multiply the cost by a factor of 1.6 for the states containing high load, accounting for the higher impact a derailment could bring, i.e. $c(i,H,0) = 2c(i,L,0)$ for $i = 1,2,3$.

We assume that other costs, such as damage to the image of the company, or lives lost are included in the cost of a derailment. Such costs can be adapted to account for different risk levels.

Furthermore, when $a = 1$, $c(s,a)$ includes the inspection and repair costs. For purposes of this research, we assume the inspection cost to be linearly dependent on the inspection length, and the repair cost to be an increasing function of the expected number of defects. Hence, it is also increasing in the state.

$$c(s,1) = E[\$ \text{ insp.}] \quad + \sum_{i=\{red,yel.\}} E[\#\ \text{def.}_i|s] \cdot E[\$ \text{ rep./def.}_i].$$

In this equation, the cost of inspection accounts for crew time, equipment used, and delays caused, and the cost of repair accounts for parts, manpower, and delay costs. The expected number of defects can be estimated from the regression models described in Section 5. Each segment may have specific cost parameters, and should, therefore, have a segment-specific policy to optimize the scheduling as well.

The solution for this Markov decision process (MDP) obtained from the dynamic programming equation provides the optimal action for every state. However, MDPs

assume the knowledge of the current state to be able to forecast the future states. Consequently, the "state" definition should include all the information one needs to forecast the next state. In the case of segments that were not inspected during the previous period, the decision maker faces the problem of not being able to observe the current state. But, in the inspection problem, the state information is available with a time delay, since each segment is observed within a finite time interval.

In order to provide the full state information, we augment the state space by the number of days since last inspection (Bertsekas, 2005). If there was an inspection yesterday, the inspection revealed red defects, and the load observed was High, then the augmented state becomes $\mathbf{s} = (3, H, 0)$. We limit the time delay to at most 9 days with no inspection, since that is the maximum delay we have observed in the data. The new state space is, therefore, the set of all combinations of deterioration levels (1,2 or 3), load information (low or high), and days since the state was last observed (0-9), i.e., $\mathbf{s} = (i, j, k)$, for $i = 1, 2, 3,\ j = H$ or L, and $k = 0, 1, \ldots, 9$. The cardinality of the state space, $|\mathcal{S}|$, is $3 \cdot 2 \cdot 10 = 60$. Then, the new transition matrices are,

State	$(\cdot,\cdot,0)$	$(\cdot,\cdot,1)$	$(\cdot,\cdot,2)$	$\cdots$	$(\cdot,\cdot,8)$	$(\cdot,\cdot,9)$
$(\cdot,\cdot,0)$	P_{insp}	0	0	$\cdots$	0	0
$(\cdot,\cdot,1)$	$P_{insp} \cdot P_{not}$	0	0	$\cdots$	0	0
$(\cdot,\cdot,2)$	$P_{insp} \cdot P_{not}^2$	0	0	$\cdots$	0	0
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$(\cdot,\cdot,8)$	$P_{insp} \cdot P_{not}^8$	0	0	$\cdots$	0	0
$(\cdot,\cdot,9)$	$P_{insp} \cdot P_{not}^9$	0	0	$\cdots$	0	0

Table 6.3: Augmented P-Matrix for the action 'inspect',

State	$(\cdot,\cdot,0)$	$(\cdot,\cdot,1)$	$(\cdot,\cdot,2)$	$\cdots$	$(\cdot,\cdot,8)$	$(\cdot,\cdot,9)$
$(\cdot,\cdot,0)$	0	P_{not}	0	$\cdots$	0	0
$(\cdot,\cdot,1)$	0	0	P_{not}^2	$\cdots$	0	0
$(\cdot,\cdot,2)$	0	0	0	$\cdots$	0	0
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$(\cdot,\cdot,8)$	0	0	0	$\cdots$	0	P_{not}^9
$(\cdot,\cdot,9)$	0	0	0	$\cdots$	0	P_{bound}

Table 6.4: Augmented P-Matrix for the action 'do not inspect'.

where P_{insp} and P_{not} are the transition matrices displayed in Tables 6.3 and 6.4. P_{bound} is the boundary matrix with all states transitioning to state 6 with probability 1 ($[\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{1}]_{(6x6)}$). The new cost vectors also need to be updated. The cost of the action 'do not inspect' for state s, is the same as before, but accounts for the expected next state. For instance, if the cost vector of not inspecting states $s \in \{1,2,\ldots,6\}$, where s contains the deterioration and load information is $\mathbf{c_{not}} = (c(1,0), c(2,0), \ldots, c(6,0))^T$, the cost vector associated with $(\mathbf{s},t)$, where t represents the time since the last inspection, is $P_{not}^t \cdot \mathbf{c_{not}}$. The extended cost of inspecting is constructed similarly across all states $(\mathbf{s},t)$.

6.1.1 Optimal Policy

The solution to equation 6.2 reveals that a threshold-type policy is optimal. In fact, the results inform the decision maker how many days should elapse before inspecting in case the current state has an optimal action of 'do not inspect'. We present two computational examples; the first assumes costs taken from the literature, and the latter uses a higher inspection cost to exemplify how a more diverse policy may arise.

The first experiment is performed for a segment assuming that the cost of a derailment is \$1.5M, and the cost of an inspection is \$1500 (assuming a 10km inspection at \$150/km (Soleimanmeigouni et al., 2016; Transportation Economics and Management Systems Inc, 2018)). Repairs are assumed to cost \$1500 for red defects and \$1300 for yellow defects (He et al., 2015). All red defects require immediate repair. Analyzing the data, we have that both red defects and some yellow defects require immediate repair, while 80% of yellow defects that are categorized as superficial may not. Hence, we assume that only more severe yellow defects will be repaired.

In the first setting (with the inspection cost of \$150/km), the optimal MDP solution is to inspect the segment right away, no matter in which state the segment currently resides. This suggests our data contains segments that are prone to defects. However, we know that it may not always be feasible for companies to inspect every segment daily. In the case of limitations such as that, we provide an approach for the constrained case in the next section.

For the second experiment, we increase the inspection cost to \$5000 (\$500/km) so that we can visualize a more diverse optimal policy. In this scenario, the threshold type policy prescribes an inspection rule in such a way that it is optimal to inspect every time the segment has red defects under high load; to wait 1 day when the system has i) red defects and is under low load, or ii) yellow or no defects and is under high load; and to wait 3 days when the system has yellow or no defects and is under low load (see Figure 6.1). It should be noted that each country or railway company may have different regulations on the inspection frequency. In case of regulation violations, the scheduling problem becomes a constrained problem that could be solved using the Linear Programming (LP) formulation of discounted MDPs given in Kallenberg (1989).

The results are compared with a benchmark policy to measure the efficiency of

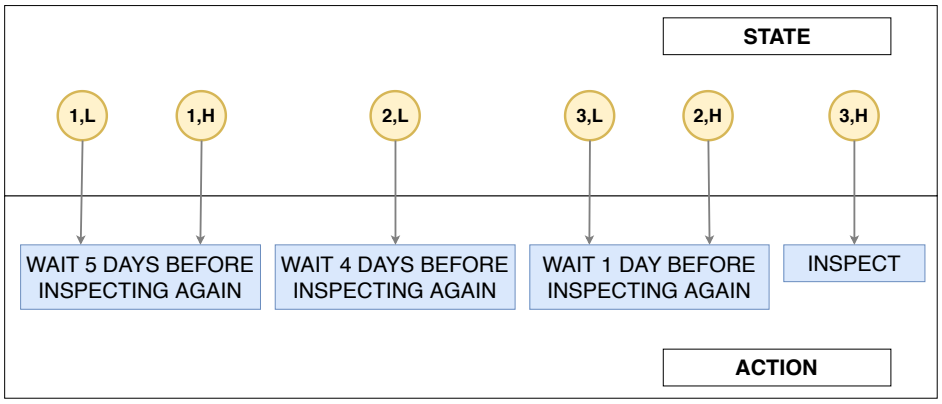
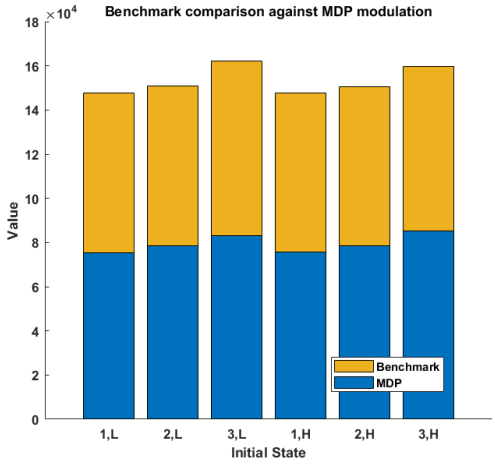


Figure 6.1: Inspection Scheduling Policy Representation from costs described in section 6.1.1 and $\alpha = 0.95$.

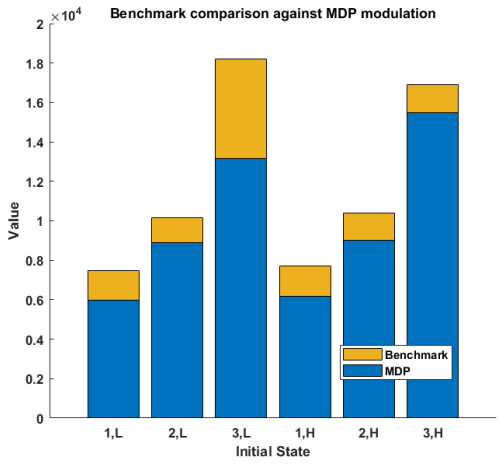
the policy obtained by MDPs and are demonstrated in Figures 6.2a and 6.2c. We construct our benchmark policy based on the inspection schedule guidelines provided with the dataset. Inspections are performed twice a week on major segments and once a week on the other segments. Therefore, for the benchmark policy, we define a 7-day wait for states with low load (1, 2 and 3), and a 4-day wait for states with high load (4, 5, and 6). Note that neither the number nor the type of defects found previously have an impact on the policy for the benchmark policy.

