Fleet planning under Demand Uncertainty: a Reinforcement Learning Approach

Mathias de Koning and Bruno F. Santos[1]

Air Transport and Operations, Faculty of Aerospace Engineering, Delft University of Technology, The Netherlands

Abstract

This work proposes a model-free reinforcement learning approach to learn a long-term fleet planning problem subjected to air-travel demand uncertainty. The aim is to develop a dynamic fleet policy which adapts over time by intermediate assessments of the states. A Deep Q-network is trained to estimate the optimal fleet decisions based on the airline and network conditions. An end-to-end learning set-up is developed, where an optimisation algorithm evaluates the fleet decisions by comparing the optimal fleet solution profit to the estimated fleet solution profit. The stochastic evolution of air-travel demand is sampled by an adaptation of the mean-reversion Ornstein-Uhlenbeck process, adjusting the air-travel demand growth at each route for general network-demand growth to capture network trends. A case study is demonstrated for three demand scenarios for a small airline operating on a domestic US airport network. It is proven that the Deep Q-network can improve the prediction values of the fleet decisions by considering the air-travel demand as input states. Secondly, the trained fleet policy is able to generate near-optimal fleet solutions and shows comparable results to a reference deterministic optimisation algorithm.

Keywords: Fleet Planning Problem, Dynamic Fleet Policy, Deep Q-network, mean-reversion Ornstein-Uhlenbeck

Nomenclature

$\mathcal{T}$ = set of discrete time periods in the finite time horizon T, $\{0, 1, ..., T\}$;
$\mathcal{E}$ = set of episodes, E being the total number of episodes;
$\mathcal{S}$ = set of infinite states representing the airline resources and network conditions;
$\mathcal{R}$ = continuous set of possible rewards, $r \in \mathbb{R}$;
$\mathcal{A}$ = set of actions/fleet decisions, A being the total number of fleet decisions;
$\mathcal{N}$ = set of routes in the network, N being the total number of routes in the network;
$\mathcal{M}$ = set of market in the network, M being the total number of markets in the network;
$\mathcal{K}$ = set of aircraft types, K being the total number aircraft types available;
$\mathcal{I}$ = set of airport in the network, I being the total number of airports available;

[1] Corresponding author - b.f.santos@tudelft.nl

1. Introduction

The fleet planning process of an airline is widely considered the most important long-term decision to ensure future profitability (Belobaba et al., 2009; ICAO, 2017). It is a strategic decision process which is concerned with the acquirement or disposal, the quantity, and the composition of the fleet in future years. In the airline business, fleet planning is considered to be the first step in the airline planning process and addressed via either one of two approaches: the *top-down* or the *bottom-up* approach. The first approach, top-down, estimates the required fleet based on a high-aggregate level analysis. Forecasters estimate the most-likely expected aggregate demand growth and the expected gap in capacity. The future gap in capacity is the amount of aircraft capacity needed to maintain profitable operations. This method is the most common approach in the contemporary airline business as it is a simple method and can be calculated using basic spreadsheets (Belobaba et al., 2009). However, this method is very sensitive to the forecast of aggregated demand; it does not account for the possible deviations from the estimated demand and accompanied fleet.

An alternative method to the top-down approach is the bottom-up approach. This approach uses a detailed modelling method to match the expected demand on a route-level to an operational flight network. By matching the demand and supply on a more detailed level, the fleet size and fleet composition can be modelled more accurately and efficiently. The bottom-up modelling approach can incorporate multiple variables and complexities to represent a more detailed planning model where the fleet decision can be tailored more precisely to the expected demand. Adding these complexities increases the size of the bottom-up models as they grow exponentially with the addition of more variables. Increasing the amount of detail has proven to be a fruitful approach to increase future profits, operational efficiency, and robustness. As a result, this thesis project will elaborate on bottom-up approaches. The remainder of this section will provide an introduction to the fleet planning problem using bottom-up approaches.

Background

In the 1950s Operational Research (OR) gained a lot of momentum as a research field led by the need of industry to optimize production methods to increase the efficient usage of resources (Hillier, 2012). Some of the first authors to address the operational fleet planning were Kirby (1959) and Wyatt (1961). They elaborated on short-term leasing of rail-cars to fulfil the temporal shortage of fleet due to excessive demand. Furthermore, Dantzig and Fulkerson (1954) and Bartlett (1957) investigated the number of ships needed to operate in a fixed schedule. The OR models were still very simple and only considered the fleet planning for a single time period. With the advent of increasing computational power in the 1970s, the size and complexity of the fleet planning problem increased. Shube and Stroup (1975) introduced the first multi-period bottom-up fleet planning model to generate long-term strategic decisions over multiple years using a Mixed Integer Linear Programming (MILP) model. This method is relatively simple to implement and further increases in computational power, meaning larger networks and multiple replacement strategies were able to be modelled using this technique (Bazargan and Hartman, 2012). However, the air-travel demand coefficients are deterministically fixed, and the solution to the problem is therefor an optimal solution for one evolution of demand. Ignoring the stochastic nature of variables such as air-travel demand can result in unfit fleet planning solutions, and thus should be characterized as probability distributions (Kall and Wallace, 1994).

Stochastic modelling allows us to capture one or more variable's stochastic nature into the fleet planning model. In two-stage stochastic models, a sub-set of the decision variables is chosen, and after the uncertainty is revealed the remainder of the variables are determined. A recourse action is performed which can be seen as a corrective measure against the infeasibility that arises due to the realisations of the uncertainty (Sahinidis (2004)). Several publications demonstrate that two-stage stochastic modelling is a suitable approach to model uncertain perspectives into the fleet planning problem. Oum et al. (2000) utilized the two-stage stochastic approach to model the optimal mix between leased aircraft and owned aircraft under demand uncertainty. In a case study, the researches obtained data from 23 different international airlines and proved that a mix of leased and owned aircraft mitigates the cost of an airline under uncertain demand at an optimal mix of around 40-60% leased aircraft of the total fleet. List et al. (2003) employed the stochastic modelling method to create a more robust fleet plan under two uncertain parameters: future demand and future operating conditions. In the recourse function, a partial moment of risk is incorporated to decrease the effect of extreme outcomes of the optimal solution. The authors show that a trade-off between the fleet decisions and the accompanying risk of insufficient resources can induce high cost on airline operations. However, having only considered three markets and a homogeneous fleet, the authors report a high computational effort to compute fleet decisions. Listes and Dekker (2002) argue that fluctuation of demand induces low load factors across the network. They propose a model which tackles the operational flexibility through a demand scenario aggregation based approach to find a fleet composition which appropriately supports dynamic allocation of capacity. The fleet composition problem is modelled as a two-stage stochastic multi-commodity flow problem across a time-space network. Despite its advantages, only one period with a fixed flight schedule is optimized, limiting the planning horizon of usage as a long-term strategic planning tool. Moreover, all aforementioned two-stage stochastic models are limited to a single period planning horizon, and therefor not suitable to create long-term fleet policy.

Multi-stage stochastic programming, such as Dynamic Programming (DP), allows a sequential decision problem to be subdivided into sub-problems which are solved recursively (Bertsekas et al., 1995). In the airline industry, Hsu et al. (2011) use a DP model to simulate the optimal replacement strategy subjected to a Grey topological demand model. The planning model incorporates the trade-off between ownership, leasing of aircraft, and aircraft maintenance to increase the flexibility and better match short-term variations in demand. In a follow-up research Khoo and Teoh (2014) argued that the Grey topological demand model is too limited, as it does not capture disruptive events influencing the air-travel demand. They propose a novel model incorporating the stochastic demand index (SDI) in a step by step procedure using Monte Carlo simulations. In a more recent study, Repko and Santos (2017) employ a different approach to modelling a multi-period fleet plan under uncertainty. The problem is modelled in a scenario tree approach where the nodes represent the fleet decisions and the branches the demand revelations. Using MILP the ideal fleet composition is derived for each scenario, and with the accompanied probability of each scenario, the optimal fleet composition for each period is calculated. Sa et al. (2019) take a more detailed approach on the generation of the fleet compositions and future demand. They propose the generation of a set of demand scenarios which are derived by sampling long-term travel demand from the mean-reverting Ornstein-Uhlenbeck process. Secondly, fleet portfolios are used to compare different fleet composition to demand scenarios. However, as the size of the airline's network, planning horizon, and the number of demand revelations increases, the amount of states grows exponentially. As a result, solving the portfolio, or the scenario tree backwards, becomes too computationally expensive; Powell (2011) refers to this as the *curse of dimensionality.*

Approximate Dynamic Programming (ADP) brings a novel solution to the curse of dimensionality by stepping forward through the scenario tree and calculating the values of the state-action transitions recursively. ADP approximates the value-function of decision nodes in the scenario tree using the Bellman equation (Bellman, 1954). ADP, or often referred to as Reinforcement Learning (RL), has been implemented by multiple research communities (e.g., Control Theory, Artificiality Intelligence, and Operations Research) (Powell, 2009). Consequently, different names are adopted to describe the same process. However, in this paper a distinction is made between the two names: RL will refer to the *model-free* approach and ADP to the *model-based* approach. The ADP method has already proven to be useful in solving resource allocation, and vehicle routing problems in the transport industry. Lam et al. (2007) used ADP to model operational strategies for the relocation of empty containers in the sea-cargo industry. Novoa and Storer (2009) examine ADP approaches in modelling single-vehicle routing problem with stochastic demands. And Powell (2009) uses ADP to build an optimizer simulator for locomotive planning. With ADP, near-optimal solutions are found for the fleet size and planning of the vehicles. The model-free, reinforcement learning, approach has found a recent surge in popularity due to the novel improvement made in long-term decision making. Moreover, the use of deep neural networks as function approximators in RL has proven highly beneficial in term of performance and practicability to learn decision processes (Mnih et al., 2015).

Paper Contribution

This research contributes to the development of dynamic policies by the generation of a model-free reinforcement learning program in a fleet planning environment subjected to uncertainty. This will be achieved by developing a model where (a) an agent learns the optimal dynamic fleet policy, by interaction with (b) an artificially created feedback environment, (c) under uncertain air-travel demand.

The agent gives the optimal decisions in acquiring or disposing a number of aircraft at a given time using the current state and the values of the state-action transition. The environment converts the decision of the agent into the next state and calculates a meaningful reward of the state-action combination in a reasonable time. The agent learns from the experiences saved from previous agent-environment interactions. The agent-environment interaction is submitted to a change in air-travel demand over time which resembles the real-life nature of demand evolvement in the physical world.

The paper offers three main contributions:

1. Although state-of-the-art stochastic modelling methods have previously been used to explore and solve operations optimisation problems, we use a reinforcement learning approach to create a dynamic fleet planning policy. This work is the first to employ a model-free learning algorithm to learn the optimal strategic long-term airline fleet policy under air-travel demand uncertainty.
2. In an End-to-end learning set-up, a neural network is trained using stochastic samples of air travel demand to predict the impact of the fleet decision on the future operational profit. As a result, the trained neural network can be used easily as a predictive model for forecasters and managers without retraining.
3. The work proposes a new sampling strategy of the air-travel demand adapted from the Ornstein-Uhlenbeck forecaster employed by Sa et al. (2019). The air-travel demand growth at each route is adjusted for general network-demand growth to capture network trends.

Report Structure

The remainder of this paper is organized as follows: Section 2 describes and formulates the fleet planning problem as a Markov decision process. Section 3 elaborates on the reinforcement learning process. The air-travel demand sampling using the adapted Ornstein-Uhlenbeck mean reversion process is described in Section 4. Section 5 presents the training environment and reward generation. In Section 6 a proof-of-concept case study for the fleet planning problem is proposed, and in Section 7 the training and testing results are presented. Finally in section 8 the concluding remarks are depicted.

2. Problem Formulation

In the airline business, one of the main factors of success can be measured by matching the supply and demand as closely as possible (Dožić and Kalić, 2015). Consequently, the optimal fleet to fulfil the future air-travel demand across the operating network is crucial to ensure profitable future airline operations. Unfortunately, aircraft are high-capital investments and require intensive usage to capitalize a profit. Moreover, aircraft are not directly at the airline's disposal in times of high demand and must be acquired/disposed long in advance. A careful planning decision process is therefor needed to predict the most feasible fleet decisions based on the possible evolutions of air-travel demand.

The aim is to develop an optimal dynamic solution tool, which allows a re-assessment of the fleet decisions as time progresses. As a result, the fleet planning process can be represented as a sequential decision process with discrete time intervals, formalized as Markov Decision Process (MDP). The fleet planning problem is represented as a finite horizon MDP where at every discrete time-step $t \in \mathcal{T}$, a state $s \in \mathcal{S}$ is observed, a decision-maker takes an action $a \in \mathcal{A}$, and the state transitions to a new state $s' \in \mathcal{S}$ under a stochastic process, and a reward to the decision is generated $r \in \mathcal{R}$. To create the optimal fleet decision plan under the given airline and network conditions, a policy π is developed which represents a probability distribution over the fleet decisions or actions given the state.

2.1. State Space

The state of the fleet planning problem at a given period t is a collection of parameters or features, which holds the information of the airline and network for the decision-maker to estimate the optimal action a_t. In this paper, two ensembles of features are created to train two different policies: a *static* fleet policy and a dynamic fleet policy. The static fleet policy will be trained 'blind' to the air-travel demand. This means that a fleet policy will be generated which is static over time, and independent on the intermediate evolution of demand. The policy is referred to as being Stochastic Static (SS) as it is static and trained on stochastic air-travel demand.

