university of groningen

faculty of science and engineering

Finding optimal containment policies to balance GDP and mortality in a SEIARDS-V model using reinforcement learning.

Francesca Perin

university of groningen / faculty of science and engineering

University of Groningen

Finding optimal containment policies to balance GDP and mortality in a SEIARDS-V model using reinforcement learning.

Master's Thesis

To fulfill the requirements for the degree of
Master of Science in Artificial Intelligence
at University of Groningen under the supervision of
Prof. Dr. Davide Grossi (Computer Science/Artificial Intelligence, University of Groningen)
and
Dr. Harmen A. de Weerd (Computer Science/Artificial Intelligence, University of Groningen)

Francesca Perin (s2865300)

December 13, 2022

Contents

Acknowledgments

First of all, I would like to express my deepest gratitude to my supervisor Prof. Dr. Davide Grossi for the support and insight that he as provided me during the entire process of this thesis, and for his excellent teachings during my bachelor and master. Many thanks should also go to the second supervisor Dr. Harmen de Weerd for his feedback and guidance during the ending stage of this project. Furthermore, I would like to mention the University of Groningen for being my home during my studies and for all the teachings provided throughout the years.
To my fiance' Davide, thank you for always being there, in the successful and stressful moments. Early days and late nights. For always being the first person to believe in me, to motivate me, help me focus and reason, and being always so excited to share opinions on different topics. Thank you for always being by my side and being so proud of me, to you I owe everything.
A big thank you goes to my father, my father's family, and my mother, for their unconditional emotional support. They all showed me how proud of me they are, at every step of the way, but in particular during this thesis. My father, in particular, incited me and allowed me to take many opportunities during the years, which lead me, in the end, to pursue this MSc degree. My mother always took time to listen to me, validate my feelings, and comforted me when missing my family and my home. Without them, I would not be the person I am today.
Finally, I would like to thank my best friend and my fiance's family, for always keeping me in their thoughts, for being interested in my studies and progress, and for being a part of my emotional support.

Abstract

The recent pandemic of COVID-19 renewed interest in epidemic models, with the scope of prediction and finding strategies to reduce mortality. Many models focus on single countries, since containment policies are implemented at national level. However, country demographic and other factors may affect the model greatly.
In this thesis, we define a SEAIRDS-V epidemiological model and combine it with a Gross Domestic Product (GDP) economic model. Both models are dependent on country demographics data, consisting on contact rates between different age groups, that simulate 26 countries in the European Union and United Kingdom. Furthermore, we include population migration between countries to simulate travelling. Migration is modelled int two different ways. One of them being a fixed percentage of the population for all countries, and one based on aviation data from the Eurostat database.
Defined this environment, we use Reinforcement Learning (Temporal Difference and Actor-Critic model) to determine if an optimal policy to contain both mortality and loss of GDP is detected across countries. A policy consists of a series of values representing the level of containment (from complete freedom to full lockdown).
From our baseline experiments, with no reinforcement learning and no containment policy, it was found that there is no significant difference in mortality between the two models of migration. Using country demographics resulted in lowered mortality and delayed peak of infection. This results stayed consistent when reinforcement learning is applied, while showing furthered reduction in mortality and delay (consistent with the stricter containment policies), especially if country demographics is included. Looking at the containment values, an optimal policy pattern is not detected across experiments. However, when countries demographics are included results show that 24 out of 26 countries apply stricter containment policies at the peak of infection, while other countries take advantage to maintain lower containment policy with no repercussion on mortality.

1 Introduction

During the past two years many epidemiological models have been used to simulate and understand the spread of COVID-19 and the effectiveness of containment policies in different countries. However, most applications of these approaches have been used to model single countries, due to containment measures and policies being implemented nationally. This because each country has different demographics (such as average age and population density) which influence how the virus affects the specific country.
Economic models have also been made to investigate the impact of COVID-19 and they suffer from a similar problems to the epidemiological ones. Using in-depth economic models requires country specific data to be fitted for each country. On the other hand, more general economic models suffer from the opposite issue being able to model only a generic country.
The aim of this project is to combine a epidemiological model that uses contact matrices to simulate the demographics of different countries within a continental section, to an adapted general economic model. This approach allows the general economic model to gain information from the contact matrices (demographics) of the countries without the need of additional data. On top of this environment, reinforcement learning will be used to see if an optimal strategy can be found and to determine containment policies, instead of investigating predetermined strategies.

1.1 Research Questions

During the pandemic, countries have mostly acted individually when setting containment policies. Focusing the model at continental section level, we can investigate what the optimal policy would be for different countries, considering the demographics and other factors.

Q1. What is the best policy according to our reinforcement learning approach? Can any pattern be seen in such a policy? Can it be compared with a baseline?

Q2. Is the containment policy found by the reinforcement learning approach successful in containing the simulated epidemic?

Q3. What are the effects, on the policy itself, of country data and different approaches of modelling interactions between nations?

1.2 Thesis Outline

The first chapter of this thesis contains an introduction to the problem at hand by providing general background information on why research on this topic is needed and the research questions that we aim to answer. Chapter 2 focuses on relevant literature that inspired this study and provides an introduction to topics that will be relevant for our methodology. In Chapter 3 we define the core methods used. The data used, parameters used in the methods, topology and partial testing are reported in Chapter 4. Chapter 5 presents the experiments and the results obtained, followed by Chapter 6 which discusses the conclusions drawn from the study, it's limitations and possible future research.

2 Background Literature

This project combines three different topics: epidemic modelling, economic modelling, and the use of reinforcement learning to try and find the optimal way to solve the lockdown problem. The COVID-19 global pandemic resulted in renewed interested in research on epidemic simulations and models, in order to understand, predict and learn containment policies to reduce the impact of the virus [1, 2]. Economic models also have been used to forecast economic activity, propose economic policies and in finance for trading [3, 4]. In some cases, reinforcement learning has been used as a mean to solve both epidemic [5, 6] and economic models [7]. Some research has taken place in combining these three elements to analyze the efficacy of containment policies on mortality and it's economic cost. However, this studies are often at region (or country) level, or simulate short periods of time. In this chapter we are going to discuss relevant research in epidemic modelling and economic modelling in a epidemic setting. We then provide a theoretic background of reinforcement learning, and we provide a summary of a study that combined multi-agent reinforcement learning in a compartmental epidemic model, from which our study was based on [8].

2.1 Epidemic modelling

Mathematical epidemic modelling simulates the mechanism used by the disease to spread within the population, with the aim of predicting possible outbreaks and to find the best strategy in order to contain such spread. This is achieved by looking at microscopic level how the disease spreads from a sick individual to a healthy individual. This is then applied on a macroscopic level. There are 2 main models currently used: stochastic methods and deterministic/compartmental models. Stochastic (or random) models, in general, estimate the probability distributions of different potential outcomes, provided some initial conditions. In epidemic modelling, for instance, an example of randomized initial conditions are the exposure risk, transmission, and recovery. In compartmental models population is assigned to groups and a set of differential equations (which can be combined with stochastic inputs) is used to define how population move between these groups. We are going to explore more in depth compartmental models, as these are the methods that we will focus in this study.
There are two basic models: SIS model and SIR model, each of which can be with, or without, vital dynamics [9, 10]. In the SIS model (Susceptible-Infected-Susceptible) individuals start as susceptible (to the disease), progress to being infected and after they are recovered from the infection they return to the susceptible group, due to immunity not being granted. In the SIR model (Susceptible-Infected-Removed) the principle is the same, however, immunity is permanently granted after individuals are recovered, thus they do not return to the susceptible group but move to the removed compartment. The inclusion of vital dynamics, in this models, adds natural births and natural deaths in the model, which is especially essential in SIR model to replenish the susceptible population over time.
On top of these models, numerous others have been built by adding more compartments and more parameters in the differential equations. One these models is the SEAIRD model also used by Aspri, Beretta, Gandolfi and Wasmer [11]. In this model, the population after being exposed to the virus becomes either asymptomatic (A) or infected (I), and after can either recover (R) form the virus, or die (D). A further extension of this model, a SEAMHQRD-V was described by Oraby et al. [12]. In this model the infected compartment is split in sub-groups (mild infected, hospitalized, quarantined) and an environment factor V. The interesting aspect of their research is their embedding of contact matrices from the study of Prem, Cook and Jit. [13] in the epidemic model to simulate different countries. These matrices summarize the contact patterns, of a country, based on age (children, adults and seniors) and location (work, home, school, other and environment). The use of these contact

matrices (and therefore the country social network) in a disease transmission model is an interesting addition, as it allows for a more realistic model and a better understanding of the effect of different containment policies.

2.2 Economic modelling in epidemic setting

Many papers investigate economy models during a pandemic or a combination of epidemic and economic models. This is important as strict containment policies can be put in place to contain and limit the spread of the virus and thus, reduce the number of deaths. However, such policies (i.e. full country lock-downs) can result in huge impacts on the country's economy as working force and production are reduced.
One study from Lars Jonung and Werner Roeger [3] as part of the European Commission, investigated the potential consequences of a pandemic on the EU economy. The simulation involves using a QUEST model (owned by the Directorate General for Economic and Financial Affairs). This is constructed on precise micro-economic foundations, on this basis the model can then be calibrated to be used as a global macroeconomic model. In the study, a pandemic is then introduced in the simulation with a fixed morbidity rate and mortality rate for 3 weeks. The study looks at the supply and demand effect on gross domestic product (GDP) and the time required for GDP to recover.
Another example of combining epidemic and economic modelling, is the study from Carletti et al. [4]. The paper focuses on estimating the drop in profits and equity shortfall in a sample of firms in Italy, as result of the COVID-19 pandemic. This is achieved by taking 2018 balance sheet data summarizing operating revenues, profits, taxes, and labor costs of the firms. On this data the effects of COVID-19 are applied, by simulating both affected and unaffected firms for a duration between 1 and 6 months. The predictions show that a hypothetical 3 months lockdown would potentially result in a total annual profit drop of € 170 billion.
The QUEST model [3] while allowing to model different countries (or entire EU) is a complex model that requires specific calibration. On the other hand, the study of Carletti et al. [4] is a simple model but it is specific for Italy (and requires country specific data). Furthermore, in both cases the simulation of the pandemic is static or derived from historical data, rather then dynamic, with containment measure changed during the simulation.
The study by Aspri, Beretta, Gandolfi and Wasmer [11], previously mentioned, combines a simple compartmental epidemic model (SEAIRD) with a basic GDP economic model. This paper defines how the two models can be combined to affect each other. An increase in infected population results in a loss of working force and production, thus, reducing the GDP of the country. This allows to study the effects of different containment policies while considering both mortality and GDP. The model described in their research result is a very versatile model aimed at modelling a generalized country. This allows to add country specific data to the model to simulate singular countries while using the same generalized model.

2.3 Reinforcement Learning

Reinforcement learning (RL) is one of the three branches of machine learning alongside supervised and unsupervised learning. The general idea in RL is to have an agent (or multiple ones) placed in an environment, which the agent is able to interact with by taking actions. Depending on which action the agent chose ,it gets a reward (a real number either positive or negative) [14]. Thus, with this reward the agent learns the optimal action to take (through trial and error) based on it's current state in the environment.

Environments in RL are often defined as Markov decision process (MDP), which in general summarizes the problem as having:

- a set of agents
- a set $\mathcal{S}$ of states, either belonging to the agent/and or environment
- a set $\mathcal{A}$ of actions, that describe how the agents navigates trough the world
- a transition function $T(s' \mid s,a)$, which delineates the probability of ending in state $s' \in \mathcal{S}$ after being in state $s \in \mathcal{S}$ and performing action $a \in \mathcal{A}$
- a reward $R(s,a)$ obtained by performing action $a \in \mathcal{A}$ at state $s \in \mathcal{S}$

The MDP describes the environment or, in other words, the problem at hand, while reinforcement learning aims at finding the solution to the problem.
At each time step, the agent is in a certain state in the state space and needs to preform an action in the environment. Depending on the current state and the action performed, the agent will end up in a next state and perceive a reward with different probabilities. This is a single step interaction and is called a *transition*, formally a tuple $\langle s,a,r,s' \rangle$. This process is then repeated in the successive state either to infinity or until and ending state (if existing) is reached. A sequence of experience transitions of an agent is called an *episode*.
The solution, or policy, $\pi(a \mid s)$ is a probability distribution which given a state s returns the probability of action a to perform in said state. The objective, in an MDP, is to find the policy that maximise the total sum of all rewards the agent experiences. We can better formalize this objective by defining the return G_t at time step t as:

$$G_t = r_t + \gamma \cdot r_{t+1} + \gamma^2 \cdot r_{t+2} + \gamma^3 \cdot r_{t+3} + ... \tag{1}$$

where r_t is the reward at time-step t and $\gamma \in [0,1]$ is the discount factor. The latter is applied to the sum of rewards for two reasons. The first, is to have a finite upper limit on the sum of the rewards even if the length of the episode is infinite. The second, is to give more weight to future rewards closer to the current time step. Notice that G_t is a random variable as it depends on the transitions the agent will experience which are stochastic. However, in an MDP, the value of the current time step should not be of importance. All the information necessary in order to take the optimal action is the current state s_t. We thus define the value of an arbitrary state s as the expected return from s onwards. Formally, this is encapsulated by the value function

$$V(s) = \mathbb{E}[G_t \mid s_t = s] \tag{2}$$

This value represents how beneficial (moving forward) being in the current state is for the agent. The value function also allows to compare states, where $V(s_i) > V(s_j)$ indicates that being in s_i will be more beneficial in the long run for the agent than being in s_j.
It is evident that the values of different states may be related to each other. For instance, in the aforementioned example s_j may be one of the possible states reachable from s_i. For consecutive

states this relation can be made more explicit by rewriting Equation 2 with the following steps:

$$V(s) = \mathbb{E}[G_t \mid s_t = s] \tag{3}$$
$$= \mathbb{E}\left[r_t + \gamma \cdot r_{t+1} + \gamma^2 \cdot r_{t+2} + \gamma^3 \cdot r_{t+3} + ... \mid s_t = s\right] \tag{4}$$
$$= \mathbb{E}[r_t \mid s_t = s] + \gamma \cdot \mathbb{E}_{s',a|s}\left[\mathbb{E}\left[r_{t+1} + \gamma \cdot r_{t+2} + \gamma^2 \cdot r_{t+3} + ... \mid s_{t+1} = s'\right]\right] \tag{5}$$
$$= \mathbb{E}[r_t \mid s_t = s] + \gamma \cdot \mathbb{E}_{s',a|s}\left[\mathbb{E}\left[G_{t+1}) \mid s_{t+1} = s'\right]\right] \tag{6}$$
$$V(s) = \mathbb{E}[r_t \mid s_t = s] + \gamma \cdot \mathbb{E}_{s',a|s}\left[V(s')\right] \tag{7}$$

