Deep reinforcement learning for global network management

van der Wilk, M.A.

Award date:
2021

EINDHOVEN UNIVERSITY OF TECHNOLOGY

Department of Industrial Engineering and Innovation Sciences
Operations, Planning, Accounting & Control group

Deep reinforcement learning for global network management

Master Thesis

M.A. (Max) van der Wilk
0993078

University Supervisor
Dr. W.L. (Willem) van Jaarsveld
Dr. A. (Ahmadreza) Marandi

Company Supervisor
Ir. L.G. (Luke) van de Bunt, Den Hartogh Logistics

Eindhoven
October 1, 2021

Eindhoven University of Technology
Department of Industrial Engineering & Innovation Sciences
Series Master Thesis Operations Management and Logistics

Keywords: *tank container industry, global network management, container imbalances, dynamic pricing, price elasticity, deep reinforcement learning*

Abstract

This research proposes a proof of concept of deep reinforcement learning for the global network management of a tank container operator. Tank container operators are faced with the challenge of tank container stock imbalances which are caused by fluctuating global demand. Global imbalances are minimized with dynamic pricing. DRL was able to develop pricing policies that generated more profit than the baseline policy. A sensitivity analysis was conducted to evaluate the impact of various factors on performance. In addition, a cross comparison analysis is conducted between DRL agents that are trained on different scenarios (i.e. MDP instances). This provides insights into the robustness of DRL policies for various uncertainty factors in the global network management. This research has provided valuable insights into the aspects of the global network management that can be improved. The information about the expected demand of contracted customers seems to be critical for the network management, while factors such as profit margin, penalty costs or price elasticity tend to have limited impact. Moreover, determining prices more accurately (e.g. distinguishing contracts by duration or expected number of tanks) is expected to result in a substantial increase in profits.

Executive summary

Den Hartogh experiences difficulty in forecasting both short-term demand as well as long-term demand. This leads to high uncertainty in how the global network will evolve. The global network management team tries to control the network imbalances with a market correction. The value of the market correction influences the probability of winning a contract. However, it is difficult to predict the impact of new market corrections without having a formalized understanding of the price elasticity. The ambiguous effects of the market correction result in more hesitance in the timing and value of the market correction. This ultimately leads to regional surpluses and shortages of tank containers that could have been prevented by adequate dynamic pricing policies. Deep reinforcement learning (DRL) is a method that can deal with complex problems with high uncertainty. The agent can learn from trying different actions in many different states in order to learn the dynamics of the global network without having direct access to parameters such as costs and price sensitivity. The following main research question is formulated to get new insights that can be used for the global network management:

Can deep reinforcement learning help to find pricing policies for Den Hartogh's global network management?

The model is designed to investigate whether global network management policies can be developed with DRL, instead of being the most accurate representation of the network. The policies cannot be directly translated to the new market corrections for the global network management due to lacking data of external factors such as competitors' prices, economic factors, and more importantly: because the orderbook is not available. Therefore, this project is aimed to develop a proof of concept for the usability of deep reinforcement learning for the network management.

The global network is modelled consisting of 6 global regions and 19 lanes that connect the regions. Every month, potential new contracts arrive in the model, representing the potential demand that could be won. Den Hartogh determines the market correction for each region. As a result, all potential new contracts on a lane receive the same market correction. The market correction results in a higher (for negative MC) or lower (for positive MC) probability of winning a potential new contract. If a contract is won, the expected tank container orders are added to the orderbook that keeps track of the expected number of orders for every lane and month. Tank container orders are satisfied from the regional stock point when the simulation arrives in a new month. The tank container arrives at its destination region after the in-transit leadtime and can be used for new orders.

The performance of DRL is compared with the baseline algorithm, which represents Den Hartogh's network management policies. DRL as well as the baseline are able to manage the global network. Though, the DRL is able to generate 1.52% more profit than the baseline algorithm. With a sensitivity

analysis, various factors of the global network management can be analyzed. From the analysis of model factors such as seasonality fluctuations, profit margin, penalty (i.e. missed opportunity) costs, price elasticity and contract lane demand, it appeared that the price elasticity has a relatively large impact on the performance. Although, the DRL agent was also able to manage the network, even without an accurate estimation of the price elasticity. This implies that the DRL policies are robust since they can deal with quite some uncertainty.

This research investigated also the effect of applying a separate market correction for spot and tender contracts. Spot contracts have a short duration (up to 2 months), while tender contracts have a longer duration. Short-term imbalances (regional surpluses and shortages) can be effectively solved with spot contracts without disturbing the long-term balances. DRL was able to achieve 20% more profit with a separate market correction compared to a single market correction for all contracts.

Based on the results, it can be concluded that DRL is able to find good pricing policies to manage the global network. Although, the policies cannot be directly used in practice due to missing necessary information. Nevertheless, the dynamics of the DRL have shown that frequently adapting the market correction for the new network state results in higher profits. In addition, the market corrections can be utilized to generate additional profit.

Finally, from the experiments with DRL, the following key insights are derived for Den Hartogh:

- Saving the monthly market corrections and saving which market correction was applied and whether the contract has been won, for each potential new contract. Future research to the global network management will benefit if Den Hartogh starts storing all quote data centrally. Moreover, the expected monthly tank container demand of a quote should be registered. Combining the quote data with the applied market correction is expected to provide new insights into the price elasticity of lanes. The relation between the price elasticity and the expected monthly demand can also be analyzed since spot and tender contracts may react differently to price changes.
- More accurate determination of the market correction in relation to the expected network value of a contract. The experiment with a separate MC for spot and tender contracts showed a substantial increase in profit compared to a single MC for all contracts. Therefore, calculating market corrections for contracts more accurately is expected to improve the global network management and achieved profit. Distinguishing by the duration and volume of contracts can also improve the MC revenue. Changing more frequently the MC (if the network state has changed) also improves the MC accuracy, and thus the network management.
- Implementing an orderbook in the global network management in order to keep track of the expected orders from won contracts. The main overall insight developed during this research is that the modelled global network can be balanced quite good with DRL or baseline algorithm, even with an inaccurate estimation of the actual price elasticity. Most plausible reason for this is that the model includes an (accurate) orderbook that provides critical information for developing good pricing policies. Customers should be required and stimulated to provide accurate demand forecasts. This minimizes customers to cherry-pick the cheapest tank operator by using multiple price contracts. Moreover, this demand forecasts gives Den Hartogh an accurate and objective demand forecast, which is required for determining adequate market corrections for new quotes.

Preface

This master thesis is the last phase of the master 'Operations Management & Logistics' at Eindhoven University of Technology. It is also the final requirement of my master's program and thus marks the end of my five-year student life. This amazing period had very different phases that I have great memories of. I would like to use this opportunity to thank some of the people who were part of this wonderful journey.

First of all, I would like to express my gratitude to Willem van Jaarsveld, my first supervisor from the TU/e, for his time and support in the challenges I faced. You provided me the opportunity to develop my programming skills as well as to get familiar with using the deep reinforcement learning framework during the research project and this graduation project. In each of our meetings, the valuable discussions we had, directly improved the quality of this thesis. Furthermore, I would like to thank my second university supervisor, Ahmadreza Marandi for his valuable insights. Your constructive feedback greatly contributed to the quality of the report.

I am thankful for the people who made my research at Den Hartogh Logistics possible. My special thanks go to, my company supervisor, Luke van de Bunt. I learned a lot from the discussions during our meetings. These discussions were valuable not only for this project, but also for my entire career.

Finally, I would like to express my gratitude to my family, girlfriend and friends for their unconditional support during my thesis, and academic career. I feel blessed that I grew up in this supportive environment in which they encourage me to be the best version of myself. Thank you all!

Max van der Wilk,
October 2021

Contents

List of Figures

List of Tables

Abbreviations

ANN: Artifical Neural Network

BL: Baseline

DRL: Deep reinforcement learning

MDP: Markov decision process

MCL: Model-based controlled learning

NN: Neural Network

NMT: Network management team

PNC: Potential new contract

R0: Region0

R1: Region1

R2: Region2

R3: Region3

R4: Region4

R5: Region5

R6: Region6

RL: Reinforcement Learning

SF: Sub-figure

TC: Tank container

Chapter 1

Introduction

Increasing globalization has led to a strong increase in international trade. Many companies relocated to specific areas in the world, mostly for cost efficient reasons. Logistics service suppliers play a key role in facilitating the international trade and managing the global trade imbalances. This research is conducted at the company Den Hartogh Logistics, one of the leading bulk logistics service providers for the chemical, gas, polymer and food industry. Section 1.1 provides background information of the company Den Hartogh Logistics and Section 1.2 describes the global network management of Den Hartogh relevant to this research. Subsequently, Section 1.3 discusses the problem that will be addressed and Section 1.4 presents the relevance and the intention of this research project. Thereafter, Section 1.5 discusses the goal of this research and the research questions that will be investigated. Section 1.6 describes the scope of this research. Lastly, the outline for this report is presented in Section 1.7.

1.1 Den Hartogh Logistics

Den Hartogh Logistics is a family-owned organisation established in The Netherlands in 1920. Den Hartogh has a presence in every region of the world, with premises or offices in 47 locations within 26 countries. The workforce of Den Hartogh consists out of 1,800 people. Furthermore, the fleet contains 20,000 tank containers which are suitable for liquids or gasses, and for multimodal transport. More than 6,100 containers are used to transport dry bulk chemicals and foodstuffs. Den Hartogh also owns 350 trucks with a fixed tanker and 650 trucks that can transport tank containers. This makes Den Hartogh the seventh largest tank operator in the world (ICTO, 2021).

Den Hartogh is divided into four business units in order to separate the different logistic activities. The liquid logistics business unit accounts for more than 50% of the total revenue and is responsible for all intra Europe liquid transports. The global logistics business unit is responsible for all global liquid logistics including interregional orders in Europe and represents 25% of the total revenue. The dry bulk business unit focuses on food related transports and accounts for 20% of the group's revenue. The smallest business unit is gas logistics which accounts for only 5% of the total revenue.

During the more than 100 year of Den Hartogh's existence, the company has only been active worldwide since 2007. Den Hartogh has build up a strong network that includes nowadays more than 377 hubs and 2276 lanes due to several acquisitions. The global network is divided into seven global regions: Region0

(R0), Region1 (R1), Region2 (R2), Region3 (R3), Region4 (R4), Region5 (R5) and Region6 (R6). At least one Den Hartogh office is located in each global region and is responsible for all logistics activities within that specific region. Each general manager (of a region) is also responsible for the availability and occupancy of the tank containers in their region to ensure a minimum service level and sufficient income.

All global orders are served with ISO tank containers due to their suitable size for all different modes of transport (e.g., ship, train, and truck). Besides the trucks that are mostly based in Europe, Den Hartogh global logistics uses other logistic service suppliers to ship their tank containers around the world. Den Hartogh has long-term service contracts on most lanes with shipping companies to provide more security that Den Hartogh can achieve a certain service level for the calculated prices. The price contracts with other logistic service suppliers are usually on a "only pay for what you use", which is a quite common method in this industry. Usually, a tank of Den Hartogh is delivered to a customer, where they charge the tank before it will be transported to its final location. After that, the tank is delivered to a depot where it needs to be cleaned before it can be used for a new order. The journey of regular tank container job is visualized in Figure 1.1

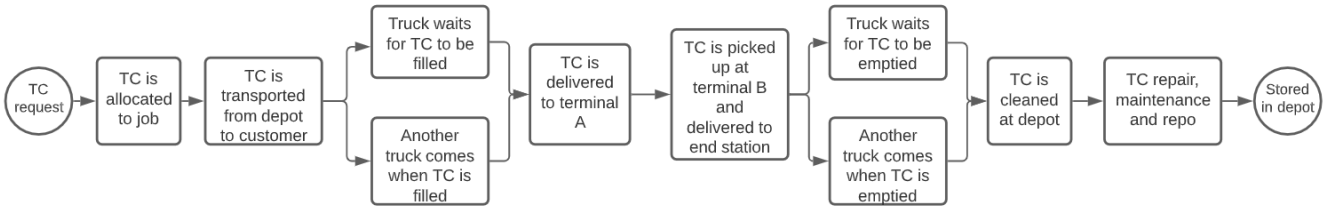

Figure 1.1: Global logistics tank container (TC) journey

1.2 Overview of global network management

The global network management refers to the management of the ISO tank container fleet that is allocated to the global business unit. The network management team motto is: "get the tanks in the right place, at the right time, for the right price". As this motto suggests, the network of Den Hartogh is very flexible to the needs of customers in order to maximize profit. Dynamic pricing can be used to increase the flow of tanks to shortage regions and decrease the flows to surplus regions. The next session discusses the factors that are considered for a new quote. Thereafter, it is discussed how the pricing process works and how it is used to manage the global network.

1.2.1 Pricing model

A customer demand for a single tank container transportation between a specific origin and destination is referred to as a job or order. The pricing model that is used to make a quote for a (potential) customer is shown in Table 1.1. A quote always has a maximum validity. Some customers have regularly orders on the same lane and prefer a longer validity of the quote. These types of quotes are called tenders, because they usually have a minimum duration of one quarter. Another type of quotation is for spot business, this has a shorter validity and the lane is often not (yet) customary for the customer. The following elements are taken into consideration for each quote.

- *Direct order costs*

Table 1.1: Pricing model

Direct order costs	
Equipment costs	
Fixed overhead costs	
Repositioning contribution	
(Repositioning saving)	
Market correction origin	
Market correction destination	
(Expected demurrage revenue)	+
= Default rate	
Quoted network margin	+
= Sales rate	

The direct order costs consist of the cost components directly linked to the execution of the order (e.g. handling, leakage check, the transportation costs, etc). The amount depends upon the requirements of the order, the characteristics of the to be transported commodity, order origin and destination.

- *Equipment costs*
  A fixed amount charged for each day that the tank is allocated to the customer in order to cover the hiring cost for leased-in tank containers and the depreciation, maintenance, and repair cost for owned tank containers.
- *Fixed overhead costs*
  A fixed amount per order to cover the indirect costs that cannot be directly linked to an order (e.g. personnel salary and offices). The amount is similar for each order, regardless of its characteristics.
- *Repositioning contribution and saving*
  Some hubs or regions have structural export or import surpluses. Empty repositionings are unavoidable for these places to balance the empty tank inventory. The expected empty repositioning costs are added to the cost price for orders destined for net import hubs or originating from net export hubs. The expected empty repositioning savings are subtracted from the cost price for orders destined for net export hubs or originating from net import hubs. The amount of contribution or saving for a single order is based on historical data. It is calculated what ratio of tanks were repositioned from or to a hub using the historical repositioning activities in the network. Based on the empty repositioning directions, the expected cost can be estimated for a future empty tank reposition to or from a certain hub. A customer receives a penalty or saving depending on the origin and destination of an order.
- *Market correction origin and destination*
  The market correction is a network steering tool to influence the expected imbalances. Similar to the repositioning contribution and saving, the market correction is also either a penalty or a discount which is based on an order's origin and destination. The main difference is that the repositioning contribution or saving is based on historic flows, whereas the market correction is aimed to anticipate on future flows and imbalances. The market correction for an order consists of the market correction origin and market correction destination. These market corrections can differ per hub, but hubs in the same global region tend to have similar market corrections. When

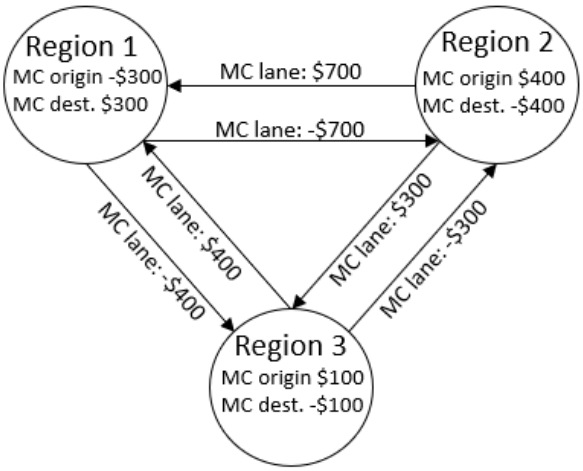

Figure 1.2: Market correction (MC) calculation example

a region is expected to run out of tanks, the market correction origin can be increased to limit export and the regions market correction destination is decreased to stimulate import. The absolute magnitude of the market correction origin and destination is equal, since this tool is only aimed to avoid shortages or surpluses within a location. It is important to remember that the total market correction included in a quote is equal to the sum of the market correction origin of the quote's origin region and the market correction destination of the quote's destination. This sum is equal to the market correction origin of origin region minus the market correction origin of the destination region, because of the mirrored market correction origin and destination of a region. An example of the market correction lanes for a network with three regions and six lanes is shown in Figure 1.2.

$$\begin{aligned} \textit{MC lane} &= \textit{MC origin (of origin region)} + \textit{MC destination (of destination region)} \\ &= \textit{MC origin (of origin region)} - \textit{MC origin (of destination region)} \end{aligned} \tag{1.1}$$

- *Expected demurrage revenue*
  Each quote includes the number of days that a customer may have the tank in their possession for free. Customers should pay a fee for having the tank longer during the exceeding period. This fee is named demurrage revenue and is often in the range of $50 per day. The expected demurrage revenue is based on historical order data of the customer, order origin and destination. The expected demurrage is subtracted from the actual price that a customer has to pay for an order.
- *Quoted network margin*
  The quoted network margin is a prediction of the value of a single order in the network. It can be seen as the gross profit that an order is expected to generate from a network perspective. It depends on the economic situation and other tank container operators what quoted network margin can be achieved by the commercial department. Besides, the expected number of tanks during the agreement can be incorporated in the quoted network margin. Normally, the more tanks a customer is expected to ship, the higher the discount on the quoted network margin.
- *Default rate versus sales rate*
  The default rate indicates the expected costs of a customer order. In general, this is a lower limit

of an order's sales price. In special cases, the commercial department can choose to neglect this lower limit (e.g. to attract new large customers). However, the sales rate is usually higher than the default rate as the commercial departments seeks to maximize the quoted network margin.

1.2.2 Global pricing process

After the quotation team has made a quote for spot or tender business, the customer can accept or decline the offer. If the quote is accepted, the customer is expected to request tanks from Den Hartogh during the agreement. However, customers only have to pay for the tanks they actually ship with Den Hartogh. The customer should request the tank(s) just 1 or 2 weeks before the tank(s) need(s) to be loaded. The global pricing process, visualized in Figure 1.3, is initiated in order to deal with (long-term) demand uncertainty. This monthly reviewing process starts with the global forecast. Each regional manager must provide a forecast for expected exports to all global regions for each of the next nine months. This should be an objective forecast for all tanks that could be sold with a positive margin without taking operational constraints (e.g. sufficient inventory) into consideration. The global network management team uses this forecast to predict regional shortages and surpluses of tank containers.

During the second phase of the pricing process, all regional managers discuss their expectations on the demand flows and regional imbalances. This should result in a consensus on the action required to balance the global network.

Thereafter, the regional manager discusses with the regional team what exact pricing action must be in case of change in the market correction. It is important that the regional team predicts the impact of a certain market correction on the region's import and export. Both flows will be impacted by a change in market correction as they are always opposite. The impact of market correction on the demand is called price elasticity. This is often defined as the measurement of the change in consumption of a product in relation to a change in its price. In this research, the price elasticity concept is used slightly different as the total change in sales rate is not taken into consideration. Therefore, price elasticity is defined as the percentage of change of the hit ratio divided by the absolute change in market correction. The market corrections are defined per hub. Though, since the repositioning of tanks within regions is relatively efficient and incorporated in the price, the market corrections are correlated within regions. Many empty tank containers are repositioned within a region, but interregional repositioning is often too expensive. Therefore, dynamic pricing is used to balance the tank containers over the different global regions.

The implementation of the new market correction is central to the final stage of the global pricing process. The pricing team uses the most recent market corrections for new quotes. Besides the execution of the pricing call outcomes, the effect of the new market corrections are also analyzed. These new insights will be used in the following monthly pricing calls.

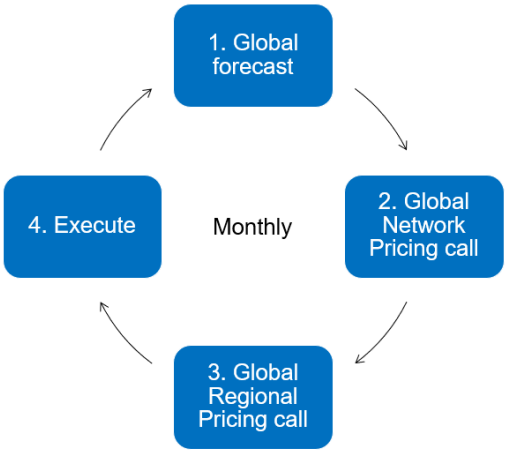

Figure 1.3: Global pricing process

1.3 Problem statement

Den Hartogh's global tank container network is impacted by demand and supply uncertainty. At the supply side, Den Hartogh's network is mainly impacted by the demurrage time. Many customers of the global business unit have tanks in demurrage to deal with their own supply or demand uncertainty and to prevent that their factory runs out of (raw) material inventory. Most global logistics orders have long transport times that can easily be two months. Therefore, customers may prefer to have more inventory which can result in longer demurrage times. Though, the supply uncertainty has small impact on the global network compared to the demand uncertainty.

At the demand side, the tank network is vulnerable to macroeconomic disruptions and customer demand volatility. Demand forecasting is used to anticipate on global imbalances. However, long-term forecasts ($>$ 3 months) are subject to macroeconomic disruptions that could cause unexpected flow imbalances or changes in global trade of chemicals. Short-term forecasts ($<$ 3 months) are impacted by demand volatility of contracted customers. As a result, the export forecasts made by regional managers are often not accurate, even the one-month ahead forecast. This may indicate that either Den Hartogh does not keeping track of how many contracts have already been won or the demand volatility of contracted customers is very high.

Besides the complexity of predicting future imbalances, it is also difficult to determine what actions are needed to anticipate future imbalances. Dynamic pricing with market corrections is used to influence the number of contracts won. Orders from existing (price) contracts are not impacted by new market corrections. The market corrections that were valid in the month in which a contract was won will apply all order from that contract. This, in combination with the effect of other (exogenous) factors (i.e. competitors' prices) makes it hard to measure and quantify the effect of price changes.

This is made even more difficult by the dynamic behavior of the actors (i.e. competitors, customers, shipping companies, etc.) in the global network. As a result, Den Hartogh cannot accurately determine the price elasticity, which is needed to predict the effect of price changes on the volume of won contracts. The unknown price elasticity hinders the speed of decision making in the global network management team for taking adequate actions. Moreover, the actions are primarily based on the view of some process experts. This makes the pricing process labor intensive, vulnerable for organizational changes and not

scalable. Looking to the future, where Den Hartogh's network will grow and become even more complex, the global network management needs to be formalized and even more data-driven.

In summary, Den Hartogh experiences difficulty in forecast both short-term demand as well as long-term demand. Short-term demand is mainly caused by demand volatility of contracted customers, where the long-term demand is mainly caused by the uncertainty in amount of new contract that will be won. These factors together result in high uncertainty in how the global network will evolve. After the regional managers have reached consensus on most plausible scenarios, they try to steer the network by using market corrections. However, it is complex to predict the impact of new market corrections without having a formalized understanding of the price elasticity. The ambiguous effects of the market correction result in more hesitance in the timing and value of the market correction. This ultimately leads to regional surpluses and shortages of tank containers that could have been prevented by adequate dynamic pricing policies.

1.4 DynaPlex project

Den Hartogh has a long and valuable connection with Eindhoven University of Technology. Den Hartogh is one of the forerunners data-driven decision making in logistics. The progressive ambitions of Den Hartogh has resulted in many collaborative projects with master students and researchers from TU/e. Den Hartogh has committed itself to the recently started DynaPlex project of Dr. Willem van Jaarsveld. Dynaplex stands for deep reinforcement learning (DRL) for data-driven logistics that aims to make artificial intelligence accessible for the planning of logistics and transport. This thesis is aimed to analyze the suitability of deep reinforcement learning for the global network management. In other words, this project is a proof of concept for the network management team.

Why (deep) reinforcement learning for global network management?
Den Hartogh's global network management problem can be seen as a sequential decision-making problem. Market corrections are reviewed every month and can be changed if the network expectations have changed. This problem can be well modelled as a Markov decision process as this is a standard framework for addressing the problem of planning and learning under uncertainty. MDP is a typical problem definition for RL with a finite set of states, action set, reward function and transition probability function (Sutton & Barto, 2018). The MDP is the base for the simulated environment the agent aims to learn a policy to effectively select actions in states, that leads to the highest cumulative reward trajectories with the highest cumulative reward. Reinforcement learning (RL) is expected to be a good method, since the agent can learn from trying different actions in many different states in order to learn the dynamics of the global network. For instance, the agent will learn itself the preferred inventory levels and service levels based on the rewards. This might even outperform algorithms or heuristics, especially in highly dynamic environments. The difference between regular RL and DRL is the usage of artificial neural networks (ANN), which is further elaborated in Section 2.1. DRL is especially suitable for high dimensional states, as the complex inputs (state spaces) are mapped to complex outputs (actions) (Serrano, 2019). An additional advantage is that the DRL agent can find a actions for states that have never been seen before, as the neural network finds the best matching state which has been seen before.

1.5 Research goals

This research project was initiated to create a proof of concept for the application of DRL as a decision support for Den Hartogh's global network management. Based on the problem statement and DynaPlex project, the following main research question (RQ) is formulated:

Can deep reinforcement learning help to find pricing policies for Den Hartogh's global network management?