Expected gains are on the magnitude of the several of thousands of dollars per year for each track segment, thus, showcasing the importance of reviewing how scheduling is currently being assessed. Percent-wise, the benchmark policy is approximately 100% and 23% more expensive than the policy resulting from the MDP approach for the respective cases in which inspection cost is set to \$150/km and \$500/km.

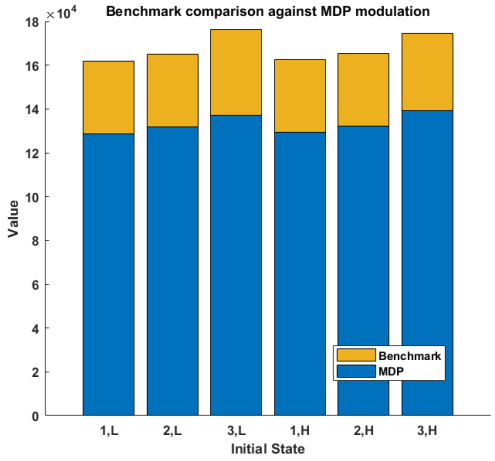
Figures 6.2b and 6.2d depict the results when the discount factor is set to $\alpha = 0.5$. It represents the short term savings by shifting to the MDP policy from the benchmark policy. As expected, the values are a lot more dependent on the initial state, as the discount factor has a higher impact on the convergence. The benchmark costs are approximately between 12% and 35% more expensive than the MDP policy in the



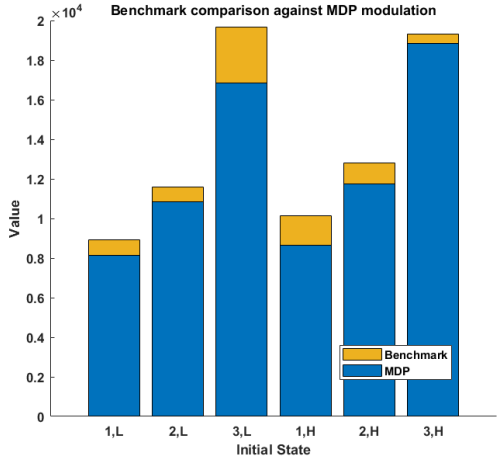
(a) Inspection cost=\$150/km, $\alpha = 0.95$



(b) Inspection cost=\$150/km, $\alpha = 0.5$



(c) Inspection cost=\$500/km, $\alpha = 0.95$



(d) Inspection cost=\$500/km, $\alpha = 0.5$

Figure 6.2: Values using the Benchmark Policy and the Optimal Values Obtained from the MDP Formulation.

scenario considering the inspection cost to be \$150/km and between 2% and 18% in the scenario for which the inspection cost is set to \$500/km .

6.2 Segments as Restless Bandits

The results obtained in Section 6.1 provide decision makers with a detailed policy they should follow when deciding whether to inspect or not to inspect a segment. However, crew limitations may cause the policy to be inadmissible. In this case, the problem can be formulated as the following constrained mathematical program

$$\lim_{N\to\infty} \min\left\{ E\left[\sum_{t=0}^{N} \alpha^t \sum_{i=1}^{n} c_i(s_i(t), a_i(t)) \right] \right\}$$
$$\text{s.t.} \sum_{i=1}^{n} a_i(t) \leq m, t \in \{1, 2, ..., N\}$$
$$a_i \in \{0, 1\} \forall i,$$

where $s_i(t)$ denotes the state of the ith segment at time t, $a_i(t)$ denotes the binary variable representing if segment i is inspected or not at time t, and $c_i(\cdot, \cdot)$ denotes the expected cost incurred at segment i when the state and action information are given.

Bandit problems are mathematical models to optimally allocate limited efforts in various competing projects so that maximum reward or minimum cost is achieved under uncertainty (Gittins et al., 2011). Originally, the problem assumes no change of state and zero rewards or costs when the passive action is taken. In the inspection scheduling problem, this is not the case. A variation of the original bandit problem, called the restless bandit problem, allows for such evolution of state and passively received reward or cost (Gittins et al., 2011). Furthermore, this framework continuously updates the transition matrices, maintaining and improving the system robustness.

Assume there are n segments to be inspected, and each segment has an initial augmented state $s_i \in \mathcal{S}$, with $|\mathcal{S}| = 60$. Each state represents the level of deterioration of the rail segment at the beginning of the period and the time passed since the last inspection, as described in Section 6.1. Let us further assume that there are m crews available to inspect n segments. Other parameters include $\mathbf{a}$, a vector of actions

for all states, and $\alpha \in [0,1]$, the discount factor. Finally, let us assume the initial probability transition matrices for the actions 'inspect' and 'do not inspect' are the ones obtained in section 6.1. If m, the number of crews available, is larger than n, all segments can be inspected and the problem becomes unconstrained. On the other hand, if $m < n$, this problem can be characterized as a restless bandit problem.

The restless bandit formulation chooses the best n of m competing segments optimally by using a priority index called Whittle index if the problem is indexable. We now prove the indexability of this scheduling problem.

The constraint on the crew number can be rewritten as

$$\sum_{t=0}^{\infty} \alpha^t \sum_{i=1}^{n} a_i(t) \leq \frac{m}{1-\alpha}, t \in \{1, 2, ..., N\}.$$

Relaxing the activation constraint, we write the Lagrangian dual function as

$$\begin{aligned}
\mathcal{L}(\lambda) &= \min_{a_i} \left\{ E\left[\sum_{t=0}^{\infty} \alpha^t \sum_{i=1}^{n} c_i(s_i(t), a_i(t)) + \lambda \cdot \left(\sum_{t=0}^{\infty} \alpha^t \sum_{i=1}^{n} a_i(t) - \frac{m}{1-\alpha} \right) \right] \right\} \\
&= \min_{a_i} \left\{ E\left[\sum_{t=0}^{\infty} \alpha^t \sum_{i=1}^{n} \left(c_i(s_i(t), a_i(t)) + \lambda a_i(t) \right) \right] \right\} - \lambda \left(\frac{m}{1-\alpha} \right) \\
&= \min_{a_i} \left\{ \sum_{t=0}^{\infty} \alpha^t \cdot E\left[\sum_{i=1}^{n} \left(c_i(s_i(t), a_i(t)]) + \lambda a_i(t) \right) \right] \right\} - \lambda \left(\frac{m}{1-\alpha} \right) \\
&\quad \text{s.t. } a_i \in \{0,1\} \forall i.
\end{aligned}$$

This problem can be decoupled for each segment, disregarding the last term, which is a constant.

$$\begin{aligned}
\mathcal{C}_i(\lambda) = \min \left\{ E\left[\sum_{t=0}^{\infty} \alpha^t (c_i(s_i(t), a_i(t)) + \lambda a_i(t)) \right] \right\} \\
\text{s.t. } a_i \in \{0,1\} \forall i.
\end{aligned}$$

The set of all states for which it is optimal to choose $a = 0$ when λ is fixed increases monotonically as λ increases. When $\lambda = 0$, all actions should be 'inspect' (based on the result from Section 6.1). As λ increases, there is a

point at which it is better not to inspect. Therefore, this problem is indexable (Gittins et al., 2011). The Whittle index is defined as

$$W_i(s_i) = \inf\left\{\lambda : E\left[c_i(s_i,0) + \sum_{t=1}^{\infty}\alpha^t(c_i(s_i(t),a_i(t)) + \lambda \cdot a_i(t))\right] < E\left[c_i(s_i,1) + \lambda + \sum_{t=1}^{\infty}\alpha^t(c_i(s_i(t),a_i(t)) + \lambda \cdot a_i(t))\right]\right\}. \quad (6.3)$$

Inspection decisions involve a risk-based approach that incorporates the trade-off between the risk cost and the repair cost. The Whittle index, a well-established measure for restless bandits, ranks the importance of this trade-off.

To compute the Whittle indices from equation 6.3, we solve several MDP for iteratively increasing λ. Because the cardinality of both the state and action spaces is finite, value-iteration or policy-iteration algorithms can be used to find the optimal cost function for each MDP efficiently (Bertsekas, 2007). After calculating an index for each segment in the candidate segments as identified in Section 6.1, and sorting them from the highest to lowest, the decision maker should pick the first m segments with the highest indices for inspection. Finally, the transition kernels are updated for these m segments using the maximum likelihood estimation.

6.2.1 Example

We provide a simplified example in which 5 segments from the same cluster are being considered for inspection, but only two crews are available. Of course, this methodology also applies to more general cases in which segments come from different clusters.

The cost parameters are assumed to be the same as the ones given in Section 6.1.1, but with minor perturbations to characterize the particularities of each segment. In

this scenario, inspections in segments B, and C cover 10 km of their length, while inspections in segments A, D, and E cover 7 km of their length at the rate of $150 per km inspected.

Segments have been inspected t days ago and their last observed state is described in Table 6.5. Suppose that the initial belief for the transition matrices of this cluster is that they follow the ones in Figures 6.1 and 6.2.

Segment	A	B	C	D	E
Derailment ($)	2M	2M	2M	1.5M	1.5M
Inspection ($)	1050	1500	1500	1050	1050
Red Repair ($)	1500	1500	1500	1400	1400
Yel. Repair ($)	1300	1200	1300	1300	1200
Last State	(3,L)	(3,H)	(2,L)	(2,H)	(1,L)
t	1	0	4	1	6

Table 6.5: Costs and state information for each segment.

Inspection policy

The calculated Whittle indices are presented in Table 6.6. Indices are calculated in increments of 100. Providing such a table allows decision makers to quickly assess which segments should be prioritized in a situation where only a limited number of crews is available. The decision maker should choose the segments with highest indices.

Table 6.6 suggests that inspecting segments B and E is the best decision in this scenario. The new states for segments B and E are observed, and t is set to 0. States for segments A, C, and D are updated to (3,L,2), (2,L,5), and (2,H,2), respectively.

Segment	Augmented State	Whittle Index
A	(3,L,1)	9600
B	(3,H,0)	10100
C	(2,L,4)	2100
D	(2,H,1)	2500
E	(1,L,6)	10200

Table 6.6: Whittle Indices for each segment.