The resulting equation is also known as Bellman equation. The advantage of defining the value function as Bellman equation, over the definition in Equation 2, is that the former function only requires to sample the state at next time step s_{t+1} and the resulting reward r_t instead of an entire episode.
To learn the value function we can thus either exploit Equation 2 or 7. From the former (Equation 2) we can approximate $V(s)$ by sampling (partial) episodes starting at s and averaging their returns. This method(s) is called Monte-Carlo [14]. On the other hand, we can enforce Equation 7 as a constraint on function approximator estimating $V(s)$, this methods are called Temporal difference (TD) [15, 14].
Once the value function is learned, a series of methods can be used to extract a viable policy from it. It can be noted that knowing the state value function $V(s)$ does not provide enough information to determine a strategy (what action to take in s). This because V only provides information whether a state s_j is more valuable than another state s_i (if $V(s_j) > V(s_i)$). However, it cannot be known from V alone what is the action to take that will transition to the most valuable state. To overcome this issue, the state-action function Q is introduced. This function has a very similar definition to V in that it is also the expected value of the return, however, the Q-value is the expected return for a given starting state and a given first action performed. Formally:

$$Q(s,a) = \mathbb{E}[G_t \mid s_t = s, a_t = a] \tag{8}$$
$$= \mathbb{E}[r_t \mid s_t = s] + \gamma \cdot \mathbb{E}_{s',a'|s,a}\left[Q(s',a')\right] \tag{9}$$

Note that, we can approximate the expectation with respect to r_t and s' by sampling, as this data is given by a transition (i.e. a sample). However, we now have introduced the problem of having to sample a' (i.e. the action taken in the next time step). There are several methods, that provide a solution to this problem, the most commonly used being Q-learning. This method, proposes to sample the action that the greedy policy would take in s', that being the one that maximises the expected value of the next state. Formally, $a' = \underset{a_j}{\operatorname{argmax}}\, Q(s', a_j)$.
Neither Monte-Carlo nor bootstrapping methods are strictly better than then other. On the one hand, the advantage of Monte-Carlo methods is that they have zero-bias, as they take true return samples. On the other hand, TD methods provide lower variance with respect to Monte-Carlo as their samples are more homogeneous. This usually result in TD algorithms converging faster while having a bias, as the returns are approximated and not sampled [14].
Other solutions to MDPs try to learn the policy directly without the need of learning the value function, these are called Policy gradient methods [16]. The advantage of using these methods is that the policy can be approximated using any function, including those that provide a probability distribution over continuous actions, as opposed to Value-function based method which almost exclusively require the action space to be finite and discrete. Assuming that we have a reward that is maximised by a policy π dependent on parameter θ, the aim is to find the optimal θ to maximise said reward. To do so we use gradient ascend (or descend). This is achieved by updating the parameters in small

steps according to their gradient. However, calculating the gradient ($\nabla \mathbb{E}_{\pi_\theta}[G_t]$) of the expected return requires knowledge of the full MDP (including transitions and reward functions). This because, when the return is expressed in this form, in order to know the changes resulting from increasing or decreasing θ, one would need to run a full episode and record the delta with the resulting return. To solve this problem, the policy gradient theorem states the following:

$$\nabla \mathbb{E}_{\pi_\theta}[G_t] = \mathbb{E}_{\pi_\theta}\left[\left(\sum_{t=1}^{T} G_t \nabla \log \pi_\theta(a_t \mid s_t)\right)\right] \tag{10}$$

From Equation 10, we can see that although the right hand still requires the return to be computed, it does not require to differentiate it. In fact, when accounting for the expectation on the right hand the simplest way to deal with the return is to sample it exactly how it would be done in aforementioned Monte-Carlo methods. As mentioned before, however, Monte-Carlo methods have a high variance and hence may require a large amount of data to converge.
One may trade some bias, in exchange for reduced variance by approximating the return with TD methods. These methods are called Actor-Critic methods. In these methods, the Actor is responsible for choosing an action to take according to a certain policy. The Critic needs to learn to asses the policy used by the actor, which is done by evaluating the value function and calculating the TD error. This error is then passed to the actor so that the policy can be changed as necessary in order to minimize the value.
In the policy gradient methods there are several different ways to compute the return. The most straight forward way, is to calculate the total reward of the trajectory (as in Monte-Carlo) or by evaluating the total reward after taking action a, with or without using a baseline. Other methods consist in using a state-action value function, an advantage function, or by using TD residual. We focus on the last method as this will be used in our actor-critic algorithm. In the case of TD residual, the return is calculated using the following equation:

$$G_t = r_t + V^\pi(s_{t+1}) - V^\pi(s_t) \tag{11}$$

where V^π represents the value function when using π as policy to act in the environment.
These are called actor critic methods [17]. In these methods, the actor dictates the probability distribution over actions (dependent on the current state). Therefore, the actor is simply another term for the policy π. On the other hand, the critic predicts how valuable was an action performed (by the actor) in a certain state. Equation 11 represents the critic, as it calculates the difference between the predicted value of s_t and what the actual of s_t is predicted to be once action a_t was performed. The former is simply $V^\pi(s_t)$, while the latter is approximated by $r_t + V^\pi(s_{t+1})$ (using the Bellman equation). Note that, in the last formula, the effect of a_t is taken in account within r_t.
The component V^π of the critic, is an approximation of the real value function of π which needs to be learned (along with π itself). In TD residual, the square of the approximation of G_t (see Equation 11) is the loss of the value function approximation. Notice that, in TD residual, both the actor and the critic's losses depend only on a single sample transition. Thus, the data used for training are batches of individual transitions and not episodes.
Besides Equation 11 there are several other methods to approximate and learn the critic [18], however we will focus only on TD residual as is the only method used in these research.

On-/Off- Policy Another important distinction is between off- and on- policy methods. Value based and policy gradient methods can be either off- or on- policy methods. This distinction is important since one of the peculiarities of reinforcement learning is that the distribution over the data changes

over time (as it depends on the policy which is updated while learning). Thus, one may want to use a different policy for behaving in the environment (behaviour policy) from the policy that is learned (target policy). For instance, one may desire the target policy to be the optimal greedy policy while having a behaviour policy that favours exploration. Methods that allow to learn a different behaviour policy from the target policy are called off-policy methods. On the other hand, one may prefer that during learning the type of behaviour policy used is taken into account. These are labeled as on-policy methods as they learn the optimal policy with respect to the type of behaviour policy being used.
An example of an off-policy is the previously introduced Q-learning algorithm. It is categorized as such because it computes the state-action value obtained by following the greedy strategy independently from what the behavioural policy is.

2.4 RL and Epidemic models

One of the main papers that has influenced this project explores the use of a Q-learning algorithm (see Section 2.3) in a multi-agent epidemic model [8]. The epidemic model used is a SEIRS compartmental epidemic model. This model consists of a SEIR model (Susceptible-Exposed-Infected-Recovered), with the variation of having part of the population losing immunity to the virus with time (this being the second S component in the name of the model). The model and population N values describe the state s of the agent. In the simulation, multiple generalized countries (or agents) are used, and agents interact togheter via migration of population from one country to another. A network topology (triangle, star, infinite) is used to define the migration between countries. The scope of the study is for each agent to use reinforcement learning to find the optimal lockdown policies through time in order to optimize their own objectives (minimize mortality) interacting trough the epidemic model and migration. In this study, the containment policies are a discrete set of values varying from complete lockdown to complete freedom. Furthermore, in the paper the ending goal for the simulation is defined as stable state, in other words, a state that does not change for 50 iterations. Due to having a discrete action space and a set ending goal for the simulation, the reinforcement learning approach used was Q-learning. This method was also chosen by the author as a simple algorithm would be a better fit in a pilot study to see if the combination of the two components could lead to promising results. Langhorst study [8] successfully showed that reinforcement learning can be applied to multi-country (or multi-agent) epidemiological model. Furthermore, increasing the complexity of the topology between agents (more agents interacting together) prove to stabilize the simulation leading to the idea that if countries would co-operate in their efforts it would mitigate the epidemic faster or maintain it under control for longer periods of time.

3 Methods

In this section we are going to first discuss in detail the different methods used in the epidemic and economic environment and how they are combined. Secondly we are going to discuss in more detail how reinforcement learning was used to solve the problem of finding the best state imposed closure strategy to minimize mortality and maximise gross domestic product.

3.1 Epidemic and economic environment

Our environment, as a general idea, simulates an epidemic disease which spreads in a number of countries that interact togheter. The spread of the epidemic disease is simulated at country level by using a SEAIRDS-V epidemiological model. Furthermore, it is possible to use (in the simulation) the patterns of social contacts between age groups of a country as mean to simulate more accurately the spread of the epidemic. Countries interact together by having people move from one another, this can either be a fixed percentage of the departing country population or by using the country aviation data. Finally, as part of the environment, we also simulate the country Gross Domestic Product (GDP), this is done to monitor also the economic status of the country through the epidemic and how it is affected by containment policies.

3.1.1 SEAIRDS-V model

In this paper we combine a SEAIRD model [11] with the V element of the SEAMHQRD-V model [12]. This is done to maintain a simple model which only considers Infected (I) individuals and does not differentiate between Mildly infected (M), Hospitalized (H) and Quarantined (Q) since this differentiation is not the main focus of this model. However, from the SEAMHQRD-V we take the use of the V component, as it does model interaction between different age-groups more in detail, something that is taken into account in this research, leading to a SEAIRD-V model.
Furthermore, in the final SEAIRDS-V model, part of the population that is recovered from the virus is then added back in the group of Susceptible due to loss of immunity (see Figure 1). This is not necessary, and depends on the virus characteristics, but it does allow the epidemic disease model to run for longer periods of time.

In the model the total population of a country is divided in separate groups: S (susceptible), E (exposed), A (asymptomatic) I (infected), D (dead), V(environment). One important constrain of this model is that S+E+A+I+R+D=1, where each value is a normalized between 0 and 1, thus depicting the fraction of the population in each group. The environment group (V) is left out from the constrain as it does not represent a population group. Instead, it represents an environment that has been contaminated and could spread the virus (e.g., the handle of a shopping cart or the railing on the stairs of a public space).
Figure 1 provides a schematic representation of the model. At the beginning most (if not all) of the population starts in the S group and then shifts to a different group with a certain rate following the arrows in the diagram. It is important to notice that our SEAIRDS-V model is implemented at country level.
In SEAIRDS-V models it also possible to model the use of containment policies, which are aimed at reducing the spread of the virus. This policies are modeled by a value α of a country which can regulate the spread of the virus between the people who did not have contact with the virus (Susceptible) and those who did (Exposed, Asymptomatic and Infected individuals). For α will use continuous

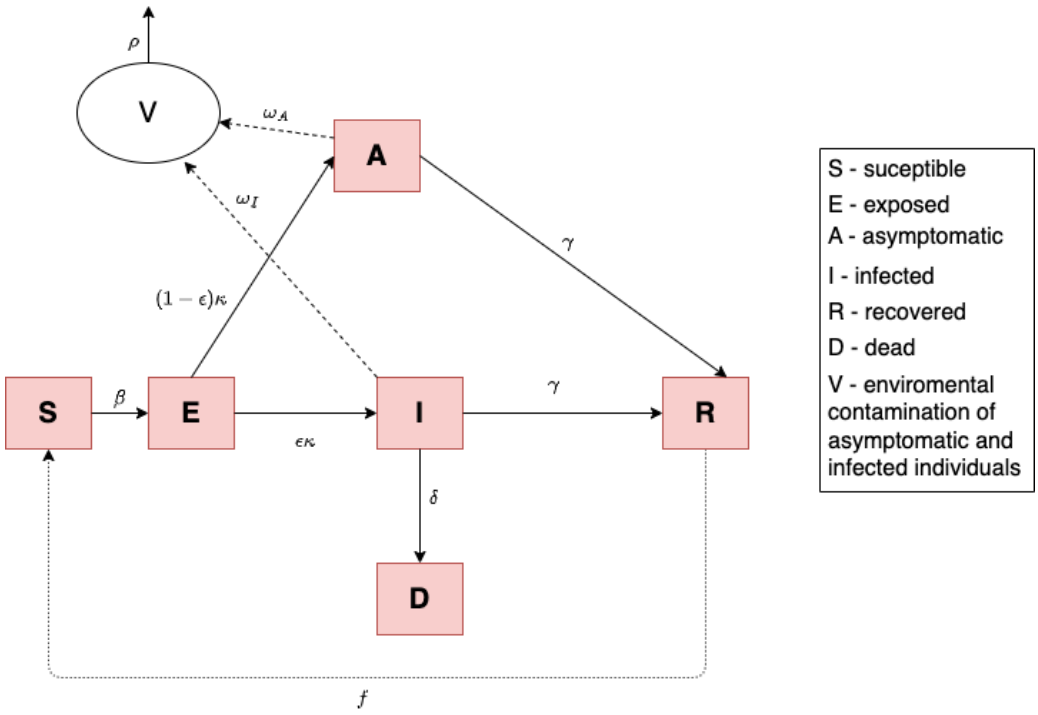

Figure 1: Graphical representation of SEAIRDS-V model and transmission rates, where arrows indicate the progression of population trough different groups.

values ranging from 0 indicating complete freedom and 1 indicating complete lock downs. State imposed containment measures can be followed or disregarded. The ratio of people following measures is indicated as p_v (with regard to environment) and p(with regard to person to person interaction). We aggregate the two aforementioned values as $\tilde{p}$= $\alpha \cdot p$ and $\tilde{p}_v$= $\alpha \cdot p_v$.

$$\frac{dS}{dt} = n \cdot (1-D) + f \cdot R - \beta \cdot S \cdot ((1-\tilde{p}_v) \cdot V + (1-\tilde{p}) \cdot \frac{A+I}{N}) - n \cdot S \tag{12}$$

In this equation, $n \cdot (1-D)$ is responsible for replenishing the population from the natural deaths occurring in each group (except for D as those are by definition deaths caused by the epidemic and thus not replenished). The second factor moves a fraction f of the Recovered population to the Susceptible group, this is due to loss of immunity after a certain period of time. The third factor $-\beta \cdot S \cdot ((1-\tilde{p}_v) \cdot V + (1-\tilde{p}) \cdot \frac{A+I}{N})$ accounts for population shifting from Susceptible to Exposed. Susceptible individuals have a certain possibility of becoming exposed due to environment contamination, this corresponds to $\beta \cdot S \cdot ((1-\tilde{p}_v) \cdot V)$. Similarly one can become exposed by having contact with an Asymptomatic or Infected person which is modeled by $\beta \cdot S \cdot ((1-\tilde{p}) \cdot \frac{A+I}{N})$. The final term, $-n \cdot S$ corresponds to the natural deaths (not due to the virus) occurring in the Susceptible group population. In the case of the Exposed group, the population that was detracted from the Susceptible group due to environment contamination and contacts with Asymptomatic and Infected, thus moving from Susceptible to Exposed, corresponding to factor, $\beta \cdot S \cdot ((1-\tilde{p}_v) \cdot V + (1-\tilde{p}) \cdot \frac{A+I}{N})$, is added to the E group. Furthermore the factor $-\kappa \cdot E$ removes Exposed population (moving from E to another group) with rate κ, which simulates the incubation period. Finally, the factor $-n \cdot E$ removes the natural deaths occurring in this population group. These principles are formalized in equation 13

$$\frac{dE}{dt} = \beta \cdot S \cdot ((1-\tilde{p}_v) \cdot V + (1-\tilde{p}) \cdot \frac{A+I}{N}) - \kappa \cdot E - n \cdot E \tag{13}$$