The main RQ aims to provide insights whether DRL has the potential to be used as a decision support tool for the network management team. As the global network of Den Hartogh contains a lot of uncertainty and restrictions, the precise value of the market corrections might not be easily implementable in business. The main insights from this research are focused the relative market correction of regions instead of the absolute values of the market corrections. More information of exogenous factors such as competitors' prices and demand are required in order to determine appropriate market corrections. The behavior of the management policies found with DRL is compared with the network management policies of Den Hartogh in order to interpret differences. The following sub-research questions (SRQ) will be answered in consecutive order to answer the main research question. The relevant section where the question is answered is indicated for every SRQ.

1. ***How can the global network management problem be modelled as a Markov decision process?***

   This SRQ is focused on determining the key features and dynamics of Den Hartogh's global network management and converting this into a Markov decision process (MDP). Moreover, the answer to this question should also elaborate which variables, parameters and assumptions are required to model the global network. In addition, the current network management shall be represented by a baseline algorithm in order to ultimately evaluate the performance of DRL. This SRQ will be discussed in Chapter 4.

2. ***What is the impact of factors that play an important role in the global network management?***

   The global network management is impacted by many endogenous and exogenous factors. The impact of some factors (e.g. price elasticity, penalty costs, demand volatility) is difficult to quantify and therefore to accurately include in the model. A sensitivity analysis is conducted in order to understand the impact of various factors on the network. For the sensitivity analysis and further analysis, various experiments are conducted. DRL agents shall be trained for these experiments. Hyperparameter tuning will necessary to find the best performing DRL agent. The experiments for the sensitivity analysis shall also be used to validate the model. This SRQ will be addressed in Chapter 5.

3. ***How do deep reinforcement learning policies perform in the global network?***

   After the definition phase of Den Hartogh's global network management and the baseline agent, several experiments should be conducted to observe the performance of the network management policies developed with DRL. Besides the experiments from the sensitivity analysis, also other experiments and visualization shall lead to improved understanding of the DRL polices. For instance, the market correction actions, lost sales and inventories can be analyzed for both the baseline and DRL agent. This SRQ will be answered in Chapter 6.

4. ***What insights can be obtained for the global network management?***

   The final SRQ is intended to provide a reflection on the obtained research results. This shall discuss the new insights on the applied DRL framework for Den Hartogh's global network management. In addition, experimenting with the model and the results is expected to provide new insights into the global network management and the current pricing process. Moreover, some insights and recommendations for for the global network management were gathered during various meetings and conversations. These are also used to answer this SQR. As the answer of this question is based on both insights from the model and from personal experiences, it will be addressed in Chapter 7 called: 'conclusions and managerial insights'.

1.6 Scope

The scope of this research is investigate the potential value of using DRL for the global network management. Though, this research is only intended to develop a proof a concept since the using DRL for decision making requires substantial organizational changes and investments. Den Hartogh's large network of many hubs is aggregated to a global level where every hub represents a single region, since the global network team also determines action on regional level. From each region, one flow to every other region is possible. A flow in the network is defined as a one directional stream of tanks that arrive at the destination after the lane's in-transit leadtime. As a result, two flows are possible between two regions, which are called lanes. This research is aimed to balance the global network with the help of dynamic pricing. Although, there are also other ways to manage the network. For instance, the commercial department can actively search for orders on specific lanes. Internal factors (i.e. commercial activities) that can be used to balance the network other than dynamic pricing are beyond the scope of this study. At last, operational constraints are also not included in the scope in order to ease modelling the network. Operational issues are related to problems to can concern local inventory level. For instance, it is not cost efficient to satisfy an order from every local stock point in a region. Therefore, balancing tanks within a region is also important, but not taken into consideration in this research. Moreover, customers can demand a specific type of ISO container (e.g. with safety rack left or right or certain connections), but this is neither included in the model. The relaxation of operational constraints is not expected to undermine this research, because the network management team does not explicitly take operational constraints into account. This would result in a too complex decision-making process. Besides the problem scope, improving or analyzing the model-based controlled learning algorithm that is used in the DRL framework is neither included in the scope. This research is solely focused on developing a proof of concept for the current DRL framework for Den Hartogh's global network management. Ultimately, this research can provide suggestions for improvements of the DRL framework regarding the applicability in a realistic case.

1.7 Outline

This research starts with describing the background of DRL and the used framework, which is presented in Chapter 2. Subsequently, Chapter 3 describes the methodology of the research concerning the model formulation, experiments and analysis. Thereafter, Chapter 4 formulates the model of the global network and the baseline algorithm that is aimed to represent the current network management. Chapter 5

describes the experiments and provides the results of the sensitivity analysis. Furthermore, Chapter 6 presents a detailed analysis of DRL performance and evaluates a potential improvement of the pricing process. Finally, the conclusion and managerial insights are discussed in Chapter 7.

Chapter 2

Background

This chapter provides the theoretical background of deep reinforcement learning and its commonly used modelling framework called Markov decision process. After the introduction of the theoretical concepts about DRL in 2.1 and MDPs in section 2.2, the model-based controlled learning (MCL) algorithm that is developed for the DRL framework is discussed in section 2.3. This literature section provides only a basic theoretical background of DRL and the MCL algorithm because this research is not focused on the algorithms behind DRL. Instead, the DRL framework with the MCL algorithm is considered as a given. This allowed to investigate the global network problem in more detail and how it can be converted to a model.

2.1 Deep reinforcement learning

2.1.1 Reinforcement learning

'Reinforcement Learning is the problem faced by an agent that must learn behavior through trial-and-error interactions with a dynamic environment' (Kaelbling et al., 1996). The power of this method lies within the ability to solve complex and large-scale Markov Decision Processes near optimally whereas the classical dynamic programming methods such as value or policy iteration fail (Gosavi, 2009). Reinforcement learning is an optimization problem with an unknown reward function $\mathcal{R}_a(s, s')$. The agent is not given the reward function, nor is it given any training examples. Instead, the agent has the ability to choose different actions $a(s, s')$ that lead to new state s' and observe their resulting reward $\mathcal{R}_a(s, s')$. The agent will perform an exploratory search by computing different actions, and it will learn from its experience by remembering the rewards obtained from each of these values (Sutton & Barto, 2018). This process where the agent learns from interacting with the environment is visualized in Figure 2.1. Over time, the agent will favor the actions that yield the most reward in a specific state. However, in order to find this optimal value, the agent has to exploit what it has already experienced, and also explore new values in order to discover better solutions over time. For that reason, reinforcement learning is an associative learning process that relies on previous experiences and trial-and-error to generate new knowledge.

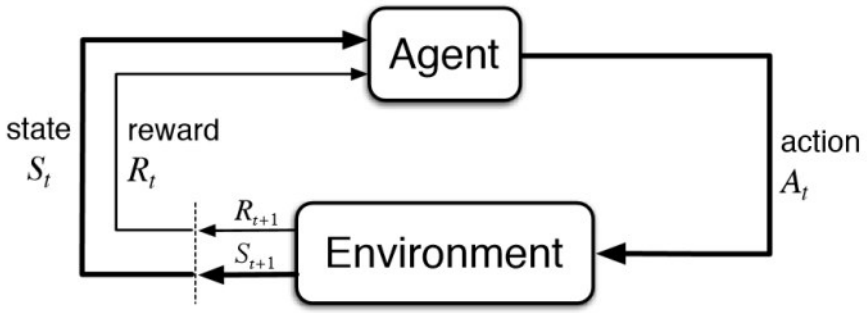

Figure 2.1: Agent environment interaction in an MDP (Sutton & Barto, 2018)

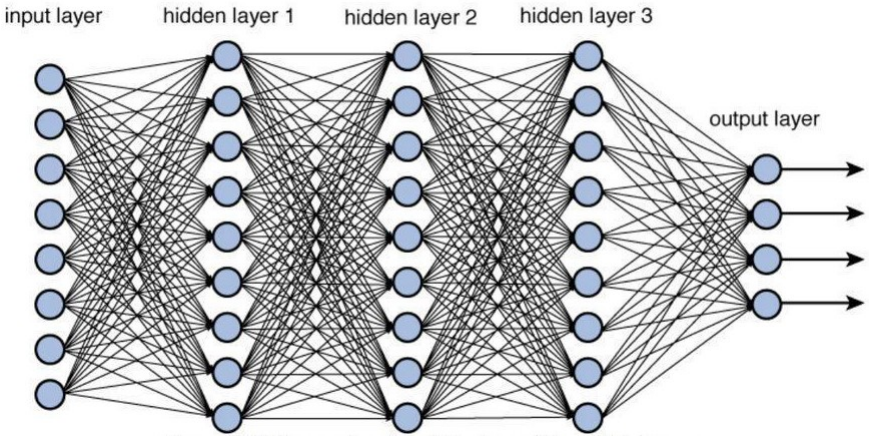

Figure 2.2: Example of artificial neural network

2.1.2 Deep reinforcement learning

Compared to reinforcement learning that provides a general-purpose framework for decision making, Deep Learning provides the general purpose framework for learning (Gijsbrechts et al., 2019). With minimal domain knowledge and given an objective, it is able to learn directly from the raw input to achieve the objective.

Basically, deep reinforcement learning refers to the application of artificial neural networks (ANNs) in reinforcement learning to approximate the value functions, the (stochastic) policy, or a combination of both (Boute et al., 2021). A neural network consists of an input layer of neurons (or nodes), hidden layers of neurons, and a final layer of output neurons. Figure 2.2 shows an example of an ANN. Each connection is associated with a numeric number called weight. The optimal number of neurons should be carefully weighted. The network could have insufficient functional complexity to capture the high dimensional state space when too few neurons are included. On the other hand, too many neurons might result in high training time due to required computation power.

2.2 Markov Decision Processes

A Markov decision process (MDP) is a mathematical framework used in (deep) reinforcement learning for modeling sequential decision problems under uncertainty. The model allows agents to determine

the ideal behavior within a specific environment, in order to maximize the model's ability to achieve a certain state in an environment. This MDP policy is applied to the agent's actions, depending on the environment, and MDP seeks to optimize the steps taken to achieve such a solution. This optimization uses a reward feedback system, where different actions are weighted depending on the expected reward related to each action.

An MDP can be defined as a problem with a finite set of states $\mathcal{S}$, action set $\mathcal{A}$, reward function $\mathcal{R}_a(s, s')$ and transition probability function $P(s, s')$. The transition probability distributions are seldom (fully) known in real life problems. A simulator can be used to model the MDP implicitly by providing samples from the transition distribution. Through this way the transition probability function could be estimated. The model also fulfills the Markov property which means that the next state only depends on the current state and the current action. This has the advantage that not all states in the trajectory have to be considered. The model should be adequately translated to a Markov Decision Process formulation in order to analyze the problem with Reinforcement Learning (RL). The Markov Decision Process could be partially defined with the following terms:

- State space $\mathcal{S}$, which shall include all information that is required for performing within the process.
- Action space $\mathcal{A}$, which shall contain all possible decisions available within the process.
- Transition probability function $\mathbb{P}(s, s')$, which indicates the transition probabilities between two consecutive states.
- Reward function $\mathcal{R}_a(s, s')$, which provides feedback on the received reward from action $a \in \mathcal{A}$ in state $s \in \mathcal{S}$ leading to new state $s' \in \mathcal{S}$.

2.3 Model-based controlled learning algorithm

Van Jaarsveld (2020) designed a model-based controlled learning (MCL) algorithm for the DRL framework. The DRL framework (programmed in c++) is aimed to be suitable for users that only have basic knowledge about DRL. Such an "easy to use" framework aims to contribute in the adoption of DRL in industry. The requirements and assumptions of the MCL algorithm are discussed in this section.

The model-based controlled learning algorithm has demonstrated to be an efficient algorithm for learning good MDP policies that overcome the shortcoming of model-free algorithms when dealing with highly stochastic MDPs (Van Jaarsveld, 2020).

A few policy improvement steps on the initial policy often result in a good policy. Exact policy improvements is computationally too hard for large state spaces $|\mathcal{S}|$. Therefore, a few approximate policy improvements steps will be made that involve Monte Carlo simulation and the neural network classifier training. A single step of the approximate policy improvement involves collecting sample actions from an approximate, simulation-based approximation of an improved policy, and finally using that data to train a neural network. This shall result in an iteration which improves the policy. Van Jaarsveld (2020) can be consulted for more detailed information about the policy iteration method.

The MCL algorithm assumes that the randomness in transitions and costs are caused by uncertain factors or events that affect the trajectory. Moreover, it assumes the randomness exists independently of the actions. This is not uncommon in operations management as the randomness is often related

to exogenous demand or macroeconomic factors which cannot be controlled. Determining which action is preferred may require too many replications, since the sample path costs can have high variance in stochastic operations management. To solve this problem, Van Jaarsveld (2020) introduces a composite random variable that comprises the uncertain factors that may influence the trajectory. This random variable is fixed to objectively compare different actions for the same starting state. The sample path costs no longer contains randomness that is not controlled for by the composite random variable. Thus, various actions can be evaluated under the same randomness. For this reason, the MCL algorithm assumes that the actions have a deterministic impact on the state variables. Otherwise, not all model's randomness is captured in the compound random variable.

In Den Hartogh's global network model, the composite random variable includes the potential new contract demand. This implies that the agent can compare the trajectories (the expected reward) of different actions with the same potential new contract sequence.

More the uncertainty included in the model is expected to lead to more samples of the composite random variable needed to determine the best action. For example, if random demand sequences do not differ much, it will soon become clear which action is the best in a certain state. However, large variance randomness could lead to high computation power requirements to determine good actions for all samples of randomness. The MCL algorithm included two hyperparameters (initial roll-outs and max roll-outs) to reduce the required computation power. After computing the best policy for a number of randomness (i.e. demand sequences) samples, equal to the initial roll-outs, the agent can estimate which actions have not a high potential to be the optimal action in a certain state. In further randomness samples, only the remaining actions will be considered and actions with low probability to be optimal will be excluded. The algorithm terminates if only a single action remains or when the maximum number of roll-outs (samples) have been analyzed. In the latter case, the algorithm chooses the action with the highest probability of being the best action.

Chapter 3

Methodology

The purpose of this chapter is to briefly describe the structure and methodology of this research. First, the model formulation approach is described in Section 3.1. Thereafter, the experiment design is described and finally the how the experimental results will be analyzed.

3.1 Model formulation

The objective of this research is to develop a proof of concept for the use of the DRL framework that incorporates the MCL algorithm for Den Hartogh's global network management. Den Hartogh's problem involves several uncertainties, many of which are caused by exogenous factors. Developing a reliable model capable of covering all situations and uncertainties is difficult and therefore unlikely in this research project. The model must be sufficiently representative for Den Hartogh's network and yet be possible with limited computational effort. Therefore, the network size has to be restricted as well as the types of contracts (demand) and forecast horizon. After the most important concepts of the global network management are modelled, the model will be defined as a Markov decision process. This should clarify how the states, actions and reward function is designed. Thereafter, the initial model state and the baseline algorithm are elaborated. The baseline algorithm represents Den Hartogh's current network management policies and will be used to asses the performance of the DRL policies.

3.2 Experiments

There exists some uncertainty in modelling some parameters or variables because of changing problem dynamics and lacking data. For instance, the profit margin and the demand of tanks is constantly changing over time. Modelling just one scenario carries the risk of questioning the reliability of the model if any of those factors have changed. This reflects the need to examine the effect of fluctuating factors on network policy. Therefore, the different experiments will be designed to assess the impact of one factor on the performance of the model. One base scenario is designed that is best representation of Den Hartogh's global network management. Other scenarios have one deviating factor from the base model in order to observe and evaluate the impact of this factor on the model.

Finally, one major global network management improvement is analyzed. The agent in this experiment uses a separate market correction for spot and tender contracts instead of only one market correction for all contracts. This shall give insights into the effect of valuating contracts more individually. A separate market correction for spot and tender contracts would significantly change Den Hartogh's pricing process, but it is feasible to implement. Investigating the potential benefits and downsides in a simulated environment offers an opportunity to analyze disruptive process changes.

3.3 Analysis

With the various scenarios, a sensitivity analysis can be conducted to either validate the model, asses the relations between factors and analyse the DRL policies. The purpose of the sensitivity analysis is to determine the relative influence of parameters, initial conditions, and alternative assumptions on model output. In each of the comparison runs, all parameters are held constant except for the parameter being examined (Kerr & Goethel, 2014). Figures 3.1 and 3.2 show the conceptual analysis approaches. The MDP scenarios in which a experiment is executed is visualized at the top of the figures. Two DRL agents will be trained on a different MDP scenario. The baseline agent is uses a generic algorithm to determine actions and needs no training. The performance of each agent in a specific scenario is indicated with PX with X ranging from 1 to 4. in the Figure. The differences between two performances is analyzed for different scenario-agent combination. The comparison between two agents is indicated with A and B.

The sensitivity of various factors is analyzed in Chapter 5. In Figure 3.1 can be observed that the performance differences are analyzed between on the one hand DRL agent 0 and the baseline agent in MDP 0 (yellow P1.A&B), and on the other hand DRL agent 1 and baseline agent in MDP 1 (purple P4.A&B). This provides insights into the strengths and weaknesses of the DRL agent (e.g. in what scenario does it outperform the baseline agent). The baseline is designed as a generic algorithm that is specifically trained for one specific situation. The performance of the baseline agent can be compared to the DRL agent's performance. This may provide new insights into the strengths and weaknesses of DRL in a specific scenario. In addition, the performance of DRL agent 0 in MDP 0 can be compared with the performance of DRL agent 1 in MDP 1 (orange P2.A&B). This may indicate which scenario results in higher profit, which can also be used to validate the model. Finally, the performance of the baseline agent can also be compared between MDP 0 and MDP 1 (red P3.A&B). This comparison can also be used to assess the baseline algorithm's validity.

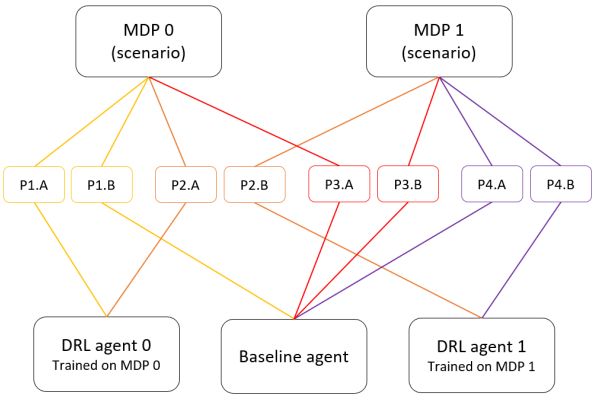

Figure 3.1: Conceptual sensitivity analysis example

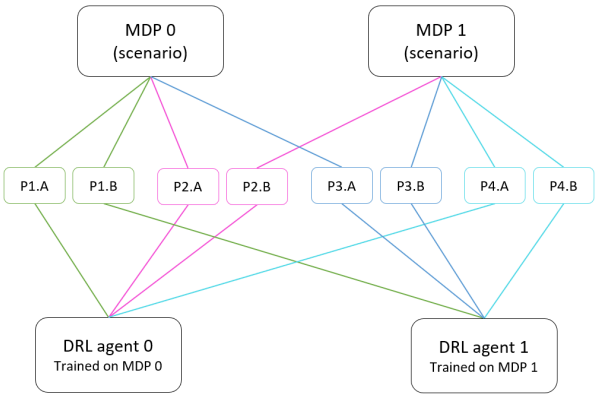

Figure 3.2: Conceptual cross comparison analysis example

In the Chapter 6, a cross comparison is made between the DRL agents. A conceptual cross comparison example is illustrated in Figure 3.2. The performance is evaluated of DRL agent 0 and DRL agent 1 in both MDP 0 (green P1.A&B). The cross comparison is also made the other way around, DRL agent 0 and DRL agent 1 in MDP 1 (turquoise P4.A&B) The cross comparison provide insights into the robustness of a DRL agent for another scenario. Models always have the risk of not representing the problem exactly. It would therefore be useful to know which factor impact the performance of the DRL agents. For example, if two scenarios are designed that have a different price elasticity, the performance of both agents is compared for the same price elasticity. This may indicate whether the performance is expected to be higher for an overestimation or underestimation of the price elasticity. Cross comparing agents in different scenarios is not often used in the analysis of DRL agents, which makes this research unique. Moreover, the performance of a agent is also compared between two different scenarios (pink P2.A&B and blue P3.A&B). In robustness of a DRL policy can be assessed in these experiments.

After the cross comparison analysis in Chapter 6, the network management of the DRL agent are analyzed in more detail. Various statistics are compared to understand which market corrections are applied in each region and how this affects the network performance. Moreover, a potential network management improvement is investigated. In this experiment is the effect of a separate market correction for spot and tender contracts analyzed.

Chapter 4

Model formulation

This chapter is intended to convert Den Hartogh's global network (management) into a model. The goal of the model is not to develop the most accurate representation of Den Hartogh's global network, but to develop a practical model that includes the most important network elements and dynamics. The model must be able to demonstrate the advantages and disadvantages of DRL for the global network management. First, in Section 4.1 the conceptual model is visualized and briefly discussed to introduce the main concepts. This would make the detailed model formalization easier to understand, which is provided in Section 4.2. In the problem formalization are the concepts individually analyzed in order to find the input parameters values and probability distributions. Thereafter, the model is defined in terms of a Markov decision process in Section 4.3. Next, Section 4.4 describes the most important features of the network and also its initial state. This chapter ends with the formulation of the baseline algorithm that aims to represent Den Hartogh's network management in Section 4.5.

4.1 Problem conceptualization

This section presents the main concepts and their mutual relationships. This is visualized with a model example containing three regions with six lanes. The network is modelled as a discrete-time Markov model that evolves every month. The key concepts of the network is addressed below:

- *Network*
  A network consists of regions and lanes. A region is a place where empty tanks are stored in inventory until needed for new orders. On the lanes are tanks moving from their origin region to their destination region. We use an example with 3 regions and 6 lanes to explain the dynamics of the network. Actual experiments are run on a network with 6 regions and 19 lanes.

- *Inventory*
  Each region has an inventory of empty tank containers that are used to satisfy customer orders from that region.

- *In-transit*
  In-transit (IT) represents the tanks containers that are shipped between regions. The model has a monthly review period which implies that all orders are shipped at the same time each month.

The in-transit can be seen as a pipeline with a length equal to a lane's in-transit leadtime. Tanks arrive at their destination region after the in-transit leadtime where they are stored in inventory until needed for satisfying a new order

- *Orderbook*
  In the orderbook (OB) are all orders stored and in which month these orders must be execute (i.e. when the tanks must depart). Each lane has it own orderbook for all months in the forecast horizon.

- *Potential new contracts*
  Events are generated every month. An event represents an incoming potential new contract (PNC). Den Hartogh gives a quote for each potential new contract. If this price is accepted, the potential new contract is converted to the orderbook where the orders from that contract are stored.

- *Market correction*
  The market correction is the dynamic component in the price that is given to a potential new contract. Each month, a new market correction (MC) origin is determined for each region. The market correction origin applies to order that depart from that region, and the market correction destination applies to orders that arrive in that region. A region's market correction destination equals ($-1\times$ MC origin). The market correction that is valid on a particular lane consists of its MC origin and MC destination. The higher the lane's market correction, the lower the probability that a potential new contract is won this lane.

Figure 4.1 visualizes the dependencies of the of concepts. For example, the inventory of a region in $t+1$ is positively dependent on the region's inventory in t, positively impacted by the in-transit tanks to that region in and negatively impacted by the outgoing orders from the orderbook. The market correction can both negatively or positively impact the relation between potential new contracts and the orderbook. This depends on the the value of the lane's market correction and the price elasticity. The concepts that are inside the large box can be observed in the network and are explicitly modelled. The 'random demand' is an exogenous factor and 'action' is the control mechanism of the agent.

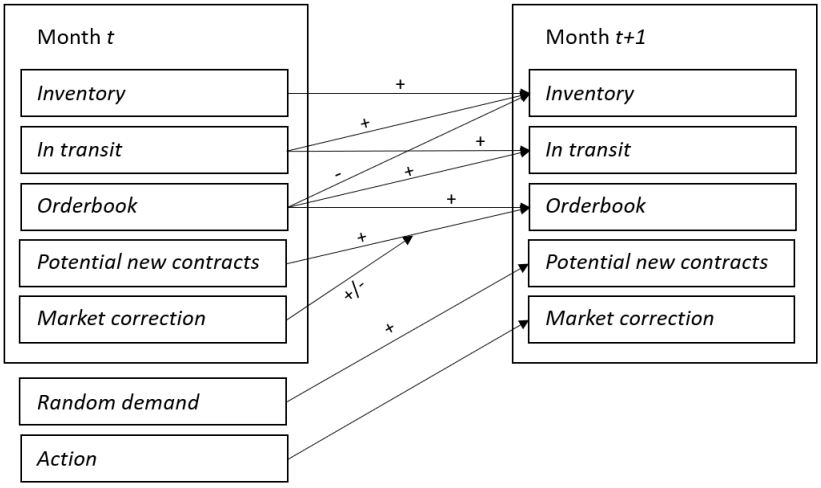

Figure 4.1: Conceptual relations

Now, when the most important model components are briefly discussed, the numerical example in Figure 4.2 can be explained. However, the reader is not expected to fully understand this example as the rest

of the chapter will elaborate on the modelling decision in more detail. The conceptual model consists of six sub-figures (SF) that show the modelling steps. In each sub-figure the changed variables and values are shown in bold. The modelling steps are discussed below:

1. In this example are 966 events generated that result in 966 potential new contracts. In the example, a potential new contract can have a duration of 1 or 2 months. A one-month contract results in one tank container order. A two-month contract results in two tank container orders, one in each month of the contract duration. For the experiments, a contract duration of up to 8 months is considered.

