Master Thesis

Reinforcement Learning to optimize the DAA of an insurer

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in Quantitative Finance and Actuarial Science

Tilburg University

Tilburg School of Economics and Management

Author: **Arjan van Vuuren**

SNR: **2047720**

Supervisors: **dr. N.F.F. Schweizer (Tilburg University)**
ir. D.P. Kroon (Ortec-Finance)
M.H.P van Joolingen MSc. (Ortec-Finance)

Tilburg, September 2022

Abstract

This research investigates the use of Reinforcement Learning (RL) to construct the Dynamic Asset Allocation (DAA) strategy for an insurer. More specifically, we implement the Proximal Policy Optimization (PPO) method introduced by Schulman et al. (2017) and consider an insurer based in Singapore which has to comply with the Risk Based Capital (RBC) requirements as determined by the Monetary Authority of Singapore (MAS). In this research, we extend the traditional risk-return trade-off by incorporating RBC requirements and preferences into the optimization problem. We test the RL strategies by comparing the performance to one realistic Strategic Asset Allocation strategy and two traditional Mean-Variance (MV) strategies. More specifically, we consider a MV Sharpe-Ratio and Minimum-Volatility strategy. The main results of this research can be summarized as follows: (1) RL strategies using PPO can achieve equal or higher expected returns compared to the benchmark strategies; (2) including RBC requirements into the RL model can significantly decrease the probability of violating the RBC requirements and preferences compared to the SAA and MV benchmark strategies; and (3) the RL model is able to determine DAA strategies which are more practically applicable compared to the MV strategies.

Acknowledgement

First of all, I want to thank Dr. Nikolaus Schweizer for his great supervision during my master thesis. His ideas, suggestions and expertise really helped me with constructing this thesis. Furthermore, I want to thank David Kroon and Maurits van Joolingen for providing me with the opportunity to write my master thesis at Ortec-Finance. I also want to thank David Kroon for his guidance and support during our weekly supervisor meetings. Our discussions during these meetings really contributed to my thesis. Lastly, I want to thank Sander Dekker for sharing his knowledge and expertise on Risk Based Capital and other insurance-related topics.

Contents

List of Tables

List of Figures

1 Introduction

During the last decades, insurers and other institutional investors typically have optimized their asset allocation using modern portfolio theory or Mean-Variance (MV) analysis, as first introduced by Markowitz (1952) and Markowitz (1959). This method attempts to find a portfolio such that it achieves the investor-specific optimum in the risk-return trade-off. At the same time, insurers may wish to include other considerations in the optimization problem, such as capital requirements. Including such nonlinear objectives or constraints in the optimization problem is complex and will likely yield a problem-specific solution, which is only worsened when adding more considerations. A set of possible solution methods that has relatively recently gained popularity is Reinforcement Learning (RL). RL is a branch of machine learning that is concerned with finding optimal dynamic strategies for Markov Decision Processes (Sutton and Barto, 2018). Many financial decision-making problems are concerned with devising asset allocation strategies for continuously changing environments. Examples of this changing environment are market fluctuations and varying interest rates. Ideally, the model used to devise the portfolio allocation strategy considers and makes use of these circumstances.

Compared to other solutions to this type of problem, i.e. analytical solutions or dynamic programming, RL is scalable and flexible. In general, analytical solutions and dynamic programming may require over-simplification of the problem, are problem-specific, or are very computationally expensive. In contrast, Reinforcement Learning provides generic tools by training an agent through repeated interactions with its environment - teaching it to take actions based on its observations such that it maximizes cumulative rewards (Sutton and Barto, 2018). RL has shown to be capable of relatively efficiently tackling large-scale problems and to be flexible in obtaining successful dynamic strategies for varying complex underlying stochastic processes (Botvinick et al., 2019). Furthermore, as stated by Botvinick et al. (2019), Deep Reinforcement Learning (DRL) methods, where neural networks are combined with RL, have been able to outperform humans in domains ranging from a game of Go, to Dota 2, to no-limit poker (Silver et al., 2016;Berner et al., 2019;Xu et al., 2021).

To conclude, the goal of this research paper is to provide insight into how non-linear objectives can be incorporated into the optimal Dynamic Asset Allocation (DAA) of an institutional investor by using a Reinforcement Learning model. We demonstrate this by incorporating Risk Based Capital (RBC) requirements into the DAA of an insurer based in Singapore. In addition, this paper provides arguments why RL is a suitable method to solve this problem. Lastly, this research provides insight into the RL model, i.e. whether or not an RL model behaves according to an investors' intuition. The resulting research question which is answered in this research is:

> *How can insurers optimize their Dynamic Asset Allocation, incorporating Risk Based Capital requirements in addition to the risk-return trade-off?*

This research question will be answered as follows. First, a review of relevant literature concerning RL and DAA is provided in Section 2. Second, the data used in this project is explained and relevant descriptive statistics are provided in Section 3. Thereafter, the extensive methodology used to develop the RL model is presented in Section 4 and subsequently, the results and performance are discussed in Section 5. Lastly, a discussion is provided and recommendations for future research are given in Section 6. For reference, Appendix A contains a glossary of the acronyms which are frequently used throughout this paper.

2 Literature review

This section is dedicated to the study of relevant literature. First, the theoretical framework and general literature of RL is discussed. Then, previous research concerning Dynamic Asset Allocation by using RL will be investigated.

2.1 Reinforcement Learning

In this section, the theoretical framework of Reinforcement Learning (RL) is discussed. As for the mathematical notation, this paper follows Sutton and Barto (2018).

2.1.1 The basic Reinforcement Learning model

The basis of the RL model can be visualized as in Figure 1. In essence, the basic RL model consists of an agent which interacts with the environment to optimize sequential decisions to maximize cumulative rewards (Grondman et al., 2012; Sutton et al., 1999).

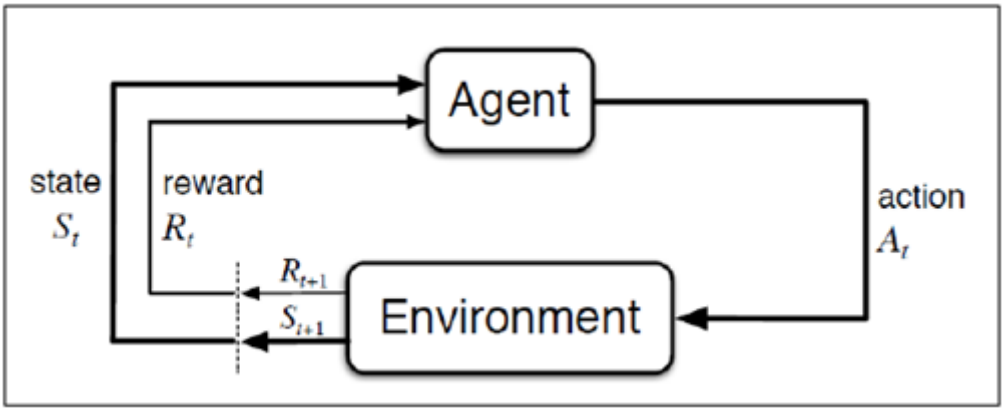


Figure 1: Basic RL model.

More formally, as defined by Sutton and Barto (2018), the agent and the environment interact at each discrete time step for all values $t \in \{0, 1, 2, ..., T\}$. At time t, the state $S_t \in \mathcal{S}$ and reward $R_t \in \mathcal{R}$ serve as the input for the agent, where $\mathcal{S}$ and $\mathcal{R} \subseteq \mathbb{R}$ are the sets containing all possible states and rewards respectively. Using this input S_t, the RL agent chooses an action $A_t \in \mathcal{A}(S_t)$, where $\mathcal{A}(S_t)$ denotes the set of all possible actions available in state S_t. After the agent has chosen its action, the transition is made towards the next time step and the environment changes. As a consequence, there will be a new reward R_{t+1} and state S_{t+1} available which will serve as input to the agent in the next time step. Then, the loop starts over for $t + 1$. In essence, the goal of the agent is to maximize the accumulative discounted long-run reward formulated as in Equation (1):

$$G_t = R_t + \gamma R_{t+1} + \gamma^2 R_{t+2} + ... + \gamma^{T-t} R_T \tag{1}$$

Here, γ denotes the discount factor and is considered to be a hyperparameter.

2.1.2 Fundamental Reinforcement Learning concepts

In this subsection, the following fundamental RL concepts are explained in more detail, namely: the policy function, value function and advantage function.

In order to determine the action at time t, the agent utilizes a mapping from the current state to a probability of selecting each possible action. This mapping is called the agent's policy and is denoted by P_t (Achiam, 2018). Provided that the agent is following policy P_t, the probability of choosing $A_t = a$ given that $S_t = s$ is denoted by $P_t(a|s)$. Whatever the application of RL, the ultimate objective is to find the optimal policy P_t^* which maximizes the discounted accumulative expected reward.

There exist several RL methods which specify how the agent adjusts its policy based on its experience. In order to update the policy of the agent, there is a trade-off between exploration and exploitation. On the one hand, exploration concerns exploring the action-space to learn usable information and explore new policies. On the other hand, exploitation relates to utilizing the information already known about the environment, in order to maximize return (Sutton and Barto, 2018). As one might expect, purely exploring might restrict the RL agent from finding the optimal policy. However, purely exploiting might push the agent's learning process towards a local optimum, limiting the RL agent from finding the global optimum (Yogeswaran and Ponnambalam, 2012).

In addition, the value function is the discounted accumulated expected reward of the agent over time, given that the agent starts in some state and follows policy P afterwards. As denoted by Sutton and Barto (2018), the value function indicates the state's desirability. Within RL, there exist mainly two different kinds of value functions. On the one hand, the value function can be a state-action value function denoted by $Q^P(a_t, s_t)$, depending on both the action and state at time t. Thus, $Q^P(a_t, s_t)$ can be interpreted as the discounted accumulated expected reward of the agent, given that the agent starts in state s_t, chooses action a_t and acts according to policy P afterwards. On the other hand, the value function could be a state value function denoted by $V^P(s_t)$, providing the expected reward depending only on the current state and not on the action at time t (Achiam, 2018). Thus, $V^P(s_t)$ denotes the discounted accumulated expected reward, given that the agent starts in state s_t and acts according to policy P. The relationship between $Q^P(a_t, s_t)$ and $V^P(s_t)$ is displayed in Equation 2:

$$V^P(s_t) = \mathbb{E}_{a\sim P}[Q^P(a_t, s_t)] \tag{2}$$

Thus, the state value function $V^P(s_t)$ can be interpreted as the expected value of the

state-action value function $Q^P(a_t, s_t)$ over all possible actions.

Lastly, the advantage function $A^P(a_t, s_t)$ is a combination of the two preceding value functions. Achiam (2018) defines the advantage function as a function indicating to what extent one action is an improvement compared to all other possible actions the agent could have taken by acting according to policy P. This results in defining the advantage function as in Equation (3).

$$A^P(a_t, s_t) = Q^P(a_t, s_t) - V^P(s_t) \tag{3}$$

2.1.3 Types of RL algorithms

Within literature, there exist a vast amount of different RL algorithms. However, an important distinction can be made between model-based and model-free RL algorithms (Achiam, 2018; Filos, 2019; Sutton and Barto, 2018). Within model-based RL algorithms, the agent tries to model the environment, enabling the agent to include predictions about future states in its decision making. On the other hand, model-free RL algorithms do not try to model the environment. Since model-free approaches perform well when a sufficient amount of data is available and this research makes use of simulated data from the Economic Scenario Generator of Ortec-Finance, a model-free approach is used.

Model-free RL can be divided into three sub-categories, namely critic-only, actor-only and actor-critic methods. These three categories differ in their usage of explicit value and/or policy functions (Grondman et al., 2012). Critic-only methods are the most published methods in finance (Fischer, 2018). These methods only rely on approximating the value function and do not use an explicit policy function (Konda and Tsitsiklis, 2003). The main drawback of critic-only methods is that these require discrete action-spaces and since we consider a continuous action-space, these methods are not applicable. The most common example of the critic-only approach is Q-learning (Watkins and Dayan, 1992).

The actor-only methods are the second most common methods in finance literature. In actor-only methods, no value function is used, thus the agent only learns the policy, i.e. the direct mapping from states to actions. The major advantage over critic-only methods is that actor-only methods can make use of a continuous action-space. A common example of actor-only methods is REINFORCE, as proposed by Williams (1992).

Lastly, the least common approach within finance literature is the actor-critic approach. This approach tries to combine the advantages of the critic-only and actor-only methods. In essence, the actor determines an action given the current state and the critic evaluates this selected action (Fischer, 2018). Although this potential major advantage over the previous two methods, the actor-critic approach has not been researched intensively in the finance domain. However, in other applications of RL, actor-critic methods have

proven to be successful in various applications in large-scale problems (Grondman et al., 2012). Frequently used examples of the actor-critic approach are Deep Deterministic Policy Gradient (DDPG) (Lillicrap et al., 2015), Twin-Delayed DDPG (TD3) (Fujimoto et al., 2018), Advantage Actor-Critic (A3C) (Mnih et al., 2016), Trust Region Policy Optimization (TRPO) (Schulman et al., 2015a) and Proximal Policy Optimization (PPO) (Schulman et al., 2017).

Aboussalah et al. (2022) have shown that out of eight policy-based algorithms, PPO and DDPG have the most promising performance in asset allocation problems. In addition, Schulman et al. (2017) have shown that PPO has better performance on continuous control tasks compared to TPRO. Moreover, as a consequence of the lack of research and the promising advantages of the actor-critic approach in finance, we have adopted this method in this research. More specifically, we implement the PPO algorithm as proposed by Schulman et al. (2017). More information about the exact implementation can be found in Section 4.4.3.

2.2 Dynamic Asset Allocation using Reinforcement Learning

This section is dedicated to the literature concerning Dynamic Asset Allocation together with Reinforcement Learning.

Dynamic Asset Allocation (DAA) optimization is considered to be one of the most challenging problems in the field of finance (Markowitz, 1959). Traditionally, asset allocation strategies are optimized by utilizing dynamic programming and convex optimization as introduced by Bellman (1966) and Boyd et al. (2004) respectively (Infanger, 2008; Merton, 1969; Neuneier, 1995). The main drawback of these approaches is that they require discrete action-spaces and consequently, suffer from the 'curse of dimensionality'.

In recent years, RL also found its way into the field of Dynamic Asset Allocation (DAA) within institutional investors. RL has shown to be capable of producing portfolio management algorithms for wealth maximization (Almahdi and Yang, 2017; Moody et al., 1998). Moreover, several studies have shown that dynamic investment strategies can be efficiently derived by using an RL approach. Guo et al. (2018) proposed a model-based RL agent which optimizes the asset allocation in a multi-asset investment environment. They showed that their RL trading algorithm outperformed traditional trading strategies on several back tests. Moreover, Yu et al. (2019) proposed a model-based RL agent that was trained by making use of an innovative RL architecture, consisting of a generative adversarial data augmentation module (DAM), an infused prediction module (IPM) and a behavior cloning module (BCM). Yu et al. (2019) demonstrated that their RL model is profitable, risk-sensitive and robust, as compared to the performance of the used

baseline trading strategies. Lastly, Li and Forsyth (2019) proposed a data-driven Deep Reinforcement Learning optimization framework to determine the optimal asset allocation during the accumulation phase of a defined contribution pension scheme. The proposed solution was validated by comparing the DRL solution to the optimal solution of the Hamilton-Jacobi-Bellman (HJB) equation, where the DRL solution achieved near optimal performance.

However, a great shortcoming of the previous mentioned papers is that they only consider the traditional risk-return trade-off, while as mentioned before, institutional investors may have other considerations like capital requirements. Therefore, the main contribution of this research is to show that RL is capable of incorporating these non-linear objectives, while still considering the traditional risk-return trade-off.

3 Data

In this section, the data used within this research will be elaborated. First, the used data sources are discussed. Then, the structure of the data will be looked at in more detail and lastly, descriptive statistics of both the historic and scenario data from the Economic Scenario Generator will be reviewed.

3.1 Data sources

In this paper, we use 15000 simulated scenarios generated by the Economic Scenario Generator (ESG) of Ortec-Finance. The idea of the ESG is to generate scenarios in a consistent and transparent way. This is achieved by decomposing economic and financial indicators into long-, medium-, and short-term components. Using these components, the ESG constructs factor models for each time series. Then, the long-, medium-, and short-term factors are merged to achieve consistency across different time horizons. For further reading about the ESG of Ortec-Finance, please consult Steehouwer (2016).

The ESG generates scenarios which are contained in the Ortec-Finance Scenario set (OFS). The OFS contains scenarios for more than 700 economic variables and asset classes. These asset classes can be divided into currencies, equities, fixed income and alternatives. The economic variables consist of GDP, unemployment, inflation and exchange rates indices (Ortec-Finance, 2022).

Besides the data from the OFS, we will make use of input and views from experts in the field of insurance from both Ortec-Finance and a client of Ortec-Finance.

3.2 Data structure

In this research, every scenario contains the time series of the returns of 12 asset classes and 5 macro-economic variables, namely: (1) Equity - Worldwide, (2) Private Equity, (3) Equity Asia excluding Japan, (4) Equity Singapore, (5) Direct Real Estate, (6) Infrastructure, (7) Government Bonds Singapore, (8) Emerging Market Debt, (9) Credits Singapore A, (10) Credits Asia A, (11) Credits Singapore BBB, (12) Credits Asia BBB, (13) Inflation Singapore, (14) GDP Singapore, (15) Inflation US, (16) GDP US and (17) Unemployment US. All time series contain monthly data points for a time span of 60 months. Within this paper, the Ortec-Finance Scenario set (OFS) of December 2021 is used, which implies that the scenarios are generated by the ESG with the economic outlook of December 2021. Moreover, the OFS of May 2022 is used to check the robustness of the proposed model. Thus, the December 2021 OFS is used for training, validation and testing, while the May 2022 OFS is solely used for robustness testing. In addition, the December 2021 OFS is split up into a training set (70%), validation set (15%) and test set (15%). The training

set is used to train the Reinforcement Learning algorithm, the validation set is used to calibrate the hyperparameters of the model and the test set is solely used to acquire results. Thus, the May 2022 OFS serves for 100% as a test set.

As mentioned before, the OFS contains macro-economic and returns data. Therefore, we will provide the descriptive statistics of the development of the macro-economic variables and the returns of all asset classes over time in the following section.

3.3 Descriptive statistics

This section is dedicated to the descriptive statistics of the data used throughout this paper. First, the historic returns will be analysed. Then, the returns of all asset classes will be looked at in more detail. Lastly, the development of the macro-economic variables will be investigated. Notice that only the descriptive statistics of the OFS of December 2021 are provided in the following section, the descriptive statistics of May 2022 can be found in Appendix B. Moreover, the individual fan charts of all asset classes are provided in Appendix C.

3.3.1 Historic returns

Figure 2 displays the historic returns of the 12 asset classes. First note that there is a significant difference between the volatility of the equity asset classes compared to the other asset classes. Moreover, the Fixed-Income (FI) asset classes, which consist of bonds, EMD and credits, show very little variation over time. In addition, note that a significant drop in returns is visible around April 2020 for almost all asset classes, which corresponds to the start of the Covid pandemic.

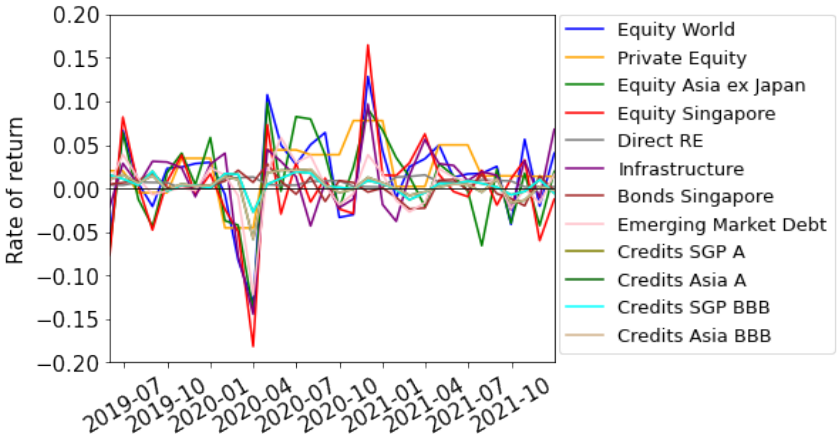


Figure 2: Historic returns of all asset classes.

3.3.2 Scenario returns

Figure 3 shows the average returns of the December 2021 OFS of all asset classes for every month. First observe that the first months show a significant decrease in returns, especially within the equity classes. Second, notice that the asset class Bonds Singapore shows very little variation over time and that, in the long run, this also results in the lowest expected return compared to other asset classes. This is consistent with literature as government bonds are seen as one of the safest investment opportunities, consequently having low expected return. In addition, all credits asset classes show a similar pattern as government bonds Singapore, although these asset classes exhibit a bit more variance and consequently, have a higher average return. Another noteworthy aspect in Figure 3 is the course of Credits SGP BBB, which seems to exhibit a significant yearly trend. According to the Scenario and Asset Valuation (SAV) department of Ortec-Finance, this is caused by the increase in the credit spread of Singapore from the first to the second year. This sudden increase results in a lower expected return for Credits SGP BBB. This effect is even more noticeable in the May 2022 OFS displayed in Appendix B.