Of the Exposed individuals removed at rate κ from the previous equation, a percentage $1-\varepsilon$ will not develop symptoms, and thus are added to the Asymptomatic group. Furthermore, at rate γ the already Asymptomatic population will recover form the virus, therfore being removed from A. Finally, we remove the factor $-n \cdot A$ corresponding to natural deaths in this group. This is formalized in equation 14

$$\frac{dA}{dt} = (1-\varepsilon) \cdot \kappa \cdot E - \gamma \cdot A - n \cdot A \tag{14}$$

For the Infected group, the equation is very similar to the Asymptomatic one. Population form the exposed group is removed at rate κ and ε percentage of the population is moved to the Infected group. Of the previously Infected population a portion is removed with rate γ who are individual that recover, and a second portion is removed with rate δ, which are deaths due to the virus, and finally n individuals are removed for natural death causes. See equation 15

$$\frac{dI}{dt} = \varepsilon \cdot \kappa \cdot E - \gamma \cdot I - \delta \cdot I - n \cdot I \tag{15}$$

In the Recovered group are added at rate κ both the population of the Asymptomatic and Infected group, respectively the first and second factor of the equation. The natural deaths in the Recovered group are removed at rate n, and finally at rate f part of the Recovered population is inserted back in the Susceptible group due to loss of immunity. See equation 16 for a formalization of these rules.

$$\frac{dR}{dt} = \gamma \cdot A + \gamma \cdot I - n \cdot R - f \cdot R \tag{16}$$

The individuals removed from the Infected group with rate δ are added in the Dead group, since this factor counts the deaths due to the virus. This group strictly counts only epidemic related deaths and not natural deaths, also it is important to note that virus related deaths in the population are not replenished only the natural ones. This change is modelled by equation 17.

$$\frac{dD}{dt} = \delta \cdot I \tag{17}$$

The Environment group emulates the public environment and does not consist of population like the other groups. In this equation, the first and second term correspond to the environment contamination of respectively Asymptomatic and Infected individuals, where ω_A and ω_I are constants but $\tilde{p}$ is a value summarizing the state imposed containment policies and how followed said policies are by the public. The last term, $-\rho \cdot V$ is the natural decontamination of the environment, due to the virus. These rules are encoded in equation 18.

$$\frac{dV}{dt} = \omega_A \cdot (1-\tilde{p}_v) \cdot A + \omega_I \cdot (1-\tilde{p}_v) \cdot I - \rho \cdot V \tag{18}$$

Table 1 provides a summary of all the parameters of the aforementioned equations and SEAIRDS-V model.

3.1.2 Migration

The migration from an arbitrary country A to a country B consists of a number of individuals, a fraction of the population of A, moving from A to B. Each individual belongs to a certain group in the SEAIRDS-V model. With respect to country A, the migration is obtained simply by subtracting the population that is emigrating from the population of the country. Because the emigrating population

Parameter	Description
β	rate of transmission
κ	rate of exposed to either infected/asymptomatic
ε	fraction of k individual that from exposed move to infected
ω_I, ω_A	environment contamination
γ	infected/asymptomatic recovery rate
δ	infected die rate
f	lost of immunity ratio
ρ	removal of environment contamination
α	state imposed closures
p, p_v	ratio of people following state imposed closures

Table 1: Summary of the parameters used in the equations of the SEAIRDS-V model

is a sample of the entire population the values of the SEAIRDS-V model do not change. Country B, which is receiving the immigrant of country A, adds them in it's own population. However, country B may have differing values in the SEAIRDS-V model, therefore the immigrant population needs to be added in the model according to the group in which they belong and then the values need to be recomputed to re-normalize the SEAIRDS-V model values for country B.
In the environment the migration value τ (the fraction of the population of A moving to B) can be calculated in two ways:

- *Fixed migration:* in this case the emigration/immigration population from country A to B is a percentage (pop_perc) of country A. Having a fixed migration allows us to have a baseline. It enables a controlled migration needed for spread of the virus in the epidemiological model, while keeping a simple model and only adding one variable fixed for all countries.
- *Aviation data migration:* to have a more accurate representation on the average traffic between countries it is possible to use aviation data that summarizes the number of individuals transported in a year from country A to country B. Aviation data alone does not account all individuals coming in and out of a country but it is one transportation method that is traceable, in particular if there is a lack of border control.

In both cases, we also add noise to the emigration/immigration value. To do so we assume that the noise is normally distributed, with mean 0. However, the amount of data is not enough to have a correct estimation of the true variance. Thus, the variance is set to a small arbitrary value, which is a percentage ($percSTD$) of the estimated mean (or the fixed migration value) τ. Formally:

$$\tau_n \sim N(\cdot \mid \tau, \text{percSTD} \cdot \tau) \tag{19}$$

where $percSTD \in [0,1]$. Furthermore, the state imposed containment policies also have an influence on the number of people able to move from countries, therefore in both the cases the migrating population is multiplied by the inverse of the state imposed containment policy value (1-$\alpha_B(t)$).

3.1.3 Gross Domestic Product

Strict state imposed containment policies such as a country lockdown can be useful solutions to prevent the spread of the epidemic. However, such solutions do also result in negative side effects, for

example on the economy of a country. Here we consider Gross Domestic Product (GDP) as measure and summary of the economy of a country, resulting in a basic economic model. This will allow to see the effect of containment policies both on the epidemiological model and the economic impact (to a certain degree).

To combine the GDP calculation with the SEAIRDS-V model we follow the paper from Aspri, Beretta, Gandolfi and Wasmer [11], while rewriting the formulas to account for our model.
Given an agent with previous state $State_t = (S_{t-1}, E_{t-1}, A_{t-1}, I_{t-1}, R_{t-1}, D_{t-1}, V_{t-1}, GDPloss_{t-1})$ and after migration has occured, we calculate for the state of the agent at timestep t. Firs the SEAIRDS-V values (S_t,E_t,I_t,A_t,R_t,D_t,V_t) are computed by using the model equations. After the $GDPloss_t$ is calculated from the former values as follows:

$$G(\alpha, \theta) = (1-\alpha)^{\theta} \tag{20}$$

$$P(State_t) = G(\alpha, \theta) \cdot (S_t + E_t + A_t + R_t) \tag{21}$$

$$V(P, \sigma) = -\frac{P^{1-\sigma} - 1}{1-\sigma} \tag{22}$$

$$\frac{dGDPloss}{dt} = -e^{-rt} \cdot (V(P(State_t)) + a \cdot D_{t-1}) \tag{23}$$

Equation 20 captures the link between the containment policy and the GDP value, this formula originally used α instead of $1-\alpha$. This change was necessary as in this paper 0 indicates complete freedom, and in the original study [11] it indicates complete closure. Equation 21 represents the labor availability, so Infected and Dead groups are not included in the working population. Furthermore this simulation does not include testing, therefore Asymptomatic population is also included in the working force. Once the labour population is calculate Equation 22 is used to calculate the loss of production according to the current time step. Finally Equation 23 evaluates the current GDP loss value according to loss of production, virus related deaths and the GDP cost of reducing mortality (certain percentage of virus related deaths are avoided by containment policy at a certain cost).

3.1.4 Age groups and Social contact matrices

Part of the project consists in the use of the social contact matrices from the research by Prem, Cook and Jit. [13], which summarize the patterns of social contacts between age groups in a specific country. For each country, the patterns are summarized for the interaction of 3 age groups: child (c), adult (a), and senior (s). Each group may interact with another in 5 different locations: household (h), school (sc), work (w), other (o) and environment (v). Given a location $k \in \{h, sc, w, o, v\}$ we have the social contact matrix $C^k \in [0,1]^{3\times 3}$, which represent the relative frequency of contacts between the different age groups in that particular location. Formally:

$$C^k = \begin{pmatrix} C^k_{cc} & C^k_{ca} & C^k_{cs} \\ C^k_{ac} & C^k_{aa} & C^k_{as} \\ C^k_{sc} & C^k_{as} & C^k_{ss} \end{pmatrix}$$

where C^k_{ij} is the relative frequency of contact of age group i with group j in location k.
From these matrices we calculate two social contact matrices, $\tilde{C}$ and $\tilde{C}^v$, which depend from the location-specific contact matrices C^k defined above, but also on the relative number of people following containment procedure, indicated by $\tilde{p}(t)$ which is defined as follows:

$$\tilde{p}^k(t) = p^k \cdot \alpha(t) \tag{24}$$

for $k = sc, w, o, v$, where $\alpha(t) \in (0, 1]$, as previously discussed, represents nation imposed restrictions at time t. The factor $p^k =< p_c^k, p_a^k, p_s^k >$ represents the percentage of population willing to follow said state imposed containment policies for each age group. Given these two components, we can now define the equations for $\tilde{C}$ and $\tilde{C}^v$ for each age-group pair $i, j \in \{c, a, s\}$ as follows:

$$\tilde{C}_{ij} = C_{ij}^h + \sum_{k \in \{sc,w,o\}} C_{ij}^k \cdot (1 - \tilde{p}_j^k(t)) \cdot (1 - \tilde{p}_i^k(t)) \tag{25}$$

and

$$\tilde{C}_{ij}^v = C_{ij}^v \cdot (1 - \tilde{p}_i^v(t)) \tag{26}$$

In the definition $\tilde{C}_{ij}$, we scale the social contact matrix value C_{ij}^k with respect to the factor $(1 - \tilde{p}_j^k(t)) \cdot (1 - \tilde{p}_i^k(t))$ for $k \in \{sc, w, o\}$. This is done to have a social contact matrix value that is directly proportional to how high the contact between the two age groups i and j is at location k (C_{ij}^k), While being also inversely proportional to the containment policy and how followed said closures are by either group i or group j. Notice that scaling is not applied to the home location. This is due to the fact that no containment policies can be applied in this case, and therefore neither how much those policies are followed. $\tilde{C}_{ij}$ takes in consideration all locations with the exception of the environment which is treated separately in $\tilde{C}_{ij}^v$. In this case, C_{ij}^v is only scaled with respect to i by $1 - \tilde{p}_i^k(t)$ this is due to the fact that we don't have two groups interacting as in the previous case but only one group i interacting with the environment.

Rewriting equation for age groups

We defined the summary contact matrices $\tilde{C}_{ij}$ and $\tilde{C}_{ij}^v$ which account for age-group, different locations, and current state imposed closure $\alpha(t)$. These matrices can now be implemented and used in our environment. Since our environment consists in a SEAIRDS-V model, migration, and GDP model, contact matrices are integrated in the formulas previously defined for each environment component.

We have the following, for $i = c, a, s$:

- SEAIRDS-V:

$$\frac{dS_i}{dt} = n \cdot (1 - D_i) + f \cdot R_i - \beta_i \cdot S_i \cdot \Big(\sum_{j \in \{c,a,s\}} \tilde{C}^v_{ij} \cdot V_j + \sum_{j \in \{c,a,s\}} \tilde{C}_{ij} \cdot \frac{A_j + I_j}{N_j} \Big) - n \cdot S_i, \tag{27}$$

$$\frac{dE_i}{dt} = \beta_i \cdot S_i \cdot \Big(\sum_{j \in \{c,a,s\}} \tilde{C}^v_{ij} \cdot V_j + \sum_{j \in \{c,a,s\}} \tilde{C}_{ij} \cdot \frac{A_j + I_j}{N_j} \Big) - \kappa \cdot E_i - n \cdot E_i, \tag{28}$$

$$\frac{dA_i}{dt} = (1 - \varepsilon) \cdot \kappa \cdot E_i - \gamma \cdot A_i - n \cdot A_i \tag{29}$$

$$\frac{dI_i}{dt} = \varepsilon \cdot \kappa \cdot E_i - \gamma \cdot I_i - \delta \cdot I_i - n \cdot I_i, \tag{30}$$

$$\frac{dR_i}{dt} = \gamma \cdot A_i + \gamma \cdot I_i - n \cdot R_i - f \cdot R_i, \tag{31}$$

$$\frac{dD_i}{dt} = \delta \cdot I_i, \tag{32}$$

$$\frac{dV_i}{dt} = \omega_A \cdot (1 - \tilde{p}_{v_i}) \cdot A_i + \omega_I \cdot (1 - \tilde{p}_{v_i}) \cdot I_i - \rho \cdot V_i \tag{33}$$

- Migration:

$$M_i = (1 - \alpha_B) \cdot N_i * \tau + \tau_n \tag{34}$$

  - Fixed: in this case, τ is a fixed percentage of the population
  - Aviation:

$$\tau = \frac{aviation}{\sum_{j \in \{c,a,s\}} N_j \cdot 365} \tag{35}$$

- GDP loss:

$$G(\alpha, \theta) = (1 - \alpha)^\theta, \tag{36}$$

$$P(agent) = G(\alpha, \theta) \cdot (S_{a_t} + E_{a_t} + A_{a_t} + R_{a_t}), \tag{37}$$

$$V(P, \sigma) = -\frac{P^{1-\sigma} - 1}{1 - \sigma} \tag{38}$$

$$\frac{dGDPloss}{dt} = -e^{-rt} \cdot (V(P(State_t)) + a \cdot D_{a_{t-1}}) \tag{39}$$

  it is important to notice that in the case of distinction between age groups, the *GDPloss* value is calculate on the basis of the adult group (a) only. This is because child and senior population do not belong to the labour working force.

3.1.5 Interaction between environment and agent

After having explained the SEAIRDS-V model, migration and gross domestic product as single elements, is important to understand how environment and agents interact.
Firstly, each agent is initiated with a certain SEAIRDS-V state, population, GDP loss and a set α value. Because α is set to a certain value and does not change, we can calculate $\tilde{p}$ and $\tilde{p}_v$, and furthermore determine the contact matrices if age groups are used. These values will not change in the simulation since they are dependent on α.
At each time step t, corresponding to one day, each agent will perform the migration to all connected agents. After all migrations are performed for time step t, each agent will calculate the new population, SEAIRDS-V values and GDP loss, which will provide information to set the new agent state. After this the simulation time step t is concluded and the simulation will progress to time step $t + 1$.

3.2 Reinforcement Learning

In this section we define the Markov decision process, previously introduced (see Section 2.3),in terms of the task at hand. Furthermore, we discuss in more detail the reinforcement learning method chosen to solve the task, this method being an actor-critic critic model and temporal difference.

3.2.1 Markov decision process

We translate the task at hand following the requirements of an MDP as follows:

- **Agents**: n agents, where each agent represents a country
- **State**: is defined as $State_t = (S_t, E_t, A_t, I_t, R_t, D_t, V_t, GDPloss_t)$, no goal state will be present. The state-space is continuous with the S,E,A,I,R,D,V values at time t being between 0 and 1.
- **Actions**: continuous actions in the set $[0,1]$ which represent the level of containment policies. Were 0 indicates complete freedom, while 1 represents full lockdown measure.
- **Reward**: Because we aim to minimize mortality and the value of the GDP loss (i.e. to maximize the delta GDP) we change the sign of the latter, thus having $GDPloss_t = -GDPloss_t$. After this the reward is calculated, as follows :

$$R(t) = \begin{cases} GDPloss_t \cdot D_t & \text{if } GDPloss_t < 0 \\ GDPloss_t \cdot (1 - D_t) & \text{if } GDPloss_t \geq 0 \end{cases} \tag{40}$$

  Which formula is used depends on if the GDP loss is positive or negative. From the SEAIRDS-V constrain D must be positive or zero, on the other hand $GDPloss$ can be either positive or negative

- **Transition function**: given the current state and an action the next state is computed by using the SEAIRDS-V model equations and the GDP loss equations. In the model as discussed we also have migration between countries which is also taken into account during the transition. The country receiving the immigrant population is going to use the state of the country of origin and the number of people to update their own state (see Section 3.2.3).
- **Policy (solution)**: the optimal policy is going to be determined by the actor-critic algorithm which will discuss in Section 3.2.2.