2. Thereafter, the market correction origin is determined for each region. The new market corrections are displayed in SF2.

3. Next, the MC lane can be calculated by combining its MC origin and MC destination. The MC lane and the price elasticity influence how many potential new contracts are won. The third figure shows the number of won new contract (WNC) from the potential new contracts.

4. After a contract is won, the first tank can be ordered after one month. Therefore, the won new contracts at time t are added to the orderbook from $t + 1$. This is visualized in the fourth figure. The first element of the orderbook is not influenced by the current won contracts. The orderbook for the first lane in month $t + 1$ had already 12 tanks at the beginning of the month. From the potential new contracts are 10 contracts won with a one-month duration and 5 with a two-months duration. As a result, the 10 and 5 tanks are added to the orderbook for $t + 1$ and 5 tanks are added for $t + 2$.

5. After this, the in-transit tanks (that were shipped their in-transit leadtime periods ago) arrive at their destination. The in-transit items shift one month forward for lanes that have a longer than one month in-transit leadtime. This is shown in the fifth sub-figure for the lanes between region 2 and 3.

6. In the sixth and last sub-figure are the first items of the orderbook shipped from the origin inventory. These orders are now in-transit to their destination. This concludes the month cycle that is repeated every month.

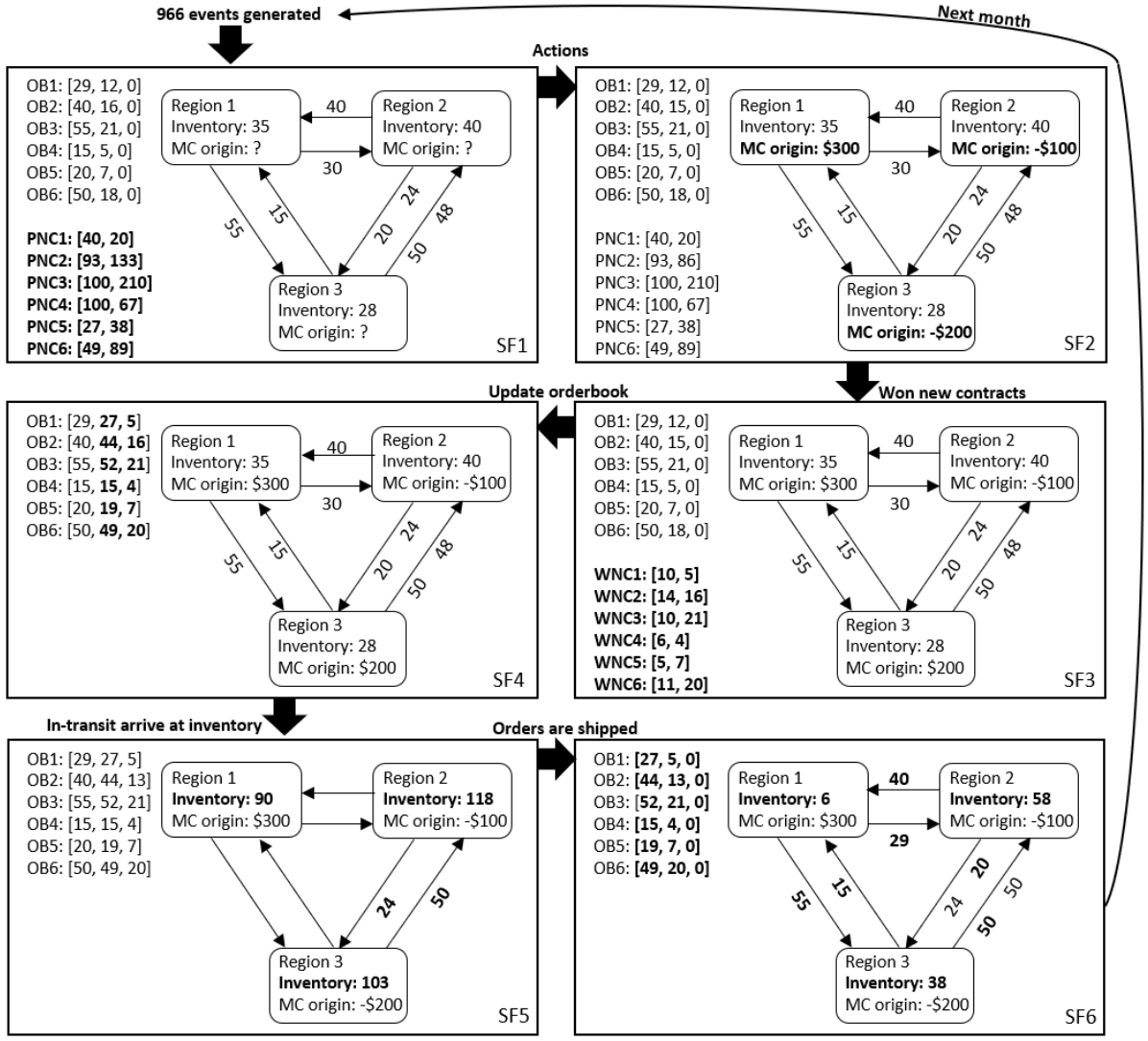

Figure 4.2: Conceptual model and steps visualized with 6 sub-figures (SB)

4.2 Problem formalization

The previous section introduced the main concepts and network dynamics with an example. This section explains the modelling decisions, and gives the actual network structure and dynamics. Moreover, the available data is analyzed to determine the parameter values and variable distributions that are used in the model.

4.2.1 Global network structure

The layout and dynamics of the network are critical elements in the analysis of Den Hartogh's global network management. This subsection is aimed to determine the lay-out of the modelled network. Currently, global network includes approximately 377 hubs which are located in one of the seven global regions. The level of detail for this research is regional, which means that only the seven regions are distinguished in the global network management. This is briefly discussed in the research scope (Section 1.6), but main task of the global network management team is to ensure that every region has enough tanks, and the intraregional tank distribution is managed by the regional team. Therefore, the global network consists out of 7 regions (i.e. nodes) which are fully interconnected, which means that there are 42 (one-directional) interregional lanes and 7 intraregional lanes. The intraregional trade can be a large part of a region's orders. For example, 28% and 40% of the orders from the Region2 and Region6 are intraregional. Therefore, it would be meaningful to include the intraregional flows to evaluate Den Hartogh's performance and required (empty) tank inventory. Moreover, the intraregional flows are critical to balance the local empty tank inventories within a region. Intraregional flows consume tanks that cannot be used for interregional demand while the tanks are in-transit. However, the market correction cannot be used to steer these flows as the sum of the market correction origin and destination is zero. Intraregional flows have also a minor impact on Den Hartogh's global network management as these flows cannot be (directly) utilized to balance the global network. Therefore, Den Hartogh has separate inventory control policies for these intraregional flows. As a result of the exclusion of intraregional flows, the number of tank containers in the model has to be adjusted. The model incorporates less demand which leads to lower inventory requirements (to achieve the same service level). The required number of tank containers to achieve a similar tank utilization in the model is addressed in Section 4.4.

Table 4.1 shows the number of interregional tanks shipped with a loading date between 1st of July 2020 and 30th June 2021. The number of lanes grows quadratically as the number of regions increases for interconnected networks. The network size has an large impact on the required computation power to train a DRL agent. The increased need for computation power is the result of increased action and state spaces. The state space even grows exponentially for the number of regions. Therefore, the global network must be critically evaluated to include only the most important regions and lanes.

Den Hartogh's main concern about decreasing the network size is that the network should still be recognizable to ensure employees' confidence in the model. Regions are very different based on their export and import volume. Region0 (R0) exported only 1302 tanks, where the Region1 exported 18118 tank containers in the same period. R0 exports amounted to only 2% of the total number of global orders. The ability to manage the global network is likely to be negligible because the import from R0 is only 9% of R5's total import and for other regions even less. Therefore, R0 and all lanes from and to R0 are excluded from the modelled network to increase the feasibility for efficient learning. R0's exclusion

Table 4.1: Interregional tank flows (July 2020 - June 2021)

from \to	**R0**	**R1**	**R2**	**R3**	**R4**	**R5**	**R6**	**Total export**
R0		296	12	92	92	384	426	1302
R1	1184		4020	7486	2524	1886	1018	18118
R2	130	2716		760	488	302	1228	5634
R3	98	5492	1024		4056	1144	702	12526
R4	64	5462	414	3064		126	3544	12674
R5	270	2302	406	524	710		46	4358
R6	2	1080	48	566	3204	314		5214
Total import	1748	17348	5924	12492	11074	4176	6964	59726

leads to a remaining network with 6 regions and 30 lanes.

Moreover, low-volume lanes have minor impact on region's import and export. Low-volume lanes have limited effectiveness to solve global imbalances. For instance, when a lane represents 5% of a region's export and the flow of this lane is significantly increased (e.g. +25%), the absolute effect on the region's imbalance is barely noticeable. The computational efficiency (i.e. controlling the network versus required computation power) would increase if only lanes are included that have a noticeable impact on a region's balance. Table 4.2 shows the distribution of export volume for all outgoing lanes per region. A network with less than 20 lanes is expected to be computationally feasible and still a good representation of the global network. Excluding all lanes that are less than 5% of a regions export would only lead to a elimination of 5 lanes (from the 30 remaining lanes). A 10% export cut-off value results in a exclusion of 11 lanes, which implies that the network has 19 lanes left (only the lanes with 10% or higher out of Table 4.2). However, this approach only considers the lane's impact on its origin region. Considering the lane's impact on its destination region would result in 3 lanes that have more than 10% of the import for their destination region. Although, the 10% export cut-off rule is used to reduce the network size to below the 20 lanes. It is unavoidable to negatively impact the network's representation when decreasing the network size. Moreover, the model is not aimed to be the best representation and accuracy, but to investigate whether global network management policies can be developed with DRL.

In summary, Den Hartogh's global network contains 7 global regions. Only the six largest regions are included in the model. In addition, intraregional flows are excluded because Den Hartogh has separate inventory control policies to manage intraregional flows. Moreover, the effect of this exclusion is compensated by reducing the number of tanks in the model that result in comparable tank utilization. Only the interregional flows that represent more than 10% of a regions export are kept in the model, which leads to a network with 6 regions with 19 lanes.

4.2.2 Market corrections

This (sub)section addresses the different market correction options for all regions. As the problem statement already described, the network management team (NMT) determines the market corrections (MCs) that are used for quotes. Figure 4.3 displays all MCs origin for all hubs (excluding in Region0) in the global network. The MCs in the figure tend to be correlated for hubs in the same region. Small local differences can emerge from chemical clusters that can imbalance a hub. The model's scope does not go

Table 4.2: Lane importance for regional export (July 2020 - June 2021)

from \to	R0	R1	R2	R3	R4	R5	R6
R0		23%	1%	7%	7%	29%	33%
R1	7%		21%	42%	14%	11%	6%
R2	2%	50%		13%	8%	5%	21%
R3	1%	44%	8%		32%	10%	6%
R4	1%	43%	3%	24%		1%	28%
R5	6%	54%	9%	12%	17%		1%
R6	0%	21%	1%	11%	61%	6%	

further than one MC for the whole region. This is not expected to have a noticeable effect because most hubs in a region have the same MC. An additional advantage of the regional MCs is that the action space is considerably smaller.

Den Hartogh usually uses only MC intervals of $100. They do not prefer to use small MC intervals because the impact of the change in the MC is small. This makes it more difficult accurately estimate of the change in demand due to the new MC. From Figure 4.3 can be observed that extreme MCs beyond $500 or -$500 are rare. Therefore, the MCs used in the model are limited to intervals of $100 en are bounded to -$500 and $500. This leads to 11 possible MCs per region which is realistic and computational feasible, mainly because of the reduced action space.

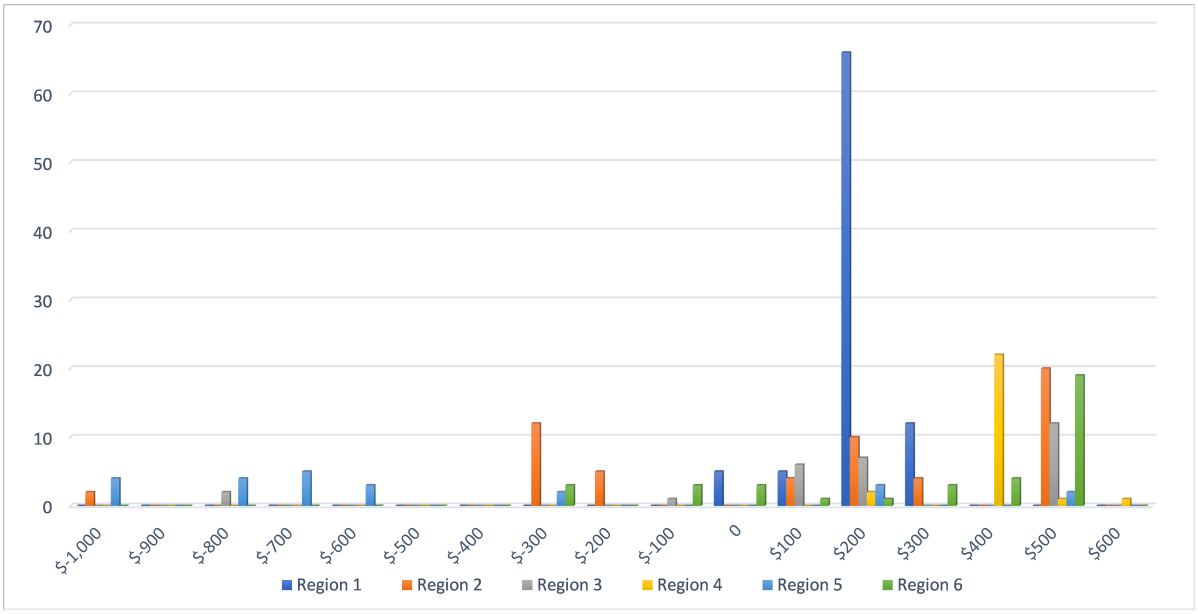

Figure 4.3: Market correction origin for all global hubs (August 2021)

4.2.3 Price elasticity

The price elasticity is neither explicitly visualized in the conceptual model nor in the conceptual relations. Although, the price elasticity effects the relationship between the market correction and the probability of winning potential new contracts. This subsection discusses the price elasticity and how it is incorporated in the model. The main discussions in the NMT are usually not about which flows should be increased

Figure 4.4: Flowchart potential new contract (PNC) to first order

or decreased, but what effect a specific MC will have on the network. Den Hartogh experiences difficulty in quantifying the price elasticity because the chemical market is highly dynamic and impacted by many external sources. Price elasticity is here defined as the percentage change in hit ratio divided by the absolute change in market correction. The hit ratio indicates the number of quotes won compared to the total number of quotes made. The price elasticity distribution is required to calculate the effect of market corrections on the flow of tanks. Unfortunately, Den Hartogh has no data available of historic MCs and number of contracts won in that period. Therefore, the price elasticity needs to be estimated with the help of expert knowledge from the NMT. The price elasticity distribution(s) is discussed in the experiment design (Section 5.2). Based on historical quote (i.e. potential new contract) acceptance data can be concluded that the hit ratio for quotes is quite comparable all over the world. The basis hit ratio (for MC = \$0) is not expected to have a major impact on the model's dynamics and validity. Therefore, the hit ratio is modelled as 0.20 for all lanes if the MC is \$0, because 0.20 is also approximately Den Hartogh's average hit ratio. The price elasticity distribution indicates the adjusted hit ratio for MCs other than \$0.

4.2.4 Unit of time

The unit of time used in the model is month. This time unit is chosen because pricing actions are only taken once per month. Having a weekly time unit would result in higher accuracy in measuring lost sales and stock levels. Though, it makes the model more complex. In addition, the state space would also increase significantly if the in-transits and orderbook have to keep track of weekly shipments. In order to reduce the required computation power, the aggregation level of time is kept on monthly for all variables.

4.2.5 Contract actualization time

The time between a potential new contract is won and the first month where the customer orders tanks is called actualization time. Spot contracts usually have a short actualization time while large tenders prefer to assess potential new contracts a few months in advance. The actualization time is assumed to be 1 month for all won potential new contracts. This means that customers order tanks in month $t + 2$, if Den Hartogh's price for the potential new contracts was accepted at start of month $t + 1$. This is visualized in Figure 4.4.

The model also assumes that customers immediately accept or decline Den Hartogh's price for a potential new contract. For instance, the potential new contract has arrived before month $t + 1$, Den Hartogh determines the MCs at the start of month $t + 1$ and the customer immediately accepts of declines the (MC) price for the potential new contract. If the potential new contract is accepted, it is converted to the orderbook and the first tank container will be ordered in month $t + 2$.

4.2.6 Demand distribution

This subsection is describes how the demand of potential new contracts is modelled. The section Problem conceptualization (4.1) describes that the potential new contracts are generated from random demand that is exogenous. In addition, a potential new contract has two attributes: lane and duration. This section addresses how the lane and duration (probability) distributions are found. Thereafter, the two distributions are combined to a contract probability distribution. Moreover, the logic is explained how potential new contracts are be converted to orders that are stored in the orderbook.

Background of Den Hartogh's demand

The transported chemicals are often part of long supply chains that is subject to disruptions. For instance, the corona pandemic initially disrupted the demand side of the supply chain (Nikolopoulos et al., 2021). Later in the pandemic, mainly the supply side of the chain was disrupted, causing bullwhip effects and shipping space issues that have never seen before in this industry. As a result, Den Hartogh experienced increased uncertainty in their tank container demand. One of the major uncertainties in Den Hartogh's global network management is the volume of potential new contracts per lane. Potential new contracts (i.e. potential demand) can be described as the total demand that Den Hartogh has if they would win every quote they make. Usually, chemical companies ask Den Hartogh to make a quote (i.e. price for a potential new contract) for a specific type of orders. If a customer accepts the quote, it accepts the price for that specific type of orders and promises to request those tanks at Den Hartogh. A type of order is characterized in the model by one certain origin and destination and the number of months where the tanks are ordered. A quote is indicated as a potential new contract in the model.

In practice, Den Hartogh distinguishes potential new contracts for spot and tender business. The criteria for tender business is not necessarily the validity period of the quote, but also whether the business is recurring. Tender quotes often include many different potential new contracts, which can be won individually. So, the lanes for which Den Hartogh offers the lowest price will be granted to Den Hartogh, while other potential new contracts in the tender can be granted to competitors. Spot business is usually not recurring, more ad hoc and includes usually less lanes than tender quotes. The difference between spot and tender quotes resulted in a different quotation process for spot and tender business. The tender quotes are not centrally stored, which means that they cannot be included in the quotation analysis. Moreover, the spot quotation process in not standardized for all global regions which negatively impacts the quotation data reliability. Therefore, the potential demand probability distributions need to be estimated without historical quotation data.

Instead, the historical order data is used estimate the potential new contracts distribution. The order data represent all Den Hartogh's satisfied demand. Two steps must be taken to convert the order data to potential new contracts demand. First, the historical order data can be combined with the hit ratio to estimate the potential order demand. A hit ratio of 0.20 leads to a 5 times higher potential order demand than the order data. This step is visualized in the first calculation step of Figure 4.5 and explained in detail under the heading "potential orders distribution". Second, the potential order demand needs to be converted to potential new contract demand. The expected number of orders per contracts depends on the contract duration, thus the contract duration distribution is used in calculation. This calculation step is visualized in the second calculation in Figure 4.5, and explained further under the heading "contract duration distribution". The third calculation step in Figure 4.5 represents the final step to combine the potential order distribution with the contract duration distribution to determine the potential new

contract distribution. To potential new contract probability distribution is given by dividing the potential new contract of a lane by the total number of global contracts. This step is described in "potential new contract distribution"

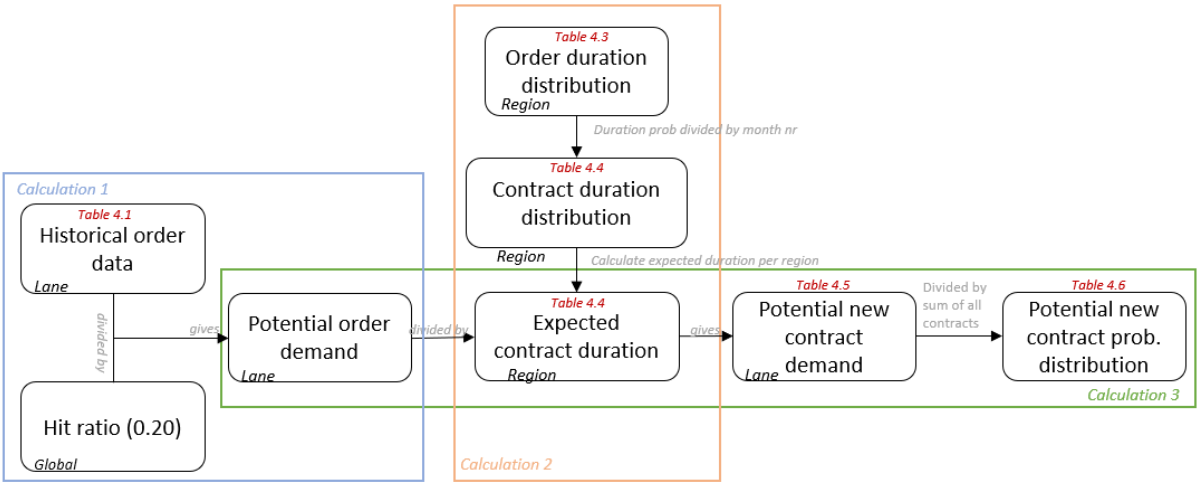

Figure 4.5: Calculation flowchart to determine potential new contract probability distribution (shown in Table 4.4)

Potential order demand distribution

The historical order data from July 2020 to June 2021 is used to estimate the potential new contract distribution because Den Hartogh has not sufficient quote data available. First, the historical order data that is shown in 4.1 can be divided by the hit ratio to determine what potential order volume. The hit ratio indicates the average probability that a quote is won. With an average hit ratio of 0.20, the potential orders are 5 times higher than the actual number of shipped orders. In conclusion, the potential order demand for each lane is equal to the lane's historical order data divided by the global hit ratio that equals 0.20.

Contract duration distribution

In practice, the number of tanks is not necessarily dependent on the duration, but the tank volume distribution cannot be analyzed due to lacking data. Den Hartogh does not have standardized data of the number of tank container order per contract. Therefore, the tank container order volume per contract is assumed to linearly increase as the contract duration increases. For example, the expected number of tanks from a 4 month contract is 4 times higher than an 1 month contract and 2 times higher than a 2 month contract duration. As a result, contracts with a longer duration have more impact on the network because longer contract lengths lead to more tank container orders. The contract duration distribution impacts the modelled network uncertainty. For instance, a network consisting of a few very large contracts is exposed to more uncertainty than a network consisting of many small contracts. To ease the agent's learning, the monthly demand per contract is assumed to be constant. This implies that demand volatility (i.e. variance in the number monthly orders per contract) is not incorporated in the model.

The quote data set has been analyzed in order to estimate the contract duration distribution. However, too many tenders were missing in the quotation data to estimate the distribution of contract lengths. Therefore, the contract duration probability distribution is estimated with the help of a process expert from the network management team. Although the contract duration is theoretically unbounded, the

Table 4.3: Percentage of orders originating from a contract with specific duration

Months	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
R1	12.5%	12.5%	12.5%	12.5%	12.5%	12.5%	12.5%	12.5%
R2	25%	25%	8.3%	8.3%	8.3%	8.3%	8.3%	8.3%
R3	12.5%	12.5%	12.5%	12.5%	12.5%	12.5%	12.5%	12.5%
R4	75.0%	3.6%	3.6%	3.6%	3.6%	3.6%	3.6%	3.6%
R5	25%	25%	8.3%	8.3%	8.3%	8.3%	8.3%	8.3%
R6	75.0%	3.6%	3.6%	3.6%	3.6%	3.6%	3.6%	3.6%

Table 4.4: Contract duration probability

Months	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**E[D] (months)**
R1	36.8%	18.4%	12.3%	9.2%	7.4%	6.1%	5.3%	4.6%	2.94
R2	52.5%	26.2%	5.8%	4.4%	3.5%	2.9%	2.5%	2.2%	2.10
R3	36.8%	18.4%	12.3%	9.2%	7.4%	6.1%	5.3%	4.6%	2.94
R4	92.4%	2.2%	1.5%	1.1%	0.9%	0.7%	0.6%	0.6%	1.23
R5	52.5%	26.2%	5.8%	4.4%	3.5%	2.9%	2.5%	2.2%	2.10
R6	92.4%	2.2%	1.5%	1.1%	0.9%	0.7%	0.6%	0.6%	1.23

modelled contract duration is bounded from 1 to 8 months with an 1 month interval because of two reasons. First, Den Hartogh uses forecast horizon of 9 months in for global network management. If the maximum duration is assumed to equal 8 months with a contract actualization time of 1 month, the horizon in the model equals also 9 month forecast. This ensures that the model is an appropriate representation of the global network management. Second, most contracts have a duration that is shorter than a year and limiting the contract duration reduces the forecast horizon and thus also the state space. The state space decreases because the number of possible contract types is lower and the orderbook is only filled up to 9 months ahead. In addition, the potential new contracts have 8 different duration possibilities which also lead to a state space reduction.

