University of Twente

Master Thesis Industrial Engineering and Management

Smart refueling decisions using a reinforcement learning approach: a case study at Nijhof-Wassink

By Myrthe Kruit

March 16, 2024

First Supervisor: Dr. A. Asadi (Amin)
Second Supervisor: J.P.S. Piest (Sebastian)
Company Supervisor: J. Lenselink (Jordi)

Management Summary

This research dives into the intricate challenge of reducing fuel costs within the transportation industry, particularly by optimizing truck refueling decisions along a fixed route. The research adopts a comprehensive approach, considering factors like fuel price uncertainty, limited access to real-time fuel prices, and the rapid expansion of the problem size. What sets this research apart is the incorporation of price uncertainty into the model through a forecast of fuel prices. The proposed solution method involves an innovative application of Reinforcement Learning (RL), framed within a sequential decision-making framework. The ultimate goal is to present a solution framework that is both academically innovative and practically relevant. A case study at Nijhof-Wassink's Dry Bulk Logistics (DBL) department validates the RL approach's effectiveness in a real-world context.

In examining the current situation at Nijhof-Wassink Group and the existing literature, several noteworthy findings emerge, emphasizing both the practical and academic relevance of this research. The analysis of Nijhof-Wassink Group's current situation, highlights the substantial impact of fuel costs, constituting 21% of total operating expenses in the DBL sector for 2022. A 1% reduction in fuel costs corresponds to an impressive 8% increase in profits, underscoring the financial significance of efficient refueling decisions. The absence of a dedicated decision support tool accentuates the need for a comprehensive solution. On the academic front, the literature review positions our research within the transportation sector. The review identified existing approaches, providing valuable insights into different facets of the refueling problem. To our knowledge, none of the works incorporated price variability, thus neglecting the stochastic nature of fuel prices over time. The contribution of this work lies in introducing a novel element by incorporating fuel price uncertainty through a Machine Learning algorithm for price forecasting within a Markov Decision Process (MDP). This integration of uncertainty in fuel prices and the application of RL techniques set the approach in this research apart and contribute to advancing the understanding of optimal refueling strategies in the transportation industry.

The sequential decision-making framework with an RL approach is validated by a case study at Nijhof-Wassink. A subset of 21 routes is selected and historical data is collected, prepared, and constructed as input for RL. The subset of routes covers 30% of the total trips driven and 28% of the total kilometres driven in the period of half a year. For the RL algorithm, both a Q-Learning algorithm and a SARSA algorithm are tested. Evaluations favored the Q-Learning algorithm, showcasing a marginal 0.02% performance advantage over SARSA. Exposing the RL algorithm to 10 new routes demonstrated the generalizability and wider applicability of the RL algorithm. A thorough comparative analysis highlighted the RL algorithm's impressive performance, with a 0.3% deviation from the deterministic optimal solution, a 3% improvement over the benchmark heuristic, and an 11% cost reduction compared to historical decisions. Realizing this cost decrease within the DBL department, saves between 660 and 990 thousand euros, increasing their profits by an impressive 88%. The computational findings for the Q-Learning algorithm, strongly support the effectiveness of RL in addressing the Fixed Route Vehicle Refueling Problem (FRVRP). Furthermore, these findings underscore the potential for fuel cost

reduction within the industry and a successful innovative implementation of stochastic fuel prices.

In addition to validating the framework's effectiveness, the study also yielded significant insights across various decision-making levels. Firstly, in the realm of predicting fuel prices, an analysis showed substantial variability influenced by independent variables, underscoring the importance of accounting for fuel price stochasticity. Notably, the analysis revealed that predicted fuel prices are lowest in Belgium and within network groups associated with fuel company Y. Operationally, the RL algorithm's policies showcased refueling cost differentials along routes and how the polices are generated for varying initial fuel levels. Tactical examinations delved into detour trade-offs, indicating a €0.005 reduction in net fuel price per extra kilometer for profitable detours on a typical 417 km route. Strategically, the study emphasized the profitability of prioritizing refueling decisions over negotiating extra discounts, as the latter showed minimal potential profit gains. These insights collectively underscore the efficacy and versatility of the framework in addressing challenges within the transportation industry.

Concluding, this research proves the contribution of RL in optimizing complex refueling decisions and its potential for real-world implementation. Furthermore, this research provides a framework for sequential-decision making under uncertainty that optimizes the refueling decisions for fixed routes on an operational level.

Recommendations for the company include continuing the project to reduce fuel costs and focusing on improving data structure and quality. While the ideal scenario involves a live connection with the trucks' board computers for real-time advice, it is recommended first to enhance data structure and quality, gradually progressing from descriptive analytics to more complex prescriptive analytics like an RL model.

The thesis acknowledges certain limitations and suggests avenues for future research. It underscores possible improvements including extending the model's evaluation to longer routes and venturing into an infinite MDP framework to enhance decision-making capabilities for unseen routes. Recommendations include expanding the case study with more historical data or applying the policies in real life to strengthen and validate the model's practical contribution. In the realm of future research, changes to the MDP formulation are proposed to enhance applicability. These include broadening the action space, integrating driving-rest time regulations, adding AdBlue levels to the state space, and increasing the drivers' freedom. These changes aim to make the model more adaptive, comprehensive, and aligned with real-world refueling strategies.

In summary, this research contributes a sequential decision-making framework under price uncertainty, which defines the FRVRP as an MDP and solves it with RL. The computational results prove the contribution of an RL approach in optimizing complex refueling decisions and its potential for real-world implementation. Furthermore, the RL approach is validated through practical application and lays the groundwork for future enhancements and implementation within the transportation industry.

Contents

List of Figures

List of Tables

List of Algorithms

1. Introduction

This master thesis explores the performance of a novel decision-making framework that aims to reduce the refueling costs for trucking companies within the transportation industry. The framework should advise at which gas station a truck should refuel along a fixed route. This advice should take into account the complexities of the problem such as the variability of the fuel prices, lack of access to the fuel prices and the fast-growing problem size. To tackle these challenges we propose a novel RL approach. This section begins by outlining the problem context and research motivation (Section 1.1), followed by a description of the research aim (Section 1.2). Subsequently, the solution framework is presented (Section 1.3), followed by an introduction about the company under study (Section 1.4). Afterwards, the problem definition (Section 1.5) and the research questions (Section 1.6) are discussed. Finally, the problem-solving approach is outlined (Section 1.7) and the structure of this research is introduced (Section 1.8).

1.1. Problem Context and Research Motivation

The transportation industry plays a pivotal role in ensuring that everything we see around us, from the items in this room to the clothing we are currently wearing, reaches its destination. This industry relies on different transportation modes such as trucks, ships, trains, and planes to move goods from various corners of the world to their destination. Trucking companies play a major role, accounting for around 77% of the total inland freight transportation in Europe (EuroStat, 2023). This statistic highlights the extensive distances these trucks cover annually, and the fuel consumption required to cover these distances. Inherent to the transportation sector are the low-profit margins and high costs. Especially, refueling costs emerge as a major contributor to the costs, representing approximately 20% to 30% of the operational costs in both Europe and the United States (Mckinnon, 2023; Persyn et al., 2019; Leslie and Murray, 2022). Hence, research on reducing refueling costs has been a topic of interest within the transportation domain in recent years (Lin et al., 2007; Suzuki, 2008, 2009; Rodrigues Junior and Cruz, 2013; Lin, 2014; Suzuki and Lan, 2018). On top of that, the increase and fluctuations in fuel prices only further heightened the need for research on this topic (Bureau of Transportation Statistics, 2022; NOS, 2023). The complexity of the refueling problem can be devoted to several factors. Firstly, the decision whether to refuel must be made at each gas station. This decision involves a trade-off between factors such as the current fuel level, the remaining distance to the destination, the trade-off between the costs of a detour vs. the benefits of a lower price, and the price of the upcoming gas stations. The last factor, the prices of the upcoming gas stations, adds an additional layer of complexity as these prices are unknown and are influenced by many factors such as the region, brand and date. Moreover, as the route length or the number of potential actions increases, the problem size grows rapidly, leading to potentially unacceptable long solution times. To effectively address these challenges and optimize cost reduction in refueling decisions, a suitable framework is required to deal with the stochastic prices and growing problem size. This framework should deal with these limitations and capture the intrinsic complexities of the system, which often involves uncertainty under sequential and online decision-making. The current body of literature on refueling problems does not incorporate the complexity of stochastic and unknown prices resulting in a research gap this thesis aims to cover.

1.2. Research Aim

This research aims to reduce the refueling costs for trucking companies in the transportation sector by developing and validating an online sequential decision-making framework with uncertain fuel prices. This framework contributes to the existing body of literature by addressing a previously overlooked aspect: The uncertainty associated with fuel prices. The framework is set out to consider the complexities of the refueling decisions including factors such as the current fuel level, the remaining distance to the destination, the costs of a detour, and the stochastic price of the current and upcoming gas stations. The fuel prices have a stochastic nature that can vary depending on factors such as the region, the location, the owner company of the gas station, the time of the day, the crude oil price, and the possible price agreements the logistics company made for a discount on the liter price. These price fluctuations can impact the decision and that decision can in turn impact the following chain of decisions. This research aims to capture this stochasticity by deploying a machine-learning approach. In the end, this research should provide an online sequential decision-making framework to find the best decisions for a logistics company, aiming to minimize the refueling costs over its routes.

1.3. The Framework

This research proposes to express the refueling problem under sequential decision-making as a Markov Decision Process (MDP). MDP is a framework for modeling stochastic dynamic programs by defining the state, actions, transition function and probabilities, and rewards (Powell, 2011). Our MDP has a finite horizon that covers the start to the end of the route. In Section 4 the MDP is defined and the mathematical formulation is presented. The solution of the MDP is a policy that determines which action to take in each state so that the total expected reward, which is the negative counterpart of the refueling cost, is maximized. We note that the refueling problem can become large-scale as the number of states (e.g., amount of fuel in the tank per truck per route) and the feasible actions (i.e., the amount of fuel to put in the tank per truck per route) as the level of granularity and the transportation network grows. Hence, deriving the transition probabilities between states and incorporating them to solve such a large-scale problem is intractable as it suffers from the curses of dimensionality (Chang et al., 2013; Powell, 2011). Hence, we apply RL, which is promising for complex, online, stochastic, and large-scale problems such as refueling decisions (Abdullah et al., 2021; Giannoccaro and Pontrandolfo, 2002; Nazari et al., 2018). Reinforcement Learning involves discovering optimal actions by learning how to map situations to actions in order to maximize a numerical reward signal. Instead of being explicitly instructed which actions to take, the learner explores different actions to determine which ones yield the highest reward (Sutton and Barto, 2018). While there is an increasing body of literature in the transportation domain recognizing the potential of RL to tackle industry challenges, there is a notable lack of information on RL as an approach to the refueling problem (Farazi et al., 2021; Winder, 2021; Yan et al., 2022; Nazari et al., 2018). Therefore, this research seeks to bridge this gap by expressing the refueling problem as an MDP and employing RL as its solution method.

1.4. Company Introduction

In practice, a minor improvement on the operational cost per tour can decide whether a freight services company is profitable or not. Therefore, the refueling problem is relevant

for all trucking companies within the transportation industry seeking to reduce their fuel costs. Thus, the optimization of costs has key importance in the operation of such companies. To evaluate and demonstrate the RL approach a case study is executed at Nijhof-Wassink Group. The Nijhof-Wassink Group comprises multiple companies and activities that collectively provide a unique comprehensive offering within the logistics sector. Two companies within the Nijhof-Wassink Group could benefit from a decision-support tool for refueling decisions namely Wemmers and Nijhof-Wassink. Wemmers provides transportation services for the food industry across Europe and Nijhof-Wassink transports chemical and compound feed across multiple countries. The scope of the case study is put on Nijhof-Wassink. Nijhof-Wassink has been a specialist in bulk transport by road, rail, and water since 1967. Their fleet consists of around 800 trucks that are active in the chemical and compound feed markets. The chemical sector typically involves longer trips, with trucks transporting a single product from origin A to destination B. In contrast, the feed sector utilizes trucks with multiple compartments, enabling a single trip to involve deliveries from origin A to client B, then to client C, and finally returning to origin A. The chemical sector comprises of three different sub-sectors Dry Bulk Logistics (DBL), Liquid Bulk Logistics (LBL), and Fuel. This research focuses on the chemical sector and more specifically the sub-sector DBL. In the DBL sector, the trips are characterised by longer distances and covering diverse countries and regions. These characteristics lead to a heightened level of price variability. The impact of choosing the most economical gas station is more significant due to this fluctuation, as the decision-making process becomes more challenging. As a result, guidance for these trips carries increased importance. More details about DBL and the routes they drive can be found in Section 2.3.

1.5. Problem Definition from Company Perspective

In the previous section, Nijhof-Wassink is introduced as a trucking company in the Netherlands. Nijhof-Wassink aims to decrease fuel costs as they represent around 21% of the operational costs. The internal budget of 2022 shows that decreasing the fuel costs is crucial to the company as for each 1% decrease in fuel costs in the DBL sector, the profit of DBL increases by 8%. Aiming for continuous innovation, they seek to explore how the increasingly popular AI algorithms can contribute to solving the refueling problem. Therefore, besides the academic contribution, this project aims to practically contribute to Nijhof-Wassink by providing insights into the possibilities of an RL model, how it can solve the refueling problem, and provide advice to the drivers. In the end, this research provides a decision-making tool for the company that shows the potential fuel savings of implementing such a model.

1.6. Research Questions

From the problem context and definition, the following main research question is derived:

"Can we provide near-optimal solutions for the refueling problem with stochastic fuel prices by using a Reinforcement Learning approach, framed within a novel sequential decision-making framework?"

To answer the main research question, 10 research questions are answered. This section describes per research question (RQ), the relevance, the aim, and if applicable the data gathering method and the data analysis method.

1. *How are refueling decisions currently made at Nijhof-Wassink?*

   The aim of answering this research question is to get an overview of the current situation before starting on the case study. It is also done to compare the result of this research to the current situation and see improvements. Lastly, it is important to understand the current situation and see how a solution would fit in the current environment. This knowledge is obtained through walk-in interviews (Chapter 2).

2. *What approaches for solving the refueling problem are present in the literature and how is the framework we propose a contribution to the existing body of literature?*

   The aim is to provide an overview of the refueling models present in the current body of literature. We want to know their features and to which extent the complexity of the refueling model is captured. Ultimately, we aim to point out the novelty of our research by comparing our problem-solving approach to the existing approaches. This knowledge question is exploratory where qualitative information is found in the literature (Chapter 3).

3. *What is the MDP formulation of the fixed-route vehicle refueling problem with stochastic fuel prices?*

   The purpose of answering this research question is to develop a mathematical representation of the refueling problem that describes the relationships between decision variables, objectives, and constraints. The result of this research question is a model that can be used as input for a computer program together with the modeling assumptions and simplifications (Chapter 4).

4. *How can we solve the MDP of the refueling problem and gain (near-)optimal results?*

   Once the problem statement is developed a solution method is needed. This research question aims to present a suitable RL algorithm to solve the MDP and to present 3 other solution methods to compare our approach with (Chapter 5).

5. *What data is needed as input for the case study and how is the data collected, prepared and constructed?*

   The relevance of this research question is to document what input data is needed to execute the case study and how this data is collected and prepared. With the data, the parameters of the model are defined (Chapter 6.1).

6. *Which machine learning regression model do we select to forecast the uncertainty in the fuel prices?*

This research question deals with the novel component of our framework, the price uncertainty. Answering this research question involves selecting a regression model to incorporate the stochastic nature of the fuel prices. The result of this question is a file with forecasted fuel prices that is used as input for the RL algorithm (Chapter 6.2).

7. *How are the parameters of the RL algorithm tuned to provide near-optimal results?*

   Algorithms are dependent on careful tuning of their parameters. The aim of this research question is to find settings that optimize the performance of the algorithm by executing experiments and tests (Chapter 6.3).

8. *What are the results of the different solution methods, and what insights do the results provide regarding the performance of the RL algorithm?*

   This question aims to quantitatively analyze the performance of RL and compare it to three other solution methods. Proving RL performs close to the optimal solution or outperforms a heuristic can validate its performance. Furthermore is comparing the result of the RL model with the current situation useful for providing an estimation of the cost reduction. This is relevant information for the management of Nijhof-Wassink (Chapter 6.4).

9. *What insights does this model in combination with this solution method provide to the practical and academic communities?*

   This research question aims to analyse the results of the RL algorithm to provide insights to the practical and academic communities on operational, tactical and strategic levels. (Chapter 6.5).

10. *Does this model have the potential to be further expanded and deployed in Nijhof-Wassink?*

    This question is relevant for Nijhof-Wassink. The answer to this question should provide them with insight into how and if the model should be deployed, and what still needs to be done to further integrate the model and what obstacles still need to be overcome (Chapter 7.4).

1.7. Problem Solving Approach

To address the research questions and guide the problem-solving process, the Cross-Industry Standard Process for Data Mining (CRISP-DM) framework is employed (Wirth and Hipp, 2000). This methodology provides a structured approach to solving complex problems in data science and machine learning. The CRISP-DM framework consists of six major phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment (see Figure 1).

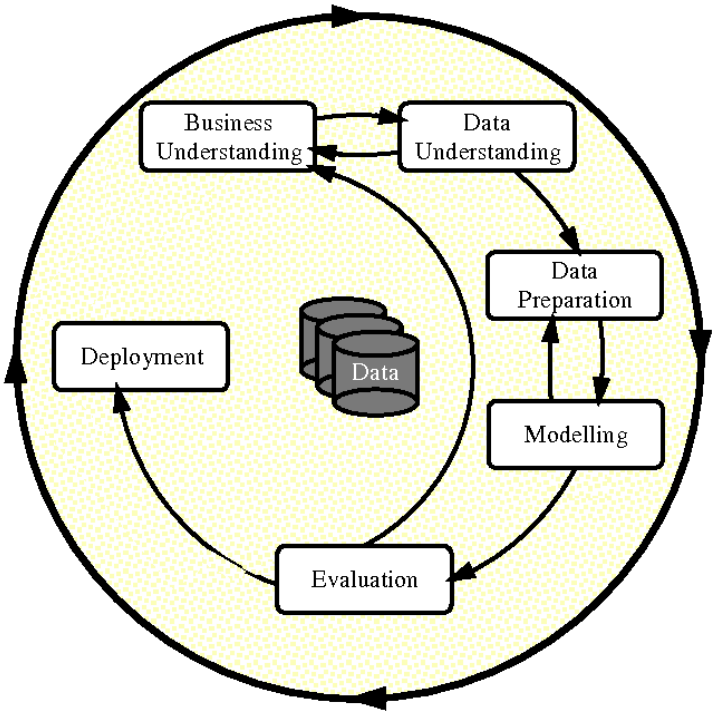

Figure 1: Phases of CRISP-DM process model for data mining (Wirth and Hipp, 2000)

The Business Understanding Phase focuses on understanding the objectives and requirements of the project. This phase is addressed in Chapter 1 by providing the problem context, the research motivation and the initial research questions. The business understanding phase continues in Chapter 2 by delving into the finances and decision-making within Nijhof-Wassink.
The Data Understanding Phase drives the focus to identify, collect, and analyze the data sets that can help accomplish the project goals and gain a deeper understanding of the problem. In this research, it is also vital to gather information from the literature and the current situation. Chapters 2, 3, and 4 focus on this phase and consequently gather information about the current refueling decision-making process at Nijhof-Wassink (RQ 1), explore existing refueling models in the literature (RQ 2), and gather data and expert knowledge (RQ 3).
The Data Preparation Phase prepares the final data sets and formulates the data requirements for the case study. This includes what data is needed, how it is collected, and how it is prepared. In this phase, we answer RQ 5 and describe it in Chapter 6.1.
The Modeling Phase is dedicated to building and assessing various models based on several modeling techniques. This phase is executed by answering RQ 4, 6 and 7 in Chapter 5, 6.2, and 6.3. This phase involves the selection of an appropriate algorithm for the refueling problem, presenting different modeling techniques for the prediction model of the fuel prices, and tuning the RL parameters.
The Evaluation Phase examines the results of the model, and if it meets the business

objectives. RQ 8 and RQ 9 address this phase and quantitatively analyze performance, compare it to the current situation at Nijhof-Wassink, and derive insights for both practical and academic communities. The elaboration is presented in Chapter 6.4.
The Deployment Phase is about documenting a plan for deploying the model. In this phase, we present what the next steps are for Nijhof-Wassink. Furthermore, this phase summarizes the practical and academic contributions, communicates insights to relevant communities, and explores the limitations of the RL model. These topics are covered in RQ 10 in Chapter 7.

1.8. Outline

This thesis follows a structured approach, starting with Chapter 2 devoted to *Business Understanding*. This chapter forms the basis for understanding the stakeholders involved, the transportation network, the impact of fuel costs, the composition of the fuel prices, and how current refueling decisions are taken. Moving on to Chapter 3, the *Related Work* section provides a comprehensive review of existing literature, addressing different approaches to fuel cost reduction. Moreover, this chapter is set to validate the novelness of the framework and to address other areas of research that adopted sequential decision-making under uncertainty. The subsequent two chapters dive into defining and presenting the framework. First, in Chapter 4, the *Mathematical Model* is introduced, elucidating the MDP with its decision epochs, states, actions, transitions, and rewards. Second, Chapter 5 presents the *Solution Methods* for the MDP including two RL algorithms. Furthermore, three other methods are introduced to compare the RL approach against. After the general framework is presented, we continue with validating it by executing a *Case Study* in Chapter 6. The case study consists of four parts (1) understanding, preparing and constructing the data (2) developing an RL regression model for the prediction of the fuel prices (3) tuning the parameters of the RL algorithms, and (4) executing the solution methods and presenting the results and insights. The research concludes with Chapter 7, *Conclusion, Future Research, and Recommendations.*

2. Business Understanding

This section describes the current situation and aims to answer the research question *"How are refueling decisions currently made at Nijhof-Wassink?"*. Furthermore, we provide more background information to understand the problem in more depth. First, Section 2.1 explains the business objectives of Nijhof-Wassink. Second, Section 2.2 presents an analysis of the refueling decisions including the factors influencing the refueling decisions, the stakeholders involved and the current refueling behavior at Nijhof-Wassink. Third, Section 2.3 explains the transportation network and corresponding terms. Lastly, Section 2.4, elaborates on the price agreements between Nijhof-Wassink and fuel companies.

2.1. Business Objectives

As mentioned before, the ultimate goal of Nijhof-Wassink is to reduce fuel costs. The financial budget of 2022 highlights the importance, showing Nijhof-Wassink's total fuel cost ranged between 21 and 25 million euros. These fuel costs represent a substantial 21% of the operating costs, which include the cost price and direct sector costs. Focusing on the Dry Bulk Logistics (DBL) sector, the area of our research, fuel costs were in the range of 6 to 9 million euros, also accounting for 21% of the operating costs. These figures underscore the significant financial investment Nijhof-Wassink makes in fuel costs. To better understand the potential impact on profitability, it is essential to assess how a reduction in fuel costs translates into increased profits. Internal documents reveal that each 1% decrease in fuel costs results in an impressive 8% increase in profits within the DBL sector. This underscores the importance of managing and potentially reducing fuel costs to enhance the company's financial performance. To evaluate the feasibility of achieving a 1% reduction in fuel costs, we examine the required decrease in the price per liter to realize this reduction. Analyzing fuel transactions from the previous year reveals that the average price per liter was approximately €1.63. Therefore, refueling at an average price of €1.61 would already lead to a 1% decrease in refueling costs.

Nijhof-Wassink has been thinking about reducing fuel costs in the past and intending to keep innovating, they want to know how the increasingly popular AI algorithms can contribute to solving this refueling problem for them. In the ideal situation envisioned by Nijhof Wassink, they have a model with real-time connectivity to trucks that would receive live data on fuel levels, locations, and destinations. The model would then generate real-time advice on an operational level. The refueling advice is then sent directly to the driver's on-board computer. Although there is recognition of potential challenges related to restricting drivers' autonomy, this ideal model could significantly enhance decision-making efficiency. This research aims to provide them with an RL algorithm within a sequential decision-making framework that provides policies for refueling decisions.

2.2. Assessment of Refueling Decisions

For the development of a model, it is crucial to gain a deeper understanding of which factors and stakeholders are involved in the decision-making of refueling decisions. First, we gather all the factors that influence the refueling decisions. Second, we present the stakeholder analysis of the refueling decisions. Lastly, an analysis is provided of the current refueling behavior.

2.2.1. Factors Influencing the Refueling Decisions

Defining a "good" refueling decision is one of the challenges identified by Nijhof-Wassink. A refuel decision is defined as deciding at which gas station(s) to refuel and how much. A question they raise is which factors they should take into account. Since their goal is to reduce fuel costs we aim to find factors that influence the total refuel costs or limit the refueling decision by practical constraints. Interviews find these factors and include the price per liter, which involves considerations of both the unknown pump price and existing price agreements influencing the direct fuel costs. More on the price agreements in Section 2.4. Furthermore, the current fuel level and the time it takes to refuel impact the refueling decision. Additionally, the option of combining refueling with necessary breaks can influence the total refuel costs as the fixed cost for stopping is reduced. Also, evaluating detour costs against lower fuel prices is important to determine the best gas station. Another factor is taking into account the refueling of AdBlue to determine if it is cost-efficient to combine refueling AdBlue and Diesel or not. Other considerations encompass the avoidance of refueling before loading to prevent increased weight, drivers' personal preferences such as saving stamps or socializing with colleagues, and the characteristics of gas stations, such as high-speed pumps, truck-friendly facilities, and the availability of shops. Together with stakeholders, we identified the most impactful factors: the net price per liter, the detour costs, the time costs for refueling and the fuel level. Figure 2, visualizes these factors and their components. Including remaining factors increases complexity or is not possible due to data or time constraints. They can be part of future research.