3.2.2 Temporal Difference and Actor-Critic model

The MDP modelling our simulation has continuous stochastic actions. For this reason, using a policy method is the default since these methods are effective in continuous action spaces and are able to learn stochastic policies [19]. As mentioned in Section 2.3, one disadvantage of policy methods, however, is that they have high variance (slow learning and thus computationally inefficient). This because it's generally hard to find a score function to correctly assess the policy. Given the reduced window of time in which the SEAIRDS-V model is reliable, high variance during training may result in the agent not converging properly within the aforementioned window. To overcome the disadvantage of high variance, we use Temporal Difference learning with Actor-Critic. Another advantage of using this method is that the agent can be trained on individual transition rather then entire episodes. This allows the agent to learn during a simulation rather then having to wait for the simulation to end

before learning takes place. Moreover, online learning is also useful as the simulation does not have a specific goal (end state), instead the agent is thus able to learn even if the simulation is stopped after an arbitrary number of time steps.
Finally, this training method is on-policy which allows us to apply arbitrary operations on top of the policy and take them into account during training. We will expand on this in Section 3.2.3.

3.2.3 Interaction between environment and agent with reinforcement learning

In the case where reinforcement learning is used, the main change in the interaction between environment and agent is that α is not fixed. This entails a few extra steps in the procedure described in Section 3.1.5.
Initialization remains the same, with each agent having a starting SEAIRDS-V state, population, GDP loss, and initial α value. At each time step t, we first compute α for each agent. These values are recomputed from each agent's policy every n time steps, otherwise the last α, calculated by an agent's policy, is used for that particular agent. In other words, this is equivalent to freezing the value of α for n time steps, for each agent. Afterwards, given α, the agent recomputes $\tilde{p}$ and $\tilde{p}_v$ (and contact matrices if age groups are used).
Then, all migrations between connected agents are performed. Once all migrations are executed for time step t, each agent calculates the new population, SEAIRDS-V values and GDP loss, to set as the new agent's state. The latter (new state), the value of α used, and the reward computed according to Equation 3.2.1 are pushed (memorized) into each agent's replay buffer. With a period of T_{learn} before ending time step t, each agent's policy is updated using the training procedure outlined in Section 2.3. Finally, the end of time step t is reached and the simulation moves at time step $t+1$.
As previously mentioned one of the advantage of TD residual is the fact that is an on-policy method. Now it may be more evident why it is more important for the proposed method, as being on- policy would mean that when training the policy should take into account that α may be frozen.

4 Experimental Setup

In this section we present the data used in the model described in the methods Section (see Section 3). We introduce the datasets used, how the social contact matrices were derived, and a summary of the parameters used for the environment. Furthermore we present the topology of the model which indicates how the agents are connected and initialized. Finally, we present the neural networks and parameters used in the Reinforcement Learning portion of the methods and preliminary testing done on this part alone.

4.1 Datasets

In this Section we provide the source and post-processing of all the datasets for the population, migration, and household composition, together with some analysis of the aforementioned data.

4.1.1 Population data

Population data for a total of 217 countries was retrieved from the World Bank dataset [20] which provides information of the total population per country in 3 different age groups (0-14, 15-64, 65 and older) together with the total population per country. The population data was taken from year 2016 to 2020 (included) and averaged to obtain more complete data. In our methods (see Section 3) the total population is referred to as N (in Section 3.1.1) or N_i if using the age groups with $i \in \{c, a, s\}$ respectively child (0-14), adult (15-64) and senior (65+)(see Section 3.1.4).

4.1.2 Air passenger transport data

Data of international intra-EU air passenger transport between a combination of countries from the Eurostat database [21] was used as approximation of people travelling between countries. This data set contains information of the annual number of passengers who departed or arrived in a country from a EU partner country. Data was taken from the years 2015 to 2021 (included) and averaged to ensure data completeness.
From this data given country A we have all passengers departed for country B, this value is referred to as $aviation$ and used in Equation 35.
In Figure 2 we provide a heatmap representing in each square the $aviation$ value of people flying from country A (along the y-axis) to country B (x-axis)

4.1.3 Household composition

Data of the household composition per country was obtained from the United Nations, Database on Household Size and Composition [22]. From the data column 'average number of household members of selected ages' we obtain information regarding the number of under 15 among all households, the average number of people of 20-64 years among all households and we calculate the average number of people of 65+ years by taking the average household size and subtracting average of 20-64 and under 15. This data does not take in account the age group between 15 and 20 years old, which is, in this simulation, usually included in the 15 to 64 years old age group. This partial incompleteness of the data, however, does not result in a relevant difference in the overall simulation.

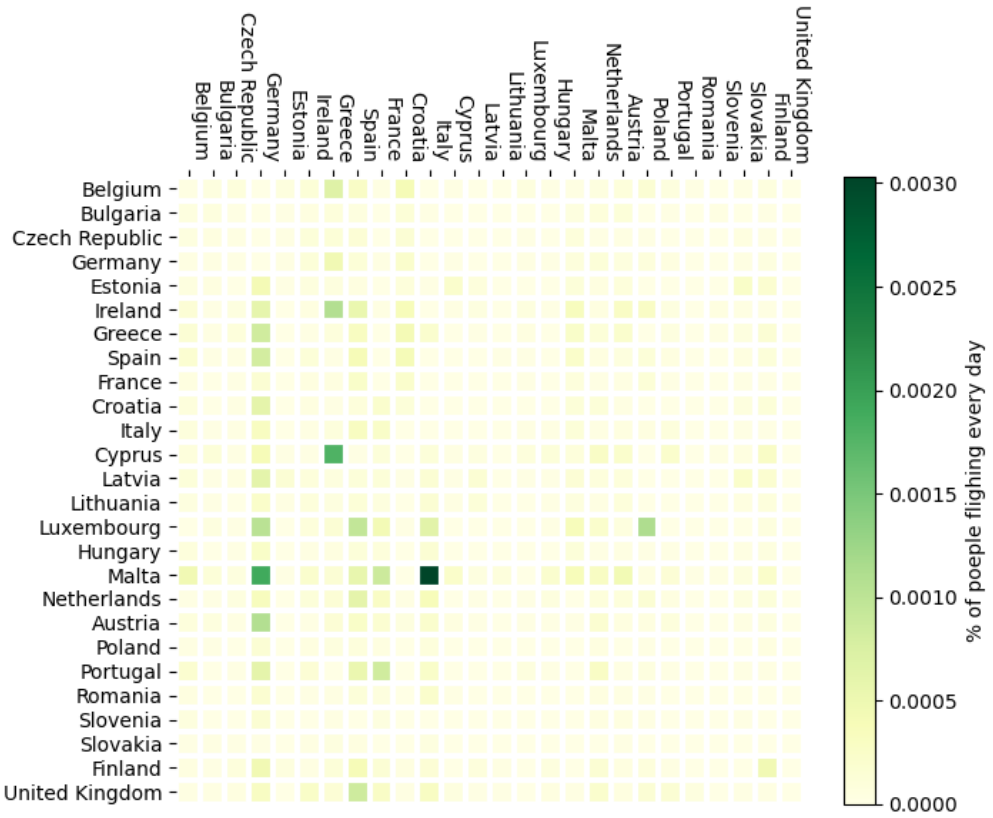

Figure 2: Heatmap of aviation data, each cell represents the population percentage of people flying every day from country A (y-axis) to country B (x-axis)

4.2 Contact matrices

The initial contact matrices were calculated and made public from the research by Prem, Cook and Jit. [13]. In this study data sourced from POLYMOD, Demographic Household Surveys (DHS) and United Nations Population Division was used to estimate contact rates at specific locations (home, work, school and other) and age (0 to 80 with 5 year intervals) for a total of 152 countries. This was achieved by first applying a Bayesian hierarchical model on the POLYMOD contact data (8 countries) and estimating the age and location specific contact rates for those countries. The posterior distribution was estimated using Markov chain Monte Carlo simulation on the contact rates. Due to a difference between the age intervals of the original study, the contact matrices were readjusted accordingly to our age groups (0-14 child (c), 15-64 adult(a)[1], 65+ senior (s)). This was done following the method used by Oraby et al. [12], which consists of summing the values of the age group corresponding columns and then averaging across rows (always according to this study age groups). This was repeated for all locations obtaining C_{ij}^{k} for $k \in \{h, sc, w, o, all\}$ respectively home, school, work, other for all locations, and $i, j \in \{c, a, s\}$). Furthermore the resulting home contact matrix C_{ij}^{h} is normalized by using the household composition per age group data (see Section 4.1.3). Oraby et

[1]Since the original data has 5 years intervals, and an age of majority needs to be set, we chose 15 (over 20) to be said limit. We justify this choice as 15 is closer to the average legal working age in Europe and the Gross domestic product is a significant part of the simulation environment.

al. [12] also include a new location called environment (which will be used for the V portion of the SEAIRDS-V model), which is a fraction r_v of the all location contact matrix, thus, $C_{ij}^{v} = \frac{C_{ij}^{all}}{r_v}$.

4.3 Parameters

Table 2 illustrates the different parameters used in our simulation regarding the SEAIRDS-V model, the population migration between countries, the social contact matrices and finally the economic model. When multiple values are given (for age groups) they are in the order c, a, s respectively.

	Parameter	No age group	Age group
SEAIRDS-V	β	0.35	[0.13, 0.35, 0.35]
	κ	0.2	0.2
	ε	0.75	0.75
	ω_I, ω_A	1	1
	γ	0.14	0.14
	δ	0.0028	0.0028
	f	0.0035	0.0035
	ρ	0.33	0.33
	n	0.00003	0.00003
	p, p_v	0.5	[1, 0.5, 1]
Migration	percSTD	0.002	0.002
Contact matrices	r_v	6	6
Economy	θ	0.33	0.33
	σ	2	2
	a	18000	18000
	r	0.0001	0.0001

Table 2: Parameters used during the experiments for the SEAIRDS-V model, migration, contact matrices, and economy equations. When multiple values are given (for age groups) they are in the order c, a, s respectively.

Most of the parameters remain the same for both the no age group and age group setting. However, two parameters to take important notice are p and p_v. If the population is not divided in age groups, 50% of the entire population will follow the state imposed containment policies, both when interacting with other people ($p = 0.5$) and the environment ($p_v = 0.5$). However if the population is divided in age groups, 100% of the children and senior citizen will follow the state imposed containment policies (both with people and environment), while for the adult population we apply the same 50% for both types of interaction.

4.4 Topology

In the simulation we chose to simulate European Union, and United Kingdom, with a total of 26 agents where each agents correspond to a country. The list of countries is the following: Austria, Belgium, Bulgaria, Croatia, Cyprus, Czech Republic, Estonia, Finland, France, Germany, Greece,

Hungary, Ireland, Italy, Latvia, Lithuania, Luxembourg, Malta, Netherlands, Poland, Portugal, Romania, Slovakia, Slovenia, Spain, United Kingdom of Great Britain. All 26 countries are connected with each other and interact trough the migration of population from one country to another.

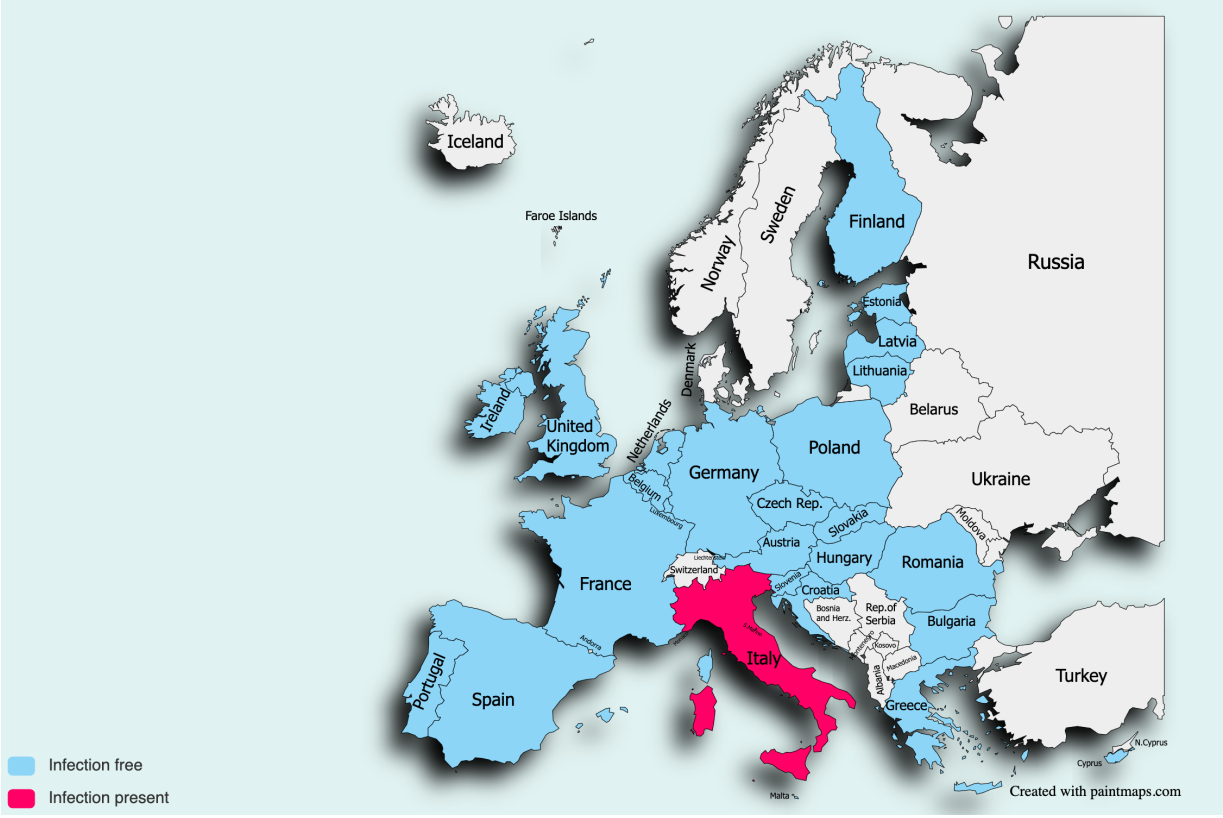

Figure 3: Europe map showing the countries simulated in the experiments. Countries colored in blue have no contact with the virus, country in red has contact with the virus, countries in white are not present in the simulation.

At the beginning of the simulation 25 countries out of the 26 are free from the virus, and are depicted in Figure 3 in blue. Italy, colored in red, is the only country with the virus present at the beginning of the simulation. The white countries are not present in the simulation due to a lack of data.
Initializing whether an agent is free from virus or the virus is present is achieved through the starting SEAIRDS-V model values.
For agents that are free from the virus, all population is susceptible, as none of the population has been in contact with the virus. On the other hand, agents that are in contact with the virus are initialized having a very small percentage of the population being exposed to virus (0.0001% of the population), while the remainder of the population is susceptible (see Table 3).