The duration distribution for tank container orders is estimated with the help of a process expert of the network management team. The distribution of orders that are related to a specific contract length is provided in Table 4.3. However, the duration distribution for contracts is required for the model. An tank container order that is related to a contract with a duration of 8 months is eight times more likely to notice than an order related to an 1 month contract if both contracts have more equal probability to occur. This is because eight times more orders result from an 8 month contract than from an 1 month contract. For instance, 125 orders are expected to come from an 8 month contract if 1,000 orders are observed from R1. 125 orders are related to 15.625 (125/8) contracts for an monthly order demand of 1 tank container. The total number of contracts related to 1,000 orders in R1 is 340 contracts. So, the probability of an 8 month contract equals 0.046 (15.625/340). The contract duration distribution is shown in Table 4.4. The duration distribution is the same for all outgoing lanes from a region.

Potential new contract (probability) distribution

The potential new contract distribution is build on the potential order demand distribution and contract duration probability distribution. Contracts have two varying attributes; lane and duration. The number of different contracts is equal to number of lanes multiplied by number of duration possibilities

Table 4.5: Estimated potential new contracts (tank order / expected contract volume) (July 2020 - June 2021)

from/ to	R1	R2	R3	R4	R5	R6
R1		6828.6	12716.2	4287.4	3203.6	
R2	6470.9		1810.5			2925.6
R3	9329.0			6889.8	1960.2	
R4	22158.0		12430.0			14377.2
R5	5484.4		1248.4	1691.6		
R6	4381.4		2296.2	12997.9		

of contracts. This results in (19×8) 152 types of contracts. The order data from July 2021 until June 2021 is used to calculate the contract lane probability distribution. Although, the order data shows the distribution for orders while the contract distribution is needed. The number of orders transported per is divided by the expected number of orders per contract (on that lane) in order to determine the expected number of contracts per lane. The expected number of orders per contract only depends on the duration of a contract due to the assumption of a fixed monthly order volume for all contracts. Table 4.5 shows the estimated number of historic potential new contracts for all lanes that are included in the model. In order to convert the potential new contract distribution to a potential new contract probability distribution, the potential new contract for a lane is divided by the total number of contracts. This gives the lane probability distribution for the 19 lanes. Multiplying the contract's lane probability with the contract's duration probability from Table 4.4 gives the probability of a contract on lane l with duration d. This contract probability distribution is provided in Table A.1 in the Appendix.

Based on the the total number of potential new contracts per year, the average number potential new contracts is 5562 (66743/12). This number will also be used in the simulation to generate on average 5562 contracts per month.

4.2.7 In-transit leadtimes

The average time that an order on lane l takes is called transport time. This time includes all activities from the loading date until the return-to-pier date (i.e. the tank has returned to a depot where it can be allocated to a new order). The order data is used to determine the average time between loading date and return-to-pier date for orders with a loading date between January 2019 until December 2020 until and a return-to-pier date before June 2021. The average number of days that a tank an order takes is rounded to the nearest number of months. The rounded transport times are shown in Table 4.6. Rounding to an integer in months is needed because the unit of time is months in the model. This means that inventories are checked and orders are shipped only once per month. Therefore, the transport times are rounded to months and constant in order reduce the model's complexity.

4.2.8 Key assumptions summary

- Den Hartogh's global network is represented by a network that includes the 6 largest regions and 19 lanes that consists more than 10% of a region's export.

Table 4.6: In-transit leadtimes

from/ to	R1	R2	R3	R4	R5	R6
R1		2	2	2	3	
R2	2		3			2
R3	2			2	2	
R4	2		2			1
R5	2		2	2		
R6	2		3	1		

- Market correction for a region's origin or destination are within the interval [-500$, 500$] and a market correction per lane which is the sum of a lane's market correction origin region plus the destination region market correction is within the interval [-1000$, 1000$].
- The hit ratio for potential new contracts is 0.20 for all lanes if the market correction is $0.
- The relationship between the market correction lane and hit ratio is described as the price elasticity.
- The actualization time of a contract is one month. This means that the period between winning a contract and receiving the first order is 1 month.
- The potential new contract (demand) lane probabilities are estimated with the actual tanks shipped divided by the hit ratio.
- Contract volume depends only on the contract duration with a linear relationship. Thus, the monthly volume is equal for all contracts.
- Minimum contract duration is 1 month and maximum contract duration is 8 months with an interval of 1 month.
- Contract duration distribution is the same for all export lanes in a region. Thus, the duration distribution is based on a lane's origin region.
- The forecast horizon is equal to the sum of maximum contract length and actualization time, which is 9 months.
- The in-transit leadtimes are assumed to be constant.

4.3 Markov Decision Process formulation

Now, when the model is described, it needs to be adequately translated into a Markov Decision Process (MDP) formulation, so it can be used for DRL. MDP is a well-suited framework for the definition of the global network management problem, as it is aimed at structural decision making within a stochastic environment. Markov Decision Process shall be defined with the following terms:

- Action space $\mathcal{A}$, which shall cover all possible decisions available within the process.
- State space $\mathcal{S}$, which shall contain all the information required for performing a decision within the process.

- Reward function $\mathcal{R}_a(s, s')$, providing the feedback on the performance of the taken action $a \in \mathcal{A}$ in state $s \in \mathcal{S}$ leading to the new state $s' \in \mathcal{S}$
- Transition probability function $\mathbb{P}(s, s'|a)$, is the probability of moving from state $s \in S$ to state $s' \in S$ when the agents perform actions given by the vector a, respectively. This transition model is stationary, i.e., it is independent of time. The transition probability is not explicitly defined, but it is underlying the simulation model discussed in the previous section.

Action space $\mathcal{A}$

Action space $\mathcal{A}$ of the problem shall cover all possible decisions available within the process. Most MDPs require one action per time unit. However, in Den Hartogh's network management needs each region a market correction at the start of the month. All MCs can be determined in one action, but this results in an action space that is equal to $|MC|^{|R|}$. This results in a lot of options for a network with 11 different MCs and 6 regions. The model-based controlled learning algorithm uses roll-outs to determine the best action. With so many actions, the MCL algorithm cannot efficiently search for the best action.

In order to limit the state space and effectively use MCL, an action will be made for each region separately. After all events are generated, the agent determines for each region separately the best market correction. The agent makes these actions in the same sequence each month. The model-based controlled learning algorithm uses roll-outs to determine the best MC for each action. In the roll-out analysis of one action, the 11 MC options are tested while the MC for the other regions / actions is fixed to the best MC based on the estimated neural network values. After all actions are made to determine the best MC for every region, the MCL have tested all actions in the action space which is equal to $|MC| \times |R|$.

Initially, the action sequence is based on the alphabetical order of region abbreviations (R1, R2, R3, R4, R5 and R6). The impact of the action sequence is evaluated in Section 5.4.5.

State space $\mathcal{S}$

State space $\mathcal{S}$ in the given model provides all necessary information about the features of the global network. These features are related to the regions, lanes and time. This information is presented in vectors with lengths equal to the number of regions or lanes. Simply put, it should contain all the information, related to the main mechanisms of simulation mechanics: (expected) demand, inventory, (expected) supply and shipments. The state variables are visualized in Figure 4.6.

- Potential new contracts $PNC_{l,d}$ for all lanes $l \in L$ and possible contract duration $d \in D$. The agent could incorporate the volume of contracts, and thus the potential demand on every lane, in its decision making. For instance, if the volume on a lane is relatively low while the inventory in the origin region is quite high, the agent can make the market correction for that region even lower to increase the regional outflow. The potential contracts are grouped per lane and contract duration d can range from 1 month to the maximum contract length which is chosen to be 8 months (discussed in Section 4.2).
- Orderbook $OB_{l,t}$ for all lanes $l \in L$ in month $t \in T$. This keeps track of the contracts that Den Hartogh already has won for each lane to calculate the expected demand. The orderbook is also based on the number of lanes in the network, and specifies the expected number of tanks for each month until the end of the forecast horizon. The history of the arriving potential new contracts and the market correction is not saved in state neither elsewhere in the model. Therefore, all future orders resulting from won potential new contracts are stored in the orderbook. The forecast horizon is 9 month because a contract can have a maximum duration of 8 months and the actualization

time is 1 month. This implies that the winning an 8 month contract in month t leads to orders up to and including month $t + 9$.

- Inventory on-hand IOH_r for all regions $r \in R$. The inventory on-hand indicates the number of empty tanks that are currently stored in each region, and can be allocated to new orders.

- In-transit $IT_{l,x}$ for all lanes $l \in L$ during the in-transit leadtime for a lane $x \in X$. Global orders that are shipped need some time before they arrive at their destination. Frequently, tanks are kept longer in demurrage and need to be cleaned before they can be used for a new order. The total time that an tank cannot be used for another order is called in-transit time. The number of tanks in-transit to a region can give information about the volume and timing of new tank inventory of a region.

- Market correction origin MC_r for all regions $r \in R$. The MC is determined for each region origin and destination. Order contain a region origin and region destination which are used to calculate the applied market correction. The optimal market corrections per lane can be made if the market corrections of the other regions can be incorporated in decision making.

  The agent has to make an action for every region separately. None of the other state variables change between the first and the last action of the time period, except the market correction. Based on the market correction, the agent can observe for which region he determined already a market correction in month t and for which regions not. In addition, from the second action, the agent can observe which market corrections are already set. The agent takes this market correction(s) into consideration for determining the remaining actions. For instance, a network of two regions and two lanes that has MCs of \$100 and \$200 results in the same flows as \$300 and \$400, because the the MC of the lanes are in both situations -\$100 and \$100. Therefore, the market corrections are necessary in the state space. After a new month has started, the MCs of the previous month are removed to indicate that the agent has to make a new action for every region.

- Current Month t. It is necessary to keep track of the time to anticipate on seasonal effects or trends. For this reason is the month is included in the state.

Month		IOH(0)	IOH(1)	IOH(2)	IOH(3)	IOH(4)	IOH(5)	
IT(0,1)	IT(0,2)	IT(1,1)	IT(1,2)	IT(2,1)	IT(2,2)	IT(3,1)	IT(3,2)	IT(3,3)
IT(4,1)	IT(4,2)	IT(5,1)	IT(5,2)	IT(5,3)	IT(6,2)	IT(6,1)		
IT(7,1)	IT(7,2)	IT(8,1)	IT(8,2)	IT(9,1)	IT(9,2)			
IT(10,1)	IT(10,2)	IT(11,1)	IT(11,2)	IT(12,1)				
IT(13,1)	IT(13,2)	IT(14,1)	IT(14,2)	IT(15,1)	IT(15,2)			
IT(16,1)	IT(16,2)	IT(17,1)	IT(17,2)	IT(18,1)	IT(18,2)			
PNC(0,1)	PNC(0,2)	PNC(0,3)	PNC(0,4)	PNC(0,5)	PNC(0,6)	PNC(0,7)	PNC(0,8)	
::	::	::	::	::	::	::	::	
PNC(18,1)	PNC(18,2)	PNC(18,3)	PNC(18,4)	PNC(18,5)	PNC(18,6)	PNC(18,7)	PNC(18,8)	
OB(0,1)	OB(0,2)	OB(0,3)	OB(0,4)	OB(0,5)	OB(0,6)	OB(0,7)	OB(0,8)	OB(0,9)
::	::	::	::	::	::	::	::	::
OB(18,1)	OB(18,2)	OB(18,3)	OB(18,4)	OB(18,5)	OB(18,6)	OB(18,7)	OB(18,8)	OB(18,9)
MC(0)	MC(1)	MC(2)	MC(3)	MC(4)	MC(5)			

Figure 4.6: State variables

Reward function $\mathcal{R}_a(s, s')$
The agent aims to take actions in an environment in order to maximize the cumulative reward. The following factors are included in the reward function:

- *Profit*: a satisfied customer order is expected to have a higher revenue than costs, which results in a profit margin that is the same for all orders. The more contracts are won the more profit can be generated. This profit excludes the bonus or discount of the market correction on the final price.
- *Penalty cost*: Den Hartogh aims to fulfill the order from contracted customers. When a region has not enough tanks, Den Hartogh can reposition empty tanks to the right place or disappoint the customer. Both cases result in additional direct and indirect costs. Therefore, the penalty cost serves as a negative reward to ensure that the agent incorporates the service level into decision making.
- *Holding cost*: a tank is not generating revenue for Den Hartogh when it is waiting to serve a new customer order. The longer the idle time, the less jobs a tank can do in a year which results in less less profit, less overhead contribution and more depot storing costs. Therefore, the holding cost stimulates win more quotes.
- *Market correction revenue*: the market correction that is valid in the month that a contract is won is applied on all orders of that contract. The effect of a market correction is noticeable for a long time. If more contracts are won with a negative market correction, this can have a negative impact on Den Hartogh's result as the price per tank and the marginal revenue is lower. Therefore, it is necessary to include the market correction discount and surcharge affect the operating result. The market corrections are included in the reward in the month that a contract is won. In case that a region has insufficient inventory to satisfy all orders, the market correction penalty or bonus cannot be returned due to modelling requirements. Although, this is not expected to have large impact on the model because market corrections can be positive and negative, which may outbalance the effect on the profit.

As a result, the monthly rewards can be calculated using Equation 4.1:

$$\textit{Reward} = \textit{Profit} - \textit{Penalty cost} - \textit{Holding cost} + \textit{Market correction revenue} \tag{4.1}$$

4.3.1 Relations

The key concepts are mostly interrelated with each other. The relations between the key concept indicate the dynamics of the global network MDP. The relations between some key concepts that are not described earlier are discussed in this section. Moreover, the modelling sequence is briefly described below.

The events, actions and evolve sequence in the model is visualized in Figure 4.7. The events which represents potential new contracts (PNC) are collected between $(t, t+1)$. Then, an action is made for each region to determine the market correction origin. Thereafter, multiple actions occur during the evolve step. First, the market corrections influence how many potential new contracts are won and can be added to the orderbook. As the orderbook keeps track of the won contracts, it can be used to generate the demand of tanks for each lane for month $t+1$. The in-transit tanks arrive in inventory that were shipped at time $(t+1) - x$ with x being equal to a lane's in-transit leadtime. Thereafter, all demand

Table 4.7: Example of calculation potential new contracts to won contracts to additional orderbook

Contract duration	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	
Potential new contracts lane l	15	6	0	9	0	1	0	5	
Won contracts (rounded) lane l	3	1	0	2	0	0	0	1	
Additional orderbook time	**t+1**	**t+2**	**t+3**	**t+4**	**t+5**	**t+6**	**t+7**	**t+8**	**t+9**
Additional orders lane l	0	7	4	3	3	1	1	1	1

must be immediately satisfied from inventory at the beginning of the month. A penalty must be paid for each order that cannot be satisfied. If there is sufficient inventory, holding cost are incurred for the remaining empty tanks. The tank containers that are just allocated to a new job are put on status "in-transit" and will be placed in a pipeline that has the length equal to the lane's in-transit leadtime x. If the expected in-transit time is one month, the tanks will arrive at the start of $t + 2$.

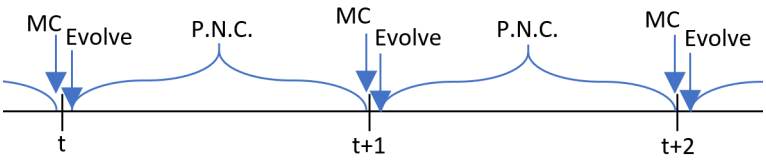

Figure 4.7: Events, actions and evolve sequence

During the month, a specified number of events are generated. An event represents an arriving potential new contract. The event probabilities need to be constant over time and independent of the state due to DRL framework restrictions. Although, the number generated events may fluctuate per month. Potential seasonality in the volume of demand is discussed in the experiments (Section 5.2). After all events are generated, the agent needs to make an action for every region where it determines the market correction origin. When the number of actions made is equal to the number of regions, the market correction for each lane is calculated.

The potential new contracts are translated to a potential new orderbook format. Take for example the potential new contracts for a certain lane in month t, shown in Table 4.7. With a hit ratio of 0.20 in month t, the rounded number of contracts won is 1/5 of the potential contract per lane. This example also assumed that all won contracts result in 1 tank container order per month for the duration of a contract. So, a contract of 1 month results in 1 order, and a contract of 4 months results in 4 orders. The won contracts can be converted to the additional orderbook which indicates what new orders can be added to the real orderbook that keeps track of all won contracts. Because the potential contracts came in month t and actualization time is 1 month, the first orders can arrive in month $t + 2$ at earliest.

When a region has an insufficient empty tank container inventory to satisfy all demand, the lost sales are evenly spread over all outgoing lanes. If the remainder of the shortage of a region divided by the number of outgoing lanes is not equal to zero, the remainder is chronologically spread over the outgoing lanes based on their lane ID number.

4.4 Network characteristics

The modelled network characteristics to the capacity of the network in terms of tank containers. Including the correct number of tanks in the model is important to evoke similar behavior in the model as in Den Hartogh's network. The number of tanks in the model is determined by the initial state which shows the number of tanks that are in-transit or empty waiting for an order. Therefore, the MDP's initial state that is also the same for all experiments should be proportionate to the actual tank container network.

4.4.1 Number of tank containers in the model

At the first of July, Den Hartogh's global network included 6908 tanks of which 5062 (74%) tanks had the in-transit status and 1846 (26%) tanks were empty. Tanks can have two different statuses; empty (waiting for new order assignment) and in-transit which indicates that the tank is already assigned to a task. The total number of tanks allocated for global logistics can variate due to the ISO tank container pool that is shared with the liquid logistics business unit. The number of tank is the model must be adjusted for the modelled flows to achieve the same network dynamics, tank utilization and service level. Moreover, the model includes less variability (e.g. only 6 inventory location and constant in-transit leadtimes) and no operational constraints leading to fewer tanks required.

In-transit tanks Therefore, the demand distribution in the model is combined with the modelled transportation times of all lanes to determine the average number of tanks that are in-transit. The expected number of tanks in-transit based on the demand without market correction influences is 3,988 in the model. This is much less than the 5062 Den Hartogh's in-transit tanks. One reason is that fewer lanes are included in the model, resulting in 26% less demand in the network. In addition, the modelled in-transit leadtime is lower than the expected cycle time that can be obtained from the network snapshot. The snapshot cycle time can be obtained with the in-transit and throughput. Den Hartogh's cycle time of a job can equals 2.9 months where the modelled cycle time almost 1 month shorter. This difference may be caused by the snapshot of the tank statuses in July where the transport times were obtained from 12 months order data. Moreover, tanks are also repositioned to more favorable inventory locations which leads to more in-transit tanks while the number of orders does not increase. In conclusion, the in-transit tanks of the initial model state deviates from the in-transit snapshot of Den Hartogh's network. Though, this difference seems reasonable given the mentioned reasons.

Inventory tanks The starting empty inventory is distributed over the regions based on their monthly export demand. The larger a region's export is, the more empty inventory (safety stock) is required to absorb variance. For a safety stock level of 75% of a region's average export, the ratio in-transit and empty tanks in the model equals roughly the ratio in Den Hartogh's network (5062/1846). This results in 1,553 empty tank distributed over the six regions, which results in a total of 5,541 tanks in the modelled network. Although, the number of tanks in the model is less than in the actual network, it is expected that the ratio in-transit tanks is important and that this number of tanks is appropriate for the fewer lanes and shorter cycle times.

Table 4.8: Begin inventory, in-transit and orderbook

Region	Empty inventory	Lane	In-transit	Orderbook
R1	480	**R1-R2**	160-160	160-106-70-47-28-15-7-2-0
R2	138	**R1-R3**	304-304	304-203-135-89-54-31-14-4-0
R3	330	**R1-R4**	104-104	104-65-43-29-17-9-4-1-0
R4	370	**R1-R5**	72-72-72	72-48-32-20-12-6-3-1-0
R5	97	**R2-R1**	110-110	110-57-31-19-12-6-3-1-0
R6	138	**R2-R3**	26-26-26	26-13-5-4-2-0-0-0-0
		R2-R6	48-48	48-23-13-8-4-2-1-0-0
		R3-R1	224-224	224-148-100-65-40-22-11-3-0
		R3-R4	168-168	168-108-73-48-30-16-8-2-0
		R3-R5	48-48	48-30-20-12-8-4-1-0-0
		R4-R1	226-226	226-40-26-18-10-6-3-1-0
		R4-R3	123-123	123-20-14-8-4-2-1-0-0
		R4-R6	145	145-24-15-11-6-2-1-0-0
		R5-R1	88-88	88-45-24-15-9-5-2-0-0
		R5-R3	16-16	16-8-4-2-0-0-0-0-0
		R5-R4	26-26	26-12-5-3-2-0-0-0-0
		R6-R1	40-40	40-6-3-1-0-0-0-0-0
		R6-R3	17-17-17	17-0-0-0-0-0-0-0-0
		R6-R4	128	128-22-14-8-5-2-1-0-0

4.4.2 Initial tank container distribution for inventory and in-transit

The begin state that will also be used for all experiments is shown in Table 4.8. The begin empty inventory is based on the region's export. The in-transit tanks that arrive in the coming month(s) originates from the demand distribution for all lanes. Moreover, the model starts with a filled orderbook to ensure that the first months do not have a large disturbing effect on the reaming simulation. The number of tanks in the orderbook for the first month is equal to the average number of tanks that are requested, because of the actualization time of one month which means that new contracts have an affect from $t+2$. As the previous section already described, the simulation starts with generating events (potential new contracts) after which the market correction action determines how many new contracts can be added to the orderbook. After this, the last place $(t+9)$ of the orderbook may also be larger than zero.

4.5 Baseline algorithm

This section describes the baseline algorithm (i.e. agent) that will be used to compare the performance of DRL policies for the global network management. The DRL policies cannot be easily compared to the actual or historic management policies of Den Hartogh because the model differs to much from reality. The baseline algorithm is aimed to represent Den Hartogh's current global management. Though, it is complex to convert Den Hartogh's network management that is mainly based on the export forecast, implicit market information and expert knowledge. The baseline algorithm is also designed to handle all

MDP scenarios that are evaluated in the next chapters.

The baseline agent determines the best market correction for each region independently to avoid a too complex algorithm. The baseline agent analyzes only the expected inventory levels of the region for which it has to determine a market correction. The algorithm calculates 11 inventory intervals for each region. One market correction is allocated to each inventory interval. The optimal inventory intervals depends on many factors, such as the total number of tanks in the system, demand variance, in-transit leadtimes, holding costs, penalty costs and profit per satisfied order. Determining the optimal intervals is computationally complex and is not in the focus of this research. The baseline algorithm is aimed to represent the management policies of Den Hartogh as much as possible. Therefore, the most appropriate (i.e. resulting in highest profit) inventory intervals are found by a trial-and-error search.

The baseline algorithm consists of mainly three components which are elaborated below:

- Target inventory level: this indicates the desired regional inventory level for the baseline agent. The market correction origin (and destination) would be zero dollar if the expected inventory level is close to the preferred inventory level (interval). When the expected inventory will be less than the desired interval, a positive market correction origin should decrease the export which should lead to a higher inventory level. On the other hand, a positive market correction origin should lead to a decrease in inventory when the expected inventory is higher than the desired inventory level.
- Interval width: this determines the maximum margin of expected inventory level from the desired inventory level. The larger the intervals, the more the expected inventory level is allowed to deviate from the target before larger market corrections are applied. In other words, smaller intervals will result in an earlier and more powerful market correction for deviating inventory levels.
- Expected inventory level: There are many possible policies for using the rolling inventory forecast for the next nine months. For example, the policy could be focused on short-term, long-term or a combination of both. The optimal choice depends on characteristics of the global network. For instance, a network with long in-transit leadtimes would benefit from a more long-term approach because actions need more time before they have an impact the network's balance (e.g. it takes longer for import tanks to solve a shortage if the in-transit leadtimes are long).

The three components that mostly affect the baseline algorithm are interrelated and cannot be optimized independently. Therefore, an extensive search has been executed to find the optimal combination of the three factors that results in the highest profit. First, the average monthly export is calculated for each region, based on the number of potential new contracts per month, contract lane and duration probability distribution and hit ratio. The average regional export is an important factor because higher exports require more inventory and safety stock.

The expected inventory level should balance the short- and long-term effects of actions. Actions focused on short-term can avoid penalty and holding cost. Though, being too focused on short-term could lead to suboptimal behavior in the long term. Den Hartogh is constantly searching for the optimal balance between the (certain) short-term and (uncertain) long-term forecasts. The inventory forecast that is used in the first month that market corrections must be set, is shown in Table 4.9. What can be noticed is that the long-term (e.g. $t + 9$) forecasts tend to be much higher or much lower than the short-term forecast. This is a result of the imbalanced network that is given as the model's input. The expected monthly export and import without the influence of market correction can be calculated based on the

Table 4.9: Inventory forecast

	t+1	t+2	t+3	t+4	t+5	t+6	t+7	t+8	t+9	Exp Inventory
R1	526	550	573	599	629	660	692	731	768	676.7
R2	117	81	44	9	-24	-53	-80	-103	-127	-56.0
R3	389	456	519	586	654	723	795	861	930	747.4
R4	298	233	170	109	49	-17	-81	-144	-211	-39.2
R5	84	71	57	41	29	18	7	-4	-17	15.1
R6	143	148	148	155	163	171	178	183	187	171.9

Table 4.10: Average monthly export and import

	Import	Export	Import - Export
R1	1067	995	+72
R2	252	294	-42
R3	776	669	+107
R4	656	1006	-99
R5	191	221	-32
R6	299	303	-5

demand probabilities, shown in Table 4.10. It is important to have a strong long-term focus because most in-transit leadtimes are between the two and three months and the imbalanced flows in the network. Adequate actions in time are necessary to balance the network. However, imbalances are not expected to have extreme disturbances on the network with a network where the demand can be backordered. In the evaluation of the baseline algorithm also appeared that the long-term forecast are often too pessimistic as the agent effectively balances the network. Therefore, the baseline agent makes a careful trade-off between short- and long-term, just as Den Hartogh does. The baseline agent includes all nine inventory forecasts, but the forecasts are weighted linearly based on how many months they are in the future. Consequently, the 1-month forecast counts only once, the 2-month forecast counts double, the 3-month forecast counts three times more than the 1-month forecast, and so on until the 9 month forecast. This calculation determines the expected inventory per region, which is also provided in the last column of Table 4.9.