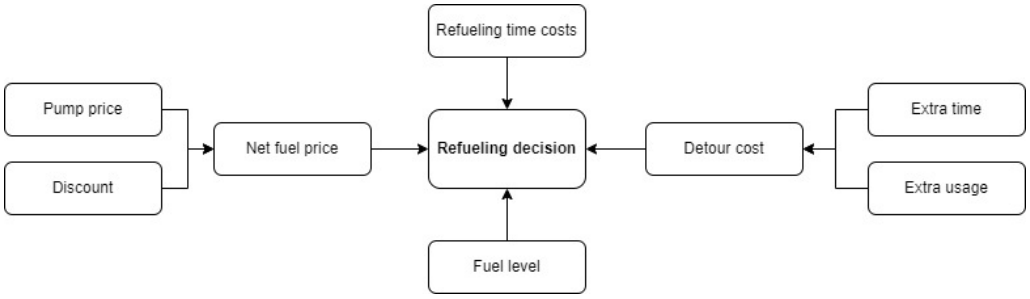

Figure 2: Overview of factors influencing the refueling decision

2.2.2. Stakeholders of Refueling Decisions

While the previous section delved into the factors influencing refueling decisions, this section explores the stakeholders affected by or impacting these decisions across various planning levels within the organization. Organizational planning typically consists of three levels: strategic, tactical, and operational planning. The strategic level, encompassing long-term planning and decision-making, aligns with Nijhof-Wassink's management goal to reduce costs through optimized refueling decisions. The purchasing department, as an internal stakeholder, plays a crucial role at this level by negotiating price agreements with fuel companies every three years. These price agreements influence the net fuel price through discounts on the liter price. Moving to the tactical planning level, focused on medium-term strategies spanning weeks to months, the Behavior Based Safety (BBS) department plays a role by analysing driver behavior and promoting safe practices

on the road. They can influence the refueling behavior of the drivers by introducing generalized rules for refueling decisions. Currently, the BBS department does not have an active role in guiding the refueling decisions. At the operational level, involving day-to-day decision-making, the refueling decision on when and how much to refuel is made. In the present situation, the drivers make these decisions. However, in an ideal scenario, an RL algorithm would guide these decisions and present the policy to drivers through the board computer. Additionally, fuel companies impact operational decisions by changing pump prices daily, introducing an uncertain element beyond the organization's control.

2.2.3. Current Refueling Behavior

To understand the current refueling behavior of the drivers we examined the amount the drivers refuel, at which fuel level they refuel, and the price discount for refueling. For this analysis, we used all trips and refuels done within the first half year of 2023 in the DBL sector. Figure 3 displays how much fuel drivers usually add at a time. The graph suggests that drivers occasionally refuel in small amounts, possibly not filling up the tank entirely or choosing to refuel when there is still enough fuel left. Figure 4 shows that most drivers refuel during the trip if their fuel level is between 40% and 50% at the start. A significant amount of refuels are done when the fuel level is above 50%, their range is still around 636 kilometers, making refueling not necessary yet. From this graph, we can conclude that drivers often refuel earlier than needed. Due to the time it takes to refuel we assume that frequent small refuels result in higher overall costs due to more stops. However, we do not have the data to confirm this. Besides the fuel amount, the use of price agreements is also analyzed (Figure 5). Out of all the times drivers got fuel for their trucks in the Netherlands, Germany, and Belgium, only 6% of them paid the regular price without any discount. Around 42% of the time the amount of discount at which drivers refueled was the maximum discount available. This does not mean they got the absolute lowest price, but it does demonstrate that drivers frequently make use of advantageous discounts. Making sure more drivers refuel at the highest discount could help cut down on fuel costs.

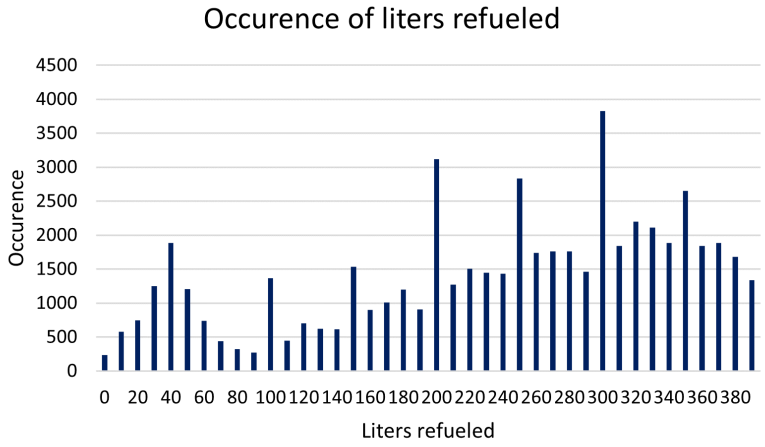

Figure 3: Occurence of how many liters drivers refuel each time

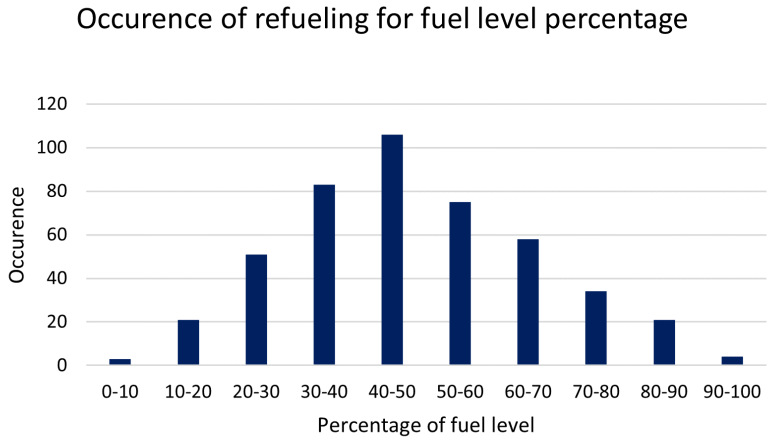

Figure 4: Occurence of refueling at different fuel levels

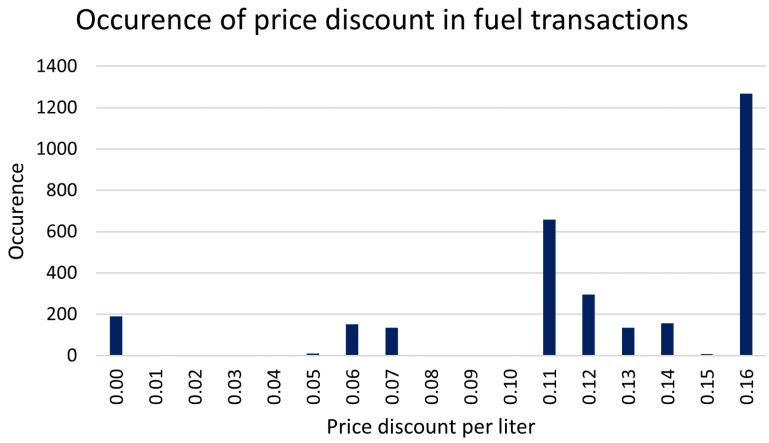

Figure 5: Count of each price discount level over the transactions

2.3. Transportation Network

In the context of this thesis, it is important to understand the four terms that are employed to indicate distinct aspects of the transportation network: lanes, shifts, trips and routes. A "lane" is defined as a distinctive route traversed by a truck, typically extending from a point of loading to an ultimate point of unloading. These lanes represent the specific paths undertaken during transportation operations and are reoccurring over time. The term "shift" is used to indicate the traveled path from the moment the driver starts working to the time he leaves this truck to go home. A shift is a sequence of lanes arranged consecutively, and it may span across a single day or extend over multiple days. Since the shifts are created by planners, each shift is distinct and does not often

reoccur over time. A "trip" is characterized by a specific lane, executed on a certain date, by a particular truck. Thus, each trip is unique, marked by the combination of a specific lane, date, and assigned truck. This distinction ensures the individualization of each transportation event within the company. Hence a trip is not reoccurring, but the trip data can be used for validation and selection of routes. The term "route" is used to indicate the path that is optimized with the model. The model will optimize the refueling decisions based on the gas stations along that path. A route, in this context, consists of one or more consecutive lanes. The complexity of determining optimal refueling points increases with the length of the route, so the recommendation has more impact when combining lanes. However, combining lanes results in more unique routes and routes that are travelled less frequently. Training an RL model for a route that occurs two times is not efficient as the training phase takes long and the policy can only be used twice. To address this, we analyzed shifts over the past six months. Our objective was to determine the optimal number of lanes included in our routes. We focused on lanes exceeding 100 kilometers, and assessed the frequency and total count of unique routes. The decision is informed by considerations of both optimization and efficiency, ensuring that the model is trained on routes with sufficient occurrence to enhance its practical application. The analysis led us to the conclusion that configuring routes to consist of a single lane is currently the most effective approach. When including two lanes, the distance of the most frequent routes did not increase significantly compared to including one lane, however, the frequency decreased and the number of unique routes increased. Including three lanes resulted in longer routes, but the frequency was insufficient to validate the model and impractical for real-world use later on.

2.4. Price Agreements

As demonstrated fuel costs are a significant part of operational expenses for trucking companies like Nijhof-Wassink. To mitigate these costs, transportation companies negotiate price agreements with fuel suppliers, securing discounts per liter of fuel. These agreements vary across different gas stations, making it important to find out at which gas stations they want the highest discount when discussing the price agreements and afterwards refuel at those stations to maximize savings. Nijhof-Wassink has price agreements with two anonymous fuel companies, hereafter referred to as Fuel Company X and Fuel Company Y. Each fuel company has divided their gas stations into network groups based on the location relative to the highway. Fuel Company X has network groups Economy, Super economy, Coverage, Non-core, and Motorway in the Netherlands, Belgium, and Germany. The discounts differentiate per country but are the same for all network groups. Company Y has different discounts for network groups NF1, NT1, NT2, NT3, and NT4 in the Netherlands. In Belgium, the discount varies for different brands. In Belgium, the gas stations are divided into network groups BF1, BT1, BT2, BT3, and BT4. The exact discounts are confidential but typically range from 0 to 16 cents per liter, emphasizing the potential for cost savings. However, a challenge arises as truck drivers are often unaware of the specific price agreements negotiated by the company. Additionally, the pump price, or the price customers pay at the pump, can vary significantly between gas stations. This complexity creates a dilemma for drivers, as they might choose to refuel at a gas station with a seemingly low pump price, assuming that the station has the lowest net price. However, the challenge lies in the fact that the net price, calculated as the pump price minus the discount, could be cheaper at another gas

station since both the pump price and the discount are not known to the driver. For instance, a driver might prefer refueling at "gas station A" with a pump price of €1.40 and a 5 cents discount, while overlooking the better option at "gas station B" with a pump price of €1.46 and a 15 cents discount. Consequently, the model needs to account for these price agreements and prefer the gas station with the lowest net price.

2.5. Take Away

The business understanding phase has provided crucial insights into the context of the research conducted. The business objective of Nijhof-Wassink is to reduce the fuel costs, which can impact the profits significantly. To achieve this they want to explore how increasingly popular AI algorithms can contribute to solving their refueling problem. The ideal situation includes a model that would generate real-time advice on an operational level. Therefore, this research focuses on providing a decision-making framework and evaluating the effectiveness of RL for refueling decisions. In the assessment phase, we aimed to answer the research question *"How are refueling decisions currently made at Nijhof-Wassink?"*. We identified the most impactful factors on the refueling decision being the fuel level, the net price, the detour costs and the time costs for refueling. Furthermore, we mapped out different stakeholders that impact or are impacted by the refueling decision on the different planning levels. This showed that the drivers are responsible for deciding where to refuel. Resulting, in the refueling decisions being made on the experience of the driver without a data-driven approach.

3. Related Work

This research is executed within the logistics sector, focusing on decreasing the fuel costs for freight companies and improving their profits. With this literature review, we aim to answer the knowledge question (RQ2): *"What approaches for solving the refueling problem are present in the literature and how is the framework we propose a contribution to the existing body of literature?"*. The review is conducted according to the guidelines presented in Kitchenham (2004) and the process can be referenced in Appendix A.

The scope of this literature review includes research that aims to reduce fuel costs of trucking companies with a mathematical model. The works are classified by problem domain, meaning each class aims to reduce fuel costs according to a different problem formulation. For each class, different relevant papers are presented and the differences with our work are highlighted to point out the novelty of this research. Within research on reducing refueling costs two problem domains are identified. We have problems that indirectly impact fuel costs by minimizing the distances or CO_2 emissions (Section 3.1), and problems that directly impact refueling costs by advising on where to refuel (Section 3.2). For the problems that directly optimize the refueling costs we have the variable-route vehicle refueling problem (VRVRP) that optimizes refueling decision along a variable route (Section 3.1.2). Within this class there is a subset of problems that consider refueling decisions along a fixed route, fixed-route vehicle refueling problems (FRVRP)(Section 3.1.1). We also see that there are two different methods to incorporate the prices of the fuel. One approach is to have a fixed price over all gas stations, this is usually the case for problems that aim to indirectly decrease the fuel costs. Another approach is to apply varying fuel prices to different gas stations. To our knowledge, all these approaches are assuming deterministic prices. Our problem aims to minimize the direct fuel costs over a set of fixed routes where the prices over gas stations vary. A new element that we could not discover in the existing literature is that our problem adopts sequential decision-making under uncertainty. To the best of our knowledge, we are the first to incorporate varying fuel prices by predicting the prices for the objective function with a Machine Learning (ML) algorithm. The section proceeds by first presenting the work closest to ours, and finishing with relevant work in other areas such as recharging problems (Section 3.3), problems with sequential decision-making under uncertainty (Section 3.4), and MDPs (Section 3.5).

3.1. Problems Directly Reducing Fuel Costs

First, we discuss research that aims to optimize the direct fuel costs by advising where to refuel on a fixed (Section 3.1.1) or variable route (Section 3.1.2). Within this domain, the closest work to ours is Ottoni et al. (2021). They were the first to define the refueling problem as a Markov Decision Process (MDP) and solve it with RL. Ottoni et al. (2021) points out that no study in the current body of literature approaches the refueling model as an MDP with RL, even though this approach has proven to be effective in combinatorial optimization problems (Ottoni et al., 2021). However, the work of Ottoni et al. (2021) is distinctly different from ours on the three following points. Firstly, the model does not incorporate the uncertainty of the prices while the ever-changing prices do contribute to the complexity of the actual environment. Introducing the price as a stochastic element in our problem better reflects reality and can lead to more accurate

predictions. Our contribution is to forecast the prices with a Machine Learning (ML) model and use this predicted price as input for the reward function. Secondly, the work of Ottoni et al. (2021) falls under the class of variable-route vehicle refueling problems (VRVRP) meaning they integrate routing into the solution. Including routing is not relevant for trucking companies who mainly drive long shifts from A to B. Thirdly, our reward function is more complete because, besides the predicted prices, we also incorporate some important aspects such as detour costs and price agreements in addition to the fuel costs.

3.1.1. Fixed-Route Vehicle Refueling Problems

Our problem belongs to the class of fixed-route vehicle refueling problems (FRVRPs) that considers reducing direct fuel costs by optimizing the refueling decisions along a fixed route. FRVRPs are a subclass of variable-route refueling problems where the refueling decisions are made along a variable route. Two relevant works in the area of FRVRPs are Lin et al. (2007), which solves the fixed-path gas station problem with a polynomial time algorithm, and Khuller et al. (2007), which presents a linear-time greedy algorithm for finding optimal refueling policies. Both works are different from ours as they do not take into account the uncertainty in the prices and the detours to the gas stations. Later, Suzuki (2008) expanded the simple FRVRP model by including the minimum purchase quantity and the detour costs but neglecting the stochastic nature of fuel prices. The problem is modelled as a mixed-integer linear program and solved to provide optimal solutions. In recent years, several studies developed variants of the FRVRP of Suzuki (2008). These include (1) a decision support system solved with a heuristic where drivers are free to choose where to refuel (Suzuki, 2009), (2) a heuristic approach that incorporates the dynamic load of a truck into the fuel consumption (Suzuki et al., 2014), and (3) a heuristic approach that successfully reduces the costs by taking into account travel cost by avoiding refueling stops in front of hills and mountains (Suzuki and Lan, 2018). These studies all provide (near) optimal results for the FRVRP but fail to validate and test the model with real data. Therefore, Rodrigues Junior and Cruz (2013) performed a case-study to prove the effectiveness of using the exact FRVRP as proposed in Suzuki (2008) by showing a 2.3% decrease in total fuel costs. Suzuki (2014) points out that exact methods have long run times and can not deal with large instances because of the NP-hardness of the problem. This is a problem as the refueling problem often requires quick near-to real-time solutions as many routes are traveled every day by a trucking company. Suzuki (2014) introduced a variable reduction technique for solving the FRVRP with exact methods. Also, the most recent research of Schulz and Suzuki (2023) presents an efficient exact method for large instances. Both papers deal with big problem instances but none of the variants of the FRVRP deal with varieties in the price. Within the area of FRVRP an RL algorithm was never explored to solve this problem even though it promises to deal with big solution spaces and provide quick solutions once trained. After examining the current body of knowledge on fixed-route vehicle refueling problems we conclude that to the best of our knowledge, we are the first to develop a fixed-route vehicle refueling problem with sequential decision-making under stochastic prices that is solved with Reinforcement Learning.

3.1.2. Variable-Route Vehicle Refueling Problems

The other class of research within optimizing the direct fuel costs studies optimizing the refueling decisions along a variable route. The variable-route vehicle refueling problem (VRVRP) is a variant of the classic vehicle routing problem (VRP) that aims to solve the trade-off between reducing travel distance and saving on fuel costs. In practice, this means a route can be created that is longer in distance but visits cheaper gas stations. Within this class, there are two approaches. The first approach solves the problem by first determining the route and then the refueling decision. Khuller et al. (2007) and Lin et al. (2007) provide examples of this approach, proving the VRP with refueling decisions is NP-hard and solve the problem with heuristics. The second approach tries to improve the solution by combining routing and refueling decisions in one algorithm. One example is the work of Lin (2008) who expands the work of Lin et al. (2007). All studies mentioned before assume that every gas station is located on the route resulting in the work of Suzuki (2009), who created a more comprehensive VRVRP including detours and driver freedom. The work of Suzuki (2009) is one of the few works that mention that price fluctuations cause uncertainty. However, they deal with these fluctuations by solving the model for real-time prices. This approach is not very suitable as the refueling problem can become computationally extensive and fast solutions are required. More recent research is from Neves-Moreira et al. (2020) and Bousonville et al. (2011) who create a multi-period planning-horizon algorithm that integrates routing and decision making. Suzuki and Dai (2013) and Lin (2014) both introduced the VRVRP with a dual objective of minimizing both travel distance and fuel costs as other works tend to cut fuel costs in exchange for increased vehicle miles leading to unwanted policies. Most works that consider a variable route do have a fixed start and end point, however, Farkas and Csehi (2017) aims to optimize the route and refueling decisions by matching instant demand to a network of trucks. When comparing the problems within VRVRP with our problem we find that all the works within VRVRP use different prices for gas stations along the route, but none incorporate the variability and uncertainty of the prices. Furthermore, this class of research is mainly focusing on a planning horizon where the vehicle visits several customers on one trip. For our research, we focus on transportation companies that travel long distances between the start and end point making the FRVRP more suitable. We can conclude that our research is the first to take into account the stochasticity of fuel prices within the body of literature that addresses direct fuel costs.

3.2. Problems Indirectly Decreasing Fuel Costs

Another problem domain within refueling problems aims to indirectly decrease fuel costs by reducing fuel consumption. Kuo (2010) developed a VRP that aims to reduce fuel consumption and takes into account the speed and loading weight. Solving this problem with simulated annealing resulted in a reduction in fuel consumption of 24.61%. Another study conducted by Zhang et al. (2015) integrated fuel consumption into a VRP, revealing not only a reduction in fuel consumption but also a trade-off relationship among fuel consumption, carbon emissions, and vehicle operating costs. Furthermore, the research of Xiao et al. (2012) is a good example of how taking into account fuel consumption by adding the fuel consumption rate as a load-dependent function, can reduce fuel costs by decreasing consumption by 5%. It is important to highlight that the works in this

class assume one single price for fuel while in practice vehicles can refuel at different prices. Therefore, we want to point out that minimizing fuel consumption is not the same problem as minimizing fuel costs for the same fuel consumption. Neves-Moreira et al. (2020) shows that solving problems that indirectly minimize fuel costs by using a single fuel price leads to sub-optimal solutions when the vehicle can refuel at different prices. In our work we not only deal with the regional price differences but also with the price differences caused by the price agreements from trucking companies. We believe the price differential between gas stations is so significant that taking into account both the location of fueling stations and the price of gas is not only reasonable but pivotal for achieving efficient operations. On top of that, we incorporate the variability of the fuel prices in the reward function, by forecasting them with ML.

3.3. Problems Similar to Refueling Problems

There is a stream of similar research to ours that studies the application of RL algorithms for the recharging decision problem of Electric Vehicles (EVs). From the review paper Abdullah et al. (2021), we conclude that the problems with single agents and the objective to minimise the charging costs for the EV owner are closest to our work. For instance, Li et al. (2020) and Chiş et al. (2017) are particularly relevant as they both applied RL successfully to find a constrained charging/discharging scheduling strategy to minimize the charging cost for the driver as well as guarantee the EV can be fully charged. They also take into account price uncertainty by predicting electricity prices for the upcoming days. Chiş et al. (2013) executes similar research to optimize profit for owners of Plug-in EVs. A difference with our work is the focus is on which day to recharge instead of which charging station. Shi and Wong (2011) and Zhang et al. (2021) study the real-time EV control problem under price uncertainty, meaning they provide real-time charging advice to EV drivers. Both works incorporate uncertain electricity prices and formulate the problem as an MDP. The results of the discussed works show that RL algorithms can work effectively in an environment with uncertain prices and that they can increase profit significantly.

3.4. Sequential Decision-Making Under Uncertainty

For our refueling problem, we develop a framework for sequential decision-making under uncertainty. The framework incorporates that at each point in time when a driver arrives at a detour point to a gas station, he stands for the decision to refuel or not. Sequential decision-making under uncertainty has been widely applied in operations research, ML, and computer science (Diederich, 2001). In sequential decision-making, "control theory" typically handles problems with continuous states and decisions, while problems, including ours, with discrete states and decisions in discrete time fall under "Markov decision processes" (Powell, 2011). Discrete-time MDPs, hereafter just referred to as MDPs, consist of some key elements namely the state space, the action space, the transition probabilities and the reward function. The state space evolves over discrete time periods, where at each observed state an action is chosen from the action set. After this action, two things will happen (1) the system receives the reward for that action, and (2) the system evolves to the next state at the next time period with a transition probability (Puterman, 1994; Powell, 2011; Sutton and Barto, 2018). MDPs can be straightforward to formulate, but solving them is another matter. In Section 5 we present the different exact and approximate solution methods and their relevant applications.

3.5. Related Work for Markov Decision Processes

Examples of works for MDPs in operations research can be seen in transportation-, inventory-, and energy problems. Inventory problems are closest to our problem as the fuel in the tank can be seen as our inventory level and in each decision epoch the new inventory level is decided on by buying more products. An example of an inventory problem with sequential decision-making is the work of Shin and Lee (2015), who defined the inventory management problem with multiple suppliers and supply and demand uncertainty as an MDP. The case study showed that exact value iteration and approximate methods reduced the costs compared to the popular method that disregarded the uncertainty and consideration of multiple criteria. Another work that uses the MDP framework for an inventory problem is from Ahiska et al. (2013). They solve the stochastic inventory control problem with unreliable sourcing showing a performance gain compared to conventional methods.

Lin et al. (2007) and Suzuki (2014) also mentioned the FRVRP can be viewed as a special case of the single-item, single-resource capacitated lot-sizing problem (CLSP), with inventory bounds and fixed lot size. Some efficient algorithms exist that can solve the deterministic CLSP with inventory bounds (Gutiérrez et al., 2003) or the CLSP with minimum lot sizes (Okhrin and Richter, 2011). Atamtürk and Küçükyavuz (2005) performs a polyhedral computational study on the CLSP with inventory bounds and fixed costs. However, we are not aware of any efficient algorithm that can solve the CLSP with both the inventory bound and the minimum lot size. An overview of deterministic CLSPs can be found in Gicquel et al. (2008). Within the CLSP there are also works that address stochastic single-item dynamic lot sizing problems and incorporate uncertainty in demand. The review paper of Sox et al. (1999) shows that stochastic CLSPs are often solved to optimality with linear models and near-optimal with heuristics. The works of van Hezewijk et al. (2022) and Dellaert and Melo (1996) approach the problem differently and formulate the stochastic CLSP as an MDP. van Hezewijk et al. (2022) solves the MDP with deep RL and Dellaert and Melo (1996) solve the MDP with heurstics.