The dynamic fleet policy refers to a MDP where the air-travel demand features are included and the fleet decision are determined based on the current air-travel demand in the network. The policy is referred to as Stochastic Dynamic (SD) due to its dynamic fleet decisions and training on stochastic air-travel demand. The states of the MDP can be defined as:

$$s_t^{SS} = \left(t, ac_{own}^t\right) \tag{1}$$

$$s_t^{SD} = \left(t, ac_{own}^t, d_t\right) \tag{2}$$

Where $t \in \mathcal{T}$ is the period of the fleet planning problem. $ac_{own}^t = \left[ac_{own}^{t,k}\right]_{k \in \mathcal{K}}$ the amount of aircraft k owned in period t, and $d_t = [d_{m,t}]_{m \in \mathcal{M}}$ is a vector listing the market's demand value at period t. The market m replaces two Origin-Destination routes with the same airports, as it is assumed that air-travel demand is similar because the majority of passenger book round trips.

The size of the state vector S^{SS} and S^{SD} is dependent on the number of aircraft types considered and one entry for the time period. For the SD policy, the state vector size is extended with number of markets in the network:

$$S^{SS} = 1 + K \tag{3}$$

$$S^{SD} = 1 + K + M \tag{4}$$

2.2. Action Space

The actions of the fleet planning problem are defined by the decision to either acquire or dispose aircraft, the amount of aircraft, and of which type. The action vector can be defined as:

$$a_t = (ac_{acq}^t, ac_{dis}^t) \tag{5}$$

With $ac_{acq}^t = \left[ac_{dis}^{t,k}\right]_{k \in \mathcal{K}}$ the amount of aircraft k acquired in period t, $ac_{dis}^t = \left[ac_{acq}^{t,k}\right]_{k \in \mathcal{K}}$ the amount of aircraft k disposed in period t. Let's assume that at each stage t for each aircraft type k, aircraft are either acquired, disposed, or no action performed. Furthermore, the amount of aircraft bought or disposed is limited per aircraft type f_{max}^k. The size of the action space initially grows exponentially with increasing amount of aircraft types K and the maximum number of aircraft acquired or disposed per aircraft type f_{max}^k: $A = \prod_{k=1}^{K} 2 \cdot f_{max}^k + 1$. In order to keep the action space from growing to large, it is assumed that only one fleet action is taken for a single aircraft type each period t. The discrete action space and size therefor becomes:

$$\mathcal{A} = \left\{0, [-F^k, F^k]_{k \in K}\right\} \tag{6}$$

$$A = (2 \cdot f^k_{max} \cdot K) + 1 \tag{7}$$

Where $F^k = [1, ..., f^k_{max}]$, and each action $a_t \in \mathcal{A}$ is mapping of the input states.

2.3. Transition function

The transition function defines how the system transitions from state s_t to state s_{t+1}. In the fleet planning problem the demand growth of each market Δ^m_{t+1} is the stochastic variable which defines the demand growth at each market at the next state. As a result, the transition function and next state are defined as:

$$s_{t+1} = \begin{bmatrix} t+1 \\ ac^{t+1}_{own} \\ d_{t+1} \end{bmatrix} = \begin{bmatrix} t+1 \\ ac^t_{own} - ac^t_{dis} + ac^t_{acq} \\ d_t \cdot \Delta_{t+1} \end{bmatrix} \tag{8}$$

where, $d_{t+1} = [d_{m,t} \cdot \Delta_{m,t+1}]_{m \in \mathcal{M}}$. The observed change of demand growth in the market is the realization of the uncertain parameter in the MDP. The evolution of the growth of the demand in each market is the result of independent sampling, and is outlined in Section 4.

2.4. Value-based Optimisation

With MDP tuples the fleet planning problem can be optimized by finding the optimal policy π^* which maximizes the expected reward. The reward r_t is an evaluation of the action a_t given the state s_t, the next state s_{t+1}, and mapped through a reward function $r_t = R(s_t, a_t, s_{t+1})$. The goal of the decision maker is formalized in terms of the reward received (Sutton and Barto, 2018). In Section 5 a more in depth analysis of the reward is given. For now, the discounted return G_t of the finite horizon fleet problem is defined as a sum of all future rewards from t to T of the MDP interactions:

$$G_t = R_{t+1} + \gamma R_{t+2} + \cdots + \gamma^{T-1} R_T = \sum_{k=0}^{T} \gamma^k R_{t+k+1} \tag{9}$$

The discount factor γ $(0 \geq \gamma \geq 1)$ discounts future returns. Returns which are expected to contribute in the future are considered to be less value than current rewards. In order to learn the optimal sequence of actions resulting in the highest cumulative reward, a value function $V^\pi(s)$ is introduced. The value function approximates the value of being in a state s under a policy π as the expected discounted in state s, depicted in Equation 10. From the value function, optimal expected return or optimal value function $V^*(s)$ is defined in Equation 11.

$$V^\pi(s) = E_\pi\left[G_t | s_t = s\right] = \mathbb{E}\left[\sum_{k=0}^{T} \gamma^k r_{t+k} | s_t = s, \pi\right] \tag{10}$$

$$V^*(s) = \max_{\pi \in \Pi} V^\pi(s) \tag{11}$$

The policy which maximizes the value function at state s is the optimal policy π^*. To create meaningful rewards, the goal of the fleet planning process must be first identified. Airlines can choose to optimize their fleet for maximization of transported passengers, adaptability of fleet in the network, minimize cost, etc. In this work it is assumed that the goal is to maximise the profit of future operational years. Thus, the policy represents the fleet decisions of the fleet planning problem which maximizes the airlines' current- and future profits.

3. Deep Reinforcement Learning

To optimise the goal and policy of the fleet planning problem, a learning algorithm is devised. In this chapter, a Reinforcement Learning (RL) process is proposed to learn the optimal policy by iteratively explore the MDP and learn to make the correct sequence of fleet decisions given airline and network conditions. The model-free RL method defines itself by fully separating the transition function from the agent's optimisation process. Contrary to model-based programming, where the transition function is known and used to estimate the optimal policy, the model-free agent is only able to observe the transitioned state after the action is taken.

3.1. Q-learning

Q-learning is the most widely known form of reinforcement learning technique developed by Watkins (1989). The Q-value function $Q(s, a)$ is introduced by extending the value function with the action value. The Q-value function represents the expected value (or total discounted reward) of performing an action a given the state s. An agent updates the Q-table with experiences $< s, a, s', r >$ gathered from interaction with an environment. At every step t, an action is chosen based on the current Q-value function, and the result of that action is observed as an experience. With this experience, the Q-value corresponding to the state-action pair is updated using Equation 12.

$$Q^{\text{new}}\left(s_t, a_t\right) \leftarrow Q\left(s_t, a_t\right) + \alpha \left[r_{t+1} + \gamma \max_a Q\left(s_{t+1}, a\right) - Q\left(s_t, a_t\right)\right] \tag{12}$$

The optimal policy of a Q-learning model can be deducted from the optimal Q-value function as it is the action which maximizes the expected return over time, represented by Equation 13. However, if greedy-policy is maintained throughout the learning process, the policy always exploits the current value function to maximize immediate reward. Higher Q-values remain hidden behind unexplored state-action pairs and the model quickly gets stuck in a local optimum. To incentivize the exploration of unvisited state-actions, an ϵ-greedy algorithm is introduced. At every decision with a probability of ϵ, a random action from the action set equal probabilities is chosen. During the learning process, initially the ϵ parameter will be high to investigate the state-action space, and progressively will decay.

$$\pi^*(s) = \underset{a \in \mathcal{A}}{\operatorname{argmax}} Q^*(s, a) \tag{13}$$

A Q-table provides a simplistic method for assessing and storing the Q-values. However, as the number of actions and states increases, lookup-tables require an increasing amount of memory storage. As a result, more functional approximators are better suited for control problems of modern-day size. Requeno and Santos (2018) observed in a similar fleet planning problem that the relationship between the airline profits and the number of aircraft owned is a non-linear concave function. Consequently, a non-linear function approximator will be most suitable to approximate the value function, such as a neural network.

3.2. Deep Q-network

In deep reinforcement learning the Q-function is approximated by a neural network (NN). The parameters θ of the NN are trained using stochastic gradient descent to minimize a loss function. Deep Q-network (DQN) (Mnih et al., 2015) is a value-learning deep neural network which brought novel solutions to the shortcomings of the original Neural Fitted Q-learning which suffered with slow and unstable convergence properties (Goodfellow et al., 2016). DQN learns the optimal Q-value function with a neural network approximation by minimizing the loss function $L_i(\theta_i)$:

$$L_i(\theta_i) = \mathbb{E}_{<s,a,s',r>\sim\mathcal{U}(\mathcal{M})}\left(r + \gamma Q\left(s', \arg\max_{a'} Q(s', a'; \theta); \theta_i^-\right) - Q(s, a; \theta_i)\right)^2 \tag{14}$$

Next to a policy network with parameters θ, a target network is created with parameters θ^-. The parameters of the target network are used to estimate the expected target values and are updated every C amount of episodes, to minimize the divergence of the estimation and the updated parameters. Secondly, a replay buffer is created where the experiences of the previous iterations are stored. At the time of training, a batch of random experiences are uniformly sampled from the buffer and used to optimize the policy network's parameters. This technique increases the learning speed of the parameters and allows for less variance. Finally, DQN clips the rewards between -1 and $+1$ to ensure more stable learning (François-Lavet et al., 2018).

In Figure 1, a representation of the reinforcement learning model is depicted. At the start of a training episode e, the demand growth d_t is sampled for the periods $T + 1$. This information is available to the environment but not the agent. The reinforcement learning loop is visible as the interaction between the agent (DQN with replay buffer) and the environment. The RL loop is iterated T-times until the finite-horizon is reached. After the RL loop is terminated, if the final episode E is not reached, the RL parameters are reset to the initial state, a new demand growth trajectory sampled, and the process repeated.

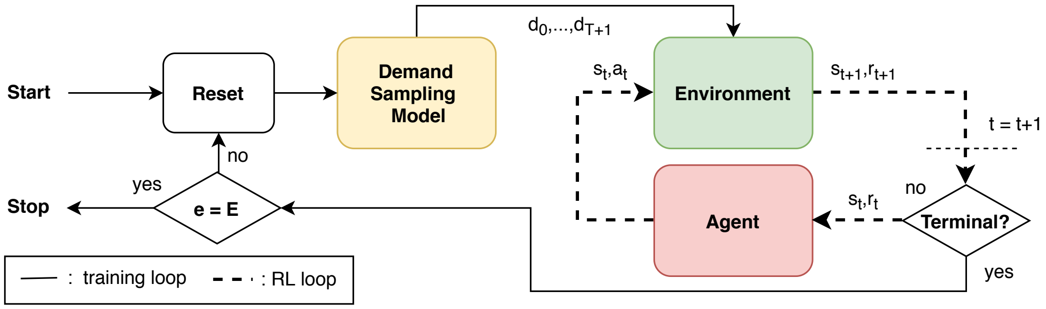

Figure 1: Diagram illustrating the architecture reinforcement learning loop (agent-environment) and the training loop for E episodes.

4. Sampling Stochastic Air-travel Demand

The stochastic nature of the MDP is formalized by a model sampling the air-travel demand values in every episode e. The sampling of a market m is a trajectory of air-travel demand growths from $t = 0, ..., T + 1$, based on the historical characteristics of the air-travel demand.

The mean-reverting Ornstein-Uhlenbeck (OU) process (Uhlenbeck and Ornstein, 1930; Vasicek, 1977) is an autoregression model which has found applications in modeling evolutionary processes in biology (Martins, 1994), diffusive processes in physics (Bezuglyy et al., 2006), and simulating volatility models of financial products Barndorff-Nielsen and Shephard (2001). Recently, using the mean-reverting time-dependent process as a sampling model for demand was successfully implemented by Sa et al. (2019) for a portfolio-based fleet planning model, and will form the basis of our demand sampling model.

4.1. Mean-Reverting Ornstein-Uhlenbeck process

In this section, the OU process as a sampling model for demand growth introduced following the notation Sa et al. (2019). The mean-reverting process is a stochastic differential equation of the growth of the air-travel demand x_t, which can be discretised to estimate the next air travel demand growth x_{t+1}:

$$x_{t+1} = x_t + \lambda\left(\mu - x_t\right) + \sigma dW_t \tag{15}$$

The term $\lambda\left(\mu - x_t\right)$ describes the mean reversion process often called the *drift term.* λ is 'the speed of the mean reversion' describing how fast deviation reverts back to the mean, and μ, 'long term mean growth rate', can be interpreted as the mean air-travel demand growth which the model will approach in the long term. Finally, the process becomes stochastic by the inclusion of $dW = W_{t+1} - W_t$ which has the Normal$(0, 1)$ distribution, and is referred to as the 'shock term'. The term σ influences the impact of the disruptions and can be interpreted as the volatility of the change in the growth of demand.

The μ, λ, σ parameters are referred to Ornstein-Uhlenbeck parameters and are deducted from historic data by approximation of the linear relationship between growth in demand x_t and the change of the growth in demand y_t with linear regression fitting. The linear regression of the historic data x_t, y_t reveals the regression coefficients for the slope b and the intercept a of the fitted data to calculate λ, μ, and σ (Chaiyapo and Phewchean, 2017).