The desired inventory level that equals 118% of a region's export and an MC interval width of 26% of a region's export. This policy with the expected inventory calculation leads to the largest profit for the baseline agent for the base scenario (which will be discussed later). This results in the inventory level intervals that are shown in Table 4.11. The baseline agent chooses the smallest market correction origin for which the expected inventory is smaller than the upper bound of the interval. For example, the MC origin in R1 is 500$ for an expected inventory of -10 tanks, or 0$ for 865 tans, or -400$ for 1450 tanks, or -500$ for 1800 tanks.

A potential downside of the baseline algorithm is that the market corrections are determined independently of each other. Only the expected inventory level of the region itself is taken into consideration. For example, all market corrections can be roughly equal to each other if most or even all regions have tank shortages or surpluses. As a result, the combined lane market corrections will be close to zero as the MC origin and destinations will cancel out each other and the network cannot be effectively controlled.

Table 4.11: Inventory level intervals to determine regional MC origin

	500	**400**	**300**	**200**	**100**	**0**	**-100**	**-200**	**-300**	**-400**	**-500**
R1	<0	<174	<348	<522	<697	<871	<1045	<1220	<1394	<1568	>1568
R2	<0	<51	<102	<153	<204	<255	<306	<358	<409	<460	>460
R3	<0	<117	<235	<352	<470	<588	<705	<823	<940	<1058	>1058
R4	<0	<132	<264	<397	<529	<662	<794	<927	<1059	<1192	>1192
R5	<0	<38	<77	<116	<155	<194	<233	<271	<310	<349	>349
R6	<0	<53	<106	<159	<213	<266	<319	<372	<426	<479	>479

However, the network could be controlled more effective if the regional market corrections are not determined independently from each other. Then, the actions can focus on several regions that have the worst prospects compared to other regions.

Chapter 5

Sensitivity analysis

The model and baseline algorithm are described in the previous chapter. This chapter uses this model to evaluate the performance of DRL against the baseline agent. Two purposes can be identified for this chapter. First, describing the experiments that are used in both this chapter and the next one. Second, discussing the result of the experiments for the sensitivity analysis. As a result, this chapter does not solely focus on the sensitivity analysis. Although, the chapter is called sensitivity analysis since it plays an important role in this research. The background of the sensitivity analysis is described first. Next, the different scenarios are described that are used in the sensitivity analysis and in the cross comparison analysis in Chapter 6. Thereafter, the experimental design and setup are discussed. Finally, the results of the sensitivity analysis are evaluated in order assess the impact of various factors on the performance. This chapter is aimed to answer SRQ2: what is the impact of factors that play an important role in the global network management?

5.1 Background

The purpose of a sensitivity analysis is to determine the relative influence of parameters, initial conditions, and alternative assumptions on model output. In each comparison run, all parameters are kept constant, except for the parameter being examined (Kerr & Goethel, 2014). The scenarios will be designed such that various important factors of the model can be analyzed. The sensitivity analysis focuses on factors such as price elasticity, demand seasonality and profit-costs ratio. In addition to the base scenario, eight other scenarios are compared with the base. These analyses serves two purposes: model validation and network analysis. The performance of the agents should be logical in a valid model. For instance, the costs are expected to increase for higher inventory holding costs and the profit is expected to decrease for increasing uncertainty. In addition, the sensitivity analysis shall be used to gain insights into various factors and their relation to the performance. This may provide insights into questions as: is it better to over- or underestimate costs? For example, if the performance of the agent that is trained in a high-cost scenario is relatively better in a low-cost scenario than vice versa, it can be concluded that overestimation of costs is less harmful in comparison to underestimation. In conclusion, the sensitivity analysis is expected to provide insights into the factors of the global network management as well as in the reliability of the model.

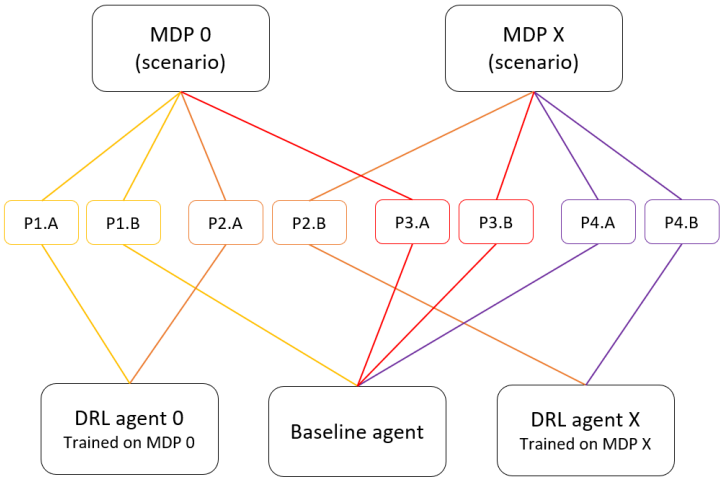

Figure 5.1: Conceptual sensitivity analysis example

The comparisons between the baseline algorithm and DRL agents are visualized in Figure 5.1 with MDP X ranging from 1 to 8. Eight DRL agents are trained, on the base or one of the seven alternative scenarios. The agents' performance is compared between DRL and the baseline for a specific scenario (yellow P1.A&B and purple P4.A&B). In addition, the performance of DRL can also be compared between two different scenarios (orange P2.A&B). The performance of the baseline algorithm in two different scenarios can be evaluated in the same way (red P3.A&B). Based on these two comparisons can be concluded in which scenario DRL or baseline can perform better. The performance between other scenarios are not directly compared in order to limit the comparisons to only the base scenario (MDP 0) and one other scenario (MDP X).

5.2 Analysis scenarios

It is important to clearly establish some directions in analysing the experiments for the experiments required for sensitivity analysis. These directions form the basis in formulating problem instances to experiment with. The scenarios should be designed in such a way that potential differences can be clearly demonstrated. The scenarios used in the sensitivity analysis are also used in the cross comparison analysis, which is provided in the next chapter. First, a base scenario is established which is used to compare deviating scenarios with.

Scenario (MDP) 0: Base scenario
The base scenario represents the most realistic MDP instance of Den Hartogh's global network (given the necessary assumptions). Some MDP variables and parameters are determined with the help of basic data analysis. Other variables and parameters that are more complex and less critical are estimated with the help of network experts.

- *Potential new contracts seasonality*
  Den Hartogh experiences quite some demand uncertainty in their network. This is caused on the one hand by fluctuations in the potential new contracts (quote) volume and on the other hand by tank container order volume variability. In the model's assumptions of Chapter 4 is already discussed that the order volume variability is neglected in the model in order to decrease the

uncertainty. However, the quote variability could be easily implemented in the model, which will make it much more representative for the global network. The quote volume relative fluctuation compared to the average number of monthly quotes is visualized in Figure 5.2. The quote data from 2019 and 2020 show some clear fluctuations in the number of quotes. However, a reliable monthly seasonality analysis cannot be conducted because data of many years is required. The data of earlier years is less complete and thus less reliable. Moreover, the exact monthly seasonality pattern goes beyond the purpose of analyzing the behavior of DRL policies. Therefore, the quote seasonality is modelled with a high-season with 110%-125%-110% and a low-season with 90%-75%-90% of the average monthly demand. This seasonality is expected to have an observable result in the agents' performances. In addition, the DRL is expected to learn which months can have high impact on the network, and policy can be adequately adapted to this.

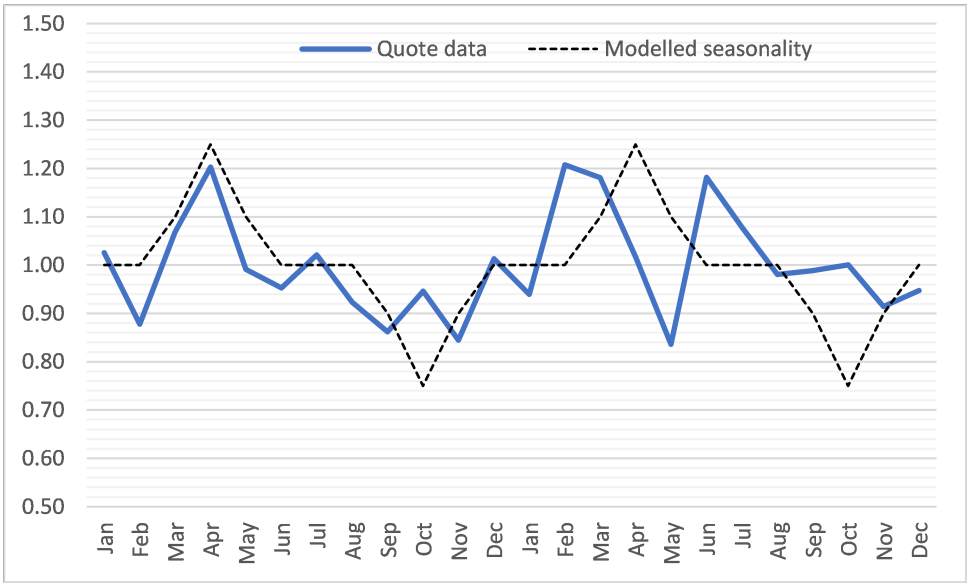

Figure 5.2: Quote seasonality

The price elasticity of tank container is a highly important but quite uncertain factor for Den Hartogh's global network management. The price elasticity is complex to quantify in a single variable through very limited data availability of this exogenous factor. Therefore, the price elasticity is roughly estimated to be linear and will influence the hit ratio with maximal -50% or +50% for a lane market correction of $1000 or -$1000. The relationship between the market correction and change in hit ratio is visualized in Figure 5.3.

- *Profit margin*
  The average profit margin per order is fluctuating over time. Some months can have negative profits and others can have high profit margins. For simplicity, the model uses a profit margin of $200. The MC revenue or discount is not included in this profit margin. In the sensitivity analysis, it may become clear what the effect of the profit margin is on the management policies.

- *Penalty costs*
  The average penalty costs are complex to calculate, because insufficient inventory impacts indirectly many factors. For example, insufficient inventories are expected to result in decreasing customer and employee satisfaction, and increased repositioning and transport costs. These costs are difficult to quantify into a single number. Therefore, the penalty costs are estimated to approximately $1800.

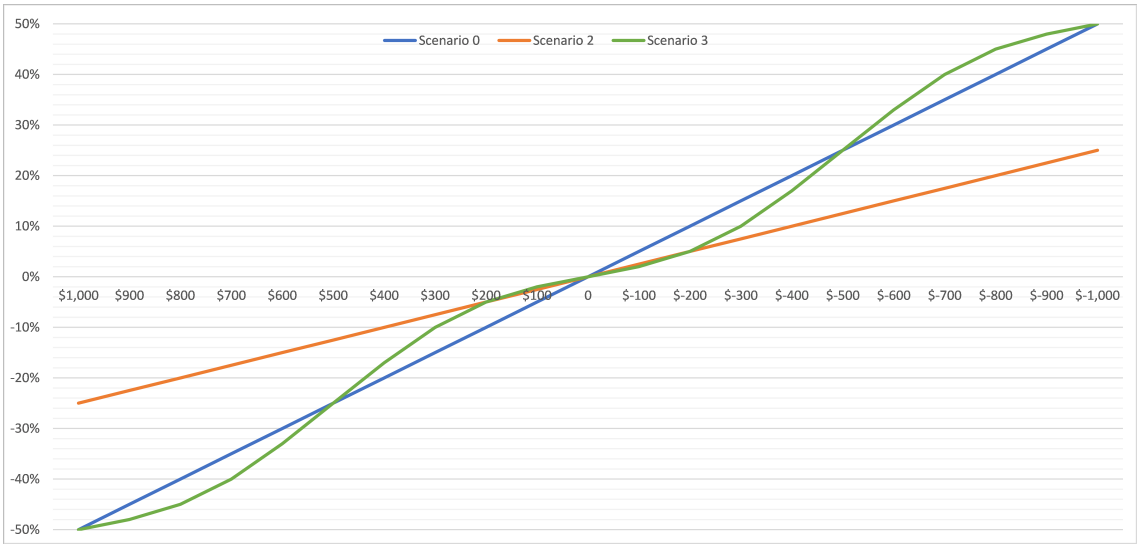

Figure 5.3: Price elasticity scenarios

Table 5.1: Hypothetical contract lane distribution

from \ to	**R1**	**R2**	**R3**	**NEA**	**R5**	**R6**	**Import surplus**
R1		4020	7486	2524	1886		7%
R2	2716		760			1228	-17%
R3	5492			4056	1154		14%
NEA	5462		3064			3544	-15%
R5	2302		524	710			-16%
R6	1080		566	3204			-2%

This penalty should be high enough to get a DRL agent that balances the network properly. This penalty should not be too high either, because lost sales do not have to be avoided at any cost.

- *Holding costs*
  The model uses a holding cost of $15.

- *Contract lane distribution*
  The model is formulated in the previous chapter, Section 4.2.6 provided the contract distribution regarding the lane and duration. Naturally, this distribution is used in the base scenario, but an experiment will be designed to analyze policies' sensitivity to changes of the contract lane probabilities. Changing the lane probabilities leads to a different balanced network. For instance, a net import region can change to a net export region.

- *Action sequence*
  The agent determines the best market correction for each region in a separate action. In the base scenario, the sequence of actions for each region goes in alphabetical order. This results in the following sequence: R1, R2, R3, NEA, R5 and R6. With this experiment, it can be analyzed how sensitive the DRL performance is for the action sequence. Though, this experiment is not intended to compare with the baseline agent since the baseline agent determines the MC for each region independently of other MCs. Therefore, the performance of the baseline agent is not impacted by the action sequence.

Scenario (MDP) 1: Stationary demand

The first alternative scenario is designed to analyze the impact of monthly seasonality in the number of potential new contracts generated. The base scenario includes a demand peak that has a duration of three months with a maximum of 125% and a demand dip of the same duration and a minimum of 75% of the average number of contracts. The demand is modelled stationary during the year in this scenario.

Scenario (MDP) 2: Lower price elasticity

It is important to experiment with price elasticity, because this price elasticity is a large uncertainty factor in Den Hartogh network management. Therefore, it would be interesting to see the effect of a lower price elasticity. À lower price elasticity gives the agent less control to steer the network. The price elasticity is half the initial price elasticity, thus ranges from -25% to +25% which is also shown in Figure 5.3. In order to assess the model validation, the network MDP with a lower price elasticity is expected to have less profit.

Scenario (MDP) 3: Non-linear price elasticity

It is likely that the actual price elasticity shows non-linear behavior, because tank prices are relatively close to each other. This means that if Den Hartogh's price exceeds others, almost all quotes will be declined and further price increases does not significantly impact the hit ratio. This scenario aims to investigate whether the DRL agent can observe the non-linearity of the price elasticity and incorporates this into decision making. This price elasticity is lower for small and large market corrections ($MC \leq |200|$ and $MC \geq |800|$). Though, moderate market correction ($|300| \leq MC \geq |700|$) have a larger price elasticity.

Scenario (MDP) 4: Lower penalty costs

A scenario with lower penalty costs is analyzed to gain insights into the importance of estimating the penalty cost correctly. Because penalty costs are difficult to quantify, it would be meaningful to known what effect the penalty costs have on the network management policies. This scenario uses penalty costs of $800 instead of $1800.

Scenario (MDP) 5: Lower profit margin

This scenario is almost similar to the penalty costs scenario. However, the effect of the assumed profit margin is analyzed in this scenario. As already explained, the profit margin is highly variable month to month. Therefore, automating the market correction decisions would be easier if profit margin has less effect on the DRL policies. Otherwise, Den Hartogh should put more effort in estimating and predicting the profit margins. The profit margin per order is halved to $100 compared to the base scenario.

Scenario (MDP) 6: Hypothetical contract lane probabilities

Den Hartogh's global network is highly dynamic and the volume of flows can significantly fluctuate from month to month. It is most likely that the historical flows do not precisely represent the network's balance. This scenario is designed to analyze the impact of changing the geographical demand. This implies that the total global volume does not change, but the lanes' volumes change. As a result, net exporting regions can turn into a net importing region and the other way around. The hypothetical lane distribution is provided in Table 5.2. R1 has turned into a net export region while R5 and R6 have turned into a net import region. This scenario tends to be more challenging, because in the base model R1 and R3 have import surpluses while R1, R4, R5 and R6 have export surpluses. This may be relatively simple to solve since R1 and R3 are also very well connected to the other region. With the hypothetical lane distribution, the imbalances appear to be more complex as the regional since the largest regions

Table 5.2: Hypothetical contract lane distribution

from \ to	R1	R2	R3	R4	R5	R6	Import surplus
R1		2200	3800	1450	1050		-11%
R2	1300		380			600	-4%
R3	2628			2028	650		16%
R4	2200		1600			2000	-7%
R5	950		262	355			8%
R6	553		283	1600			6%

(R1, R3 and R4) show larger imbalances.

Scenario (MDP) 7: Action sequence based on descending regional volume
It would be interesting from a technical point of view to assess the impact of the action sequence on the DRL agent's performance. The base scenario uses an action sequence based on the alphabetical order of the regions' abbreviations. Another action sequence can be based on the regions' demand volume. The greater the export and import of a region, the greater the impact on Den Hartogh's result. This may be an argument to prioritize larger lanes and determine their market correction first. The regional action sequence for this experiment is: R1, R3, R4, R2, R6 and R5.

Scenario (MDP) 8: Action sequence based on ascending regional volume
This scenario is the exact opposite of the previous scenario. The agent's action sequence starts with the smallest region and ends with the largest region based on tank container volume. These two experiments best demonstrate the sensitivity to the sequence of action. The regional action sequence for this experiment is: R5, R6, R2, R4, R3 and R1.

5.2.1 Overview of key variables and parameters of each scenario/MDP

Table 5.3: MDP experiments input

	Profit	p	h	Price elasticity	Seasonal	Lane probs	Sequence
MDP 0	200	1800	15	Linear (+/-50%)	Yes	Historical	Alphabetical
MDP 1	200	1800	15	Linear (+/-50%)	No	Historical	Alphabetical
MDP 2	200	1800	15	Linear (+/-25%)	Yes	Historical	Alphabetical
MDP 3	200	1800	15	Non-linear (+/-50%)	Yes	Historical	Alphabetical
MDP 4	200	800	15	Linear (+/-50%)	Yes	Historical	Alphabetical
MDP 5	100	1800	15	Linear (+/-50%)	Yes	Historical	Alphabetical
MDP 6	200	1800	15	Linear (+/-50%)	Yes	Hypothetical	Alphabetical
MDP 7	200	1800	15	Linear (+/-50%)	Yes	Historical	Volume descending
MDP 8	200	1800	15	Linear (+/-50%)	Yes	Historical	Volume ascending

Table 5.4: Fixed experiment parameters

Number of regions	6
Number of lanes	19
Actualization time	1 month
Maximum contract length	8 months
Horizon length	9 months
Hit ratio for MC(0)	0.20
Average number of events per month	1390
Number of potential new contracts per event	4
Average number of potential new contracts per month	5560 (1390×4)
Episode length	60 months

5.3 Experimental design and setup

5.3.1 Experiment parameters

In addition to the variable experiment parameters described below, the parameters in Table 5.4 stay constant for all experiments. Most parameters are already discussed in the problem formalization (Section 4.2). The input distributions for the transport times for all lanes, contract lane probability distribution and duration probability distribution for origin region for all lanes are also already provided in the problem formalization section. The initial state of all experiments is set equal to the initial status discussed in the section network characteristics. Other experiment parameters that have not been discussed yet are the number of tanks per month of the contract length and episode length. Four of the same contracts are generated per event in order to make simulation faster and required memory size smaller. Each month, thousands of contracts need to be generated to equal Den Hartogh's monthly volume. Generating so many contracts is computational burdensome. Moreover, generating multiple contracts per event increases the variability in the potential new contracts. This variability is needed to challenge the agent more and the result may be more easily noticeable. A potential downside is that contracts that have a very small probability are generated every month as the number of events decreases. Four contracts per event is chosen as it reduces the computation power and needed memory significantly and it increases the variability. At the same time, low probability contracts are also generated in most months, because still 1390 events are generated on average. The episode length is relation the duration of one simulation. This is set equal to 60 months in order to have sufficient simulation length to analyze the agents' actions and to restrict the required memory size.

5.3.2 Hyperparameters

An extensive hyperparameter search has been conducted to find the best performing combination. Hyperparameters are related to the training procedure of the DRL agent and to the lay-out of the artificial neural network. The DRL framework that uses the model-based controlled learning algorithm has various hyperparameters. The number of samples is related to the number of simulations (i.e. number of generated demand (potential new contracts) sequences) that the model uses to train itself. Too many samples may result in overfitting, where too few samples may result in underfitting (Van der Aalst et

al., 2008). The number of initial roll-outs and maximum roll-outs are related to the number of samples (demand sequences) that will be used to analyze which action has the highest reward in a specific state. The actions that have the lowest probability to be the optimal action will already be excluded during the initial roll-outs. The remaining actions that have a good chance of being "optimal" are analyzed until the maximum number of roll-outs is reached or only one action remains. The optimal number of roll-outs are likely to increase as the variability or number of actions in the model increases. The number of generations indicates how many training cycles the DRL agent will have. The hyperparameter that is related to the design of the artificial neural network determines the 'complexity' of the ANN. The more hidden layers and nodes, the more calculations are required to update the values of the nodes in the ANN. The higher the number of features included in the modelled problem, the more information must be translated into actions. However, the optimal ANN design depends also on the other hyperparameters and is very problem specific, so not a general rule can be applied. Gijsbrechts et al. (2019) described that hyperparameter search may be require many trial and error runs, since because there is not a rule or heuristic that easily finds the optimal combination.

An extensive search has been conducted to find the best performing hyperparameter combination for the base scenario. More than 30 different hyperparameter combinations were tested with a trial-and-error approach. The mathematical notation of the hyperparameters is adopted from Van Jaarsveld (2020). It appeared that the optimal hyperparameter combination consists of $K = 100,000$ samples, with $\underline{n} = 200$ initial roll-outs and $\bar{n} = 250$ maximal roll-outs. The training of one generation of neural network that has one hidden layer with 128 nodes. The following three parameters were directly adopted from Van Jaarsveld (2020). The training uses a minibatch size of 64, $\epsilon = 0.05$ value of bandit optimizer and $\beta = 0.05$ as the fraction of random actions.

This DRL agent generates on average 1.52% more profit than the baseline agent in one episode. A total of 250 simulations were replicated to calculate the average performance difference with an appropriate confidence interval. Each simulation consists of a 60-month period over which the profit is maximized. This hyperparameter combination will also be used to train DRL agents for other scenarios. This hyperparameter combination may not be optimal for these scenarios. Although, it provides a fair comparison for all DRL agents and saves a lot of tuning time and computational effort. Moreover, the expected differences between various close-to-optimal hyperparameter combinations are small.

5.3.3 Computation power consumption

DRL requires usually a lot of computation power to experience enough situations (i.e. states) in order to learn policies that are "good". A total of 100,000 samples are collected to provide the agent with sufficient situations in the experiments. In addition, the episode length of 60 months causes the required time for generating one sample to be 1.3 seconds. The long computation time is caused by the high number of events that must be generated each month. In addition, this research has not focused on efficiently programming the global network (management) since this was not expected to be a bottleneck.

The whole learning process of one DRL agent takes approximately 42 minutes with 40 nodes and 24 cores per node. In addition, roughly 400GB of memory is utilized for training, which is also a result of the large number of events that must be generated per episode. In the experiments, on average 1390 events per month are generated, thus 83,400 per episode of 60 months. For this research, parallel computation is executed on the Cartesius cluster of SURFsara.

5.3.4 Performance evaluation

The most common performance measure for the DRL agent is the absolute profit, because the reward corresponds directly to the objective of the optimization problem. The DRL framework returns every month the total profit. The function that compares the performance of two agents can only show the mean profit and its standard deviation after the episode has completed. The number of replications depends on the performance variability and desired confidence interval. For the experiments in this research, 250 replications are used to achieve a reliable estimation of the performance of agents in a scenario. In the next section, the performances are shown of DRL agent trained on that specific scenario. The cross comparison between agents is described in the next chapter. Unfortunately, the DRL framework does not yet provide simple possibilities to analyze the performance and statistics (e.g.lost sales, inventory on-hand, MC correlations) in detail. The framework can only return the cumulative reward of one episode. The reward function is adjusted to return the value of a KPI that is analyzed. For instance, a reward of \$1 is returned to determine how often a specific MC origin is applied for a region. The reward is zero for months with a different MC origin. Eleven simulations of 250 replications are conducted to determine the MC origin distribution for one region. Besides the fact that this takes a lot of time, there can also be randomness in the replications. The analysis of various statics is mainly discussed in Chapter 6.