4.5 Reinforcement learning

Because there are many agents interacting together the neural network topology was kept relatively small. We implemented the actor and the critic on two separate networks both with 3 linear layers, and with learning rate set to 0.001, with a discount factor set to 0.99.
Finally, the action space corresponds to $[0, 1]$, being all possible (continuous) lockdown measures of a country. Therefore, the policy needs to return a probability distribution over $[0, 1]$ for a given state

Group	Virus free	Virus present
S	1.00	0.999999
E	0	0.000001
A	0	0
I	0	0
R	0	0
D	0	0
V	0	0

Table 3: Initial SEAIRDS-V model values for agents depending on whether the virus is present in the country or not.

in this setup. To achieve this, the policy network will output the parameters of a beta distribution over the interval $[0,1]$, as we assume the noise to be normally distributed, but the values to be bounded.

4.5.1 Testing

The actor-critic method was separately tested to check that learning is occurring and that this part of the model was implemented correctly. To do so the actor-critic method was tested in an arbitrary environment. We used the Cart Pole from the Classic Controls environments of OpenAI Gym [23] that corresponds to the cart-pole problem described by Barto, Sutton, and Anderson [24]. It consists of a pole attached to a cart, that moves left and right without friction. The aim is to keep the pole in the upright position, which is achieved by moving the cart left and right. The action space in this environment is discrete with actions being $\{0,1\}$ respectively left and right. Because in our method the action space is continuous (between the values of 0 and 1) the actor critic method needed to be slightly adjusted to account for this difference. This is achieved by using a multinomial distribution instead of a beta distribution. In Figure 4 we show that the running average return increases over time. This indicates that the implemented actor-critic algorithm is able to learn to perform the cart-pole task being able to maintain the pole in equilibrium fro over 200 time steps at the end of the experiment.

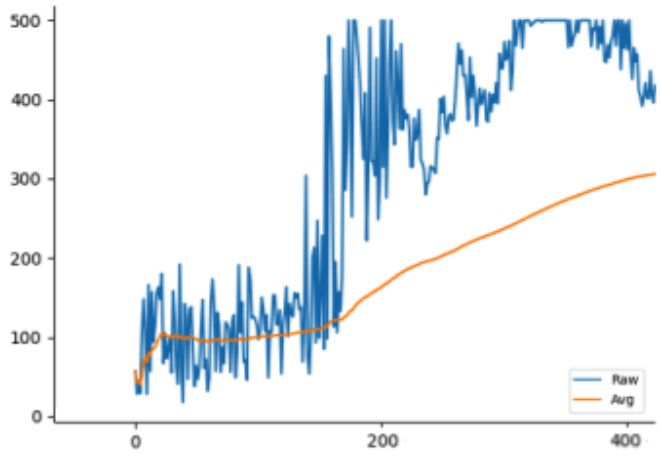

Figure 4: The plot shows in blue the return per episode, and in orange we show the running average of the return

5 Results

Our SEAIRDS-V environment simulates the spread of a virus within a population. For the baseline of our environment (when no reinforcement learning method is applied), the state does not impose any containment measures, thus, the virus is free to spread between all simulated countries, we will refer to this method as NoRL. On the other hand, when reinforcement learning is applied, each state applies a containment policy, varying from total freedom (represented with a value of 0) to complete closure and lockdown (equal to 1). We investigate the containment policy values used by the different countries and their effect compared to the baseline. Furthermore, we also explore the effect of having a fixed amount of people migrating between countries compared to data retrieved from aviation data. Finally, we also compare if simulating age groups does benefit the simulation.

5.1 Experimental procedure

For each different setting in Table 4 we computed 5 runs of the experiment, and averaged the values to have a more accurate result. All plots show the averaged valued and standard error, although standard error is not visible for the NoRL method experiments. This is due to the fact that there is no randomness in our environment.

Method	Iterations	Fixed migration	Age groups	Freeze period	Reference label
No RL	1000	True	False	1	NoRL1
	1000	True	True	1	NoRL2
	1000	False	False	1	NoRL3
	1000	False	True	1	NoRL4
RL	1000	True	False	1	RL1
				15	RL2
	1000	True	True	1	RL3
				15	RL4
	1000	False	False	1	RL5
				15	RL6
	1000	False	True	1	RL7
				15	RL8

Table 4: Table summarizing the experiments executed without (No RL) and with reinforcement learning (RL). The setting of having fixed migration or not and age groups or not are explored in all possible combination for both methods. Furthermore for the RL method we also explore containment measures possibly changing every day or every 15 days.

5.2 Results for no containment policies

First we compare our baseline results in particular to see the effects of the two types of migration and whether having the population split into age groups affects the results.

5.2.1 Migration

To see the effects of the two types of migration in the baseline we look at the experiment where the population is not divided in groups. Thus, we compare experiments NoRL1, which has fixed migration and NoRL3 which has migration based on data.

Scuscptible population: Firstly we are going to compare the Susceptible group.

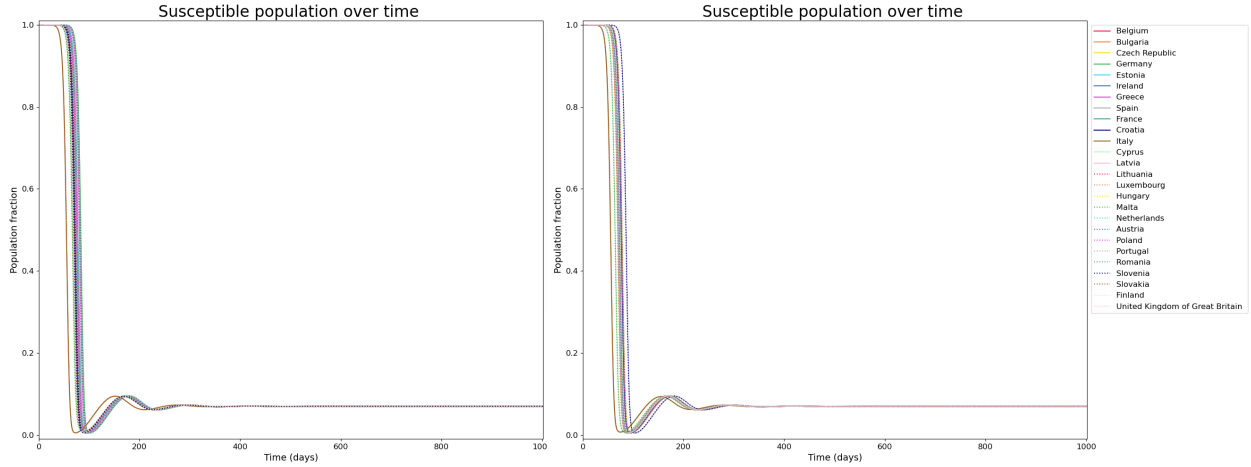

Figure 5: Susceptible population fraction (no age group differentiation) over time for fixed migration (left) and aviation based data (right) for the baseline of no reinforcement learning.

From the results in Figure 5 we can see that the pattern of the susceptible population over time is the same for both types of migration. The population starts at 1 (all population), and reaches the minimum value of around $0.0047\% \pm 6E-06$ around day 94 for all countries and in both migration conditions. Italy, the first country infected results in the highest minimum, around $0.0069\%E-05$. Furthermore, because the virus spread starts from this country, the minimum value is reached at approximately the 72 days mark. In both cases the susceptible population plateaus at around $0.07\% \pm 3E-08$. Although the pattern is the same, if migration is not fixed we can that the is less variance between countries, with the exception of two countries standing out from the other, those being Malta (green dotted line close to Italy) and Slovenia (last infected country).

Exposed population: Looking at the results from the Exposed compartment (see Figure 6) , we can notice again that both experiments show the same patterns for both groups. Aviation based migration (on the right) shows less variance between countries. The number of Exposed individual has the first peak at day 79 with a value of $0.20\% \pm 2E-05$ for all countries (except Italy which is shifted a few days prior). There is also a second peak at day 202 with a value of $0.018\%1-08$. The peak of the exposed population is a result of the drop in the Susceptible group, as people move from one group to the other. The Susceptible population decreases of approximately 0.95% while the Exposed population only increase of approximately 0.20%. This is due to the fact that the population

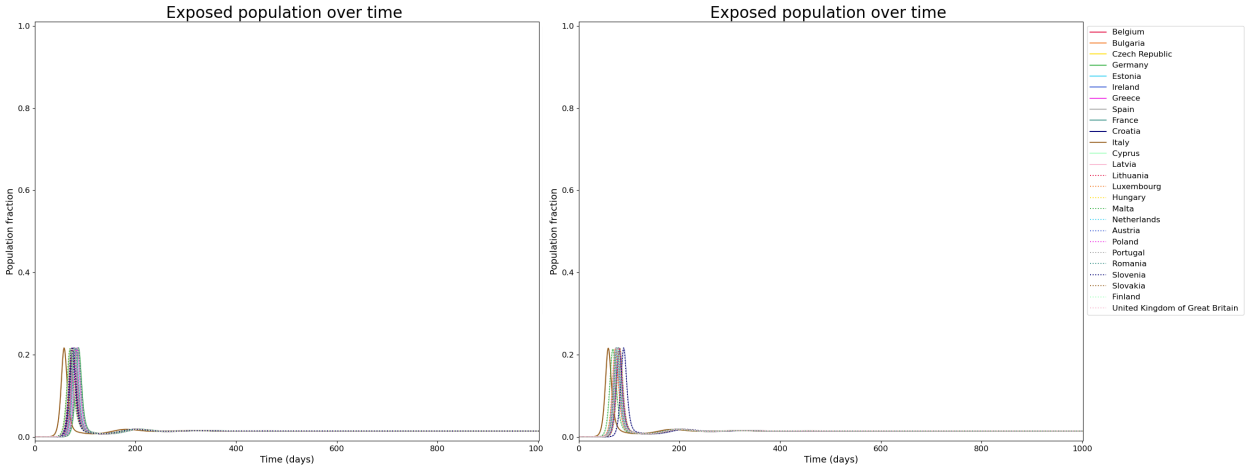

Figure 6: Exposed population fraction (no age group differentiation) over time for fixed migration (left) and aviation based data (right) for the baseline of no reinforcement learning.

is removed from the Susceptible population in a cumulative pattern, while in the Exposed category, the increase is differential.

Infected and Asymptomatic population: From the Exposed category population will progress to either the Infected or Asymptomatic group. At the peak of the virus spread the Exposed category reaches a value of 0.20%. We know form the model parameter that, at a certain rate, 75% of the Exposed population moves to Infected an 25% to Asymptomatic. This is reflected in the results seeing that the peak for Infected reaches a value of $0.14\% \pm 8E-05$ (see Figure 7) and $0.048\% \pm 2E-05$ for Asymptomatic (see Figure 8).

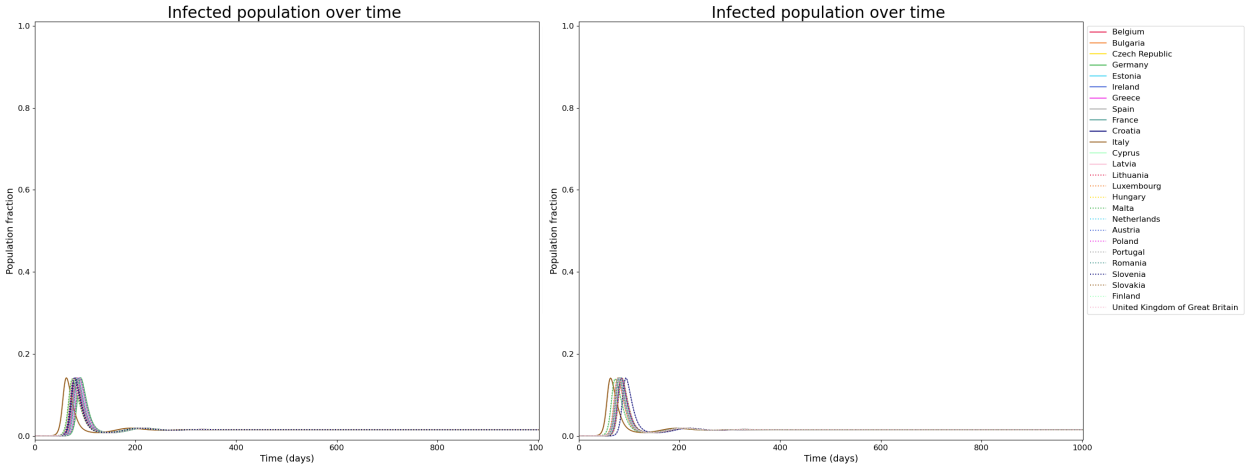

Figure 7: Infected population fraction (no age group) over time for fixed migration (left) and aviation based data (right) for the baseline of no reinforcement learning.

In this case, the ratio is maintained since the population is transitioning between groups with a certain rate, unlike what was previously explained for the Susceptible-Exposed transition. The first peak takes place in all four experiments at around day 84, while the second smallest peak is around day 210. It is important to notice that the Infected and Asymptomatic population from day 400 to 1000 never reaches a 0 value, instead it stabilizes at a constant value of 0.014% and 0.0049% respectively.

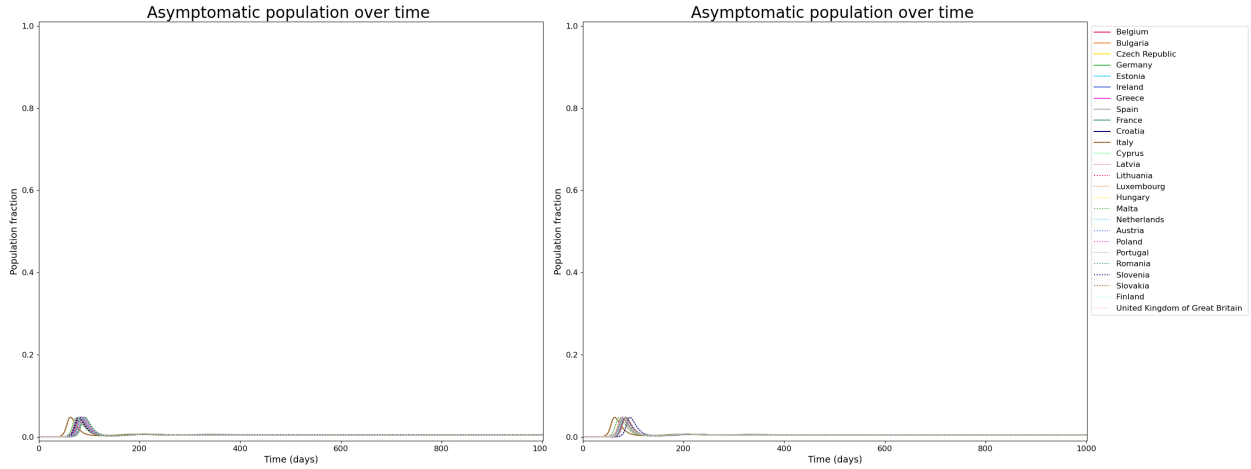

Figure 8: Asymptomatic population fraction (no age group) over time for fixed migration (left) and aviation based data (right) for the baseline of no reinforcement learning.