Once the Ornstein-Uhlenbeck parameters are calculated for every route using the historic demand growth, numerous growth trajectories can be estimated for every route by independently and iteratively estimating the next demand growth using Equation 15.

4.2. Adapted Demand Sampling Model

Sa et al. (2019) assume that because air-travel demand tends to correlate to the GDP, the same way the stock of financial markets correlates to GDP, the OU process is a proven predictor of air-travel demand. In the research of Sa et al. (2019), the air travel demand trajectories for multiple routes are sampled independently from each other. Due to the normal distribution of the shock term, the aggregate shock term over multiple routes will converge to zero. As a result, the stochastic element is visible on individual markets; however, this effect is mitigated in the cumulative demand growth of the network and behaves therefor as a linear extrapolation. Secondly, the OU process assumes a fixed long-term drift rate. Consequently, the long-term cumulative demand growth always converges to a single value. The result is a demand trajectory which converges to quasi-identical solution each sampling, and does not fully represent the stochastic behaviour of the real-life air-travel demand.

In this work, it is assumed that an overarching network demand growth exists which influences all routes equally. This growth is influenced by economic growth, fuel prices, airline branch reputation, etc. These factors influence all market demand growths. Secondly, changes in emerging economic markets or demographics of the area locations of airports, influence the willingness to fly certain markets. These factors are specific to independent market growths. Finally, it is assumed that the long-term mean growth is not a fixed value. For the same reason yearly growth changes, long-term mean growth can also diverge from the intended path due to numerous economic, environmental, and political factors.

An adapted demand forecasting model is proposed, where overarching network demand growth prediction δ_{t+1} is added to each market growth prediction $x_{m,t+1}$. Before sampling air-travel demand trajectories on a market level, a cumulative air-travel demand growth trajectory is sampled using the OU parameters derived from the cumulative historic air-travel demand in the network. By averaging the predicted network growth rate (δ_t^n) -which is equal for all markets in period t- and the predicted growth market rate, a semi-independent market demand growth $x_{m,t+1}$ for market m, is sampled:

$$\delta_{t+1} = \delta_t + \lambda\left(\mu' - \delta_t\right) + \sigma dW_t) \tag{16}$$

$$x_{m,t+1} = \frac{1}{2}(\delta_{t+1} + x_{m,t} + \lambda_m\left(\mu'_m - x_{m,t}\right) + \sigma_m dW_{m,t}) \tag{17}$$

$$\begin{aligned} \text{where} \quad & \mu' \sim \mathcal{N}(\mu, \sigma_\mu^2) \\ & \mu'_m \sim \mathcal{N}(\mu_m, \sigma_{m,\mu}^2) \end{aligned} \tag{18}$$

The normal distribution of the long-term mean growth μ' and μ'_m is depicted in Equation 18. Every simulation of an air-travel demand trajectory, a realisation of the normal distribution is drawn. The μ and μ_m represent respectively the estimated long-term mean growth for the network, and for every market m. The σ_μ^2 and $\sigma_{m,\mu}^2$ are the variance terms of the long-term mean growth of respectively the network and markets s. The latter variance terms represent how much dispersion is present in the mean growth over time. However, for the network and every market, only one historical long-term demand is present; hence, it is impossible to calculate the variance. In Section 6 three demand scenarios are represented to test the sensitivity of the variance parameters.

The air-travel demand sampled using Equation 17 is sampled for every time step t which consists of n_{years} years. However, the adapted OU sampling model is discretized in yearly demand growth predictions. As a result, there is a misalignment in time horizon between the RL loop and the demand sampling model. Consequently, the time horizon for the sampling model is changed to $Y = n_{years} \cdot (T+1)$ years, and the periodical change in growth $\Delta_{m,t}$ is related to the yearly demand growth $x_{m,t+1}$ in Equation 19. Note that sampling horizon $(T+1)$ is longer than the time horizon T to accommodate for the revealing of the final air-travel demand d_{T+1} and evaluation of the final action a_T in the RL loop.

$$\Delta_{m,t} = \prod_{i=1}^{n_{years}} (1 + x_{m,y+i}) \quad \text{where} \quad y = t \cdot n_{years} \tag{19}$$

5. Training Environment

The RL agent is tasked with learning the environment's dynamics in order to maximize the expected future reward. The purpose of the environment is thus twofold: transitioning to the next state s_{t+1}, and evaluating the action a_t in the form of a reward r_t. In the fleet planning problem, the environment represents the airline's resources and the air-travel demand network. The transition to the next state is explained in Section 2 and the accompanied revealing of the uncertainty in air-travel demand in Section 4. The remainder of this section will explain the training strategy of the agent-environment interaction and generation of the reward r as an evaluation of the fleet decisions.

5.1. Training Strategy

To successfully train the RL model, a meaningful reward function $R(s_t, a_t, s_{t+1})$ needs to be created which measures the goal of the fleet planning problem: maximise the expected profit. Every fleet decision influences the future profitability of the airline, which can be measured using a Fleet Assignment Model (FAM). The FAM optimizes the airline's profit (C_{FAM}) in period t by matching the flight frequency of aircraft types in the fleet to the sampled air-travel demand. However, the profit under the fleet decision can not be transformed directly into a reward as there is not a frame of reference to say how good or bad this decision was compared to other possible fleet decisions. Moreover, the sampling of demand influences the height of the optimal achievable profit severely, thus making the optimal profit fluctuate throughout the episodes.

An *oracle* is introduced as the optimal fleet decision and profit over the network and period to evaluate the periodical fleet decisions. This larger optimisation algorithm, Fleet Planning Optimisation Model (FPM), optimizes the frequency of flights in conjunction with the optimal fleet. At the start of each episode, after sampling the air-travel demand, the FPM calculates the optimal frequency, fleet composition, and consequently the accompanied profit (C^t_{FPM}) in each period as an upper bound to the FAM.

Figure 2 shows a schematic representation of the full reinforcement learning model with all sub-component interactions. At the start of the training process, the OU parameters are estimated, and the RL model parameters reset to $t = 0$. After initialization (or resetting after each episode), the air-travel demand for each market m is sampled for the full time horizon $T + 1$, resulting in a set of demand vectors $d_0, ..., d_{T+1}$. With the sampled future demand, the optimal fleet and periodical profit C^t_{FPM} can be calculated in the FPM. In addition to this, the optimal fleet decisions can be stored as experiences in the replay buffer to increase the size of the replay buffer and learning experiences of the model.

Now, RL loop is initiated the initial state s_0 is used to predict the first fleet decision a_0 in the policy network. With the fleet decision, and the first vector of the demand matrix d_0, the FAM is optimized for the optimal fleet usage over the network and the profit C_{FAM}. In the reward function, the reward is calculated by comparing the profit of the fleet decision to the optimal profit in the corresponding period C^t_{FPM}. The resulting experience is stored in the replay buffer. The RL model transitions to the next period. If the final period T is reached, or the action a_t is infeasible, the RL loop is terminated and the network parameters are updated.

The dotted lines show the training process of the policy Q-network. For every experience sampled from the replay buffer the DQN loss is computed as depicted in Equation 14, and the weights are optimized by using stochastic gradient descent to minimize the loss. The target values of the loss function are estimated using a one-step Temporal Difference (TD(1)) method, and every $C = 10$ episodes the target network is updated with the learning parameters from the policy Q-network.

If the terminal episode E is reached, then the training process is stopped and the learned parameters of the DQN-model can be used to estimate the optimal dynamic fleet policy.

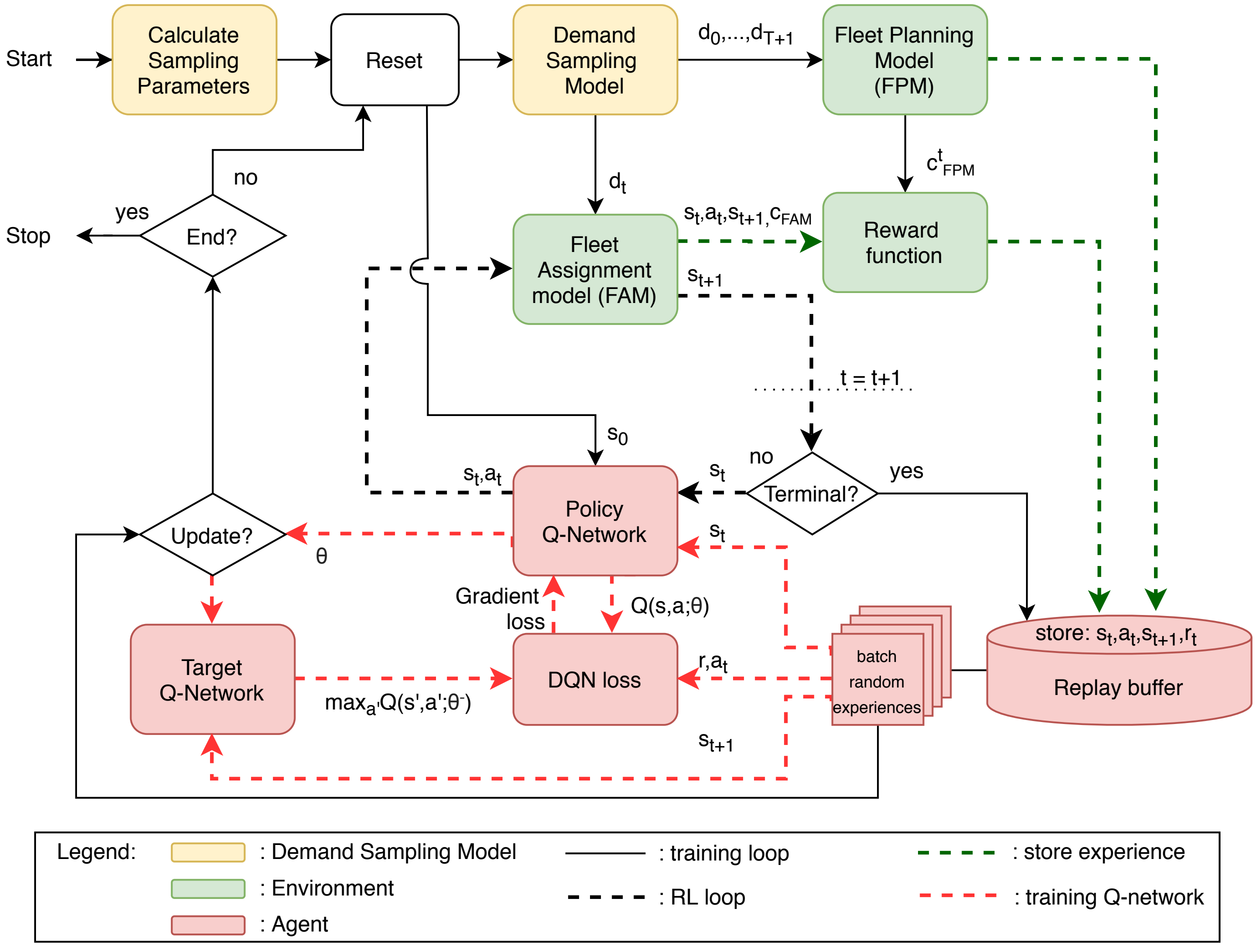

Figure 2: A diagram of the interaction of the Demand Sampling Model (yellow), the Environment (green), and the Deep Q-learning Model (red).

5.2. Fleet Assignment/Planning Model

The Fleet Planning Model is in practice an extension of the Fleet Assignment Model over multiple periods and with the fleet decisions as extra decision variables. A hub & spoke network with the possibility for point to point operations is assumed to be the operational network of the airline. The FPM and FAM are explained in conjunction in the remainder of the section. At every stage, the differences between the two models will be highlighted in the text. The FPM and FAM are both MILP's, which are adapted from Santos (2017).