5.4 Sensitivity analysis results

5.4.1 Seasonality

Table 5.5 shows the performance of the baseline and DRL agents that are trained on that MDP instance. The baseline agent can achieve an average profit of 26.31 million dollars in a scenario with seasonal demand of potential new contracts. The profit generated without seasonality is slightly more, 26.31 million dollars. This follows the logical expectation, because the baseline algorithm does not include a mechanism to detect and anticipate on seasonality. The DRL agent shows a similar increase in profit for MDP 1, but the difference between the two scenarios is significantly less compared to the baseline agent (shown in bottom row "Difference 1"). In addition, the last column "Difference" shows that the difference between BL and DRL is 1.52% in the favor of DRL with seasonality. On the other hand, the difference without seasonality is only 1.29% between BL and DRL. This may indicate that the DRL agent experiences less difficulty than the baseline agent in anticipating on seasonal demand.

Table 5.5: Seasonality experiments (profit $\times \$1,000,000$)

	BL	**DRL**	**Difference**
MDP 0	26.31	26.71	1.52%
MDP 1	26.39	26.73	1.29%
Difference 1	0.30%	0.07%	

5.4.2 Price elasticity

From Table 5.6 can be concluded that the baseline performs 0.53% worse with a lower price elasticity, but 0.38% better for a non-linear price elasticity (price elasticities visualized in Figure 5.3). This suggests

that the agent experiences more difficulty in balancing the network with less change in the hit ratio for a specific market correction. This seems reasonable because the agent has less control over the network. On the other hand, the baseline agents perform better with a non-linear price elasticity. This might be a result of the market correction on the profit. From a simulation appeared that the baseline agent loses \$245600 due to market correction in the base scenario. The baseline agent loses only \$163500 due to market correction with the non-linear price elasticity in scenario 3. In conclusion, the baseline agents give less discounts, because it is shown that the price elasticity is higher for moderate market corrections.

The DRL agent's performance increases when the price elasticity is lower or non-linear. The increase in performance for lower price elasticity seems strange as the agent has less control over the network. The increase in the non-linear price elasticity can be explained by the fact that the agent has more impact on the hit ratio for moderate market corrections compared to the basis linear price elasticity. What also appeals is the fact that the DRL agent performs better compared to the baseline agent, when there is a low price elasticity (2.48%) compared to the higher price elasticity (1.52%). This suggests that the DRL agent can control the network with higher profit when the price elasticity is lower.

Table 5.6: Price elasticity experiments (profit $\times \$1,000,000$)

	BL	**DRL**	**Difference**
MDP 0	26.31	26.71	1.52%
MDP 2	26.17	26.82	2.48%
MDP 3	26.41	26.79	1.44%
Difference 2	-0.53%	0.41%	
Difference 3	0.38%	0.30%	

5.4.3 Profit and costs

The results of the scenarios with different penalty costs and profit margins are shown in Table 5.7. The baseline agent achieves a slightly (0.04%) higher profit with penalty costs of \$800 instead of \$1,800. This small difference in performance indicates that the baseline agent has a strong focus on preventing lost sales. The average profit decreases with 53.04% if the fixed profit margin per tank is halved to \$100. This shows the dominance of the profit margin in the total network profit. If the profit margin is less dominant, as in MDP 5, the DRL agent can relatively outperform the baseline agent more easily, 3.31% performance improvement compared to 1.52%. The DRL agents show similar performance differences between the different scenarios. In conclusion, the small performance improvement for \$1000 decrease in penalty costs indicate that lost sales a not frequently occurring in both scenarios. Moreover, from the results of scenario 5 can be concluded that the profit margin revenue is very dominant in the model, which makes improvement in penalty, holding and market correction costs relatively difficult to observe. Therefore, 1.5% performance improvement in the base scenario has a different value as in scenario 5.

5.4.4 Contract lane distribution

Table 5.8 shows the model's sensitivity of the contract lane distribution. Both the BL and DRL agents achieve a lower profit with the new distribution in MDP 6 compared to the base MDP. The performance of the baseline algorithm decreases with 4.07% while the DRL performance decreases with 2.95%. Moreover,

Table 5.7: Price elasticity experiments (profit $\times\$1,000,000$)

	BL	DRL	Difference
MDP 0	26.31	26.71	1.52%
MDP 4	26.32	26.74	1.60%
MDP 5	12.36	12.76	3.31%
Difference 4	0.04%	0.11%	
Difference 5	-53.04%	-52.21%	

the DRL outperforms the baseline with 2.69% in MDP 6, but only with 1.52% in MDP 0. This may indicate that DRL is even more useful in scenarios with more complex imbalances.

Table 5.8: Lane distribution experiment (profit $\times\$1,000,000$)

	BL	DRL	Difference
MDP 0	26.31	26.71	1.52%
MDP 6	25.24	25.85	2.69%
Difference 6	-4.07%	-2.95%	

5.4.5 Action sequence

The performance of the agents for the action sequence scenario are shown in Table 5.9. The baseline performance is not impacted by the action sequence, because the actions are determined independently. Though, the performance of the DRL policies is impacted by the action sequence. The difference between the base scenario sequence and the largest region first sequence is negligible, only 0.01% in favor of the base sequence. This may be explained by the fact that the alphabetical sequence does not differ that much from the sequence based on descending order volume. On the other hand, the performance of the DRL agent decreases with 0.12% if the MC of the smallest region is determined first. This confirmed the hypothesis that it will be advantageous to determine MC of the most important regions and take these into consideration for the smaller regions.

Table 5.9: Action sequence experiment (profit $\times\$1,000,000$)

	BL	DRL	Difference
MDP 0	26.31	26.71	1.52%
MDP 7	26.31	26.71	1.51%
MDP 8	26.31	26.68	1.40%
Difference 7	0%	-0.01%	
Difference 8	0%	-0.12%	

5.4.6 Model validation

A sensitivity analysis is often used to validate the model. The central question here is: does the model show logical behavior or performance if one factor is changed? The developed scenarios highlight various elements in the model. For this validation, the performances of the baseline and DRL agents are evaluated

and are shown in the Tables 5.5, 5.6 and 5.7. The performance of both the baseline and DRL for the base and eight other scenarios is visualized in Figure 5.4. The performance of both the baseline and the DRL agents is higher for the scenario with stationary demand compared to seasonal demand. This demand fluctuation can increase the probability of lost sales in high-season, and excess inventory in low-season. Therefore, the performance difference seems reasonable.

Theoretically, the performance is expected to decrease when the model has less control over the network. Therefore, the performance increases for the DRL agent with a lower price elasticity seems questionable. A possible explanation for this might be that the DRL agent tends to overreact for potential shortages and surpluses. A lower price elasticity can reduce the DRL agent's overreaction. Although, the baseline agent's performance decreases for lower price elasticity. Both agents can generate more profit for the non-linear price elasticity in scenario 3. This may be reasonable since the agent gets a "bigger bang for the buck" for moderate market corrections. Especially if moderate lane MCs are more frequently occurring than very large or small MCs. This will be analyzed in next chapter. Altogether, the price elasticity experiment shows valid model behavior. Only the performance ambiguity for scenario 2 can be seen as negative for the model validity.

The performance differences for lower penalty costs or lower fixed profit margin are exactly how it would be expected from a theoretical point of view. Therefore, the cost and profit components indicate that the model and agents' behaviors are valid. The agents' performance for the scenario 5 with the halved profit margin is not shown in the figure, because it would make the y-axis to long for good visibility of the differences.

The performance for the baseline and the DRL agent decreases for the new geographical demand distribution. The action sequence has a small impact on the performance of the DRL agent. Although, the small difference in performance follow the expectation that starting with the largest regions results in more profit.

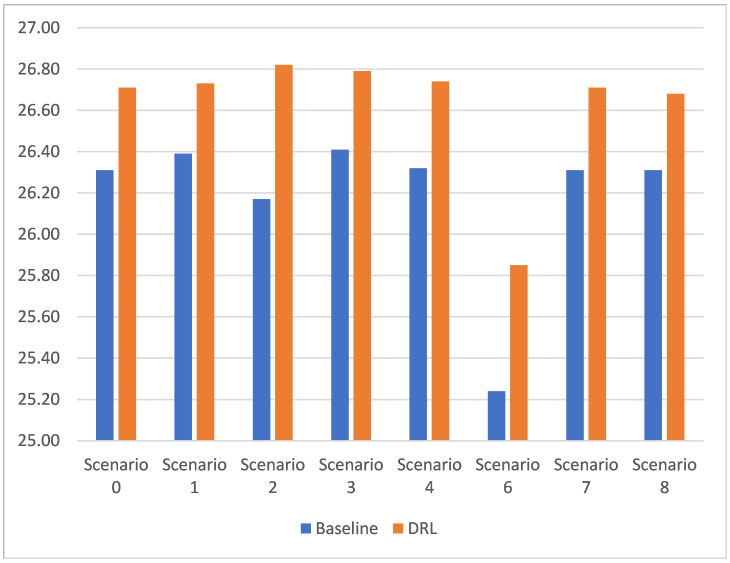

Figure 5.4: Baseline and DRL profit performance for first 5 scenarios (profit $\times \$1,000,000$)

Chapter 6

DRL policy analysis

This chapter discusses the performance of DRL policies for the global network management. First, the scenarios will be analyzed with cross comparisons between DRL agents. The same experiments as described in Section 5.2 and the experiment parameters provided in Section 5.3.1 are used in the cross comparison (except for the action sequence scenarios). Thereafter, Section 6.2 analyzes the DRL policy for the base scenario in more detail with the help of various statistics (i.e. market correction distribution). The chapter ends with a model extension in which the effect of a separate MC for spot and tender contracts is analyzed. This Chapter formulates an answer for SRQ3: how do deep reinforcement learning policies perform in the global network?

6.1 Cross comparison scenario analysis

In the cross comparison scenario analysis are DRL agents trained in a specific Markov decision process that includes the specific parameter values that are unique for that scenario. Thereafter, the trained agent has to manage a modelled network that is slightly different than the MDP in which the agent is trained in. For example, the agent could be trained in a specific MDP with seasonal demand. Then, the trained agent has to interact with an environment that does not have seasonal demand. The performance of this agent can be compared with an agent that has been trained in the environment without seasonal demand. Such a cross comparison is conceptually visualized in Figure 6.1 (pink P2.A&B and blue P3.A&B). X indicates a different scenario which can range from 1 to 6. The cross comparisons can provide new insights into questions such as whether a DRL agent can easily adapt to new demand patterns. In addition, the difference in the performance of two agents can be compared for a scenario. The change in performance may provide new insights into the robustness of a DRL agent (green P1.A&B and turquoise P4.A&B).

6.1.1 Seasonality

Table 6.1 shows that DRL0 performs better than DRL1 in MDP0 and DRL1 performs better than DRL0 in MDP 1. It may be interesting that the DRL agent trained in MDP 0 (called DRL0) performs 0.11% better than the agent trained in the MDP without seasonality. On the other hand, DRL1 performs only

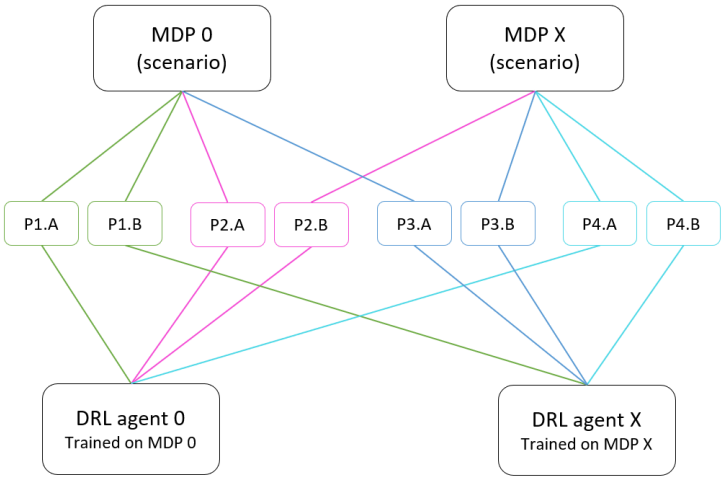

Figure 6.1: Conceptual cross comparison analysis example

Table 6.1: Seasonality experiments (profit $\times \$1,000,000$)

	DRL0	**DRL1**	**DRL0 vs DRL1**
MDP 0	26.71	26.68	-0.11%
MDP 1	26.72	26.73	0.04%
Difference 1	0.04%	0.19%	

0.04% better than DRL0 in MDP 1. This may indicate that the DRL agent can adapt more easily to less demand seasonality than the other way around. Knowing this, an advice for Den Hartogh would be that demand seasonality should not be underestimated.

6.1.2 Price elasticity

Comparing the performance of agents DRL0 and DRL2 gives interesting insights which are shown in Table 6.2. DRL2 is trained with a price elasticity that is half as responsive as in MDP 0, performs 14.34% worse than DRL0 in MDP 0, while DRL0 performs only 2.76% worse than DRL2 in MDP 2. When DRL2 has to control the network of MDP 0, it is likely to overreact because the agent expects less effect on the hit ratio for market correction. On the other hand, DRL0 might underreact in MDP2 because it expects a higher impact on the hit ratio. The experiments show that overreaction of DRL2 results in a performance decrease of 14.34% compared to DRL0 in MDP 0, while underreaction of DRL0 results in only a performance decrease of 2.76%, compared to DRL2 in MDP 2. This indicates that overreacting on network changes is more costly than underreacting.

The same performance differences appear in the comparison between DRL0 and DRL3. DRL0 performs 2.02% better than DRL3 in MDP 0 and DRL3 performs only 0.30% better than DRL0 in MDP 3. This suggests that assuming a linear price elasticity while the actual price elasticity is non-linear is better than the other way around. However, this does not have to apply for all non-linear price elasticities as the effect may depend on type of non-linearity. Moreover, the performance of DRL0 agent is not expected to change for the non-linear scenario as the profit stays equal to $26.71 million. The policies developed by DRL0 are therefore robust for potential non-linear price elasticity as long as that the absolute maximum change of the hit ratio is 50%. If the absolute maximum change is lower than 50%, the performance can

decrease with 2.38%. Based on these cross comparisons, determining the price elasticity range ((-50%, 50%) vs (-25%, 25%)) is more important than differentiating between linear or (slightly) non-linear price elasticity.

Table 6.2: Price elasticity experiments (profit $\times \$1,000,000$)

	DRL0	**DRL2**	**DRL3**	**DRL0 vs DRL2**	**DRL0 vs DRL3**
MDP 0	26.71	22.88	26.17	-14.34%	-2.02%
MDP 2	26.10	26.82		2.76%	
MDP 3	26.71		26.79		0.30%
Difference 2	-2.28%	17.22%			
Difference 3	0.00%		2.37%		

6.1.3 Profit and costs

The last two columns of Table 6.3 indicate a minimal difference between agents that are trained in a different profit or penalty environment. For the decreased penalty costs of \$800 in MDP 4 can be concluded that lost sales are avoided as much as possible. Penalty costs of \$800 tend to be high enough to aim for approximately the same service level as for penalty costs of %1,800. The same effect can be observed for the halved fixed profit margin per tank. The DRL5 agent that is trained with a profit of \$100 is able to generate as much profit as DRL0 in MDP 0 with a profit margin of \$200. The same applies to DRL0 that performs almost equally well as DRL5 in MDP 5.

Table 6.3: Profit and costs experiments (profit $\times \$1,000,000$)

	DRL0	**DRL4**	**DRL5**	**DRL0 vs DRL4**	**DRL0 vs DRL5**
MDP 0	26.71	26.70	26.71	-0.04%	0.00%
MDP 4	26.72	26.74		0.07%	
MDP 5	12.76		12.76		0.02%
Difference 4	0.04%	0.11%			
Difference 5	-52.22%		-52.21%		

6.1.4 Contract lane distribution

Table 6.4 shows the cross comparison between the agents DRL0 and DRL6. From this experiment can be concluded that the agent trained on the base scenario performs 2.47% better than DRL6 in MDP 0. On the other hand, DRL6 performs only 0.28% better in MDP 6, in which it is trained. It is not obvious which contract lane distribution is more complex or more difficult to balance. This makes it difficult to indicate which (geographical) demand changes impact the performance of the DRL agent. At least, it can be concluded that DRL policies are robust for changes in demand, because the performance of both agents did not decline drastically in a new scenario.

Table 6.4: Lane distribution experiments (profit $\times \$1,000,000$)

	DRL0	**DRL6**	**DRL0 vs DRL6**
MDP 0	26.71	26.05	-2.47%
MDP 6	25.85	25.92	0.28%
Difference 6	-3.22%	-0.50%	

6.2 DRL policy analysis

The DRL policy is analyzed in this section. Visualizing the DRL policy is complicated because there are many variables that represent a state. Moreover, variables as regional inventory, in-transit pipeline and orderbook can take tens or hundreds of different values. Finding relations between specific states and actions is not expected to be fruitful. Therefore, mainly the statistics of the market corrections, flows and performance is analyzed. The analysis starts with a broad view and deepens into the interesting aspects.

6.2.1 Regional market corrections

In this section the DRL agent's policy is further analyzed for the base scenario (i.e. MDP 0). The goal is to obtain a better understanding of the differences between DRL policy and the baseline policy. Figure 6.2 shows the average MC origin distribution (based on 250 replications) for both the baseline and the DRL agent during one episode. Both agents tend to use more positive than negative MC origins. Although, the distribution of DRL agent is slightly more shifted towards the lower MC origins. The MC origin probability distributions for both agents are shown in the Appendix A in Tables B.1 and B.2. Figures 6.4 and 6.5 show the number of times an MC origin occurs on average for a region in one episode of 5 years. Both figures show a regional MC origin distribution that resembles a normal distribution. The average regional MC origin can be calculated from the regional MC distribution. The average MC origin is plotted against the regions relative import surpluses in Figure 6.3. This figure shows that the DRL agent uses a lower MC origin for every region. R1 and R3 have the largest absolute and relative differences compared to baseline agent: \$69 (-48%) and \$72 (-112%), relatively. It is no coincidence that these are the two import surplus regions. It makes sense stimulating regional exports by lowering the MC origins.

Another appealing aspect of these four numbers is that the MC origin is mainly focused on the positive domain. For the baseline agent, it can be argued that the number of tanks in the model is too low. The baseline agent only takes the inventory position of one region into account. If all regions are low in inventory, all regions will get a high MC origin to discourage export and encourage import. However, applying the same MCs for all regions will not influence the hit ratio as the MC for the lanes are \$0. A possible explanation for the positive biased MC origin distribution of DRL agent is that it might prefers to have more room/options for using negative MC origins. A lower MC origin results in higher export flows and lower import flows. This may imply that the DRL agent prefers to have more control and freedom in decreasing regional inventory than increasing inventory (which is caused by more positive MC origins). Another noticeable result of Figure 6.3 is that R6 and R4 have the highest regional MC origins while they do not have the largest export surpluses. Actually, R6 is the most balanced region since the import demand is roughly equal to the export demand. The market corrections must be analyzed on

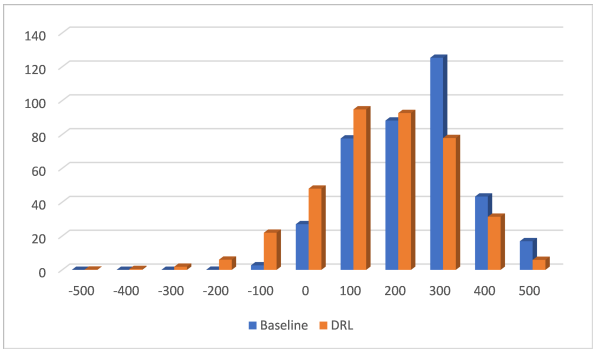

Figure 6.2: Global MC origin distribution

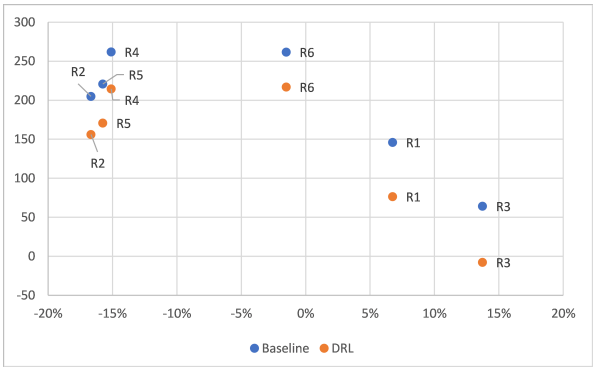

Figure 6.3: Average MC origin and import surplus

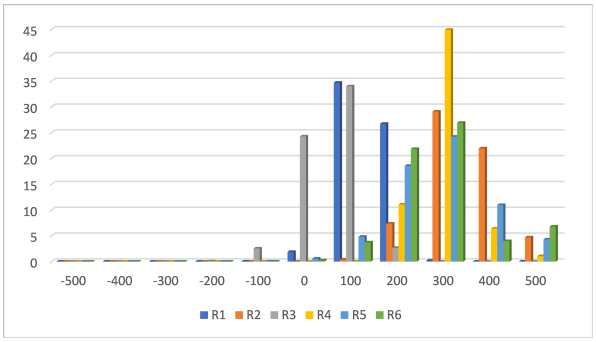

Figure 6.4: Regional MC origin distribution BL

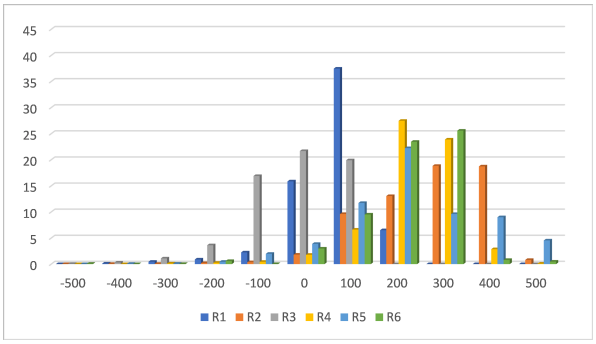

Figure 6.5: Regional MC origin distribution DRL

lane level in order to find an explanation for the (relatively) high MC origins of R4 and R6.

6.2.2 Lane market corrections

Considering only the MC origin of all regions does not give a complete picture, because the actual MC lane impacts the monthly hit ratio. In other words, the timing of a regional MC origin is important in relation to the other regional MCs. Figure 6.6 shows the relation between the average MC lane and the lane's balance. A lane's balance is indicated to be the origin region's import surplus plus the destination region's export surplus. The larger the lane's balance, the higher the expected need for more orders on this lane to increase the origin region's export and the destination region's import. The bubble sizes are defined to be the average volume for each lane, indicating the lanes' impact on the network. The figure tends to show a negative linear relationship between the lane's balance and the average MC lane. The data points are not on one straight line, but seem to lie on within a linear bandwidth. For instance, the lanes R4-R6 and R6-R4 have a quite different balance while both have an MC lane close to zero. This implies that the larger the sum of import surplus origin and export surplus destination, the lower the expected MC lane. Given that lanes move tanks from a region that have a larger import surplus than the destination's export surplus improve the network's balance and is therefore logical behavior.

In Figure 6.6 can also be observed that lanes originating in R1 and R3 often have a negative MC. This negative MC is aimed to increase the export of these two net import regions. The MC on the lanes between R4 and R6 approach the zero dollar. This may already indicate why R6 has such a high average MC origin: to avoid a disturbance of the flows between R4 and R6. Figure 6.7 displays a more detailed lane analysis. The second and third column show the regional dependencies of a lane. For example, the

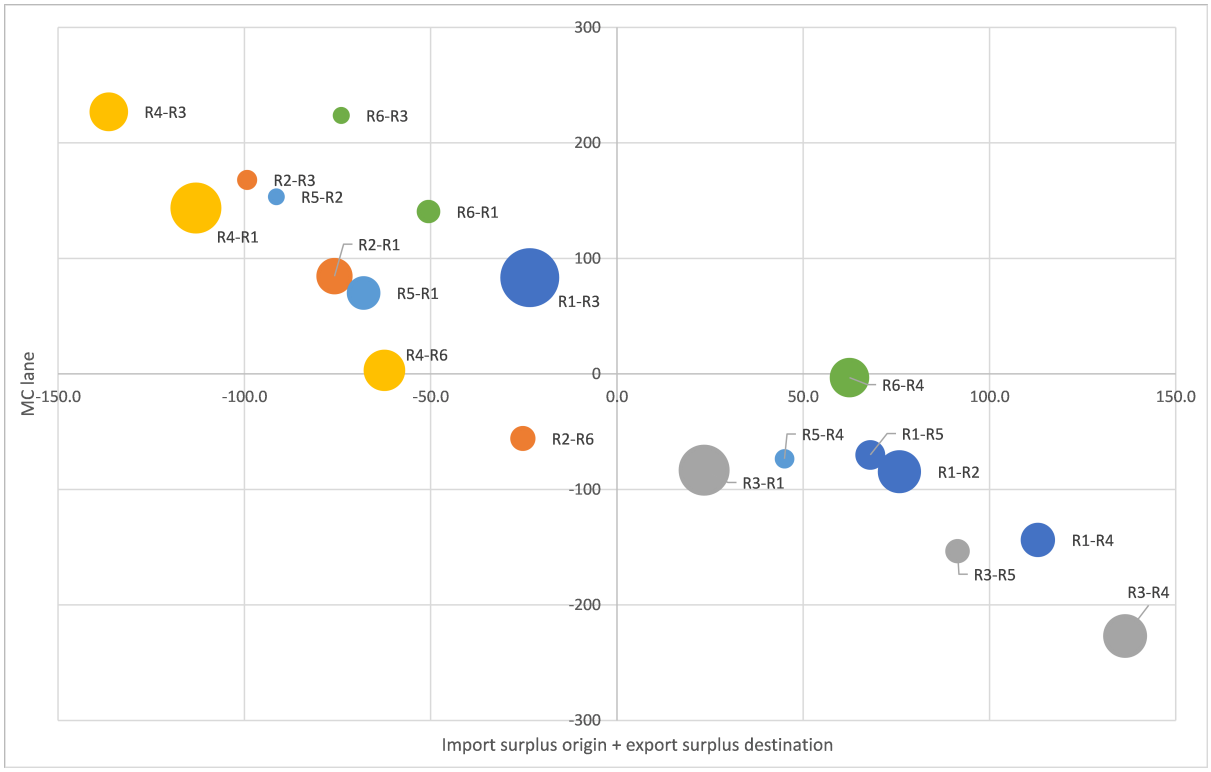

Figure 6.6: Average MC lane ($) compared to lane's balance with bubble size as avg lane volume

first lane is equal to 25% of R1's import and 100% of R2's import. The importance of this lane to balance R1 is limited compared to the importance to ensure sufficient flow to R2. R6 is highly dependent on R4 for its import (74%) and export (66%), while R4 is less dependent on R4. R4 exports a large volume of tanks to R1 and R3, which both have import surpluses. Therefore, R4 is a perfect region to decrease its export in order to decrease the import surplus of R1 and R3. However, R6 is also negatively impacted by R4's higher MC origin while R6 has hardly an imbalance. In other words, R6 needs to compensate R4's high MC origin with even a higher MC origin in order to have an attractive price for demand between R4 and R6. This is the most likely reason for the high MC origin of R6. The example of R6 tends to be typical for low volume regions. The main regions like R1, R3 and R4 make sure that the largest lanes will balance the network. Thereafter, the smaller regions like R2, R5 and R6 will base their MC on the prominent regions.