Updating transition matrices

Lastly, the transitions matrices for segments B and E need to be updated, after the inspection is finished and the new state is known. Suppose the new deterioration and load state after the inspection is 5 and 2, respectively. Since the new t is 0, given that they have just been inspected, we will update the frequency of the corresponding transition. The corresponding augmented states are 5 (2,H,0), and 2 (2,L,0).

With two new inspections, we can update the frequencies of the augmented transition matrices. We normalize the matrices by the total number of inspections, including the last inspections. Finally, we compute the updated probabilities for the transitions matrices following the procedure in Section 6.1, but directly using the augmented matrices instead (using MLE on the updated frequencies). These new transition matrices are now the updated ones for the corresponding cluster. Due to the strong law of large numbers, they will tend to the true ones as more and more data are accumulated.

Chapter 7

Local Maintenance and Inspection Scheduling via Search Games

This chapter introduces search games in which Nature plays the role of the Hider, hiding defects in tracks, and an inspection crew plays the role of the Searcher, trying to find as many of the defects as possible. The problem is discretized by dividing the section the segment into smaller sub-segments. Each subsegment is assumed to be small enough so that at most one defect can be hidden in it. Furthermore, searches in subsegments containing defects always find the hidden defect.

1. Games in which both players know the number of defects. The Searcher does not decide on how many subsegments to inspect, but only which segments to inspect subject to constraints.

   (a) **Length Finding Game (LFC):** This game is the only continuous one. The objective for the Searcher is to find the most defects, given a fixed inspection length.

   (b) **LFC with Inspection Length Choice:** The Searcher aims to minimize the sum of the cost of inspecting and the penalty paid for defects not found.

The cost to inspect each sub-segment is the same, and so is the penalty cost for defects not found.

(c) **Finding all objects hidden in multiple Locations:** The main difference is that each subsegment has a different inspection cost. The objective for the Searcher is to find all hidden defects while spending as little in inspection costs as possible.

2. Games in which the searcher chooses which subsegments to inspect.

(a) **Deciding how many and which subsegments to inspect:** The main difference is that each subsegment has a different inspection cost. The objective for the Searcher is to decide how many and which segments to inspect. (solved 2x2)

(b) **Searching for Objects in Multiple locations with Inspection Length Choice:** In this scenario, the Hider chooses how many defects to hide, and such number remains unknown to the Searcher. The Searcher pays for subsegments inspected not containing defects, as well as for subsegments not inspected containing defects. The objective for the Searcher is to minimize the total costs of inspecting defect-free subsegments and not inspecting subsegments with defects.

7.1 Constrained Decision Set for the Searcher

7.1.1 Length Finding Game

The length finding game is a simple game that can be used by a decision-maker that has very little information about the segment to be inspected. It is also a starting point for more detailed games. At the end of this section, we include some additional

remarks.

In this opening game, we assume the track to be a segment. The Hider and the Searcher know how many defects, k, will be hidden, and the fraction of track α to be inspected is fixed. Since the goal of the Searcher is to maximize the number of defects found given the constraints, he pays a cost, c_1, for each defect not found (outside of the searched area). This game is a zero-sum game.

Ruckle (1983) proposes a similar game, but with the payoff obtained by the Searcher being 1 if all defects are found, and 0 otherwise. This game is called the Lenght Hiding Game (LDG)

Proposition 3. *From Ruckle (1983) (text adapted): "The value of the LHG game for the searcher is* $(1-\alpha)^k$*. An optimal strategy for the Hider is to choose k points in the segment independently by a uniform distribution on the segment. For each* $\epsilon > 0$*, the Searcher has the following* ϵ*-optimal strategy: choose n so large that (by the Law of Large Numbers)*

$$\frac{\binom{n-k}{\lfloor \alpha \cdot n \rfloor + 1}}{\binom{k}{\lfloor \alpha \cdot n \rfloor + 1}} \geq (1-\alpha)^k - \epsilon.$$

Divide the segment into n subintervals $[(i-1)/n, i/n) : i = 1, 2, 3, ..., n$*; and search one of the* $\binom{n}{\lfloor \alpha \cdot n \rfloor + 1}$ *unions of* $\lfloor \alpha \cdot n \rfloor + 1$ *with probability* $\binom{n}{\lfloor \alpha \cdot n \rfloor + 1}^{-1}$*."*

We use this proposition as a base for the LFG game solution.

Proposition 4. *The value of the Length Finding Game is* $\nu = c_1 k(1-\alpha)$*, and the optimal strategy for the Hider is to choose k points uniformly at random to hide the defects. The optimal strategy for the Searcher is to divide the segment into n intervals, with n small enough that the probability of more than one defect hidden in a particular interval is negligible, and then choose* $\alpha \cdot n$ *segments uniformly at random.*

Proof. Let the two players, Hider and Searcher, play this game on the segment $I = [0, 1]$. $\mathcal{H}$ consists of all subsets of I with exactly k points; $\mathcal{S}$ consists of all subsets of I with a total length exactly $\alpha < 1$. The Hider chooses a set $H \in \mathcal{H}$ and the searcher chooses a set $S \in \mathcal{S}$. The payoff to Hider is $S \cap H$ or the number of points chosen by Searcher located within the subset chosen by Hider.

Assuming an optimal strategy for Hider is to place the k points in I independently by a uniform distribution on I. The Searcher chooses an arbitrary strategy. The value of the game for the Hider is, then, given by the expected value of the number of defects not found within the subset chosen by Hider. Each defect has a probability of $1 - \alpha$ of not being found. Clearly, the Hider wants to maximize such payoff. The payoff for the Hider is:

$$V_{hider} = c_1 \sum_{m=0}^{k} m \binom{k}{m} (1-\alpha)^m \alpha^{k-m} = c_1 k(1-\alpha).$$

On the Searcher's side, consider the uniform strategy described in Proposition 3. Let us calculate the expected payoff against an arbitrary pure strategy of Nature, the Hider. A given defect j gets chosen, giving a reward of c_1 if and only if its location $f(j)$ is not in the Searcher's chosen subset. This has probability $\frac{n-\lfloor \alpha \cdot n \rfloor + 1}{n}$ for the first defect found. As more defects are found, the chances of finding defects go down; alternatively, when no defects are found, the chances of finding go up. This system is a binomial without replacement. If we choose n, so that $\alpha \cdot n$ is an integer, the probability of not finding the first defect becomes $1 - \alpha$. Therefore, the probability of matching a certain number of locations also follows a hyper-geometric distribution. The probability that the Searcher

finds m defects out of the k placed by the Hider is given by:

$$P_m = \frac{\binom{k}{m}\binom{n-k}{\lfloor \alpha n \rfloor + 1 - m}}{\binom{n}{\lfloor \alpha n \rfloor + 1}}.$$

The payoff is, therefore, given by $c_1 \cdot (k-m) \cdot P_m$ with probability P_m, for $m = \{0, 1, 2, 3, ..., k\}$ and with expected value given by:

$$V_{searcher} = c_1 \cdot k \cdot \frac{n - \lfloor \alpha \cdot n \rfloor - 1}{n} = c_1 k(1-\alpha),$$

when $\alpha \cdot n$ is an integer. □

Remark. *Although the mean depicted above is the same one as the binomial strategy played by the Hider, its variance is smaller, given that it comes from a hyper-geometric distribution.*

Corollary 1. *If we loosen the constraint of having at most one defect per subset, the expectation remains as the sum of the indicator functions for each point found. This probability is* $(1-\alpha)$*, yielding a value of the game equal to the one we proved:* $V_{game} = k(1-\alpha)c_1$*. However, the distribution will not be binomial or hypergeometric; it will depend on the strategies chosen by the Hider.*

7.1.2 Length Finding Game with Inspection Length Choice

For many scenarios, the inspection crew is allowed to choose how much of the track should be inspected. Clearly, the more the track is inspected, the more expensive it is for the company to pay for the crew. We can assume initially that the increment in cost is linear to the increment in the length of track inspected.

Let us now look at the game described in section 7.1.1 but letting the Searcher choose α. This is a trivial game if no cost is associated with α because he would choose $\alpha = 1$ for any scenario. We, therefore, add a cost to how long α is. Logically,

the Searcher wants to minimize the cost of the chosen α plus the costs associated with the points not found.

Proposition 5. *The optimal strategy for the searcher is to choose $\alpha = 0$ if $c_2 > c_1 \cdot k$, and $\alpha = 1$, otherwise. The hider can choose any strategy.*

Proof. Following the same strategies as the LFG, when the Hider plays the uniform strategy, the payoff for the Searcher will then be

$$V_{searcher} = c_1 k(1-\alpha) + c_2\alpha = c_1 k + \alpha(c_2 - c_1 k),$$

where c_1 represents the cost of each non-found point and c_2 represents the cost of the length α chosen. Remember the searcher wants to minimize such payoff. If we assume c_2 to be a constant, therefore making the increase in cost to α linear, it is trivial to see that the only two possible options for α are 0 or 1, depending on the relationship $(c_2 - c_1 \cdot k)$. If $c_2 > c_1 \cdot k$, then $\alpha = 0$. Otherwise, $\alpha = 1$.

□

7.1.3 Finding All Objects Hidden in Multiple Locations

Another consideration the decision-maker might want to include in the game is that some locations are more likely to have defects than others. Possible causes include load that passes by a section of the track, temperatures that a track section is subject to, the age of a segment track section, among many others.

Let us look at a similar game to the LFG. In this scenario, Nature has an incentive to place defects in specific locations at the expense of others. Therefore, we will add different costs for the Searcher to inspect different subsections of I. In this problem, we assume the game only ends when the inspection crew finds all defects.

Let $\mathbf{c}$ be a vector $[n \times 1]$ of costs for each i subsection of an arbitrary n total subsections of same size s in the line $I = [0, 1]$. The size of each subsection is assumed to be small enough that no more than one defect may arise there. Therefore, $s \cdot n = 1$.

We can now look at this problem as a discrete search game, in which the Hider hides k points among the n subsections. The Searcher wants to minimize the total cost paid to find all points. There are two approaches to solving this game. The first consists of developing a matrix version of the game, where each strategy is a combination of k out of the n subsections. The latter was proposed by Lidbetter (2013), who proved the following theorem:

Theorem 2. *An optimal strategy for the Searcher is to choose k subsegments that he will inspect initially, given that he will need to inspect at least k subsegments if he wants to find k defects. If there are still missing defects after these initial inspections, he inspects further subsegments randomly. The Hider hides defects with the same strategy the Searcher chooses the initial k subsegments.*

Matrix representation of the game

The payoff for the Hider will be the sum of the cost for the chosen k subsections he initially opens, the cost of the subsegments containing defects, and the cost of remaining subsegments being inspected times the probability he inspects it before the game ends.