The decision variables:
There are four types of decision variables: the direct passenger flow, the transfer passenger flow, the number of flights with a specific aircraft type, and the decision variables related to the fleet decisions which include the amount of aircraft owned, amount of aircraft disposed, and the amount of aircraft acquired. The first three types are used in the FAM without the periodical dependency. All decision variables are used in the FPM as described below.

x_{ij}^{t} passenger flow non-stop from origin airport i to destination airport j in period t
w_{ij}^{t} passenger flow from airport i to airport j that transfers at the hub in period t
$z_{ij}^{k,t}$ amount of aircraft operating from from airport i to airport j in period t
$ac_{own}^{k,t}$ amount of aircraft owned of type k in period t
$ac_{dis}^{k,t}$ amount of aircraft disposed of type k in period t
$ac_{acq}^{k,t}$ amount of aircraft acquired of type k in period t

The non-decision variables:
Next to the decision variables, a set of non-decision variables are required to define the airline environment and to create the objective function and constraints. All variables below are the fixed values; meaning they do not vary over episodes e.

c_{var}^{k} variable cost of operating an aircraft of type k per flown km in [dollar/miles]
c_{own}^{k} cost of owning an aircraft of type k each year in [dollar]
c_{dis}^{k} cost of disposing an aircraft of type k each year in [dollar]
n_{week} number of operating weeks per year
n_{year} number of years in one period
OT_{ij} average time to fly leg ij in [hours]
TAT_{k} turn around time of aircraft type k in [hours]
OH_{k} the maximum operation hours of an aircraft type k per week in [hours]
D_{ij}^{t+1} demand between airport i to airport j per year in period $t+1$ [pax]
$dist_{ij}$ distance between airport i to airport j [miles]
s^{k} seats in aircraft type k [pax]
g_{ij} $g = 0$ if a hub is located at airport i, j, $g = 1$ otherwise
R^{k} range of aircraft type k in [miles]
F_{init}^{k} initial owned fleet for type k at initial period $t = 0$

The objective function
As established at the beginning of this section the objective function is to maximise the yearly profit of the airline in future years, which is a combination of maximizing the revenue while

minimizing the expenditures of future operations. The objective function implemented in the FPM is depicted below in equation 20. The revenue is assumed to be solely due to ticket sales on direct and indirect flights. The costs are broken down into operational costs, including those of operating the flights $z_{ij}^{k,t}$, and fixed costs related to the ownership and disposal of aircraft. The fleet costs are unrelated to the actual operational costs of the network, in other words the airline has these costs whether the aircraft is operated or not. The disposal costs of aircraft can be interpreted as fine due to a breach of the lease contract.
The FAM uses a similar objective function. However, the summation over periods is removed, and the fleet costs become a fixed value.

$$\text{Maximized Profit} = \text{Maximized Revenues} - \text{Minimized Costs}$$

$$\begin{aligned}\text{Maximized profit} \;=\;& \underbrace{\sum_{t\in\mathcal{T}}\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{I}}\left[Fare_{ij}\cdot(x_{ij}^{t}+w_{ij}^{t})\right]}_{\text{Revenue}} \underbrace{-\sum_{t\in\mathcal{T}}\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{I}}\sum_{k\in\mathcal{K}}\left[C_{var}^{k}\cdot dist_{ij}\cdot z_{ij}^{k,t}\right]}_{\text{Operational Costs}} \\ & \underbrace{-\sum_{t\in\mathcal{T}}\sum_{k\in\mathcal{K}}\left[C_{own}^{k}\cdot ac_{own}^{k,t}+C_{dis}^{k}\cdot ac_{dis}^{k,t}\right]}_{\text{Fleet Costs}}\end{aligned} \tag{20}$$

The constraints:
The objective function is subjected to a set of constraints:

$$x_{ij}^{t}+w_{ij}^{t}\leq D_{ij}^{t+1} \quad ,\forall i,j\in\mathcal{I},t\in\mathcal{T} \tag{21}$$

$$\sum_{j\in N} z_{ij}^{k,t}=\sum_{j\in N} z_{ji}^{k,t} \quad ,\forall i\in\mathcal{I},k\in\mathcal{K},t\in\mathcal{T} \tag{22}$$

$$w_{ij}^{t}\leq D_{ij}^{t}\times g_{ij} \quad ,\forall i,j\in\mathcal{I},k\in\mathcal{K},t\in\mathcal{T} \tag{23}$$

$$\sum_{i\in N}\sum_{j\in N}(OT_{ij}+TAT_{k})\cdot z_{ij}^{k,t}\leq OH_{k}\cdot ac_{own}^{k,t} \quad ,\forall k\in\mathcal{K},t\in\mathcal{T} \tag{24}$$

$$x_{ij}^{t}+\sum_{m\in\mathcal{I}} w_{im}^{k,t}(1-g_{ij})+\sum_{m\in\mathcal{I}} w_{mj}^{k,t}(1-g_{ij})\leq\sum_{k\in\mathcal{K}} z_{ij}^{k,t}\cdot s^{k} \quad ,\forall i,j\in\mathcal{I},t\in\mathcal{T} \tag{25}$$

$$\sum_{k\in K} z_{ij}^{k,t}\leq a_{ij}^{k}=\begin{cases}1000, & \text{if } dist_{ij}\leq R^{k}\\ 0, & \text{otherwise}\end{cases} \quad , \quad \forall i,j\in\mathcal{I},k\in\mathcal{K} \tag{26}$$

$$ac_{own}^{k,t}+ac_{dis}^{k,t}-ac_{acq}^{k,t}=ac_{own}^{k,t-1}, \quad \forall t=(1,\ldots,T),k\in\mathcal{K} \tag{27}$$

$$ac_{own}^{k,0}+ac_{dis}^{k,0}-ac_{acq}^{k,0}=F_{init}^{k}, \quad \forall k\in\mathcal{K} \tag{28}$$

$$x_{ij}^{t}\in\mathbb{R}^{+},w_{ij}^{t}\in\mathbb{R}^{+},z_{ij}^{k,t}\in\mathbb{Z}^{+},ac_{own}^{k,t}\in\mathbb{Z}^{+},ac_{dis}^{k,t}\in\mathbb{Z}^{+},ac_{own}^{k,t}\in\mathbb{Z}^{+} \tag{29}$$

The first set of Constraints (21) dictates that the sum of direct and indirect passengers transported over a route ij does not exceed the demand in period $t+1$. The demand matrix $D = \{d_1, ..., d_{T+1}\}$ is shifted one time step. As these sets represent the revealed air-travel demand in the environment. The set of Constraints (22) ensures the aircraft balance in the airports because the operated network is repeated on a weekly basis. At the beginning and end of each week, the amount of arrived aircraft in each airport must equal the amount of departed aircraft. The set of Constraints (23) allows indirect passengers only to be transferred via the hub airport. The fourth set of Constraints (24) limits each type of aircraft k to be operated no more than the maximum allowance of operating hours per week. The amount of aircraft owned of type k will define how many times routes can be operated. The set of Constraints (25) ensures that the amount of non-stop and transfer passengers in each route is lower than or equal to the maximum amount of seats on a route. Similarly to (24), the amount of available seats is dependent on how many aircraft of type k operate that route. The range of Constraints (26) ensures that flights can only be operated by aircraft types which have sufficient range capability. The set of Constraints (21,22,23,24,25,24,26) are used in both the FPM and FAM, although the dependency of periods is removed in the FAM. Because every set of constraints is repeated for $t \in \mathcal{T}$ in the FPM, the FAM will have $T+1$ times fewer constraints.

The set of Constraints (27 and 28) are added to the FPM to extend the single period fleet planning model to a multiple period fleet planning model. The set of Constraints (27) dictate that the amount of aircraft owned, plus aircraft disposed, and minus aircraft acquired in period t must equal the aircraft owned in the previous period $t-1$. Constraint (28) initializes the period zero with the initial fleet F^k_{init}.

To conclude, the integrality and non-negativity Constraints (29) are depicted. Both of the decision variables related to the transported passengers (x_{ij} and w_{ij}) are assumed to be positive real numbers. As a result, these variables are continuous, because the influence on the profit is marginal, and the computational load is decreased. The decision variables related to the aircraft fleet and flight frequency are positive natural numbers.

5.3. The reward function:

The set of solutions $(C^0_{FPM}, ..., C^T_{FPM})$ for every optimal action a^*_t in $\mathcal{T}$, represents the upper bound of the profit C_{FAM} for an air-travel demand trajectory $\{d_0, ..., d_{T+1}\}$ in an episode. In the RL loop, after the agent's fleet decision, the FAM optimizes profit (C_{FAM}) given the fleet decision a_t and revealed demand d_{t+1}. The reward function transforms the calculated profit into a reward.

If the profit of the FAM is equal to the profit of the FPM ($C^t_{FPM} = C_{FAM}$), the fleet decision a_t of the agent is the optimal fleet decision a^*_t. Consequently, the accompanying reward is $+1$. However, a meaningful reward needs to be established for all sub-optimal fleet decisions and profits. Obviously, the worst decision should receive a reward of -1, but calculating the profits to all possible fleet decisions in the action space $\mathcal{A}$ to establish a profit range is a cumbersome task that is too computationally expense. Moreover, the worst fleet decision in the action space $\mathcal{A}$ can often yield a profitable airline. Hence, a lower bound $(lb) \in (0,1)$ is established which is a percentage of the optimal profit. The rewards of the sub-optimal fleet decisions are obtained by mapping the profit C_{FAM} on a linear function between the optimal profit (C^t_{FPM}) where $r_t = +1$, and lower bound ($C^t_{FPM} \cdot lb$) where $r_t = 0$.

If the profit C_{FAM} is lower than $C^t_{FPM} \cdot lb$, the reward immediately becomes zero. Finally, the agent is punished for fleet decisions which are infeasible. If the fleet decision by the agent disposes more aircraft than available, the reward is (-1).

$$r_t = \begin{cases} -1, & \text{if } ac^{k,t}_{own} - ac^{k,t}_{dis} < 0 \\ \dfrac{C_{FAM} - lb \cdot C^t_{FPM}}{(1 - lb)\, C^t_{FPM}}, & \text{elseif } C_{FAM} > lb \cdot C^t_{FPM} \\ 0, & \text{else} \end{cases} \tag{30}$$

By increasing or decreasing the lower bound the aggressiveness of the rewards can be tuned. However, the learning of the agent can be very sensitive to the lower bound. If the lower bound is too low, all fleet decisions will have a reward close to +1, and the agent will be unable to learn the optimal policy. If the lower bound is too high, the rewards will be very sparse. As a result, the agent would have little rewards to learn from and have trouble converging to an optimal policy. The lower bound is therefor a hyperparameter of the model which is tuned to achieve the best learning.

6. Experimental set-up

6.1. Case Study

As proof of concept, a case study is conducted using the proposed methodology. The aim is to mimic the real life fleet planning process of a small airline operating on a domestic airport network in the United States (US). Ten major US airports are included in the network comprising 90 possible routes. Two aircraft types will be considered in the case study: A Boeing 737-800 (BOE738) and a Boeing 757-300 (BOE753). These aircraft are typical examples of narrow-bodies aircraft commonly chosen by airlines to operate shorter domestic flights. A planning horizon of 10 years is assumend with a fleet decision every 2 year resulting in 5 time periods for the RL loop. In Table 1 the case study parameters are depicted.

Table 1: Case study parameters

Notation	Definition	Value
E	# Episodes	5000
T	# Time horizon	5
Y	# Planning horizon	10
N	# Routes in network	90
M	# Markets in network	45
K	# Aircraft types	2

In Table 2, the aircraft parameters are displayed for two aircraft types commonly operated in domestic networks. The BOE738 is a newer narrow-body aircraft with a higher ownership cost. The BOE753 is an older type of aircraft thus having a lower ownership cost, yet it is more expensive to operate per flown mile due to higher fuel costs. There is not a high initial cost for the acquirement of the aircraft, as it is assumed the aircraft are leased on a yearly basis. If a lease contract is broken, the airline is assumed to pay an extra year in the form of a disposal cost. In addition, it is assumed that all flights have a load factor of $LF = 85\%$, and the aircraft operations are continued for $n_{week} = 50$ per year.

Symbol	s^k	v_c^k	$range^k$	OH^k	TAT^k	c_{var}^k	c_{own}^k	c_{dis}^k
Units	$-$	$\frac{hour}{week}$	$miles$	$\frac{hour}{week}$	$hour$	$\frac{USD}{mile}$	$\frac{USD}{year}$	$\frac{USD}{year}$
BOE738	162	543	3582	77	1	0.13	3.05E+06	3.05E+06
BOE753	243	530	3357	80	1.5	0.14	2.4E+06	2.4E+06

Table 2: Aircraft-related parameters

6.2. Demand Model parameters

The air-travel demand on routes is very difficult to measure as it is dependent on various parameters such as the air-fare, time of year, special events, etc. Except from surveying, the only data which resembles the air-travel demand on routes is the number of passengers who actually travelled. Consequently, it is assumed that the historical transported passengers represents the historical air-travel demand, and can be used as a predictor of future air-travel demand values.

The historical travelled passenger data is extracted from the Bureau of Transportation Statistics (BTS), part of the United States Department of Transportation (US DOT). In the TransStat database, the T-100 Domestic Market (U.S. Carriers) database contains the historical monthly market data of all US airlines. It is important to note that the historical travelled demand data is the market data between two airports, meaning it contains all passengers transported between two airports directly and indirectly by all US airlines.

With the adapted demand sampling model, infinite variations of demand matrices based on estimated parameters can be sampled. The variance of the shock terms ($dW_t, dW_{m,t}$) are defined by the Wiener process and therefor initially defined as N $\sim$ $(0, 1)$. However, the estimation of OU parameters on the markets shows very high historical estimation errors σ due to historical variability in transported passengers. As these estimation errors produce unreasonable shocks and demand growths in the prediction of air-travel demand, the standard deviation σ of the normal distribution's disruption effect is lowered my multiplying it with a predefined smoothing factor $\eta \in (0, 1)$ and $\eta_m \in (0, 1)$ for respectively the network and market prediction. In Equation 31 and 32, the adaptations for the Ornstein-Uhlenbeck sampling model are visualized.

$$\delta_{t+1} = \delta_t + \lambda \left(\mu' - \delta_t\right) + \eta\sigma dW_t) \tag{31}$$

$$x_{m,t+1} = \frac{1}{2}(\delta_{t+1} + x_{m,t} + \lambda_m \left(\mu'_m - x_{m,t}\right) + \eta_m\sigma_m dW_{m,t}) \tag{32}$$

The normal distributed long-term growth rates ($\mu' \sim \mathcal{N}(\mu, \sigma_\mu^2), \mu'_m \sim \mathcal{N}(\mu_m, \sigma_{m,\mu}^2)$), replace the previously fixed long-term mean growth rates. In Section 4, it was established that the variance terms ($\sigma_{\mu g}^2, \sigma_\mu^2$) define the amount of dispersion of the long-term mean growth, and are difficult to define because of the lack of data. Due to the uncertainties in the variance terms of the long-term mean growth and the smoothing factor of the shock term, a set of demand scenarios generated to observe different sampling behaviours and investigate the sensitivity to the DQN-model. Three demand scenarios are generated: Average Demand (AD), Dominant Network Demand (DND), and Dominant Market Demand (DMD). In Table 3 the variance values and smoothing factors for the demand scenarios are depicted.