3.6. Take Away

In this literature review, we answered the knowledge question: *"What refueling models are present in the literature, what features do they have, and how is the framework we propose a contribution to the existing body of literature?"*. To answer this question a comprehensive overview of refueling models within the logistics sector, focusing on reducing fuel costs for freight companies is presented. We classified existing research into two main problem domains: those indirectly impacting fuel costs by minimizing distances or emissions, and those directly optimizing refueling decisions. Our work falls within the latter, specifically the FRVRP, aiming to minimize direct fuel costs along fixed routes with varying fuel prices. We identified a gap in the literature, as existing approaches often assume deterministic prices, overlooking the stochastic nature of fuel prices. Deterministic prices decreases the complexity of the problem, however, a deterministic model might not capture the range of possible outcomes and make overly positive or negative predictions. Our proposed sequential decision-making framework entails defining the FRVRP as an MDP and solving it with RL while incorporating a novel approach using machine learning for price prediction.

4. Mathematical Model

This section answers the research question *"What is the MDP formulation of the fixed-route vehicle refueling problem with stochastic fuel prices?"* by modeling the refueling problem as an MDP. The MDP for the refueling problem aims to optimize refueling decisions along a set of fixed routes, considering varying fuel prices at different gas stations, to minimize the overall refueling costs for a truck. This type of problem is called the Fixed Route Vehicle Refueling Problem (FRVRP) (Lin et al., 2007; Suzuki, 2008; Khuller et al., 2007).

4.1. Markov Decision Process

An MDP is a mathematical framework for addressing problems involving sequential decision-making under uncertain conditions. It is well-suited for addressing the refueling problem, where uncertainties in fuel prices and the need for sequential decisions along a route come into play. Due to minimal overlap between routes and the natural start and termination point of a route, we present a finite horizon MDP that can model the refueling problem for all the independent routes. The MDP considers the variable refueling costs at each gas station, the different detour costs to reach a gas station, the benefits of price agreements with certain gas stations, and the fixed costs for stopping to refuel. Importantly, the uncertainty and variability of the prices are incorporated as explained in Section 6.1.3. In the MDP model, we define the state as the fuel level of the truck in liters and the action is whether to refuel or not at a nearby gas station. Given the present state and the taken action, we can find the future state using the state transition function. The solution of the MDP is a policy that determines which action to take in each state so that the total expected reward, which is the negative counterpart of the refueling cost, is maximized. All sets and parameters and their notation can be found in Table 1 at the end of this paragraph. The problem is formulated as follows.

4.1.1. Decision Epochs

The decision epochs represent the discrete points when decisions are made. The total number of decision epochs depends on the route. Therefore, to define the set of decision epochs for routes, we use the set $T = \{1, 2, \ldots, G\}$ where G is the last fuel station along the route r. In this model, the decision epochs are detour points to gas stations on the routes. The choice of decision epochs may lead to having non-equidistant times between two consecutive decision epochs.

4.1.2. States

The state of the system at point t, $s_t \in S = \{L, L + \delta, L + 2\delta, \ldots, U\}$, for all decision epochs $t \in T$. The state indicates the fuel left in the tank (in liters) at decision epoch t, where L and U denote the lower bound and upper bound of the fuel level in the tank, respectively. The increment level is δ, which is used to discretize the state space of the system.

4.1.3. Actions

The action of the system at each decision epoch $t, a_t \in A = \{0, 1\}$, where action 0 means no refueling and action 1 means refuel up to the max level. We note that in theory, it is possible to define the larger action space such that $a_t \in A_t = \{0, U - s_t\}$. However, inspired by the drivers' behavior that makes their tank completely full when they stop on long-distance trips, we made this choice that helps to reduce the size of the problem as well.

4.1.4. Transitions

The state of a system transitions from state s_t to s_{t+1} as a function of present state s_t, and action a_t, according to Equation (1). The transition is deterministic as the element of uncertainty is absent in the state and action. We note that the element of uncertainty is the price that impacts the reward of the system, which is discussed later. The fuel level of the next state plus the fuel needed for the detour to the next gas station can not be lower than L. The amount of fuel needed from the present detour point to the next detour point is usage (u_t), which is subtracted from the present fuel level (s_t). In case the refueling action is taken, then the amount of fuel used inside the detour d_t also needs to be subtracted (in addition to u_t) from the max level of fuel in the tank to account for the fuel needed to return to the route.

$$s_{t+1} = max\big(L, (1 - a_t)(s_t - u_t) + a_t(U - u_t - \frac{1}{2}d_t)\big) \tag{1}$$

4.1.5. Reward

The goal is to maximize the total expected reward. To calculate the total expected rewards we need the immediate rewards of taking action a_t in state s_t at decision epoch t. The immediate reward is defined as the negative of the refueling costs at decision epoch t, which is a function of s_t, s_{t+1}, a_t, and the uncertainty element $\hat{p}_t$. The refueling costs are determined by the total usage for the trip $(\sum_{t=1}^{G} u_t)$, the length of the detour (d_t), the time costs of refueling (C), the time costs for the detour (b_t), and the discount per liter (k_t). As can be noted, we only charge the liters that are used during the trip while the first logical calculation that comes to mind is to include all fuel that is refueled because you pay for it. However, this may result in the model preferring refueling at the beginning of the route because the tank level is higher, and therefore fewer liters are charged. This can result in refueling earlier on the trip at stations with higher fuel costs, while later cheaper options are available. Alternatively, we could deduct the remaining fuel inside the tank that remained unused at the end of the route, but this option is less preferred from the practical perspective as the destination of one route could be the origin of another route. If the next fuel level turns out to be less than L, a huge penalty, $-M$ is assigned as the truck cannot continue the trip to the destination point. We calculate the immediate rewards using Equation (2).

$$r_t(s_t, a_t, s_{t+1}, \hat{p}_t) = \begin{cases} -(a_t((\hat{p}_t - k_t)(\sum_{t=1}^{G} u_t + d_t) - C - b_t)) & \text{if} \quad s_{t+1} > L, \\ -M & \text{if} \quad s_{t+1} \leq L. \end{cases} \tag{2}$$

Where $\hat{p}_t - k_t$ is the net fuel price per liter at decision epoch t, d_t is the amount of fuel consumed to reach the gas station inside the detour, C is the fixed time cost for stopping

to refuel, and b_t is the time costs for a detour. The fuel price $\hat{p}_t$ is exogenous and forecasted by a function dependent on the date, country and location. Further details on the forecast are provided in Section 6.1.3. The time costs for a detour are taken into account to cover the salary of the driver for making a longer detour to a gas station. The terminal reward is the same as immediate rewards in other decision epochs. The total reward can be calculated by using the state and time-dependent decision rules. We denote the state and time-dependent decision rules, $h_t(s_t) : s_t \rightarrow A_{s_t}$, and use them to indicate which action a_t to select when in state s_t at decision epoch $t \in T$. A sequence of these decision rules denotes a policy π. The expected total reward of policy π is $v_G^\pi(s_1)$ and equals

$$v_G^\pi(s_1) = \mathbb{E}_{s_1}^\pi \left[r_t(s_t, a_t) \right] \tag{3}$$

The aim is to find a policy π^* with the maximum expected total reward.

Table 1: Summary of model sets and parameters

Notation	**Definition**
Sets	
T	Set of decision epochs representing detours points to gas stations
S	Set of possible states with lower bound L, upperbound U, and step-size δ
A	Set of actions that can be taken
Parameters/variables	
t	Decison epoch in set T
s_t	State from set S the system is in at epoch t
a_t	Action from set A that is chosen at time t
r_t	Immediate reward for taking action a_t, in state s_t
u_t	Usage to go from decision epoch t to $t+1$
d_t	Usage to go to the gas station at decision point t
$\hat{p}_t$	Predicted price at the gas station located at decision epoch t
k_t	The discount per liter at the gas station located at decision epoch t
b_t	The costs for the time it takes to make the detour at decision epoch t
C	Constant costs for the time it takes to stop for refueling
$-M$	Big penalty assigned when the truck runs out of fuel

4.2. Assumptions and Simplifications

When defining the model several assumptions and simplifications were made. These are listed and explained below:

i Start and end of route at first and last detour point.
In practice, the route starts from the origin to the destination. Without loss of generality, we consider the starting point of the routes to be from the first detour located close to the first gas station on each route, and the ending point of the routes to be from the last detour point.

ii Deterministic fuel usage.
To simplify our model, we assume a deterministic fuel usage pattern for the trucks. This means that the amount of fuel consumed by the trucks is considered to be constant and predictable. While this simplification facilitates the modeling process, it is essential to be aware of its deterministic nature when interpreting the results.

iii The time to refuel is constant at every gas station and for each refuel size.
While in reality some stations have faster fuel lanes for trucks and refueling more liters takes longer, the model assumes each refueling takes the same amount of time.

iv The truck has to be refueled completely.
In practice a driver can choose to refuel the amount he desires. However, for the sake of reducing the size of the problem, the refueling decision is simplified to not refueling or refueling up to the max level.

v The fuel level of the truck is discretized per liter.
To facilitate computational efficiency and practical implementation, we discretize the fuel level of the truck per liter. This discretization approach allows for a more manageable representation of fuel quantities in our model, striking a balance between computational complexity and accuracy.

4.3. Take Away

The MDP for the FRVRP captures the refueling behavior of drivers by aligning the decisions epochs with detour points to gas stations and the possible actions at each epoch being no refueling or refueling up to the max level. The state of the MDP is denoted by the fuel level. We keep things straightforward by assuming fuel usage between decision epochs is predictable. The MDP transitions from one state to another by subtracting the fuel usage from the fuel level and adding the refuel amount (if applicable). Immediate rewards, composed of fuel and detour costs, quantify the financial implications of each action in each state and navigate the model to an optimal policy. In the reward function, forecasted fuel prices inject a crucial element of uncertainty. The total expected reward of a policy is the sum of the immediate rewards over all decision epochs.

5. Solution Methods

The first part of this section dives into various solution approaches for MDPs and additional topics such as policies, step-sizes, and the exploration-exploitation trade-off (Section 5.1). The subsequent part of this section is dedicated to addressing the research question *"How can we solve the MDP of the refueling problem and gain (near-)optimal results?"* (Section 5.2). This section introduces four distinct solution methods: a reinforcement learning algorithm, an approach for establishing an upper bound, a benchmark heuristic, and lastly, an examination of the historical decisions made by drivers.

5.1. Introduction of Solution Methods for Markov Decision Processes

MDPs can be straightforward to formulate, but solving them is another matter. Dynamic programming offers a framework in which MDPs can be solved. A standard solution method, which is an exact algorithm, is backward induction (Puterman, 1994). However, the state size of the refueling problem can grow large and the stochasticity of the fuel prices introduces an extra layer of complexity. Additionally, exact methods within dynamic programming presuppose complete knowledge of the MDP, encompassing transition probabilities and rewards, which may not always be obtainable. Thus, as is common with large and complex problems, we divert to approximate dynamic programming for a suitable solution approach (Powell, 2011). For solving the MDPs with approximate methods, we discuss three algorithmic strategies. Next, we introduce the policy for choosing an action, and how we approximate the value function. Lastly, we elaborate on how we tackle the challenges of exploration vs. exploitation and the learning rate.

5.1.1. Approximate Algorithmic Strategies for Solving Finite Markov Decision Processes

Approximate dynamic programming is a powerful tool for addressing difficulties in solving large problems. However, its applicability extends to smaller problems that become challenging when lacking a formal model of the information process or when uncertainty exists about the transition function. For instance, in our problem, we have observations of changes in fuel prices but we do not have a mathematical model that describes these changes. Approximate dynamic programming involves progressing forward in time, posing two challenges to overcome. First, we need sample paths of fuel prices, that forecast what might happen in the future (See Section 6.1.3). A sample path (ω^n) refers to a particular sequence of exogenous information generated for episode n of the algorithm. Secondly, we require a policy to determine the way of decision-making as we advance to the next state (See Section 5.1.3). As for solving an approximate dynamic program, we present three major algorithmic strategies that have evolved. The first approximate strategy is Reinforcement Learning (Sutton and Barto, 2018). RL is suitable for large-scale problems, especially when not all elements of the mathematical model are fully known or captured (e.g. the full transition probability is hard to be determined). The second strategy is a Real-Time Dynamic Program (RTDP). RTDP is a variation of a dynamic program designed to converge and make real-time decisions in small state-action space problems. The third strategy is Approximate Value Iteration (AVI). AVI relaxes the assumption of computing the one-step transition matrix, allowing for problem-solving in situations with challenging transitions (Powell, 2011). Another method to solve the finite

MDP is with heuristics. Heuristics are algorithms that involve relatively simple problems with practical strategies. The performance of a heuristic depends on the problem.

5.1.2. Reinforcement Learning

From the three approximate strategies, the RL framework proved ideal for sequential decision-making in unknown environments with large amounts of data (Gupta et al., 2022; Sutton and Barto, 2018) and it has shown to be an effective tool for complex, online, stochastic, and large-scale problems such as refueling decisions (Abdullah et al., 2021; Giannoccaro and Pontrandolfo, 2002; Nazari et al., 2018). Also, when we look at large-scale inventory problems, Shakya et al. (2022) shows that an RL approach for multi-period inventory with stochastic demand outperformed other methods. Reinforcement learning is a machine learning training method based on rewarding desired behaviours and penalizing undesired ones. In RL, the goal is to learn policy π that optimizes the expected rewards, where a policy is a sequence of actions based on the state in time-step i (Sutton and Barto, 2018). Within RL, SARSA and Q-learning are two popular algorithms used to learn optimal policies for agents in an environment (Powell, 2011; Sutton and Barto, 2018). While the algorithms share similarities, a key distinction is that SARSA learns on-policy meaning the action selected for updating the value function is used in the next decision epoch and Q-Learning is off-policy meaning the action selected for the update function may be different than the action chosen in the next state. More elaboration on both algortims follow in Section 5.2.1. The actual performance difference between the two algorithms can vary based on the specific characteristics of the environment, so experimenting might be needed to determine which algorithm performs better. When looking at applications of these both algorithms in relevant work we see that Q-learning is more commonly used for RL-based EV charging management systems (Abdullah et al., 2021), but no papers compare both algorithms. The traveling salesman problem with refueling from Ottoni et al. (2021), shows that for different problem instances, both algorithms work equally well. Therefore, in this research, we test both SARSA and Q-Learning to see which algorithm performs better.

5.1.3. Policy and Approximation of the Value Function

When solving MDPs with an approximate solution method such as RL, we need to establish the policy that dictates which action to choose in a given state. Powell (2011) discusses four different policies: a myopic policy, a lookahead policy, a policy function approximation, and a value function approximation. When adopting a myopic policy, the action is based solely on the current state, without considering the long-term consequences or future rewards associated with its actions. Therefore, if future rewards hold significance, this policy may prove less suitable. On the contrary, a lookahead policy considers the future by explicitly simulating or considering potential future states and their consequences. This is however computationally expensive. The two remaining policies are value- and policy-function approximation, where value-function approximation focuses on estimating the value of states (V values) or state-action pairs (Q values), and policy-function approximation focuses on the value of a complete policy π. Value function approximation is the most commonly used policy for approximate solution methods (Puterman, 1994; Powell, 2011). In RL, the model often aims to find the state-action values (Q values), which is why we use value function approximation as our policy (Sut-

ton and Barto, 2018; Powell, 2011).

Several strategies exist for how these values are stored and used in the value function approximation: look-up tables, parametric representations, and non-parametric representations. Look-up tables store the value approximations per state-action pair. This is a straightforward but effective method. However, the matrix can become large when the state-action space grows significantly. Parametric representations solve this issue by estimating value functions using regression methods. However, the drawback is that they are only effective if you can design an effective parametric model. Non-parametric representations avoid the challenge of parameter tuning and offer tremendous potential, but significant hurdles remain regarding complexity, data availability, and good approximations (Powell, 2011). The size of our state-action space allows us to use a look-up table, so we select this simple but effective approach.

5.1.4. Learning Rate and Exploration vs. Exploitation Trade-off

Using an approximate method to solve the MDP presents several challenges, with the learning rate and the exploration-exploitation trade-off being among the most significant ones (Sutton and Barto, 2018; Powell, 2011; Mes et al., 2017). First, the learning rate is important as the core of approximate dynamic programming is some form of iterative learning. In value function approximation, the approximations are updated with every episode, but how much of the newly observed value is used for this approximation is determined by the learning rate. A high learning rate puts more weight on the new observations and, conversely, a low learning rate puts more weight on past observations. Both settings have their benefits and drawbacks. A too high learning rate introduces instability, while a too low learning rate causes slow convergence. The selection of an appropriate learning rate is paramount for an RL algorithm's success, influencing the overall learning process (Powell and Ryzhov, 2012). Typically, an RL algorithm requires different levels of the learning rate during training. Powell (2011) presents various step sizes to dynamically adjust the learning rate during training, including deterministic, stochastic, and optimal step sizes. The deterministic step size that is used in this research is the straightforward yet efficient Generalized Harmonic Step Size (See Figure 6). This step size is suitable for value iteration and RL, because of its larger step size (See Equation (4)).

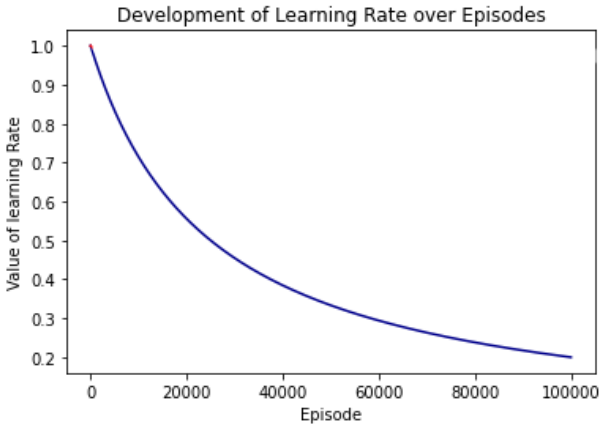

Figure 6: Development of learning rate under harmonic step size with $\Delta = 22500$

$$\alpha_{n-1} = \frac{\Delta}{\Delta + n - 1} \tag{4}$$

In navigating the exploration-exploitation trade-off for optimal performance, both algorithms employ the epsilon-greedy method. This strategy, outlined in Powell (2011), involves pure exploration ϵ percent of the time and pure exploitation the remaining $1 - \epsilon$ percent. The epsilon value is dynamically adjusted every episode using the Generalized Harmonic Step Size.

5.2. Outline of Selected Solution Methods

This section continues by outlining the different solution methods we adopt to determine optimal or near-optimal policies for the FRVRP. In the first step, we present the RL algorithms used for solving the MDP. Then, three methods are introduced to compare the RL algorithm with. We start by presenting a method that provides an upper bound, then we introduce a benchmark heuristic and, lastly, we compare the RL model with historic decisions. The purpose of developing different solution methods is to assess the performance of RL for the refueling problem.

5.2.1. Reinforcement Learning

This section introduces the Q-Learning and SARSA RL algorithms for the FRVRP. As discussed before, both algorithms adopt an ϵ-greedy method for action selection and a Generalized Harmonic step size that adjusts the ϵ-greedy and learning rate over the training period. We opt to test both algorithms, recognizing that their performance depends on specific problem characteristics, and the literature lacks a consensus on which one outperforms the other in a similar context.

Algorithm 1 and Algorithm 2 present the general pseudo code for SARSA and Q-Learning, respectively. While the algorithms share similarities, a key distinction arises from Q-Learning being off-policy and SARSA being on-policy. This distinction manifests in two aspects: the update function for the state-action approximation and the timing of action selection for the next state. Equations (5) and (6) illustrate the update functions for Q-Learning and SARSA. Q-learning converges to the optimal policy as it maximizes rewards using a greedy action selection strategy for the highlighted part of the update function (Equation (5)). In contrast, SARSA employs an ϵ-greedy strategy, allowing it to include exploration in the highlighted part (Equation (6)). Furthermore, the update functions show that the value approximation depends on both the action taken in the current state (a_t) and the action selected in the next state (a_{t+1}). In SARSA, the action for the next state is determined in line 9 before updating the state-action value in line 10. Conversely, in Q-Learning, the next action is not determined before updating the state-action value in line 9. In Q-Learning, the update function selects the next action based on the one that yields the highest expected reward. This distinction reflects the off-policy nature of Q-Learning, where the learning policy may differ from the target policy and the on-policy nature of SARSA, where the learning policy may not differ from the target policy.

$$Q_{n+1}(s_t, a_t) = Q_n(s_t, a_t) + \alpha[r(s_t, a_t) + \gamma \max_{a_{t+1}} Q_n(s_{t+1}, a_{t+1}) - Q_n(s_t, a_t)] \quad (5)$$

$$Q_{n+1}(s, a) = Q_n(s_t, a_t) + \alpha[r(s_t, a_t) + \gamma Q_n(s_{t+1}, a_{t+1}) - Q_n(s_t, a_t)] \quad (6)$$

Algorithm 1 SARSA algorithm

1: Set the parameters: α, γ, and ε
2: Initialize the matrix $Q(s, a) = 0$ for all pairs s,a
3: Observe the state s_t
4: Select the action a_t using ε-greedy method
5: **repeat**
6: Take the action a_t
7: Receive immediate reward $r(s_t, a_t)$
8: Observe the new state s_{t+1}
9: Select the new action a_{t+1} using ε-greedy method
10: Update $Q(s_t, a_t)$ with (6)
11: $s_t = s_{t+1}$
12: $a_t = a_{t+1}$
13: **until** stopping criterion is satisfied

Algorithm 2 Q-Learning algorithm

1: Set the parameters: α, γ, and ε
2: Initialize the matrix $Q(s, a) = 0$ for all pairs s,a
3: Observe the state s_t
4: **repeat**
5: Select the action a_t using ε-greedy method
6: Take the action a_t
7: Receive immediate reward $r(s_t, a_t)$
8: Observe the new state s_{t+1}
9: Update $Q(s, a)$ with (5)
10: $s_t = s_{t+1}$
11: **until** stopping criterion is satisfied

Now that the differences are highlighted we explain SARSA (Algorithm 3) and Q-Learning (Algorithm 4) for the FRVRP line by line. Both algorithms start by initiating the look-up table $Q(s, a)$, the discount factor γ, the learning rate α, epsilon-greedy exploration parameter ϵ, and the number of episodes N (lines 1-2). The algorithm is an iterative process that continues until the total number of episodes N is reached (line 3). Within each episode, a new sample path of prices for the gas stations in the route is forecasted (line 4). Also, the learning rate α and epsilon-greedy parameter ϵ are updated using the Generalized Harmonic Step size, and an initial state is chosen (lines 5-7). Additionally for SARSA, the first action is determined at this stage (line 8). Both SARSA and Q-Learning then proceed by looping through the decision epochs. In the decision epochs, Q-Learning first selects an action (line 9) and then both algorithms continue by executing the chosen action, observing its reward and calculating the next state (lines 10-12). Subsequently, Q-Learning updates the lookup table (line 11). In contrast, SARSA chooses the action for the next state in line 11 before updating the lookup table in line 12. After updating the lookup table, both algorithms prepare variables for the next decision epoch by setting the next state to the current state. For SARSA, the next action is also set as the current action. This process continues until all episodes are executed, resulting in a lookup table capturing learned values that can be used to derive the optimal policy for both algorithms.

Algorithm 3 SARSA Algorithm for the FRVRP

Initialize the look-up table $Q(s, a)$ with zeros
Initialize discount factor γ, α for learning rate, ϵ for epsilon greedy, and number of episodes N
for n = 0, ..., N-1 **do**
    Get sample price path ω^n from forecast function
    Update learning rate: $\alpha = \frac{\Delta}{\Delta+n-1}$
    Update epsilon greedy: $\epsilon = \frac{e}{e+n-1}$
    Select initial state s_0^n
    Choose action a_0^n from s_0^n using ε-greedy method
    for t = 0, ..., G-1 **do**
        Take action a_t^n
        Observe reward $r(s_t^n, a_t^n, \hat{p}_t^n)$ Eq. (2)
        Transition to next state s_{t+1}^n Eq. (1)
        Choose action a_{t+1}^n from s' using ε-greedy method
        Update look-up table: $Q(s, a)$ Eq. (6)
        Update action: $a_t^n = a_{t+1}^n$
        Update state: $s_t^n = s_{t+1}^n$
    end for
end for

Algorithm 4 Q-Learning Algorithm for the FRVRP

Initialize the look-up table $Q(s, a)$ with zeros
Initialize discount factor γ, α for learning rate, ϵ for epsilon greedy, and number of episodes N
for n = 0, ..., N-1 **do**
    Get sample price path ω^n from forecast function
    Update learning rate: $\alpha = \frac{\Delta}{\Delta+n-1}$
    Update epsilon greedy: $\epsilon = \frac{e}{e+n-1}$
    Select initial state s_0^n
    for t = 0, ..., G-1 **do**
        Choose action a_0^n from s_0^n using ε-greedy method
        Take action a_t^n
        Observe reward $r(s_t^n, a_t^n, \hat{p}_t^n)$ Eq. (2)
        Transition to next state s_{t+1}^n Eq. (1)
        Update look-up table: $Q(s, a)$ Eq. (5)
        Update state: $s_t^n = s_{t+1}^n$
    end for
end for

5.2.2. Optimal Solution

The first method adopted to demonstrate the performance of the RL algorithm is an exact method. We aim to find the upper bound for maximizing the negative costs of the refueling problem by assuming the future prices are known to attain the optimality gap of the RL algorithm. We can find the optimal solution by solving an Integer Linear Program (ILP), calculating the reward for each starting fuel level. The solution space is represented as $S \times A^G$. Since the action space (A) has size 2 and the state space (S) has size 370, a problem instance where $G = 130$, has a solution space of $370 \times 2^{130} = 5 * 10^{41}$. However, given that all routes fall under 1161 km and the model is required to fill up the tank completely, the truck can always complete the route without requiring a second refueling. Furthermore, a second refueling is never preferred due to the additional costs it incurs. Consequently, we can add a constraint that restricts the number of refuels on a route to reduce the problem size. The exact solution is calculated for every price sequence and compared to the result of the FRVRP RL algorithm.