Recovered and Deceased population: Finally we look at the last two compartments, Recovered and Deceased. These groups can also be considered cumulative as the population remains in that group, similarly to the Susceptible category. This is true for the Deceased compartment, which simply sums the number of deaths due to the virus (and not natural deaths). The Recovered group is not strictly cumulative as a small percentage of the Recovered population is fed back into the Susceptible group after a period of time due to loss of immunity.

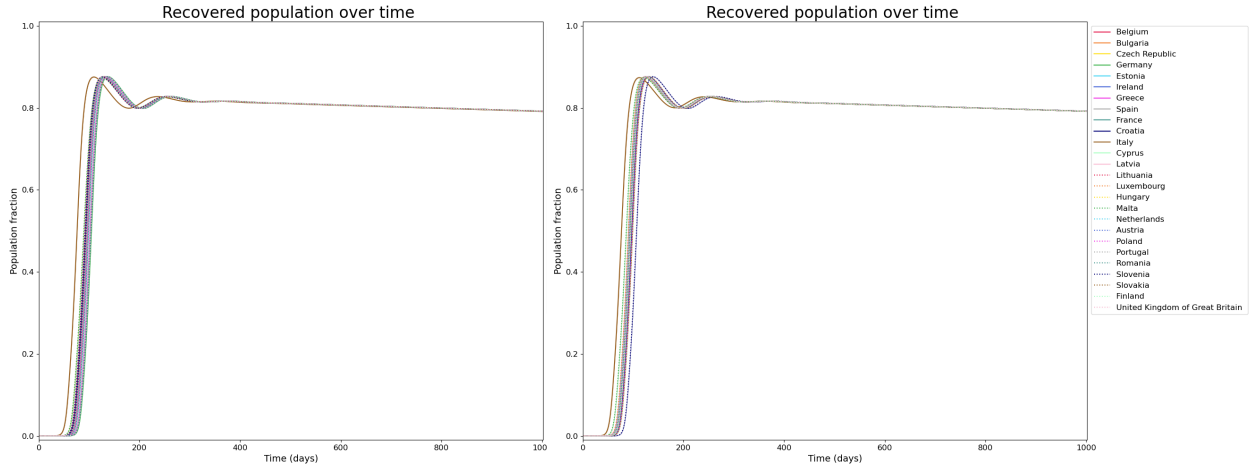

Figure 9: Recovered population fraction (no age group) over time for fixed migration (left) and aviation based data (right) for the baseline of no reinforcement learning.

This is visible in both plots in the Figure 9 as a negative slope starting from day 400 to 1000. On the other hand, the Deceased group after an initial peak resulting from the peak of Infected has a steady increase from 400 to 1000 (see Figure 10), this is resulting from the Infected and Asymptomatic groups never reaching a value of 0 and thus the virus continuing to affect the population.

5.2.2 Age groups

Although aviation data does not affect results greatly, to isolate the effect of age groups in the population we are going to analyze the experiments with fixed migration only. Thus, we are going to

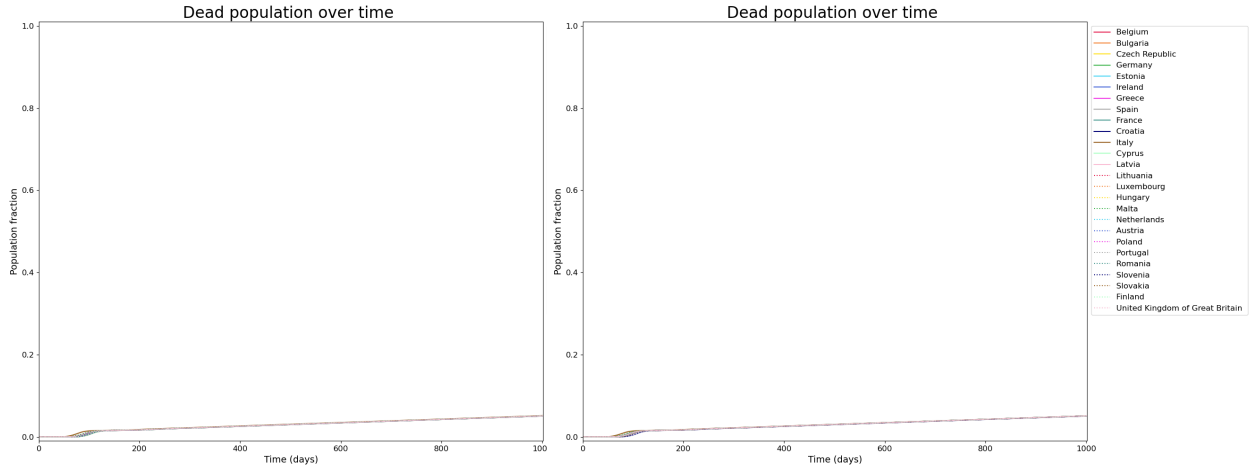

Figure 10: Dead population fraction (no age group) over time for fixed migration (left) and aviation based data (right) for the baseline of no reinforcement learning.

compare experiments labeled as NoRL1 and NoRL2. Please be aware that plots with age groups have varying scales along the y axis for the different age groups, however scale for the total group (bottom plot) is maintained constant (from 0 to 1). This was done to allow the patterns to be more visible.

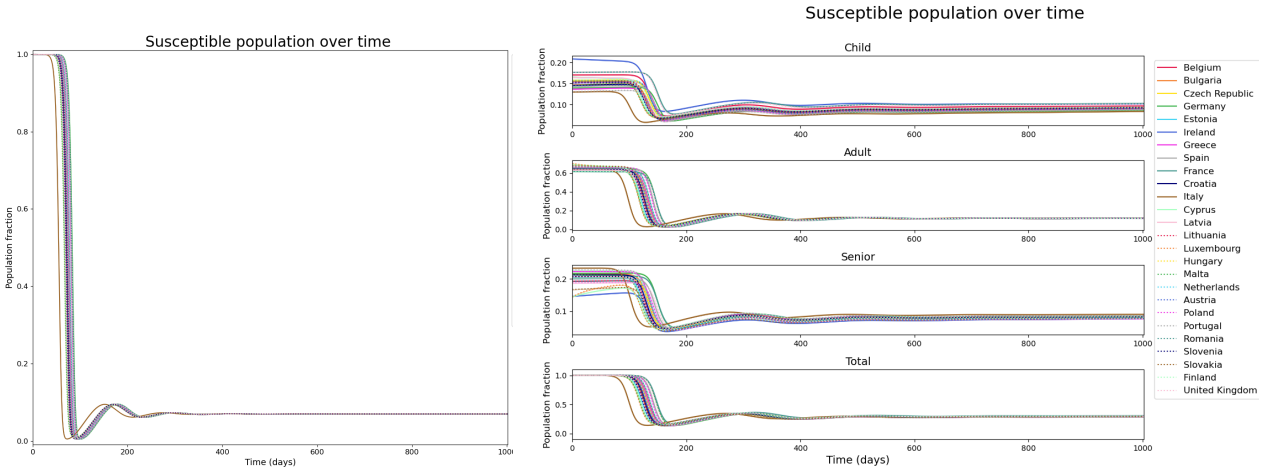

Figure 11: Susceptible population fraction(s) over time for fixed migration with no age groups (left) and with age groups (right)

Susceptible population: Comparing the Susceptible group we can see from Figure 11 right that the value decreases less drastically and values only start to decrease around day 100 (with the exception of Italy) with the minimum value of 0.14% reached at day 164 compared to a value of 0.047% around day 94 in the left. We can also see the second peak being at day 302 with a value of 0.34%. In both cases the function plateaus at the latest iterations. In the case where age groups are used the last value is $0.29\% \pm 1E-05$ (for the total group) which is significantly higher compared to the no age group setting reaching a value of $0.07\% \pm 1E-08$.

Exposed population: For the Exposed group with no age group differentiation the peak value for the total population is $0.21\% \pm 1E-04$ (see Figure 12 left), while, for the other case the maximum

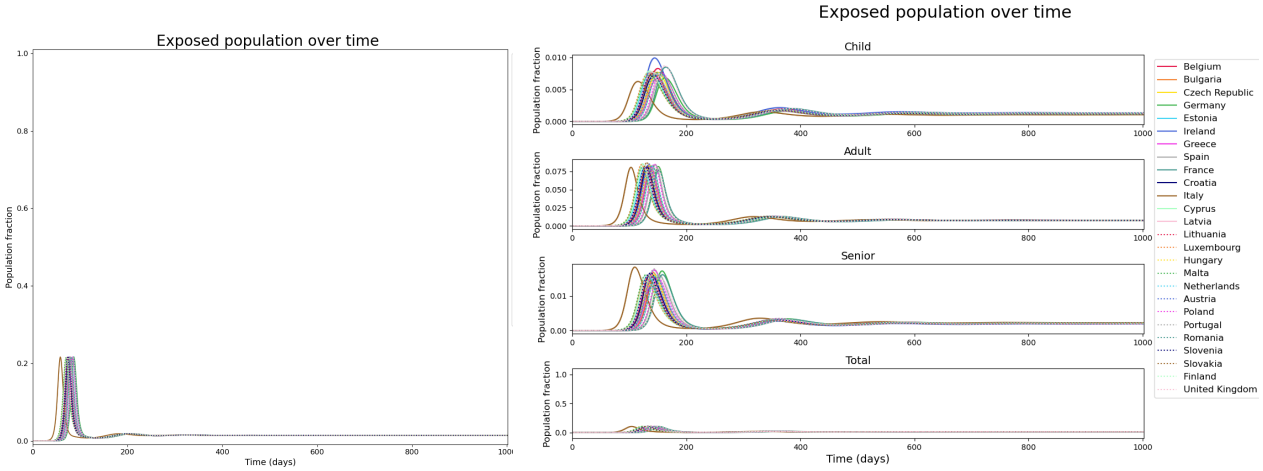

Figure 12: Exposed population fraction(s) over time for fixed migration with no age groups (left) and with age groups (right)

peak is $0.1\% \pm 1E-05$. Looking at both plots and the values, the general effect of the addition of age groups on the SEAIRDS-V model is that the spread of the virus is slowed down, as values are smaller in comparison and the peaks span over a longer period of time.

Group	Metric	No age group	Age group (total)
Asymptomatic	max	$0.048\% \pm 1E-05$	$0.030\% \pm 1E-06$
	min	$0.0\% \pm 0$	$0.0\% \pm 0$
	last	$0.0050\% \pm 1E-08$	$0.0038\% \pm 1E-07$
	first peak	83 (70-110)	145 (115-200)
Infected	max	$0.14\% \pm 5E-05$	$0.089\% \pm 5E-06$
	min	$0.0\% \pm 0$	$0.0\% \pm 0$
	last	$0.014\% \pm 5E-08$	$0.011\% \pm 5E-08$
	first peak	89 (75-120)	158 (128-208)
Recovered	max	$0.88\% \pm 2E-06$	$0.74\% \pm 3E-06$
	min	$0.0\% \pm 0$	$0.0\% \pm 0$
	last	$0.79\% \pm 2E-06$	$0.60\% \pm 1E-06$
	first peak	127 (83-205)	200 (130-310)
Dead	max	$0.051\% \pm 2E-06$	$0.036\% \pm 2E-06$
	min	$0.0\% \pm 0$	$0.0\% \pm 0$
	last	$0.051\% \pm 2E-06$	$0.036\% \pm 2E-06$
	bigger slope	92 (78-n.a.)	192 (140-n.a.)

Table 5: Summary of averaged maximum, minimum and last value obtained with and without age group distinction for different compartments. We also record the average day of the first peak (or bigger slope) to see if any change is present.

Summary of remaining compartments: For the Asymptomatic, Infected, Recovered and Dead compartments we summarize the most representative approximated values in Table 5. It is visible how the values are consistently higher when population is not split in group. Furthermore, looking at the day of the peak and duration (interval in which the peak takes place) there is a significant delay in the start of infection if age groups are used This is what was also reflected in the previous results.

5.2.3 Migration and age groups

Now that we investigated separately the effects the different types of migration and of population distinction, we can explore whether any interaction effects are present between the two settings. This is achieved by comparing NoRL 4 to NoRL2 (aviation based migration and no age groups) and NoRL3 (fixed migration and age groups).

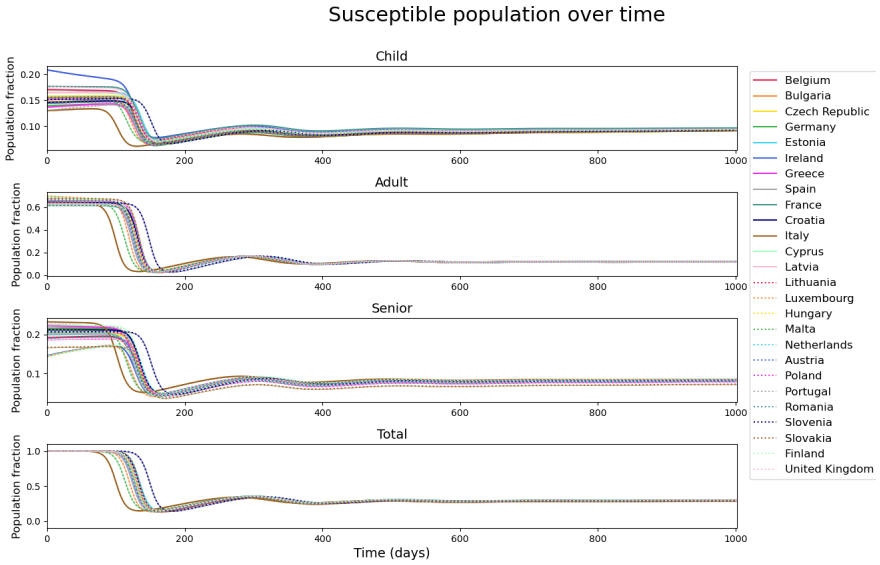

Figure 13: Susceptible population fraction(s) over time for aviation data based migration and age group distinction.