Let $\mathcal{N}$ be the set $[1, 2, 3, ..., n]$ of all n subsegments from I. Let us assume the searcher chooses a k-subset H from the set $\mathcal{B}^{(k)}$ of all possible k-subsets of $\mathcal{N}$. Let $\Pi(H) = \sum_{i \in H} c_i$, which represents the cost of inspecting the k initial subsections.

Now, since we are inspecting the remaining subsegments randomly following a uniform distribution, the chance that the Searcher inspects a certain defect-free subsegment before he finds all defects is the same for all subsegments.

Let $f(r)$ be the probability of inspecting a certain subsection not containing a

defect before finding all remaining defects, where r denotes the number of remaining defects. r can be obtained as the cardinality of the set $A \setminus \{A \cap H\}$. Let us order subsegments so that the ones containing defects are first, and focus on the sequence of the first $r+1$ subsegments. The probability of not inspecting that segment without a defect is the probability that all segments with defects are inspected first. This probability is $\frac{r!}{(r+1)!} = \frac{1}{r+1}$. The probability of inspecting the subsegment without a defect is, then, the complement, given by $\frac{r}{r+1}$. Because we can arbitrarily order subsegments, this is the same probability for all subsegments not containing a defect.

Each cell of the matrix represents a pair of the strategy $H \in \mathcal{B}^{(k)}$ played by the Searcher, and a strategy $A \in \mathcal{B}^{(k)}$ played by the Hider. The cell values representing the payoffs for each pair of strategies will be:

$$C_{HA} = \Pi(H) + \Pi(A) - \Pi(A \cap H) + \sum_{i \notin H,A} c_i f(r).$$

For instance, if $n = 3$, $k = 2$, and $c = [1, 2, 3]$. Then $\mathcal{B}^{(k)} = \{(1,2), (1,3), (2,3)\}$. If $H_1 = \{2, 3\}$, then $\Pi(H_1) = 2+3 = 5$. If $A_1 = \{1, 2\}$, $\Pi(A_1) = 3$, and $\Pi(A_1 \cap H_1) = \Pi(\{2\}) = 2$. Therefore: $C_{H_1 A_1} = 5 + 3 - 2 = 6$. This makes sense, because the searcher will need to inspect all three subsegments to find all k defects.

With the matrix populated for every possible combination of H and $A \in \mathcal{B}^{(k)}$, we now have a zero-sum matrix game for which we can find the optimal equilibrium strategy for both players. One approach is to use linear programming (Young and Zamir, 2014). Assume the set $\mathcal{B}^{(k)}$ has cardinality ϕ. The payoff matrix can be represented as follows:

Now, we know that the Searcher wants to minimize his payoff ν. Assume the Hider plays a certain pure strategy A_n. We know the payoff for the searcher in this case will be given by $\sum_{i=1}^{\phi} P_{Hi} C_{H_i A_n}$. Additionally, we also know he wants to pay as minimum as possible, even though he knows the Hider will force him to pay as much as possible. Therefore, he wants to find the minimum-maximum he will have to play.

		Hider		
		A_1	...	A_ϕ
Searcher	H_1	$C_{H_1A_1}$	...	$C_{H_1A_\phi}$
	...	...	...	...
	H_ϕ	$C_{H_\phi A_1}$	...	$C_{H_\phi A_\phi}$

We can then formulate the problem as

$$\begin{aligned}
&\min \nu \\
&\text{s.t.} \sum_{i=1}^{\phi} P_{Hi} C_{H_i A_n} \leq \nu, \forall n = 1, 2, \ldots, \phi \\
&\sum_{i=1}^{\phi} P_{Hi} = 1, \\
&P_{Hi} \geq 0, i = 1, 2, 3, ..., \phi, \nu \text{ urs.}
\end{aligned}$$

In a parallel way, the hider wants to maximize the payoff, knowing that the searcher will minimize it. In a sense, he wants to maximize the minimum payoff paid by the searcher. It turns out the linear programming model for the hider is the Dual of the problem for the searcher:

$$\begin{aligned}
&\max \omega \\
&\text{s.t.} \sum_{i=1}^{\phi} P_{Ai} C_{H_n A_i} \geq \omega, \forall n = 1, 2, \ldots, \phi \\
&\sum_{i=1}^{\phi} P_{Ai} = 1, \\
&P_{Ai} \geq 0, i = 1, 2, 3, ..., \phi, \omega \text{ urs.}
\end{aligned}$$

Clearly, by duality, the payoff for both the Hider and the Searcher is the same, $\nu = \omega$.

Closed-form solution

An alternative solution proposed by Lidbetter (2013) is a closed-form solution for the strategies of both players. The advantage of this method is that it avoids the exponential increase of the matrix size when the number of subsegments increases. The payoff of the game is the same obtained via the method using linear programming to solve the matrix representation of the game. Lidbetter (2013) was the first to solve this problem, and he shows that an optimal hider choice is the distribution for k-subsets that is proportional to $\pi(H) := \prod_{i \in H} c_i$.

7.2 Unconstrained Decision Set for the Searcher

This section loosens the constraint that the Searcher is bounded by the number of subsegments he needs to inspect. At first, we still keep the hider set of strategies constrained, but then we also relax this constraint for the second game in this section.

7.2.1 Searching for Known Number of Defects with Inspect Length Choice

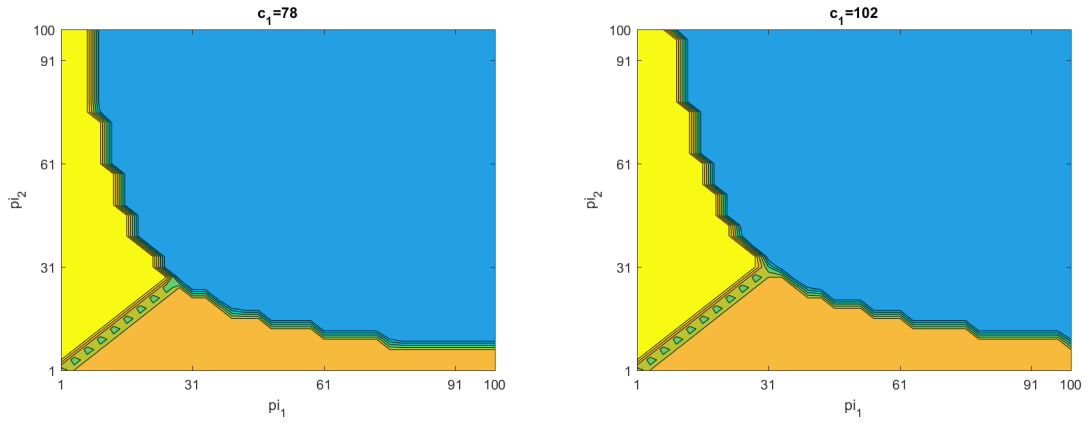
This version assumes both players know the number of defects to be hidden. Therefore, the Hider is limited to how many segments will be included in her strategy. On the other hand, the Searcher is allowed to choose to inspect however many subsegments he wishes. Again, for each subsegment inspected containing a defect, the cost is zero; for each subsegment inspected not containing a defect, the cost is c_i; and, for each subsegment containing a defect that is not inspected, the Searcher pays π_i.

At first, let us analyze the case for two subsegments with one hidden defect. The matrix is then (Hider in columns and Searcher in rows):

In this game, one can see that the set of possible strategies for the Searcher has different cardinality than the one for the Hider. For instance, in the case where

	(1)	(2)
$\emptyset$	π_1	π_2
(1)	0	$c_1 + \pi_2$
(2)	$c_2 + \pi_1$	0
(1,2)	c_2	c_1

there are two subsegments and one defect to be hidden, the Hider can choose to mix from the set $\mathcal{S}_h = \{(1), (2)\}$ and the Searcher can choose to mix from the set $\mathcal{S}_s = \{\emptyset, (1), (2), (1, 2)\}$. In a sense, the Searcher commits to what boxes he opens before playing, no matter what happens, making this a non-sequential game. Figure 7.1 represents how the optimal set of strategies change according to the choices of c_i s and π_is. From this representation, we conjecture that the boundary is given by the equations $c_1c_2 = \pi_1\pi_2$ and $\pi_1 = \pi_2$.



(a) $c_1 = 78$ (b) $c_1 = 102$.

Figure 7.1: Regions for which the Set of Optimal Strategies for the Searcher Remains Optimal when $c_2 = 10$.

From this conjecture, we can prove the solution for the game.

Proposition 6. *Claim: We can to subdivide this game in four cases:*

1. $\pi_1\pi_2 \leq c_1c_2$, $\pi_1 \leq \pi_2$

The hider should play strategy y where he places the item in box (1) with probability $\frac{\pi_2}{\pi_2+c_2}$ *and in box (2) with probability* $\frac{c_2}{\pi_2+c_2}$*, and for the searcher to play strategy x where he mixes between* $\emptyset$ *with probability* $\frac{c_2+\pi_1}{\pi_2+c_2}$ *and (2) with probability* $\frac{\pi_2-\pi_1}{c_2+\pi_2}$*.*

2. $\pi_1\pi_2 \leq c_1c_2$*,* $\pi_2 \leq \pi_1$

The hider should play strategy y where he places the item in box (1) with probability $\frac{c_1}{\pi_1+c_1}$ *and in box (2) with probability* $\frac{\pi_1}{\pi_1+c_1}$*, and for the searcher to play strategy x where he mixes between* $\emptyset$ *with probability* $\frac{c_1+\pi_2}{\pi_1+c_1}$ *and (1) with probability* $\frac{\pi_1-\pi_2}{c_1+\pi_1}$*.*

3. $\pi_1\pi_2 \geq c_1c_2$*,* $c_2 \geq c_1$

The hider should play strategy y where he places the item in box (1) with probability $\frac{\pi_2}{\pi_2+c_2}$ *and in box (2) with probability* $\frac{c_2}{\pi_2+c_2}$*, and for the searcher to play strategy x where he mixes between* (1) *with probability* $\frac{c_2-c_1}{\pi_2+c_2}$ *and (1,2) with probability* $\frac{\pi_2+c_1}{c_2+\pi_2}$*.*

4. $\pi_1\pi_2 \geq c_1c_2$*,* $c_2 \leq c_1$

The hider should play strategy y where he places the item in box (1) with probability $\frac{c_1}{\pi_1+c_1}$ *and in box (2) with probability* $\frac{\pi_1}{\pi_1+c_1}$*, and for the searcher to play strategy x where he mixes between* (2) *with probability* $\frac{c_2-c_1}{\pi_1+c_1}$ *and (1,2) with probability* $\frac{\pi_1-c_2}{c_1+\pi_1}$*.*

Proof of proposition 6 by conjecture. The strategy set for the Hider is $\{(1),(2)\}$. The strategy set for the Searcher is $\{(\emptyset),(1),(2),(1,2)\}$; In a sense, the Searcher commits to what boxes he opens before playing, no matter what happens, making this a non-sequential game.