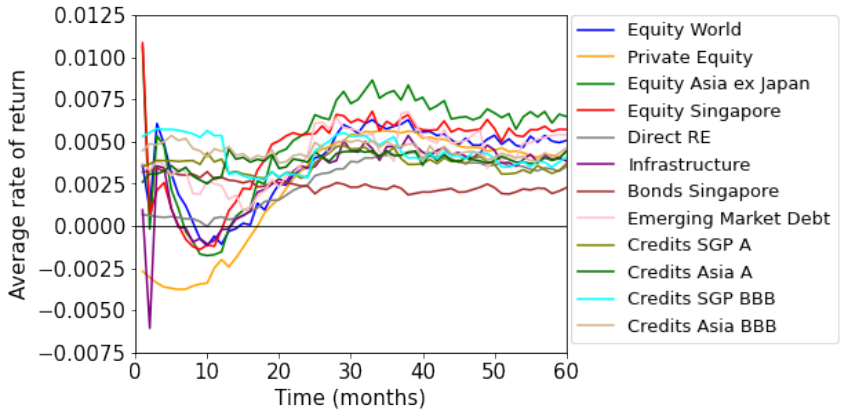


Figure 3: Average nominal scenario returns of all asset classes in December 2021 OFS.

3.3.3 History macro-economic variables

Figures 4 and 5 display the historic development of the macro-economic variables. First note that at the start of the pandemic, both GDP's and Unemployment US experienced a large decrease and increase respectively. However, where GDP Singapore and US quickly recovered from this significant decrease, Unemployment US needed much more time to restore to its long-run mean. Furthermore, notice that Inflation Singapore and US did not seem to be effected by the pandemic. However, notice that a slight upward trend for both Inflation Singapore and US is visible in the last months.

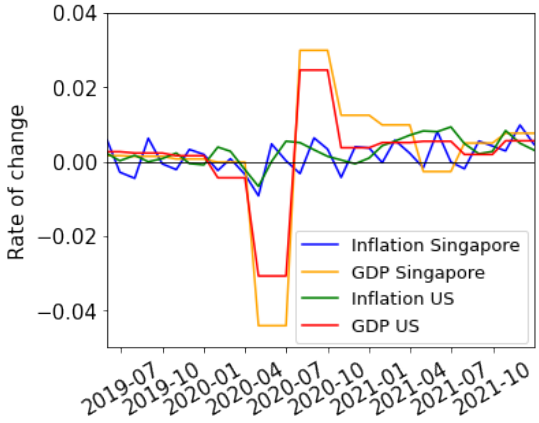


Figure 4: History of macro-economic variables, excluding Unemployment US.

Figure 5: History of macro-economic variables, only Unemployment US.

3.3.4 Macro-economic variables

Figure 6 shows the average change of the four macro-economic variables and Figure 7 displays the average level of unemployment in the US over all scenarios. First, note the sudden drop of both Inflation and GDP Singapore in the first month in Figure 6. Second, notice that in the December 2021 OFS it is assumed that the growth factor of the GDP of Singapore will slow down in the first 18 months. A similar pattern is visible for GDP US in the first 10 months. After these periods, the growth factors increase slightly and then converge to their long-run mean. On the other hand, the growth factor of Inflation US is first expected to increase slightly and then converge to its long-run mean. When looking at Figure 7, one notices that the average unemployment level first increases to approximately 5% and then remains equal.

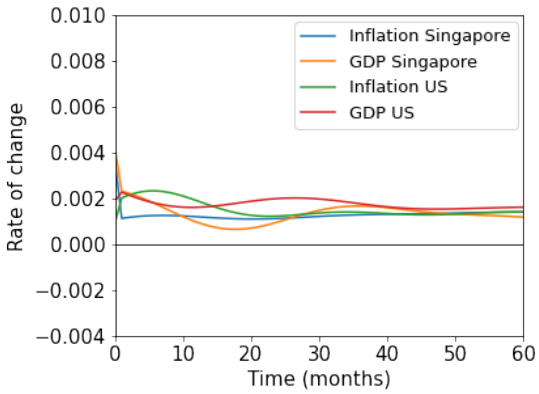


Figure 6: Average change of macro-economic variables over all scenarios in the December 2021 OFS.

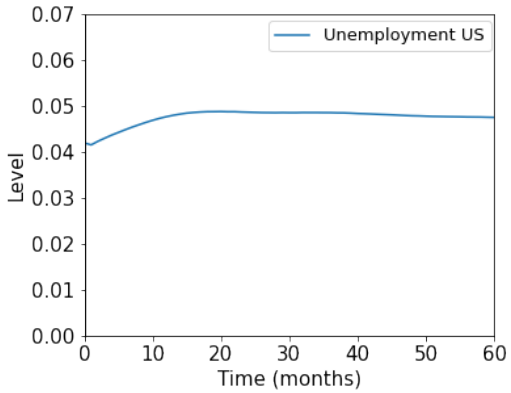


Figure 7: Average level of Unemployment US over all scenarios in the December 2021 OFS.

4 Methodology

This section is dedicated to the methodology used within this research. First, the problem definition and objective are clarified. Then, the DAA framework needed for this research is provided and the used RBC framework is elaborated. Thereafter, the RL model implementation is explained in more detail and the benchmark models are elaborated. Lastly, the performance metrics and hyperparameter tuning are discussed.

4.1 Problem definition

In this research, we consider an insurer based in Singapore. The main goal of this insurer is to maximize returns while conforming to the Risk Based Capital (RBC) framework provided by the Monetary Authority of Singapore (MAS). The MAS is the central bank of Singapore. The RBC framework forces insurers to properly define and calculate their risk capital. The National Association of Insurance Commissioners (NAIC, 2022) states that 'the purpose of RBC requirements is to identify weakly capitalized companies, which facilitates regulatory actions to ensure that policyholders will receive the benefits promised without relying on a guarantee association or taxpayer funds. In essence, the RBC formula calculations are critical thresholds that enable timely regulatory intervention'. The most crucial concept within the RBC framework is the Capital Adequacy Ratio (CAR_t), which determines how solvent an insurer is. A detailed description on how the CAR_t is determined is provided in Section 4.3.

The considered time horizon is 5 years where monthly time steps are used. The starting capital is set to 33 Bln SGD and the starting liability value is 21 Bln SGD. The only actions the insurer can take is to adjust its portfolio allocation and we consider both 0.15% and no transaction costs. We choose to consider 0.15% transaction costs since this is a representative number for most multi-asset funds (Petraki, 2020). The objective of this research was to train an RL agent which is able to find the DAA which maximizes the expected net return, while conforming to the Risk Based Capital requirements set by the MAS. First, lets define the net return at time t as NR_t, more information about the precise definition of NR_t can be found in Section 4.2. Then, this results in the following general optimization problem:

$$\max_{\pi_i(s_t)} \quad \mathbb{E}[\sum_{t=1}^{T} \gamma^t u_t(\text{NR}_t, \text{CAR}_t)] \tag{4a}$$

$$\text{Subject to:} \quad \sum_{i=1}^{p} \pi_i(s_t) = 1, \qquad \forall t \in \{1, 2, ..., T\} \tag{4b}$$

$$l_i \leq \pi_i(s_t) \leq u_i, \qquad \forall i \in \{1, 2, ..., p\}, \forall t \in \{1, 2, ..., T\} \tag{4c}$$

where γ denotes the discount factor and $u_t(\text{NR\_t}, \text{CAR}_t)$ denotes some utility function of NR_t and CAR_t. Moreover, note that the asset weights $\pi_i(s_t)$ explicitly depend on the state of the environment s_t via the policy P. More information about the state s_t can be found in Section 4.4.1. In addition, both NR_t and CAR_t depend on $\pi_i(s_t)$ and the state s_t. Thus, the objective in Equation 4a can be interpreted as maximizing the expected discounted cumulative utility from net returns and the CAR. Then, solely focusing on maximizing net returns while conforming to the RBC requirements and setting $\gamma = 1$, results in the following intuitive objective function:

$$\max_{\pi_i(s_t)} \quad \mathbb{E}[\sum_{t=1}^{T} \text{NR}_t - \alpha_2 \mathbb{1}_{\text{CAR}_t \leq 1}] \tag{5a}$$

$$\text{Subject to:} \quad \sum_{i=1}^{p} \pi_i(s_t) = 1, \qquad \forall t \in \{1, 2, ..., T\} \tag{5b}$$

$$l_i \leq \pi_i(s_t) \leq u_i, \qquad \forall i \in \{1, 2, ..., p\}, \forall t \in \{1, 2, ..., T\} \tag{5c}$$

In consultation with the client of Ortec-Finance, we concluded that the objective function in Equation 5a did not fully capture the preferences of an insurer. Therefore, conforming to the RBC requirements has been interpreted as follows. On the one hand, an insurer wants to minimize the probability of having the CAR_t below 1. On the other hand, an insurer wants to maximize NR_t provided that the Capital Adequacy Ratio at time t is between a given range. Then, putting everything together, this corresponds to the following optimization problem:

$$\max_{\pi_i(s_t)} \quad \mathbb{E}[\sum_{t=1}^{T} \text{NR}_t \mathbb{1}_{\alpha_0 \leq \text{CAR}_t \leq \alpha_1} - \alpha_2 \mathbb{1}_{\text{CAR}_t \leq 1}] \tag{6a}$$

$$\text{Subject to:} \quad \sum_{i=1}^{p} \pi_i(s_t) = 1, \qquad \forall t \in \{1, 2, ..., T\} \tag{6b}$$

$$l_i \leq \pi_i(s_t) \leq u_i, \qquad \forall i \in \{1, 2, ..., p\}, \forall t \in \{1, 2, ..., T\} \tag{6c}$$

First note that α_0, α_1 and α_2 are preference parameters. Here, α_0 and α_1 denote the preferred lower and upper bound of the Capital Adequacy Ratio respectively. From literature and in consultation with a client of Ortec-Finance, the preferred Capital Adequacy Ratio range is set to $(1.75 - 2.50)$ (Milliman, 2019). Although these boundaries are arbitrary, some intuition can be extracted from it. The lower bound of 1.75 ensures that there is a sufficient buffer available for when the portfolio value decreases rapidly, for in stance

during an economic recession. In addition, the upper bound of 2.5 is related to the notion that when the CAR is above this level, one has too much safety capital available. Here, having too much safety capital implies that an insurer could have increased its liabilities, for instance by selling more insurance products. Therefore, having an extremely high CAR is not preferred. Furthermore, note that in Equation (6a) there is no explicit penalty for volatility as this is implicitly captured within the CAR_t. That is, a volatile portfolio will result in heavy fluctuations in the CAR_t, resulting in an increased probability of not meeting the RBC requirements. Therefore, a trade-off has to be made between on the one hand, staying within the specified bounds of the CAR_t, implying a less risky portfolio and lower returns, and on the other hand maximizing expected return, implying a more risky portfolio. Therefore, the indicator function ensures that the agent only gets a reward if the CAR preferences are met. Lastly, a penalty is introduced for when the CAR_t is lower than 1, as this would imply regulatory intervention. Using α_2, one can increase or decrease the weight of this penalty according to the preferred probability of (not) meeting the RBC requirements. In addition, the insurer is only allowed to invest all of its capital, resulting in the constraint displayed in Equation (6b). Moreover, the asset weights must adhere to the lower and upper bound l_i and u_i respectively. Within this research research, these boundaries are set to 0 and 1 for all asset classes, implying that only short selling is not allowed. Then, the optimization objective which is considered throughout this research can be written as follows:

$$\max_{\pi_i(s_t)} \quad \mathbb{E}[\sum_{t=1}^{60} \text{NR}_t \mathbb{1}_{1.75 \leq \text{CAR}_t \leq 2.5} - \alpha_2 \mathbb{1}_{\text{CAR}_t \leq 1}] \tag{7a}$$

$$\text{Subject to:} \quad \sum_{i=0}^{11} \pi_i(s_t) = 1, \qquad \forall t \in \{1, 2, ..., 60\} \tag{7b}$$

$$0 \leq \pi_i(s_t), \qquad \forall i \in \{0, 1, ..., 11\}, \forall t \in \{1, 2, ..., 60\} \tag{7c}$$

4.2 Dynamic Asset Allocation fundamentals

In this section, the fundamental concepts and elements of DAA are discussed. For this purpose, the mathematical notation is largely based on the lecture notes of Munk (2017).

First, let us define the price of an asset (class) as $P_{i,t}$, then the nominal return over one period can be denoted as:

$$R^n_{i,t+\Delta t} = \frac{P_{i,t+\Delta t} - P_{i,t}}{P_{i,t}} \tag{8}$$

Since only real returns are considered within this research, the real returns are defined as

in Equation (9).

$$R^r_{i,t+\Delta t} = \frac{1 + R^n_{i,t+\Delta t}}{1 + I_{t+\Delta t}} \tag{9}$$

where I_t denotes the inflation rate at time t. Throughout this research, real return $R^r_{i,t+\Delta t}$ is also denoted as $R_{i,t+\Delta t}$ to improve readability. At time t, the agent chooses a portfolio which will remain unchanged until time $t + \Delta t$, holding $M_{i,t}$ units of asset i. Then, current wealth W_t is given by:

$$W_t = \sum_{i=1}^{p} M_{i,t-\Delta t} P_{i,t} \tag{10}$$

where p denotes the number of asset classes within the portfolio. It is assumed that the insurer is not allowed to add or withdraw wealth to the fund. Thus, the portfolio choice at time t must satisfy the following budget restriction:

$$\sum_{i=1}^{p} (M_{i,t} - M_{i,t-\Delta t}) P_{i,t} = 0 \tag{11}$$

Equation 11 is commonly known as the budget restriction for a self-financing portfolio. Now let us define the amount invested in asset i at time t as $\theta_{i,t} = M_{i,t} P_{i,t}$. Then, the portfolio weight of asset i at time t is given by:

$$\pi_{i,t} = \frac{\theta_{i,t}}{W_t} \tag{12}$$

As stated before, it is assumed that all capital must be invested, this results in the following constraint:

$$\sum_{i=1}^{p} \pi_{i,t} = 1, \qquad \forall t \in \{0, 1, ..., T\} \tag{13}$$

In addition, the asset weights should adhere to the following constraint:

$$\pi_{i,t} \geq 0, \qquad \forall i \in \{1, 2, ..., p\}, \forall t \in \{1, 2, ..., T\} \tag{14}$$

Moreover, transaction costs are denoted by β and will be paid over the difference in wealth allocation per asset type when the RL agent adjusts the asset allocation at time t. Let TC_t denote the transaction costs at time t, then TC_t is calculated by using equation 15.

$$\text{TC}_t = \sum_{i=1}^{p} \beta |M_{i,t} - M_{i,t-\Delta t}| P_{i,t} \tag{15}$$

Combining all of the above, the wealth dynamics can be described as:

$$W_{t+\Delta t} = W_t + \sum_{i=1}^{p} \theta_{i,t} R_{i,t+\Delta t} - \text{TC}_{t+\Delta t} \tag{16}$$

Then, the Total (real) Return obtained in period t (TR_t) is calculated by:

$$\text{TR}_t = W_t - W_{t-\Delta t} \tag{17}$$

$$= \sum_{i=1}^{p} \theta_{i,t-\Delta t} R_{i,t} - \beta |M_{i,t} - M_{i,t-\Delta t}| P_{i,t} \tag{18}$$

Where Equations (15) and (16) are used to go from Equation (17) to (18). Subsequently, the net return over period $(t - \Delta t, t)$ is denoted as NR_t and calculated by:

$$\text{NR}_t = \frac{W_t - W_{t-\Delta t}}{W_{t-\Delta t}} \tag{19}$$

$$= \frac{\sum_{i=1}^{p} \theta_{i,t-\Delta t} R_{i,t} - \beta |M_{i,t} - M_{i,t-\Delta t}| P_{i,t}}{W_{t-\Delta t}} \tag{20}$$

Then, using Equation (12), Equation (20) can be rewritten as:

$$\text{NR}_t = \sum_{i=1}^{p} \pi_{i,t-\Delta t} R_{i,t} - \beta \frac{|M_{i,t} - M_{i,t-\Delta t}| P_{i,t}}{W_{t-\Delta t}} \tag{21}$$

Here, the first part of Equation (21) can be interpreted as the real returns over the portfolio allocation and the second part as the correction for the transaction costs as a percentage of the previous wealth level. Thus, the net returns NR_t can be interpreted as the real returns of the portfolio corrected for transactions costs. Throughout this research, the goal of the insurer is to maximize NR_t while conforming to the RBC requirements and preferences.

4.3 Risk Based Capital fundamentals

This section is dedicated to the fundamental concepts of RBC and subsequently, the implementation of RBC requirements in this research. Here, the RBC guidelines published by the Monetary Authority of Singapore (MAS) are used (MAS, 2022).

As mentioned before, the Capital Adequacy Ratio (CAR_t) is a crucial concept within the RBC framework. This CAR_t consists of two components, namely Financial Resources (FR_t) and Total Risk Requirement (TRR_t). The FR_t are simply calculated as the difference between the total wealth (W_t) of the insurer and its liabilities (L_t), thus the financial resources at time t are calculated by $\text{FR}_t = W_t - L_t$. In practice, a regulatory adjustment is made to the FR_t, however, this is outside the scope of this research. In order to model the liabilities, we use an example liability structure provided by Ortec-Finance. This specific liability structure contains a set of potential monthly cash flows of a closed-book

insurer for 100 years. These cash flows have been discounted to calculate the present value for all $t \in \{0, 1, ..., 60\}$. The resulting present value at time t is then used as the value of the liabilities at time t. The second component of the CAR_t, the TRR_t is less simple and will be discussed in more detail below. In essence however, the TRR_t represents the minimum amount of capital which has to be reserved, based on the amount of risk the insurer takes.

Now, an insurer based in Singapore has to calculate their Capital Adequacy Ratio at time t according to Equation (22).

$$\text{CAR}_t = \frac{\text{FR}_t}{\text{TRR}_t} \tag{22}$$

In order to calculate the TRR_t, the following formula has to be used:

$$\text{TRR}_t = \sqrt{C_{1,t}^2 + C_{2,t}^2} + \text{OR}_t \tag{23}$$

where:

$$C_{1,t} = \sqrt{C_{1,\text{Life},t}^2 + C_{1,\text{General excl. A\&H},t}^2} + C_{1,\text{General (A\&H only)},t} \tag{24}$$

$$C_{2,t} = \sqrt{C_{2,\text{Market},t}^2 + C_{2,\text{Default},t}^2 + 2\rho_{\text{M,D}} C_{2,\text{Market},t} C_{2,\text{Default},t}} \tag{25}$$

$$\text{OR}_t = \min(0.005 L_t, 0.1(C_{1,t} + C_{2,t})) \tag{26}$$

where $C_{1,t}$ and $C_{2,t}$ denote the diversified life and general liability risks and diversified market risks respectively. Moreover, OR_t denotes the Operational Risk at time t. The main focus of this research is in the $C_{2,t}$ element of the RBC requirements, since this element dominates the contribution to the TRR_t and this is the only element which directly depends on the asset allocation of an insurer. As a simplification, a fixed percentage of the total value of liabilities is assumed for the $C_{1,t}$ risk requirements. Throughout this research, this percentage is set to 2.5%. In addition, another simplification in this research is that the $C_{2,\text{Default},t}$ risk is set to 0. This is the consequence of using the ESG of Ortec-Finance, where only sub asset classes are defined and consequently, no proper default risk can be assigned to these classes. Moreover, the contribution of $C_{2,\text{Default},t}$ to $C_{2,t}$ is commonly less than 1% and therefore, will not influence the results significantly (MAS, 2020). Therefore, only the detailed methodology with regard to the calculations of $C_{2,\text{Market},t}$ and OR_t are provided below. For more details about the proper calculations of $C_{1,t}$ requirements, see MAS (2022).

The $C_{2,Market,t}$ requirements are calculated according to equation (27).

$$C_{2,\text{Market},t} = \sqrt{\sum_{i,j} \rho_{i,j} \text{Market}_{i,t} \text{Market}_{j,t}} \tag{27}$$

where $\text{Market}_{i,t}$ denotes the shock for market i at time t. According to MAS (2022), the market shock has to be calculated as the sum of the shocks of every asset class which falls into this market category. The shock for every asset class separately is simply calculated as a percentage of the total amount invested in this asset class. Thus, the shock for market j at time t has to be calculated as follows:

$$\text{Market}_{j,t} = \sum_{q \in \mathbb{S}_j} s_q \sum_{i \in q} \theta_{i,t} \tag{28}$$

where $\mathbb{S}_j$ denotes the collection of sub-markets which fall into market j and q denotes the collection of asset classes in sub-market q. Additionally, s_j denotes the shock which has to be calculated over asset class i. As can be seen in Table 1, every sub-market has a different shock percentage since different asset classes are assumed to carry different amounts of risk. In order to clarify these calculations, a simple example is provided. Consider an insurer which invested 1 Bln SGD into Direct Real Estate US and 2 Bln SGD into Infrastructure. Then, the shock for market Property will be $0.3(1 + 2) = 0.9$ Bln SGD.