Figure 6.7, it can also be seen that the lanes that go from net import to net export regions are fully utilized to balance the network. These lanes are indicated with "perfect" network feasibility. Worst feasibility is for lanes that go from net export to net import regions. Other possibilities are that a lane goes from net import to net import ("destination import surplus") or from net export to net export ("origin export surplus"). These two types solve only the problem of one region. The column "actual flow" shows the relative difference between the actual flow after the MC and the initial flow without MC interference. It shows that the DRL agent increases the flows for "perfect" lanes with 8.1% to 12.9%. The lanes that are "origin export surplus" or "destination import surplus" are changed with -4.4% to 4.0%. The lanes labelled as "worst" are changed with -8.1% to -15.9%. This shows that the DRL agent consistently influences the flow of every lane based on the lane origin's balance and the lane destination's balance.

In Figure 6.8, relative change of a lane's flow plotted against the lane's balance. The relationship also

Lane origin-destination	Export region dependence	Import region dependence	Exp won demand (0.2)	DRL won demand	Actual flow (relative change)	Avg MC lane	Origin import surplus	Destination export surplus	Lane's balance	Netto imbalance
R1-R2	25%	100%	167.9	185.2	10.3%	$ -85	47.3	28.5	75.8	Perfect
R1-R3	47%	60%	313.1	299.7	-4.3%	$ 83	47.3	-70.8	-23.4	Destination overflow
R1-R4	16%	24%	106.5	117.0	9.8%	$ -144	47.3	65.7	113.0	Perfect
R1-R5	12%	62%	80.7	87.2	8.1%	$ -70	47.3	20.7	68.0	Perfect
R2-R1	58%	16%	115.9	103.0	-11.1%	$ 85	-28.5	-47.3	-75.8	Worst
R2-R3	16%	6%	34.9	29.3	-15.9%	$ 168	-28.5	-70.8	-99.3	Worst
R2-R6	26%	26%	54.9	52.5	-4.4%	$ -56	-28.5	3.3	-25.3	Origin outflow
R3-R1	51%	32%	229.8	238.9	4.0%	$ -83	70.8	-47.3	23.4	Destination overflow
R3-R4	38%	39%	169.3	191.2	12.9%	$ -227	70.8	65.7	136.4	Perfect
R3-R5	11%	38%	51.7	56.5	9.3%	$ -153	70.8	20.7	91.4	Perfect
R4-R1	45%	32%	231.0	212.2	-8.1%	$ 144	-65.7	-47.3	-113.0	Worst
R4-R3	25%	25%	131.7	114.2	-13.3%	$ 227	-65.7	-70.8	-136.4	Worst
R4-R6	29%	74%	151.8	150.6	-0.7%	$ 3	-65.7	3.3	-62.4	Origin outflow
R5-R1	65%	13%	99.2	90.6	-8.7%	$ 70	-20.7	-47.3	-68.0	Worst
R5-R3	15%	4%	24.2	21.0	-13.2%	$ 153	-20.7	-70.8	-91.4	Worst
R5-R4	20%	7%	32.6	31.7	-2.8%	$ -74	-20.7	65.7	45.0	Origin outflow
R6-R1	22%	6%	47.6	43.6	-8.4%	$ 141	-3.3	-47.3	-50.6	Worst
R6-R3	12%	5%	25.1	22.1	-12.0%	$ 224	-3.3	-70.8	-74.0	Worst
R6-R4	66%	31%	137.5	137.8	0.2%	$ -3	-3.3	65.7	62.4	Origin outflow

Figure 6.7: Lane MC analysis

tends to be linear within a certain bandwidth. In general, the lower the MC lane, the higher the relative change of won orders. Besides the fact that all data points are within a certain bandwidth, the lanes with the same origin (and color) tend to be on a regional linear line. For instance, the line can be drawn through the lanes originating from NEA which is quite different than the lines through lanes from R5 or R2. Moreover, the larger lanes tend to be more centered around the upper part of the bandwidth. The upper site of the bandwidth results in a higher MC lane and a more positive change of the order won. A higher MC lane results in a higher MC revenue compared to the lower site of the bandwidth. Therefore, the MC revenue can be increased by aiming at the upper bound of the bandwidth for the larger lanes.

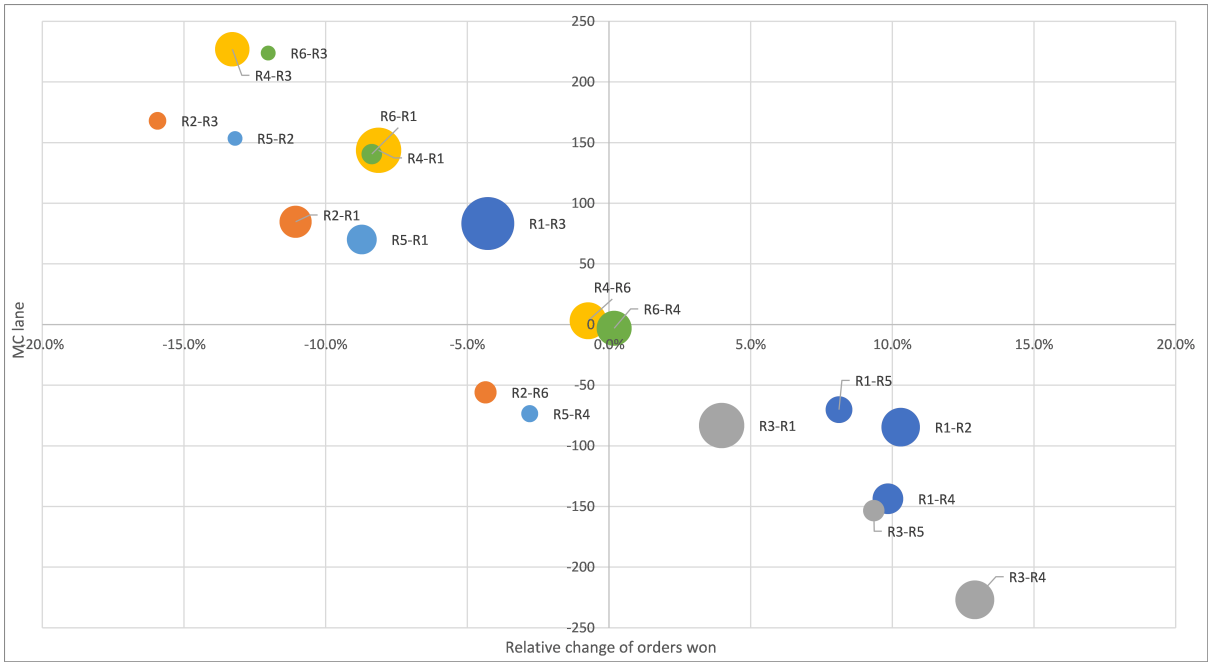

Figure 6.8: MC lane compared to relative change in flow with bubble size as avg lane volume

6.3 Model extension: separate MC for spot and tender contracts

The MDP formulation in Chapter 4 is aimed to represent Den Hartogh's current global network management. Investigating potential process improvements can provide Den Hartogh with new insights. Some improvements in the global network management are complex to test in practice and the support for process changes is limited if they are not extensively investigated in advance. This research can contribute to quantifying a possible network management improvement. This section investigates the effects of a separate market correction for spot and tender contracts instead of the same market correction for all contract lengths.

In this experiment, spot contracts are categorized as one- and two-month contracts. Tender contracts can have a duration between three and eight months. Theoretically, it would make sense to value a spot contract differently from a tender contract in terms of market correction. The market correction is defined as the expected compensation that a tank is worth in a network perspective. For instance, the market correction origin increases when regional inventory decreases in order to discourage the export and to encourage the import of tanks. A spot contract that imports tanks to this region for two months has a different value than a tender contract that is importing tanks for eight months. The tank shortage in a region is not expected to remain for a long period (i.e. six months) since Den Hartogh's network is highly dynamic. Therefore, the import of tanks with a spot contract is expected to have a different network value per tank compared to a tender contract.

The reason why Den Hartogh does not use a separate market correction for spot and tender contracts is because of the existing risk that too much money can be lost on the market corrections. In the current situation, tanks entering a region with a specific MC destination compensate for the tanks leaving the region with MC origin. The only risk is that the volume of import and export can differ

per month, which can result in either more MC origin discount/surcharge or more MC destination discount/surcharge (depending on the MC origin and destination). When there is a separate MC for spot and tender contracts, tanks can enter with MC spot and leave with an MC tender which is not restricted to be mirrored. Moreover, regions can variate on the ratio spot and tender contracts which increases the uncertainty in the MC revenue.

Although, Den Hartogh's assumption that there is a balance between the market correction discount and surcharge also not plausible. Negative market corrections will increase the ratio won contracts (and thus actual demand) while positive market corrections will decrease the hit ratio. Therefore, negative market corrections are expected to be applied more frequently than positive market corrections which results in a negative net MC revenue. Though, this depends on the initial demand flows in the network before the flows are controlled with market corrections. This theory of negative net MC revenue is supported by the base model experiment. Based on 250 replications, the expected MC revenue of the baseline agent is -$337,172.

6.3.1 Adjusted Markov decision process

The agent needs to determine a market correction for spot and for tender contracts. The MC origins for both spot and tender contracts is still bounded to the interval (-$500, $500) with a $100 interval. The agent needs to make twelve actions instead of six. As a result, the state space has now two vectors with the length equal to the number of regions, instead of only one vector. The MC origins for both spot contracts are stored in a vector and the MC origins for tender contracts are stored in a vector. In this experiment, contracts with a one- or two-month duration are categorized as spot and thus longer duration is classified as tender contracts. In addition, the tender MC is also applied to orders in the first two months of a tender contract. All other MDP parameters and distributions are equal to those in the base scenario (MDP 0) which is provided in Section 5.2.

6.3.2 Results

A new DRL agent is trained on this new MDP with the same hyperparameters as the training for the base model in order to obtain a fair comparison. The baseline algorithm does not include an option to determine separate MCs for spot and tender contracts. Therefore, the DRL performance of only the initial MDP is compared to the DRL performance of the adjusted MDP. Table 6.5 shows the performance achieved by the DRL agents in the normal situation where there is only one MC for each region, and an MDP where the agent determines the MC for spot contracts and the MC for tender contracts. The DRL agent is able to generate 20.65% more profit if it can use different market correction for spot and tender contracts. An interesting aspect of this large increase of performance is that it is not caused by the profit margin due to increased number of tanks shipped, neither lower penalty or holding costs. However, the DRL agent managed to increase the market correction revenue from $101,693 to $7,192,850.

Revenues and costs

Table 6.6 shows the MC revenue from spot and tender contracts. Out of $7,192,850 MC revenue, $5,384,107 comes from spot contracts and $1,808,740 from tender contracts. In the standard MDP where only one MC is applied to all contracts, $3,029,563 is generated from contracts with a one- or two-month duration while $2,927,870 is lost on tender contracts. The DRL agent can increase its profit

Table 6.5: Separate MC spot and tender contracts

	DRL profit	Contract margin	MC revenue	Penalty	Holding
MC(spot & tender)	$ 26,716,969	$ 27,892,700	$ 101,693	$ 17,984	$ 1,259,440
MC(spot) & MC(tender)	$ 32,234,954	$ 26,601,700	$ 7,192,850	$ 107,096	$ 1.452,500
Difference	+20.65%	-4.63%	+6973%	+495%	+15.33%

Table 6.6: Expected MC revenue for spot and tender contracts

	MC revenue spot	MC revenue tender
MC (spot & tender)	$ 3,029,563	$ -2,927,870
MC(spot) & MC(tender)	$ 5,384,107	$ 1,808,740

from spot contracts with $2.35 million and from tender contracts with $4.73 million with a separate MC for spot and tender contracts.

Lane market correction

The DRL agent uses the market correction in order to increase its profit in the new MDP with a separate MC for spot and tender contracts. This may be logical since the profit margin per tank is $200 while the market correction for a tank ranges from -$1,000 to $1,000. Winning tanks with a market correction below -$200 do not directly generate profit, but no penalty and holding costs need to be paid. The price elasticity tends to play a key role in this experiment as the hit ratio decreases with 0.02 for an additional MC of $100. This implies that the profit increases more compared to the decrease of the demand, relatively. Though, the agent cannot increase all MCs in order to maximize the profit, because balancing the network can be still important. In Figure 6.9 can be observed that the DRL agent uses more extreme (i.e. further from $0) market corrections for spot contracts. A possible explanation for this phenomenon is that tender MCs are more conservative since imbalances are expected to be temporary. In addition, the expected network state has more uncertainty in the long-term which may lead to less extreme market corrections. Spot contracts only impact the next few months, so their network value can be estimated with more accuracy and if inventory allows, the network can be controlled with more extreme MCs. Moreover, the substantial difference between spot and tender market correction may also provide evidence for the hypothesis that the network value is not equal for all contract lengths.

Lane order volume

Figure 6.10 visualizes the number of orders won from spot and tender contracts for each lane in the new MDP. This is compared to the number of orders won from spot and tender contracts in the standard MDP. The DRL agents has for almost all lanes a lower order volume in the new MDP. Though, this decline in flow is almost negligible for the lanes between R1 and R3 (1 and 7). This may seem strange because both regions have large import surpluses. It would be logical that these regions would mainly increase the export of tanks to net export regions. Though, the DRL agent may use its two largest flows to increase its tank container utilization. As a result, the empty tank inventories of R1 and R3 will decrease and the number of in-transit tanks will increase. Moreover, the increase of flow is entirely due to increased spot orders. Having a separate MC for spot contracts provides the agent more flexibility in increasing some flows, when the inventory might become too high.

Regional market correction revenue

No explanation could yet be found from the above MC lane analysis for the high MC revenue for separate

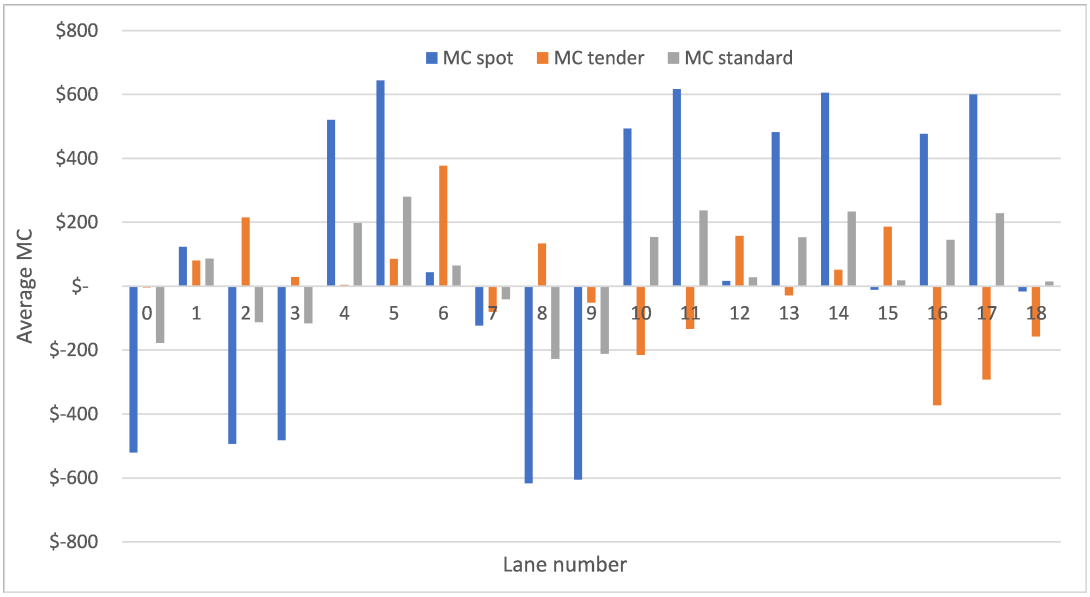

Figure 6.9: MC for spot & tenders and standard MC

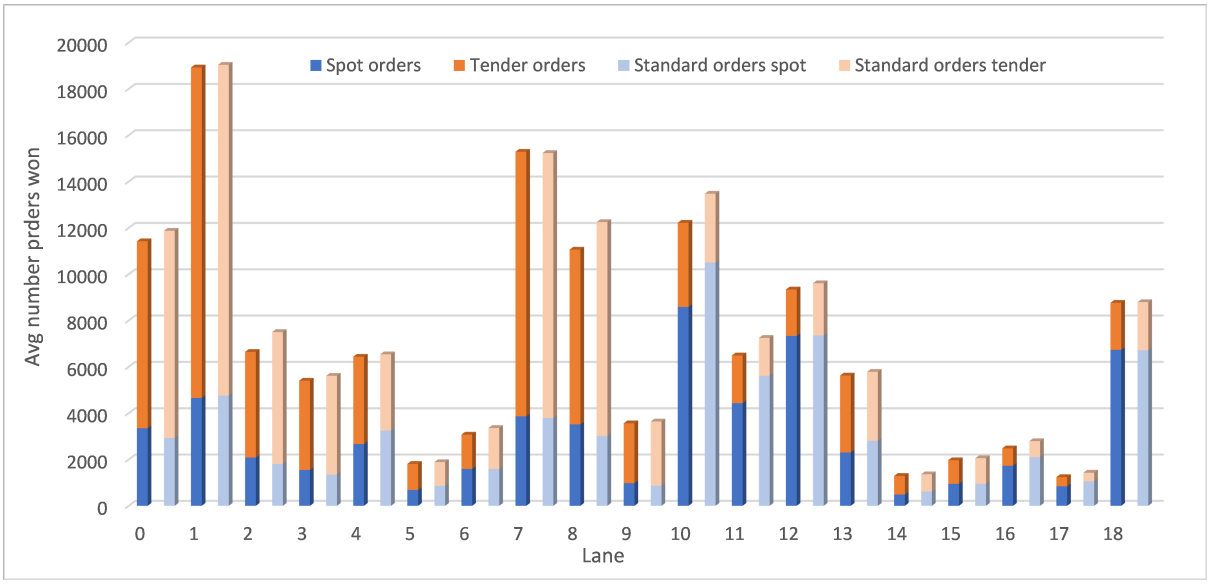

Figure 6.10: Order volume for spot & tenders and standard

Table 6.7: MC revenue for regional export

	Revenue spot	Revenue tender	Total revenue	Mean MC
R1	$ -2,913,439	$ 2,211,491	$ -701,948	$ -16.56
R2	$ 1,885,890	$ 702,639	$ 2,588,529	$ 69.21
R3	$ -3,204,834	$ -75,825	$ -3,280,659	$ -161.71
R4	$ 7,017,444	$ -646,399	$ 6,371,045	$ 381.11
R5	$ 1,390,057	$ 202,890	$ 1,592,946	$ 59.89
R6	$ 1,208,989	$ -586,056	$ 622,933	$ 19.95

Table 6.8: MC revenue for regional import

	Revenue spot	Revenue tender	Total revenue	Mean MC
R1	$ 7,032,784	$ -1,874,374	$ 5,158,410	$ 122.71
R2	$ -1,730,150	$ -15,052	$ -1,745,202	$ -152.81
R3	$ 4,506,309	$ 1,014,875	$ 5,521,184	$ 185.61
R4	$ -3,290,903	$ 1,811,454	$ -1,479,449	$ -52.04
R5	$ -1,317,299	$ -7,981	$ -1,325,280	$ -147.87
R6	$ 183,366	$ 879,818	$ 1,063,184	$ 85.67

spot and tender MCs. The regional MC revenue from exports and imports are shown in Tables 6.7 and 6.8. Based on these tables can be concluded that the highest regional MC revenue is generated in R1 and R4. The high MC revenue for R1 originates from the import while it originates from the export for R4. This is evident as the R1 has an import surplus and R4 has an export surplus, which can be decreased with positive market corrections. As a result, the lane from R4 to R1 generates $3,470,628, which is the highest lane revenue from the market correction. On the other hand, the DRL agent loses 'only' $89,321 on the lane from R1 to R4. The mean MC spot and tender needs to be analyzed in order to explain this result.

The mean MC spot for the lane R1-R4 equals $-490, while the MC tender equals $203. This is the opposite for the lane R4-R1: MC spot equals $490 and MC tender equals $-203. The answer for this remarkable difference for spot and tender MC can be found in the regional contract duration (i.e. spot vs tender) distribution, which is shown in Table 4.4 in Section 4.2.6. The actual achieved ratio of spot and tender orders from EU to R4 is roughly 1 spot for 2.2 tender orders. For the lane R4-R1, this ratio is 1 spot for 0.4 tender orders. This substantial difference in the distribution spot and tender results in a flow from R1 to R4 that consists of mainly tender orders, while the flow from R4 to R1 consists of mainly spot orders. This provides the opportunity to use a profitable MC tender for the lane R1-R4 while the inevitable negative effect (recall that spot or tender MCs are still mirrored) on the lane R4-R1 is limited. Furthermore, the MC spot should be more profitable for the lane R4-R1 than for R1-R4, because relatively more spot orders are shipped from R4 compared to tender orders. The applied mean market corrections follow this reasoning since the MC tender on R1-R4 is positive and the MC spot for R4-R1 is positive. Using the different spot and tender distributions, the upside of the positive market corrections can be utilized while the downside of the mirrored MC lane that flows the other way is limited.

Chapter 7

Conclusions and managerial insights

The aim of this thesis was to develop a proof of concept for DRL that can contribute in the decision making process of the global network management. For this purpose, a model was developed that represents the most important elements of the global network. Also, the performance of DRL was compared with the baseline algorithm and is analyzed in detail. This final chapter draws conclusions on the insights obtained in this thesis and reflects on the conducted research. Section 7.1 summarizes the conclusions for all sub-research question, after which the main research question is discussed. Subsequently, limitation and further research directions are formulated based on insights obtained from this thesis in Section 7.2. In addition, this section provides a brief discussion of the possible improvement of the DRL framework. This thesis ends with an overview of recommendations for Den Hartogh in Section 7.3. This Chapter is also aimed to address SRQ4: what insights can be obtained for the global network management?

7.1 Conclusions

Den Hartogh experiences high uncertainty in the global network management. This uncertainty consists of mainly three factors. The first factor concerns the uncertainty in the effect of market corrections. Second, the demand volatility of contracted customers leads to uncertainty in the short-term. At last, global exogenous factors affect the network mainly in the long-term. All of these uncertainties result in a laborious and stiff pricing process for the market corrections that are used for the global network management. The purpose of this research was to develop proof of concept for DRL can contribute decision making in the global network management. An additional advantage of this research is to gain more insights into the question how the current network management and pricing process can be improved to enhance formalization and automation. Therefore, the main research question was identified:

Can deep reinforcement learning help to find pricing policies for Den Hartogh's global network management?

To answer this (main) research question, four sub-research questions were formulated. Combining the answers to these sub-research questions provides the necessary insights for answering the main research question. Recall the first sub-research question:

SRQ1: How can the global network management problem be modelled as a Markov decision process?

This research primarily focused on the first type of uncertainty to gain new insights into the relationship between market corrections and the network performance. Every month, potential new contracts arrive in the global network model. Den Hartogh determines the market correction that is applied for the monthly potential new contracts. The market correction results in a higher (for negative MC) or lower (for positive MC) probability of winning a potential new contract. If a contract is won, the expected tank container orders are added to the orderbook that keeps track of the expected number of orders for every month and lane. Tank container orders are satisfied from the regional stock point when the simulation arrives in a new month. The tank container arrives at its destination region after the in-transit leadtime and can be used for new orders.

A baseline algorithm is developed to represent Den Hartogh's current network management. The algorithm uses an export forecast to determine regional market correction. The market correction for a region is determined independently from other regions in order to keep the algorithm relatively simple and easy to interpret. The second research question is designed in order to investigate the relations between internal factors and the achieved profit by DRL and baseline algorithm:

SRQ2: What is the impact of factors that play an important role in the global network management?