5.2.3. Benchmark Heuristic

The second method used to reflect on the performance of the RL algorithm is a benchmark heuristic. A benchmark heuristic refers to a heuristic that is used as a baseline or reference point for comparison with other algorithms or methods. In the context of optimization problems, a benchmark heuristic is often a simple heuristic that can provide a reasonable solution quickly. It serves as a benchmark against which the performance of other algorithms, in this case the RL algorithm, can be compared. Our benchmark heuristic takes into account the same characteristics as the RL algorithm except that the heuristic decisions are fixed and not impacted by uncertainty that is realized after the decision is made. The logic applied is that the truck passes every gas station and at every gas station the algorithm evaluates if the fuel level is below the lower bound because then the driver needs to refuel at the current gas station. Otherwise, if the fuel level drops below the lower bound within the next Z gas stations, the algorithm will choose the cheapest of those gas stations to refuel. The algorithm can be found in the pseudo-code presented in Algorithm 5.

The algorithm begins by initializing the route, sample price path, and the number of look-ahead stations, denoted as Z (lines 1-3). The algorithm iterates through different fuel levels, starting from lower bound L up to upper bound U with an increment of δ (line 4). Within this loop, it further iterates through gas stations (line 6). The primary objective is to make decisions regarding refueling at gas stations to optimize the overall refuel cost. If the fuel level is below the lower bound L (line 8), the algorithm refuels at the current gas station (line 9), calculates the immediate reward based on Equation (2) (line 10), and proceeds to the next station. Suppose the fuel level in the next Z stations reaches below the lower bound (line 11). In that case, it explores the upcoming Z stations to find the refueling option with the lowest forecasted refueling costs, and updates the refueling station, reward, and other related variables (lines 13-19). When the refueling takes place at the current gas station the fuel level is set to the upper bound minus the fuel needed to drive back to the route (lines 22-25). The algorithm continues this process, adjusting the fuel level at each step (line 27) until all fuel levels and gas stations are considered.

Algorithm 5 Benchmark Heuristic Algorithm

Initialize route
Initialize sample price path
Initialize number of look-ahead stations Z
for $f = L, L + \delta, L + 2\delta, \ldots, U$ **do** ▷ Loop through fuel levels
    Initialize $refuel$ to False
    for $t = 0, \ldots, G - 1$ **do** ▷ Loop through gas stations
        Calculate fuel needed to reach next five gas stations: $usageZ$
        if $f \leq L$ **then** ▷ If the fuel level is under the lower bound
            Refuel at current gas station: $refuel = t$
            Calculate the $reward$ according to Eq. (2)
        else if $f-$usageZ$\leq L$ and $refuel =$ "false" **then** ▷ If fuel level in Z stations is under the lower bound
            Initialize $reward$ to $-M$
            for $station$ in range$(t, \min(t + Z, G))$ **do** ▷ Loop through next Z stations
                Calculate the $temp\_reward$ for $station$ according to Eq. (2)
                if $temp\_reward > reward$ **then** ▷ Find cheapest station
                    Set $reward$ to $temp\_reward$
                    Refuel at current gas station: $refuel = t$
                end if
            end for
        end if
        if $t =$ refuel **then** ▷ If refueling takes place, up fuel level
            Set fuel level f to upper bound minus detour usage
            Set $refuel$ to False
            Total reward $+= reward$
        end if
        if $t < G$ **then** ▷ If we are not at the last station, set new fuel level
            Set fuel to new state: $f -$ usage to next station
        end if
    end for
end for

5.2.4. Historical Decisions

The third method involves evaluating the historical decisions made by drivers to demonstrate the practical contribution of the model in making superior decisions. Utilizing historic trip data, we compile an overview that entails the refueling occurrences at the routes we trained. This overview includes the date, the fuel level at the start of the trip, and the gas station the refueling took place. To compare the outcomes of the historic decision and the RL algorithm, we determine the gas stations recommended by our RL algorithm for refueling. Please note that as we derived the historical prices from transactions, we do not have fuel prices for every possible refueling decision on a certain route at the time of refueling. Therefore, we generate different sample paths of forecasted prices for every route and gas station. Utilizing forecasted sample paths still provides valuable insights into the potential cost savings of the model.

5.3. Take Away

This section has introduced various solution methods for the MDP and aimed to address the research question, *"How can we solve the MDP of the refueling problem and gain (near-)optimal results?"*. The proposed framework in this research adopts an RL approach to yield near-optimal outcomes. Within this RL framework, the harmonic step size is employed for the learning rate, and actions are chosen using the epsilon-greedy method, effectively balancing exploration and exploitation. A value function approximation determines the expected values of actions, and these state-action values are stored in a lookup table. For the RL approach, two algorithms are chosen: the SARSA algorithm and the Q-Learning algorithm. Since each algorithm has advantages and drawbacks in different scenarios, this research evaluates both for a comprehensive understanding of their performance. The validity of the RL approach is established by comparing it to three other methods introduced in this section: an optimal solution, a benchmark heuristic, and historical refueling decisions. This comparative analysis aims to assess the effectiveness and efficiency of the RL approach in the context of the refueling problem.

6. Case Study

This case study aims to bridge theoretical knowledge with practical application, offering insights and informing decision-making processes. We build the case study following multiple steps inspired by CRISP-DM. First, we understand, collect, prepare and construct the data for the input of our model (Section 6.1). Afterwards, we select and execute the modeling techniques for the forecast of the fuel prices (Section 6.2). Then the parameters for both the Q-Learning and SARSA algorithms are tuned for optimal performance (Section 6.3). The next steps consider testing and validating the RL model (Section 6.4). We start by comparing the performance of the Q-learning- and SARSA algorithms and selecting the best-performing algorithm for further testing (Section 6.4.1). Next, to validate the robustness of the algorithm the model is exposed to 10 unseen routes and the performance is verified (Section 6.4.2). Lastly, the RL model is tested against a benchmark heuristic (Section 6.4.3), an optimal solution (Section 6.4.4), and the historic driver decisions (Section 6.4.5). The case study concludes by showcasing insights derived from the RL policies (Section 6.5).

6.1. Data

This section answers the research question *"What data is needed as input for the case study and how is the data collected, prepared and constructed?"*. First, Section 6.1.1 focuses on identifying, collecting, and analysing the data sets. Second, the final data sets are prepared for modeling in Section 6.1.2. Third, in Section 6.1.3 the data is processed to construct the datasets that can be used as input for the model.

6.1.1. Data Understanding

This section dives into the data understanding phase of CRISP-DM. First, we identify which data is still needed guided by the model parameters. Afterwards, we collect and present the raw datasets and their features. From the mathematical model, we derived the parameters and sets that still need to be defined.

1. The set of routes: $R = \{1, 2, \ldots, Z\}$
2. The set of decision epochs: $T = \{1, 2, \ldots, G\}$
3. The parameters U, L and δ for the state space: $S = \{L, L + \delta, L + 2\delta, \ldots, U\}$
4. The amount of fuel needed from the present detour point to the next detour point: u_t
5. The amount of fuel used inside the detour: d_t
6. The costs regarding extra time in the detour: b_t
7. The uncertain price element: $-\hat{p}_t$
8. The big penalty: $-M$
9. The constant costs for stopping to refuel: C

To establish the definitions for these parameters, various data files are required, and these are outlined in Table 2. First, we need a comprehensive list of gas stations where Nijhof-Wassink could potentially refuel. For this we collect raw data files r.1, r.2 and r.4. Second, we need a file with all the routes that Nijhof-Wassink's trucks drive, so that we can select the set of routes for the model. For this data file r.6 is collected. Third, we need to determine the decision epochs of the routes and therefore find the gas stations

along the selected routes. Furthermore, the detour of each gas station and the distance between each detour point should be known. For this, we need raw data files r.1, r.2, r.3 and r.6. Fourth, we need data to define the state space, the big penalty, and the constant costs, for this, we need an expert opinion. Last, we need forecasted sample paths of fuel prices for training and testing the model.

For the fuel prices, we know that the prices at gas stations vary per gas station, per day and sometimes even during the day and are thus not constant. When a driver needs to decide whether to refuel, the real-time prices of the gas stations along the route are not known, and this means the price at a gas station at a certain time is uncertain. We can train the model by generating sample paths of fuel prices for random days. To construct price paths we need the price of each gas station on route r on random days. We build a predictive model to be trained for generating sample paths. When using these sample paths, the RL agent is exposed to a range of outcomes that can result from the same actions in similar states, reflecting the probabilistic nature of the environment. This can help the RL agent learn how to make decisions in the presence of uncertainty. We use the historic fuel transactions (r.5) to train a regression model.

The raw data files are presented in Table 2. In total six documents were acquired from the procurement department, ORD and Qlik. ORD is a platform that stores trip data and provides insights by presenting the data and Qlik is a data visualization tool that stores the transaction data.

Table 2: Overview raw data

ID	File name (sheet(s) used)	Description	Dimension (r x c)	Retrieved from
r.1	20230310 Stationslijst.xlsx (1)	Gas stations of fuel company 'x'	3710 x 23	Procurement
r.2	Nijhof-Wassink stationslijst.xlsx (3)	Gas stations of NW for fuel company 'y'	24193 x 18	Procurement
r.3	platss prijzen 2022 Homebase.xlsx (1)	The price/L at the home base	438 x 10	Procurement
r.4	20221219_Price_agreements.xlsx (1-4)	The price agreements per fuel company	40 x 5	Procurement
r.5	Fuel transactions export goed.xlsx (1)	Fuel transactions in 2022	110265 x 16	Qlik database
r.6	Ritten_export.xlsx (1)	Trips DBL sector 09-22 to 03-23	53656 x 21	ORD database

6.1.2. Data Preparation

In this section, we use the raw data from Table 2 for further processing. Data preparation is a critical phase in CRISP-DM framework. This phase involves cleaning, transforming, and organizing raw data into a format suitable for analysis, ensuring accuracy and reliability. Our data preparation consists of five steps. First, we create a list of all stations including the price discounts (p.1), second, we create a list of all transactions (p.2), third we scope both lists to reduce the size (p.3 & p.4). Fourth, the transactions are prepared to create historical prices (p.5). Lastly, the trip data is cleaned and routes for the model are selected (p.6 & p.7). An overview of all prepared files is presented in Table 3.

Table 3: Overview prepared data

ID	File name (sheet(s) used)	Description	Dimensions (r x c)	Retrieved from
p.1	AllStations.xlsx (1)	All gas stations including price discount	28841 x 27	r.1, r.2, r.3
p.2	Fuel transactions export goed.xlsx (3)	Fuel transactions DBL, incl homebase prices	109456 x 16	r.3, r.5
p.3	ScopedStationList.xlsx(1)	Scoped list of gas stations	1558 x 27	p.1, p.2
p.4	ScopedTransactions.xlsx(1)	Scoped list of transactions	58719 x 24	p.1, p.2
p.5	Prices4.xlsx(1)	Transactions for input regression model	31521 x 13	p.4
p.6	RittenCSV.csv(1)	Lanes of DBL with frequency 09-22 to 03-23	3317 x 9	r.6
p.7	selected_routes.xlsx(1)	The 21 routes selected for testing the model	22 x 9	p.6

The first step of the data preparation phase involves merging files r.1 and r.2 to create one file with all gas stations. Also, the quality of the data is checked by verifying the GPS locations and checking for duplicates. Next, file r.4 is merged with the new list to determine the price discount for each gas station. This results in file p.1, which contains a comprehensive list of all gas stations including the location and the price discount.

The second step of the data preparation involves filtering the transactions, merging files r.3 and r.5 and cleaning the transactions. File r.3 contains all the transactions in 2022, however, our research focuses on the DBL sector, so the transactions are filtered to only contain transactions for the DBL sector. The next operation involves merging files r.3 and r.5. File r.3 includes the total fuel costs and the price per liter for the refuels, however, at the home base these stats are unknown. We retrieved document r.5 from the procurement department containing the daily liter price at the home base. Using these liter prices we can add the price per liter and total costs for transactions at the home base. Lastly, the transaction data is cleaned by removing outliers based on the price per liter. Which decreased 2% of the data size. The result is data file p.2.

In the third step, we scope the gas stations and the transactions for two purposes. First, to limit the solution space of the model and second, for the regression model we only want to include gas stations with at least one data point. The transactions are then scoped by removing transactions that are not matched with a gas station because they are not useful as input for the regression model. The gas stations are scoped by only including the stations where Nijhof-Wassink refueled in the last year. To find these stations we need to match the transactions of the last year to the gas stations. Figure 7 presents the logic of the matching algorithm. The result shows that 96% of the transactions can be matched to a gas station and that only 13% of the gas stations were visited in the last year. Filtering all gas stations results in file p.3, the scoped stations list. We also scope the transaction list to only include the transactions that match a station. This results in file p.4.

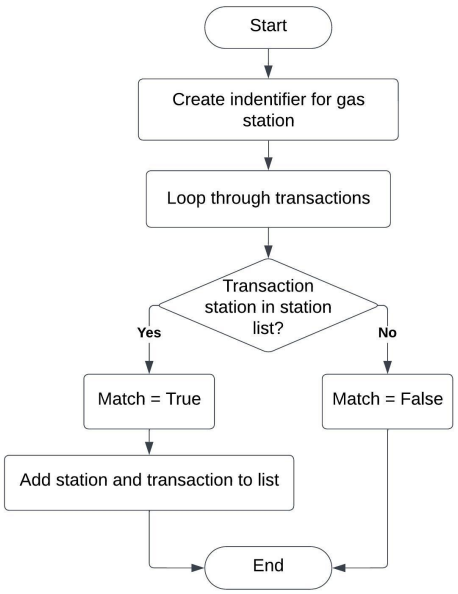

Figure 7: Logic flow matching gas stations to transactions

In the fourth step, we want to prepare a data file with historic prices as input for a regression model. A regression model can be trained to predict the value of unknown data by using other known values. It mathematically models the relationship between the unknown or dependent variable and the known or independent variable. In this case, the dependent variable is the price per liter and the independent variables are the day, month, year, country, and network group. The data needs to be prepared such that the dependent variable and independent variables are stored in a data file. Currently in file p.4, the country and network group is categorical data of string type, however, regression models require numerical data. So, we use one hot encoding for those two variables. One hot encoding is the process of converting categorical variables, such as "country" in our case, into numerical representations like 0 or 1. The next operation is to aggregate the transactions for each gas station per day, by taking the mean of the transactions on that day. Aggregating the prices for each gas station per day reduces the noise in the data and therefore the regression model will perform better. The assumption that we can aggregate the data on a day is based on prices per day per gas station not varying a lot. Therefore, one measure for each gas station per day is sufficient. To verify this, we found that the average standard deviation was €0.003, and for 94% of the prices that were aggregated, the standard deviation was below 1 cent. The aggregation resulted in a 45% reduction of the data size. The resulting data file is p.5.

In the fifth and last step, the trip data is cleaned and analysed, involving removing duplicates, removing trips of which the start or end location is unknown, and determining the routes. We aim to use the trip data to determine the routes that should be tested in the model and to calculate the gas stations along the routes. Therefore, we create a

pivot table including unique trips and their characteristics such as the occurrence of the trips, start country, end country, distance, start GPS and end GPS. The result is file p.6, containing all lanes that can be used for training and testing the model. A selection of routes is made because training and testing all 3316 lanes in this file is too time-intensive. The selected set of routes should be diverse so that we can select robust parameters and test the performance of the model for different types of routes. To provide insights into the refueling problem, the set of routes should be a good representation of all the trips driven. Taking into account these two factors we select a set of routes based on length, the number of gas stations along the route, origin, and destination. Also, a high frequency is needed to have sufficient historical data for testing. The result is a set of 21 routes with diverse characteristics that are presented in Appendix B and visualized in Figure 8. Note that some routes overlap and are therefore not completely visible. These 21 routes cover 30% of the total trips driven and 28% of the total kilometres driven in that time period. The new notation of the set of routes for the case study is: $R = \{1, 2, \ldots, 21\}$. File p.7 contains the selected routes and their characteristics.

Figure 8: Visualisation of 21 selected routes on map

6.1.3. Data Construction

In the previous section, we selected the set of routes and in this section, we need to construct the characteristics of each route including the gas stations along each route, the usage between detour points, the usage and time from a detour point to a gas station, and the time costs of refueling. An overview of data files generated for model input can be found in Table 4.

Table 4: Overview generated data

ID	File name (sheet(s) used)	Description	Dimensions (r x c)	Retrieved from
i.1	Final_routes_input.csv	For each route the gas stations and usage between detour points	22 x 6	p.3, p.7
i.2	Detours.csv	For each gas station on each route the detour usage and time	1063 x 5	i.1, p.3

For each route, the set of decision epochs T, the usage between detour points u_t, the usage of detours d_t, the time cost b_t and the time costs for refueling C need to be generated. This information is retrieved by a Python script of which the logic flow can be found in Figure 9. First, we need to generate the route coordinates by using the TomTom Routing API. The start and end points of the routes are used as input. Next, to find the gas stations along route r we compare the coordinates of the route with the coordinates of the gas stations. Gas stations within a 10 km radius of the route are added to the "along route list". The detour points of a route are found by looking for the coordinates on the route that are closest to the gas stations. Then, with the TomTom Routing API the distance between detour points, the detour distance, and the detour time is calculated. From internal documents, we find the costs per minute and the average fuel consumption of 3.14 km/l. With this information the usage u_t and detour time costs b_t can be calculated. Figure 9 shows file i.1 contains the gas stations along the route usage between detour points and file i.2 contains the detour usage and time for each station on each route. Lastly, based on expert opinions the time costs of refueling are $C = €20$,-.

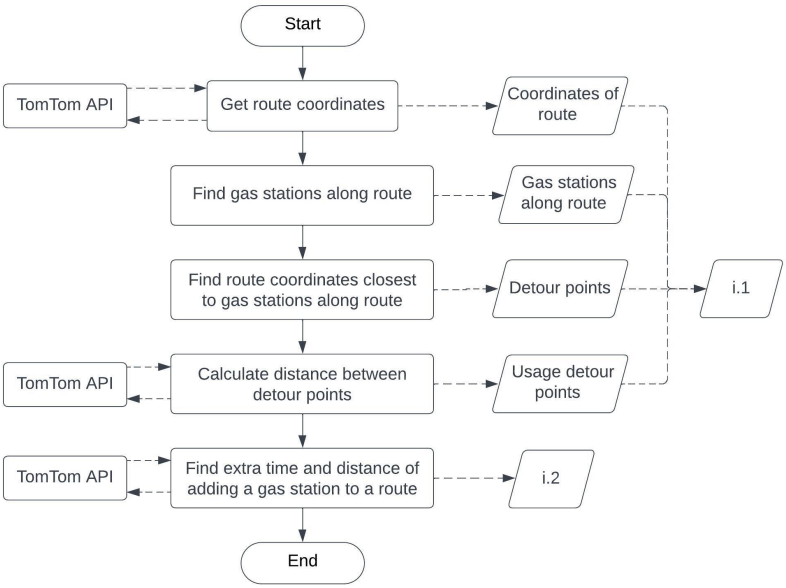

Figure 9: Logic flow route information

6.1.4. State Space and Big Penalty

In previous sections, raw data is prepared and used to generate model input. This section aims to define the following remaining parameters and constants.

- The parameters U, L and δ for the state space: $S = \{L, L + \delta, L + 2\delta, \ldots, U\}$
- The big penalty: $-M$

The state space is defined with an upper bound of 400 liters and a lower bound of 30 liters. The upper bound is the maximum capacity of the fuel tank and the lower bound is

based on the drivers' behavior and the distance between gas stations. Figure 4 in Section 2.2, shows drivers' discomfort with excessively low fuel levels. A driver will not adhere to the advice when he feels uncomfortable with driving with low fuel levels. Besides, we install a buffer for when a route is longer than expected or the gas station for refueling is closed. In some countries, the distance between fuel stations can be 60 km, resulting in additional fuel consumption of 20 liters. In this situation, the buffer of 30 liters is enough to drive to the next gas station. The fuel level is discretized per liter, yielding the following state space: $S = \{30, 31, 32, \ldots, 400\}$. Within the model, a significant penalty $M = -9999$ is assigned for running out of fuel. It is approximately ten times more negative than the reward for refueling, so this strategic penalty ensures that running out of fuel is never the optimal choice.

6.1.5. Take Away

This section answered the research question *"What data is needed as input for the case study and how is the data collected, prepared and constructed?"*. The data collection, preparation, and construction process for the case study involved a systematic approach, following the CRISP-DM framework. Various raw data sources, including gas station details, fuel transactions, and trip data, were acquired and prepared to create comprehensive datasets suitable for modeling. The data preparation included merging, cleaning, and scoping operations to refine the datasets. To construct model input, gas stations along routes, usage between detour points, detour usage, and time costs were determined. The state space and big penalty parameters were also defined. The resulting datasets provide a solid foundation, enabling the development and testing of the RL algorithm and modeling the stochastic fuel prices.

6.2. Modeling of Fuel Prices

The novel component of our sequential decision-making framework is introducing the stochastic nature of fuel prices at the pump by a predictive model for these prices. This section aims to answer the research question *"Which machine learning regression model do we select to forecast the uncertainty in the fuel prices?"*. This section starts by presenting and selecting different regression models (Section 6.2.1). Next, we generate the test design and assess the performance of different regression models (Section 6.2.2).

6.2.1. Presenting Different Regression Models

Different regression models are suitable to use, so we try a variety of regression models from the SKLearn package that uses machine learning namely: Random Forest, Decision Tree Regressor, Gradient Boosting Regressor, and the K-Nearest Neighbour Regressor. The goal is to learn a model F to predict values of the form $\hat{y} = F(x)$ where x represents the independent variable(s) and $\hat{y}$ is a prediction for the dependent variable. In this case, the dependent variable is the price per liter and the independent variables are the day, month, year, country, and network group. Equation (7) shows the prediction function for the fuel prices and Table 5 the range of values each independent variable can take.

$$\hat{p_t} = F(\text{day, month, year, country, network-group}) \tag{7}$$

Table 5: List of values the independent variables can adopt

Independent variable	**Set of possible values**
Day	{1, ..., 30}
Month	{1, ..., 12}
Year	{2022, 2023}
Country	{Netherlands, Belgium, Germany}
Network-Group	{NT1, NT2, NT3, NT4, NF1, Super Economy, Economy, Coverage, Motorway, Non-core, BT1, BT2, BT3, BT4, BT5, BF1}

6.2.2. Results and Validation of Regression Models

To train the regression model and test it afterwards, data file p.5 is split into a train and test set (80%-20%). The performance measures used to evaluate the regression models are R-squared and the Root Mean Squared Error (RMSE). The result of testing the different regression models after training is showcased in Table 6. It shows that the Gradient Booster Regressor provides the best test results for our data. The formulation of a Gradient Boosting Regressor involves combining multiple weak learners (usually decision trees) to create a strong predictive model. The general idea is to sequentially fit new models to the residuals of the previous ones, with each new model focusing on the errors made by the ensemble so far (Friedman, 2001). The model is trained by minimizing the Mean Squared Error (MSE) of the predicted value $\hat{y}$ and the observed value y.

Table 6: Performance of regression models on the test set

Regression model	Parameter	R-squared	RMSE
Random Forest	100	0.82	0.059
Decision Tree	-	0.81	0.060
Gradient Booster	100	**0.83**	**0.057**
K-nearest neighbours	5	0.80	0.062

For each route, we generate a data set of N price paths by drawing a random date for each episode n and then predict the price path consisting of prices for the gas stations along that route. This results in a different sequence of prices for each episode. For each route, we generate separate files with sample paths for training the model and for testing the model. The RL model can learn the patterns in the price data and find a solution suitable for price levels that change over time.

6.2.3. Take Away

In this modeling phase, the focus was on creating a model for forecasting fuel prices at the pump, introducing the stochastic nature of the fuel prices within the sequential decision-making framework. This is done by answering the following research question, *"Which machine learning regression model do we select to forecast the uncertainty in the fuel prices?"*. This section considered various machine learning regression models, including Random Forest, Decision Tree Regressor, Gradient Boosting Regressor, and K-Nearest Neighbour Regressor. The models were trained and evaluated using key performance measures such as R-squared and Root Mean Squared Error (RMSE). The Gradient Boosting Regressor emerged as the most suitable model for our dataset, exhibiting the highest R-squared and the lowest RMSE. The trained regression model enables the generation of diverse price paths for each route.

6.3. RL Parameter Definition

In RL algorithms, various parameters play crucial roles in shaping the algorithm's behavior, influencing the performance and the quality of the results. So, the success of an RL algorithm hinges significantly on the careful tuning of its parameters (Powell and Ryzhov, 2012). Therefore, this section answers the research question *"How are the parameters of the RL algorithm tuned to provide near-optimal results?"*. The key parameters in our two RL algorithms include: the number of episodes (N), the discount factor (γ), the learning rate (α), and the epsilon-greedy (e). First, the experimental set-up is presented in Section 6.3.1 and afterwards, the results are shared in Section 6.3.2.