We arbitrarily order the boxes so that $c_2 \geq c_1$. The matrix is then (Hider in columns and Searcher in rows):

		y_1	y_2
		(1)	(2)
x_1	$\emptyset$	π_1	π_2
x_2	(1)	0	$c_1 + \pi_2$
x_3	(2)	$c_2 + \pi_1$	0
x_4	(1,2)	c_2	c_1

Case 1. $\pi_1 \cdot \pi_2 \leq c_1 \cdot c_2$

Case 1.1. $\pi_1 \leq \pi_2$

Suppose that under the condition $\pi_1 \cdot \pi_2 \leq c_1 \cdot c_2$ and $\pi_1 \leq \pi_2$ the optimal strategy for the searcher is to mix between opening no boxes, or box (2). We will prove this is true by verifying that if the Searcher chose to play one of the other two pure strategies, the resulting expected payoff of at least v, where v is the payoff of the reduced game.

The reduced game can be displayed as:

	(1)	(2)
$\emptyset$	π_1	π_2
(2)	$c_2 + \pi_1$	0

By the difference trick, since $\pi_1 \leq \pi_2$, an optimal strategy y for the hider is to mix (1) and (2) with probabilities q and $(1 - q)$, where $q = \frac{\pi_2}{\pi_2 + c_2}$, and for the searcher to play strategy x where he mixes between $\emptyset$ and (2) with p and $1 - p$, where $p = \frac{c_2 + \pi_1}{\pi_2 + c_2}$.

The value of this game can be easily calculated by multiplying $(p, 1 - p) \cdot (\pi_2; 0)$, when the searcher plays x and the hider plays the pure strategy (2). This results in:

$$v = p\pi_2 = \frac{(c_2 + \pi_1)\pi_2}{\pi_2 + c_2}.$$

Now looking at the extended game, assume the searcher plays the pure strategy $(1, 2)$ when the hider plays strategy y, given by the mix described above. We will show that he ensures an expected payoff of at least v. The new payoff is:

$$v' = \frac{c_2\pi_2 + c_1c_2}{\pi_2 + c_2}.$$

When comparing:

$$\begin{aligned} v - v' &= \frac{(c_2 + \pi_1)\pi_2}{\pi_2 + c_2} - \frac{c_2\pi_2 + c_1c_2}{\pi_2 + c_2} \\ &= \frac{\pi_1\pi_2 - c_1c_2}{c_2 + \pi_2} \leq 0 \text{ since } (c_1c_2 \geq \pi_1\pi_2). \end{aligned}$$

Therefore, the Searcher has no incentive to ever play this strategy (1,2) under the conditions described as $v' \geq v$.

Now, looking at the payoff if the searcher plays the pure strategy (1). The new payoff is:

$$v' = \frac{c_2(c_1 + \pi_2)}{\pi_2 + c_2}.$$

This is the same payoff of playing strategy (1,2) purely. Therefore, as proven, $v' \geq v$.

Therefore, the Searcher has no incentive to ever play this strategy (1) under the conditions described as $v' \geq v$.

We now have proved that under the case $\pi_1 \cdot \pi_2 \leq c_1 \cdot c_2$ and $\pi_1 \leq \pi_2$, one of the searcher optimal strategies is to mix only between $\emptyset$ and (2).

Case 1.2. $\pi_2 \leq \pi_1$

Now looking at the case where $\pi_1 \cdot \pi_2 \leq c_1 \cdot c_2$ and $\pi_2 \leq \pi_1$, our claim is that an optimal solution for the searcher is to mix only between $\emptyset$ and (1). The proof is symmetric.

By the difference trick, since $\pi_2 \leq \pi_1$, an optimal strategy y for the hider is to mix (1) and (2) with probabilities q and $(1-q)$, where $q = \frac{c_1}{\pi_1 + c_1}$, and for the searcher to mix between $\emptyset$ and (1) with p and $1-p$, where $p = \frac{c_1 + \pi_2}{\pi_1 + c_1}$.

The value of the game is, calculated when the searcher plays x and the hider plays the pure strategy (1):

$$v = p\pi_1 = \frac{(c_1 + \pi_2)\pi_1}{(\pi_1 + c_1)}.$$

When choosing to play the pure strategy $(1, 2)$ against the strategy y of the hider, we obtain the payoff:

$$v' = \frac{c_1(c_2 + \pi_1)}{\pi_1 + c_1}.$$

When comparing:

$$\begin{aligned} v - v' &= \frac{(c_1 + \pi_2)\pi_1}{(\pi_1 + c_1)} - \frac{c_1(c_2 + \pi_1)}{\pi_1 + c_1} \\ &= \frac{\pi_1\pi_2 - c_1c_2}{c_1 + \pi_1} \leq 0 \text{ since } (c_1c_2 \geq \pi_1\pi_2). \end{aligned}$$

Finally, let us look at the payoff when the Searcher plays the pure strategy (2) against y.

$$v' = \frac{c_1(c_2 + \pi_1)}{\pi_1 + c_1}.$$

This is the same payoff of playing strategy (1,2) purely. Therefore, as proven, $v' \geq v$.

Case 2. $\pi_1 \cdot \pi_2 \geq c_1 \cdot c_2$

Without loss of generality, again we assume $c_2 \geq c_1$, as we can arbitrarily determine the order of the boxes. We claim the Searcher has an optimal strategy to mix solely between opening box (1) or opening both (1,2). The reduced game is displayed below:

	(1)	(2)
(1)	0	$c_1 + \pi_2$
(1,2)	c_2	c_1

By the difference trick, since $\pi_1 \leq \pi_2$, an optimal strategy y for the hider is to mix (1) and (2) with probabilities q and $(1-q)$, where $q = \frac{\pi_2}{\pi_2 + c_2}$, and for the searcher to play strategy x where he mixes between (1) and (1,2) with p and $1-p$, where $p = \frac{c_2 - c_1}{\pi_2 + c_2}$.

The value of this game can be easily calculated by multiplying $(p, 1-p) \cdot (0; c_2)$, when the searcher plays x and the hider plays the pure strategy (1). This results in:

$$v = (1-p)c_2 = \frac{(c_1 + \pi_2)c_2}{\pi_2 + c_2}.$$

Now looking at the extended game, assume the Searcher plays the pure strategy $\emptyset$ when the Hider plays strategy y, given by the mix described above. We will show that he ensures an expected payoff of at least v. The new payoff is:

$$v' = \frac{\pi_1\pi_2 + c_2\pi_2}{\pi_2 + c_2}.$$

When comparing:

$$\begin{aligned} v - v' &= \frac{(c_1 + \pi_2)c_2}{\pi_2 + c_2} - \frac{\pi_1\pi_2 + c_2\pi_2}{\pi_2 + c_2} \\ &= \frac{c_1c_2 - \pi_1\pi_2}{c_2 + \pi_2} \leq 0 \text{ since } (c_1c_2 \leq \pi_1\pi_2). \end{aligned}$$

Therefore, the Searcher has no incentive ever to play this strategy $\emptyset$ under the conditions described as $v' \geq v$.

Now, looking at the payoff if the searcher plays the pure strategy (2). The

new payoff is:

$$v' = \frac{\pi_2(c_2 + \pi_1)}{\pi_2 + c_2}.$$

This is the same payoff of playing strategy $\emptyset$ purely. Therefore, as proven, $v' \geq v$.

Therefore, the Searcher has no incentive to ever play this strategy (2) under the conditions described as $v' \geq v$.

We now have proved that under the case $\pi_1 \cdot \pi_2 \geq c_1 \cdot c_2$ and $\pi_1 \leq \pi_2$, one of the searcher optimal strategies is to mix only between (1) and $(1, 2)$.

By the rearranging principle, it is easy to see that, if $c_1 < c_2$, an optimal strategy for the Searcher is to mix between (2), and (1,2). □

In a more general setting, with n subsegments and k defects, a closed-form solution has not yet been found. As shown in Figure 7.2, the number of regions grows, and they do not follow as distinct functions as the two-subsegment case.

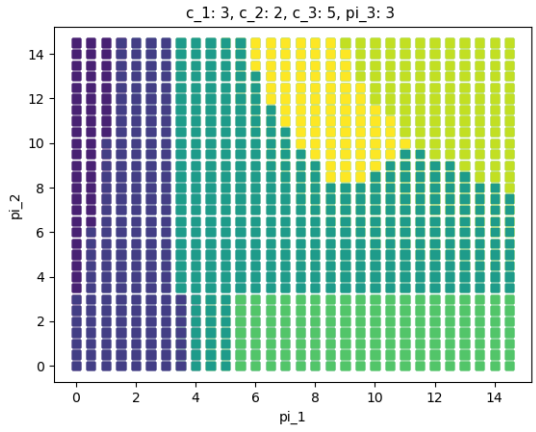


Figure 7.2: Regions for which the Set of Optimal Strategies for the Searcher Remains Optimal when $c_1 = 3$, $c_2 = 2$, $c_3 = 5$, and $\pi_3 = 3$.

For these cases, a LP framework can solve specific numerical games. This framework does not scale well to large number of defects and hiding locations, however.