Another important aspect which can be noted in Table 1 is that this research only applies currency risk to non-FI asset classes, since it is assumed that currency risk is only fully hedged for the Fixed-Income asset classes. With regard to Interest Rate requirements, another noteworthy simplification is introduced compared to MAS (2022). Since this research abstracts away from the modeling of the duration dynamics of the Fixed-Income asset classes and a flat term-structure is used, it is additionally assumed that the duration mix within all government bonds and credits stays constant over time. As a result, a fixed percentage (8) is considered for all Fixed-Income asset classes to account for interest rate risk. This is a significant simplification compared to MAS (2022) since these assumptions completely disregard the duration dynamics and the accompanying risk requirements.

To account for interaction between the markets, all markets are given correlation factors. These correlation coefficients can be found in Table 2. As can be seen, it is assumed that especially the markets Credit Spread and Property are assumed to exhibit strong correlation with the Equity market, while the Currency market displays weak correlations with all other markets. Note that in MAS (2022), the Monetary Authority of Singapore prescribes the use of two different correlation matrices, namely one for an upward and an downward interest rate scenario. As mentioned before however, this research assumes a flat term-structure and consequently, only the downward interest rate correlation matrix is used. This research choose to use the downward correlation matrix since this matrix contains higher correlation coefficients between the different markets, resulting in a higher TRR_t.

Market	Sub-market	Asset class	Shock
Equity	Listed, developed	Equity World (0)	0.35
		Private Equity (1)	0.35
		Equity Asia excl. Japan (2)	0.35
	Other	Equity Singapore (3)	0.50
Interest Rate	All*	Government Bonds SGP (6)	0.08
		Emerging Market Debt (7)	0.08
		Credits SGP A (8)	0.08
		Credits Asia A (9)	0.08
		Credits SGP BBB (10)	0.08
		Credits Asia BBB (11)	0.08
Credit Spread	Government	Government Bonds (6)	0.0000
	Rating A	Credits SGP A (8)	0.0165
		Credits Asia A (9)	0.0165
	Rating BBB	Emerging Market Debt (7)	0.0245
		Credits SGP BBB (10)	0.0245
		Credits Asia BBB (11)	0.0245
Property	Immovable	Direct Real Estate (4)	0.30
		Infrastructure (5)	0.30
Currency	Foreign Markets	Equity World (0)	0.12
		Private Equity (1)	0.12
		Equity Asia excl. Japan (2)	0.12
		Direct Real Estate (4)	0.12
		Infrastructure (5)	0.12

Table 1: Market risk shocks table.

	Equity	**Interest Rate**	**Credit Spread**	**Property**	**Currency**
Equity	1	0.5	0.8	0.8	0.1
Interest Rate	0.5	1	0.5	0.25	0.1
Credit Spread	0.8	0.5	1	0.5	0.1
Property	0.8	0.25	0.5	1	0.1
Currency	0.1	0.1	0.1	0.1	1

Table 2: Correlation table market risk.

To conclude, the RBC framework implemented in this research can be summarized as:

$$\text{CAR}_t = \frac{\text{FR}_t}{\text{TRR}_t} \tag{29}$$

$$= \frac{W_t - L_t}{\sqrt{C_{1,t}^2 + C_{2,t}^2} + \text{OR}_t} \tag{30}$$

where:

$$C_{1,t} = 0.025 L_t \tag{31}$$

$$C_{2,t} = \sqrt{\sum_{i,j} \rho_{i,j} \text{Market}_{i,t} \text{Market}_{j,t}} \tag{32}$$

$$\text{OR}_t = \min(0.005 L_t, 0.1(C_{1,t} + C_{2,t})) \tag{33}$$

4.4 Reinforcement Learning model

This subsection is dedicated to the elaboration on the RL model which we employ in this research.

4.4.1 State definition

The state specification we use is inspired by Fischer (2018), which presents a survey on RL in financial markets. The first part of the state s_t consists of the last period nominal returns of all assets, denoted by the vector x_t with length 12. The choice for only using one lag is based on the study of Corazza and Bertoluzzo (2014), who investigated the use of employing multiple lags in the state space and conclude that the best performance is achieved with the inclusion of one lag. Since Corazza and Bertoluzzo (2014) state that this problem might be problem specific, we tested and confirmed this finding. Moreover, the asset weights of the previous period π_{t-1} are included in the state to account for transaction costs (Jiang et al., 2017). In addition, current wealth W_t, liability value L_t and the cash flows CF_t are used in the state of the agent to calculate the Capital Adequacy Ratio. Additionally, the CAR_t itself is used in the state as investment decisions could depend on the level of the CAR. Lastly, the following macro-economic variables (ME_t) are incorporated into the state of the RL agent: inflation US, inflation Singapore, GDP Singapore, GDP US and unemployment US. These macro-economic variables are included since Ndlovu et al. (2018) have found evidence that they could provide information about stock returns. Moreover, previous research has shown that including macro-economic variables could result in better performance (Neuneier, 1995). To conclude, the state representation at time t is given by the vector s_t with length 34:

$$s_t = (x_t, \pi_{t-1}, W_t, L_t, \text{CF}_t, \text{CAR}_t, \text{ME}_t, t) \tag{34}$$

4.4.2 Reward function

The reward function is crucial in the learning process of the RL agent. Aboussalah et al. (2022) state that coming up with a reward function is not a difficult task. However, designing a reward function that results in the desired behaviour, while still being learnable, can be an inconvenient task. Within literature, it is very common to use a purely profit based reward function such as the Sharpe Ratio (Fischer, 2018). However, after experimentation, we quickly found out that using the Sharpe Ratio lead to undesirable results, especially regarding performance on the CAR_t. Therefore, we experimented with incorporating RBC requirements into the reward function. After testing several configurations, the reward function which best suits our goal is:

$$\text{RW}_t = \text{NR}_t \mathbb{1}_{1.75 \leq \text{CAR}_t \leq 2.5} - \alpha_2 \mathbb{1}_{\text{CAR}_t \leq 1} \tag{35}$$

Note that the reward function in Equation (35) and the objective displayed in Equation (6a) are almost identical, resulting in the alignment of the preferences of an insurer and the learning process of the RL agent. The combination of net returns with an indicator function, which ensures that the net return reward is only given when the CAR_t is between a set range, is a novel concept. Additionally, the indicator function which provides a penalty when the CAR_t is less than 1 is innovative as well. However, as it is important for an insurer to be able to minimize the probability that the CAR_t is lower than 1, it is intuitive to utilize a penalty to steer the RL agent away from this boundary. Lastly, the introduction of this penalty did improve the learning process of the RL agent considerably.

4.4.3 Proximal Policy Optimization setup

Within this research, we model both the policy and value function using Neural Networks (NN), which are denoted as the actor and critic network respectively. The dimension of these networks are hyper parameters and can be tuned according to performance. In addition, both activation functions are Rectified Linear Units (ReLU). The ReLU function can be denoted as $f(x) = \max(0, x)$.

The NN of the actor uses the state s_t as defined in Equation (34) as input. Furthermore, the output layer has a Softmax function which returns an action-vector of length 12 containing the weights for every asset class. PPO explores the action space, i.e. all possible asset weights, by sampling its actions from a normal distribution for every action separately:

$$a_i \sim N(\mu_i, \sigma_i), \quad \forall i \in \{0, ..., 11\} \tag{36}$$

Here, the initial value of σ_i is a hyper parameter and it is assumed that every asset class has the same starting value. Although the Softmax output function guarantees that the μ_i add up to 1, the sampled weights during exploration in the training phase do not have this guarantee. Therefore, the sampled weights a_i are normalized by using:

$$\pi_i = \frac{a_i}{\sum_{i=0}^{11} a_i} \tag{37}$$

During the evaluation and test phase, there is no exploration and consequently, the asset weights π_i are equal to μ_i.

The critic network has the same input layer s_t as the actor network, yet the output layer is different. Instead of a Softmax output function, the critic network has the so-called Identity output function which returns a scalar estimate of the state-value function $V(s_t)$.

4.4.4 Proximal Policy Optimization algorithm

The PPO algorithm which we are using in this research is based on the algorithm introduced by Schulman et al. (2017). Moreover, the pseudo-code which we use in this research is inspired by Achiam (2018) and is provided in full detail in Appendix D. In addition, the implementation of Raffin et al. (2021) served as the inspiration for our implementation of PPO in Python.

The first step of the PPO algorithm is to randomly split the scenarios in the train set into batches. The batch size N is a hyperparameter and as stated before, the time horizon T is 60 months. The number of iterations over the whole data set K is a hyperparameter as well and is chosen such that the learning curve of the RL has converged.

Then, the algorithm begins with generating the states, actions, rewards and value function estimates for all t for each scenario in the batch. For all the sampled actions, the log of the probability density function of the normal distribution can be obtained. All this information is then stored into the memory of the PPO agent. After completing this for the whole scenario batch, the memory of the PPO agent contains NT entries of states, actions, rewards, value estimates and log probabilities of the actions. Now, the advantage $\hat{A}_t$ for each t is estimated by using the finite version of Generalized Advantage Estimation (GAE) as proposed by Schulman et al. (2015b):

$$\hat{A}_t = \delta_t + (\gamma\lambda)\delta_{t+1} + ... + (\gamma\lambda)^{T-t}\delta_T \tag{38}$$

$$= \delta_t + \gamma\lambda\hat{A}_{t+1} \tag{39}$$

With $\gamma \in [0, 1]$ and $\lambda \in [0, 1]$, which are hyperparameters. Additionally, δ_t is computed by using the current reward and state-value function estimates of the current and next time

step:

$$\delta_t = \hat{R}_t + \gamma \mathrm{V}(s_{t+1}) - \mathrm{V}(s_t) \tag{40}$$

Hereafter, the parameters of the actor and critic network are updated. To update the policy parameter θ, we utilize the clipped loss function of Schulman et al. (2017) in our algorithm:

$$L^{\mathrm{CLIP}}(\theta) = \hat{\mathbb{E}}_t[\min(r_t(\theta)\hat{A}_t, \mathrm{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t)] \tag{41}$$

Where $r_t(\theta)$ denotes the probability ratio which is computed by:

$$r_t(\theta) = \frac{P_\theta(a_t|s_t)}{P_{\theta_{\mathrm{old}}}(a_t|s_t)} \tag{42}$$

Moreover, the clipping ratio $\epsilon \in (0, 1)$ is a hyperparameter. Here, the intuition is to increase the probability of actions which have a higher advantage. To mitigate excessive changes to the policy parameter θ, the clip function is introduced to limit the probability ratio $r_t(\theta)$ from above at $1+\epsilon$ if the advantage estimate $\hat{A}_t$ is positive and from below at $1-\epsilon$ when $\hat{A}_t$ is negative.

For both the actor and critic network, mini-batch optimization is used with the Adam solver first introduced by Kingma and Ba (2014). Note that the mini-batches used for optimization are different from the scenario batches. More specific, the mini-batches are randomly sampled without replacement from the NT entries of the agent's memory.

After the optimization is completed, the agent's memory is cleared and all steps are repeated for every scenario batch. Once all scenario batches from the training data have been used, one training-iteration of the PPO algorithm is completed. However, one can allow for multiple iterations over the training data if the learning process of the PPO agent has not been completed. Since the learning curve can be plotted, one can increase the number of iterations by visual inspection. Note that after every training iteration, all scenario's are shuffled to prevent identical scenario batches. In our case, convergence takes place after 30 iterations.

4.5 Performance evaluation

This subsection is dedicated to the performance evaluation methodology. First, an elaboration on the benchmark strategies is provided. Then, the used evaluation metrics are discussed in more detail.

4.5.1 Benchmark strategies

In order to evaluate the performance of the RL model, three benchmark strategies are used. First, the simplest but most realistic strategy, a Strategic Asset Allocation (SAA) of

a client of Ortec-Finance is used. The full allocation is presented in table 8(a). As can be seen, a substantial amount of 66% is invested in Fixed-Income, where Government Bonds Singapore has the greatest contribution within this class. The SAA is used to ensure that the RL model outperforms a real-life strategy. Moreover, the benchmark allocation is used as the starting allocation for the RL model to ensure that both models have the same Capital Adequacy Ratio at $t = 0$.

Main asset class	**Sub asset class**	**Allocation**
Equity		**29%**
	Equit World (0)	7%
	Private Equity (1)	2%
	Equity Asia exl. Japan (2)	9%
	Equity Singapore (3)	11%
Property		**5%**
	Direct Real Estate (4)	4%
	Infrastructure (5)	1%
Fixed-Income		**66%**
	Bonds Singapore (6)	29%
	EMD (7)	1%
	Credits SGP A (8)	10%
	Credits Asia A (9)	10%
	Credits SGP BBB (10)	8%
	Credits Asia BBB (11)	8%

Table 3: Strategic Asset Allocation

In addition to the SAA, we employ a basic Dynamic Mean-Variance (DMV) model as well. Here, the Mean-Variance methodology of Markowitz (1952) and Markowitz (1959) is used. Both the DAA for a Minimum Volatility strategy (DMV-MV) and Max Sharpe Ratio strategy (DMV-SR) are considered. Traditionally, the expected returns and covariance matrix are estimated by using historical data. However, Bielstein and Hanauer (2019) and Kempf et al. (2015) have shown that using forward looking information to calculate the expected returns vector and covariance matrix leads to a significant improvement in performance. Since we are dealing with scenario data, forward looking information can be used. Therefore, a forward looking DMV is used which utilizes the same training data as the RL agent. Thus, the expected returns and covariance matrix at time t are estimated with scenario data for $t, ..., t + n$, where the best performing value for n is used. More specific, the value for n is chosen in every setting, such that maximum performance is

achieved. This implies that for the DMV-MV strategy, the value for n is chosen such that the variance in the validation set is minimized. Moreover, for the DMV-SR strategy, the value for n is chosen such that the expected final wealth is maximized in the evaluation set. For the results, see Tables 5 and 6 in Appendix E. These strategies are introduced to challenge the RL agent with regard to the average return and volatility of the total wealth value W_t. However, note that these DMV models do not consider the CAR when determining the asset allocation.

4.5.2 Evaluation metrics

The evaluation metrics which are used to measure the out-of-sample performance of the RL model against the benchmark allocations are listed below.

(1) Risk vs Return

Return is measured by the average real portfolio value $\bar{W}_t$ over time. Then, risk is measured by the metrics variance, downward variance and upward variance of the real portfolio value. The downward and upward variance are defined in Equations 43 and 44:

$$\text{Var}_{t,\text{down}} = \frac{1}{n_{t,\text{down}} - 1} \sum_{s=1}^{n} \min(W_{s,t} - \bar{W}_t, 0)^2 \tag{43}$$

$$\text{Var}_{t,\text{up}} = \frac{1}{n_{t,\text{up}} - 1} \sum_{s=1}^{n} \max(W_{s,t} - \bar{W}_t, 0)^2 \tag{44}$$

where n denotes the number of considered scenarios. Moreover, $n_{t,\text{down}} = \sum_{s=1}^{n} \mathbb{1}_{W_{s,t} < \bar{W}_t}$ and $n_{t,\text{up}} = \sum_{s=1}^{n} \mathbb{1}_{W_{s,t} > \bar{W}_t}$ denote the number of scenarios at time t where $W_{s,t}$ is respectively smaller and larger than $\bar{W}_t$.

With regard to the evaluation of the performance, the RL strategy is considered to outperform a benchmark strategy if the average wealth level is higher and the variance is lower. However, an RL strategy is also considered to outperform a benchmark strategy if the downward and upward variance of the RL strategy is respectively lower and higher. We make this distinction since downward variance is not preferred, while upward variance is considered to be acceptable.

In addition to the above risk measures measures, an insurer wants to minimize the risk of having low returns. Therefore, this paper considers the Value-at-Risk (VaR) and the Conditional VaR (CVaR). The VaR_α is the $100\alpha\%$-quantile of the portfolio value distribution or more formally, $\text{VaR}_\alpha(W_t) = \max\{c | \mathbb{P}[W_t \leq c] \leq \alpha\}$ for $\alpha \in (0, 1)$. In this research we set $\alpha = 0.05$, implying that the wealth level at time t is greater than the $\text{VaR}_{0.05}$ with a confidence level of at least 95%. In addition, the CVaR_α of portfolio wealth is considered and defined as $\text{CVaR}_\alpha(W_t) = \mathbb{E}[W_t | W_t \leq \text{VaR}_\alpha(W_t)]$.

Within this paper, the RL strategy is considered to outperform a benchmark strategy if both the VaR and CVaR of the RL strategy are higher compared to the benchmark strategy. Thus, this is formalized as:

$$\text{VaR}_\alpha(W_t^b) \geq \text{VaR}_\alpha(W_t) \quad \text{and} \quad \text{CVaR}_\alpha(W_t) \geq \text{CVaR}_\alpha(W_t^b), \quad \forall t \in \{1, 2, ..., T\} \tag{45}$$

where W_t and W_t^b denote the wealth of the RL and benchmark strategy respectively.

(2) Capital Adequacy Ratio

The CAR_t, as defined in Equation (22), is used to compare the models based on its performance regarding the RBC requirements. In order to evaluate whether a model performs well, the CAR's for all scenario's at every time step are measured. First, the percentage of the scenario's for which the CAR_t is between the levels $(1.75 - 2.5)$ is plotted over time to evaluate the preference fulfillment. Additionally, the percentage of the scenario's for which the CAR_t is below 0 and 1 are plotted over time. The threshold of 1 is considered since an insurer will experience regulatory restrictions and/or interventions when the CAR is below this level. In addition, the threshold of 0 is considered as this would imply that the Financial Resources are negative, i.e. $W_t < L_t$, resulting in bankruptcy.

4.6 Hyperparameter tuning

The PPO algorithm contains multiple hyperparameters which have to be tuned for optimal performance of the RL model. Appendix F displays the candidate hyperparameters and the optimal values. The hyperparameter candidates are based on Andrychowicz et al. (2020), who present an empirical study on hyperparameters within on-policy RL. First, note that the total amount of potential configurations exceeds 640 million. Due to the substantial amount of potential configurations, we have performed a random search to determine the optimal hyperparameter values. More specific, we have performed multiple trials with 32 randomly chosen configurations, where we shrunk the hyperparameter candidate space according to visual inspection after every trial. To determine the best candidate, we have both looked at the average final wealth and the RBC violation rate. In total, we have tested 320 configurations to determine the optimal hyperparameter configuration for our setting. However, we acknowledge that due the substantial size of the candidate hyperparameter space, it could be the case that our hyperparameter setting is sub-optimal. However, time and money restricted us from testing more configurations.

4.7 Tools and programs

Finally, we give some technical details. All programming is done in Python, version 3.10. Moreover, for the RL implementation we have used the PyTorch package and the internal ofrl-package from Ortec-Finance. For the implementation of MVO in Python, the PyPortfolioOpt package is used (Martin, 2021). Lastly, Microsoft Azure is used for the tuning of the hyperparameters.

5 Results

This section contains the results of all models that we considered in this research. First, we consider the setting where transaction costs are included. In this setting, we evaluate the allocation strategy and performance of all models. Thereafter, a similar study will be performed in the setting where transaction costs are set to 0. Finally, a robustness test is performed where only the model performance will be investigated.

5.1 Results including transaction costs

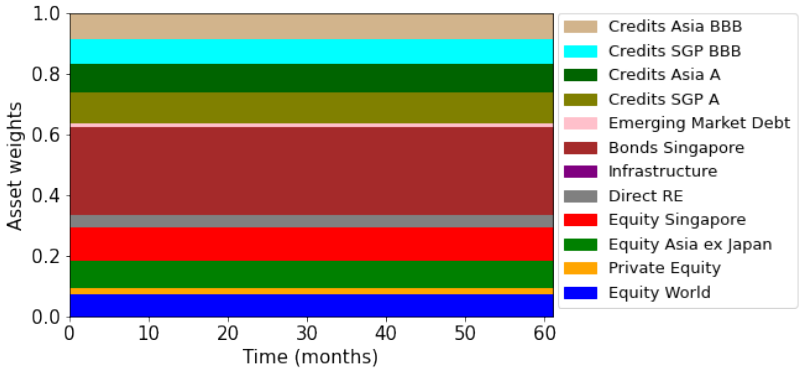
This section is dedicated to the results of the considered strategies while including transaction costs. As noted before, the transaction costs are set to 0.15%. As can be seen in Table 5 in Appendix E, the best performing DMV-SR model bases its allocation on scenarios from time $t, ..., t+2$. Moreover, the best performing DMV-MV strategy bases its allocation at time t only on the scenarios at time t.