6.3.1. Experimental Set-up

As mentioned in Section 5.1, we adopt the generalized harmonic step size for the learning rate and the epsilon greedy. We aim to find the parameter setting that works for various routes with different characteristics so that unseen routes, that are not in set $R = \{1, 2, \ldots, 21\}$, can be trained with those settings as well. Two commonly employed methodologies for parameter tuning in RL research are acknowledged (Dabney, 2014). The first approach involves manual testing of each algorithm with a relatively small collection of parameter values, reporting the best results found. The second methodology performs a large parameter optimization procedure and is more complete but more computationally- and time-intensive. However, given time constraints, we opt for the first approach. The number of episodes is found by manually experimenting and the discount factor is based on expert opinions and reasoning to see how future rewards relate to current rewards. The learning rate and the epsilon-greedy are found by designing experiments with different combinations of parameter settings. For the experimental design, we employ Latin hypercube sampling (LHS), a statistical method that enables the generation of a near-random sample of parameter values from a multidimensional distribution (Pilz et al., 2023). LHS is advantageous as it reduces the number of experiments compared to a full-factorial design that tests all possible combinations. For the experimental design of the LHS, we find upper and lower bounds for e and Δ by manually testing different parameter settings for various routes. The range for e is between 0.01 and 9.55 and for Δ between 14500 and 95500. Lowering the lower bound caused worse results and increasing the upper bound did not improve or worsen the result. Note that these settings are tailored for the total number of iterations and increasing or decreasing the number of iterations results in different bounds. In total, we generated 30 experiments using the LHS method Appendix C.

6.3.2. Tuning of Parameters

First, we discuss the discount factor and the number of episodes and afterwards, the learning rate and the epsilon-greedy. For the number of episodes, it is important to determine at which episode the algorithm converges to a good solution and for the discount factor, we need to define the importance of future rewards relative to the immediate rewards (Powell, 2011; Sutton and Barto, 2018). The number of episodes is set on N = 200,000 and the discount factor is set to $\gamma = 1$, meaning future rewards are equally important as current rewards. This makes sense because the time span of a route covers several hours and thus the costs of refueling at a later stage are equally important as refueling earlier.

For the results of the LHS experiments each route within the set R undergoes training for 200,000 episodes, with for every episode a new forecasted price path. Each route is trained for all the parameter settings determined in the 30 experiments. We aim to find a robust general parameter setting that works for all routes in set R. The result of a policy is determined by exposing the trained policy to 20 new forecasted price paths and taking the average of the reward over these 20 price paths. The performance of a parameter setting is evaluated by summing the rewards over all routes for that parameter setting. Choosing a general parameter setting may result in performance loss compared to selecting flexible parameter settings per route. The result of the flexible parameter setting is calculated by determining the best parameter setting for each route and then summing these rewards.

Table 7 shows the result for the flexible parameter settings, for both Q-Learning and SARSA. The table shows for each route, the number of decision epochs, the best parameter settings and the corresponding result. The total result is the sum of the reward of all routes and is 180163.2 for SARSA and 180194.4 for Q-Learning. The graph on the left in Figure 10, shows that routes with more decision epochs require higher values for the learning rate. A higher value for the learning rate results in a slower decreasing learning rate meaning longer routes need faster learning in the beginning to ensure a good result. Furthermore, the graph on the right in Figure 10 shows that the epsilon-greedy needs to be as small as possible meaning fast exploitation is important for good results.

Table 7: Result of SARSA and Q-Learning for flexible parameter settings

		SARSA			**Q-Learning**		
Route	#epochs	Result	Δ	e	Result	Δ	e
328	60	7101.2	41500	0.05	7101.2	41500	0.05
353	20	3414.7	23500	0.10	3414.7	23500	0.10
510	30	6171.3	23500	0.10	6171.1	68500	4.15
678	20	2715.9	23500	0.10	2715.9	23500	0.10
804	45	2121.0	23500	0.10	2121.0	41500	0.05
1236	133	34089.3	95500	0.51	34138.8	95500	0.51
1237	47	6691.7	32500	1.45	6692.8	23500	5.05
1238	34	3391.4	59500	0.42	3401.5	59500	0.08
1556	47	9884.6	23500	0.10	9884.6	23500	0.10
1675	113	20952.5	95500	0.09	20937.6	95500	6.85
1888	17	1866.0	23500	0.10	1866.0	23500	0.10
1915	113	16228.0	77500	0.03	16221.4	95500	0.09
1997	32	5837.1	23500	0.10	5837.1	23500	0.10
2621	20	1166.9	68500	0.07	1166.9	23500	0.10
2622	75	14291.5	77500	0.03	14293.1	50500	9.55
2642	101	18959.9	95500	0.09	18950.7	95500	0.51
2820	42	4846.4	23500	0.10	4846.4	23500	0.10
2910	53	9245.5	23500	0.10	9245.5	23500	0.10
2941	25	6385.3	23500	0.10	6385.3	23500	0.10
3170	20	2187.0	23500	0.10	2187.0	23500	0.10
3208	36	2615.9	23500	0.10	2615.9	23500	0.10
	Total result	**180163.2**			**180194.4**		

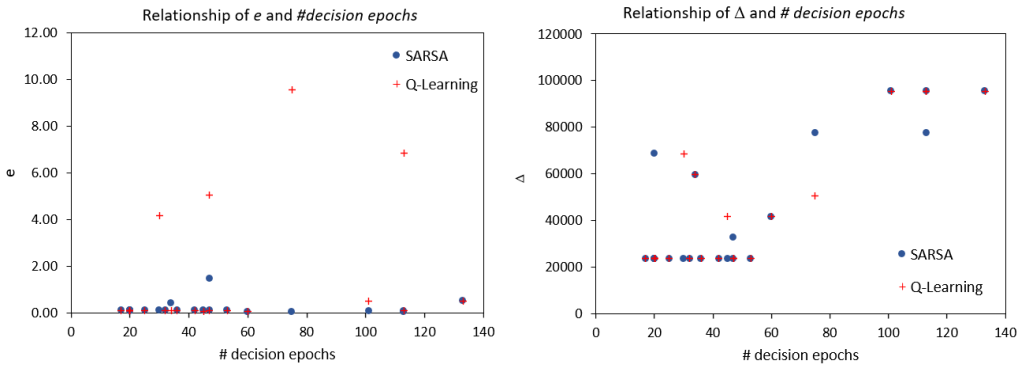

Figure 10: Relationship of Δ and e and number of decision epochs T for SARSA and Q-Learning

As mentioned before, we aim to find one general parameter setting that performs well on all routes so that the algorithm is robust and can be used on unseen routes as well. For each parameter setting, we trained all routes independently and afterwards retrieved the total result for that parameter setting by summing the reward over all routes. The best 15 total results for SARSA and Q-Learning can be found in Table 8 and 9, respectively.

For SARSA the comprehensive results are available in Appendix D.1 and for Q-Learning the complete results can be referenced in Appendix D.2.

Table 8: 15 best experiments of general parameter settings for the SARSA algorithm

Experiment	Δ	e	**Total result**	**Gap**
8	95500	0.09	180322.6	0.09%
2	86500	0.04	180565.3	0.22%
6	77500	0.03	180732.0	0.32%
1	68500	0.07	180733.7	0.32%
11	77500	0.15	180922.5	0.42%
5	59500	0.08	181259.7	0.61%
4	50500	0.01	181775.1	0.89%
3	41500	0.05	181973.8	1.00%
9	32500	0.06	182909.7	1.52%
13	50500	0.24	182929.5	1.54%
12	59500	0.42	183405.8	1.80%
0	23500	0.10	185138.7	2.76%
7	14500	0.02	185701.1	3.07%
19	86500	0.33	189979.6	5.45%
18	41500	0.69	192005.7	6.57%

Table 9: 15 best experiments of general parameter settings for the Q-Learning algorithm

Experiment	Δ	e	**Total result**	**Gap**
17	95500	0.51	180305.9	0.06%
25	95500	6.85	180409.0	0.12%
8	95500	0.09	180409.2	0.12%
22	86500	2.35	180468.5	0.15%
2	86500	0.04	180507.4	0.17%
19	86500	0.33	180534.9	0.19%
6	77500	0.03	180542.6	0.19%
11	77500	0.15	180595.0	0.22%
28	77500	3.25	180607.9	0.23%
14	68500	0.60	180769.4	0.32%
1	68500	0.07	180815.3	0.34%
23	59500	8.65	180921.7	0.40%
26	68500	4.15	180975.7	0.43%
12	59500	0.42	180999.7	0.45%
5	59500	0.08	181166.9	0.54%

To estimate the performance loss due to selecting a general parameter setting, the total result per parameter setting is compared with the total result of the flexible parameters. The gap is calculated by Equation (8) and the result is shown in the last column.

$$\text{Gap} = \frac{\text{Total result general} - \text{Total result flexible}}{\text{Total result flexible}} \times 100\% \tag{8}$$

Analyzing the results for both algorithms, we observe that for Q-Learning the learning rate (Δ) has a more pronounced impact on performance compared to the epsilon-greedy (e) (Figure 11). Interestingly, this stands in contrast to the SARSA-based algorithm's behavior where the epsilon-greedy (e) has more impact on the performance (Figure 12). However, for both algorithms, the trend reveals that higher Δ values correlate with improved outcomes, while lower Δ values result in less favorable performance. The gap between the total result per parameter setting and the total result of the flexible parameters shows that the Q-Learning-based algorithm demonstrates robust performance across various configurations, with the most significant gap being around 3% (Appendix D.2). In contrast, the SARSA-based algorithm appears more sensitive to parameter settings, showcasing a substantial 110% gap with the result of the flexible parameters settings (Appendix D.1).

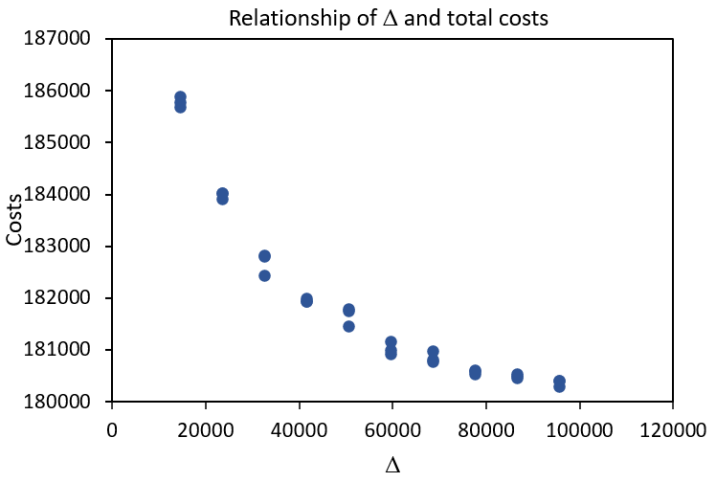

Figure 11: The positive influence of a larger Δ on the total costs for the Q-Learning algorithm

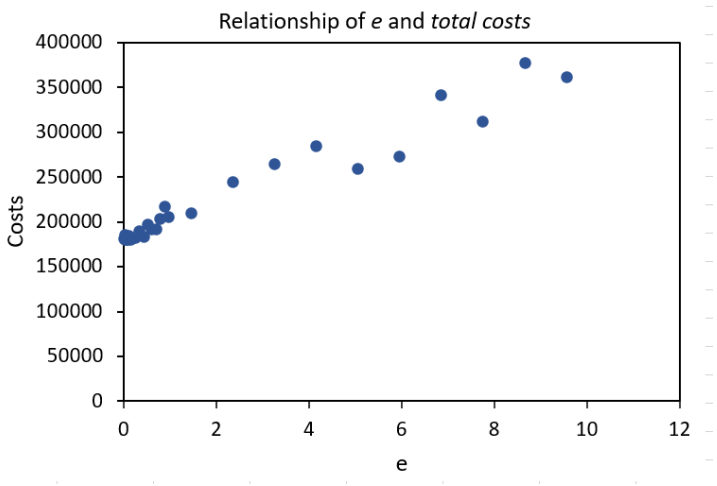

Figure 12: The positive influence of a smaller e on the total costs for the SARSA algorithm

The best results for Q-Learning are achieved with the parameter setting $\Delta = 95500$ and $e = 0.51$ and for the SARSA with the parameter setting $\Delta = 95500$ and $e = 0.09$. To assess the stability and convergence of the algorithm with these parameter settings, we observed various state-action approximations (Q-values) over the episodes. Figure 13 and Figure 14 on the next page show the Q-values for route r over the episodes, where f is the fuel level, t the decision epoch and a the action that is taken. The graphs show fast convergence in the initial phases due to higher learning rates, and towards the end, Q-values stabilize with some observed fluctuations. To assess the acceptability of these fluctuations, we analyze the Q-values over the last 20,000 episodes. The observed fluctuations remain within a 4% bound, and these fluctuations did not trigger a policy change any more. To conclude, we find the best performance is with parameter settings $\Delta = 95500$ and $e = 0.09$ for SARSA, and $\Delta = 95500$ and $e = 0.51$ for Q-Learning.

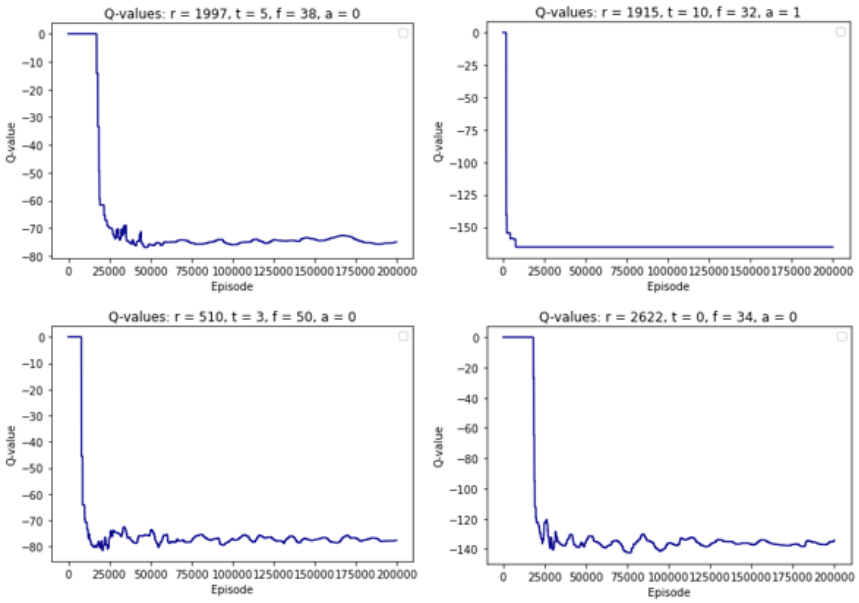

Figure 13: The convergence of the Q-values of state-action pair Q(t, f, a) for Q-Learning

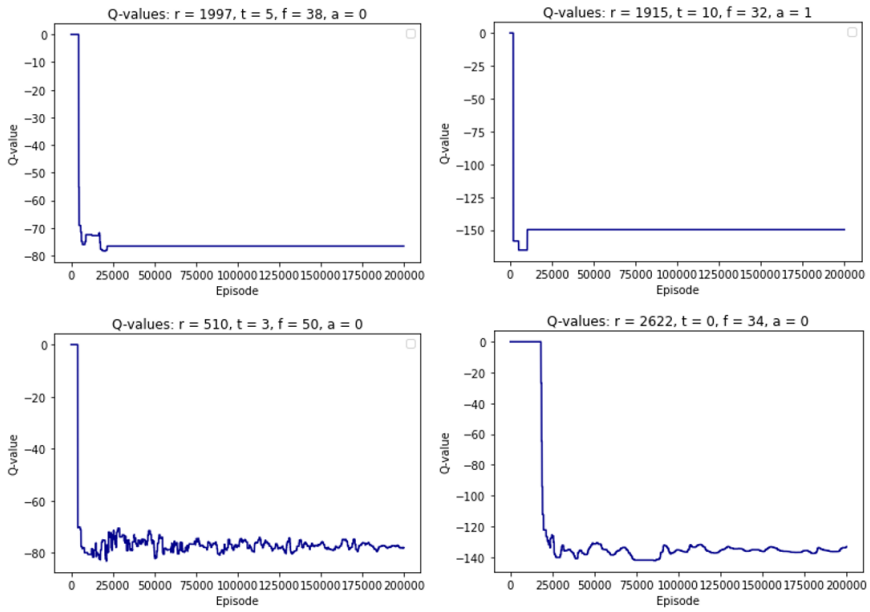

Figure 14: The convergence of the Q-values of state-action pair Q(t, f, a) for SARSA

6.3.3. Take Away

To answer the research question, *"How are the parameters of the RL algorithm tuned to provide near-optimal results?"*, we delved into the nuanced process of tuning the RL parameters. By manually testing we found the optimal number of episodes to be 200,000 and the gamma is set to 1, giving future rewards the same importance as immediate rewards. We determined boundary settings for tuning the learning rate and epsilon greedy and created 30 LHS experiments. These experiments identified general parameter configurations for both SARSA and Q-Learning that work well for all routes. The results showed that SARSA demonstrated sensitivity to the parameters settings, while Q-Learning showcased a good performance over all settings. The optimal general parameter setting for Q-Learning is $\Delta = 95{,}500$ and $e = 0.51$ and for SARSA $\Delta = 95{,}500$ and $e = 0.09$.

6.4. Computational Results

In this section, we address the research question *"What are the results of the different solution methods, and what insights do the results provide regarding the performance of the RL algorithm?"*. First, we test both SARSA and Q-Learning to determine the better-performing RL algorithm in Section 6.4.1. Second, we aim to validate the generalizability of our algorithm and the chosen parameters. Hence, we train and test a set of new unseen routes in Section 6.4.2. Thereafter, we analyse the performance by comparing the RL approach to a benchmark heuristic in Section 6.4.3, the optimal solution in Section 6.4.4, and the historical refuel decisions in Section 6.4.5. The methods are compared by using 20 sample price paths. For the RL model, the policies gained from training the model are tested by applying them to these 20 new sample paths. The result of a route is the average reward of these 20 sample paths. At the end, the results are summarized in Section 6.4.6

6.4.1. Q-Learning VS. SARSA

In this section, we evaluate the suitability of two RL algorithms, Q-learning and SARSA, for addressing the FRVRP. The results in Table 10 show the average reward over the 20 price paths for both SARSA and Q-Learning. The last column presents the performance gap as calculated by Equation (9).

$$\text{Gap} = \frac{\text{Result Q-Learning} - \text{Result SARSA}}{\text{Result SARSA}} \times 100\% \tag{9}$$

The total result is the sum of the result over all routes and the average performance gap. The average performance gap indicates a modest performance distinction, with Q-Learning demonstrating a slightly favorable outcome for the FRVRP. Given the observed sensitivity of the SARSA algorithm to the parameter settings and our objective to extend the model to additional routes, we conclude that Q-Learning is more suitable for our FRVRP.

Table 10: Q-Learning algorithm vs. SARSA algorithm

*Q = Q-Learning, S = SARSA, n.a. = not applicable

Route	Result SARSA	Result Q-Learning	Gap	Best*
2642	18959.9	18950.7	-0.05%	Q
3208	2615.9	2615.9	0.00%	n.a.
804	2121.9	2121.9	0.00%	n.a.
1997	5837.1	5837.1	0.00%	n.a.
2820	4846.4	4846.4	0.00%	n.a.
1236	34198.2	34138.8	-0.17%	Q
510	6171.4	6171.4	0.00%	n.a.
328	7101.2	7101.2	0.00%	n.a.
353	3414.7	3414.7	0.00%	n.a.
1915	16241.7	16262.3	0.13%	S
2621	1166.9	1166.9	0.00%	n.a.
2941	6385.3	6385.3	0.00%	n.a.
1675	20952.5	20983.2	0.15%	S
2910	9245.5	9245.5	0.00%	n.a.
1237	6700.8	6699.2	-0.02%	Q
1556	9884.6	9884.6	0.00%	n.a.
678	2715.9	2715.9	0.00%	n.a.
1888	1866.0	1866.0	0.00%	n.a.
2622	14308.0	14310.4	0.02%	S
3170	2187.0	2187.0	0.00%	n.a.
1238	3401.5	3401.5	0.00%	n.a.
Total result	**180322.6**	**180305.9**	**-0.01%**	**Q**

6.4.2. Parameter Validation and Generalization

To prove the generalizability of the algorithm it is critical to validate the algorithm's effectiveness on routes beyond the training set. For the training set the parameter configuration, specifically with $\Delta = 95500$ and $e = 0.51$, has demonstrated optimal performance on our training set of routes. To verify this generalizability, we selected a test set of 10 representative routes that cover various characteristics. The details of these routes, along with their characteristics, are outlined in Appendix E. These routes are trained by executing 200,000 episodes with in each episode a new forecasted price path. The result is the average reward of exposing the trained policies to 20 new forecasted price paths. Our objective is to assess the suitability of the parameters by examining the gap with the upper bound we derived from the optimal solution method (Equation (10)).

$$\text{Optimality gap} = \frac{\text{Result Q-Learning} - \text{Result optimal}}{\text{Result optimal}} \times 100\% \tag{10}$$

If the observed gap closely aligns with the optimality gap identified in the training set, it suggests that the performance on routes outside the training set remains consistent with those parameters. The upper bound of the test set is found by applying the optimal method demonstrated in Section 5.2. Table 11 reveals a 0.68% gap with the optimal solution, confirming that these parameter settings remain effective when training on

routes beyond the initial set. We do see that longer routes such as route 538 with 131 decision epochs perform worse than shorter routes, which can be explained by the increase in dimension of the problem. Consequently, we can conclude that the identified parameters are well-suited for training on a diverse range of routes and that the Q-Learning algorithm can be generalized to other problem instances.

Table 11: Results of parameter testing

Route	**# decision epochs**	**Result optimal**	**Result Q-Learning**	**Gap**
538	131	22941.8	23501.0	2.44%
1055	73	13095.5	13161.4	0.50%
2959	71	15329.8	15384.3	0.36%
2148	66	46871.7	47580.5	1.51%
1511	65	53362.7	53841.5	0.90%
287	55	24444.6	24462.5	0.07%
88	43	8329.9	8413.2	1.00%
3204	42	3275.9	3275.9	0.00%
1457	40	9292.7	9292.7	0.00%
1510	32	4243.2	4243.2	0.00%
Total result		**201187.8**	**203156.1**	**0.68%**

6.4.3. Benchmark Heuristic VS. Reinforcement Learning

To validate the performance of RL, the performance of the Q-Learning algorithm is compared against the benchmark heuristic introduced in Section 5.2. We recall that the basic idea is that the truck passes every gas station and at every gas station the algorithm evaluates if the fuel level is below the lower bound because then the driver needs to refuel at the current gas station. Otherwise, if the fuel level drops below the lower bound within the next Z gas stations, the algorithm will choose the cheapest of those gas stations to refuel. As determined in Section 6.1.4, the lower- and upper bounds of the fuel levels are 30 and 400. We set the parameter Z, which determines the number of future stations that are taken into account, at 5 because at the moment a driver decides refueling is needed we assume he thinks 5 gas stations ahead. The biggest difference between real life and the heuristic is that for the heuristic the fuel prices of these next five gas stations are given. The heuristic is thus deterministic in comparison to our RL model which takes into account that the prices of the next gas stations are unknown.

Table 12 shows the results of Q-Learning and the heuristic and presents the total result at the bottom of the table including, the sum of the costs over all routes and the average gap. The performance gap in the last column of the table is calculated by Equation (11).

$$\text{Gap} = \frac{\text{Result Q-Learning} - \text{Result heuristic}}{\text{Result heuristic}} \times 100\% \tag{11}$$

Table 12: Q-Learning algorithm versus Benchmark Heuristic algorithm

Route	Result Q-Learning	Result Heuristic	Gap	Best
2642	18950.7	19409.9	-2.37%	Q
3208	2615.9	2688.5	-2.70%	Q
804	2121.9	2208.0	-3.90%	Q
1997	5837.1	5814.3	0.39%	H
2820	4846.4	5110.9	-5.18%	Q
1236	34138.8	34640.6	-1.45%	Q
510	6171.4	6354.9	-2.89%	Q
328	7101.2	7486.8	-5.15%	Q
353	3414.7	3504.2	-2.55%	Q
1915	16262.3	17451.8	-6.84%	Q
2621	1166.9	1205.4	-3.19%	Q
2941	6385.3	6539.9	-2.37%	Q
1675	20983.2	22506.3	-6.79%	Q
2910	9245.5	9955.5	-7.13%	Q
1237	6699.2	7002.3	-4.33%	Q
1556	9884.6	10589.9	-6.67%	Q
678	2715.9	2774.8	-2.12%	Q
1888	1866.0	1913.2	-2.47%	Q
2622	14310.4	14545.0	-1.61%	Q
3170	2187.0	2248.8	-2.75%	Q
1238	3401.5	3577.0	-4.91%	Q
Total result	**180305.9**	**187528.1**	**-3.67%**	**RL**

The obtained results reveal that the RL algorithm consistently outperforms the heuristic method across a range of routes with 3.67% lower costs. Specifically, the RL model achieved a total average cost per route of 180305.9, while the heuristic resulted in a total cost of 187528.1. This result signifies an overall cost improvement when adopting RL for refueling decisions. In real-life the performance of RL would be even better as the heuristic policy does not consider the price stochasticity, possibly resulting in too optimistic or pessimistic policies. These findings underscore the effectiveness of RL in learning optimal strategies for refueling, showcasing its adaptability to dynamic conditions and superior route optimization capabilities compared to a heuristic method. When diving into the results and gaining a deeper understanding as to why the RL model performs better we find that the heuristic especially performs worse over routes where you have few gas stations with significantly lower refuel costs. If the prices are low at a few points of the route the heuristic can not always select those gas stations because it can choose from the 5 gas stations before the fuel level dips below 30. To demonstrate this we selected route 328, with fuel level 67 at $t = 0$. When the driver starts with fuel level 67 on this route, the fuel level reaches below the lower bound at the 51^{st} decision epoch. Figure 15 shows the refueling costs, calculated with a randomly selected price path, for each decision epoch. We see that the refueling cost varies between €142 and €79 highlighting the importance of a good refueling decision. Both algorithms should refuel before the dotted line at the 51^{st} station. The RL algorithm selects the station at decision epoch 17 with costs of €79.87, while the heuristics can only select a gas

station in decision epochs 46 until 50. The best option for the heuristic is station 46 with corresponding costs of €94.60. In this example, the RL outperforms the heuristic by 18% explaining the overall gap in performance between RL and the heuristic.