	η_m	η	$\sigma^2_{m,\mu}$	σ^2_{μ}
Average Demand	0.5	1	0.005	0.005
Dominant Network Demand	0.1	1	0.005	0.05
Dominant Market Demand	1	0.1	0.05	0.005

Table 3: Smoothing factor of the shock term and Variance of long-term mean growth for the three demand scenarios: Average Demand (AD), Dominant Network Demand (DND), Dominant Market Demand (DMD).

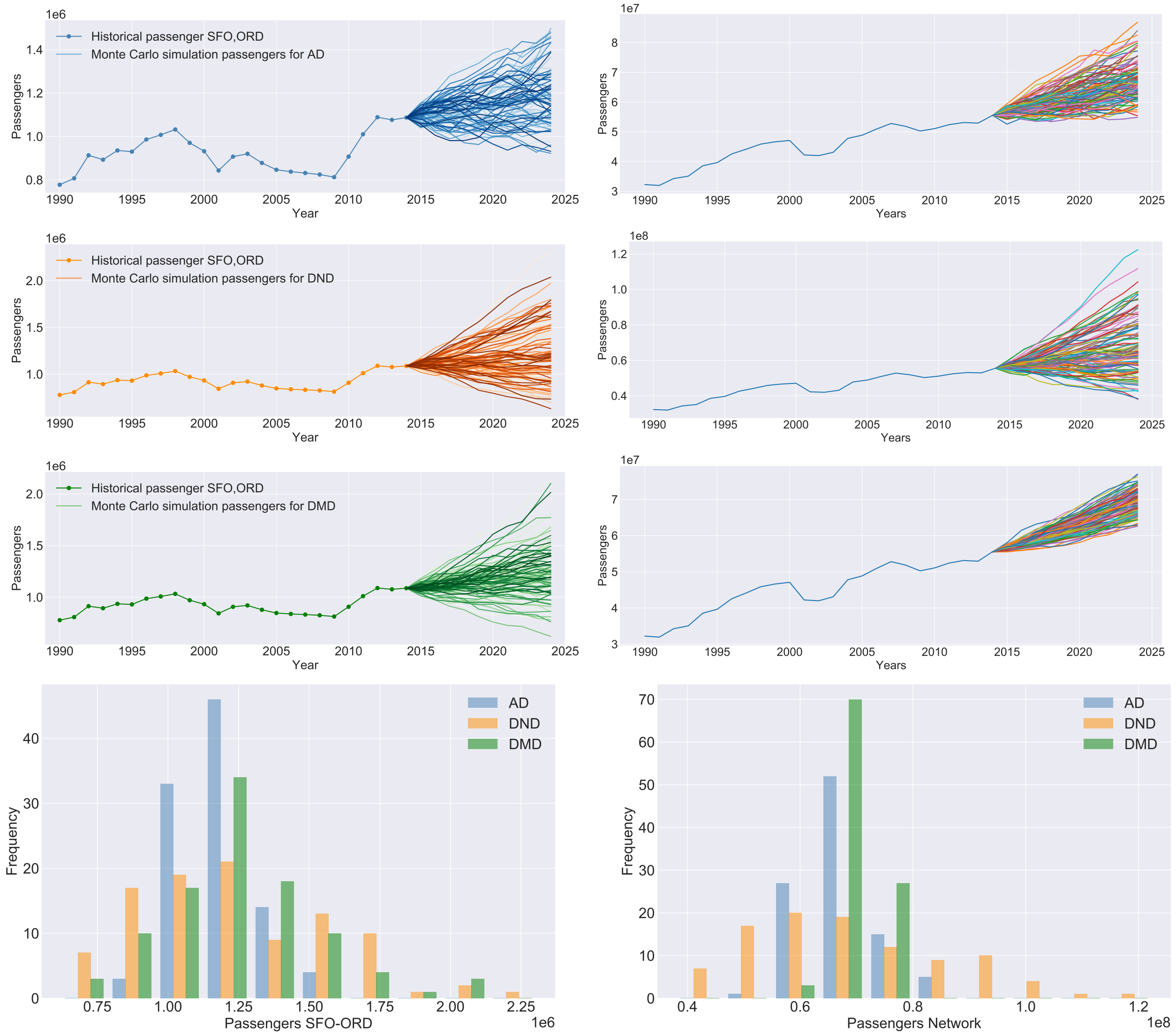

Figure 3: Simulations of stochastic demand trajectories for the market SFO-ORD for the three demand scenarios (left), and the corresponding cumulative network demand for the three demand scenarios (right); the histograms show the distributions of the predicted demand in period $T + 1$.

Figure 3 shows a simulation of the 50 trajectories sampled using the three different demand scenarios. The top row shows the simulation of the AD scenario, the second row the simulation of the DND scenario, and third row the simulation of the DMD scenario. Finally, the histogram on the bottom row shows a comparison of the demand predictions in period

$T + 1$ for the three demand scenarios. The left side shows the simulation of the market SFO-ORD, the right plots show the cumulative simulated demand in the network.

The AD scenario employs a smoothing factor $\eta_m = 0.5$ to compensate for the high shock terms in the market scenarios. Because, the historical network growth is a summation of multiple historical markets, the estimation error for the network σ is lower and does not require smoothing. The variance terms of the network and markets are assumed to equal 0.005, which translates in a long-term deviation of the mean growth of 0.5% for 68% of the demand trajectories. The latter parameters showed the most reasonable stochasticity and growth behaviour outlined in Figure 3. Both the market demand and the network demand shows a long-term mean growth which is reasonably distributed, and a stochastic behaviour both on a market level and a network level.

The DND scenario employs a higher smoothing factor for the market and a higher variance for the network long-term mean growth. The result depicted in Figure 3 (row 2). This demand scenario has a very strong stochastic behaviour and high dispersion on the network's long-term mean growth sampling, contrary to the market's long-term mean growth sampling. As a result, the market's growth is dominated by the network long-term mean growth sampling. The sampled cumulative demand trajectories behave less stochastic than the AD scenario, but the spread of the air-travel demand at the modelling horizon is large due to the increase in variance in the normal distribution sampling of the long-term mean growth rate of the network σ_μ^2.

The DMD scenario samples a future air travel demand which is highly market dependent and has little influence on network growth. The smoothing factor of the market is low and the variance is high, contrary to the network high growth smoothing factor, accompanied by a low variance term. As the influence from the network growth and stochasticity is very limited, this demand scenario resembles the use-case of Sa et al. (2019). In Figure 3 (row 3) an example of 50 sampled demand predictions of the route SFO-ORD (left), and the resulting cumulative network demand (right) is depicted. It is clearly visible that the divergent and stochastic behaviour of the market demand samples is mitigated in the cumulative network demand trajectories.

6.3. Hyperparameter Tuning

The Q-function approximator of the DQN is a fully connected feed-forward neural network. To estimate the appropriate fleet action, the state of the environment s_t is normalized, and fed to the input layer of the neural network. The input vector consists of 48 state values for the SD policy and 3 state vectors for the SS policy, and every value corresponds to a single neuron. The input layer is fully connected to two sequential hidden layers, each with 64 neurons, with a (non-linear) *elu* activation function. After the two non-linear hidden layers, an extra hidden layer with a linear activation of the weights is added. When performing a regression problem, such as approximation of the value of the action in the output layer, a conversion of the non-linear outputs of the elu layers to a real value is needed (Goodfellow et al., 2016). The final hidden layer is fully connected to the output neurons, which correspond to the size of the action vector. It is important to note that the DQN-model was tuned for the SD policy only, and for both policies the same hyperparameters were used. In Table 4 all the relevant hyperparameters of the DQN are shown:

Hyperparameter	Value
Multi-step returns	1
Learning rate [α]	0.001
Discount rate [γ]	0.95
Exploration [ϵ]	$1 \rightarrow 0.05$
Maximum memory length	100K experiences
Batch size	64 experiences
Hidden layers	3 layers
Dense size	64 neurons
Lower bound [lb]	0.985

Table 4: DQN hyperparameters

6.4. Training methodology

The RL model is run for $E = 5000$ episodes, which was determined iteratively to be a sufficient length to learn the fleet policy. Every episode, five periods are considered as five fleet decision points, and consequently RL loop is represented by five agent-environment iterations. Every period is assumed to consist of $n_{year} = 2$ years. At the beginning of each period a fleet decision is taken, with an assumed delivery time of one year. Consequently, the new fleet composition is assessed (the reward) in the two consecutive years after the first year has ended, and the aircraft delivered. For a five period time horizon with two years in each period, 11 years need to be sampled. As a result the sampling horizon becomes $Y = 1 + (n_{\text{years}} \cdot (T + 1))$. In Figure 4 a schematic representation demand revelation with actions and rewards for a finite horizon of 5 periods is illustrated.

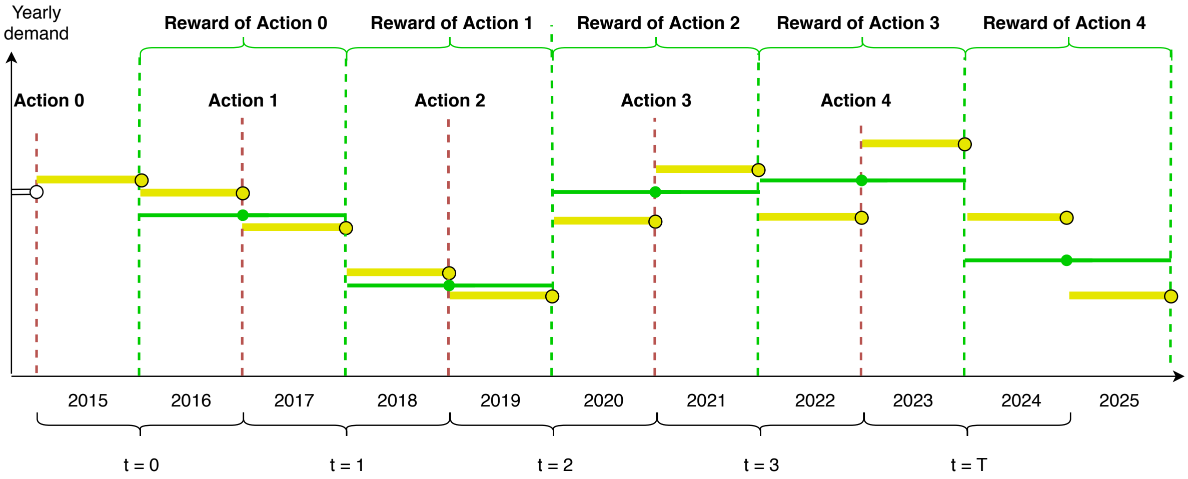

Figure 4: Example of one episode of sampled yearly demand (yellow dots) over a finite-horizon with fleet decisions (red) and calculation of the corresponding reward (green).

During the RL loop experiences are stored in the replay buffer. If the time horizon is reached, the policy NN is trained on the 64 randomly selected experiences from the replay buffer. The generation of experiences is computationally expensive due to optimisation of the FPM in the RL loop. In order to increase the efficiency and sample generation, at the start of the learning process, multiple RL loops are run for one demand sample to increase the exploration of the state-action space and the usage of the FPM optimisation. In Figure 5 it is visible that at the start of the learning process five iterations of the RL loop are performed which decreases exponentially to one iteration from episode 2500 to the end.

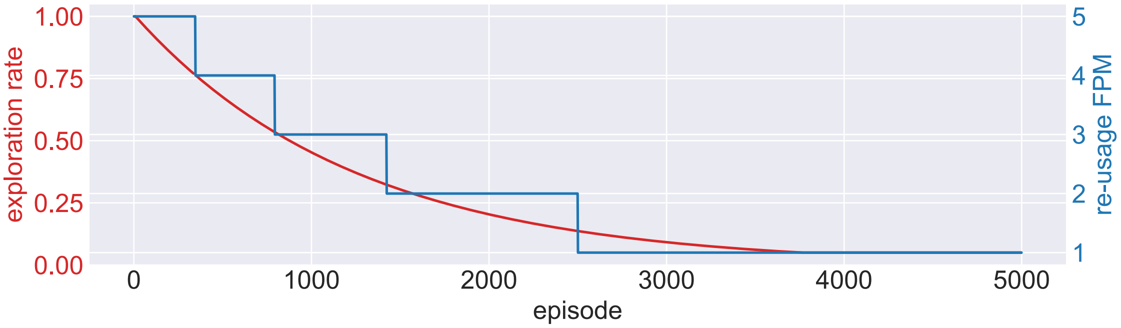

Figure 5: Exploration rate (red), and the amount re-usages of the FPM in one episode (blue).