5.1.1 Allocation strategies

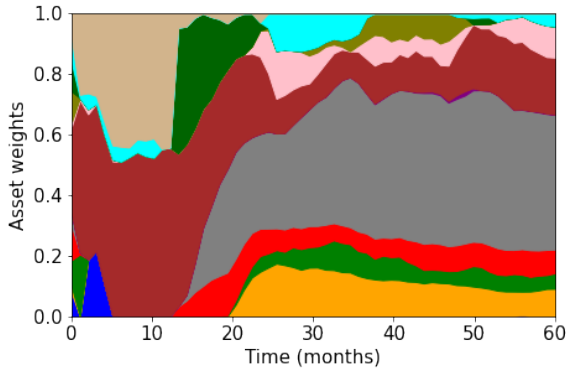
Figure 8 displays all asset allocation strategies from the different methods. Subfigure 8(a) displays the SAA, which is the most realistic and basic asset allocation strategy. Moreover, this asset allocation functions as the initial allocation for all dynamic asset allocation strategies to ensure a consistent initial CAR_0. In addition, Subfigures 8(b) and 8(c) display the Mean-Variance strategies. As can be seen, both strategies significantly differ from one another. The Sharpe-Ratio strategy fluctuates more heavily compared to the Minimum-Volatility strategy. When comparing Figures 3 and 8(b), we notice that the Sharpe-Ratio strategy mimics the estimated market development. More specifically, the Equity asset classes will perform poorly in the first 15 months and the FI asset classes perform relatively well, thus allocating more to FI asset classes in the first 15 months seems trivial. Thereafter, the Equity asset classes show an increase in expected returns while the FI asset classes remain stable and thus, substituting allocation from FI to Equity asset classes seems beneficial. Lastly, as can be seen in Figure 29 in Appendix C, the DMV-MV strategy mainly uses assets which have the least variation, namely Bonds Singapore and Direct Real Estate.

When looking at the average RL strategies displayed in Subfigures 8(d) and 8(e), we first notice that both strategies slowly increase the amount invested in Equity asset classes. This can be explained by the fact that we consider a closed-book insurer, implying that the insurer does not increase its liabilities over time when it has sufficient wealth. Since the liability value does not increase and one of the goals is to maximize returns, the RL agent can slowly increase its allocation to Equity asset classes while still being able to

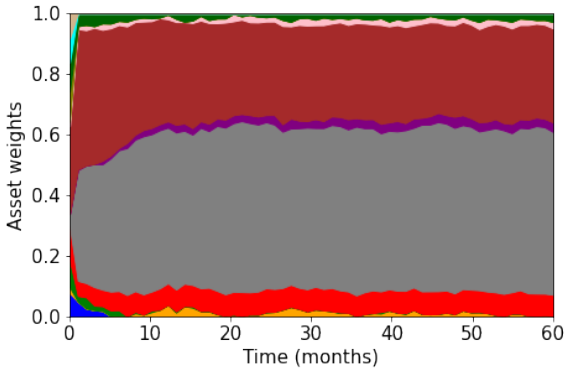
comply with the RBC requirements and preferences. Note that this observation does not necessarily hold when considering an open-book insurer, where the liabilities could increase over time.



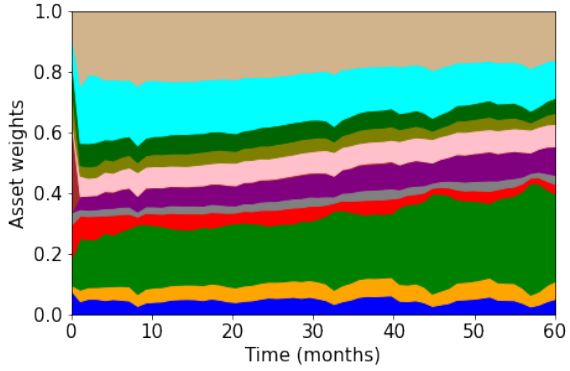
(a) SAA



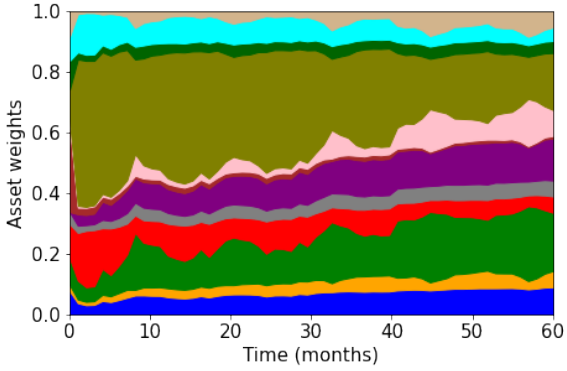
(b) DMV-SR



(c) DMV-MV



(d) RL-0.001



(e) RL-0.2

Figure 8: Asset allocation strategies including transaction costs.

Additionally, we look at the usability of the allocation strategies displayed in Figure 8. As for the SAA in Subfigure 8(a), this strategy is intended to mimic a strategy from the client of Ortec-Finance and therefore, this is the most practical strategy as well. The DMV-SR strategy in Subfigure 8(b) might be the least applicable strategies, since the

allocation substantially differs over time. Especially in the first 20 months of the DMV-SR model, the allocation strategy varies significantly and at some time points only 2 or 3 assets are used. Thus, one might question whether this proposed strategy is realistically plausible. When looking at the DMV-MV strategy in Subfigure 8(c), one notices that very few assets are used over the whole time horizon. On average, more than 90% is invested in Bonds Singapore, Direct RE and Equity Singapore. Thus, similar as with the DMV-SR strategy, one might question the usability of this allocation strategy. When considering the RL strategies displayed in Subfigures 8(d) and 8(e), one notices that especially the RL-0.001 shows similar behaviour as compared to the SAA in Subfigure 8(a). That is, the allocation is relatively stable over time compared to the other DAA strategies, which contributes to the applicability of the strategy. Most of the change in allocation over time is due to slowly increasing the allocation to Equity Asia excluding Japan and decreasing the allocation to both Credits BBB asset classes. The RL-0.2 strategy displayed in Subfigure 8(e) has substantially more variation over time compared to the RL-0.001 strategy. This variation is due to the seasonal effect where there is an exchange in allocation between Equity Singapore and Equity Asia and between Emerging Market Debt and Credits SGP A. Therefore, the RL-0.2 is less applicable compared to the RL-0.001 and SAA strategy.

Lastly, another noticeable difference between the RL and benchmark strategies is the allocation to Bonds Singapore and Equity Asia excluding Japan. While the RL strategies significantly invest in Equity Asia excluding Japan and less in Bonds Singapore, the benchmark strategies invest more heavily into Bonds Singapore and significantly less in Equity Asia excluding Japan. This difference can be explained by on the one hand the expected return of Equity Asia excluding Japan and on the other hand, the relatively low required risk capital. When looking at Figure 3, one notes that especially in the second half of the time horizon, the expected returns of Equity Asia are significantly higher than other asset classes. In addition, Table 1 shows that Equity Asia is assigned to the sub-market 'listed, developed', which implies that relatively little risk capital needs to be withheld in order to comply with the RBC requirements. Thus, from a RBC perspective, it seems that Equity Asia excluding Japan is a worthwhile asset class to invest in. On the other hand, Figure 3 shows that Bonds Singapore has the lowest expected long-run return. However, Table 1 also shows that Bonds Singapore requires the least amount of risk capital as well. Thus, although Bonds Singapore requires the least amount of risk capital, the RL agent does not utilize this asset class because of the low long-run returns.

To conclude, all allocation strategies substantially differ from one another. Moreover, the DMV-SR strategy is the least applicable strategy since it fluctuates too substantially over time. On the other hand, the DMV-MV utilizes too little assets to be practicably feasible. Then, the RL-0.2 strategy shows seasonal fluctuations which questions its practical applicability. Lastly, the SAA and RL-0.001 strategies are the most applicable since these

do not fluctuate too heavily over time and utilize a sufficient number of assets. Therefore, these strategies will be easier to understand for senior management and other stakeholders.

5.1.2 Performance evaluation

The average wealth development of all strategies are displayed in Figure 9. First note that a kink is visible around 25 months for all strategies. This kink corresponds to the turning point displayed in Figure 3, where the Equity asset classes start having a higher expected return compared to the Fixed-Income asset classes. Thus, all portfolio strategies suffer from the poor performance of the Equity asset classes in the first 25 months. In addition, note that the DMV-SR model slightly outperforms the SAA and RL strategies in those first months. However, as the Equity asset classes start to perform better, the RL strategies start having a higher expected real wealth compared to the DMV-SR strategy. Especially the RL-0.001 model with more focus on acquiring returns has superior performance over all other strategies. Lastly, note that the real wealth development of the DMV-MV strategy is worse compared to all other strategies at all t. Especially in the first 20 months the DMV-MV strategy does not acquire enough nominal return to outweigh the inflation and transaction costs, resulting in almost no increase in expected real wealth.

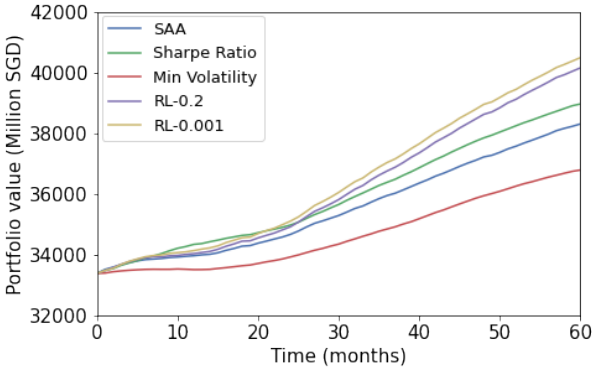
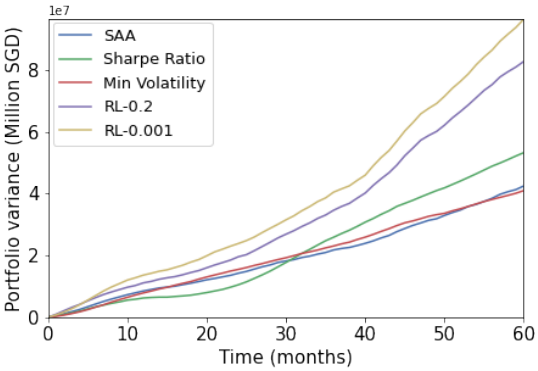


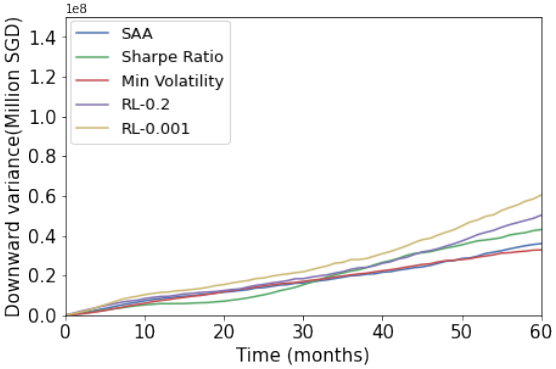
Figure 9: Average real wealth development of all strategies.

The volatility measures are displayed in Figures 10 and 11. When looking at Figure 10(a), which displays the variance of all strategies, one quickly notices that the variance of the RL strategies are substantially higher compared to the benchmark strategies. Moreover, especially the DMV-SR strategy has a noteworthy low variance in the first half of the time horizon. On the other hand, the DMV-MV strategy does not display the expected low variance one would expect in the first 30 months. However, note that in Subfigure 10(b) the DMV-MV strategy does have the lowest downward variance at the end of the time horizon. Moreover, notice that the RL learning strategies do have a slightly higher downward variance and that especially the upward variance is significantly larger. Thus, the RL

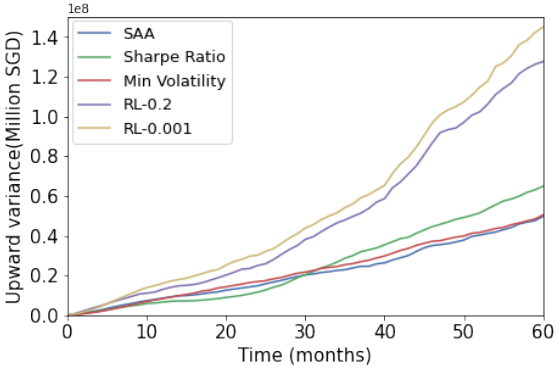
models are able to find investment opportunities which have relatively low downward variance and high upward variance.



(a) Variance.



(b) Downward variance.

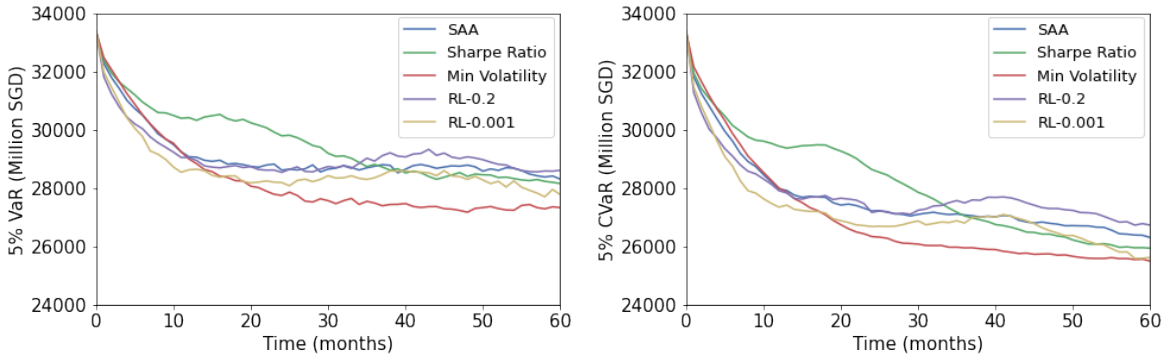


(c) Upward variance.

Figure 10: Variance plots of all strategies.

Now, when looking at Figure 11, we notice that the RL strategies have a similar VaR and CVaR development compared to the SAA. Thus, although the RL strategies have larger variance, the higher expected return and similar downward variance ensure that the VaR and CVaR are not worse compared to the SAA benchmark. Only the CVaR of the RL-0.001 strategy is slightly worse compared to the SAA and RL-0.2 strategies. In addition, the VaR and CVaR of the DMV-SR benchmark clearly outperforms all other strategies in the first 30 months, which is explained by the high expected real return and low downward volatility in this period. After 30 months, the DMV-SR VaR and CVaR are comparable to the SAA and RL strategies. Lastly, the DMV-MV strategy has a significantly worse VaR and CVaR compared to all other strategies, which is explained by the significantly lower expected net returns of the DMV-MV strategy.

Figure 12 displays the performance of all the strategies regarding the RBC requirements and preferences. When looking at Subfigure 12(a), we first notice that the preference violation percentage of the benchmark strategies increase over time, while the RL strategies



(a) 5% VaR.

(b) 5% CVaR.

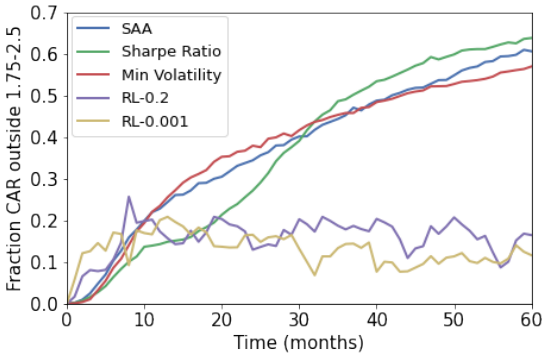
Figure 11: 5% Value-at-Risk and Conditional Value-at-Risk of all strategies.

do not exhibit this property. Moreover, the RL strategies do violate the RBC preference boundaries more often in the first 5 months, after which the violation percentage stabilizes. After 20 months, the violation rate of the RL strategies is significantly lower than the other strategies. Another noteworthy aspect of Subfigure 12(a) is that the violation rate of the RL-0.001 is comparable to the RL-0.2 model in the first 30 months and thereafter, significantly lower. This is the result of the lower value of the preference parameter α_1, which corresponds to letting the RL agent putting more emphasis on staying within the preference boundaries.

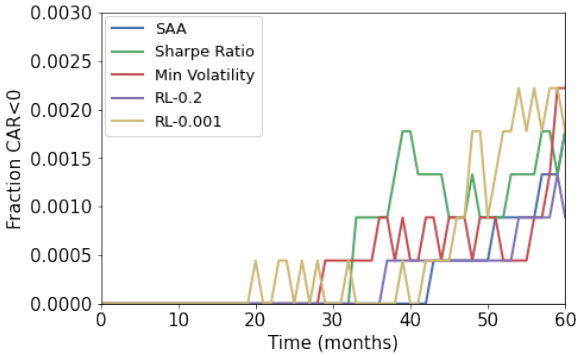
When looking at Subfigure 12(b), one should first realize that one jump corresponds to only 1 scenario. When considering the performance of the strategies, we conclude that the performance is similar across all methods. However, one could argue that the RL-0.02 model slightly outperforms the other models, although more scenarios should be considered to strengthen this statement.

When looking at Subfigure 12(c), one notes that the RL-strategies clearly outperform all benchmark strategies, especially after 30 months. However, the DMV-SR strategy has similar performance up until approximately 30 months, which can be explained by the following. On the one hand, the DMV-SR strategy has relatively high returns and low variance and therefore, the FR_t are high in this period. On the other hand, as can be seen in Subfigure 8(b), the DMV-SR strategy invests heavily in non-Equity asset classes in this period and consequently, the TRR_t are relatively low. Therefore, the CAR_t is relatively high and consequently, the probability of having a CAR_t which is below 1 is small. Another noteworthy aspect from Subfigure 12(c) is that all strategies have a very low probability of having a regulatory intervention in the first 10 months. In addition, notice that the RL-0.2 strategy outperforms the RL-0.001 strategy, which was expected since we put more emphasis on minimizing the penalty for violating the RBC requirement. Lastly, note that the DMV-SR model only outperforms the SAA benchmark in the first 40

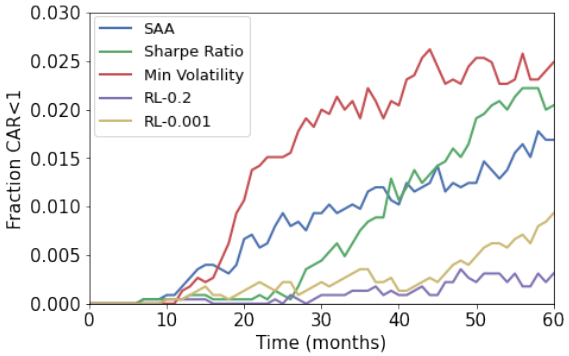
months. This can be explained by the fact that at the end of this period, the DMV-SR increases its allocation to equities, resulting in more variance and an increased probability of having a CAR which is lower than 1.



(a) CAR outside preference boundaries.



(b) CAR < 0.



(c) CAR < 1.

Figure 12: RBC requirements plots of all strategies.

To conclude, the RL strategies outperform all benchmark strategies in the long-run, with respect to both the wealth and RBC performance metrics. However, especially the DMV-SR strategy has superior performance in first 30 months when considering the wealth performance metrics. In this period, the DMV-SR strategy is able to cope with the poor performance of the Equity asset classes and to acquire high returns and low volatility by investing in non-Equity asset classes. Although the DMV strategies do not optimize over RBC requirements or preferences, the RBC performance of the DMV-SR model over this period is excellent as well due to investing in non-Equity asset classes. Lastly, the variance of the RL strategies is significantly higher compared to the benchmark strategies. However, this higher variance is mainly caused by the higher upward variance, since the downward variance of the RL strategies is similar to the benchmark strategies.

5.2 Results excluding transaction costs

This section is dedicated to the results of the considered strategies while excluding transaction costs. As in the previous subsection, we will first look at the allocation strategies and afterwards, a performance comparison will be provided.