The impact of various factors (e.g. penalty and holding cost) appeared to be relatively small due to the profit margin ($200) from won orders that tends to be quite constant. As a result, performance differences in market correction revenue or penalty and holding costs have a small impact on the total profit. The DRL is able to achieve 1.52% more profit than the baseline algorithm in the base scenario (scenario that is best representation of the global network). Though, the performance of DRL increases compared to the baseline algorithm for a scenario with lower profit margin.This may indicate that the technique like DRL can make more difference in times of lower margins and increased competition.

The price elasticity and contract lane distribution have the largest impact on the network performance in the sensitivity analysis. The performance of the baseline agent decreased with a lower price elasticity while the DRL agent's performance increased surprisingly. This may indicate that when DRL has increasing value in a network, market corrections can have less influence. In this situation it is even more important to take adequate (timely) actions for network imbalances, because the hit ratio can only be changed up to 25% instead of 50%. Therefore, potential imbalances need to be identified and addressed in a timely manner. Another contract lane distribution has also impact on the network profitability. This seems obvious since other geographical demand can result in imbalances that can be managed with more network profitability. The following sub-research question concerns the more in-depth analysis of DRL policies.

SRQ3: How do deep reinforcement learning policies perform in the global network?

The performance of DRL agents that are trained on different MDPs is analyzed in the cross comparison. Substantial differences appeared between agents that are trained on another MDP than the MDP where it has to perform in. The agent that is trained with a low price elasticity performed 14.34% worse in a high price elasticity scenario, which is probably caused by overreacting to imbalances. On the other hand, the performance decreases with 'only' 2.76% when training took place with a high price elasticity and performance is measured in a low price elasticity scenario. This decrease in performance is probably due to underreacting. Based on this experiment can be concluded that overreacting (underestimating the price elasticity) performs worse than underreacting (overestimating the price elasticity) in the global network management.

We found that the DRL agent is able to increase its profit with 20.65% when it can use a separate MC for spot and tender contracts. The increased profit is entirely attributable to the increased revenue from the market corrections. The DRL agent appears to utilize the different distributions of spot and tender contracts (related to contract length distribution). Den Hartogh's argument for not using spot and tender MCs is that the MC can get out of control. Although, it turns out that the DRL agent makes use of this property to make this work to its advantage. In the current model, the lane from R1 to R4 consists of mainly tender orders while the lane from R4 to R1 consists of mainly spot orders. This provides the opportunity to utilize the upside of the positive MC on one flow while the downside of the mirrored MC in the other direction is limited. DRL tends to be useful in determining the appropriate market corrections for contracts to generate profit with the market corrections alone. In addition, DRL has proven that determining a more accurate market correction for contracts is profitable. The fourth, and final, sub-research question concerns all relevant learnings from this research project:

SRQ4: What insights can be obtained for the global network management?

These new insights are based on the experimental results and experiences gained during the project. The answer to this question is also partially elaborated in the recommendations. The results of the DRL policy analysis showed that the DRL agent is able to manage the global network with more profit than the baseline agent. However, the relative difference between DRL and baseline is limited due to high profit margin and the number of contracts won is approximately the same. It can be concluded that the baseline algorithm is well able to manage the modelled global network. A plausible reason for this is the orderbook, that is included in the state, provides useful information in determining adequate actions to prevent major regional shortages or surpluses. The fact that Den Hartogh's export forecasts are even inaccurate for the next month indicates that appropriate information is available to predict the network's state. The market correction affects only new contracts and therefore hardly impacts next month's exports. Accordingly, improving the information related to the tank container demand of contracted customers seems to be crucial for improving the global network management. Suggestions for improving the customers' demand forecast are discussed in the recommendations (Section 7.3).

Another important insight comes from the sensitivity experiments for price elasticity. These showed that underreacting is less bad than overreacting for the total profit. This is in line with Den Hartogh's network management team which is often hesitant to change market corrections frequently. The main motivation of Den Hartogh to not change market corrections frequently is the uncertain effect of market correction changes.

The average market correction origin of all regions equals $138 for the DRL agent in the base scenario. This implies that the agent has more freedom to lower the MC origins. Decreasing the MC origin (and consequently increasing the MC destination) results in more export and less import, which leads to a lower regional inventory level. This may indicate that the DRL agent prefers to have more options (lower MC origin) to decrease the inventory if the situation allows (i.e. sufficient inventory to satisfy expected demand). R4 and R6 had the highest average MC origin. This can be explained by the fact that R4 is the most important region for the export to R1 and R3 that experienced a substantial import surplus. The MC origin of R4 was increased to probably discourage the export to those regions. An additional effect was that the export to R6 was also discouraged while R6 depends for 74% on import from R4. Therefore, R6 should follow the MC origin of R4 in order to make the MC destination of R6 attractive for demand from R4. Based on this example, it can be suggested that the MCs of large regions determine the MC of small regions. This theory is confirmed by the sensitivity analysis of the action sequence in

which also showed that it is better to determine the MC first for the largest (highly connected and less dependent) regions is better than starting with small (less connected and more dependent) regions.

A separate market correction for spot and tender contracts provides the opportunity to value contracts closer to their expected network value. Moreover, this flexibility provides the opportunity to utilize the regional differences in the spot and tender distribution. For instance, a tank can be shipped to a region with favorable MC tender and shipped from that region with favorable MC spot if most tanks arrive from a tender contract and leave with a spot contract.

The main overall insight developed during this research is that the modelled global network can be balanced quite good with DRL, but also with the baseline algorithm. This was even the case if the price elasticity was actually half the price elasticity as the DRL agent was trained with. This shows the robustness of DRL pricing policies for one of the most important factors in the network management: price elasticity. The most plausible reason for this good performance is that the orderbook provides highly useful information to determine good actions that balance the network. The orderbook ensures that the exports and resulting regional inventory levels can be predicted accurately. This enables the agent to effectively change the market corrections that result in the desired additional orders to balance the network.

Main RQ: Can DRL help to find pricing policies for Den Hartogh's global network management?

DRL has shown to be appropriate for finding policies to manage a global network problem that represents Den Hartogh's network. It is able to achieve 1.52% higher profit than the baseline algorithm and the use of a separate spot and tender market correction resulted in a performance increase of more than 20%. Though, various hurdles need to be taken before DRL can be successfully used as decision support for the global network management team.

Currently, global network management is based on many implicit information (e.g. expert knowledge) which cannot be used in the decision making of DRL. The orderbook is probably the most important element that is missing in Den Hartogh's decision making. In the model, the orderbook provided the agents necessary information to determine what price was required to win sufficient contracts. Den Hartogh's export forecast is most similar to the orderbook. However, the export forecast does not indicate how many tanks have already been won and how many potential new contracts are expected to be won. Therefore, only using the export forecast is not convenient for dynamic pricing.

Another important element in the success of DRL is the definition of the reward function. The sole objective of DRL is to optimize the reward, which makes it crucial the reward function meets all of Den Hartogh's objectives. Then, DRL might make better choices because it makes objective considerations which are not biased by personal interests. It may be difficult to quantify all factors needed in the reward function such as penalty costs that include usually many implicit factors (e.g. customer and employee satisfaction). Objective decision making is likely to be more profitable since regional managers tend to prioritize their own region (and customers). Moreover, the market corrections can be used more to increase the network flexibility (spot and tender MC) and to directly increase the profit with the MC revenue.

7.2 Reflection

7.2.1 Limitations

The research in this thesis is subject to several limitations. First, exogenous factors that impact the global network management could not be included in the model. Competitors' prices impact the relationship between Den Hartogh's hit ratio and the market correction. Den Hartogh did not have data available on the hit ratio and the market corrections in order to estimate the distribution of the price elasticity. Moreover, the tank container volume distribution of contracts could also not be estimated due to missing of data. Furthermore, the effect of fluctuating values of parameters (i.e. profit margin, penalty costs, etc.) and variables (i.e. demand distribution, seasonality, etc.) are not analyzed in the model. For example, the profit margin of tank container shipments can fluctuate per lane and month. These lacking elements negatively impact the validity of the model's outcome in this research. Therefore, the conclusions drawn from the model may be different from the network in reality.

Another limitation of this research is that the DRL behavior is difficult to analyze. Defining the reason for a specific combination of market corrections is complex for such a large problem with many state variables and large variable ranges. For instance, the orderbook contains already 171 variables (19 lanes $\times$ 9 month rolling forecast horizon) and each variable can take many possible values. It is complex to visualize the effect of various state variables on the market corrections.

Finally, in this research are many factors assumed to be known by the decision makers. For example, the orderbook (and thus future tank container demand), the potential contracts and the in-transit leadtime. DRL and the baseline algorithm were capable of managing the network with a high profit. However, DRL in practice with fewer information is expected to be less effective. Therefore, the implementation of a technique such as DRL will not be successful if the pricing process will not be changed.

7.2.2 Future research

Future research related to this project may focus on either modelling the global network more accurately. Besides, future research can also focus on the question how Den Hartogh's global network management can be formalized and made more data driven. The latter research direction is a step that must be taken before machine learning techniques can be used effectively. Machine learning requires a lot of information of the historic network parameters, but also information on how to determine the current network state accurately. The uncertainty of determining the best action is high without sufficient information of the current network state. For instance, a better organization of the information of Den Hartogh's orderbook would immediately improves the global network management without requiring complex machine learning algorithms. One of the most important aspects is that the uncertainty of the current network state needs to be decreased for improved decision making.

Decreasing the state uncertainty implies that Den Hartogh has better information of the number of contracts won and the expected tank container orders from these contracts. Currently, Den Hartogh is very flexible since customers are not obligated to conform to a tank volume. Future research should focus on the question how the demand volatility can be decreased. Several options can be investigated to decrease this uncertainty. For instance, the global network management team determines in advance how much tank containers must flow on each lane. The commercial department should use this volume requirement

to achieve the desired demand. Another option is the introduction of strict volume agreements with customers and a separate pricing mechanism for violating the agreement.

Finally, future research should also focus on improve the model of the global network management. This project only investigated the model's sensitivity of varying values of parameters such as price elasticity, profit margin and costs. However, the model's sensitivity of state variables such as the orderbook and potential new contracts are still not analyzed. Future research may be able to answer the question what the impact of various elements of the MDP is on the model's performance.

7.2.3 DRL framework with MCL algorithm

The DRL framework is used to develop a proof of concept to investigate the DRL for global network management. This framework has helped to implement DRL with only basic knowledge of DRL. Although, two potential improvements were identified during this research.

- Model-based controlled learning algorithm requires deterministic actions since all uncertainty needs to be captured by the composite random variable. However, the effect of actions on the state can also be stochastic due to action uncertainty in various dynamic environments. Incorporating stochastic actions in the framework could be valuable for problems with uncertain action outcomes.
- DRL framework does not have standardized method to analyse the taken actions by the agent. This prevents to analyse the behavior of the agents and to assess the applicability. In this study, the modify state with action had to be changed for every single statistic. For instance, to determine how often a specific region had a specific MC. Only determining the market correction distribution (Figure 6.4) leads to 66 model adjustments and runs (of 250 replication in this study) for a model with six regions and eleven market correction options. Determining the market correction distribution for all lanes required 399 model runs (19 lanes $\times$ 21 market corrections).

  The framework includes the functions "modify state with action" and "modify state with event" where can be programmed how the state changes after an action or event. New functions can be added to the network that can be called "modify KPIs with action" and "modify KPIs with event" in which the key performance indicators are stored. Some possible KPIs for the global network management are: market correction origin for each region, market correction for each lane, variance of the market correction, penalty costs, inventory on-hand. Furthermore, the correlation between actions or state variables can be easily determined with these new functions by storing more detailed information of the applied market corrections for specific inventory level intervals.

7.3 Recommendations

This thesis ends with a number of recommendations for Den Hartogh in order to improve the global network management. These recommendations are based on the results of the research in this thesis and insights obtained during the project.

- Saving the monthly market corrections and saving which market correction was applied and whether the contract has been won, for each potential new contract. Future research on the global network

management will benefit if Den Hartogh starts storing all quote data centrally. For each quote, it is important to store all price elements (e.g. material costs, repositioning cost, etc.) and in particular the market correction. Moreover, the expected monthly tank container demand of a quote should be registered. Combining the quote data with the applied market correction is expected to provide new insights into the price elasticity of lanes. Also, the relation between the price elasticity and the expected monthly demand can be analyzed since spot and tender contracts may react differently to price changes.

- More accurate determination of the market correction in relation to the expected network value of a contract. The experiment with a separate MC for spot and tender contracts showed a substantial increase in profit compared to a single MC for all contracts. Therefore, calculating market corrections for contracts more accurately is expected to improve the global network management and achieved profit. Distinguishing by the duration and volume of contracts can also improve the MC revenue. Changing more frequently the MC (if the network state has changed) also improves the MC accuracy, and thus the network management

  Losing too much money on the market corrections is currently the reason not to change the pricing process. The standardization and automation of the pricing process for global network management can be improved to reduce the risk of having a negative net MC revenue. The MC revenue can have a large impact on Den Hartogh's result, especially in times of low gross margins of tank container shipments. The MC surcharges and discounts can be included in the pricing process to minimize the risk of losing money on MCs. The MC can be combined with the export forecast to estimate the expected MC revenue. Furthermore, it is possible to further extend this model in order to analyze different MC combinations. In the first phase, regional managers can predict the impact of a MC combination on the export volumes. This MC scenario analysis may also result in an improved understanding of the price elasticity dynamics. Eventually, the MC scenario analysis can be automated if Den Hartogh can estimate the price elasticity accurately. This provides the opportunity to increase the impact of dynamic pricing on the global network management and Den Hartogh's financial performance.

- Implementing an orderbook in the global network management in order to keep track of the expected orders from won contracts. The main insight developed during this research is that the modelled global network can be balanced quite good with DRL or the baseline algorithm, even with an inaccurate estimation of the actual price elasticity. This means that there must be another reason why Den Hartogh experiences difficulty in balancing the network. This is likely due to the lack of an (accurate) orderbook.

  Currently, price contracts have little liability for customers to provide accurate demand forecasts. Many customers often use multiple tank container operators for their transports. It is possible that a customer asks for a new price at other tank operators and decides to ship tanks under a new price contract. This can lead to less demand from contracted customers if market prices have declined in the meantime. This may also work the other way around when other tank operators are increasing their prices (if they are running out of inventory). This may lead to increasing tank demand for Den Hartogh, while the prices have not been increased. As a result, Den Hartogh is selling its tanks too cheap and is also going short due to the time delay for price changes. In this case, account managers can prioritize their relationship with customers over Den Hartogh's network and profit.

A potential solution for minimizing the possibilities of customers for gaming (i.e. cherry-picking the cheapest tank operator) is to require demand forecasts from customers. This will give Den Hartogh insights into which customers do not forecast their demand accurately and stimulate them to improve this. The focus of the commercial department and account management will shift from maximizing the sales and their customer satisfaction to maximizing the global network management. In the end, improved global network management can lead to even higher customer satisfaction and profit for Den Hartogh. Moreover, Den Hartogh will be able to manage the global network effectively when future flows can be predicted more accurately. The network management is then no longer dependent on the inaccurate and subjective export forecasts of regional managers. Furthermore, macroeconomic changes can be observed in an early phase when the demand forecasts of customers can be monitored closely. This gives Den Hartogh a competitive advantage in optimizing its prices.

Bibliography

Boute, R. N., Gijsbrechts, J., Jaarsveld, W. V., & Vanvuchelen, N. (2021). Deep Reinforcement Learning for Inventory Control : a Roadmap.

Gijsbrechts, J., Boute, R. N., Van Mieghem, J. A., & Zhang, D. (2019, 1). Can Deep Reinforcement Learning Improve Inventory Management? Performance and Implementation of Dual Sourcing-Mode Problems. *SSRN Electronic Journal*. Retrieved from https://papers.ssrn.com/abstract=3302881 doi: 10.2139/ssrn.3302881

Gosavi, A. (2009). Reinforcement Learning: A Tutorial Survey and Recent Advances. *INFORMS Journal on Computing*, *21*(2), 178–192. Retrieved from http://pubsonline.informs.org.https://doi.org/10.1287/ijoc.1080.0305http://www.informs.org doi: 10.1287/ijoc.1080.0305

ICTO. (2021, 2). *Global Tank Container Fleet Survey 2021* (Tech. Rep.). Retrieved from www.itco.org

Kaelbling, L. P., Littman, M. L., & Moore, A. W. (1996, 5). Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, *4*, 237–285. Retrieved from https://www.jair.org/index.php/jair/article/view/10166 doi: 10.1613/jair.301

Kerr, L. A., & Goethel, D. R. (2014, 1). Simulation Modeling as a Tool for Synthesis of Stock Identification Information. *Stock Identification Methods: Applications in Fishery Science: Second Edition*, 501–533. doi: 10.1016/B978-0-12-397003-9.00021-7

Nikolopoulos, K., Punia, S., Schäfers, A., Tsinopoulos, C., & Vasilakis, C. (2021, 4). Forecasting and planning during a pandemic: COVID-19 growth rates, supply chain disruptions, and governmental decisions. *European Journal of Operational Research*, *290*(1), 99–115. doi: 10.1016/J.EJOR.2020.08.001

Serrano, W. (2019, 8). Deep Reinforcement Learning Algorithms in Intelligent Infrastructure. *Infrastructures 2019, Vol. 4, Page 52*, *4*(3), 52. Retrieved from https://www.mdpi.com/2412-3811/4/3/52/htmhttps://www.mdpi.com/2412-3811/4/3/52 doi: 10.3390/INFRASTRUCTURES4030052

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement Learning: An Introduction* (Second ed.). MT Press. Retrieved from http://incompleteideas.net/book/the-book-2nd.html

Van der Aalst, W. M. P., Rubin, V., Verbeek, H. M. W., van Dongen, B. F., Kindler, E., & Günther, C. W. (2008, 11). Process mining: a two-step approach to balance between underfitting and overfitting. *Software & Systems Modeling 2008 9:1*, *9*(1), 87–111. Retrieved from https://link.springer.com/article/10.1007/s10270-008-0106-z doi: 10.1007/S10270-008-0106-Z

Van Jaarsveld, W. (2020). Model-based controlled learning of MDP policies with an application to lost-sales inventory control. , 1–24. Retrieved from http://arxiv.org/abs/2011.15122

Appendix A

Contract probability distribution

Table A.1: Contract probability distribution on lane l with contract duration CD

Lane	Origin	Dest.	CD(1)	CD(2)	CD(3)	CD(4)	CD(5)	CD(6)	CD(7)	CD(8)
0	R1	R2	0.01882	0.00941	0.00627	0.00471	0.00376	0.00314	0.00269	0.00235
1	R1	R3	0.03505	0.01753	0.01168	0.00876	0.00701	0.00584	0.00501	0.00438
2	R1	R4	0.01182	0.00591	0.00394	0.00295	0.00236	0.00197	0.00169	0.00148
3	R1	R5	0.00883	0.00442	0.00294	0.00221	0.00177	0.00147	0.00126	0.00110
4	R2	R1	0.02545	0.01272	0.00282	0.00212	0.00169	0.00141	0.00121	0.00106
5	R2	R3	0.00712	0.00356	0.00079	0.00059	0.00047	0.00040	0.00034	0.00030
6	R2	R6	0.01151	0.00575	0.00128	0.00096	0.00077	0.00064	0.00055	0.00048
7	R3	R1	0.02572	0.01286	0.00857	0.00643	0.00514	0.00429	0.00367	0.00321
8	R3	R4	0.01899	0.00950	0.00633	0.00475	0.00380	0.00317	0.00271	0.00237
9	R3	R5	0.00540	0.00270	0.00180	0.00135	0.00108	0.00090	0.00077	0.00068
10	R4	R1	0.15343	0.00366	0.00244	0.00183	0.00146	0.00122	0.00104	0.00091
11	R4	R3	0.08607	0.00205	0.00137	0.00103	0.00082	0.00068	0.00059	0.00051
12	R4	R5	0.09955	0.00237	0.00158	0.00119	0.00095	0.00079	0.00068	0.00059
13	R5	R1	0.02156	0.01078	0.00240	0.00180	0.00144	0.00120	0.00103	0.00090
14	R5	R3	0.00491	0.00245	0.00055	0.00041	0.00033	0.00027	0.00023	0.00020
15	R5	R4	0.00665	0.00332	0.00074	0.00055	0.00044	0.00037	0.00032	0.00028
16	R6	R1	0.03033	0.00072	0.00048	0.00036	0.00029	0.00024	0.00021	0.00018
17	R6	R3	0.01589	0.00038	0.00025	0.00019	0.00015	0.00013	0.00011	0.00009
18	R6	R4	0.08997	0.00215	0.00143	0.00107	0.00086	0.00072	0.00061	0.00054

Appendix B

Market correction probability distributions

Table B.1: Regional MC origin probabilities for BL agent

	-$500	**-$400**	**-$300**	**-$200**	**-$100**	**$0**	**$100**	**$200**	**$300**	**$400**	**$500**
R1	0	0	0	0	0.0003	0.0298	0.5460	0.4205	0.0032	0	0.0003
R2	0	0	0	0	0	0	0.0065	0.1161	0.4579	0.3456	0.0739
R3	0	0	0	0.0008	0.0401	0.3820	0.5348	0.0419	0	0	0.0003
R4	0	0	0	0	0	0	0.0002	0.1744	0.7072	0.1011	0.0171
R5	0	0	0	0	0.0007	0.0093	0.0761	0.2923	0.3812	0.1727	0.0678
R6	0	0	0	0	0.0003	0.0038	0.0588	0.3437	0.4236	0.0627	0.1072

Table B.2: Regional MC origin probabilities for DRL agent

	-$500	**-$400**	**-$300**	**-$200**	**-$100**	**$0**	**$100**	**$200**	**$300**	**$400**	**$500**
R1	0	0.0016	0.0072	0.0140	0.0353	0.2496	0.5899	0.1023	0	0	0
R2	0.0002	0.0007	0.0009	0.0032	0.0059	0.0286	0.1513	0.2054	0.2964	0.2948	0.0126
R3	0.0007	0.0048	0.0172	0.0570	0.2659	0.3411	0.3134	0	0	0	0
R4	0	0	0.0019	0.0041	0.0069	0.0281	0.1042	0.4324	0.3759	0.0451	0.0014
R5	0	0.0009	0.0013	0.0071	0.0311	0.0611	0.1845	0.3499	0.1513	0.1415	0.0714
R6	0.0011	0	0.0006	0.0094	0.0003	0.0470	0.1497	0.3691	0.4026	0.0127	0.0075

Table B.3: MC probabilities for each lane

Lane	-$700	-$600	-$500	-$400	-$300	-$200	-$100	0	$100	$200	$300	$400	$500	$600	$700
0	**0**	0.0001	0.0024	0.0542	0.2937	0.3343	0.2052	0.0979	0.0119	0.0002	0.0001	0	0	0	0
1	**0**	0	0	0	0	0.0008	0.0511	0.3065	0.4118	0.2052	0.0241	0.0004	0.0001	0	0
2	**0**	0	0	0.0037	0.0799	0.4441	0.4185	0.0530	0.0009	0	0	0	0	0	0
3	**0**	0.0001	0.0130	0.0741	0.1495	0.2079	0.3287	0.1818	0.0409	0.0037	0.0003	0	0	0	0
4	**0**	0	0	0	0.0001	0.0002	0.0119	0.0979	0.2052	0.3343	0.2937	0.0542	0.0024	0.0001	0
5	**0**	0	0	0	0	0.0001	0.0031	0.0433	0.1092	0.2228	0.3146	0.2233	0.0756	0.0079	0.0003
6	**0**	0	0	0.0008	0.0042	0.0582	0.1332	0.2710	0.3240	0.1725	0.0321	0.0032	0.0005	0.0001	0.0002
7	**0**	0	0.0001	0.0004	0.0241	0.2052	0.4118	0.3065	0.0511	0.0008	0	0	0	0	0
8	**0**	0.0003	0.0080	0.0917	0.3358	0.4285	0.1310	0.0047	0.0001	0	0	0	0	0	0
9	**0**	0.0227	0.0708	0.1339	0.2028	0.2421	0.2331	0.0783	0.0129	0.0009	0.0005	0	0	0	0
10	**0**	0	0	0	0	0	0.0009	0.0530	0.4185	0.4441	0.0799	0.0037	0	0	0
11	**0**	0	0	0	0	0	0.0001	0.0047	0.1310	0.4285	0.3358	0.0917	0.0080	0.0003	0
12	**0**	0	0	0	0.0008	0.0140	0.2084	0.4760	0.2593	0.0396	0.0018	0.0001	0.0001	0	0
13	**0**	0	0	0	0.0003	0.0037	0.0409	0.1818	0.3287	0.2079	0.1495	0.0741	0.0130	0.0001	0
14	**0**	0	0	0	0.0005	0.0009	0.0129	0.0783	0.2331	0.2421	0.2028	0.1339	0.0708	0.0227	0.0020
15	**0**	0	0.0003	0.0020	0.0207	0.1118	0.2691	0.2677	0.1822	0.1140	0.0288	0.0032	0.0001	0	0
16	**0**	0	0	0	0.0004	0.0008	0.0107	0.1149	0.3725	0.4140	0.0795	0.0069	0.0003	0	0
17	**0**	0	0	0	0	0.0009	0.0023	0.0315	0.1744	0.3749	0.3015	0.1010	0.0120	0.0016	0
18	**0**	0	0.0001	0.0001	0.0018	0.0396	0.2593	0.4760	0.2084	0.0140	0.0008	0	0	0	0