7.2.2 Searching for Unknown Number of Defects with Inspect Length Choice

In this game, we relax the constraint of both players knowing the number of defects. The Hider has a choice of how many defects he can hide, being from no defects to n defects, given there are n subsegments of track available for hiding. As an incentive for the Searcher, he does not need to pay for subsegments inspected that contains defects but pays a fee for those inspected containing no defects, and a penalty for those subsegments containing defects, but that he decided not to inspect.

Theorem 3. *For a general n, let $p(A) := \prod_{j\in A} \frac{c_j}{\pi_j}$, with $p(\emptyset) = 1$. Then, an optimal hider strategy is to choose a subset $A \subset [n]$, where $[n] := \{1, 2, ..., n\}$, with probability p_A, given by*

$$p_A = \frac{p(A)}{S([n])},$$

where c_j is the cost of inspecting subsegment j and not finding a defect, π_j is the penalty cost of not finding a defect hidden in subsegment j, and $S(A) = \sum_{B\subset A} p(B)$.

For instance, if $n = 2$, then $[n] = \{1, 2\}$. If $A = \{1, 2\}$, then $p(A) = p(\{1, 2\}) = \frac{c_1}{\pi_1} \cdot \frac{c_2}{\pi_2}$, $S(\{1, 2\}) = p(\emptyset) + p(\{1\}) + p(\{2\}) + p(\{1, 2\})$.

On the opposite side, an optimal strategy for the searcher is to choose $B \subset [n]$ with probability $p_{\bar{B}}$.

The value of the game is

$$V_{game} = \sum_{j\in[n]} \frac{S([n] - \{j\})}{S([n])} \cdot c_j.$$

Proof. Assuming the Hider plays the strategy described. Then, the probability q_j that some subsegment $j \in [n]$ has a hidden defect is

$$q_j = \sum_{\{A \subset [n] : j \in A\}} p_A = \frac{c_j}{\pi_j} \cdot \frac{S([n] - \{j\})}{S([n])}.$$

The complementary probability $1 - q_j$ can also be written as

$$1 - q_j = \sum_{\{A \subset [n] : j \notin A\}} p_A = \frac{S([n] - \{j\})}{S([n])}.$$

Now let B be the subset of segments chosen by the searcher to inspect. If $j \in B$, then the payoff cost in incremented by $(1 - q_j)c_j$. On the other hand, if $j \notin B$, then the payoff cost is incremented by $q_j \pi_j$. Hence

$$\begin{aligned} V_{searcher} &= \sum_{\{j\} \in B} (1 - q_j)c_j + \sum_{\{j\} \notin B} q_j \pi_j \\ &= \sum_{\{j\} \in B} \frac{S([n] - \{j\})}{S([n])} \cdot c_j + \sum_{\{j\} \notin B} \frac{c_j}{\pi_j} \cdot \frac{S([n] - \{j\})}{S([n])} \cdot \pi_j \\ &= \sum_{\{j\} \in [n]} \frac{S([n] - \{j\})}{S([n])} c_j. \end{aligned}$$

□

The proof for the other side, fixing the strategy of the Searcher and verifying the payoff for the Hider, is symmetrical.

Remark. *This game can also be solved as a matrix game, in which row strategies are the ones chosen by the Hider, and column strategies are the ones chosen by the Searcher. For a small number of subsegments, this matrix remains solvable using the min-max theorem and LP. However, as the cardinality of the subsegment set increases, the matrix size increases exponentially.*

Chapter 8

Conclusion

This dissertation proposes novel methods to solve transportation decision problems. The two main focuses are in traffic congestion forecasting and railway maintenance scheduling.

In roadways, the main contributions obtained are the simplification of the solution for the distribution of traffic density described in Baykal-Gürsoy et al. (2009a), and the generalization of the model for longer segments of road. The equation developed by the literature is mathematically cumbersome and hard to implement. By analyzing valid approximations for roadway scenarios, we simplify their solution to a straightforward closed-form equation. In chapter 2 (Lopes Gerum et al., 2019b), we show this stationary distribution for the number of vehicles in the roadway in explicit form.

We validate the model using real traffic data for short segments of the road. We show that it generates accurate performance measures with explicit dependencies on traffic and incident parameters, avoiding the use of costly simulation. In chapter 3 (Lopes Gerum and Baykal-Gürsoy, 2020), we generalize the model for long segments of the road. We suggest that a sequence of random-queues can be modeled as a tandem-queue. We explicitly give the closed-form solution, derived from algebraic

manipulations of the product-form solution. With the same data, we validate the models and compute the drop in performance as the segment length increase. The results extend the validity of the solutions to much more complicated cases.

Current models in traffic management are expensive and complicated, yet the models proposed in this study are simple, intuitive, and use straightforward statistical parameters. This difference illustrates the disruptive potential our work has on the design and operation of management tools for roadway traffic and incident mitigation. With our models, decision-makers may no longer need to rely on expensive tools to forecast traffic.

In railways, understanding defect occurrence is paramount for risk assessment. Seminal work in the literature suggested that machine learning algorithms produce accurate predictions of defect generation. However, they use complex and obscure data that most companies do not have at hand. Chapter 5 (Lopes Gerum et al., 2019a) looks at defect generation under inspections subjected to random interarrivals. This randomness emerges from the dynamically changing budget that railway companies often face. In opposition to the literature, the proposed algorithm only requires reliable data that most companies have at hand. We determine that Random Forests provide similar accuracy to the literature with fewer and these simpler features. We also suggest an alternative interpretation of the results that accounts for the risk-aversion characteristic of the problem (it is a lot better to inspect one extra time than to leave a severe defect untreated).

With accurate predictions, we model the maintenance problem as a Markov decision process. Chapter 6 demonstrates that inspection and predictive maintenance has a threshold structure. Last, we consider this problem when there are not enough inspection/repair teams. In this case, we provide Whittle indices for service prioritization. Previous studies address these problems separately or only consider track

defects. We (Lopes Gerum et al., 2019a) are the first study that integrates defect prediction with the design of optimum inspection and maintenance strategies using real-world rail track defect and inspection data. The results indicate an improvement in all current estimates.

The last decision for optimal scheduling is determining which sub-section of the rail segment should be inspected by the crew. Walking inspections are often limited and cannot cover a full segment at a time. In chapter 7, we develop search games whose equilibria produce policies that minimize the worst-case scenario. Each game contains distinct rules, accommodating different user cases. For most games, we find closed-form equilibrium policies. For others, we provide a framework on how to find equilibria numerically.

Bibliography

Agarwal, S., Kachroo, P., and Regentova, E. (2016). A hybrid model using logistic regression and wavelet transformation to detect traffic incidents. *Iatss Research*, 40(1):56–63.

Alfa, A. and Neuts, M. (1995). Modelling vehicular traffic using the discrete time Markovian arrival process. *Transportation Science*, 29:109–117.

Anbaroglu, B., Heydecker, B., and Cheng, T. (2014). Spatio-temporal clustering for non-recurrent traffic congestion detection on urban road networks. *Transportation Research Part C: Emerging Technologies*, 48:47–65.

Anderson, R. T. and Barkan, C. P. (2005). Derailment probability analysis and modeling of mainline freight trains. In *Proceedings of 8th International Heavy Haul Railway Conference*, pages 88–98.

Anderson, T. W. and Goodman, L. A. (1957). Statistical inference about Markov chains. *The Annals of Mathematical Statistics*, pages 89–110.

Andrade, A. R. and Teixeira, P. F. (2016). Exploring different alert limit strategies in the maintenance of railway track geometry. *Journal of Transportation Engineering*, 142(9):04016037/1–9.

Arnesen, P. and Hjelkrem, O. A. (2018). An estimator for traffic breakdown probabil-

ity based on classification of transitional breakdown events. *Transportation Science*, 52(3):593–602.

Bartlett, M. S. (1951). The frequency goodness of fit test for probability chains. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 47, pages 86–95. Cambridge University Press.

Baykal-Gürsoy, M., Duan, Z., and Xu, H. (2009a). Stochastic models of traffic flow interrupted by incidents. *IFAC Proceedings Volumes*, 42(15):442–449.

Baykal-Gürsoy, M. and Xiao, W. (2004). Stochastic decomposition in M/M/∞ queues with Markov-modulated service rates. *Queueing Systems*, 48:75–88.

Baykal-Gürsoy, M., Xiao, W., and Ozbay, K. (2009b). Modeling traffic flow interrupted by incidents. *European Journal of Operational Research*, 195(1):127–138.

Bertsekas, D. P. (2005). *Dynamic programming and optimal control*, volume 1. Athena Scientific Belmont, Massachusetts.

Bertsekas, D. P. (2007). *Dynamic programming and optimal control*, volume 2. Athena Scientific Belmont, Massachusetts.

Bilmes, J. A. et al. (1998). A gentle tutorial of the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models. *International Computer Science Institute*, 4(510):126.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and Regression Trees*. Wadsworth.

Budai-Balke, G., Dekker, R., and Kaymak, U. (2009). Genetic and memetic algorithms for scheduling railway maintenance activities. *Econometric Institute, Erasmus University Rotterdam*, (EI 2009-30):1–23.

Camci, F. (2014). The travelling maintainer problem: integration of condition-based maintenance with the traveling salesman problem. *Journal of the Operational Research Society*, 65(9):1423–1436.

Cárdenas-Gallo, I., Sarmiento, C. A., Morales, G. A., Bolivar, M. A., and Akhavan-Tabatabaei, R. (2017). An ensemble classifier to predict track geometry degradation. *Reliability Engineering & System Safety*, 161:53–60.

Celikoglu, H. B. (2014). Dynamic classification of traffic flow patterns simulated by a switching multimode discrete cell transmission model. *IEEE Transactions on Intelligent Transportation Systems*, 15(6):2539–2550.

Celikoglu, H. B. and Silgu, M. A. (2016). Extension of traffic flow pattern dynamic classification by a macroscopic model using multivariate clustering. *Transportation Science*, 50(3):966–981.

Cheah, J. and Smith, J. (1994). Generalized M/G/C/C state dependent queuing models and pedestrian traffic flows. *Queueing Systems*, 15:365–385.