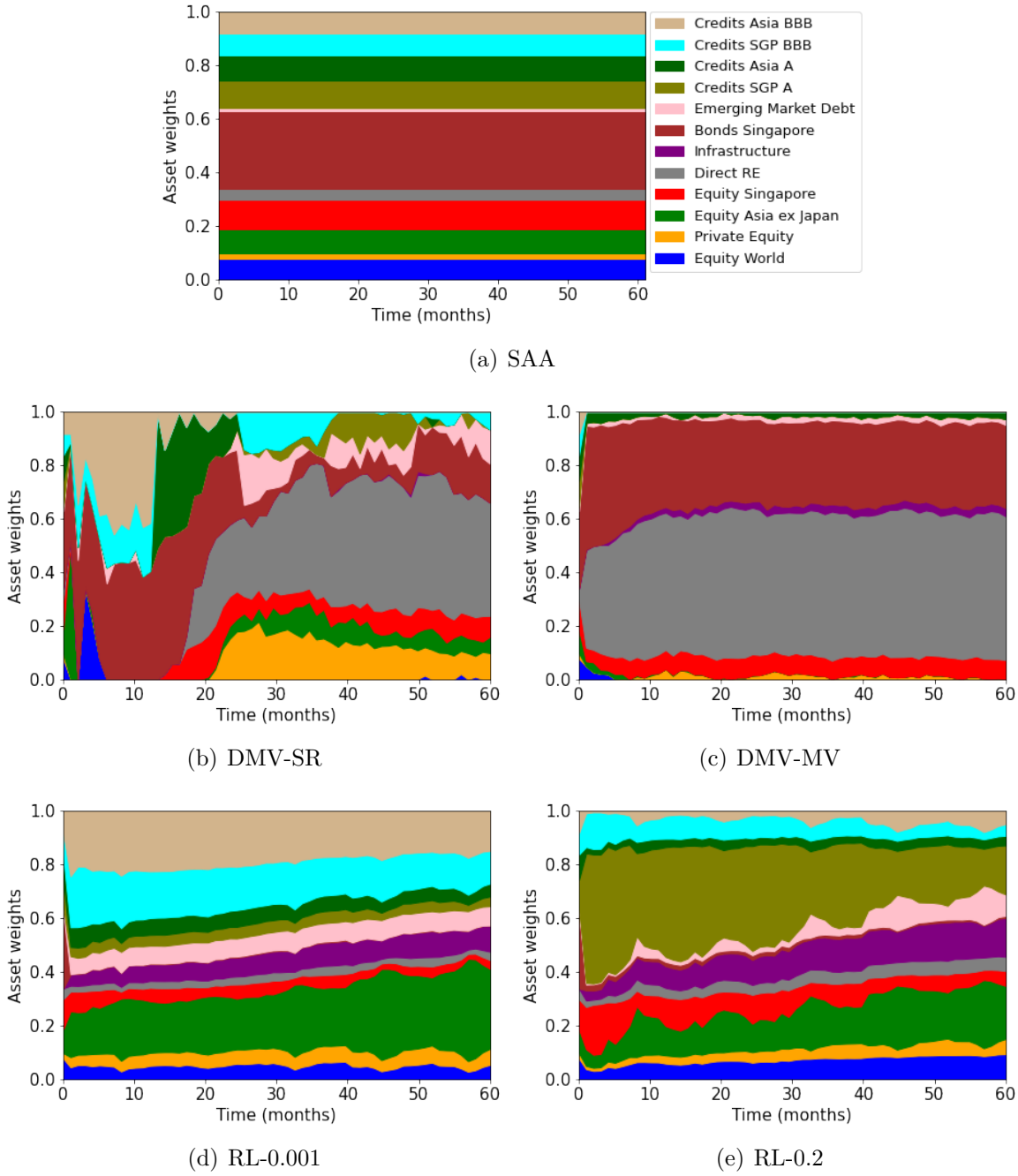
5.2.1 Allocation strategies

All asset allocation strategies are displayed in Figure 13. First note that to show the robustness of the RL model, we do not train the model separately for the case without transaction costs. In addition, as can be seen in Table 6 in Appendix E, the best performing DMV-SR and DMV-MV strategies are the ones that base their allocation at time t only on the scenarios at time t. Thus, the SAA and DMV-MV benchmark strategies are equivalent to the strategies displayed in Figure 8, where transaction costs were included. When comparing the average RL strategies in Figure 8 and 13, one might seem to notice that the strategies do not significantly differ from one another, however, the strategies do differ slightly on a scenario level. Again, when comparing Figure 3 and Subfigure 13(b), the DMV-SR strategy mimics the expected market development. In addition, as can be seen in Subfigure 13(b), the DMV-SR model is much more volatile compared to the DMV-SR strategy presented in Subfigure 8(b), although both strategies exhibit similar patterns. This volatile DAA is the best performing strategy since reallocation is not penalized by having transaction costs.

Although the DMV-SR strategy is able to mimic the expected market development, the applicability of this strategy is even less than the strategy presented in Subfigure 8(b). That is, the DMV-SR allocation strategy is too volatile to convince senior management and other stakeholders.

5.2.2 Performance evaluation

The average real wealth development is displayed in Figure 14. First note that as in Figure 9, a kink is visible around 25 months. Thus, all portfolio's still suffer from the poor performance of the Equity asset classes even when transaction costs are excluded. In addition, note that the DMV-SR portfolio dominates all portfolio strategies in these first months. Note that the outperformance with respect to the RL strategies is greater than in Figure 9, where transaction costs were included. This can be explained by the fact that the DMV-SR strategy benefits the most from the exclusion of transaction costs, since it is the most volatile strategy. However, similar as when transaction costs were considered, the RL strategies start dominating all other strategies after 30 months when the Equity asset classes perform better. Again, the DMV-MV portfolio strategy struggles with acquiring



(a) SAA

(b) DMV-SR

(c) DMV-MV

(d) RL-0.001

(e) RL-0.2

Figure 13: Asset allocation strategies excluding transaction costs in the Dec 2021 OFS excluding transaction costs.

enough nominal returns to outweigh the inflation in the first 20 months. Note that when comparing Figures 9 and 14, excluding transaction costs results in a significant increase in the expected final real wealth. This increase ranges from 0.5 Billion SGD for the DMV-MV strategy to 1 Billion SGD for the DMV-SR and RL strategies.

The volatility measures of all portfolio strategies are displayed in Figures 15 and 16. Again, one notices that the variance of the RL strategies is much higher compared to the

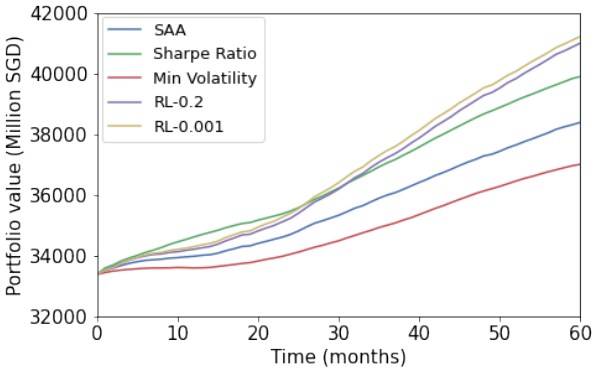
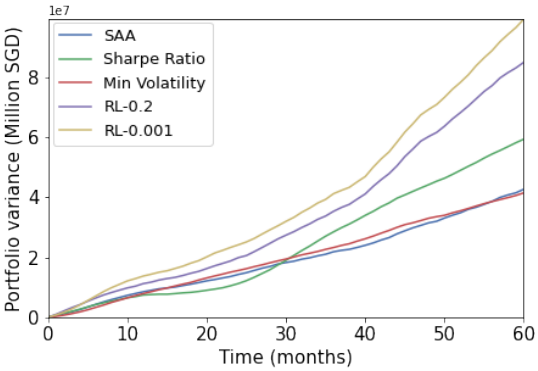


Figure 14: Average real wealth development of all strategies in the Dec 2021 OFS excluding transaction costs.

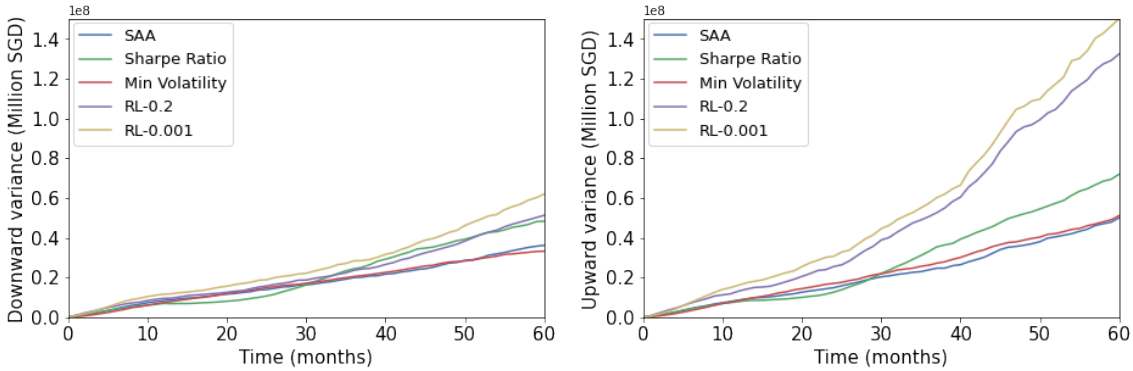
other strategies. Similar as with transaction costs, the DMV-SR model has relatively low variance in the first 30 months. In addition, the DMV-MV model does not display the expected low variance which one would expect. However, when looking at the Subfigures 15(b) and 15(c), the DMV-MV strategy does have the lowest downward variance in the long-run. Moreover, the downward variance of the DMV-SR and RL strategies are much more similar. Especially the RL-0.2 strategy has a similar downward variance, with the exception of the months $10 - 30$ where the DMV-SR dominates. In addition, the upward variance of the RL strategies displayed in Subfigure 15(c) is significantly higher than all benchmark strategies. Thus, similar as in the setting with transaction costs, the higher overall variance of the RL strategies is mainly caused by the higher upward variance and not by the downward variance.

Now, when looking at Figure 16, again the 5% VaR and CVaR of the RL strategies are comparable to the benchmark strategies. However, the DMV-SR benchmark still has a superior VaR and CVaR in the months $10 - 30$, caused by the high expected returns and low downward variance. However, when comparing Figures 11 and 16, one notices that the RL-0.2 strategy significantly outperforms the other strategies in the long-run when no transaction are considered.

Figure 17 displays the performance of all strategies regarding the RBC requirements and preferences. When looking at Subfigure 17(a), one notices that similar as in Subfigure 12(a), the RBC preference violation rate of the benchmark strategies increase over time, while the RL strategies do not exhibit this pattern. In addition, the RL strategies do violate the RBC preference boundaries more often in the first 5 months, after which the violation rate stabilizes. Again, after approximately 20 months the violation rate of the RL strategies is significantly lower compared to the benchmark strategies. Similar as the results including transaction costs, Subfigure 17(a) shows that the RL-0.001 strategy has the lowest violation rate, especially after 30 months. When looking at Subfigure 17(b),



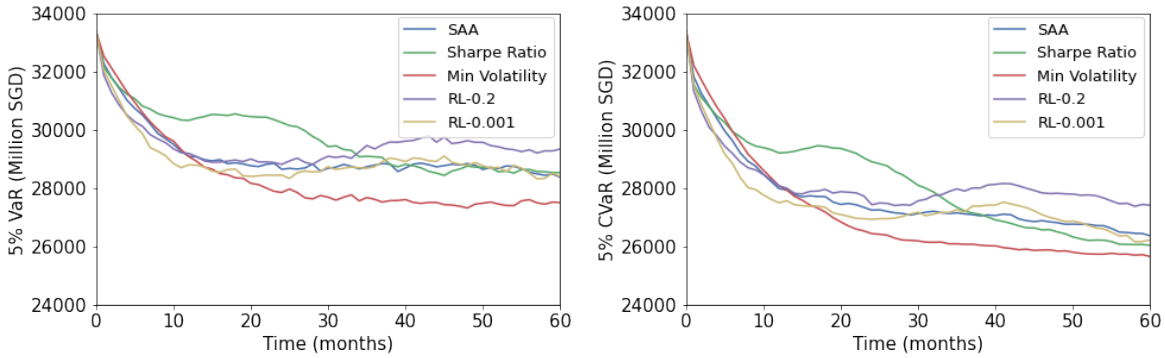
(a) Variance.



(b) Downward variance.

(c) Upward variance.

Figure 15: Variance plots of all strategies in the Dec 2021 OFS excluding transaction costs.



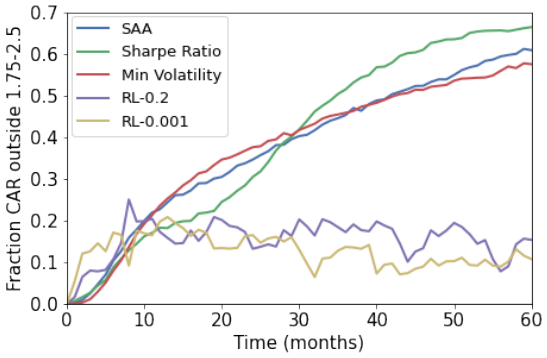
(a) 5% VaR.

(b) 5% CVaR.

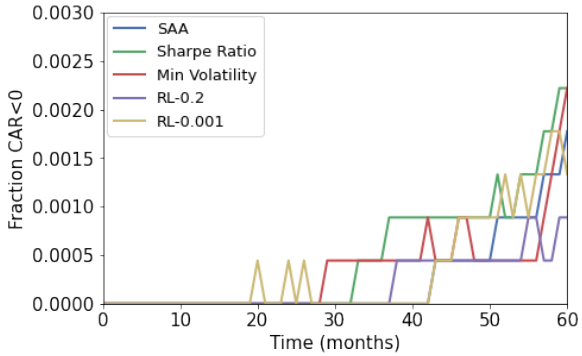
Figure 16: 5% Value-at-Risk and Conditional Value-at-Risk of all strategies in the Dec 2021 OFS excluding transaction costs.

we conclude that the performance of all strategies is again similar, mainly caused by the scarcity of the scenarios in which the CAR_t is below 0. Comparing the Subfigures 12(b) and 17(b), one denotes that having no transaction costs results in a lower probability of having a $\text{CAR}_t < 0$. This is an expected result since having less costs results in higher expected returns. Consequently, the insurer has higher FR_t and thus a higher CAR_t.

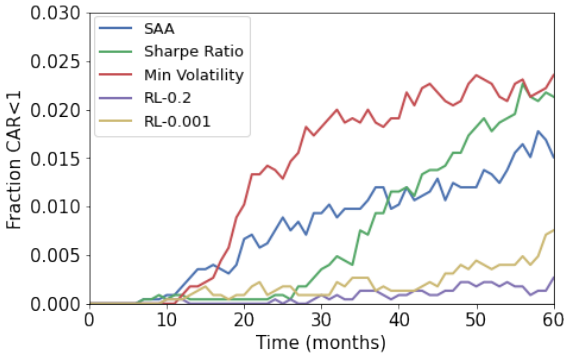
When looking at Subfigure 17(c), we denote a similar pattern as in Subfigure 12(c). First, all strategies have a very small probability (< 0.001) to violate the RBC requirements in the first 10 months. Second, all RL strategies clearly outperform all other strategies after approximately 30 months. While the DMV-SR and SAA benchmarks perform poorly after approximately 10 months, the DMV-SR strategy still has similar performance compared to the RL strategies until month 30. This result is once again explained by on the one hand, the high expected returns and low downward variance of the DMV-SR strategy in this period and on the other hand by the relatively high allocation to non-Equity asset classes. Again, notice that the RL-0.2 strategy, with more preference for a lower probability of violating the RBC requirements, outperforms the RL-0.001 strategy. Lastly, notice that the RBC requirement violation rate of the DMV-SR benchmark quickly increases after 30 months, resulting in worse performance compared to the SAA benchmark after 40 months. This result is explained by the increased allocation to equities as displayed in Figure 13(b) and consequently, having more downward variance in these periods.



(a) CAR outside preference boundaries.



(b) CAR < 0.



(c) CAR < 1.

Figure 17: RBC requirements plots of all strategies in the Dec 2021 OFS excluding transaction costs.

To conclude, the performance of the RL strategies is comparable to the previous section in which transaction costs were included. Again, all strategies outperform the

benchmark strategies in the long-run, regarding both the wealth and RBC performance metrics. However, an important difference compared to the case including transaction costs is that the DMV-SR strategy has substantially more expected real wealth. This is explained by the fact that the DMV-SR strategy is the most volatile strategy and thus, benefits the most from having no transaction costs.

5.3 Robustness test with transaction costs

This section is dedicated to the robustness test of the RL model. Throughout this section, we will make use of the May 2022 OFS. As stated in Section 3.2, the May 2022 OFS is only used for testing, not for training and validation. More specific, we use the RL model from the previous section which is trained and calibrated on the December 2021 data. On the contrary, the DMV models will be both trained and tested on the full May 2022 OFS, since these strategies use the actual scenarios at a specific time point to determine the asset allocation. Note that we will only look at the performance evaluation metrics in this section. However, the asset allocation strategies are displayed in Appendix G for completeness.

5.3.1 Performance evaluation with transaction costs

The wealth development of all strategies when considering the May 2022 OFS is displayed in Figure 18. As can be seen, a similar kink pattern as in Figure 9 is visible. An important difference however is that due to the high inflation in this period, the DMV-MV strategy has negative net returns in the first 15 months. Another important difference compared to Figure 9, is that the DMV-SR strategy outperforms the RL strategies over a longer period. Moreover, the RL-0.2 is almost completely dominated by the DMV-SR strategy over the whole time horizon. On the one hand, this result can be explained by the fact that the RL strategies have not been trained on the May 2022 OFS and thus, the performance of the RL agent is not optimal. On the other hand, the RL agent only gets a reward for acquiring returns when the CAR is between the specified preference bounds. Therefore, it could be the case that the RL agent had to invest in less risky assets to stay within these preference bounds, resulting in less expected net returns. To conclude, we note that the RL strategies have more difficulty with determining an asset-allocation which is suitable for acquiring high net returns in the May 2022 OFS. However, in the long-run it still seems that both RL strategies outperform all benchmark strategies.

In addition, the volatility measures are displayed in Figure 19, where the overall results are comparable to the ones displayed in Figure 10. The only noteworthy difference is that the variance of the DMV-SR strategy is larger in the May 2022 OFS. More specific, the downward variance displayed in Subfigure 19(b) is significantly higher compared to

Subfigure 10(b). The upward variance of all strategies displayed in Subfigure 19(c) is comparable to the December 2021 case.

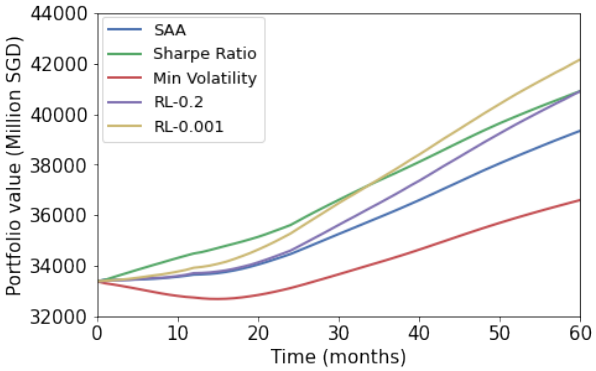
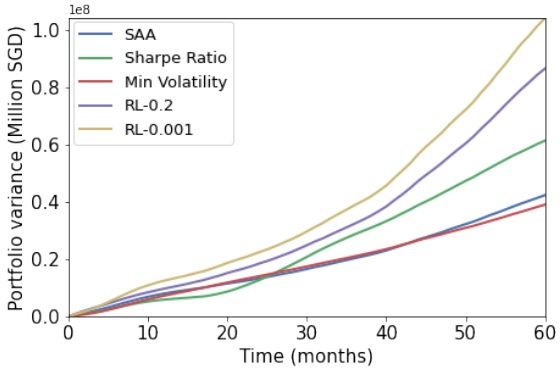
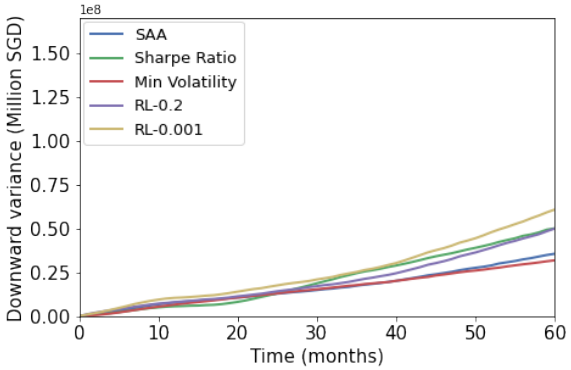


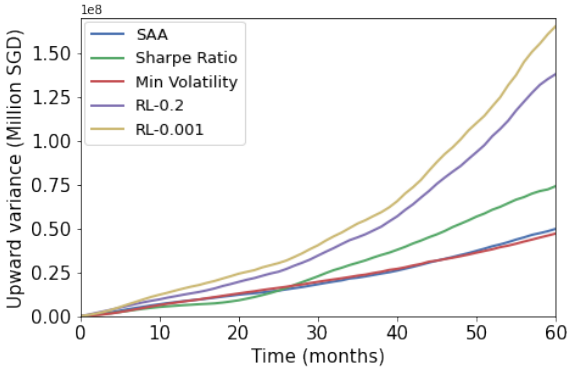
Figure 18: Average real wealth development of all strategies in the May 2022 OFS including transaction costs.



(a) Variance.



(b) Downward variance.

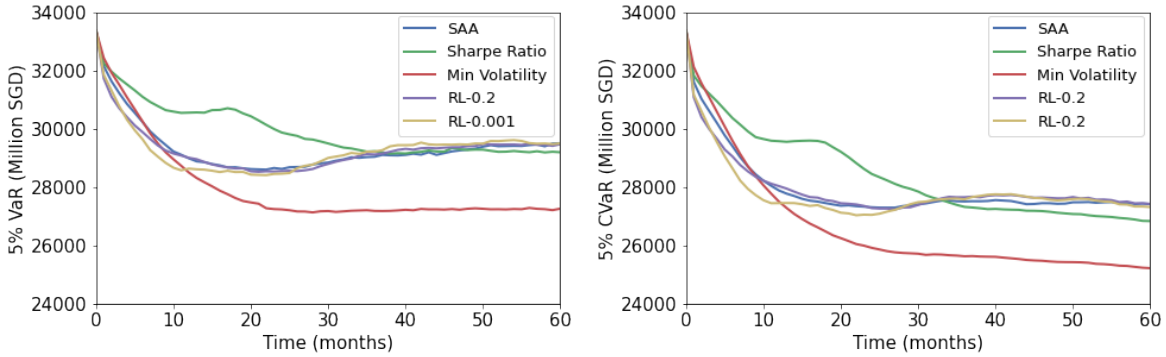


(c) Upward variance.