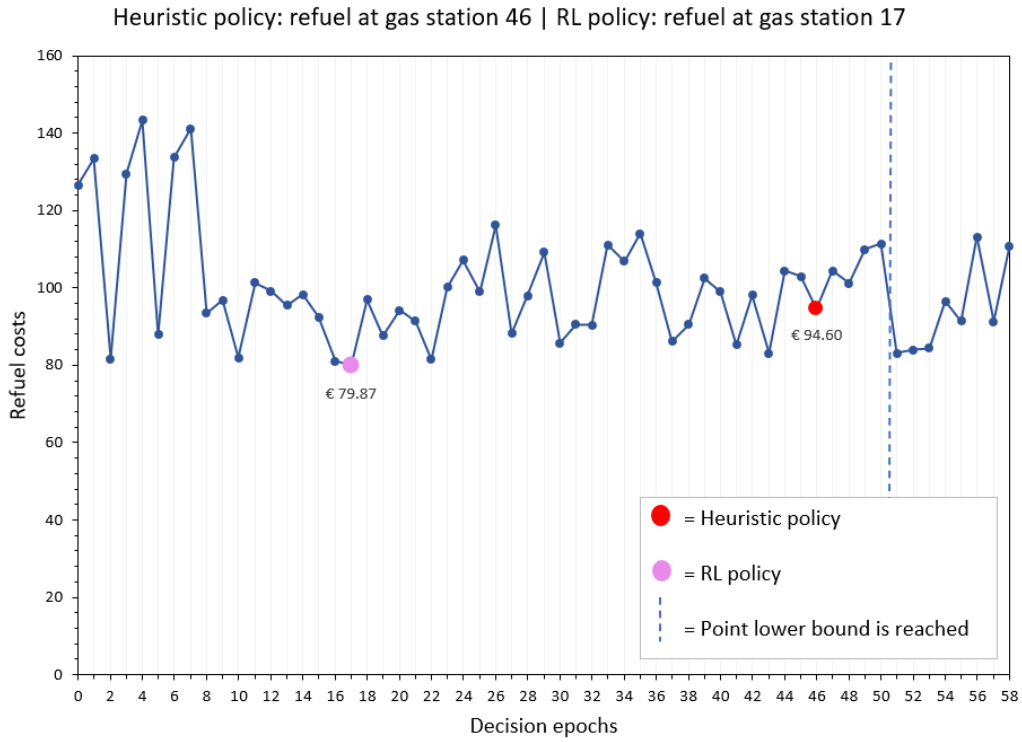

Figure 15: Refueling policies RL and heuristic for route 328 and start fuel level 67

6.4.4. Optimal Solution VS. Reinforcement Learning

This section is dedicated to demonstrating the RL algorithm's capability to yield near-optimal solutions for the FRVRP. This section aims to find an upper bound for the stochastic FRVRP, by lifting the non-anticipativity and thus allowing decisions to depend on future information. The model is transformed into a deterministic model and can be solved quickly to provide optimal solutions. The solution is used to validate the high performance of the RL algorithm over various routes and fuel levels, by calculating the optimality gap for each route (Equation (10)). Table 13 presents the results, including the total results at the bottom, representing the sum of the results over all routes and the average optimality gap.

Table 13: Q-learning-based FRVRP algorithm versus Optimal Solution

Route	# decision epochs	Result Q-Learning	Result Optimal	Gap
1236	133	34138.8	33495.0	1.93%
1675	113	20983.2	20750.7	1.13%
1915	113	16262.3	16138.4	0.77%
2642	101	18950.7	18820.0	0.70%
2622	75	14310.4	14233.2	0.54%
1237	47	6699.2	6665.1	0.51%
1238	34	3401.5	3391.4	0.29%
328	60	7101.2	7097.9	0.05%
510	30	6171.4	6168.8	0.04%
804	45	2121.9	2121.0	0.04%
1997	32	5837.1	5835.5	0.03%
353	20	3414.7	3414.7	0.00%
678	20	2715.9	2715.9	0.00%
1556	47	9884.6	9884.6	0.00%
1888	17	1866.0	1866.0	0.00%
2621	20	1166.9	1166.9	0.00%
2820	42	4846.4	4846.4	0.00%
2910	53	9245.5	9245.5	0.00%
2941	25	6385.3	6385.3	0.00%
3170	20	2187.0	2187.0	0.00%
3208	36	2615.9	2615.9	0.00%
Total result		**180305.9**	**179045.2**	**0.29%**

Observing the results, it becomes evident that for around 50% of the routes, the RL algorithm consistently provides the optimal solution. This implies that, across all tested price sequences for those routes, the RL algorithm's advised refueling strategy aligns with the optimal solution. It can be inferred that, for these specific routes, fluctuating prices do not significantly alter the optimal refueling decision. For routes where optimality is not achieved, the gap with the optimal solution remains relatively small, ranging from 0.03% to 1.93%. The overall optimality gap is calculated to be 0.29%. Notably, Figure 16 shows the RL algorithm's performance experiences a decline with longer routes, where the dimension of the problem increases. As an example, the longest route is 1236, with a length of 133, and this route exhibits the largest gap to the optimal solution. In summary, the RL algorithm consistently provides solutions close to optimal, with an overall optimality gap of 0.29%.

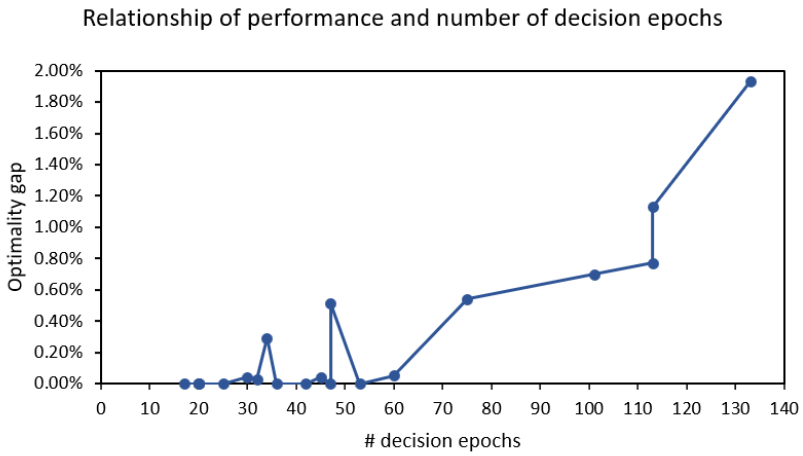

Figure 16: The decreasing performance of the RL algorithm when the problem size grows

6.4.5. Historical Refuel Decisions VS. Reinforcement Learning

This section compares the refueling decisions of the driver with the refueling advice of the model and shows potential savings in the costs. The hypothesis is that the model's advice is better and thus the refuel costs decrease compared to the driver's decision. This part of the results is crucial to show the practical contribution and indicate potential savings by implementing the RL algorithm as a decision-making tool. The refuels done by drivers on the routes in set R are retrieved from the historical data. This results in 166 refuels that can be compared with the advice of the model. In the next paragraph, we elaborate on the results and show the highlights.

As for the characteristics of the drivers' refueling behaviour, we see that the home base is the most popular location, covering 95% of all refuels used for comparison. The fuel station is considered the cheapest by the company and therefore most drivers fill up their tank after arriving at the home base so that they can depart with a full tank the next day. However, because of the finite nature of the model, it does not take into account that the truck will drive another route the next day. Therefore, when the home base is located at the end of a route the model will never advise to refuel there. The model will only refuel when the end can not be reached with at least 30 liters still in the tank. In practice, this finite nature does not matter because the next morning when the driver departs from the home base the model will advise to refuel there if it is the cheapest gas station. For the comparison, we do not include the refuels at the home base as the behavior of refueling at the end of the route can not be compared with the behavior of the model. So, we look at the other 5% of refuels that are not done at the home base. Table 14 shows the results by presenting the average costs over 20 random price sequences and the average performance gap for each refueling decision. At the bottom of the table, the summed costs over all refuelings and the total average gap are presented. From these results, we conclude that the model's advice outperforms the driver's decision by on average 10% per route and a decrease in total costs by 11%. Table 15 shows more details on prices and discounts at the gas stations. We see that in 10/10 routes the model chooses a gas

station with cheaper or equal net prices. Interestingly, all advices from the model also have an equal or higher discount. Assuming the predictions of the prices are reliable, we can say that the highest price reductions are most favorable when refueling. Note that these refuels only cover 25% of the trained routes and a few fuel levels due to the limited data availability. So, even though the results show an 11% decrease in costs, this is merely an indication because adding extra historical refueling decisions can change the total result and conclusions significantly.

Table 14: Results comparison of driver decisions (d) vs. model advice (m)

ID	Route	Fuel level	Refuel location (d)	Refuel location (m)	Cost (d)	Cost (m)	Gap
9	1236	121.6	essoheerdehetveen	bsposs	218.76	212.39	3%
142	1236	146.1	essoapeldoorndebrink	g&varendonk	212.35	187.16	13%
279	1238	228.8	stdcoevorden	shellokken	65.03	61.56	6%
280	1238	228.8	stdcoevorden	shellokken	65.03	61.56	6%
727	2622	62.4	truckshellantwerpenhaven200	g&vtruckstationbeverenwaas	145.70	124.44	17%
1260	2642	78.4	stdnijkerk	essobunnikdeforten	182.43	156.45	17%
1516	1236	150.2	truckshellarendonk	g&varendonk	207.86	187.16	11%
1924	510	94.4	truckshellgent	essogentkennedylaan	86.22	75.15	15%
2085	1238	137.6	stdcoevorden	shellokken	65.03	61.56	6%
Total result					**1248.41**	**1127.44**	**10%**

Table 15: Details on comparison of historic decisions and RL

ID	Route	*Price (d)	Discount (d)	Cost (d)	*Price (m)	Discount (m)	Cost (m)
9	1236	1.53	-0.11	218.76	1.47	-0.1625	212.39
142	1236	1.47	-0.1625	212.35	1.26	-0.3151	187.16
279	1238	1.54	-0.145	65.03	1.54	-0.145	61.56
280	1238	1.54	-0.145	65.03	1.54	-0.145	61.56
727	2622	1.47	-0.2051	145.70	1.30	-0.3151	124.44
1260	2642	1.60	-0.145	182.43	1.51	-0.1625	156.45
1516	1236	1.43	-0.2051	207.86	1.26	-0.3151	187.16
1924	510	1.44	-0.2051	86.22	1.29	-0.2801	75.15
2085	1238	1.54	-0.145	65.03	1.54	-0.145	61.56

*The price is the net price including the discount

6.4.6. Take Away

The computational results section aimed to answer the research question *"What are the results of the different solution methods, and what insights do the results provide regarding the performance of the RL algorithm?"*. The section started, by determining the best algorithm (Section 6.4.1). By evaluating the performance of both SARSA and Q-Learning, we observed that the Q-Learning-based algorithm achieved marginally better results, with only a 0.02% performance advantage. Given its consistent and slightly better performance, the Q-Learning-based FRVRP algorithm emerged as the most effective. Subsequently, we validated the RL algorithm with the parameter setting by proving the gap with the lower bound stays within bounds when tested on 10 unseen routes (Section 6.4.2). These results contribute valuable insights into the application of RL techniques for solving the FRVRP. In the next step, we conducted a comparative analysis of the RL algorithm against a benchmark heuristic (Section 6.4.3), against the optimal solution (Section 6.4.4) and the historical refuel decisions (Section 6.4.5). Notably, the RL algorithm demonstrated a 0.3% deviation from the deterministic optimal solution, showcasing its proficiency in providing solutions close to the lower bound. Furthermore,

The RL algorithm outperformed the benchmark heuristic by 3% and showcased an 11% improvement in refueling costs compared to the historical driver decisions. A noteworthy insight from the historical comparison indicated that choosing stations with high discounts leads to the most cost-effective solution.

6.5. Insights

This section aims to answer the research question *"What insights does this model in combination with this solution method provide to the practical and academic communities?"*. The objective is to dive into a comprehensive understanding of the learned policies and decision-making processes of the RL algorithm. The analysis yields insights from the reinforcement learning policies and reveals relationships between various variables. To begin, Section 6.5.1 analyses the predicted fuel prices to evaluate the impact of independent variables on forecasted pump prices. Subsequently, Section 6.5.2 illustrates and visualizes the policies at an operational planning level. Moving forward, Section 6.5.3 delves into insights on the tactical planning level. Finally, Section 6.5.4 presents strategic insights gained from the model's results.

6.5.1. Predicted Fuel Prices

A distinctive and innovative aspect of this research involves capturing the stochastic nature of fuel prices at the pump through an ML regression model. This section aims to shed light on the impact of various independent variables—day, month, year, country, and network group—on the forecasted fuel prices. To conduct this analysis, 200 prices were generated for each network group in each country over a one-year period.

Figure 17 illustrates the results, revealing two key insights. Firstly, the significant gap between the minimum and maximum fuel prices, along with an average standard deviation of 0.11 over all categories, underscores the substantial variability in prices over the course of a year. This variability implies that incorporating these price fluctuations over time can lead to more accurate and reliable policies. Secondly, the figure demonstrates the influence of network groups and countries on pump prices. For instance, the average forecasted price for Germany (DE) is 1.76, the Netherlands (NL) is 1.69, and Belgium (BE) is 1.62. It is important to note that these values represent pump prices, with the final net price being determined by the pump price minus the discount.

The analysis further reveals distinctions among network groups within each country. These network groups belong to two different fuel companies, denoted as company X and company Y. A detailed examination of price levels at the network groups of both companies indicates that company Y consistently offers lower pump prices than company X in both the Netherlands (by 5 cents on average) and Belgium (by 6 cents on average). When looking at the network groups we see that in the Netherlands, group NT1 has the lowest average pump price, in Belgium, BT1, and in Germany, Non-core. The analysis on fuel prices provides insights into the impact of independent variables on fuel prices and underscores the importance of taking into account this variability in the model.

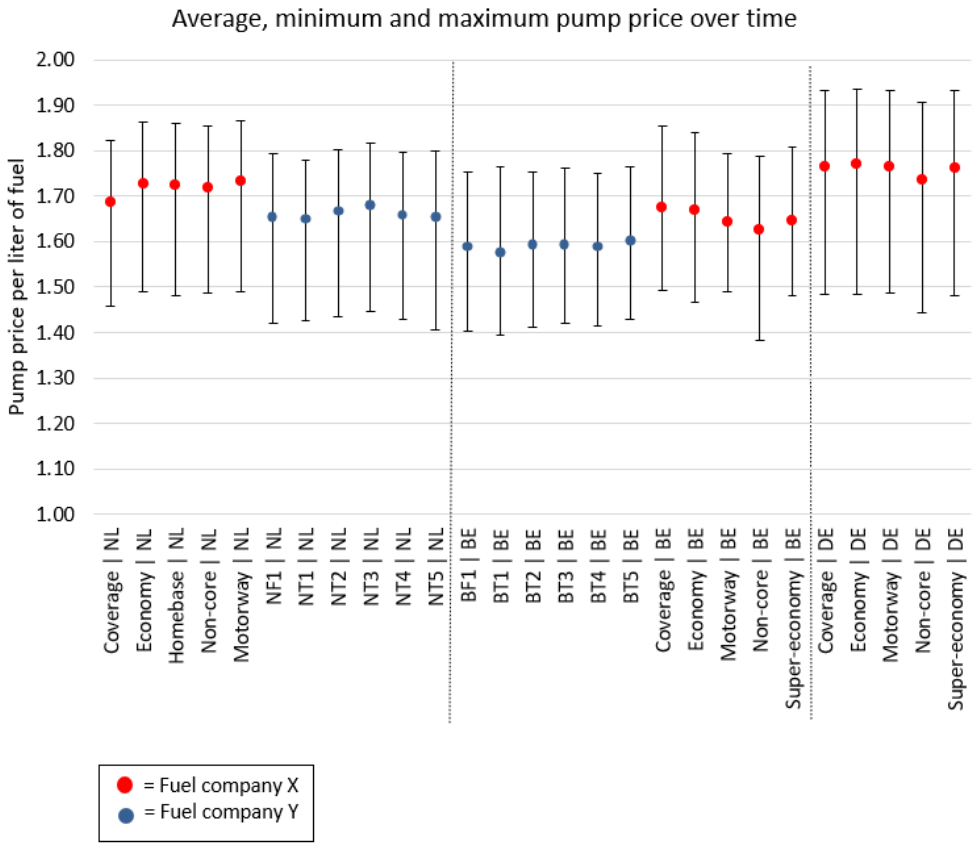

Figure 17: The minimum, maximum and average of 200 forecasted prices over time for the different countries and network groups

6.5.2. Operational Level

This section offers more insights into the policies on an operational level by visualizing and analysing policies from the RL algorithm. A policy defines which action to choose at any given time in a given state by choosing the action with the highest state-action value. Our policy is deterministic, so the agent will always choose the same action for a given state and time. Consider an illustrative example with three gas stations along a route. At the first decision epoch, the driver, with a fuel level of 38, contemplates whether to refuel or not. The lookup table provides the state-action values for both options, and the driver opts not to refuel as it yields the highest value. This decision transitions the driver to the next epoch with a new state, reducing the fuel level to 33. Subsequently, for the state-action pair in the second epoch (fuel level: 33), refueling emerges as the optimal action. The driver refuels, progresses to the next epoch with a new fuel level of 398, and at the final gas station, the best action is to refrain from refueling, logically aligning with the recent refueling. Thus, the policy for the initial state ($s_0 = 38$) is represented as $[0, 1, 0]$. The model encompasses 370 states from which the truck can commence its journey, resulting in 370 unique policies for each route. Resulting in this research generating

a total of $21 \times 370 = 7770$ policies. The subsequent exploration focuses on a detailed analysis of route 2642, which spans 260 kilometers with 110 gas stations along the route.

Figure 18 displays route 2642 with circle markers representing gas stations along the route. The average of 20 forecasted sample paths of prices is used to show the average refueling costs per station. Green markers indicate lower costs, while red markers represent higher costs. Along this route, the total refueling costs vary from €149 to €200, demonstrating a 30% difference. This emphasizes the significance of selecting an optimal refueling point. Additionally, the figure illustrates a trend where cheaper gas stations tend to be closer to the route, while more expensive ones are situated farther away. This observation suggests that increasing detours for potentially lower prices may not be profitable, as discussed further in Section 6.5.3.

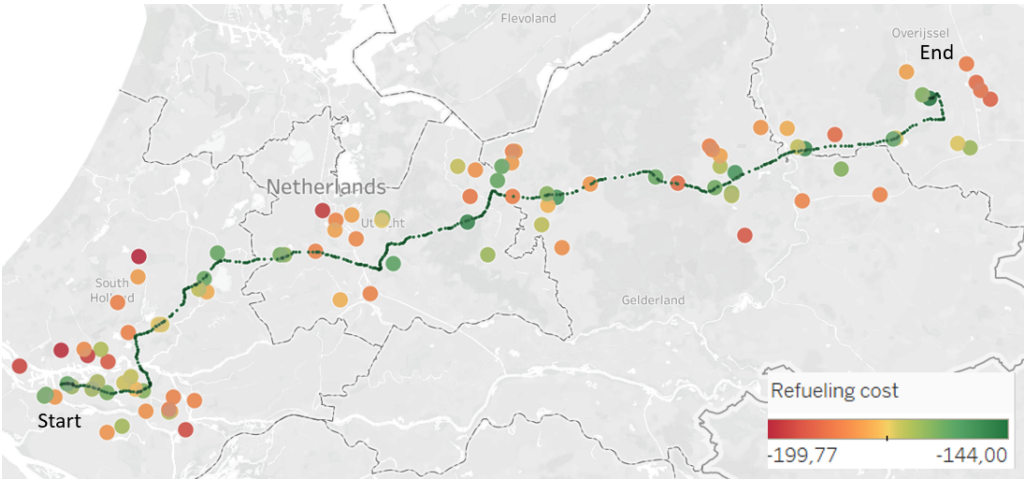

Figure 18: Gas stations along route 2642 with color indication for refueling cost

Moving on to the policies, Figure 19 showcases route 2642 with recommended gas stations. The marker size indicates the frequency of gas station recommendations. It is important to note that this representation is based on the average of forecasted price paths, and the total refueling costs in other price paths may differ. The figure reveals that the gas station with the lowest costs (€149) is the most frequently chosen. However, as the fuel level can drop below the lower bound before reaching this gas station, more expensive alternatives are selected earlier. Furthermore, after the cheapest gas station more pricier gas stations are selected as well. This is reasonable because the difference in total refueling costs can change for different forecasted price paths.

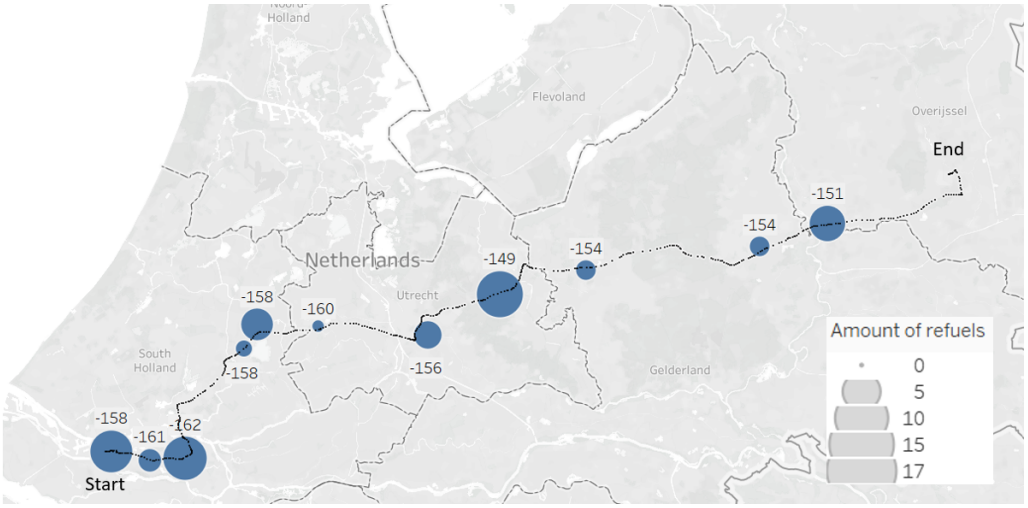

Figure 19: The gas stations that are selected for refueling over all policies for route 2642 with the size indicating the amount of advised refuelings

In Figure 20, the total refueling costs of policies from an initial fuel level of 31 to 123 are presented. Lower initial fuel levels exhibit higher total refueling costs due to an inability to reach the cheaper gas stations. Post fuel level 79, a decrease is observed because the gas station with costs of 149 is reached. For initial fuel levels higher than 115, no refueling occurs as the driver can complete the route without additional fuel. This pattern of lower total refueling costs with higher initial fuel levels is consistent across various routes. However, the cut-off point for refueling during the trip is route-specific and depends on the total fuel usage.

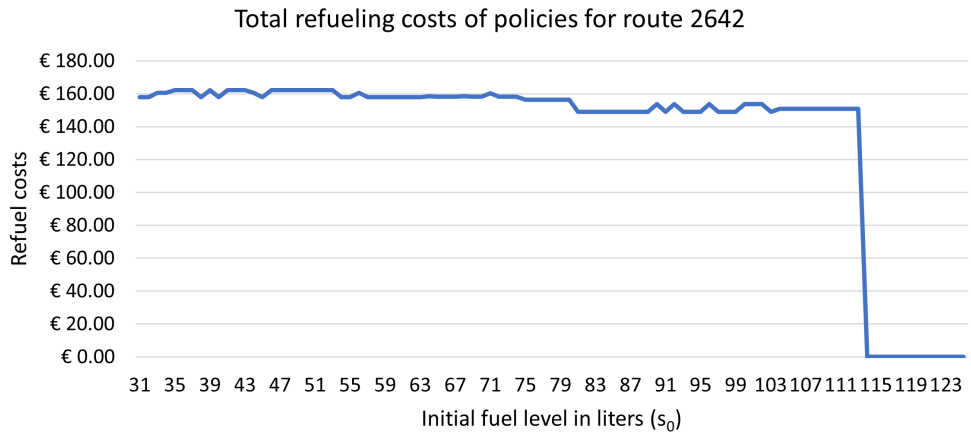

Figure 20: Costs of policies over different initial fuel levels (s_0) for route 2642

In summary, Figure 21 presents a map featuring gas stations advised across all policies

and routes. The marker size indicates the frequency of recommendations in policies. The map highlights a total of 108 different gas stations, with 83% being exclusively recommended in a single route, while the remaining 17% are advised in 2 to 4 other routes as well. Refer to Table 16 for details on these gas stations and the frequency of recommendations. The map reveals two notable hotspots for refueling locations: Zuid-Holland and Rijssen, where the company is situated. Zuid-Holland emerges as a potential area for Nijhof-Wassink to negotiate higher discounts, offering significant cost-saving opportunities.