Comparing Figure 13 to Figure 11 (right) we can see that the main effect of using age group is present. The pattern is maintained in the two plots, as the first minimum reaches an approximate value of $0.14\% \pm 3E-07$ at time 166 , compared to $\approx 0.14\% \pm 3E-07$ at day 164. This results are incredibly similar especially when compared to Figure 5(right) (which represents the effects of aviation based migration), for which the minimum value was approximately $0.047\% \pm 6E-06$ around day 94.
This holds true even when comparing Figure 14 to Figure 12 (representing age group differentiation and fixed migration) and Figure 6 (representing aviation data based migration and no age groups). The pattern of the age group is visibly maintained, however the effect of aviation based migration is more subtle. In more detail, we can see in Figure 13 and Figure 14 that type of migration when combined with age groups lowers the variance in time between the peaks.The lower variance maintained for the aviation based setting with the exception of Malta and Slovakia. The former being the second country showing the first signs of infection (after Italy) and the latter being the last. This is result for Malta and Slovakia is consistent with the results for aviation based data and no age group distinction. The remaining countries variance along the time axis seems to be reduced as peaks tend to overlap more compared to the fixed migration. This is the same pattern found when comparing fixed migration and

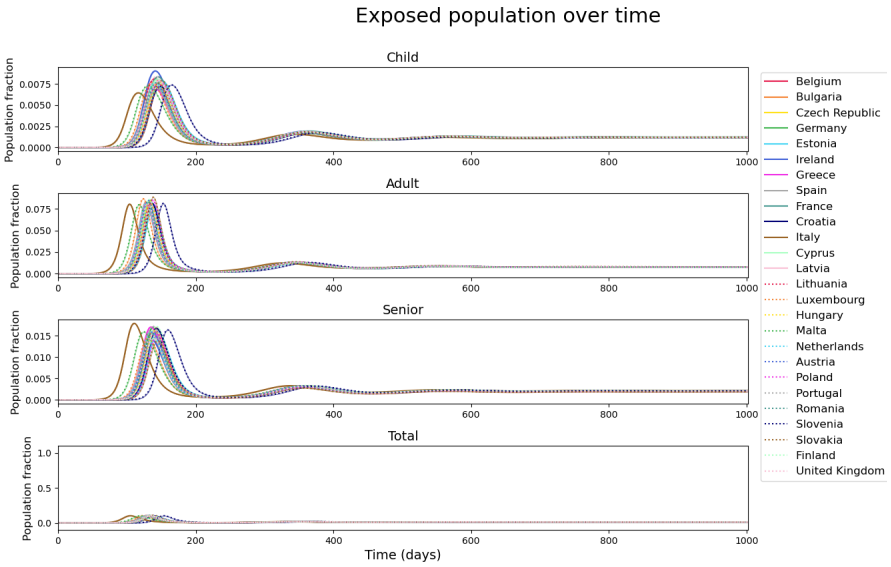

Figure 14: Exposed population fraction(s) over time for aviation data based migration and age group distinction.

aviation based migration with no age groups. Therefore there does not seem to be any significant interaction effects between age groups and type of migration.

5.3 Reinforcement learning and changing containment policies

In the previous results there were no containment policies imposed in neither of the states. Here containment policies are applied, and renewed either every day or every 15 days. After having analyzed the effect of aviation based migration, age group distinction and their interaction effects we can focus on analyzing the effect of containment policies.

5.3.1 Migration

First we look at the results comparing fixed migration and aviation based data for two conditions: containment policy changing every day or every 15.

Containment policies for fixed migration: As containment policies is one of the main changing factors we will discuss it's results first.

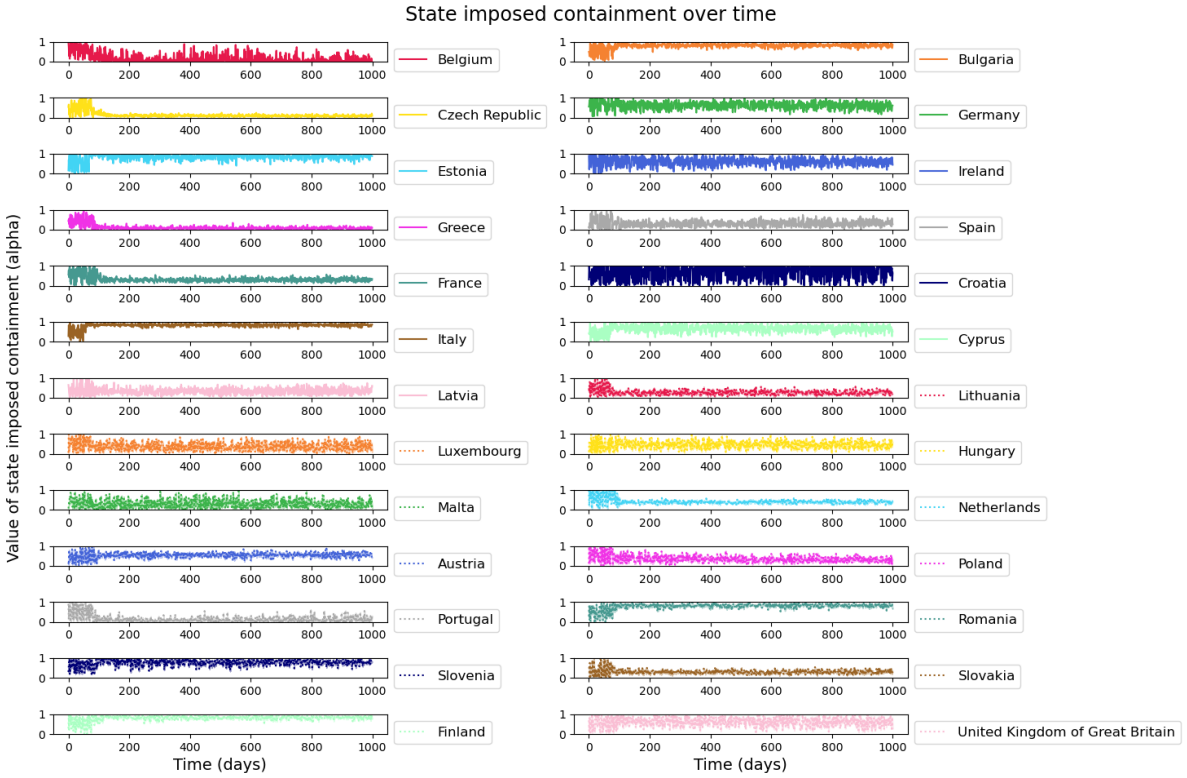

Figure 15: Containment policy over time for each country (updated every day), fixed migration and no age group distinction.

In Figure 15 we can how the containment policies (α) updated every day for each country. Overall it seems that most countries experience high variability in α from day 0 to day 100 approximately, with value ranging from 0 to 1. For most countries, from day 200 to the end of the episode, the variance of α is greatly reduced, and α tends to stabilize at a certain value. Belgium, Check Republic, Greece, France, Spain, Latvia, Lithuania, Malta, Portugal, and Slovakia (10 out of 26 countries) tend to use containment policy smaller then 0.4, thus, more geared toward freedom (closer to a value of 0). On the opposite end, Bulgaria, Estonia, Croatia, Italy, Cyprus, Romania, Slovenia and Finland (8 out of 26 countries) reach a value closer to 1 representing almost complete lockdown for those countries.

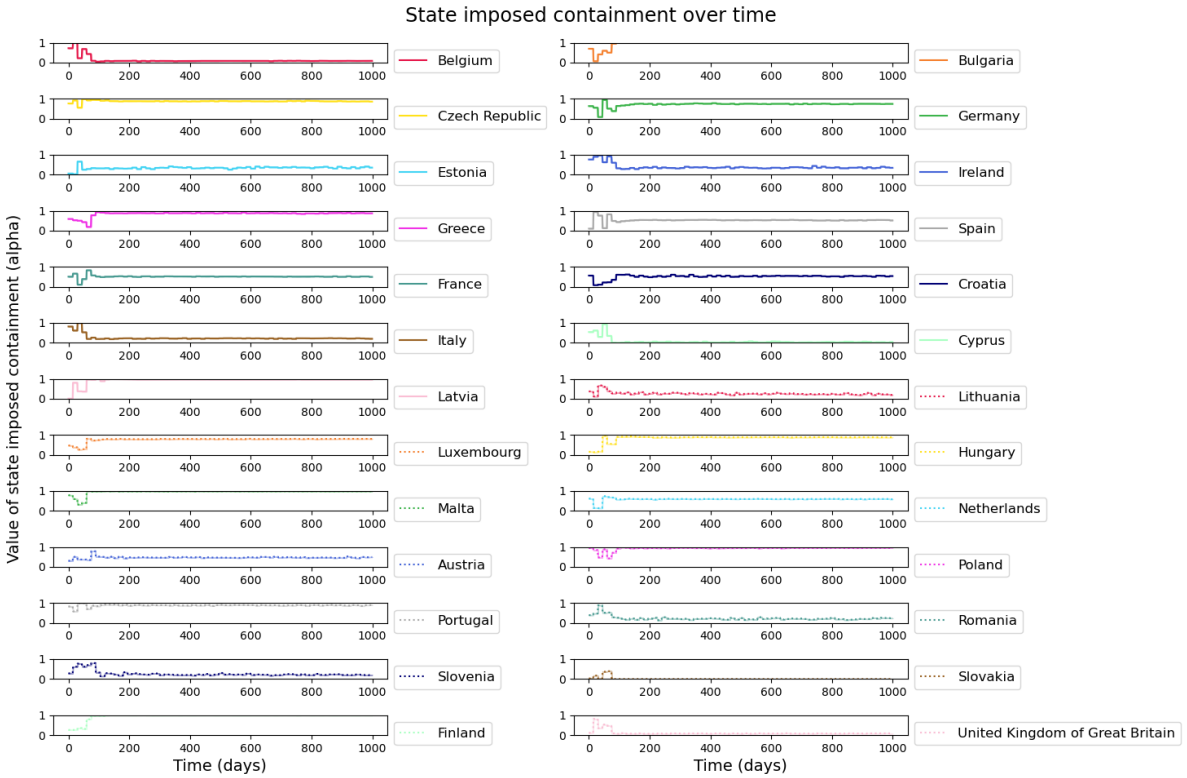

Figure 16: Containment policy used over time for each country (updated every 15 days), fixed migration and no age group distinction.

Finally Germany, Ireland, Luxembourg, Hungary, Netherlands, Austria, Poland, and United Kingdom (8 out of 26 countries) maintain a value between 0.4 and 0.6.

Days	Metric	Belgium	Bulgaria	Czech Republic	Germany	Estonia	Ireland	Greece	Spain	France	Croatia	Italy	Cyprus	Latvia	Lithuania	Luxembourg	Hungary	Malta	Netherlands	Austria	Poland	Portugal	Romania	Slovenia	Slovakia	Finland	United Kingdom
1	mean	0.5	0.6	0.4	0.7	0.7	0.6	0.4	0.4	0.5	0.6	0.6	0.5	0.4	0.4	0.5	0.4	0.3	0.6	0.4	0.5	0.5	0.6	0.6	0.4	0.6	0.5
	last	0.1	0.8	0.1	0.6	0.8	0.6	0.1	0.3	0.3	0.7	0.8	0.7	0.3	0.3	0.4	0.5	0.3	0.4	0.5	0.4	0.2	0.8	0.8	0.3	0.8	0.6
15	mean	0.5	0.6	0.8	0.5	0.3	0.7	0.5	0.5	0.5	0.3	0.5	0.4	0.6	0.4	0.5	0.5	0.7	0.5	0.5	0.7	0.8	0.5	0.6	0.1	0.5	0.4
	last	0.1	0.9	0.9	0.7	0.3	0.4	0.8	0.5	0.5	0.5	0.3	0.1	0.9	0.3	0.8	0.8	0.9	0.6	0.5	0.9	0.9	0.2	0.3	0	0.9	0.1

Table 6: Summary of values used for containment policies for the case of fixed migration. For each country we show the mean of the first 100 days of the episode (as mean) and the mean of the last 100 days (as last), these are shown for α updated every day and every 15. The table shows approximated results to $\pm 0.05\%$.

In Figure 16 and Table 6 we can compare the same experiment where containment policies are updated every 15 days. From the results, we can see that the there is a significant change. Looking at Czech Republic we can see that were before it approached a value of approx. 0.1 at the end of the episode, while if the policy is frozen the value of in now approx. 0.9. Cyprus shows the opposite effect, if the containment policy is updated every day the value at the end of the episode reaches approximately 0.7, and if updated every 15 it reaches a value of approx. 0.1.
Comparing Estonia and Latvia, who have relatively similar populations (adjacent when ordering according to population), and who both start the episode (day 1) with a low value, Latvia has an average

containment policy of roughly 0.6, and almost full lockdown is experienced from day 200 to 1000. Estonia on the contrary has a lower containment policy around 0.3 which is maintained throughout the episode. A similar point can be made when comparing Poland and Romania or Malta and Cyprus. Therefore the population of the country does not seem affect to how the policy is chosen.

Containment policies for aviation based migration: In this section, we compare, as before, containment policies updated every day and every 15 days for the case where migration is based on aviation data.

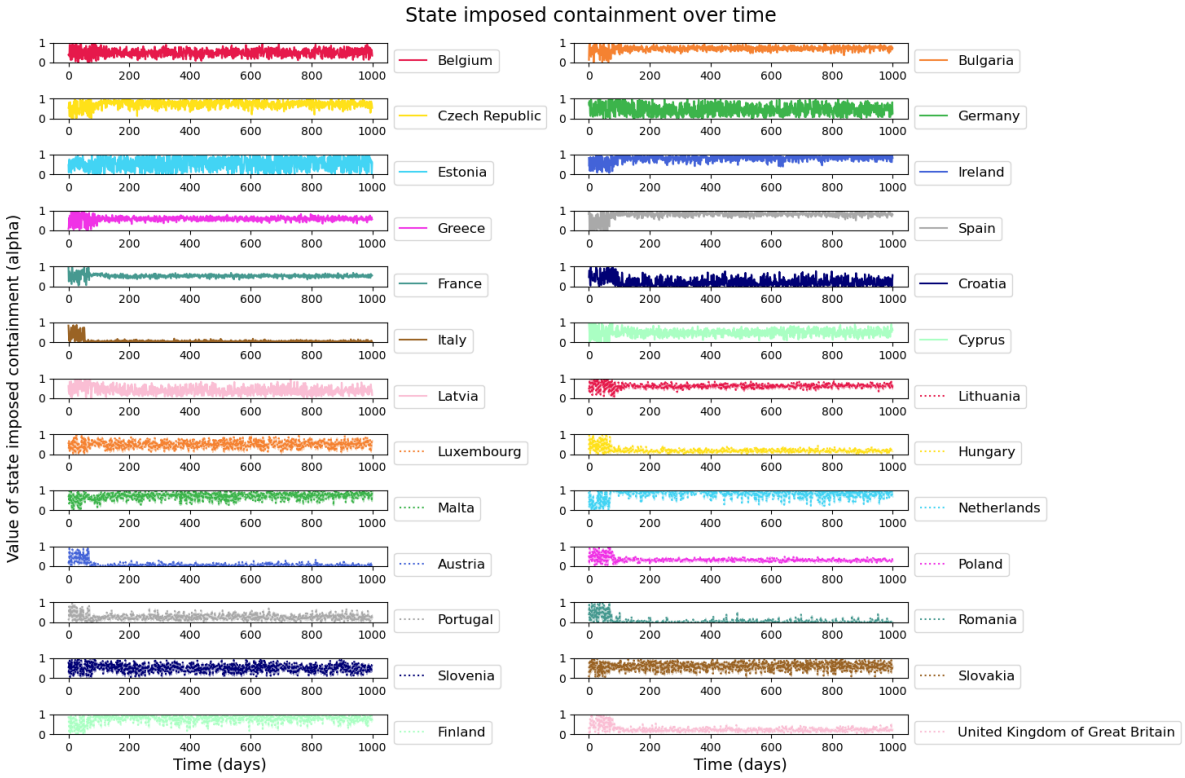

Figure 17: Containment policy over time for each country (updated every day), aviation based migration and no age group distinction.