Figure 19: Variance plots of all strategies in the May 2022 OFS including transaction costs.

When looking at Figure 20, one notices that as in the previous sections, the VaR and CVaR of the DMV-SR strategy is superior to all strategies in the first 25 months. However,

a noteworthy difference compared to the December 2021 case, as displayed in Figure 11, is that the RL strategies converge to a similar VaR and CVaR as the SAA benchmark. In addition, the CVaR of the DMV-SR strategy seems to worsen over time in the long-run.



(a) 5% VaR.

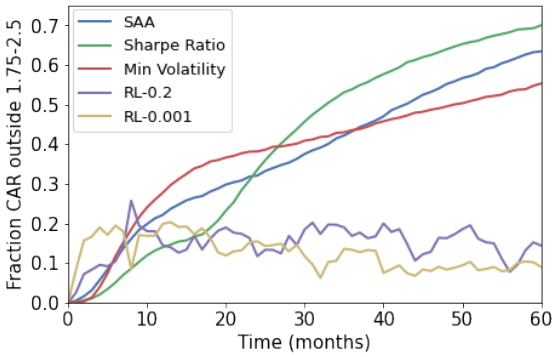
(b) 5% CVaR.

Figure 20: 5% VaR and CVaR of all strategies in the May 2022 OFS including transaction costs.

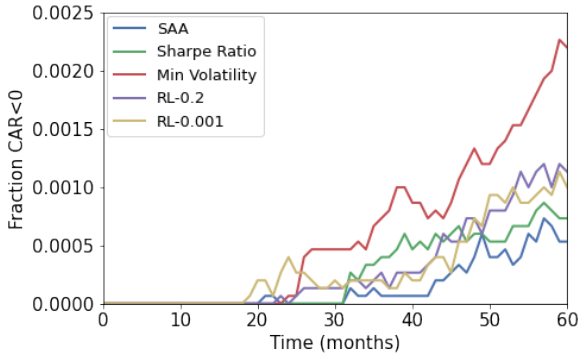
In addition, the RBC evaluation plots are displayed in Figure 21. The most noteworthy difference compared to the December 2021 OFS is that Subfigure 21(b) is significantly more informative, since we are considering substantially more scenarios. Consequently, there are more cases in which the CAR drops below 0. Similar as in Subfigure 12(b), the probability of going bankrupt in the first 14 months is small. However, a noteworthy difference compared to Subfigure 12(b) is that in Subfigure 21(b) the DMV-MV is performing very poorly, especially at the end of the considered time horizon. Moreover, the probability of going bankrupt of all other strategies is comparable, although the DMV-SR has a slightly higher probability between months $30 - 45$ and the RL strategies have a slightly higher probability at the end of the time horizon. When looking at Subfigure 21(c), we draw the same conclusion as in the December 2021 case. That is, the performance of the RL strategies is superior compared to all other strategies. Again, the RL-0.2 strategy slightly outperforms the RL-0.001 strategy in the long-run.

To conclude, most results of the May 2022 OFS are comparable to the December 2021 case. The most noteworthy difference is that the net returns of the DMV-SR strategy are significantly higher compared to the other strategies in the May 2022 OFS. Consequently, the DMV-SR strategy outperforms the RL strategies during a substantially longer period. However, the downward variance of the DMV-SR is significantly higher as well, resulting in less performance compared to the RL strategies in the long-run. With respect to the RBC requirements and preferences, the RL strategies still outperform the benchmarks on all evaluation metrics. Thus, in the case where we include transaction costs, the RL model trained on the December 2021 economy is considered to be robust when evaluating

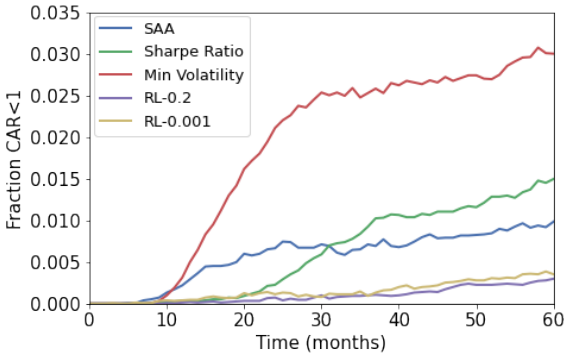
the performance on the May 2022 OFS.



(a) CAR outside preference boundaries.



(b) CAR < 0.



(c) CAR < 1.

Figure 21: RBC requirements and preferences plots of all strategies in the May 2022 OFS including transaction costs.

5.3.2 Performance evaluation excluding transaction costs

The wealth development of all strategies are displayed in Figure 22. Similar as in the previous section, the DMV-SR strategy outperforms the RL-0.2 strategy up until the end of the time horizon. Moreover, the DMV-SR strategy outperforms the RL-0.001 strategy until approximately $t = 40$. Again, the RL-0.001 strategy clearly dominates all other strategies in the long-run. Again, note that when comparing Figures 18 and 22, excluding transaction costs results in a significant increase in the expected final real wealth. This increase is similar as in Section 5.2.2 and ranges from approximately 0.5 Billion SGD for the DMV-MV strategy to 1 Billion SGD for the DMV-SR and RL strategies.

The volatility measures are displayed in Figure 23, where the overall results are similar to the previous volatility results displayed in Figures 10, 15 and 19. Again, the most noteworthy difference compared to the December 2021 OFS case is that the DMV-SR strategy has substantially more variance when considering the May 2022 OFS, caused by the increased downward variance.

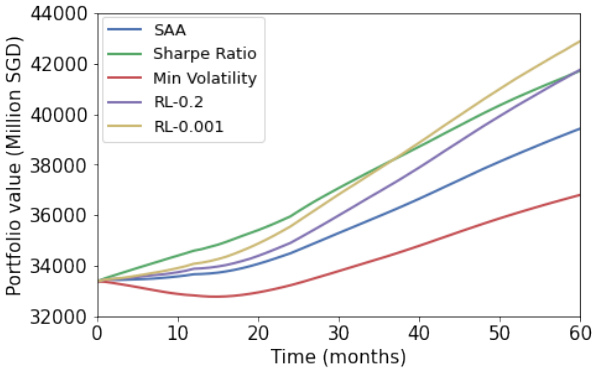
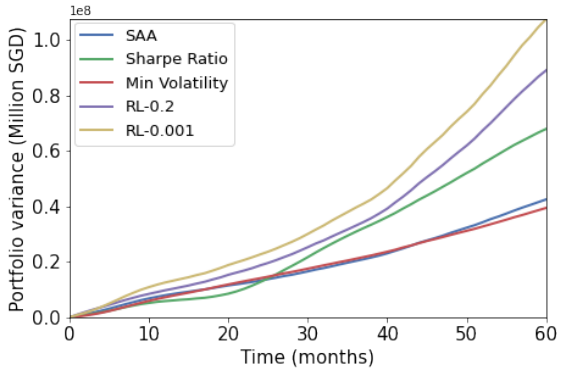
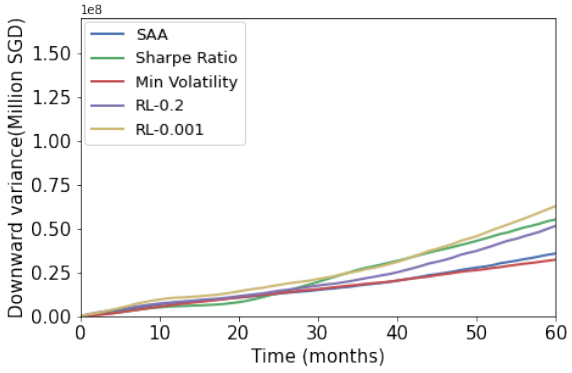


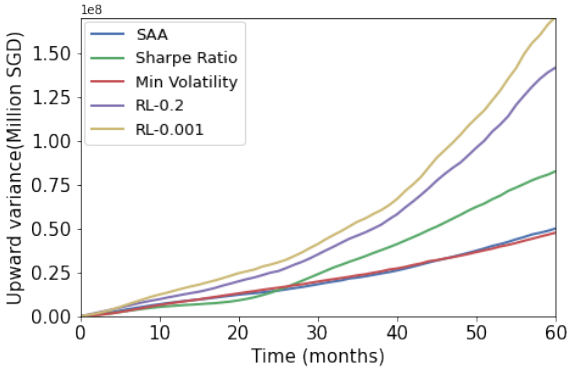
Figure 22: Average real wealth development of all strategies in the May 2022 OFS excluding transaction costs.



(a) Variance.



(b) Downward variance.

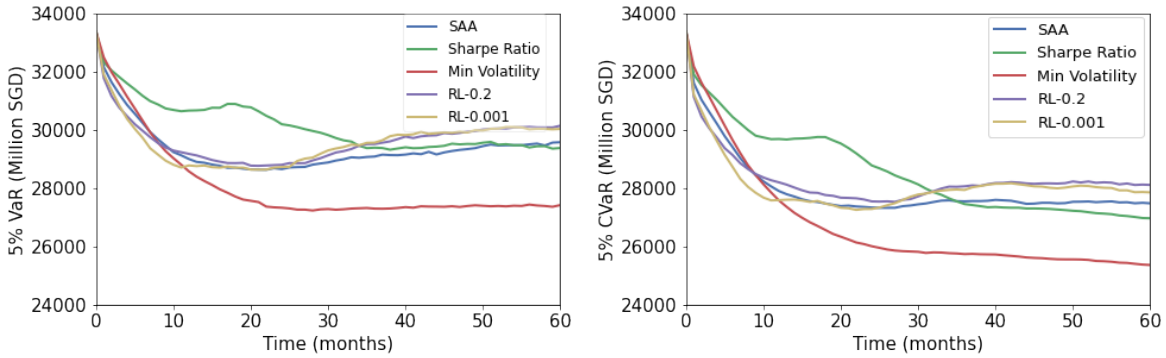


(c) Upward variance.

Figure 23: Variance plots of all strategies in the May 2022 OFS excluding transaction costs.

The 5% VaR and CVaR are displayed in Figure 24. When comparing these risk measure results to the ones displayed in Figure 16, we again notice the dominance of the DMV-SR strategy in the first 25 months. However, after approximately 35 months, both RL strategies start to dominate the benchmark strategies. Moreover, notice that when comparing the VaR and CVaR in the long-run, the SAA seems to outperform the

DMV-SR strategy.

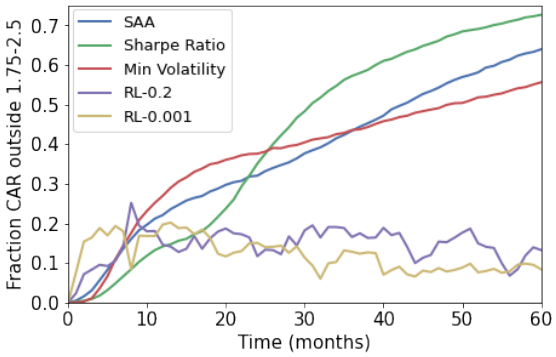


(a) 5% VaR.

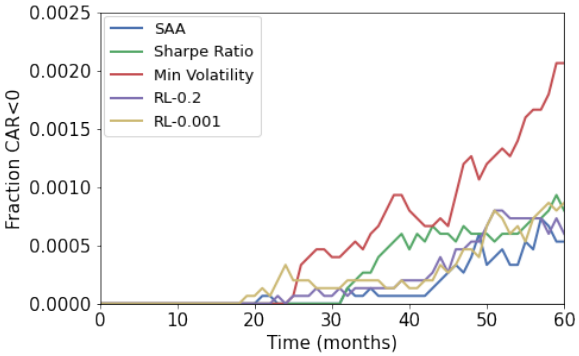
(b) 5% CVaR.

Figure 24: 5% VaR and CVaR of all strategies in the May 2022 OFS excluding transaction costs.

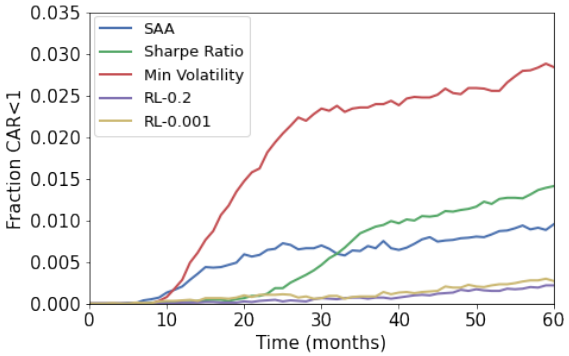
Figure 25 displays the performance of all strategies regarding the RBC requirements and preferences. When looking at Subfigure 25(a), one notices that similar as in Subfigures 12(a), 17(a) and 21(a), the RBC preference violation percentage of the benchmark strategies increase over time, while the RL strategies do not exhibit this property. Again, the RL strategies have a lower violation rate in the long-run, while having a higher violation rate in the first months. Moreover, the RL-0.001 strategy has the lowest preference violation probability in the long-run. When looking at Subfigure 25(b), one notices that the performance of the SAA, DMV-SR and RL strategies is similar. Only the DMV-MV strategy clearly deviates from the other strategies and shows significantly less performance in the long-run. As in Subfigure 21(b), the DMV-SR strategy has a slightly higher probability of going bankrupt in the months $35-50$. Subfigure 25(c) displays the expected probability of having a $\text{CAR}_t < 1$. Again, all strategies have a very small probability of not complying with the RBC requirements in the first 10 months. In addition, the DMV-SR strategy still has similar performance compared to the RL strategies until $t = 30$, while the SAA and DMV-SR benchmarks already perform poorly after approximately 10 months. Especially the DMV-MV benchmark has an increased violation probability compared to the other strategies. As before, the RL-0.2 strategy clearly outperforms all other strategies, although the difference compared to the RL-0.001 strategy is smaller than before. Lastly, similar as in the December 2021 OFS, the probability of not complying with the RBC requirements is less when we exclude transaction costs.



(a) CAR outside preference boundaries.



(b) CAR < 0.



(c) CAR < 1.

Figure 25: RBC requirements plots of all strategies in the May 2022 OFS excluding transaction costs.

To conclude, the performance of the RL strategies in the May 2022 OFS while excluding transaction costs are comparable to the performance in the December 2021 OFS. Again, the most noteworthy difference are the high net returns of the DMV-SR strategy in the short-run. However, these higher net returns are accompanied by severely more downward variance. With respect to the RBC requirements and preferences, the RL strategies still outperform the benchmarks on all evaluation metrics. Lastly, since the RL model is not trained on the May 2022 economy, the performance of the RL strategies in this section should be considered as a lower bound of the potential performance.

6 Discussion

6.1 Conclusion

In this paper, we investigated the possibilities of using a Reinforcement Learning model to optimize the Dynamic Asset Allocation of an insurer, while extending the traditional risk-return trade-off. More specifically, two RL models with different preference parameters for complying with the Risk Based Capital requirements and preferences are proposed. In addition, we considered two data sets both having 15000 simulated scenarios, one with the economic outlook of December 2021 and one with the economic outlook of May 2022. In this setting, we intended to answer the following research question:

> *How can insurers optimize their Dynamic Asset Allocation, incorporating Risk Based Capital requirements in addition to the risk-return trade-off?*

We answered this research question by introducing an RL model which is able to optimize over both expected net returns and RBC requirements and preferences. Moreover, we compared the performance of the RL strategies to three benchmark strategies, namely a Strategic Asset Allocation and two Dynamic Mean-Variance strategies. This research found some evidence that the RL agent is able to come up with a DAA which outperforms all benchmark strategies in the long-run. However, in the first 30 months, the DMV-SR strategy showed superior performance with respect to the traditional risk-return trade-off and comparable performance regarding the RBC requirements. We should, however, point out that the similar RBC performance was accidentally caused by the high allocation to and performance of the non-Equity asset classes in this period. Moreover, by allocating more wealth to the non-equity asset classes, the TRR is lower and consequently, the CAR is higher. Lastly, note that after 30 months, the performance of the DMV-SR strategy with respect to both the RBC requirements and preferences worsened very quickly in all considered settings. In this period, the performance of the RL strategies became superior to all benchmark strategies.

Additionally, this research looked at the applicability of the proposed DAA strategies. We concluded that the DMV strategies are too volatile or too non-diversified and thus, do not show potential to be used in practice. On the other hand, the SAA and RL strategies do show practical applicability.

In addition, the results when including or excluding transaction costs were not significantly different. The only difference was that the DMV-SR strategy had substantially higher expected wealth when excluding transaction costs, due to the volatile nature of the strategy and consequently, having the most benefit from having no transaction costs. In addition, note that the RL model only has been trained once, using the December 2021

OFS and including 0.15% transaction costs. We decided to only train the model once to show the robustness of the RL model in different settings. However, one could have trained a separate model for every setting to better match the market dynamics in these different environments. Thus, the performance of the RL models in especially the May 2022 OFS are a lower bound to the potential performance.

Another important finding from this research is that the relationship between the asset allocation and the CAR is rather complex and time sensitive. On the one hand, adjusting the asset allocation at time t has a direct impact on the TRR at time t, where the effect depends on the required risk capital of the new strategy. For example, substantially substituting the allocation to Equity asset classes for FI asset classes results in a direct decrease of the TRR. On the other hand however, adjusting the asset allocation at time t has both short- and long-term effects on the FR. More specifically, allocating more wealth to more volatile asset classes results in heavier fluctuations in the FR in the short-term. Therefore, also the CAR experiences heavy fluctuations in the short-term. In the long-run however, investing in more volatile asset classes usually results in higher expected returns and therefore, a higher expected FR and CAR.

To conclude, our proposed RL model shows the potential to solve the DAA optimization problem of an insurer due to the combination of high performance on the evaluation metrics and the practical applicability of the proposed strategies. In addition, RL has shown to be capable of optimizing over a complex investment environment, where the traditional risk-return trade-off is extended by non-linear objectives. However, more research should be dedicated to these type of problems to confirm the findings in this research and to make RL more attractable to financial institutions. Therefore, the following and final section of this research will present possibilities for future research.

6.2 Future research

One of the greatest limitations of this research is that we simplified the calculations for the Interest Rate market shock. More specifically, we assumed that underneath every FI asset class, there is a constant mix of different durations which cannot be adjusted by the RL agent. Consequently, the Interest Rate market shocks displayed in Table 1 are constant over time and across asset classes. Therefore, a possibility for future research could be to divide every FI asset class into multiple sub asset classes with different durations. This way, one would be able to better model the Interest Rate risk market shock. In addition, this is a relatively easy extension which would contribute to the practical applicability of this model. Unfortunately, data and time restrictions did not allow us to incorporate it in this research.

Another interesting opportunity for future research would be to extend the current

model with currency hedging dynamics. As of now, we assumed that the FI asset classes are fully hedged and that the other asset classes are exposed to currency risk. However, currency hedging is accompanied with extra costs which could potentially limit net returns. Therefore, it would be interesting to extend the current model such that the RL agent is able to decide whether the currency risk for an individual asset class will be hedged or not.

Another opportunity for future research would be to incorporate asset weight constraints to the optimization problem. In this research, we only assume that short-selling is not allowed, but no additional constraints are imposed. Although the average RL strategies are not as volatile over time as for instance the DMV-SR strategy, it could still be beneficial to investigate what the consequence of introducing asset weight constraints will be on the performance of the RL strategies. Introducing asset weight constraints could also be beneficial for the practical applicability of RL, since one would be assured that the RL agent will not provide irrational allocation strategies.

Another limitation of this research is that we only consider RBC requirements and preferences in addition to the traditional risk-return trade-off. However, as pointed out by the client of Ortec-Finance, liquidity is another crucial aspect which an insurer has to consider. More specifically, an insurer needs to allocate a sufficient amount of its wealth to asset classes which can be easily transformed into cash. Although we did not consider liquidity in this research, it would be interesting to see how the RL strategies would differ when one would incorporate it in the model. A simple extension would be to add a penalty to the reward function of the RL agent for asset classes which are very illiquid. This way, one could force the RL agent to give more preference to liquid asset classes.

Another possibility for future research would be to consider a different time horizon of for instance 10 or 30 years. Since the RL strategies start dominating the benchmark strategies at the end of the considered time horizon, it would be interesting to see whether this would continue to hold. Moreover, one could also consider quarterly or yearly time steps instead of monthly time steps.

Another limitation of this study is that we used a standard liability structure to model the liabilities. Therefore, an interesting opportunity for future research would be to consider liability scenarios or a setting in which the RL agent is able to adjust the liabilities. For instance, one could consider an RL model in which the agent is able to increase the liabilities when the CAR is above 2.5.

With regard to general RL, an interesting opportunity for future research would be to investigate the relationship between the number of iterations over the data set and the complexity of the reward/objective function. During this research, we have noticed that as we extended the implementation of the RBC requirements, the RL agent required more iterations to find an optimal policy. After the final implementation, we needed 30 iterations over the whole data set before convergence in rewards took place. Thus, it would

be interesting to investigate whether the number of iterations over the data set correlates with the complexity of the objective function.

Finally, it would be interesting to explore the implementation of RBC (or other solvency frameworks) into other actor-critic methods such as DDPG or TD3. These methods have the advantage that they are more sample efficient which might increase the performance of the RL agent. However, as mentioned in Section 2.1.3, Aboussalah et al. (2022) have shown that PPO outperforms other actor-critic methods in portfolio management applications. Nevertheless, it would be interesting to see whether this conclusion still holds when considering a more complex investment environment, such as incorporating RBC requirements and preferences.