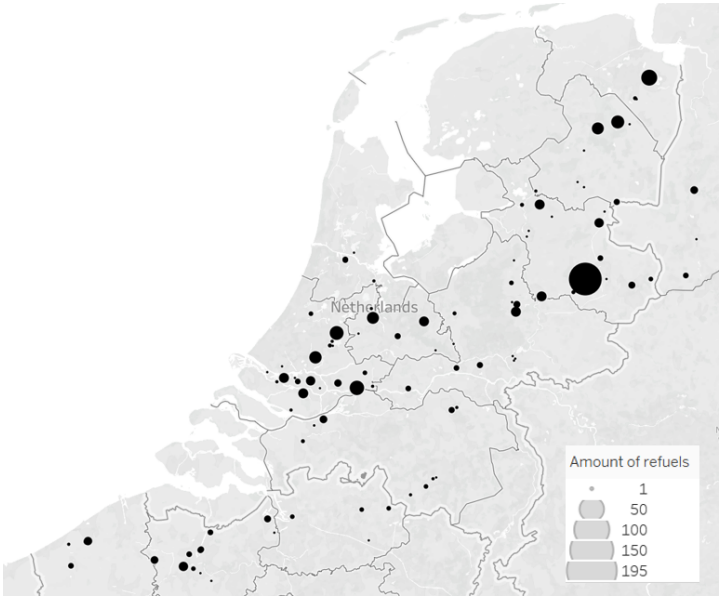

Figure 21: Gas stations advised in all policies over all routes with the size indicating the amount of advised refuelings

Table 16: Gas stations that are advised in multiple routes

Gas station	**#routes**	**#refuels**
Esso Bodegraven	4	35
Home base Rijssen	3	195
Esso Deventersiemelinksweg	3	18
BSP Apeldoorn Ecofactorij	3	9
Shellstation Struik	3	3
Esso Apeldoorn de Brink	3	19
Esso Midwolda	3	46
Gabriels Power Aaltertieltsestwg	2	12
Argos Hoogvliet	2	2
Shellstation Maatveld	2	28
Esso Bunnik de Forten	2	7
Shellstation Vondelingenweg	2	13
Esso Express Spijkenisse noord	2	2
Shell Okken	2	31
Esso Gentkennedylaan	2	7
Shellstation Portland	2	3
Esso Hardinxvelddenbout	2	38
Shellstation Rijssensestraat	2	2

6.5.3. Tactical Level

On the tactical planning level, the focus of insights shifts towards medium-term strategies spanning weeks to months. This section aims to establish relationships within the reward function and offer advice regarding the preferred network group or country for refueling by executing policy analysis.

Beginning with an examination of the variables in the reward function, see Equation (12), which consists of three parts. The first component (depicted in purple) represents fuel costs influenced by the net price and usage across the entire route. The second term (depicted in red) accounts for detour costs, incorporating factors such as detour length, detour time, net price, and fixed time costs. The third term (depicted in green) represents the fixed cost associated with the time taken to refuel. Given that usage remains constant over a route, the reward function introduces a trade-off for each route, weighing the detour length or time against the net fuel price.

$$r_t(s_t, a_t, s_{t+1}, \hat{p}_t) = \begin{cases} -(a_t((\hat{p}_t - k_t) \times \sum_{t=1}^{G} u_t + (\hat{p}_t - k_t) \times d_t - b_t - C)) & \text{if} \quad s_{t+1} > L, \\ -M & \text{if} \quad s_{t+1} \leq L. \end{cases} \tag{12}$$

To provide insight into this trade-off, Figure 22 illustrates how much the net fuel price should decrease for each additional kilometer of detour. For instance, on route 1236, a driver may extend their detour by 1 km if they can refuel for €0.005 cheaper. Similarly, on route 2622, a driver might consider a 1 km longer detour if the refueling cost is €0.007 lower. Notably, the graph reveals another relationship: the shorter the route, the lower the usage, and the more significant the impact of detour costs. Figure 22

exemplifies this phenomenon, demonstrating that for route 1236, with a usage of 133 liters, the relationship between detour length and net fuel price is less steep compared to route 2622, with a usage of 75 liters. This analysis contributes valuable insights into the nuanced dynamics of the trade-offs involved in refueling decisions across different routes.

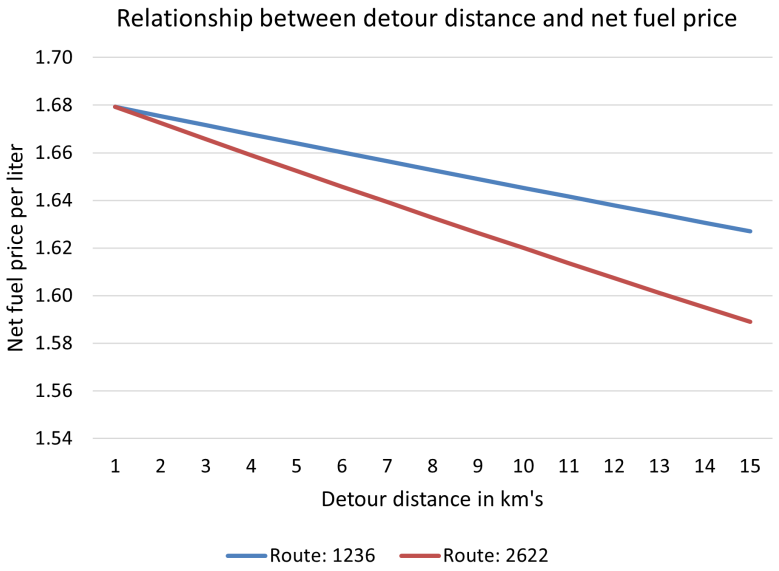

Figure 22: The relationship between the detour costs and the net fuel price

Second, the policy analysis is to find out which characteristics make gas stations preferable for refueling. First, we investigate in which country to refuel when a route traverses multiple countries. From the routes that we trained, route 1997 crosses the Dutch-German border and routes 1236 and 2622 cross the Dutch-Belgium border. For each scenario we analyze the policies, counting the instances where the recommendation advises refueling in one country over the other. In 72% of the policies, refueling in the Netherlands is recommended over Belgium. This is counter intuitive to the expectations due to the tax benefits in Belgium. However, this result can be explained by 75% of the route being in the Netherlands. For the routes crossing the Dutch-German border, Germany is recommended over the Netherlands in 58% of the policies. This is interesting as the forecasted pump prices in Figure 17 show better results for the Netherlands and the price agreements are also better in the Netherlands. The statistics can be explained by the route starting in Germany and taking place for 60% in Germany. For advice on which country is preferable it is required to test more routes that cross borders because currently, the characteristics of the route have a big impact on the result.

Delving further into the analysis of policies concerning refueling in the Netherlands and Belgium, we shift our focus to the frequent recommendations of specific network groups. Network groups categorize gas stations within a fuel company based on location and pricing, with price agreements often varying across these groups. Illustrated in Figure 23 is

the distribution of refueling recommendations across network groups in the Netherlands. The red groups represent stations affiliated with fuel company X, while the blue groups belong to fuel company Y. The figure reveals that 67% of the refueling recommendations within the Netherlands advocate refueling at company Y. To be more specific, the majority of the recommendations are within network group NT1 with 52%. Similarly, Figure 24 outlines a parallel distribution, this time focusing on network groups in Belgium. The graph illustrates a recommendation in favour of fuel company Y, with percentages standing at 90% versus 10% for company X. Notably, the optimal network groups in Belgium emerge as BT2 and BT4 with refueling recommendations of 27% and 26%, respectively.

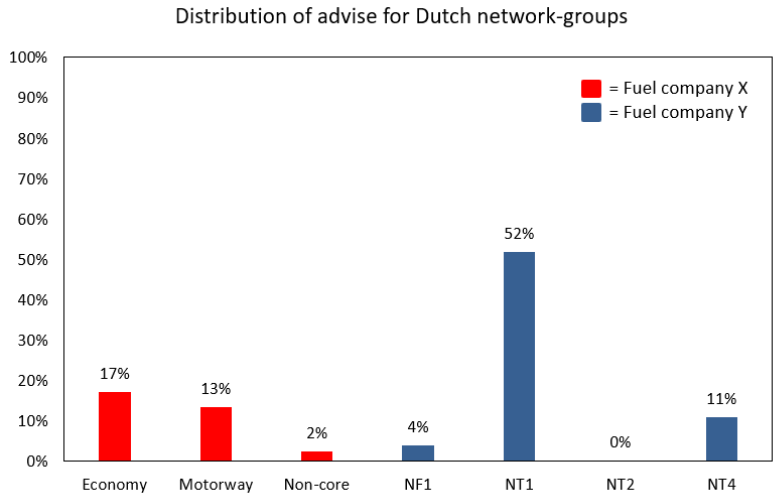

Figure 23: Representation of how often a network group is advised in a policy in the Netherlands

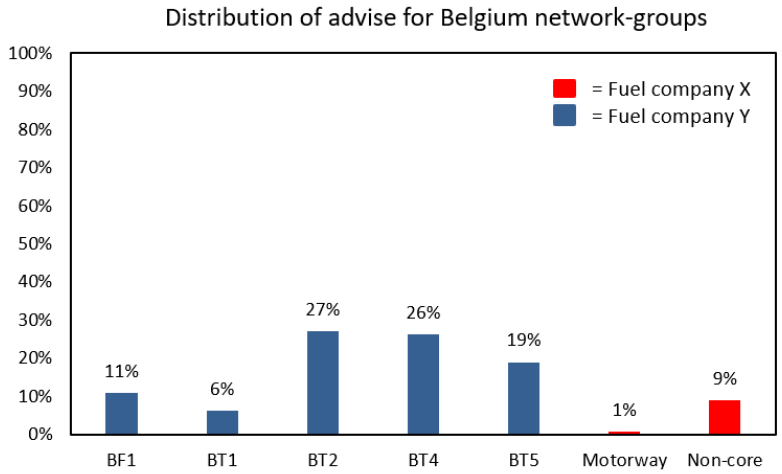

Figure 24: Representation of how often a network group is advised in a policy in Belgium

6.5.4. Strategic Level

Strategic insights center around Nijhof-Wassink's management goal of reducing fuel costs and their existing price agreements with fuel companies. For the goal of fuel cost reduction, we demonstrated the RL approach for refueling decisions can potentially decrease the total refueling costs for the DBL department by 11%, translating to an 88% increase in profits (Section 6.4.5). The refueling costs are influenced directly by the net prices, determined by discounts in price agreements and forecasted pump prices. Therefore, the subsequent analysis concentrates on delving into the dynamics of price agreements.

The analysis of price agreements involves selecting one network group and examining what happens to the total costs over all routes when an extra discount is negotiated. The network group selected is the NT1 network group of company Y because this network is recommended the most. Since this network group is frequently recommended, negotiating additional discounts for this group has the most substantial impact on total refueling costs. Figure 25 illustrates the results, revealing that, at first glance, extra discounts do not significantly impact total fuel costs, with less than a 1% decrease for a 5-cent extra discount. However, given that small changes in costs can determine the profitability of a transportation company and that a 1% decrease in fuel costs leads to an 8% increase in profit for Nijhof-Wassink, this result can still have a significant impact on the business. Table 17 provides a detailed overview of the increase in profit corresponding to extra discounts compared to the "no extra discount" scenario. Negotiating additional discounts can be challenging and, as shown, may not lead to substantial profit increases. Therefore, it is crucial to recognize that, rather than solely focusing on optimizing price agreements, emphasizing refining refueling advice through decision-making tools is a more effective strategy. Nevertheless, the RL approach proves valuable in testing various scenarios for price agreements, aiding in the determination of which approach outperforms others.

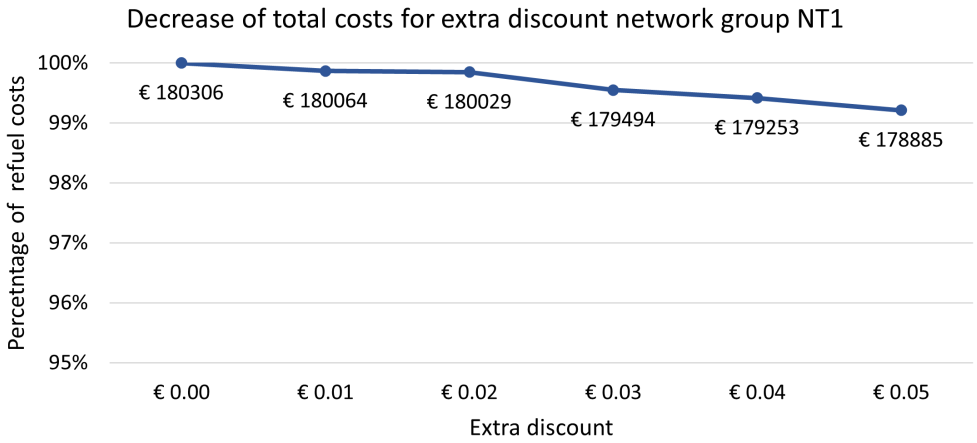

Figure 25: Relationship between the total fuel costs and extra discount

Table 17: The increase in profit for every discount level compared to the current discount level

Extra discount	**Increase in profit**
€ 0.00	0.00%
€ 0.01	1.07%
€ 0.02	1.23%
€ 0.03	3.60%
€ 0.04	4.67%
€ 0.05	6.30%

6.5.5. Take Away

This comprehensive analysis aimed to address the research question *"What insights does this model, in combination with the solution method, provide to the practical and academic communities?"*. The study unfolded valuable insights across multiple levels of decision-making. Firstly, in predicting fuel prices, the model highlighted the substantial variability in fuel prices influenced by independent variables, underlining the importance of considering the stochastic nature of the fuel prices. The insights showed that the predicted fuel prices are the lowest in Belgium and at network groups of fuel company Y. At the operational level, the RL algorithm's policies demonstrated refueling cost differentials along routes and the policies for different initial fuel levels. Tactical insights unveiled trade-offs in detour decisions for different lengths of routes. It showed for a route of 417 km, a 1 km extra detour should result in a net fuel price reduction of €0.005 to be profitable. Strategic considerations revealed potential profit gains through the negotiation of extra price discounts. The insight showed that focusing on refueling decisions is more profitable than focusing on price agreements.

7. Conclusion, Future Research, and Recommendations

This section begins with the conclusion of the research in Section 7.1, where we restate the aim, present the results, and address the main research question. After this, Section 7.2 outlines the key contributions for both academic and practical communities. Moving forward, Section 7.3 discusses the limitations and suggests directions for future research. Lastly, in Section 7.4, we provide recommendations tailored for Nijhof-Wassink.

7.1. Conclusion

The present research aimed to introduce a novel framework to reduce fuel costs for trucking companies in the transportation industry. The business understanding phase at Nijhof-Wassink underscored the practical significance of this research, showcasing a substantial 21% of total operating costs allocated to fuel and an impressive 8% increase in profits achievable through a 1% fuel cost reduction. This study revealed gaps in existing approaches for the refueling problem in the literature regarding variable and unknown fuel prices, and scalability to larger problem instances. Therefore, this research set out to develop a sequential decision-making framework that considers the complexities of the problem such as the variability of the fuel prices, limited access to fuel prices, and the fast-growing problem size. By framing the refueling problem as an MDP and leveraging RL techniques, the study seeks to offer a practical and academically novel decision-making framework. Ultimately, the goal was to answer the following main research question: *"Can we provide near-optimal solutions for the refueling problem with stochastic fuel prices by using a Reinforcement Learning approach, framed within a novel sequential decision-making framework?"*. To answer the research question, this study investigated the performance of the novel framework through a case study conducted at the DBL department of Nijhof-Wassink.

The FRVRP is defined as an MDP, where the decision epochs are the detour points to a nearby gas station, the state is the fuel level, and the action space is not refueling or refueling. The MDP considers the variable refueling costs at each gas station, the different detour costs to reach a gas station, the benefits of price agreements with certain gas stations, and the fixed costs for stopping to refuel. This research solves the MDP approximately with RL, which has proven to be an effective approach for complex, stochastic, and large-scale problems such as refueling decisions. Within RL, SARSA and Q-learning are two popular algorithms used to learn optimal policies for agents in an environment. For the FRVRP, Q-Learning demonstrated slightly better performance and less sensitivity to the parameters settings than SARSA.

The computational results of this study have not only validated RL for solving the FRVRP but also provided crucial insights into its practical efficacy. First, we ensured the generalizability of the framework, establishing the RL algorithm successfully applies to unseen scenarios. Subsequent analyses compared the RL algorithm against a benchmark heuristic, an optimal deterministic solution, and historical refuel decisions. Notably, the RL algorithm demonstrated a commendable 0.3% deviation from the deterministic optimal solution, underscoring its ability to provide solutions close to the lower bound. Moreover, it outperformed the benchmark heuristic by 3% and exhibited an 11% improvement in refueling costs compared to historical driver decisions. Realizing this

cost decrease within the DBL department saves between 660 and 990 thousand euros, increasing their profits by an impressive 88%.

Additionally, the study unfolded valuable insights across multiple decision-making levels. Firstly, in predicting fuel prices, the model highlighted substantial price variability influenced by independent variables, emphasizing the importance of considering fuel price stochasticity. The insights showed that the predicted fuel prices are the lowest in Belgium and at network groups of fuel company Y. At the operational level, the RL algorithm's policies demonstrated variations in refueling costs along routes and in policies for different initial fuel levels. Tactical analyses uncovered detour trade-offs, suggesting a €0.005 reduction in net fuel price per extra kilometer for profitable detours on a route of 417 km. Strategically, the study emphasized the profitability of focusing on refueling decisions over price agreements, showcasing minimal potential profit gains through negotiation of extra discounts.

Overall, these results not only confirm the RL algorithm's effectiveness in providing near-optimal solutions, but also shed light on its practical superiority in addressing the complexities of the FRVRP by incorporating the uncertainty of the fuel prices in the decision-making process. We conclude that the Reinforcement Learning approach, integrated into a novel sequential decision-making framework, effectively addresses the main research question with near-optimal policies resulting in an optimality gap of 0.3%.

7.2. Main Contributions

To summarize, the main academic and practical contributions of this research are (1) a data-driven mathematical framework using MDP approach for the refueling problem (2) an RL algorithm as an appropriate solution method to tackle the uncertainty and complexity of the problem with a 0.3% optimality gap, (3) an ML regression model to deal with the uncertainty of the fuel prices and (4), a case study at Nijhof-Wassink that successfully proved the practical use of this model. These contributions advance the understanding and practical implementation of optimal refueling strategies, paving the way for new research, and more efficient and cost-effective operations in the transportation industry.

The main contributions for Nijhof-Wassink are (1) indicating a possible 88% increase in profits by decreasing the refuel costs by 11% with the policies from the RL algorithm (2) new knowledge of AI and its potential use for the refueling problem (3) a well-performing RL algorithm applicable to all fixed routes within Nijhof-Wassink Group (3) insights and improvements in the data structure surrounding refueling decisions, and (4) tactical and operational insights from the RL policies to improve refueling decisions and decrease refueling costs.

7.3. Limitations and Future Research

Despite the contributions of the proposed MDP model for the Fixed Route Vehicle Refueling Problem, it is crucial to acknowledge certain limitations and provide directions for future research. First, the model's performance has been evaluated primarily on short routes, and its effectiveness under more extended routes remains uncertain. This raises

questions about the robustness of the model's decision-making capabilities for bigger problem instances. Future research can entail extending the case study to accommodate longer routes, causing multiple refueling instances. Second, when confronted with routes that have not been trained, the RL limits its ability to provide meaningful recommendations. Future research on this limitation can include venturing the refueling problem into an infinite MDP framework. For the infinite MDP we propose including all gas stations in a network where the transition from one gas station to another is probabilistic. Training the network yields a policy for every gas station enabling recommendations for unseen routes and thereby expanding the model's utility in diverse and unforeseen scenarios. Third, the validation of the model's practical contributions is hampered by the limited availability of historical data. The scarcity of extensive and diverse datasets restricts the depth of insights that can be drawn regarding the model's real-world applicability. For future research, the case study can be expanded by more historical data or applying the policies in real-life to strengthen and validate the practical contribution. Fourth, while the refueling problem is modeled as a finite MDP optimized over a single route, real-world scenarios often involve consecutive routes. In situations where the truck can complete the first route without refueling but requires refueling for the subsequent route, the current approach only considers gas stations along the second route. However, there may be cheaper refueling options available along the first route. To address this limitation, it is proposed to introduce a terminal reward that considers both the terminal fuel level and the expected reward of the route driven next. By doing so, refueling in the first route is incentivized to achieve a better reward by avoiding a high terminal reward when refueling in the subsequent route.

In the realm of future research, several changes to the MDP formulation can be explored to possibly enhance the applicability of the proposed MDP model for the FRVRP. The first change in the formulation is broadening the action space to include the decision of the amount to refuel at each station. This offers a more accurate representation of real-world refueling strategies and the possibility of refueling smaller batches. The second addition to the formulation is incorporating the fuel price prediction into the state space. This can render the model more adaptive to dynamic price conditions by providing advice based on the price level as well. The third factor that can be added to the formulation is the integration of driving-rest time regulations in the state space. This provides the opportunity to save costs by combining refueling with resting. The fourth addition is adding the fuel level of AdBlue to the state space. AdBlue is an essential additive in diesel trucks and is not available at every gas station. The policies emerging from the RL algorithm advise refueling at gas stations without AdBlue 53% of the time. Combining refueling diesel and AdBlue, can contribute to more complete and cost-effective advice on where to refuel. The fifth change in formulation is regarding the element of driver freedom. According to Suzuki (2009) the driver should be given a free choice for practical use. To incorporate this in the model the formulation can be changed to provide a list of preferred gas stations at the start of a route instead of advice to refuel in a predetermined gas station.

7.4. Recommendations for the Company

In this section we present the recommendations for Nijhof-Wassink and address the last research question *"Does this model have the potential to be further expanded and de-*

ployed in Nijhof-Wassink?". During the case study at Nijhof-Wassink Group, we gained insight into the organization and their challenges. In Section 2.1 we showed that for each percent of fuel savings, the profit increases by 8%. Subsequently, we highly recommend continuing the project to reduce fuel costs. This section aims to provide the company with the next steps on how to reduce fuel costs. At the beginning of this research, Nijhof-Wassink stated their ideal situation is having a data-driven approach to generate real-time advice to the drivers via a live connection with the board computers of the trucks. Figure 26 shows different data-driven approaches companies can adopt to support decision-making: descriptive analysis, predictive analysis, and prescriptive analysis. For all these approaches good data structure and quality are required, however, this has proven to be an ongoing challenge during this project. Besides, we see that prescriptive analytics such as the RL model provide the most value, but are complex and costly to implement. Considering the absence of any existing decision support tool, data-driven approach, or established data structure for refueling decisions, we believe that implementing prescriptive analytics at this stage might be too ambitious for Nijhof-Wassink. Our recommendation is to prioritize enhancing the data structure and quality initially. Subsequently, utilizing descriptive analytics can aid in formulating strategic decisions.

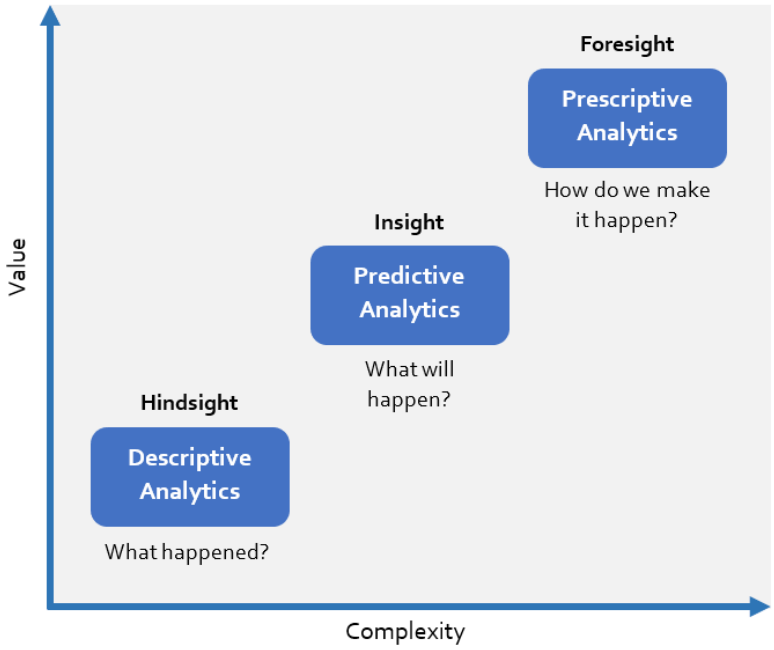

Figure 26: Different data-driven approaches for decision making in businesses

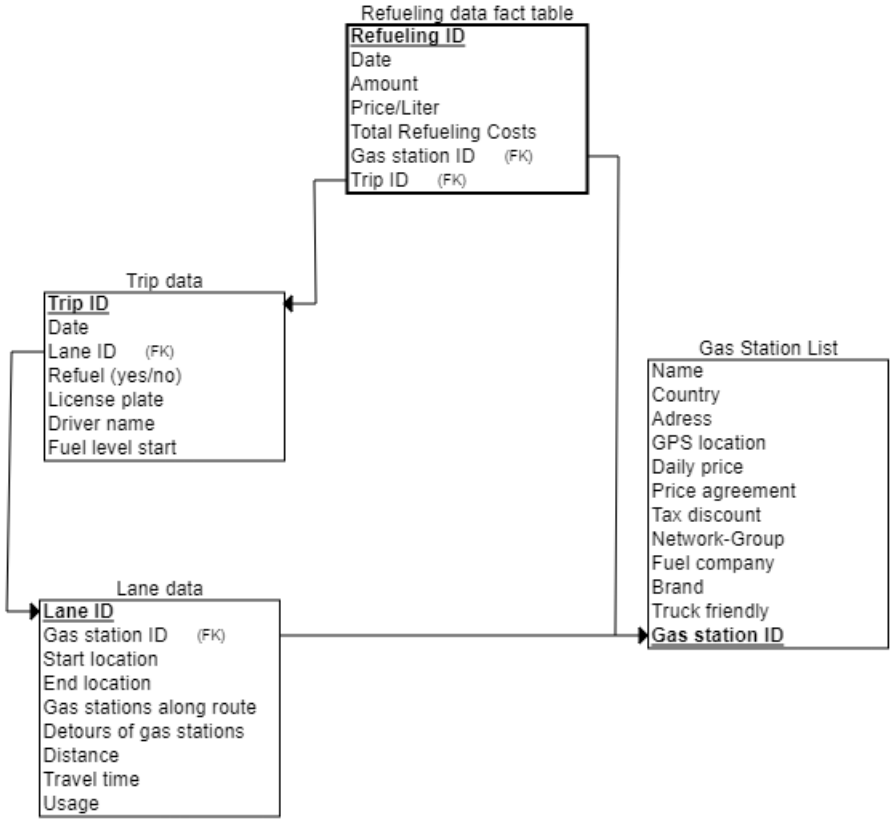

Figure 27: Visualization of envisioned data structure

For improving the data structure and quality we envison the ideal situation as depicted in Figure 27. The following actions may be taken to achieve this data structure:

1. Keep a database with all the gas stations, their correct GPS location and characteristics such as truck refueling, AbBlue, and price agreement.
2. Collect the daily fuel prices for all gas stations and add them to the gas station list.
3. Develop an identifier for each gas station so that a transaction can be matched to a gas station from the gas station list.
4. Work on improving the data quality of the fuel transactions since the price per liter that is paid shows a lot of outliers, also the price discounts do not equal the price discounts retrieved from the price agreements.
5. Work on adding the price per liter of the home bases to the transactions to complete the overview.
6. In the trip data from ORD we can see on which trip a driver refueled. However, we do not know at which gas station and for which price. Create a connection so that the refueling decision can be evaluated.