Every episode the FPM optimizes the optimal weekly operational network, with the fleet evolution and the accompanied profit. As the number of considered airports in the network increases, so does the computational time of the optimisation algorithm. Consequently, the training becomes a very slow process, as the optimisation algorithm struggles to decrease the optimality gap between the primal and dual problem of the MILP. A solution is to increase the MIP gap and let the optimisation stop earlier. As a result, during training, the MIP gap of the FPM and FAM is set to $1E-3$. A side-investigation showed that the average impact of decreasing the gap from $1E-4$ to $1E-3$ on the optimal solution's reward function is less than 1%. This error-margin is assumed to be acceptable.

The performance of the training process is measured by the training score. The training score is measured by the average reward in the RL loop, by dividing the accumulated reward over five periods by the optimal achievable accumulated reward $R_{optimal} = 5$. The final 25% of the episodes, at every target NN update, the trained model is validated using a validation set of 25 air-travel demand evolutions. The validation score of the NN shows an unbiased evaluation of the model performance on the training data. Similar to the training process, the RL loop is run and the agent predicts the optimal fleet decision for each period t; however, the air-travel demand is not sampled randomly iteratively inserted from the validation set. The accumulated reward divided by $R_{optimal} = 5$ to obtain the validation score of one demand trajectory; this process is repeated for all 25 demand evolutions to calculate an average and variance of the validation score for a NN.

After the training process, the *Average Validation Score* and *Average Variance Validation Score* shows the average score and average variance over all the validated NN. The latter is an indication of the average variance of the NN's performance on the validation set. *Variance of Average Validation Score* shows the variation of the *Average Validation Score* and is and an indication of the performance dispersion of over the trained NN's. Finally, the trained NN with the highest score on the validation set is selected as the optimal fleet predictor based on the best training score and is then used for evaluation.

7. Result Analysis

7.1. Evaluation methodology

To evaluate the trained fleet policies SS and SD, two conventional fleet planning policies, referred to as the Deterministic Static (DS) and Deterministic Dynamic (DD), are developed. The DS and DD policies represent the fleet planning problem as a deterministic bottom-up approach and utilize a deterministic sampling of the air-travel demand to optimize an FPM and predict the optimal fleet composition. The deterministic demand of the two models is estimated using the adapted Ornstein-Uhlenbeck process without the stochastic shock and stochastic drift term. The result is a quasi-linear extrapolation of the historic long-term mean growth. An example of such a sampling process is visible in Figure 6.

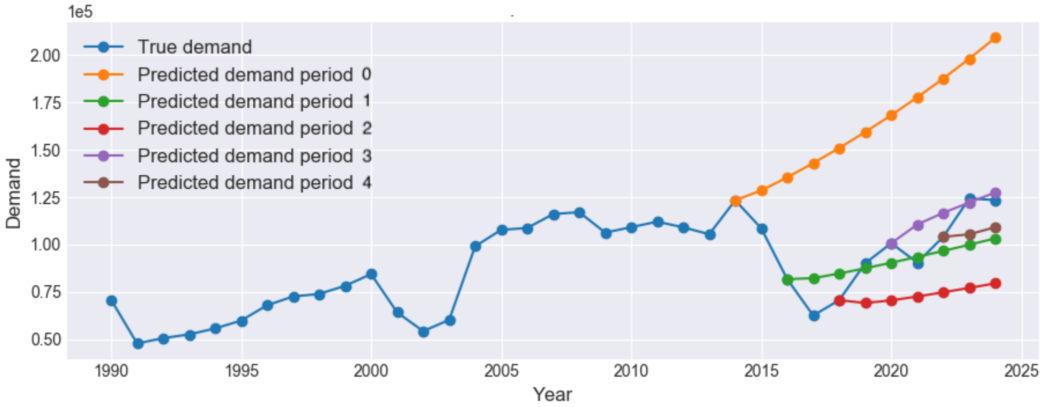

Figure 6: Example of a stochastic demand sampled for the market 'SFO-ORD' (blue), and intermediate non-stochastic demand samples throughout the episodes.

Here, the blue line consists of the historic yearly demand and a stochastic sampled demand from 2014, representing the true demand evolution of an episode. When the first fleet decision is taken in $t = 0$, the demand is predicted with the deterministic sampling policy, and the FPM calculates the optimal fleet decisions for the next 5 periods. The DS policy applies the 5 fleet decisions calculated at period $t = 0$ as a static policy, and follows that policy without adaptations to the revealed demand values. On the contrary, the DD policy allows for the fleet policy to be updated dynamically. At every period, after the true demand values of the previous period are revealed, a new deterministic demand evolution is predicted as well as an FPM optimized, spanning the remaining periods, to obtain the dynamic fleet decisions.

The DS policy is comparable to the Stochastic Static (SS) policy as both policies are not dependent on the evolution of the sampled demand. Both static policies represent a fleet planning process where a long-term fleet policy is generated which is not updated over time. The DD policy is comparable to the Stochastic Dynamic (SD) policy as both the policies represent a fleet planning process where the optimal fleet plan is re-optimized over time to generate a dynamic long-term fleet policy.

To evaluate and compare the stochastic and deterministic models three evaluation-sets (one for each demand scenario) are constructed containing 50 air-travel demand predictions including the optimal fleet decisions and profit. The trained fleet policy showing the highest Validation Score is utilized as the optimal network and is compared to their deterministic counterparts. Similarly to the validation process the *Evaluation Score* and *Variance Evaluation Score* are calculated for the evaluation data-set of 50 samples.

7.2. Average Demand

Figure 7 shows the training score and validation score of the two stochastic models. It can be noted that the training score suffers from high variance. This is attributed to the assumption that the MDP is a fully observable process. The Markov Property assumes that the current state is a sufficient statistical of the future Silver (2018). It can be argued that the current state of the air-travel demand is not predictively sufficient of future observation of air-travel demand. This is because the action of a state is evaluated after two revelations of growth in air-travel demand, and can abruptly change due to the shock term of the OU sampling process. The prediction of the action is therefor the most-likely action under the probability distributed evolution of the demand. However, the 'noise' due to unpredictable and volatile revelations of air-travel demand induces a large amount of variance in the model.

Due to the variance, the training score and validation score is smoothed out using a moving average over 20 samples to visualize the trend. In the early stages of training, the training score is low for both stochastic models. As the number of episodes increases and the exploration decreases, the NN of the DQN's are trained and more increasingly the greedy policy is employed which leads to an increasing training score. It is clearly visible that the inclusion of the current demand for markets increases the training score of the DQN-model; the validation score confirms this behaviour. This is an indication that the SD NN can detect patterns in the demand values, and the dynamic policy chooses more optimal fleet decisions compared to the static policy.

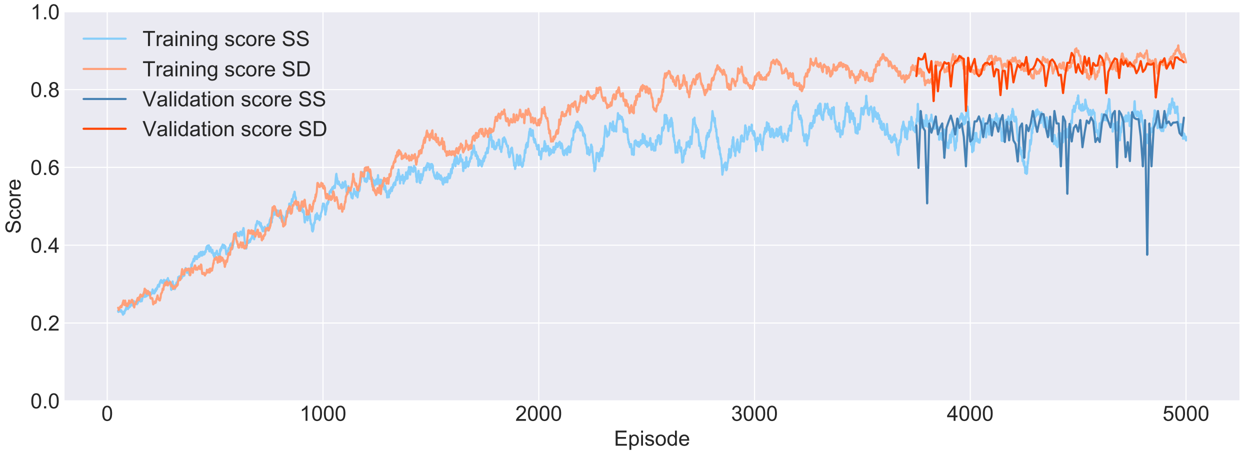

Figure 7: Training score and Validation Score of SS and SD policy network for Average Demand scenario.

In Table 5, the validation and evaluation scores are displayed. The Average Validation Score shows that the stochastic models perform slightly sub-optimal to their deterministic counterparts. Again it is clearly visible that the dynamic policies out-perform the static policies. Moreover, the Variance Validation Score and Variance of Average Score improves with the addition of intermediate evaluations of the fleet decision based on the air-travel demand features (SD policy).

The Evaluation Score shows better results than expected w.r.t. the Average Validation Scores. The SS policy outperforms the DS counterpart both on Evaluation Score and Variance Evaluation Score. The SD policy performs not considerably worse than the DD benchmark for both the Score and Variance.

	DS	SS	DD	SD
Average Validation Score	73.68%	69.44%	88.86%	85.68%
Average Variance Validation Score	5.762	4.72	0.906	1.16
Variance of Average Val. Score	-	25.68	-	6.656
Evaluation Score	73.86%	74.86%	90.59%	88.81%
Variance Evaluation Score	6.615	6.597	0.522	0.691

Table 5: Average Validation Score of Average Demand scenario

A similar behavior is visible in Figure 8 and Figure 9. The relative testing error shows the percentage difference between the profit using fleet prediction models (C_{FAM}), and the optimal (true) profit solution (C^t_{FPM}). In all four policies, the relative errors from the optimal solution are equal for the first period. The reason behind this is that the initial state is always the same state, and as a result, the policies will always predict the same fleet decision.

The SS policy shows the exact predictions and errors as the DS in periods one through three. However, in the fourth period, the mean and spread decrease slightly; wherein the fifth period, the spread of the error of the SS policy starts to increase again compared to the DS. Overall, the SS policy shows slightly better results than the DS policy.

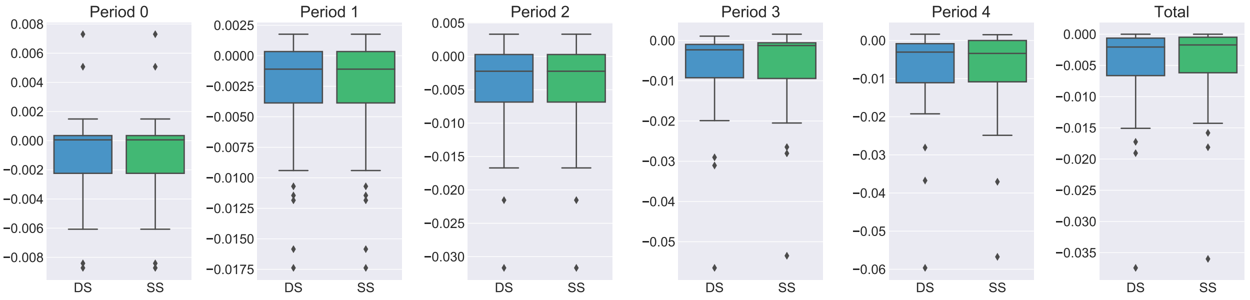

Figure 8: The relative testing error of the Deterministic Static (DS)(blue), and the Stochastic Static (SS)(green) policy to the optimal solution profit for Average Demand scenario.

Figure 9 shows the comparison of the dynamic policies. Where the stochastic policy shows improved prediction for periods three and four, the deterministic policy performs better in period two and five. The difference is more pronounced in the comparison of the total profit errors where DD outperforms SD. However, the range of the y-axis reveals that the difference in error is small therefor both policies show similar results.

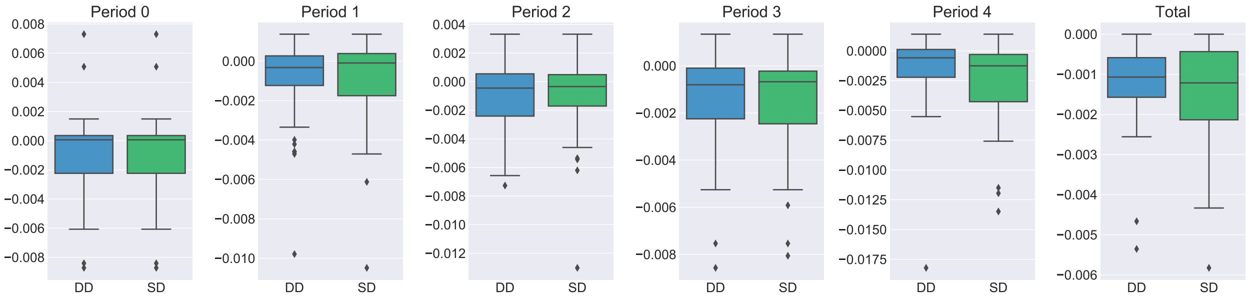

Figure 9: The relative testing error of the Deterministic Dynamic (DD)(blue), and the Stochastic Dynamic (SD)(green) policy to the optimal solution profit for Average Demand scenario.