References

Aboussalah, A. M., Xu, Z., & Lee, C.-G. (2022). What is the value of the cross-sectional approach to deep reinforcement learning? *Quantitative Finance*, *22*(6), 1091–1111.

Achiam, J. (2018). *Spinning up in deep reinforcement learning.* https://spinningup.openai.com (accessed: 08-08-2022)

Almahdi, S., & Yang, S. Y. (2017). An adaptive portfolio trading system: A risk-return portfolio optimization using recurrent reinforcement learning with expected maximum drawdown. *Expert Systems with Applications*, *87*, 267–279.

Andrychowicz, M., Raichuk, A., Stańczyk, P., Orsini, M., Girgin, S., Marinier, R., Hussenot, L., Geist, M., Pietquin, O., Michalski, M., et al. (2020). What matters in on-policy reinforcement learning? a large-scale empirical study. *arXiv preprint arXiv:2006.05990*.

Bellman, R. (1966). Dynamic programming. *Science*, *153*(3731), 34–37.

Berner, C., Brockman, G., Chan, B., Cheung, V., Dębiak, P., Dennison, C., Farhi, D., Fischer, Q., Hashme, S., Hesse, C., et al. (2019). Dota 2 with large scale deep reinforcement learning. *arXiv preprint arXiv:1912.06680*.

Bielstein, P., & Hanauer, M. X. (2019). Mean-variance optimization using forward-looking return estimates. *Review of Quantitative Finance and Accounting*, *52*(3), 815–840.

Botvinick, M., Ritter, S., Wang, J. X., Kurth-Nelson, Z., Blundell, C., & Hassabis, D. (2019). Reinforcement learning, fast and slow. *Trends in cognitive sciences*, *23*(5), 408–422.

Boyd, S., Boyd, S. P., & Vandenberghe, L. (2004). *Convex optimization*. Cambridge university press.

Corazza, M., & Bertoluzzo, F. (2014). Q-learning-based financial trading systems with applications. *University Ca'Foscari of Venice, Dept. of Economics Working Paper Series No*, *15*.

Filos, A. (2019). Reinforcement learning for portfolio management. *arXiv preprint arXiv:1909.09571*.

Fischer, T. G. (2018). *Reinforcement learning in financial markets-a survey* (tech. rep.). FAU Discussion Papers in Economics (No. 12/2018).

Fujimoto, S., Hoof, H., & Meger, D. (2018). Addressing function approximation error in actor-critic methods. *International conference on machine learning*, 1587–1596.

Grondman, I., Busoniu, L., Lopes, G. A., & Babuska, R. (2012). A survey of actor-critic reinforcement learning: Standard and natural policy gradients. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *42*(6), 1291–1307.

Guo, Y., Fu, X., Shi, Y., & Liu, M. (2018). Robust log-optimal strategy with reinforcement learning. *arXiv preprint arXiv:1805.00205*.

Infanger, G. (2008). Dynamic asset allocation strategies using a stochastic dynamic programming aproach. *Handbook of asset and liability management*, *1*, 199–251.

Jiang, Z., Xu, D., & Liang, J. (2017). A deep reinforcement learning framework for the financial portfolio management problem. *arXiv preprint arXiv:1706.10059*.

Kempf, A., Korn, O., & Saßning, S. (2015). Portfolio optimization using forward-looking information. *Review of Finance*, *19*(1), 467–490.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Konda, V. R., & Tsitsiklis, J. N. (2003). Onactor-critic algorithms. *SIAM Journal on Control and Optimization*, *42*(4), 1143–1166.

Li, Y., & Forsyth, P. A. (2019). A data-driven neural network approach to optimal asset allocation for target based defined contribution pension plans. *Insurance: Mathematics and Economics*, *86*, 189–204.

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., & Wierstra, D. (2015). Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*.

Markowitz, H. (1952). Portfolio selection. *Journal of Finance*, *7*(1), 77–91.

Markowitz, H. (1959). *Portfolio selection: Efficient diversification of investments*. Yale University Press.

Martin, R. A. (2021). Pyportfolioopt: Portfolio optimization in python. *Journal of Open Source Software*, *6*(61), 3066.

MAS. (2020). *Insurance statistics 2016-2020*. https://www.mas.gov.sg/statistics/insurance-statistics/annual-statistics/insurance-statistics-2020 (accessed: 09-08-2022)

MAS. (2022). *Notice 133 valuation and capital framework for insurers*. https://www.mas.gov.sg/regulation/notices/notice-133 (accessed: 01-07-2022)

Merton, R. C. (1969). Lifetime portfolio selection under uncertainty: The continuous-time case. *Review of Economics and Statistics*, 247–257.

Milliman. (2019). *Life insurance capital regimes in asia: Comparative analysis and implications of change, summary report*. https://us.milliman.com/-/media/milliman/importedfiles/ektron/life-capital-regimes-asia.ashx (accessed: 11-08-2022)

Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., Silver, D., & Kavukcuoglu, K. (2016). Asynchronous methods for deep reinforcement learning. *International conference on machine learning*, 1928–1937.

Moody, J., Wu, L., Liao, Y., & Saffell, M. (1998). Performance functions and reinforcement learning for trading systems and portfolios. *Journal of Forecasting*, *17*(5-6), 441–470.

Munk, C. (2017). *Dynamic asset allocation*. Lecture notes. https://sites.google.com/view/clausmunk/teaching. (accessed: 05-08-2022)

NAIC. (2022). *Risk-based capital*. https://content.naic.org/cipr-topics/risk-based-capital (accessed: 04-07-2022)

Ndlovu, B., Faisa, F., Resatoglu, N. G., & Türsoy, T. (2018). The impact macroeconomic variables on stock returns: A case of the johannesburg stock exchange. *Romanian Statistical Review*, (2).

Neuneier, R. (1995). Optimal asset allocation using adaptive dynamic programming. *Advances in Neural Information Processing Systems*, *8*.

Ortec-Finance. (2022). Asset class coverage: Overview of available asset classes, benchmarks and funds. Internal document.

Petraki, A. (2020). *The transaction costs manual: What is behind transaction cost figures and how to use them*. https://www.schroders.com/en/sysglobalassets/digital/insights/2020/october/transaction-costs/the-transaction-costs-manual/ (accessed: 13-08-2022)

Raffin, A., Hill, A., Gleave, A., Kanervisto, A., Ernestus, M., & Dormann, N. (2021). Stable-baselines3: Reliable reinforcement learning implementations. *Journal of Machine Learning Research*, *22*, 1–8.

Schulman, J., Levine, S., Abbeel, P., Jordan, M., & Moritz, P. (2015a). Trust region policy optimization. *International conference on machine learning*, *37*, 1889–1897.

Schulman, J., Moritz, P., Levine, S., Jordan, M., & Abbeel, P. (2015b). High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. (2016). Mastering the game of go with deep neural networks and tree search. *Nature*, *529*(7587), 484–489.

Steehouwer, H. (2016). Ortec finance scenario approach. https://www.ortecfinance.com/en/insights/whitepaper-and-report/ortec-finance-scenario-approach. (accessed: 01-08-2022)

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT Press. Second edition.

Sutton, R. S., McAllester, D., Singh, S., & Mansour, Y. (1999). Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, *12*.

Watkins, C. J., & Dayan, P. (1992). Q-learning. *Machine learning*, *8*(3), 279–292.

Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, *8*(3), 229–256.

Xu, J., Chen, J., & Chen, S. (2021). Efficient opponent exploitation in no-limit texas hold'em poker: A neuroevolutionary method combined with reinforcement learning. *Electronics*, *10*(17), 2087.

Yogeswaran, M., & Ponnambalam, S. (2012). Reinforcement learning: Exploration–exploitation dilemma in multi-agent foraging task. *Opsearch*, *49*(3), 223–236.

Yu, P., Lee, J. S., Kulyatin, I., Shi, Z., & Dasgupta, S. (2019). Model-based deep reinforcement learning for dynamic portfolio optimization. *arXiv preprint arXiv:1901.08740*.

Appendix A - Acronyms

Table 4 contains the glossary of the acronyms used throughout this paper.

Abbreviation	Description
CAR	Capital Adequacy Ratio
CF	Cash Flows
CVaR	Conditional Value-at-Risk
DAA	Dynamic Asset Allocation
DMV	Dynamic Mean-Variance
DMV-MV	DMV Sharpe Ratio
DMV-SR	DMV Minimum Volatility
EMD	Emerging Market Debt
ESG	Economic Scenario Generator
FI	Fixed Income
FR	Financial Resources
GAE	General Advantage Estimation
MAS	Monetary Authority of Singapore
ME	Macro-Economic variables
NN	Neural Network
OFS	Ortec-Finance Scenario set
OR	Operational Risk
PPO	Proximal Policy Optimization
RBC	Risk Based Capital
RL	Reinforcement Learning
RL-0.001	RL model with preference parameter set to 0.001
RL-0.2	RL model with preference parameter set to 0.2
SAA	Strategic Asset Allocation
SGD	Singapore Dollar
SGP	Singapore
TRR	Total Risk Requirements
VaR	Value-at-Risk

Table 4: Alphabetically ordered acronyms used throughout this paper with their descriptions.

Appendix B - Additional descriptive statistics

Figures 26, 27 and 28 contain additional descriptive statistics of the May 2022 OFS.

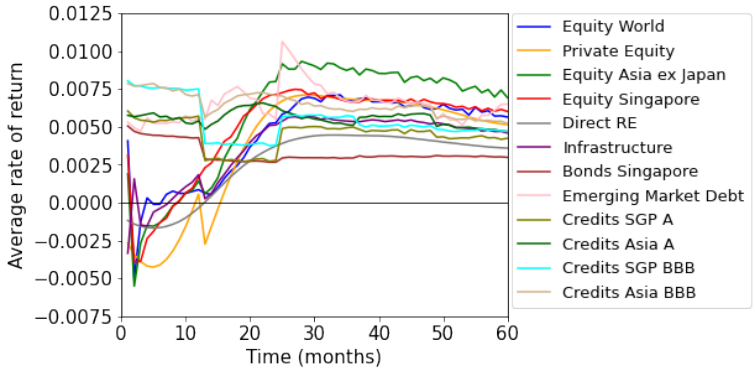


Figure 26: Average nominal scenario returns of all asset classes in the May 2022 OFS.

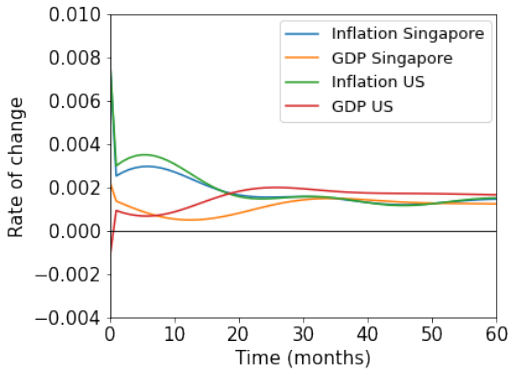


Figure 27: Average change of macroeconomic variables over scenarios in the May 2022 OFS.

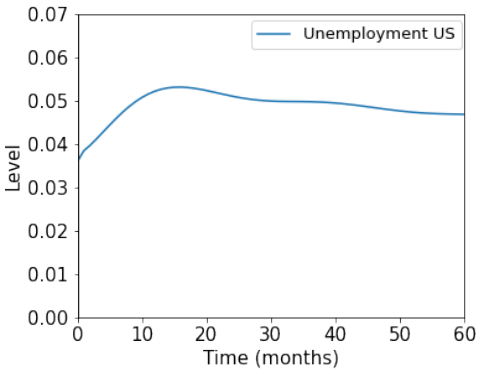
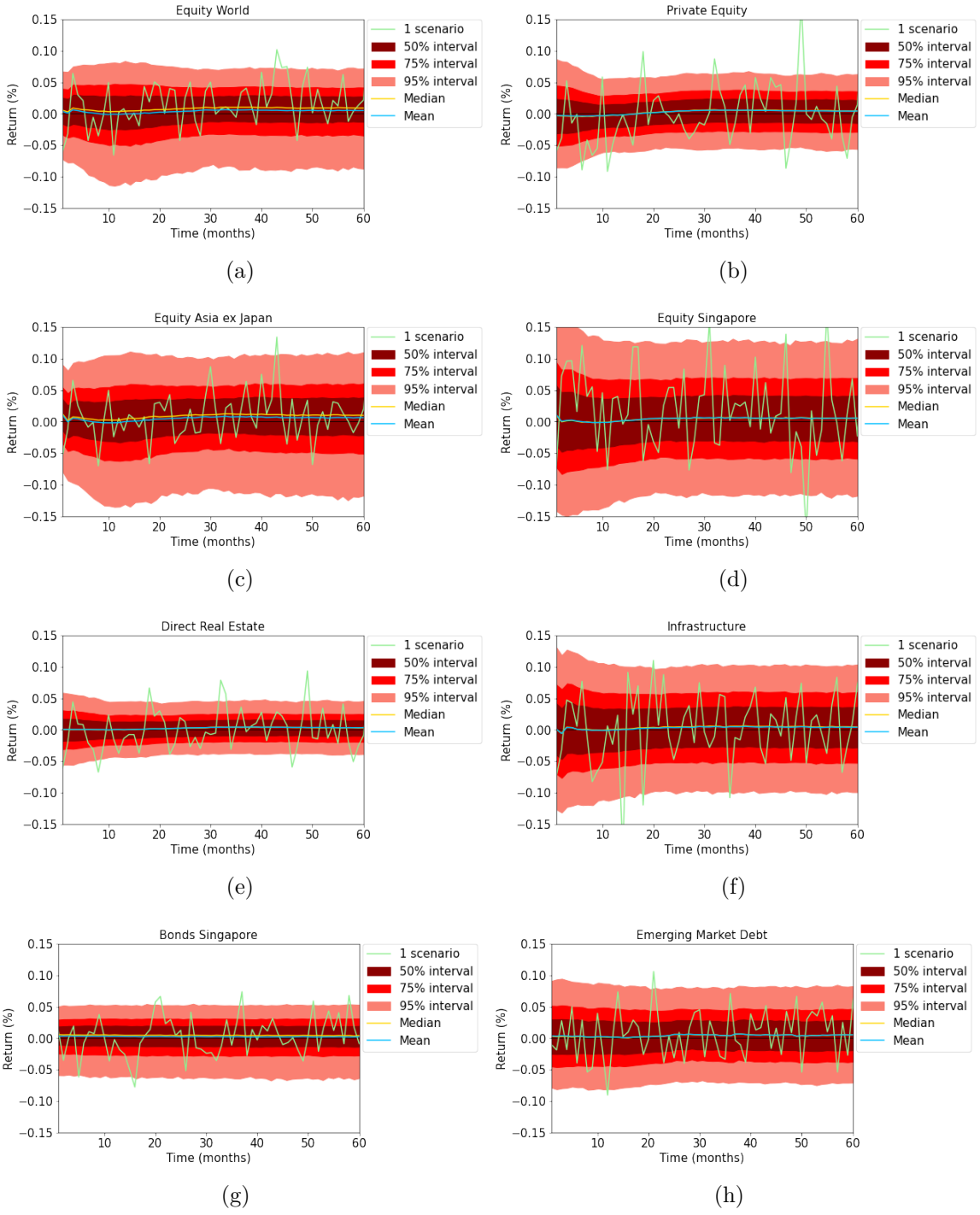


Figure 28: Average level of Unemployment US over all scenarios in the May 2022 OFS.

Appendix C - Fan charts of individual asset classes

This section contains the fan charts of all individual asset classes. The fan charts for the December 2021 and May 2022 OFS are displayed in Figures 29 and 30 respectively.



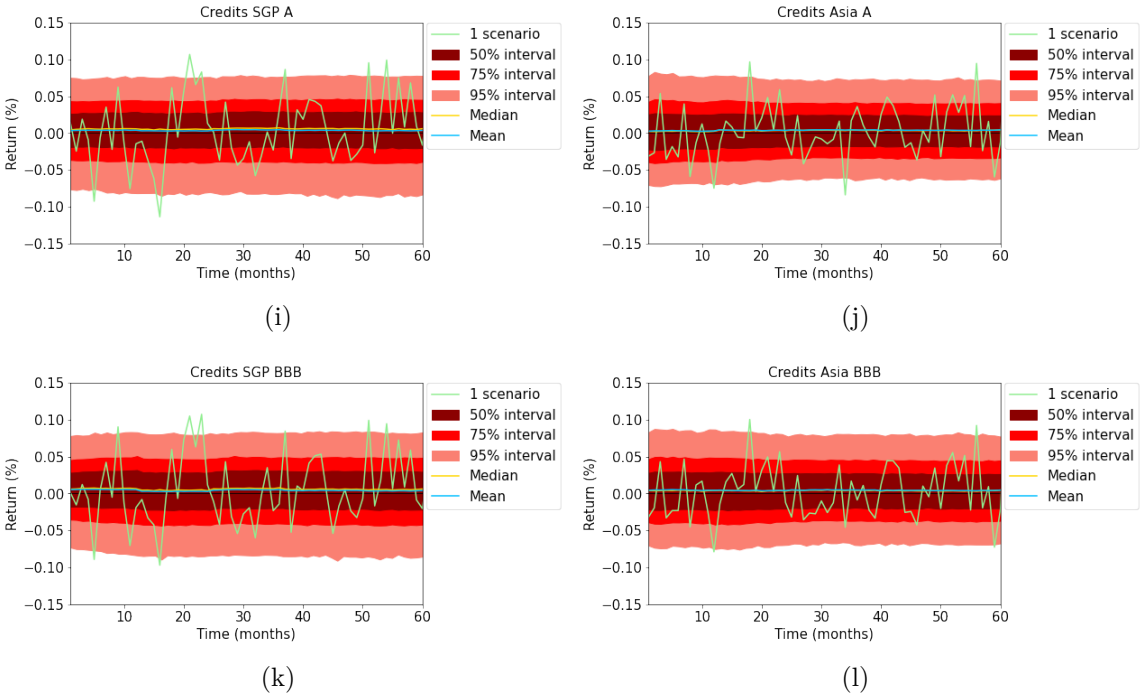
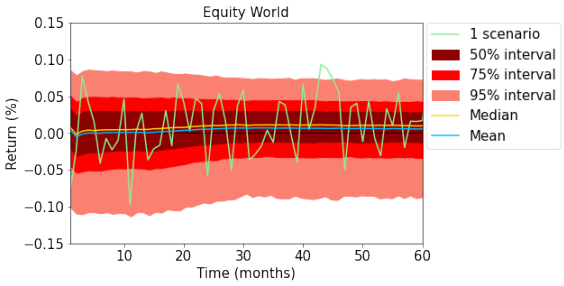
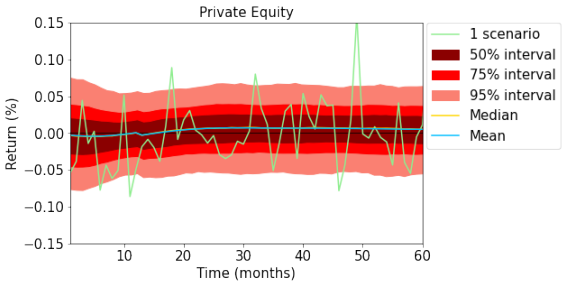


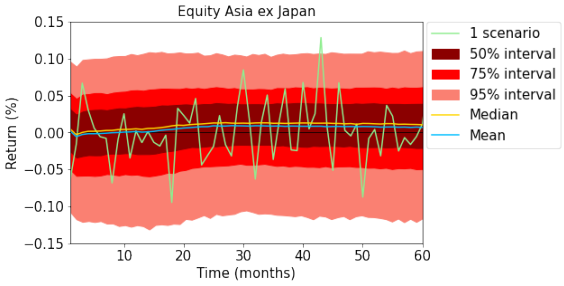
Figure 29: December 2021 OFS fan charts.