Chen, M., Yu, X., and Liu, Y. (2018). Pcnn: Deep convolutional networks for short-term traffic congestion prediction. *IEEE Transactions on Intelligent Transportation Systems*, 19(11):3550–3559.

Chen, S.-K., Ho, T.-K., Mao, B.-H., and Bai, Y. (2014). A bi-objective maintenance scheduling for power feeding substations in electrified railways. *Transportation Research Part C: Emerging Technologies*, 44:350 – 362.

Chen, Z., Liu, X. C., and Zhang, G. (2016). Non-recurrent congestion analysis using data-driven spatiotemporal approach for information construction. *Transportation Research Part C: Emerging Technologies*, 71:19–31.

Cheng, Y., Zhang, M., and Yang, D. (2015). Automatic incident detection for urban expressways based on segment traffic flow density. *Journal of Intelligent Transportation Systems*, 19(2):205–213.

Chiou, S.-W. (2016). A robust urban traffic network design with signal settings. *Information Sciences*, 334:144–160.

Chrobok, R., Kaumann, O., Wahle, J., and Schreckenberg, M. (2004). Different methods of traffic forecast based on real data. *European Journal of Operational Research*, 155(3):558–568.

Clark, R. (2004). Rail flaw detection: overview and needs for future developments. *NDT & E International*, 37(2):111 – 118.

Consilvio, A., Febbraio, A. D., and Sacco, N. (2016). Stochastic scheduling approach for predictive risk-based railway maintenance. In *2016 IEEE International Conference on Intelligent Rail Transportation (ICIRT)*, pages 197–203.

Daganzo, C. F. (1994). The cell transmission model: A dynamic representation of highway traffic consistent with the hydrodynamic theory. *Transportation Research Part B, Methodological*, 28(4):269–287.

Daganzo, C. F. (2006). In traffic flow, cellular automata= kinematic waves. *Transportation Research Part B: Methodological*, 40(5):396–403.

Dailey, D. J. (1999). A statistical algorithm for estimating speed from single loop volume and occupancy measurements. *Transportation Research Part B: Methodological*, 33(5):313–322.

D'Ariano, A., Meng, L., Centulio, G., and Corman, F. (2019). Integrated stochastic optimization approaches for tactical scheduling of trains and railway infrastructure maintenance. *Computers & Industrial Engineering*, 127:1315 – 1335.

Dia, H. (2001). An object-oriented neural network approach to short-term traffic forecasting. *European Journal of Operational Research*, 131(2):253–261.

Dogru, N. and Subasi, A. (2018). Traffic accident detection using random forest classifier. In *2018 15th Learning and Technology Conference (L&T)*, pages 40–45. IEEE.

Dong, J. and Mahmassani, H. S. (2012). Stochastic modeling of traffic flow breakdown phenomenon: Application to predicting travel time reliability. *IEEE Transactions on Intelligent Transportation Systems*, 13(4):1803–1809.

Dougherty, M. S., Kirby, H. R., and Boyle, R. D. (1993). The use of neural networks to recognise and predict traffic congestion. *Traffic engineering & control*, 34(6).

Dunne, M. C. (1967). Traffic delays at a signalized intersection with binomial arrivals. *Transportation Science*, 1:24–31.

Fan, P., Seungmo, K., Xiaopeng, L., Yanfeng, O., Kamalesh, S., and Dharma, A. (2011). A heuristic approach to the railroad track maintenance scheduling problem. *Computer-Aided Civil and Infrastructure Engineering*, 26(2):129–145.

Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning*, volume 1. Springer series in statistics New York, NY, USA:.

Gazis, D. C. (2006). *Traffic theory*, volume 50. Springer Science & Business Media.

Ghofrani, F., He, Q., Goverde, R. M., and Liu, X. (2018). Recent applications of big data analytics in railway transportation systems: A survey. *Transportation Research Part C: Emerging Technologies*, 90:226 – 246.

Gittins, J., Glazebrook, K., and Weber, R. (2011). *Multi-armed bandit allocation indices*. John Wiley & Sons.

Güler, H. (2014). Prediction of railway track geometry deterioration using artificial neural networks: a case study for Turkish state railways. *Structure and Infrastructure Engineering*, 10(5):614–626.

Hamid, A. and Gross, A. (1981). Track-quality indices and track degradation models for maintenance-of-way planning. *Transportation Research Board*, 802:2–8.

Han, Y. and Ahn, S. (2018). Stochastic modeling of breakdown at freeway merge bottleneck and traffic control method using connected automated vehicle. *Transportation research part B: methodological*, 107:146–166.

He, Q., Li, H., Bhattacharjya, D., Parikh, D. P., and Hampapur, A. (2015). Track geometry defect rectification based on track deterioration modelling and derailment risk assessment. *Journal of the Operational Research Society*, 66(3):392–404.

Heidemann, D. (1996). A queueing theory approach to speed-flow-density relationships. In *Proceedings of the 13th International Symposium on Transportation and Traffic Theory*, France.

Heidemann, D. (2001). A queueing theory model of nonstationary traffic flow. *Transportation Science*, 35:405–412.

Heinicke, F., Simroth, A., Scheithauer, G., and Fischer, A. (2015). A railway maintenance scheduling problem with customer costs. *EURO Journal on Transportation and Logistics*, 4(1):113–137.

Jain, R. and Smith, J. (1997). Modeling vehicular traffic flow using M/G/C/C state dependent queueing models. *Transportation Science*, 31:324–336.

Jiang, Y., Zhang, Z., Voorhis, T. V., and McCalley, J. D. (2003). Risk-based maintenance optimization for transmission equipment. In *In Proc. of 12th Annual Substations Equipment Diagnostics Conference*, pages 1191–1200.

J.N. Darroch, G. N. and Morris, R. (1964). Queues for vehicle-actuated traffic light. *Operations Research*, 12:882–895.

Kallenberg, L. (1989). Markov decision processes. *European Journal of Operational Research*, 39(1):1 – 16.

Keilson, J. and Servi, L. (1993). The matrix M/M/∞ system: Retrial models and Markov modulated sources. *Advances in Applied Probability*, 25:453–471.

Kerner, B. S., Klenov, S. L., and Wolf, D. E. (2002). Cellular automata approach to three-phase traffic theory. *Journal of Physics A: Mathematical and General*, 35(47):9971.

Khalouli, S., Benmansour, R., and Hanafi, S. (2016). An ant colony algorithm based on opportunities for scheduling the preventive railway maintenance. In *2016 International Conference on Control, Decision and Information Technologies (CoDIT)*, pages 594–599.

Kulkarni, V. G. (2016). *Modeling and analysis of stochastic systems*. Crc Press.

Kurzhanskiy, A. A. (2009). Set-valued estimation of freeway traffic density. *IFAC Proceedings Volumes*, 42(15):271–277.

Kwon, J., Mauch, M., and Varaiya, P. (2006). Components of congestion: Delay from incidents, special events, lane closures, weather, potential ramp metering gain, and excess demand. *Transportation Research Record: Journal of the Transportation Research Board*, 1959(1959):84–91.

Kwong, K., Kavaler, R., Rajagopal, R., and Varaiya, P. (2010). Real-time measurement of link vehicle count and travel time in a road network. *IEEE Transactions on Intelligent Transportation Systems*, 11(4):814–825.

Laharotte, P.-A., Billot, R., and El Faouzi, N.-E. (2017). Detection of non-recurrent road traffic events based on clustering indicators. In *ESANN*.

Lannez, S., Artigues, C., Damay, J., and Gendreau, M. (2015). A railroad maintenance problem solved with a cut and column generation metaheuristic. *Networks*, 66(1):40–56.

Lehoczky, J. (1972). Traffic intersection control and zero-switch queues. *J. of Applied Probability*, 9:382–395.

Lidbetter, T. (2013). Search games with multiple hidden objects. *SIAM Journal on Control and Optimization*, 51(4):3056–3074.

Lidén, T. and Joborn, M. (2017). An optimization model for integrated planning of railway traffic and network maintenance. *Transportation Research Part C: Emerging Technologies*, 74:327 – 347.

Lighthill, M. and Whitham, G. (1955). On kinematic waves: II. a theory of traffic on long crowded roads. In *Proc. Roy. Soc. London Ser. A 229*, pages 317–345.

Lin, K. Y. and Singham, D. I. (2015). Robust search policies against an intelligent evader. Technical report, Naval Postgraduate School Monterey United States.

Liu, X., Saat, M. R., and Barkan, C. P. (2012). Analysis of causes of major train derailment and their effect on accident rates. *Transportation Research Record*, 2289(1):154–163.

Lopes Gerum, P. C., Altay, A., and Baykal-Gürsoy, M. (2019a). Data-driven predictive maintenance scheduling policies for railways. *Transportation Research Part C: Emerging Technologies*, 107:137–154.

Lopes Gerum, P. C. and Baykal-Gürsoy, M. (2020). Predicting congestions with tandem-queues. *working paper*.

Lopes Gerum, P. C., Benton, A. R., and Baykal-Gürsoy, M. (2019b). Traffic density on corridors subject to incidents: models for long-term congestion management. *EURO Journal on Transportation and Logistics*, pages 1–37.

Luan, X., Miao, J., Meng, L., Corman, F., and Lodewijks, G. (2017). Integrated optimization on train scheduling and preventive maintenance time slots planning. *Transportation Research Part C: Emerging Technologies*, 80:329 – 359.

Maerivoet, S. and De Moor, B. (2005). Cellular automata models of road traffic. *Physics reports*, 419(1):1–64.

Martey, E. N., Ahmed, L., and Attoh-Okine, N. (2017). Track geometry big data analysis: A machine learning approach. In *Big Data (Big Data), 2017 IEEE International Conference on*, pages 3800–3809. IEEE.

Massey Jr, F. J. (1951). The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American statistical Association*, 46(253):68–78.

May, A. and Keller, H. (1967). Non-integer car-following models. *Highway Res. Rec.*, 199:19–32.

Merrick, J. R. and Soyer, R. (2015). Semiparametric bayesian optimal replacement policies: application to railroad tracks. *Applied Stochastic Models in Business and Industry*, 33(5):445–460.