7.3. Dominant Network Demand

In Figure 10, the training score and validation score of the two stochastic models are presented. Both the SS and SD training score increases over the episodes but show a much lower score compared to the training score of the AD scenario (Figure 7). Moreover, the difference between the dynamic and static training score is much higher than in the AD scenario. This behaviour is attributed to the DND scenario's air-travel demand which -at the end of the modelling horizon T- has much larger dispersion than the AD scenario. As a result, the static policies, which are not updated iteratively on the evolution of air-travel demand, perform inferior to the dynamic counterpart. Moreover, the experiences at the end of the time horizon will have very contradicting actions to the same state, thus inducing a lot of noise and training variance. This is visible in the validation score of the SS method which shows a high variance compared to the SD approach (Variance of Average Val. Score).

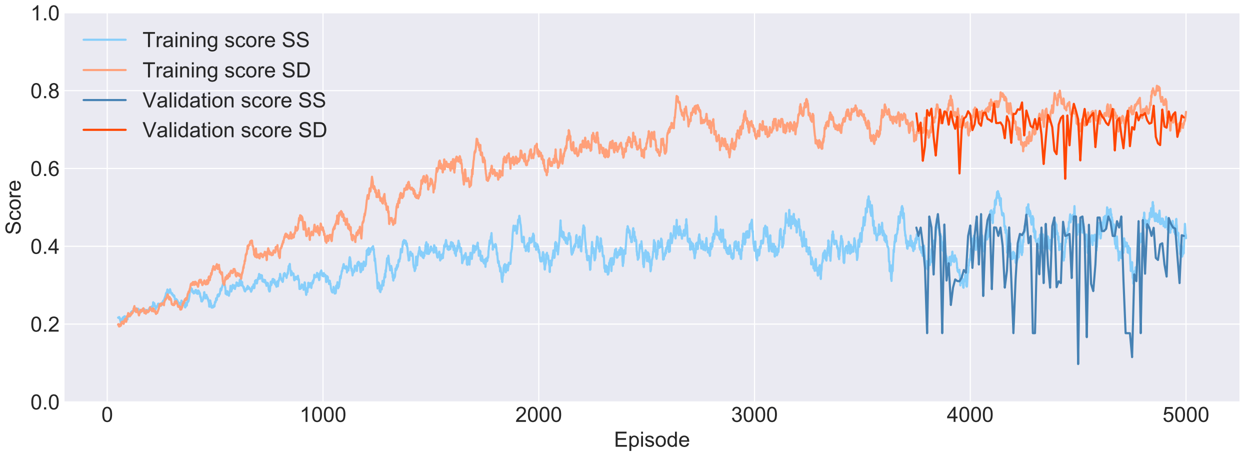

Figure 10: Training score and Validation Score of SS and SD policy network for Dominant Network Demand scenario.

In Table 6 the validation and evaluation scores and variances are presented. Next to the high Variance of Average Validation Scores, also the Average Variance Score is notably larger than in the AD scenario. The Average Validation Score of the DS approach seems to outperform the average validation score of the SS method by 10%. However, the Evaluation score of SS outperforms the DS evaluation score, but with a higher variance. This again shows high variance in the NN training and dependence on the model selection as the final predictor of the fleet.

The dynamic policies too show a high variance as well as lower validation and test scores compared to the AD demand scenario. However, the SD trained NN outperforms the DD policy both in the validation- and evaluation score and variance. Especially the Average Variance Score is notable lower for the SD policy. Because the DD policy samples the demand based on the average growth of the historic transported passengers (as depicted in Figure 6), it overestimates (or underestimates) the mean growth of demand in the extreme demand trajectories. The SD policy adapts better to the demand trajectories and is able to outperform the DD policy both on evaluation and validation score.

	DS	SS	DD	SD
Average Validation Score	47.24%	37.84%	64.91%	71.39%
Average Variance Validation Score	12.363	8.26	7.63	3.35
Variance of Average Val. Score	-	88.027	-	13.761
Evaluation Score	43.09%	46.45%	68.44%	71.2%
Variance Evaluation Score	7.491	8.075	5.471	3.557

Table 6: Average Validation Score of Dominant Network Demand scenario

In Figure 11 and 12, the most notable trend is the lower variance of the SS and SD policy in the early stages of the planning horizon. The lower variance is a direct result of training the stochastic policy on sampled evolutions of demand. Training on this evolutions yields initial fleet decision which is better over a wider range of demand evolutions. Towards the end of the planning horizon, the error unfortunately increases for the SS policy. As a result, the total relative error from the optimal solution is very similar for both static methods with a slightly increased variance for the SS model.

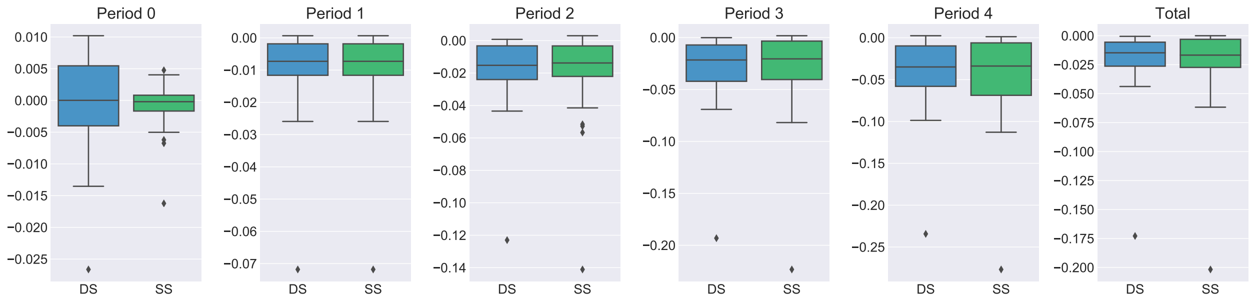

Figure 11: The relative testing error of the Deterministic Static (DS)(blue), and the Stochastic Static (SS)(green) policy to the optimal solution profit for Dominant Network Demand scenario.

Figure 12 shows a significantly improved performance for all periods except period two, which could be attributed to an over-fitting of the NN in that period. The total relative error of the SD is again similar to the DD with a lower variance for the SD.

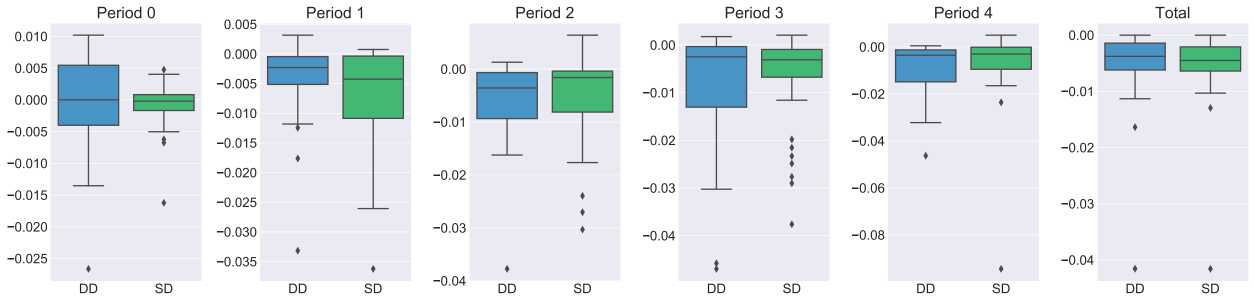

Figure 12: The relative testing error of the Deterministic Dynamic (DD)(blue), and the Stochastic Dynamic (SD)(green) policy to the optimal solution profit for Dominant Network Demand scenario.

7.4. Dominant Market Demand

In Figure 13 the training score of the RL model for both the SS and SD policy under the DMD scenario is shown. It is immediately notable that both the SS and SD policy converges to a very high and a similar training score. Moreover, both the training and the validation scores show low variance over the validated networks. The difference in performance scores between the static (SS) and dynamic (SD) policy is the lowest of all three tested demand scenario's. Because of the mitigating effect of sampled multiple markets, the dispersion of cumulative demand in the network is low, and the optimal fleet decision does not deviate much. Consequently, the static policy is as almost as good an estimator as the dynamic policy.

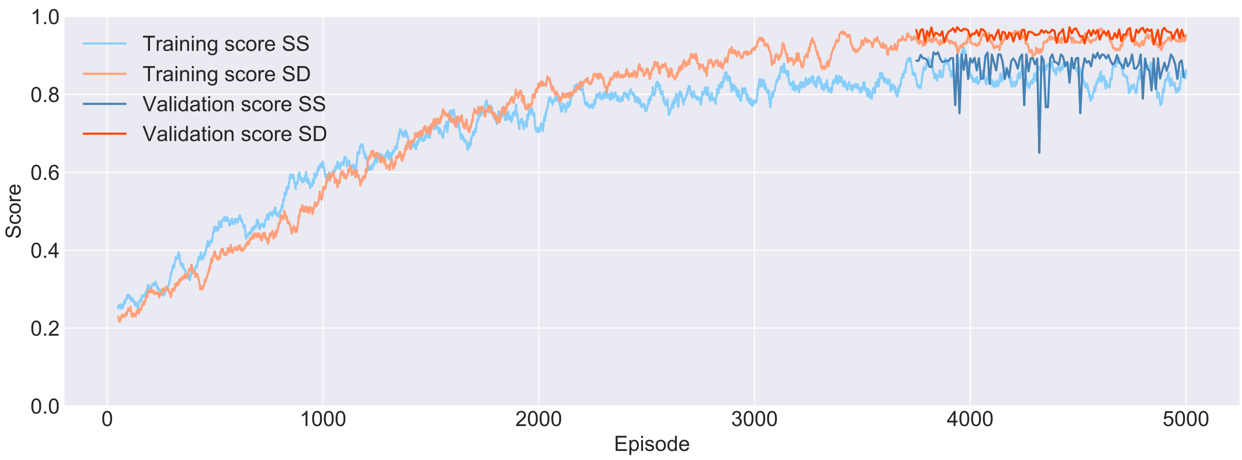

Figure 13: Training score and Validation Score of SS and SD policy network for Dominant Market Demand scenario.

In Table 7 the validation and evaluation scores and variances are shown. The stochastic trained policies show compatible results to the deterministic counterparts. The SS model's Average Validation Score is practically equal to the DS score but with a lower variance score. The SD shows a slightly sub-optimal Average Validation Score and Variance to the DD model. The evaluation of the models displays similar results, with the stochastic trained DQN-models matching the deterministic optimized method on both Average Score and Variance Score. The comparable results between the stochastic and deterministic are again related to the demand scenario. In DMD, the influence of the growth sampling of the network is very little on the episodic sampling of air-travel demand, compared to the sampling of independent market growths. The resulting divergence of network growth over these samples is therefor small, and the optimal fleet decision converges to a small set of actions which always score high. This is attributed to the fact that a small growth in one market can be compensated with growth in another market.

	DS	SS	DD	SD
Average Validation Score	87.18%	87.34%	97.78%	95.61%
Average Variance Validation Score	2.158	1.41	0.065	0.09
Variance of Average Val. Score	-	14.983	-	1.229
Evaluation Score	85.83%	87.83%	96.74%	96.86%
Variance Evaluation Score	2.382	2.519	0.123	0.054

Table 7: Average Validation Score of Average Demand scenario

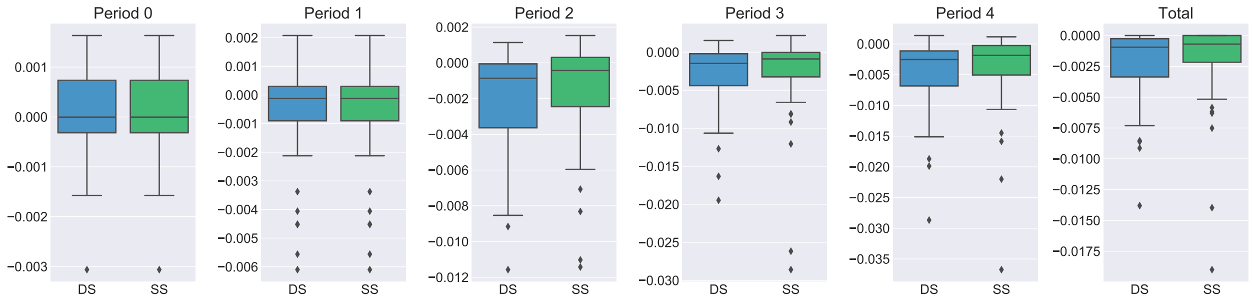

Figure 14: The relative testing error of the Deterministic Static (DS)(blue), and the Stochastic Static (SS)(green) policy to the optimal solution profit for Dominant Market Demand scenario.

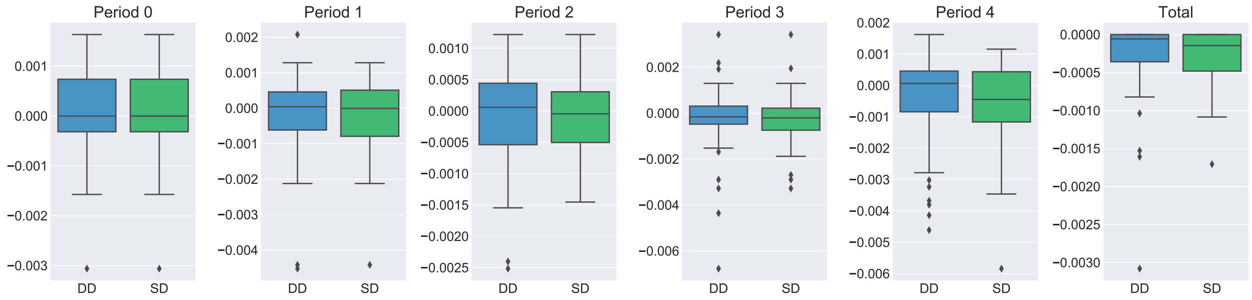

Figure 15: The relative testing error of the Deterministic Dynamic (DD)(blue), and the Stochastic Dynamic (SD)(green) policy to the optimal solution profit for Dominant Market Demand scenario.