Equity World
Return (%)
Time (months)
1 scenario
50% interval
75% interval
95% interval
Median
Mean

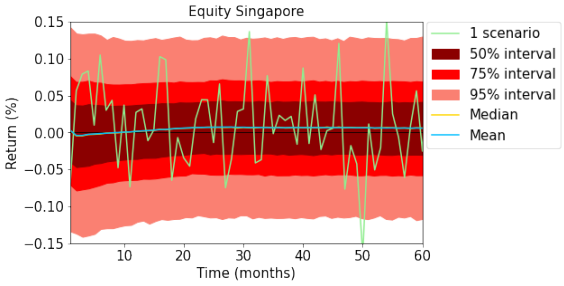
(a)


Private Equity
Return (%)
Time (months)
1 scenario
50% interval
75% interval
95% interval
Median
Mean

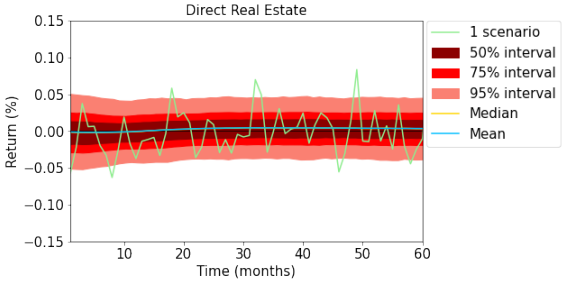
(b)


Equity Asia ex Japan
Return (%)
Time (months)
1 scenario
50% interval
75% interval
95% interval
Median
Mean

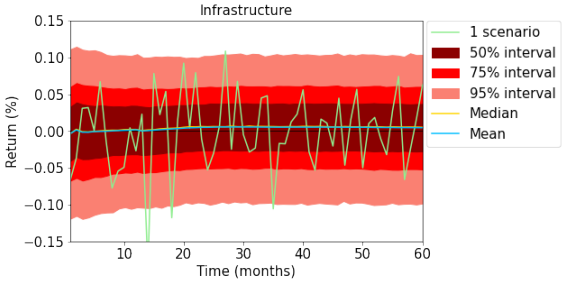
(c)


Equity Singapore
Return (%)
Time (months)
1 scenario
50% interval
75% interval
95% interval
Median
Mean

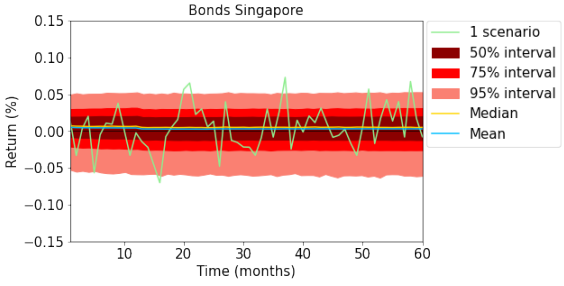
(d)


Direct Real Estate
Return (%)
Time (months)
1 scenario
50% interval
75% interval
95% interval
Median
Mean

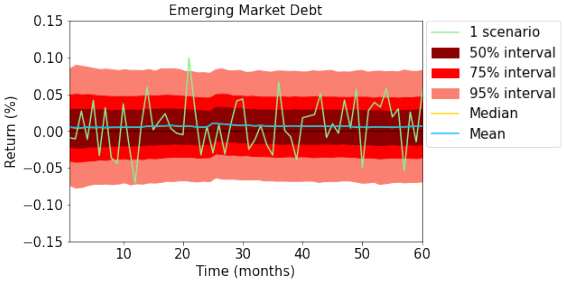
(e)


Infrastructure
Return (%)
Time (months)
1 scenario
50% interval
75% interval
95% interval
Median
Mean

(f)


Bonds Singapore
Return (%)
Time (months)
1 scenario
50% interval
75% interval
95% interval
Median
Mean

(g)


Emerging Market Debt
Return (%)
Time (months)
1 scenario
50% interval
75% interval
95% interval
Median
Mean

(h)

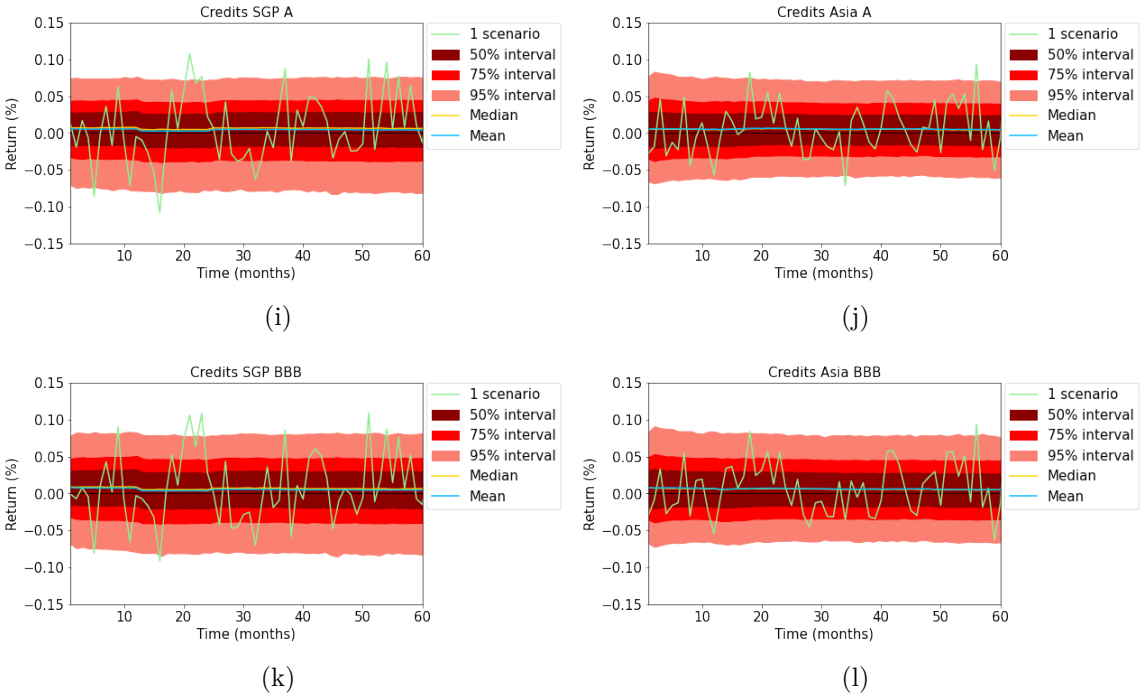


Figure 30: May 2022 OFS fan charts.

Appendix D - Pseudo-code

The pseudo-code of our PPO implementation is displayed in Algorithm 1. This pseudo-code is our extension to the pseudo-code presented by Achiam (2018).

Algorithm 1: Proximal Policy Optimization algorithm pseudo-code.

Input: initial policy parameters θ_0, initial value function parameters ϕ_0, the total number of iterations over the training data set K, the batch size B and the time horizon T. In addition, γ, λ and ϵ are hyperparameters.

Let e denote the number of policy parameter updates (epochs), set $e = 0$

for *k=1,2...,K* **do**

Shuffle scenarios in the train set

for *all scenarios batches of size B in the train set* **do**

for *i=1,2,...,B* **do**

for *t=1,2,...,T* **do**

Run policy π_{θ_e} to obtain states $s_{i,t}$, actions $a_{i,t}$ and value estimates $\mathrm{V}(s_{i,t})$ and calculate the reward $\hat{R}_{i,t}$

Add the state, action, log probability of the action, reward and value estimate to the memory.

end

end

Estimate the advantage $\hat{A}_{i,t}$ using the BT data entries in the memory:

$$\hat{A}_{i,t} = \delta_{i,t} + \gamma \hat{A}_{i,t+1}, \quad \forall t \in \{1, 2, ..., T\}, \forall i \in \{1, 2, ..., B\}, \tag{46}$$

Where:

$$\delta_{i,t} = \hat{R}_{i,t} + \gamma \mathrm{V}_{\phi_e}(s_{i,t+1}) - \mathrm{V}_{\phi_e}(s_{i,t+1}) \tag{47}$$

Use mini-batch optimization by sampling from the agent's memory and using the Adam solver to update the policy parameters by maximizing the PPO-clip objective:

$$\theta_{e+1} = \arg\max_{\theta} \frac{1}{BT} \sum_{i=1}^{B} \sum_{t=1}^{T} \min(r_{i,t}(\theta)\hat{A}_{i,t}, \mathrm{clip}(r_{i,t}(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_{i,t})] \tag{48}$$

Where:

$$r_{i,t}(\theta) = \frac{\pi_{\theta}(a_{i,t}|s_{i,t})}{\pi_{\theta_{\text{old}}}(a_{i,t}|s_{i,t})} \tag{49}$$

Fit the value function by regression on the mean-squared error:

$$\phi_{e+1} = \arg\max_{\phi} \frac{1}{BT} \sum_{i=1}^{B} \sum_{t=1}^{T} (\mathrm{V}_{\phi}(s_{i,t}) - \hat{R}_{i,t})^2 \tag{50}$$

Clear memory of the agent and set $e = e + 1$.

end

end

Appendix E - Benchmark models results

Tables 5 and 6 contain the benchmark model results of the model specification with and without transaction costs respectively.

		Wealth				CAR_T		
Model	**Hor.**	**Mean**	**Var**	**Min**	**Max**	**>0**	**>1**	**1.75-2.50**
SAA	na	38413	4.4E+07	953814	351010	0.9987	0.9840	0.3982
SR	1	38964	5.4E+07	1142440	306005	0.9982	0.9724	0.4520
	2	39089	5.4E+07	1035313	301611	0.9991	0.9742	0.4524
	3	39122	5.4E+07	1015795	295597	0.9991	0.9747	0.4516
	4	39122	5.4E+07	1056098	288298	0.9996	0.9747	0.4467
	5	39103	5.4E+07	1082680	283068	0.9996	0.9742	0.4462
	6	39075	5.4E+07	1084122	277484	0.9991	0.9729	0.4471
	12	38782	5.5E+07	1073292	277097	0.9996	0.9693	0.4440
	60	37790	5.8E+07	1300747	368181	0.9969	0.9516	0.4569
Min Vol	1	36750	4.3E+07	1243603	280216	0.9973	0.9693	0.4520
	2	36796	4.3E+07	1265317	282225	0.9973	0.9693	0.4489
	3	36799	4.3E+07	1270726	282830	0.9973	0.9689	0.4480
	4	36796	4.3E+07	1280261	283493	0.9973	0.9689	0.4476
	5	36792	4.4E+07	1284068	284442	0.9973	0.9689	0.4471
	6	36788	4.4E+07	1284073	285613	0.9969	0.9689	0.4489
	12	36773	4.5E+07	1320408	289669	0.9969	0.9680	0.4484
	60	36761	4.6E+07	1382348	297147	0.9964	0.9671	0.4484

Table 5: Benchmark model results at time T including 0.15% transaction costs.

		Wealth				**CAR$_T$**		
Model	**Hor.**	**Mean**	**Var**	**Min**	**Max**	**>0**	**>1**	**1.75-2.50**
SAA	na	38500	4.4E+07	959313	352289	0.9991	0.9844	0.3969
SR	1	39767	5.6E+07	1190218	318890	0.9991	0.9813	0.3453
	2	39637	5.5E+07	1064282	310288	0.9991	0.9827	0.3502
	3	39600	5.5E+07	1042269	302845	0.9991	0.9818	0.3507
	6	39445	5.5E+07	1105514	282929	0.9996	0.9800	0.3556
	12	39094	5.5E+07	1091346	281562	0.9996	0.9742	0.3569
	60	37949	5.9E+07	1313431	370893	0.9969	0.9600	0.3871
Min Vol	1	36972	4.3E+07	1259576	283563	0.9973	0.9693	0.4467
	2	36966	4.3E+07	1275341	285238	0.9973	0.9693	0.4440
	3	36955	4.4E+07	1282339	285132	0.9973	0.9693	0.4458
	6	36930	4.4E+07	1293336	287960	0.9973	0.9689	0.4418
	12	36902	4.5E+07	1330201	291619	0.9969	0.9684	0.4462
	60	36885	4.6E+07	1393949	298839	0.9969	0.9680	0.4431

Table 6: Benchmark model results at time T excluding transaction costs.

Appendix F - Hyperparameter tuning

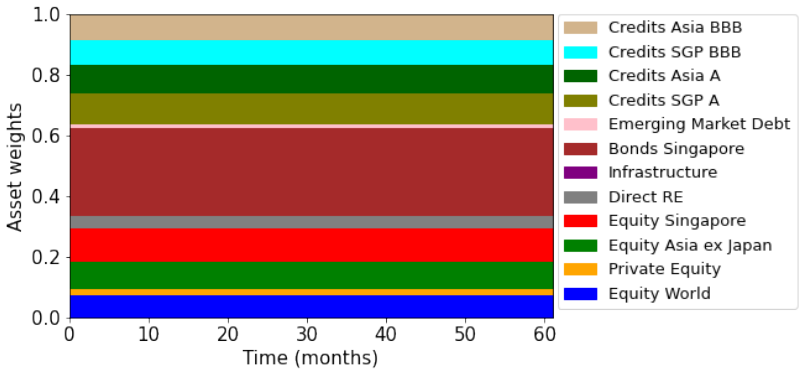
Table 7 displays the hyperparameter candidates and the optimal value.

Hyperparameter	**Description**	**Candidates**	**Optimal**
K	Number of loops over all scenarios.	$(10, 15, 20, 25, 30)$	30
N	Number of scenarios used per update.	$(25, 50, 75, 100)$	25
ϵ	Clipping parameter in PPO objective.	$(0.05 - 0.2)$	0.185
γ	Discount factor in value function.	$(0.99, 1)$	0.99
λ	Discount factor in GAE function.	(0.9)	0.9
ν_{actor}	Learning rate of the actor.	$(0.01, 0.005, 0.001, 0.0005, 0.0001)$	0.0001
ν_{critic}	Learning rate of the critic.	$(0.01, 0.005, 0.001, 0.0005, 0.0001)$	0.005
$\log(\sigma)$	Initial log standard deviation of actions.	$(-2, -0.7)$	-2
Activation	Activation function for PPO update.	(Relu)	Relu
Actor epoch	Number of epochs for actor update.	$(1, 2, 3, 4, 5)$	3
Actor size	Number of hidden layers for the actor NN.	$(2, 3)$	2
Actor width	Number of neurons per layer in the actor NN.	$(64, 128, 256, 512)$	128
Critic epoch	Number of epochs for actor update.	$(1, 2, 3, 4, 5)$	3
Critic size	Number of hidden layers for the critic NN.	$(2, 3)$	2
Critic width	Number of neurons per layer in the critic NN.	$(128, 256, 512)$	128
Mini-batch	Mini-batch size for Adam optimization.	$(32, 64, 128, 256)$	256

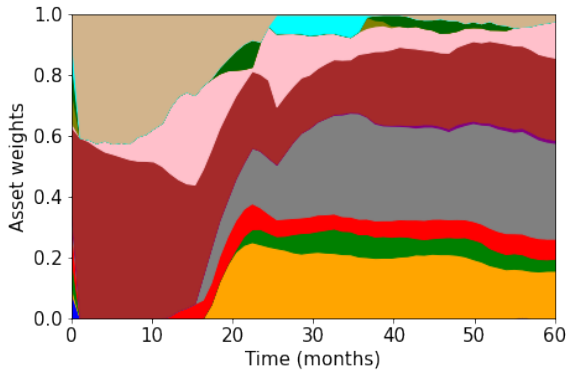
Table 7: Hyperparameter candidates and optimal values of the PPO algorithm.

Appendix G - Additional asset allocation strategies

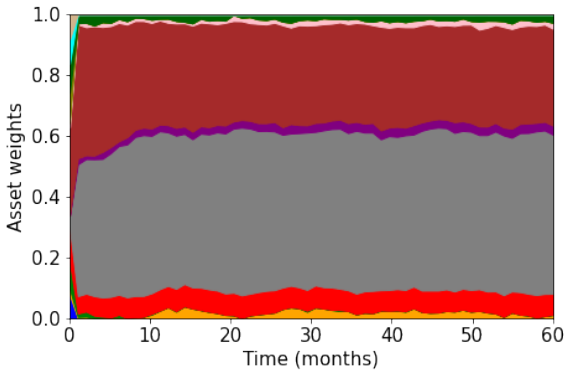
Figure 31 and 32 display the asset allocation strategies of the RL model in the May 2022 OFS with and without transaction costs respectively. Note that the RL strategies display the average allocation across all scenarios. Moreover, notice that the RL strategies in the case with and without transaction costs differ from one another on a scenario level, although it might seem from the figures that they do not differ significantly.



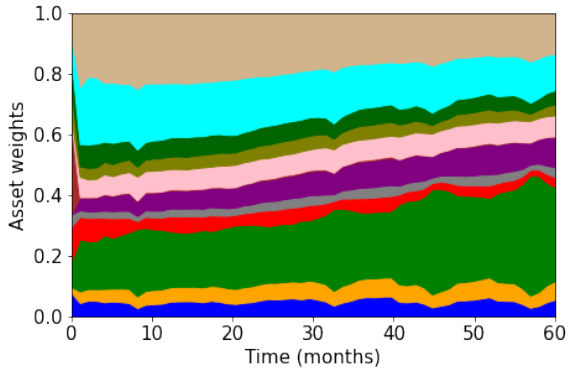
(a) SAA



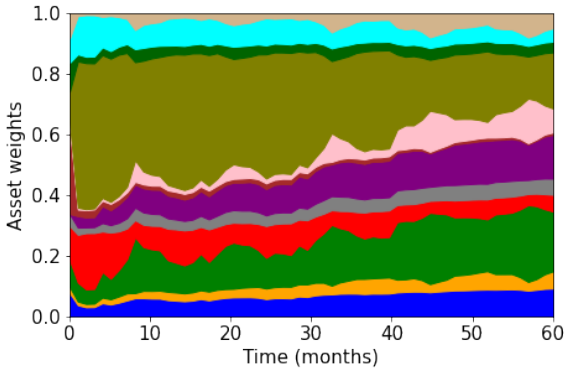
(b) DMV-SR



(c) DMV-MV

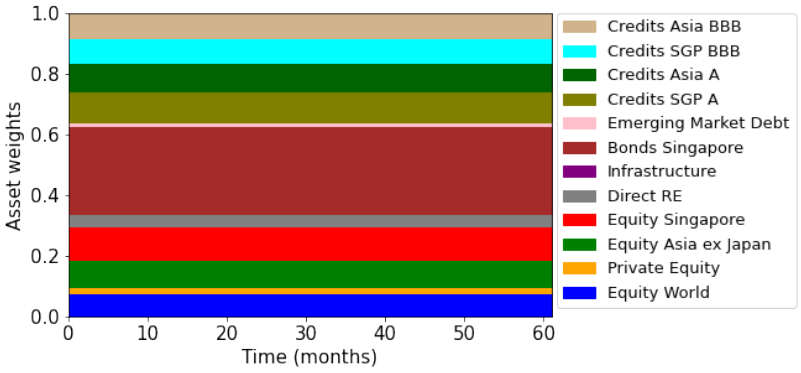


(d) RL-0.001

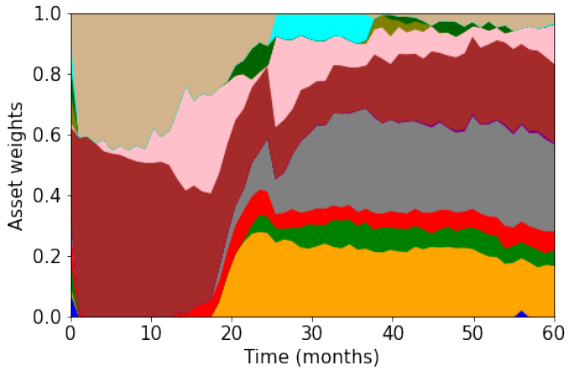


(e) RL-0.2

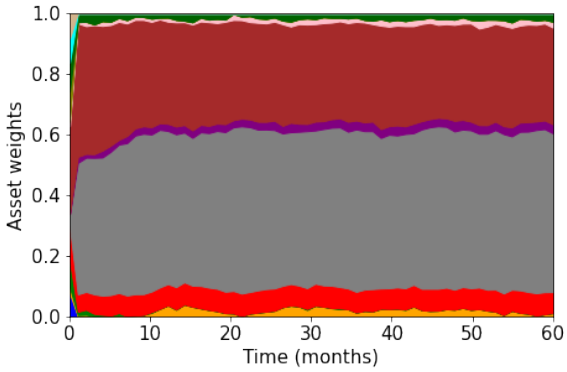
Figure 31: Asset allocation strategies in the May 2022 OFS including transaction costs.



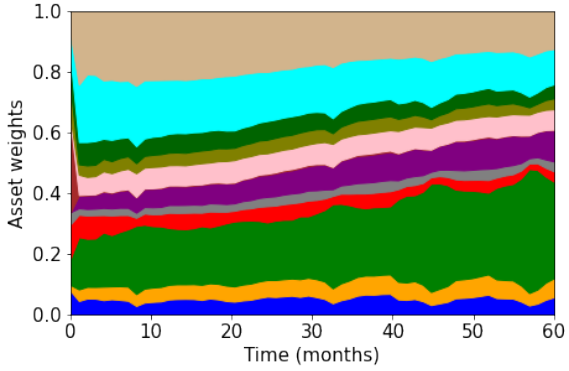
(a) SAA



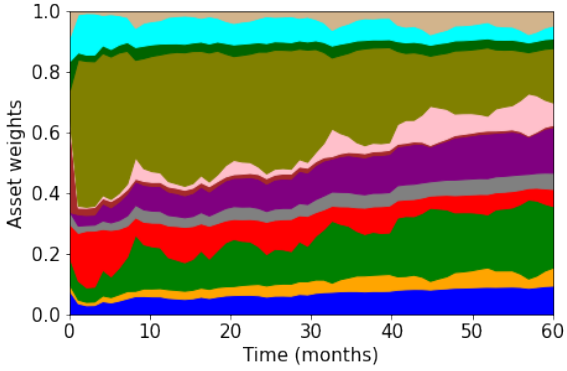
(b) DMV-SR



(c) DMV-MV



(d) RL-0.001



(e) RL-0.2

Figure 32: Asset allocation strategies in the May 2022 OFS excluding transaction costs.