7. Create identifiers for all lanes. This way you can collect the gas stations, and their detours for each lane and access this data when a shift consists of multiple lanes instead of calculating it again.

Once the data structure and quality are ensured, Nijhof-Wassink can focus on visualizing the data with descriptive analytic tools. This improves the data understanding and supports developing decision strategies. The visualization can work as a validation tool for strategic decisions. Furthermore, the visualization enables Nijhof-Wassink to estimate the potential cost savings and determine the amount worth investing in a decision-making tool. With in-house expertise in developing visualizations, this proves to be the most cost-effective next step.

When Nijhof-Wassink is ready for the next step towards prescriptive analytics, RL can be a powerful tool to solve the refueling problem, particularly if the company intends to expand the scope by incorporating longer routes, additional actions, or other extensions outlined for further research. Given the complexity of developing RL, we present alternative, simpler methods that discard the stochasticity of the fuel prices. The first option involves solving each route to optimality, with the average fuel prices over a period of time, using the optimal method presented in this research. When Nijhof-Wassink wishes to increase the length of a route so that two refuelings are needed, they can adapt the optimal method or decide to use the benchmark heuristic presented in this research.

References

Abdullah, H.M., Gastli, A., Ben-Brahim, L., 2021. Reinforcement learning based ev charging management systems–a review. IEEE Access 9, 41506–41531. doi:10.1109/ACCESS.2021.3064354.

Ahiska, S.S., Appaji, S.R., King, R.E., Warsing, D.P., 2013. A markov decision process-based policy characterization approach for a stochastic inventory control problem with unreliable sourcing. International Journal of Production Economics 144, 485–496. URL: https://www.sciencedirect.com/science/article/pii/S0925527313001370, doi:https://doi.org/10.1016/j.ijpe.2013.03.021.

Atamtürk, A., Küçükyavuz, S., 2005. Lot sizing with inventory bounds and fixed costs: Polyhedral study and computation. Operations Research 53, 711–730.

Bousonville, T., Bernhardt, A., Melo, T., Kopfer, H., 2011. Vehicle routing and refueling: The impact of price variations on tour length, pp. 83–101.

Bureau of Transportation Statistics, 2022. Record-breaking increases in motor fuel prices 2022. Website. URL: https://www.bts.gov/data-spotlight/record-breaking-increases-motor-fuel-prices-2022.

Chang, H.S., Hu, J., Fu, M.C., Marcus, S.I., 2013. Markov Decision Processes. Springer London, London. pp. 1–17. URL: https://doi.org/10.1007/978-1-4471-5022-0_1, doi:10.1007/978-1-4471-5022-0_1.

Chiş, A., Lundén, J., Koivunen, V., 2013. Scheduling of plug-in electric vehicle battery charging with price prediction , 1–5doi:10.1109/ISGTEurope.2013.6695263.

Chiş, A., Lundén, J., Koivunen, V., 2017. Reinforcement learning-based plug-in electric vehicle charging with forecasted price. IEEE Transactions on Vehicular Technology 66, 3674–3684. doi:10.1109/TVT.2016.2603536.

Dabney, W.M., 2014. Adaptive step-sizes for reinforcement learning. URL: https://api.semanticscholar.org/CorpusID:59866489.

Dellaert, N., Melo, M., 1996. Stochastic lot-sizing: Solution and heuristic methods. International Journal of Production Economics 46-47, 261–276. URL: https://www.sciencedirect.com/science/article/pii/0925527395001123, doi:https://doi.org/10.1016/0925-5273(95)00112-3. proceedings of the 8th International Working Seminar on Production Economics.

Diederich, A., 2001. Sequential decision making, in: Smelser, N.J., Baltes, P.B. (Eds.), International Encyclopedia of the Social & Behavioral Sciences. Pergamon, Oxford, pp. 13917–13922. URL: https://www.sciencedirect.com/science/article/pii/B0080430767006367, doi:https://doi.org/10.1016/B0-08-043076-7/00636-7.

EuroStat, 2023. Key figures on europe: 2023 edition. URL: https://ec.europa.eu/eurostat/documents/15216629/17177791/KS-EI-23-001-EN-N.pdf/5df7a393-8461-9270-7eaa-91a4b1c2acc6?version=2.0&t=1689583429855.

Farazi, N.P., Zou, B., Ahamed, T., Barua, L., 2021. Deep reinforcement learning in transportation research: A review. Transportation Research Interdisciplinary Perspectives 11, 100425. URL: https://www.sciencedirect.com/science/article/pii/S2590198221001317, doi:https://doi.org/10.1016/j.trip.2021.100425.

Farkas, M., Csehi, C.G., 2017. Truck routing and scheduling. Cent Eur J Oper Res 25, 791–807 doi:https://doi.org/10.1007/s10100-016-0453-8.

Friedman, J.H., 2001. Greedy function approximation: A gradient boosting machine. The Annals of Statistics 29, 1189–1232. URL: http://www.jstor.org/stable/2699986.

Giannoccaro, I., Pontrandolfo, P., 2002. Inventory management in supply chains: a reinforcement learning approach. International Journal of Production Economics 78, 153–161. URL: https://www.sciencedirect.com/science/article/pii/S0925527300001560, doi:https://doi.org/10.1016/S0925-5273(00)00156-0.

Gicquel, C., Minoux, M., Dallery, Y., 2008. Capacitated lot sizing models: a literature review. HAL URL: ffhal-00255830f.

Gupta, M.K., Hemachandra, N., Bhatnagar, S., 2022. Chapter 6 - learning in sequential decision-making under uncertainty, in: Pandey, R., Khatri, S.K., kumar Singh, N., Verma, P. (Eds.), Artificial Intelligence and Machine Learning for EDGE Computing. Academic Press, pp. 75–85. URL: https://www.sciencedirect.com/science/article/pii/B9780128240540000113, doi:https://doi.org/10.1016/B978-0-12-824054-0.00011-3.

Gutiérrez, J., Sedeño-Noda, A., Colebrook, M., Sicilia, J., 2003. A new characterization for the dynamic lot size problem with bounded inventory. Computers & Operations Research 30, 383–395.

van Hezewijk, L., Dellaert, N., Woensel, T.V., Gademann, N., 2022. Using the proximal policy optimisation algorithm for solving the stochastic capacitated lot sizing problem. International

Journal of Production Research 0, 1–24. URL: https://doi.org/10.1080/00207543.2022.2056540, doi:10.1080/00207543.2022.2056540, arXiv:https://doi.org/10.1080/00207543.2022.2056540.

Khuller, S., Malekian, A., Mestre, J., 2007. To fill or not to fill: The gas station problem, in: Arge, L., Hoffmann, M., Welzl, E. (Eds.), Algorithms – ESA 2007, Springer Berlin Heidelberg, Berlin, Heidelberg. pp. 534–545.

Kitchenham, B., 2004. Procedures for performing systematic reviews. Keele, UK, Keele Univ. 33.

Kuo, Y., 2010. Using simulated annealing to minimize fuel consumption for the time-dependent vehicle routing problem. Computers & Industrial Engineering 59, 157–165. URL: https://www.sciencedirect.com/science/article/pii/S0360835210000835, doi:https://doi.org/10.1016/j.cie.2010.03.012.

Leslie, A., Murray, D., 2022. An analysis of the operational costs of trucking: 2022 update. American Transportation Research Institute (ATRI) .

Li, H., Wan, Z., He, H., 2020. Constrained ev charging scheduling based on safe deep reinforcement learning. IEEE Transactions on Smart Grid 11, 2427–2439. doi:10.1109/TSG.2019.2955437.

Lin, S.H., 2008. Finding optimal refueling policies in transportation networks, in: Fleischer, R., Xu, J. (Eds.), Algorithmic Aspects in Information and Management, Springer Berlin Heidelberg, Berlin, Heidelberg. pp. 280–291.

Lin, S.H., 2014. Multi-objective vehicle refueling planning using mixed integer programming, in: 2014 IEEE International Conference on Industrial Engineering and Engineering Management, pp. 677–681. doi:10.1109/IEEM.2014.7058724.

Lin, S.H., Gertsch, N., Russell, J.R., 2007. A linear-time algorithm for finding optimal vehicle refueling policies. Operations Research Letters 35, 290–296. URL: https://www.sciencedirect.com/science/article/pii/S0167637706000666, doi:https://doi.org/10.1016/j.orl.2006.05.003.

Mckinnon, A., 2023. Increasing fuel prices and market distortion in a domestic road haulage market: the case of the united kingdom .

Mes, M.R., Rivera, P., Eduardo, A., 2017. Approximate Dynamic Programming by Practical Examples. Springer. Number 248 in International Series in Operations Research; Management Science, pp. 63–101. doi:10.1007/978-3-319-47766-4_3.

Nazari, M., Oroojlooy, A., Snyder, L.V., Takác, M., 2018. Reinforcement learning for solving the vehicle routing problem. CoRR abs/1802.04240. URL: http://arxiv.org/abs/1802.04240, arXiv:1802.04240.

Neves-Moreira, F., Amorim-Lopes, M., Amorim, P., 2020. The multi-period vehicle routing problem with refueling decisions: Traveling further to decrease fuel cost? Transportation Research Part E: Logistics and Transportation Review 133, 101817. URL: https://www.sciencedirect.com/science/article/pii/S1366554519306702, doi:https://doi.org/10.1016/j.tre.2019.11.011.

NOS, 2023. Weer hogere prijs aan de pomp: hoe komt het? News article. URL: https://nos.nl/artikel/2472353-weer-hogere-prijs-aan-de-pomp-hoe-komt-het.

Okhrin, I., Richter, K., 2011. An o (t3) algorithm for the capacitated lot sizing problem with minimum order quantities. European Journal of Operational Research 211, 507–514.

Ottoni, A.L.C., Nepomuceno, E.G., de Oliveira, M.S., de Oliveira, D.C.R., 2021. Reinforcement learning for the traveling salesman problem with refueling. Complex & Intelligent Systems 8, 2001 – 2015.

Persyn, D., Díaz-Lanchas, J., Barbero, J., 2019. Estimating road transport costs between eu regions. JRC Working Papers on Territorial Modelling and Analysis No. 04/2019, European Commission, Seville, JRC114409 .

Pilz, J., Melas, V.B., Bathke, A., 2023. International Workshop on Simulation and Statistics. Springer. URL: https://doi.org/10.1007/978-.

Powell, W., Ryzhov, I., 2012. Optimal learning and approximate dynamic programming. Reinforcement Learning and Approximate Dynamic Programming for Feedback Control doi:10.1002/9781118453988.ch18.

Powell, W.B., 2011. Approximate Dynamic Programming: Solving the Curses of Dimensionality. Wiley Series in Probability and Statistics. 2nd ed., Wiley, Hoboken, NJ, USA.

Puterman, M.L., 1994. Markov Decision Processes: Discrete Stochastic Dynamic Programming. 1st ed., John Wiley & Sons, Inc., USA.

Rodrigues Junior, A.D., Cruz, M.M.d.C., 2013. A generic decision model of refueling policies: a case study of a brazilian motor carrier. Journal of Transport Literature 7. URL: https://www.scielo.br/j/jtl/a/rnrfyCPMgc84ZrK63GS5nGj/.

Schulz, A., Suzuki, Y., 2023. An efficient heuristic for the fixed-route vehicle-refueling problem. Transportation Research Part E: Logistics and Transportation Review 169, 102963. URL: https://www.sciencedirect.com/science/article/pii/S1366554522003404, doi:https://doi.org/10.1016/

j.tre.2022.102963.
Shakya, M., Ng, H.Y., Ong, D.J., Lee, B.S., 2022. Reinforcement learning approach for multi-period inventory with stochastic demand, in: Maglogiannis, I., Iliadis, L., Macintyre, J., Cortez, P. (Eds.), Artificial Intelligence Applications and Innovations, Springer International Publishing, Cham. pp. 282–291.
Shi, W., Wong, V.W., 2011. Real-time vehicle-to-grid control algorithm under price uncertainty , 261–266doi:10.1109/SmartGridComm.2011.6102330.
Shin, J., Lee, J.H., 2015. Mdp formulation and solution algorithms for inventory management with multiple suppliers and supply and demand uncertainty, in: Gernaey, K.V., Huusom, J.K., Gani, R. (Eds.), 12th International Symposium on Process Systems Engineering and 25th European Symposium on Computer Aided Process Engineering. Elsevier. volume 37 of *Computer Aided Chemical Engineering*, pp. 1907–1912. URL: https://www.sciencedirect.com/science/article/pii/B9780444635761500121, doi:https://doi.org/10.1016/B978-0-444-63576-1.50012-1.
Sox, C.R., Jackson, P.L., Bowman, A., Muckstadt, J.A., 1999. A review of the stochastic lot scheduling problem1this paper is based upon work supported in part by the national science foundation under grant number dmi-9409344.1. International Journal of Production Economics 62, 181–200. URL: https://www.sciencedirect.com/science/article/pii/S0925527398002473, doi:https://doi.org/10.1016/S0925-5273(98)00247-3.
Sutton, R.S., Barto, A.G., 2018. Reinforcement Learning: An Introduction. A Bradford Book, Cambridge, MA, USA.
Suzuki, Y., 2008. A generic model of motor-carrier fuel optimization. Naval Research Logistics (NRL) 55, 737–746. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/nav.20317, doi:https://doi.org/10.1002/nav.20317.
Suzuki, Y., 2009. A decision support system of dynamic vehicle refueling. DECISION SUPPORT SYSTEMS 46, 522–531. doi:10.1016/j.dss.2008.09.005.
Suzuki, Y., 2014. A variable-reduction technique for the fixed-route vehicle-refueling problem. Computers & Industrial Engineering 67, 204–215. URL: https://www.sciencedirect.com/science/article/pii/S0360835213003707, doi:https://doi.org/10.1016/j.cie.2013.11.007.
Suzuki, Y., Dai, J., 2013. Decision support system of truck routing and refueling: A dual-objective approach. Decision Sciences 44, 817–842. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/deci.12029, doi:https://doi.org/10.1111/deci.12029.
Suzuki, Y., Lan, B., 2018. Cutting fuel consumption of truckload carriers by using new enhanced refueling policies. International Journal of Production Economics 202, 69–80. URL: https://www.sciencedirect.com/science/article/pii/S0925527318302019, doi:https://doi.org/10.1016/j.ijpe.2018.05.007.
Suzuki, Y., Montabon, F., Lu, S.H., 2014. Dss of vehicle refueling: A new enhanced approach with fuel weight considerations. Decision Support Systems 68, 15–25. URL: https://www.sciencedirect.com/science/article/pii/S0167923614002504, doi:https://doi.org/10.1016/j.dss.2014.10.005.
Winder, P., 2021. The future of transportation infrastructure: Reinforcement learning. Website article. URL: https://winder.ai/the-future-of-transportation-infrastructure-reinforcement-learning/.
Wirth, R., Hipp, J., 2000. Crisp-dm: towards a standard process modell for data mining. URL: https://api.semanticscholar.org/CorpusID:1211505.
Xiao, Y., Zhao, Q., Kaku, I., Xu, Y., 2012. Development of a fuel consumption optimization model for the capacitated vehicle routing problem. Computers & Operations Research 39, 1419–1431. URL: https://www.sciencedirect.com/science/article/pii/S0305054811002450, doi:https://doi.org/10.1016/j.cor.2011.08.013.
Yan, Y., Chow, A.H., Ho, C.P., Kuo, Y.H., Wu, Q., Ying, C., 2022. Reinforcement learning for logistics and supply chain management: Methodologies, state of the art, and future opportunities. Transportation Research Part E: Logistics and Transportation Review 162, 102712. URL: https://www.sciencedirect.com/science/article/pii/S136655452200103X, doi:https://doi.org/10.1016/j.tre.2022.102712.
Zhang, F., Yang, Q., An, D., 2021. Cddpg: A deep-reinforcement-learning-based approach for electric vehicle charging control. IEEE Internet of Things Journal 8, 3075–3087. doi:10.1109/JIOT.2020.3015204.
Zhang, J., Zhao, Y., Xue, W., Li, J., 2015. Vehicle routing problem with fuel consumption and carbon emission. International Journal of Production Economics 170, 234–242. URL: https://www.sciencedirect.com/science/article/pii/S0925527315003692, doi:https://doi.org/10.1016/j.ijpe.2015.09.031.

Appendix A. Literature Search Documentation

The systematic literature review is conducted according to the steps of Kitchenhamm (2004). For this literature review we used digital libraries, reference lists from relevant studies and the Internet. In order to construct a good search string, synonyms and closely relates terms are identified by trial and error and linked with the Boolean OR. See Table A.18.

Table A.18: Search terms

Key word	**Related terms**
Refueling OR	refuel* OR recharg* OR fueling OR charging
Optimization OR	solving OR problem OR model OR policy OR VRP
Logistics OR	transportation OR logistical OR transport OR truck OR vehicle
Approach OR	exact OR heuristic OR MDP OR RL OR linear OR deterministic OR stochastic

From these search terms we can compose our final search query that covers all topics we want to include:

(Refueling OR Refuel* OR recharg* OR fueling OR charging OR 'fuel consumption') AND (Optimization OR Solving OR problem OR model OR approach) AND (Logistics OR Transportation OR logistical OR transport OR truck) AND (Approach OR exact OR heuristic OR MDP OR Markov decision OR Reinforcement learning OR linear OR deterministic OR stochastic)

The search process documentation can be found in Table A.19. The databases that are used are chosen on whether they cover a relevant research area for our problem and the size of the database.

Table A.19: Final searches

Database	**Query**	**Hits**	**Sources retrieved**
Scopus	Final query	150	14
	(inventory AND ("sequential decision making" OR MDP OR RL)	277	3
Scholar	Final query	40.600	4
Reference lists	n.a.	n.a.	9
Internet	n.a.	n.a.	3
Books	n.a.	n.a.	3
Total			**36**

Later, additional topics such as related work in MDPs and inventory management are added. In total 35 papers are used for the literature review and several problem classes were covered.

Appendix B. Selected routes

Table B.20: Route information

Route	ID	Frequency	# Decision Epochs	Orgin	Destination
0	2642	747	101	NL	NL
1	3208	506	36	NL	NL
2	804	278	45	NL	NL
3	1997	233	32	DE	NL
4	2820	230	42	NL	NL
5	1236	223	133	NL	BE
6	510	208	30	BE	BE
7	328	207	60	NL	NL
8	353	193	20	NL	NL
9	1915	177	113	NL	NL
10	2621	176	20	NL	NL
11	2941	171	25	NL	NL
12	1675	157	113	NL	NL
13	2910	148	53	NL	NL
14	1237	126	47	NL	NL
15	1556	124	47	NL	NL
16	678	123	20	BE	BE
17	1888	111	17	NL	NL
18	2622	111	75	NL	BE
19	3170	111	20	NL	NL
20	1238	110	34	NL	NL

Appendix C. Experiment configurations

Table C.21: Experiment configurations

Experiment	Δ	e
0	23500	0.10
1	68500	0.07
2	86500	0.04
3	41500	0.05
4	50500	0.01
5	59500	0.08
6	77500	0.03
7	14500	0.02
8	95500	0.09
9	32500	0.06
10	14500	0.78
11	77500	0.15
12	59500	0.42
13	50500	0.24
14	68500	0.60
15	32500	0.96
16	23500	0.87
17	95500	0.51
18	41500	0.69
19	86500	0.33
20	50500	9.55
21	23500	5.05
22	86500	2.35
23	59500	8.65
24	41500	7.75
25	95500	6.85
26	68500	4.15
27	14500	5.95
28	77500	3.25
29	32500	1.45

Appendix D. Results: parameter tuning

Appendix D.1. Best results per experimental settings SARSA-based FRVRP algorithm

Table D.22: Result of SARSA-based FRVRP algorithm with parameter settings Δ and e

Experiment	Δ	e	**Total result**	**Gap**
8	95500	0.09	180322.6	0.09%
2	86500	0.04	180565.3	0.22%
6	77500	0.03	180732.0	0.32%
1	68500	0.07	180733.7	0.32%
11	77500	0.15	180922.5	0.42%
5	59500	0.08	181259.7	0.61%
4	50500	0.01	181775.1	0.89%
3	41500	0.05	181973.8	1.00%
9	32500	0.06	182909.7	1.52%
13	50500	0.24	182929.5	1.54%
12	59500	0.42	183405.8	1.80%
0	23500	0.10	185138.7	2.76%
7	14500	0.02	185701.1	3.07%
19	86500	0.33	189979.6	5.45%
18	41500	0.69	192005.7	6.57%
14	68500	0.60	192178.1	6.67%
17	95500	0.51	197566.7	9.66%
10	14500	0.78	203963.0	13.21%
15	32500	0.96	205407.4	14.01%
29	32500	1.45	209820.6	16.46%
16	23500	0.87	217914.7	20.95%
22	86500	2.35	244503.6	35.71%
21	23500	5.05	259213.0	43.88%
28	77500	3.25	264433.4	46.77%
27	14500	5.95	273425.0	51.77%
26	68500	4.15	285242.4	58.32%
24	41500	7.75	312017.1	73.19%
25	95500	6.85	342288.4	89.99%
20	50500	9.55	361966.6	100.91%
23	59500	8.65	377582.1	109.58%

Appendix D.2. Best results per experimental settings Q-Learning-based FRVRP algorithm

Table D.23: Result of Q-Learning-based FRVRP algorithm with parameter settings Δ and e

Experiment	Δ	e	**Total result**	**Gap**
17	95500	0.51	180305.9	0.06%
25	95500	6.85	180409.0	0.12%
8	95500	0.09	180409.2	0.12%
22	86500	2.35	180468.5	0.15%
2	86500	0.04	180507.4	0.17%
19	86500	0.33	180534.9	0.19%
6	77500	0.03	180542.6	0.19%
11	77500	0.15	180595.0	0.22%
28	77500	3.25	180607.9	0.23%
14	68500	0.60	180769.4	0.32%
1	68500	0.07	180815.3	0.34%
23	59500	8.65	180921.7	0.40%
26	68500	4.15	180975.7	0.43%
12	59500	0.42	180999.7	0.45%
5	59500	0.08	181166.9	0.54%
20	50500	9.55	181453.2	0.70%
4	50500	0.01	181748.6	0.86%
13	50500	0.24	181799.1	0.89%
3	41500	0.05	181936.8	0.97%
24	41500	7.75	181945.6	0.97%
18	41500	0.69	181996.8	1.00%
15	32500	0.96	182441.4	1.25%
29	32500	1.45	182807.8	1.45%
9	32500	0.06	182826.7	1.46%
0	23500	0.10	183915.9	2.07%
21	23500	5.05	184018.7	2.12%
16	23500	0.87	184022.0	2.12%
7	14500	0.02	185687.1	3.05%
10	14500	0.78	185785.5	3.10%
27	14500	5.95	185885.0	3.16%

Appendix E. Selected routes for testing

Table E.24: Route information parameter testing

Route	ID	# Decision Epochs	Orgin	Destination
0	1055	73	BE	DE
1	1511	65	NL	DE
2	538	131	NL	NL
3	1510	32	NL	DE
4	88	43	NL	BE
5	2148	66	BE	DE
6	1457	40	NL	DE
7	3204	42	NL	NL
8	2959	71	DE	BE
9	287	55	BE	DE