Mirasol, N. M. (1963). Letter to the editor—the output of an m/g/∞ queuing system is poisson. *Operations Research*, 11(2):282–284.

Morbidi, F., Ojeda, L. L., De Wit, C. C., and Bellicot, I. (2014). A new robust approach for highway traffic density estimation. In *2014 European Control Conference (ECC)*, pages 2575–2580. IEEE.

Moridpour, S., Mazloumi, E., and Hesami, R. (2017). Application of artificial neural networks in predicting the degradation of tram tracks using maintenance data. In *Applied Big Data Analytics in Operations Management*, pages 30–54. IGI Global.

Muñoz, L., Sun, X., Horowitz, R., and Alvarez, L. (2003). Traffic density estimation with the cell transmission model. In *Proceedings of the 2003 American Control Conference, 2003.*, volume 5, pages 3750–3755. IEEE.

Nagel, K. and Schreckenberg, M. (1992). A cellular automaton model for freeway traffic. *Journal de Physique I*, 2(12):2221–2229.

Nelder, J. A. and Mead, R. (1965). A simplex method for function minimization. *The Computer Journal*, 7(4):308–313.

Newell, G. (1965). Approximation methods for queues with application to the fixed-cycle traffic light. *SIAM Rev.*, 7:2:223–240.

Newell, G. (1971). *Applications of Queueing Theory*. Chapman and Hall, London.

Newell, G. F. (1993). A simplified theory of kinematic waves in highway traffic, part I: General theory. *Transportation Research Part B: Methodological*, 27(4):281–287.

Nishi, T., Muroi, Y., and Inuiguchi, M. (2011). Column generation with dual inequalities for railway crew scheduling problems. *Public Transport*, 3(1):25–42.

NOAA (Accesed in 2015-2016). National climatic data center.

Olver, F. W. (2010). *NIST Handbook of Mathematical Functions Hardback and CD-ROM*. Cambridge University Press.

Panichpapiboon, S. and Leakkaw, P. (2017). Traffic density estimation: A mobile sensing approach. *IEEE Communications Magazine*, 55(12):126–131.

Papadopoulou, S., Roncoli, C., Bekiaris-Liberis, N., Papamichail, I., and Papageorgiou, M. (2018). Microscopic simulation-based validation of a per-lane traffic state estimation scheme for highways with connected vehicles. *Transportation Research Part C: Emerging Technologies*, 86:441–452.

Peralta, D., Bergmeir, C., Krone, M., Galende, M., Menéndez, M., Sainz-Palmero, G. I., Bertrand, C. M., Klawonn, F., and Benítez, J. M. (2018). Multiobjective optimization for railway maintenance plans. *Journal of Computing in Civil Engineering*, 32(3):04018014/1–11.

Polson, N. and Sokolov, V. (2017). Deep learning for short-term traffic flow prediction. *Transportation Research Part C: Emerging Technologies*, 79:1–17.

Pour, S. M., Drake, J. H., Ejlertsen, L. S., Rasmussen, K. M., and Burke, E. K. (2018). A hybrid constraint programming/mixed integer programming framework for the preventive signaling maintenance crew scheduling problem. *European Journal of Operational Research*, 269(1):341 – 352.

Puterman, M. L. (1994). *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., New York, NY, USA, 1st edition.

Qiu, T. Z., Lu, X.-Y., Chow, A. H., and Shladover, S. E. (2010). Estimation of freeway traffic density with loop detector and probe vehicle data. *Transportation Research Record*, 2178(1):21–29.

Richards, P. (1956). Shock waves on the highway. *Operations Research*, 4:42–51.

Ritchie, S. G. and Cheu, R. L. (1993). Simulation of freeway incident detection using

artificial neural networks. *Transportation Research Part C: Emerging Technologies*, 1(3):203–217.

Roberts, D. and Gittins, J. (1978). The search for an intelligent evader: Strategies for searcher and evader in the two-region problem. *Naval Research Logistics Quarterly*, 25(1):95–106.

Ross, S. (1992). *Applied Probability Models with Optimization Applications*. Dover Books on Advanced Mathematics. Dover Publications.

Ross, S. (1996). *Stochastic Processes*, volume 2. Wiley New York.

Ruckle, W. H. (1983). *Geometric games and their applications*. Pitman Boston.

Schrank, D., Eisele, B., Lomax, T., and Bak, J. (2015). Urban mobility scorecard. *College Station: Texas A&M Transportation Institute and INRIX*, (August).

Sharma, S., Cui, Y., He, Q., Mohammadi, R., and Li, Z. (2018). Data-driven optimization of railway maintenance for track geometry. *Transportation Research Part C: Emerging Technologies*, 90:34 – 58.

Simson, S., Ferreira, L., and Murray, M. (2000). Rail track maintenance planning: An assessment model. *Transportation Research Record: Journal of the Transportation Research Board*, 1713:29–35.

Skabardonis, A., Petty, K., and Varaiya, P. (2003). Measuring recurrent and nonrecurrent traffic congestion. In *Proc. of the 82th Annual Meeting of the Transportation Research Board*. volume CD-ROM, Washington D. C.

Skabardonis, A., Petty, K., Varaiya, P., and Bertini, R. (1998). Evaluation of the Freeway Service Patrol (FSP) in Los Angeles, ucb-its-prr-98-31. Technical report,

California PATH Research Report, Institute of Transportation Studies, University of California, Berkeley.

Soleimanmeigouni, I., Ahmadi, A., Letot, C., Nissen, A., and Kumar, U. (2016). Cost-based optimization of track geometry inspection. In *Proceedings of the 11th World Congress on Railway Research.*, pages 1–7.

Sun, X., Muñoz, L., and Horowitz, R. (2004). Mixture kalman filter based highway congestion mode and vehicle density estimator and its application. In *Proceedings of the 2004 American Control Conference*, volume 3, pages 2098–2103. IEEE.

Tanner, J. (1953). A problem of interface between two queues. *Biometrica*, 40:58–69.

Thakur, G. S., Huiz, P., and Helmy, A. (2012). Modeling and characterization of urban vehicular mobility using web cameras. In *Proceedings of the 2012 IEEE Conference on Computer Communications Workshop*, pages 262–267. IEEE.

Thomas, J. L., Jones, G. N., Scarinci, I. C., Mehan, D. J., and Brantley, P. J. (2001). The utility of the CES-D as a depression screening measure among low-income women attending primary care clinics. *The International Journal of Psychiatry in Medicine*, 31(1):25–40.

Tomasi, C. (2004). Estimating gaussian mixture densities with em–a tutorial. *Duke University*, pages 1–8.

Transportation Economics and Management Systems Inc (2018). Northern Michigan Rail Ridership Feasibility and Cost Estimate Study. Technical report, The Groundwork Center For Resilient Communities, Grant Fiduciary: Bay Area Transportation Authority.

Treiber, M., Hennecke, A., and Helbing, D. (2000). Congested traffic states in empirical observations and microscopic simulations. *Physical review E*, 62(2):1805.

Tricomi, F. G. and Erdélyi, A. (1951). The asymptotic expansion of a ratio of gamma functions. *Pacific Journal of Mathematics*, 1(1):133–142.

Turner, C., Tiwari, A., Starr, A., and Blacktop, K. (2016). A review of key planning and scheduling in the rail industry in Europe and UK. *Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit*, 230(3):984–998.

van Erp, P. B., Knoop, V. L., and Hoogendoorn, S. P. (2018). Macroscopic traffic state estimation using relative flows from stationary and moving observers. *Transportation Research Part B: Methodological*, 114:281–299.

Vandaele, N., VanWoensel, T., and Verbruggen, N. (2000). A queueing based traffic flow model. *Transportation Research-D: Transportation and Environment*, 5:121–135.

Wang, L., An, M., Qin, Y., and Jia, L. (2018). A risk-based maintenance decision-making approach for railway asset management. *International Journal of Software Engineering and Knowledge Engineering*, 28(04):453–483.

Wang, S. and Ahmed, N. (2017). Dynamic model of urban traffic and optimum management of its flow and congestion. *Dynamic Systems and Applications*, 26(3-4):575–587.

Wang, Y. and Papageorgiou, M. (2005). Real-time freeway traffic state estimation based on extended kalman filter: a general approach. *Transportation Research Part B: Methodological*, 39(2):141–167.

Wang, Y., Papageorgiou, M., Messmer, A., Coppola, P., Tzimitsi, A., and Nuzzolo, A. (2009). An adaptive freeway traffic state estimator. *Automatica*, 45(1):10–24.

Yolmeh, A. and Baykal-Gürsoy, M. (2018). Urban rail patrolling: a game theoretic approach. *Journal of Transportation Security*, 11(1-2):23–40.

Young, P. and Zamir, S. (2014). *Handbook of game theory*. Elsevier.

Zhang, J., Yu, Y., and Lei, Y. (2016). The study on an optimized model of traffic congestion problem caused by traffic accidents. In *2016 Chinese Control and Decision Conference (CCDC)*, pages 688–692. IEEE.

Zhang, S., Wu, G., Costeira, J. P., and Moura, J. M. (2017). Understanding traffic density from large-scale web camera data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5898–5907.

Zhang, X., Onieva, E., Perallos, A., Osaba, E., and Lee, V. C. (2014). Hierarchical fuzzy rule-based system optimized with genetic algorithms for short term traffic congestion prediction. *Transportation Research Part C: Emerging Technologies*, 43:127–142.

Zhao, Z., Chen, W., Wu, X., Chen, P. C., and Liu, J. (2017). Lstm network: a deep learning approach for short-term traffic forecast. *IET Intelligent Transport Systems*, 11(2):68–75.

Zheng, F., Jabari, S. E., Liu, H. X., and Lin, D. (2018). Traffic state estimation using stochastic lagrangian dynamics. *Transportation Research Part B: Methodological*, 115:143–165.

Zheng, J. and Liu, H. X. (2017). Estimating traffic volumes for signalized intersections using connected vehicle data. *Transportation Research Part C: Emerging Technologies*, 79:347–362.

Zhong, Y., Xie, X., Guo, J., Wang, Q., and Ge, S. (2018). A new method for

short-term traffic congestion forecasting based on lstm. In *IOP Conference Series: Materials Science and Engineering*, volume 383, page 012043. IOP Publishing.