In both Figure 14 and 15, the static and dynamic approach are shown respectively. The relative error of the SS is proportional to the DS method aside from a few outliers in the SS prediction that induce a higher variance of the evaluation score. The dynamic policies show comparable relative errors and variances too. The relative error on the total profit is very small from the optimal solution with minor deviations between the DD and SD.

7.5. Discussion

In the three demand scenarios, both the Validation- and Evaluation Score/Variance of the dynamic policy using air-travel demand as a feature show improvement over the static policy. Both policies were trained in a similar method with the only difference being the number of input features in the neural network. As the Static Stochastic (SS) policy was trained on only the fleet and period features, and the Dynamic Stochastic (SD) included features about the current state of the air-travel demand in the network, it was found that re-evaluation of the current air-travel demand increases the predictability of the optimal fleet decision and thus the performance of the fleet plan.

The SD policy showed overall a lower variance than the SS policy. As the SD policy can adjust the fleet decision based on the current demand, the SS policy trains for the best policy over the aggregate of air-travel demand evolutions. Moreover, the evaluation of the SS policy outperforms its deterministic reference policy (DS) for every demand scenario.

The SS policy validation score shows a lot more variance compared to the SD policy. This attributed to the training on very conflicting experiences; where the similar input states yield different fleet actions, which increases the noise during the training process. Moreover, the SS policy and SD share the same neural network architecture which is probably over-fitted for the SS due to its small input vector and the less complex policy it is trying to approximate.

The AD scenario shows the largest stochasticity in the market and network demand evolution out of the three scenarios due to the low smoothing the shock term and low variance of the long-term demand growth. As a result, the SD and SS model have more trouble learning as the stochasticity in the air-travel demand features creates noise in the training process. As a result, only in the AD scenario, the SD policy shows near-optimal results to the deterministic reference policy (DD).

The amount of divergence of the air-travel demand influences the learning of the policies. This is clearly demonstrated in the comparison of the DMD and DND scenario performance. The DMD scenario shows a high divergence of cumulative network demand, which resulted in the lowest prediction scores of all policies. Here, growth of the air-travel demand changes rapidly due to the large variance in long-term mean air-travel growth. This affects the predictability of the optimal fleet decisions, most notably the score and variance of the static policies as they can only generate one fleet plan for all sampled demand values. The DND scenario shows the highest prediction scores for both the static and dynamic policies. Due to the sampling of multiple market demand evolutions, the mean growth always converges to the long-term mean growth. Due to the network optimisation, and the re-assignment of fleet, the optimal fleet decisions converge to a small diverging set of fleet plans.

As pointed out earlier the training time of the neural network is highly dependent on the optimization time of the FPM and FAM. Multiple measures were taken to reduce the training time by increasing the MIP gap, re-using the sampled air-travel demand trajectory, and storing the optimal experience from the FPM. In Table 8, the training time of the NN's for the two stochastic policies are shown. Here it is clearly visible that the FPM and FAM optimisation are responsible for the long runtime and not the back-propagation of the NN for the SS and SD policy.

Nonetheless, after training the NN, the generation of the fleet decisions for a sampled tra-

	Total	FPM	FAM	SS NN	SD NN
AD [min]	1130.32	467.47	627.14	16.04	18.01
DND [min]	1082.12	429.96	616.14	16.21	18.18
DMD [min]	1159.95	459.15	665.2	16.04	17.96

Table 8: Training time of the two DQN's for the SS policy and the SD policy for the three different demand scenarios.

jectory of demand takes less than a second compared to several minutes for the deterministic counterpart (dependent on the MIP gap of the optimizer). As a result, the trained neural network could become an interesting tool for fleet planners and managers in the airline business to assist the decision process and quickly asses the fleet plan for new air-travel demand predictions. However, if the airlines' operated network changes, or the aircraft/route/demand characteristics change, a retraining of the DQN's is necessary.

8. Conclusion

The aim of this research was to contribute to the development of a dynamic fleet policy by the generation of a model-free reinforcement learning program in a fleet planning environment subjected to air-travel demand uncertainty. With this research, it is demonstrated that a RL program can be used to estimate the dynamic policy based on the air-travel demand. The proposed RL program (a) learns the optimal fleet policy and aggregates for demand uncertainties over time; (b) contains a neural network that gives good approximations of the future profit of fleet decisions; (c) has a demand forecasting model that samples realistic air-travel demand trajectories and trends; (d) results in a model that may be utilized to generate the fleet prediction for unseen air-travel demand trajectories and thus act as a reliable tool for the airline business to predict the solution to the arduous long-term fleet planning problem almost instantly.

This work shows for the first time the usage of a model-free learning algorithm with a neural network as a function approximator to learn the optimal strategic long-term airline fleet policy under air-travel demand uncertainty which completely replaces the optimisation process of the fleet problem. Using an end-to-end strategy, the fleet decisions of the agent are evaluated by comparing the profit of the predicted action using a Fleet Assignment Model (FAM) optimisation to the profit of the optimal action using a Fleet Planning Model (FPM) optimisation. At every episode, a trajectory of air-travel demand is sampled for each market using an adaptive Ornstein-Uhlenbeck forecaster based on the historical demand. With a case study, three demand scenarios are created, and two fleet stochastically trained policies (Stochastic Static (SS) and Stochastic Dynamic (DS)) are developed and evaluated against two deterministic fleet planning policies (Deterministic Static (DS) and Deterministic Dynamic (SD)).

The results showed that both of the stochastically trained policies were able to predict viable fleet plans which showed comparable or better results than the deterministic optimisation methods. However, the performance of the stochastic trained model decreased slightly with increasing stochasticity. This was attributed to the increased noise due to the fact that the state is not sufficient statical of future demand. The SD policy, using the air-travel demand as in the input states, outperformed the SS policy, without the demand in the input state consistently over all the different scenarios tested. This proves that the neural network learned from the inclusion of air-travel demand as input feature.

Although the proposed methodology is not flawless, it employs a large benefit over the deterministic approach. Once the DQN-model is trained, generating new fleet decisions for demand trajectories can be generated almost instantly. As a result, airline fleet planners and managers can use this tool to quickly asses the composition of their fleet and when to acquire or dispose of aircraft given air-travel demand trajectories.

This research opens opportunities for future work. It can be argued that the current state of the air-travel demand is not predictively sufficient for future observation of air-travel demand; therefore the assumption that the MDP is fully observable may be rejected. In future work, it should be investigated on a more detailed level if POMDP representation of the fleet planning problem is a viable solution and could potentially improve fleet predictions. Secondly, this research only considers the air-travel demand as a stochastic parameter. In future work, other or more uncertain parameters (e.g. fuel price, aircraft failure, competition, etc.) should be included to better simulate the stochastic nature of the fleet planning process. Finally, the generation of the reward using the FPM and FAM optimisation process has proved to be the bottleneck for upscaling fleet planning problem to larger networks. Consequently, future work should investigate and develop faster methods to generate a meaningful reward function.

References

P. Belobaba, A. Odoni, C. Barnhart, The global Airline Industry, 2009.

ICAO, Aviation data and analysis seminar: Fleet planning and airline route evaluation, `https://www.icao.int/MID/Documents/2017/Aviation%20Data%20and%20Analysis%20Seminar/PPT4%20-%20Fleet%20Planning.pdf`, 2017. [Online; accessed 29-01-2020].

F. S. Hillier, Introduction to operations research, Tata McGraw-Hill Education, 2012.

D. Kirby, Is Your Fleet the Right Size?, Journal of the Operational Research Society 10 (1959) 252–252.

J. K. Wyatt, Optimal Fleet Size, Journal of the Operational Research Society 12 (1961) 186–187.

G. B. Dantzig, D. R. Fulkerson, Minimizing the number of tankers to meet a fixed schedule, Naval Research Logistics Quarterly 1 (1954) 217–222.

T. E. Bartlett, An algorithm for the minimum number of transport units to maintain a fixed schedule, Naval Research Logistics Quarterly 4 (1957) 139–149.

D. P. Shube, J. W. Stroup, Fleet Planning Model, Winter Simulation Conference Proceedings (1975).

M. Bazargan, J. Hartman, Aircraft replacement strategy: Model and analysis, Journal of Air Transport Management 25 (2012) 26–29.

P. Kall, S. W. Wallace, Stochastic Programming Second Edition, 1994.

N. V. Sahinidis, Optimization under uncertainty: state-of-the-art and opportunities, Computers & Chemical Engineering 28 (2004) 971–983.

T. H. Oum, A. Zhang, Y. Zhang, Optimal demand for operating lease of aircraft, Transportation Research Part B: Methodological 34 (2000) 17–29.

G. F. List, B. Wood, L. K. Nozick, M. A. Turnquist, D. A. Jones, E. A. Kjeldgaard, C. R. Lawton, Robust optimization for fleet planning under uncertainty, Transportation Research Part E: Logistics and Transportation Review 39 (2003) 209–227.

O. Listes, R. Dekker, A scenario aggregation based approach for determining a robust airline fleet composition, Technical Report, 2002.

D. P. Bertsekas, D. P. Bertsekas, D. P. Bertsekas, D. P. Bertsekas, Dynamic programming and optimal control, volume 1, Athena scientific Belmont, MA, 1995.

C.-I. Hsu, H.-C. Li, S.-M. Liu, C.-C. Chao, Aircraft replacement scheduling: a dynamic programming approach, Transportation research part E: logistics and transportation review 47 (2011) 41–60.

H. L. Khoo, L. E. Teoh, An optimal aircraft fleet management decision model under uncertainty, Journal of Advanced Transportation 48 (2014) 798–820.

M. G. Repko, B. F. Santos, Scenario tree airline fleet planning for demand uncertainty, Journal of Air Transport Management 65 (2017) 198–208.

C. A. Sa, B. F. Santos, J.-P. B. Clarke, Portfolio-based airline fleet planning under stochastic demand, Omega (2019) 102101.

W. B. Powell, Approximate dynamic programming : solving the curses of dimensionality, Wiley, 2011.

R. Bellman, The theory of dynamic programming, Bulletin of the American Mathematical Society 60 (1954) 503–515.

W. B. Powell, What you should know about approximate dynamic programming, Naval Research Logistics (NRL) 56 (2009) 239–249.

S. Lam, L. Lee, L. Tang, An approximate dynamic programming approach for the empty container allocation problem, Transportation Research Part C: Emerging Technologies 15 (2007) 265–277.

C. Novoa, R. Storer, An approximate dynamic programming approach for the vehicle routing problem with stochastic demands, European Journal of Operational Research 196 (2009) 509–515.

W. B. Powell, Approximate Dynamic Programming-I: Modeling, Technical Report, 2009.

V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, D. Hassabis, Human-level control through deep reinforcement learning, Nature 518 (2015).

S. Dožić, M. Kalić, Three-stage airline fleet planning model, Journal of air transport management 46 (2015) 30–39.

R. S. Sutton, A. G. Barto, Reinforcement learning: An introduction, MIT press, 2018.

C. Watkins, Learning from delayed rewards (1989).

L. Requeno, B. Santos, Multi-period adaptive airline fleet planning problem, Submitted to: *Transportation Science* (2018).

I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, MIT Press, 2016. http://www.deeplearningbook.org.

V. François-Lavet, P. Henderson, R. Islam, M. G. Bellemare, J. Pineau, et al., An introduction to deep reinforcement learning, Foundations and Trends® in Machine Learning 11 (2018) 219–354.

G. E. Uhlenbeck, L. S. Ornstein, On the theory of the brownian motion, Physical review 36 (1930) 823.

O. Vasicek, An equilibrium characterization of the term structure, Journal of financial economics 5 (1977) 177–188.

E. P. Martins, Estimating the rate of phenotypic evolution from comparative data, The American Naturalist 144 (1994) 193–209.

V. Bezuglyy, B. Mehlig, M. Wilkinson, K. Nakamura, E. Arvedson, Generalized ornstein-uhlenbeck processes, Journal of mathematical physics 47 (2006) 073301.

O. E. Barndorff-Nielsen, N. Shephard, Non-gaussian ornstein–uhlenbeck-based models and some of their uses in financial economics, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 63 (2001) 167–241.

N. Chaiyapo, N. Phewchean, An application of ornstein-uhlenbeck process to commodity pricing in thailand, Advances in Difference Equations 2017 (2017) 179.

B. Santos, Lecture Notes: Airline Planning and Optimization, [Accessed on: 2019/10/15], Technical Report, Delft University of Technology, Faculty of Aerospace Engineering, 2017.

D. Silver, Ucl course on rl: Lecture 2 markov decision processes, http://www0.cs.ucl.ac.uk/staff/d.silver/web/Teaching.html, 2018. [Online; accessed 18-October-2019].