From Figure 17 it is not possible to see any concrete difference with Figure 15 as they both show high variance in containment policy value between day 1 and 100. Both cases also have a reduced variance for the remainder of the duration of the experiment. Quantitatively analyzing the results from Table 7, from day 100 to 1000, 7 out of 26 countries use a containment policy larger then 0.7, 8 out of 26 countries apply containment policies below or equal to 0.3, and 11 countries use containment policies corresponding to values between 0.4 and 0.6. This result is comparable to earlier results where containment policy is updated every day. Even when looking at the average difference between mean and last value no patterns seem to be visible.
We also compare countries by also considering country size, we can see that for the 6 countries with the largest population such as: Germany, France, United Kingdom, Italy and Spain although they show different patterns the first 100 days of the experiment they do reach approximately the same value for the remainder of the experiment. The last largest country Poland, does show different patterns reaching a value closer to 0 for aviation based migration and closer to 1 for fixed migration. For the 6 countries with smallest population (Slovenia, Latvia, Estonia, Cyprus, Luxembourg and

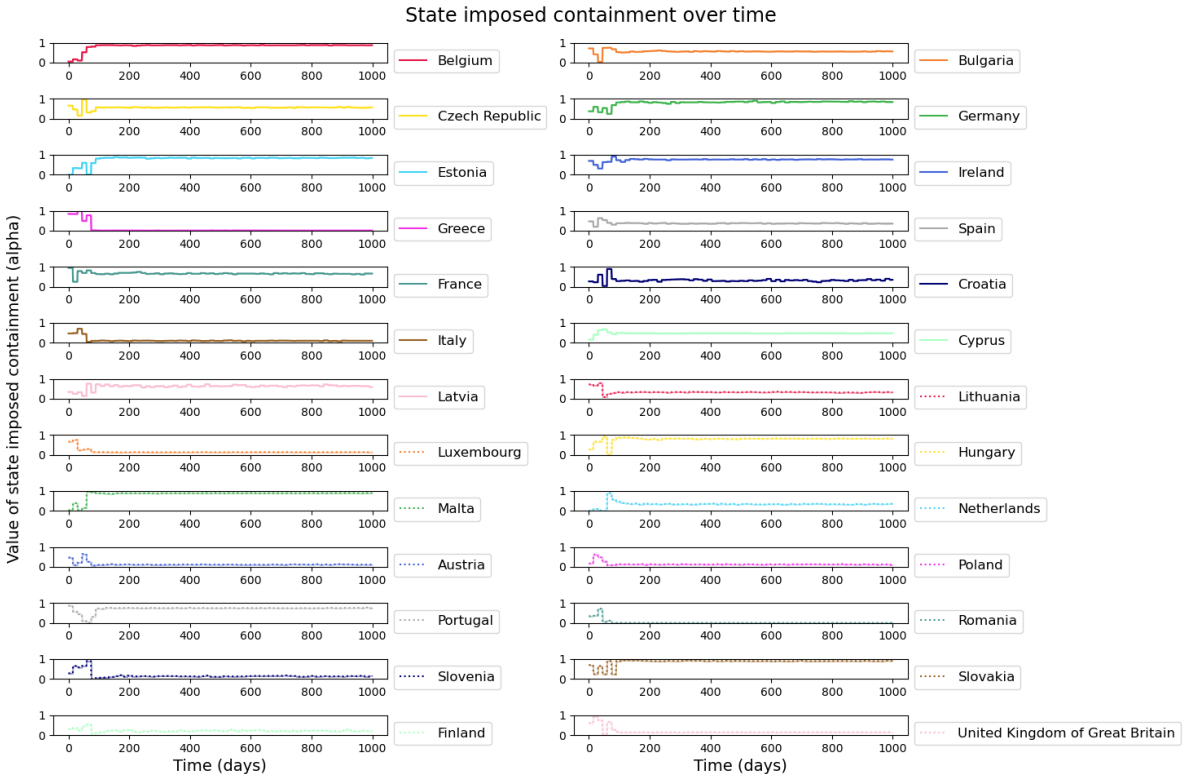

Figure 18: Containment policy used over time for each country (updated every 15 days), aviation based migration and no age group distinction.

Malta), two countries (Slovenia and Luxembourg) reach a value closer to zero while the reminder reach around 0.5 or above.

Days	Metrics	Belgium	Bulgaria	Czech Republic	Germany	Estonia	Ireland	Greece	Spain	France	Croatia	Italy	Cyprus	Latvia	Lithuania	Luxembourg	Hungary	Malta	Netherlands	Austria	Poland	Portugal	Romania	Slovenia	Slovakia	Finland	United Kingdom
1	mean	0.5	0.6	0.5	0.5	0.4	0.6	0.5	0.5	0.6	0.5	0.2	0.5	0.6	0.6	0.5	0.4	0.6	0.6	0.3	0.5	0.3	0.4	0.6	0.6	0.5	0.5
	last	0.5	0.7	0.7	0.4	0.5	0.8	0.6	0.8	0.5	0.2	0.1	0.5	0.4	0.6	0.5	0.2	0.7	0.8	0.1	0.3	0.3	0.1	0.5	0.6	0.8	0.2
15	mean	0.5	0.6	0.5	0.5	0.4	0.6	0.6	0.4	0.7	0.4	0.3	0.5	0.4	0.5	0.4	0.6	0.5	0.3	0.3	0.3	0.4	0.2	0.5	0.5	0.3	0.5
	last	0.8	0.6	0.5	0.8	0.8	0.7	0.1	0.4	0.7	0.3	0.1	0.5	0.6	0.3	0.2	0.8	0.8	0.3	0.1	0.1	0.7	0	0.2	0.9	0.2	0.2

Table 7: Summary of values used for containment policies for the case of aviation based migration. For each country we show the mean of the first 100 days of the episode (as mean) and the mean of the last 100 days (as last), these are shown for α updated every day and every 15. The table shows approximated results to $\pm 0.05\%$

We can also compare the containment policies values if updates take place every 15 days. Figure 18 and Figure 16 (respectively from aviation based migration and fixed migration). The type of migration does seem to affect the values of α as the pattern of each country does significantly change. A small number of countries do seem to show a similar pattern such as Italy, Slovenia, Romania and United Kingdom, however these are the minority, being only 4 out of 26 countries. Furthermore as for fixed migration there is also a lot of difference between the values obtained for α based on the updating time of the containment policy.

Exposed population: As a quick remainder, for the case of no reinforcement learning (no containment measures) the exposed population reached a peak of approximately 0.20% (see Figure 6). For all cases plotted in Figure 19 the average value is less the 0.20%.
Comparing the left and right plots of Figure 6 we can notice that the right plots show a higher variability in maximum value. In top-left plot the maximum value spans from approximately 0.17% to 0.2%. In top-right plot the maximum value of Exposed population ranges from approximately 0.15% to 0.21%. We can also observe that the country of Slovakia has a quite significant dip in the middle of the peak region. For the bottom section of the figure, the maximum value for the bottom left plot is around 0.17% to 0.2%. Finally, for plot at bottom right the values are approximately 0.16 to 0.21%. Overall, the lowest values were obtained for the setting with fixed migration and α updated every 15 days.

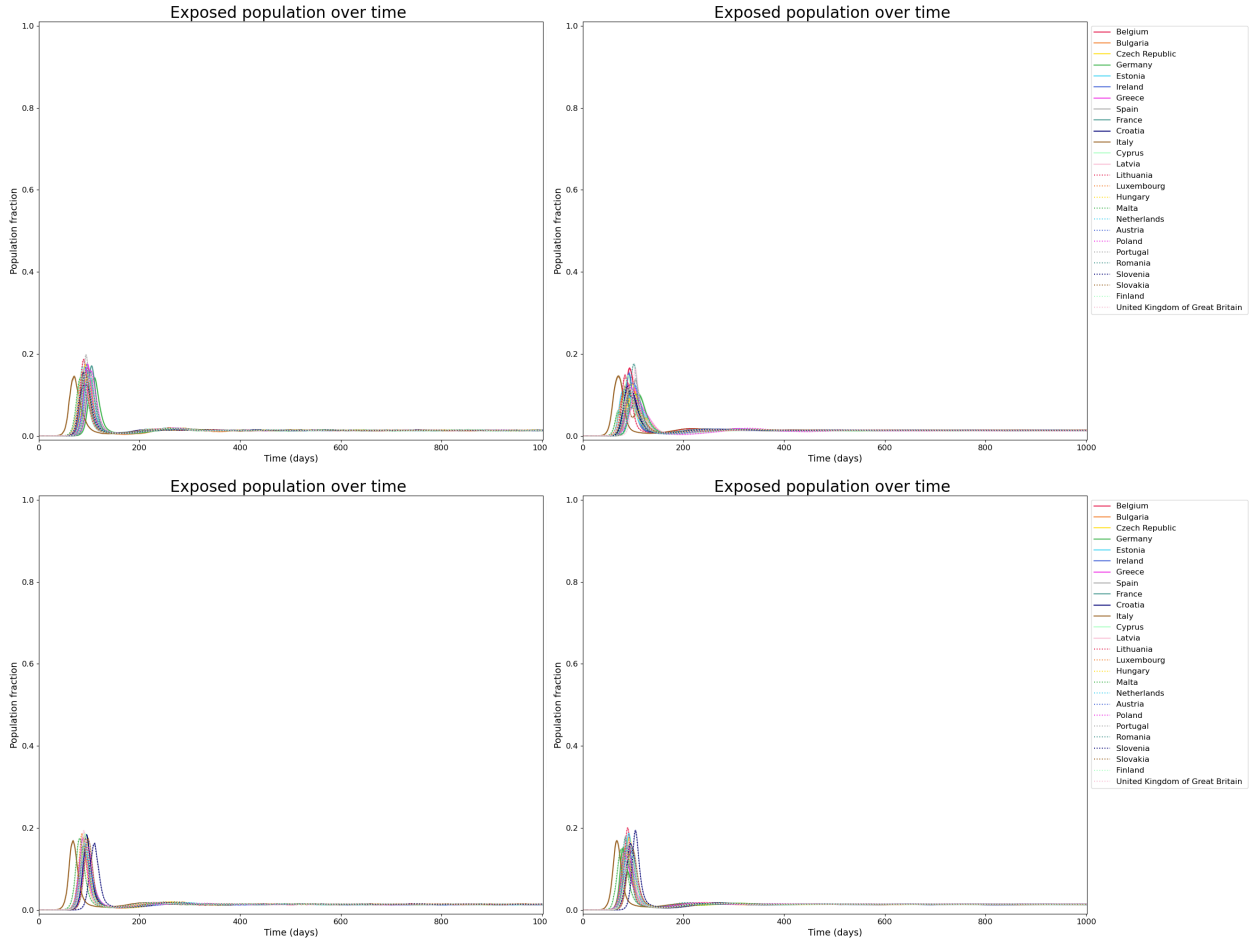

Figure 19: Exposed population fraction (no age group differentiation) over time for fixed migration (top) and aviation based data (bottom), and on the the left α updated every day and right updated every 15.

Recovered population: In Figure 20, we maintain the same disposition as previously used to make comparison easier on the reader. Comparing the left side of the image (α updated every day), we can observe how the variance between countries is smaller compared to the right side, in particular when aviation based migration is used. Overall fixed migration seems to lead to more readable and insightful trends and results. For instance, looking at the top right plot, we can see that 2 countries

seem to have consistently lower Recovered people, those countries being Poland and Finland. From the containment policies in Figure 16 we can see that these countries were experiencing complete closure. Regarding the values, all experiments reach a maximum value,at the peak, of approximately to $0.87\% \pm 0.01\%$.

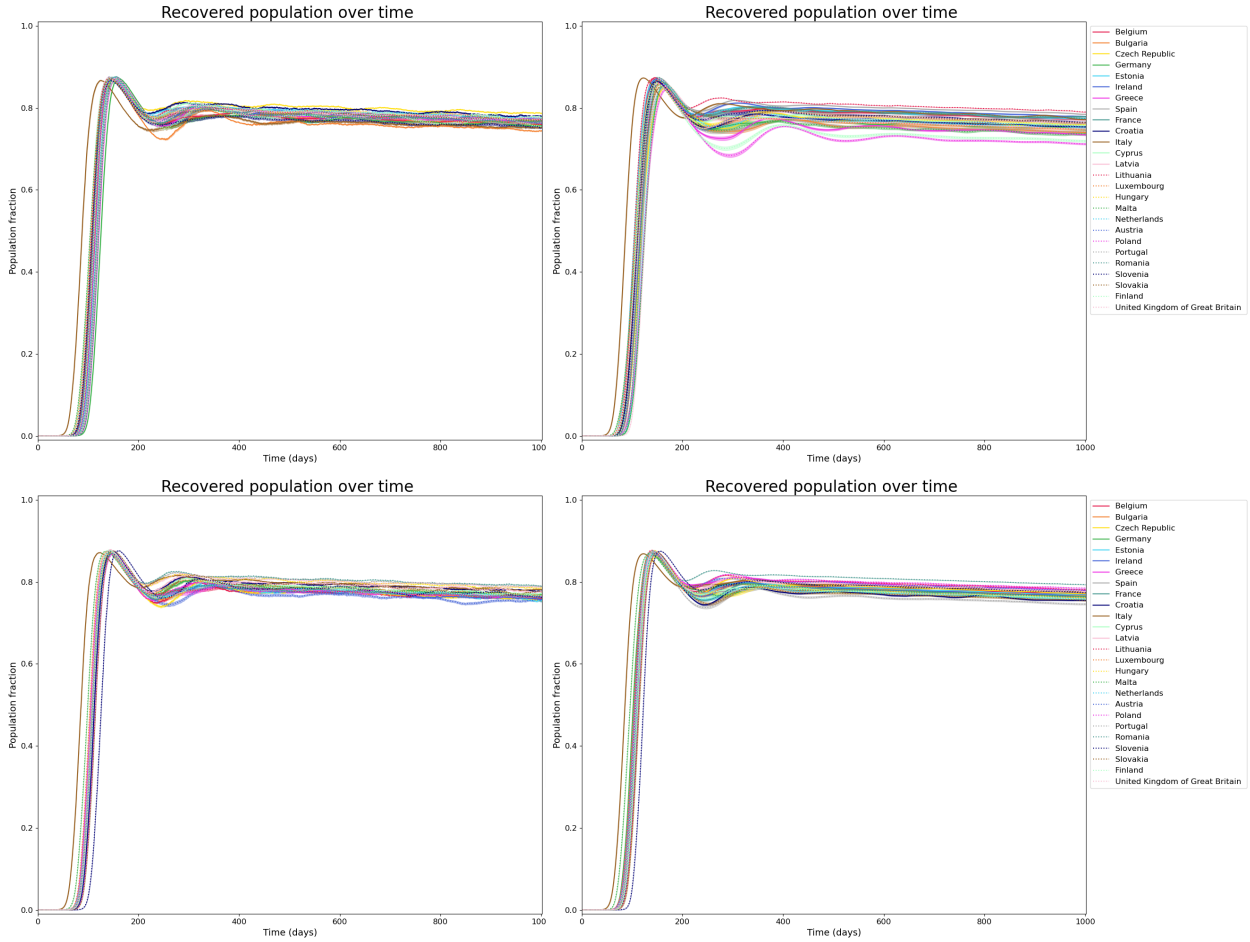

Figure 20: Recovered population fraction (no age group differentiation) over time for fixed migration (top) and aviation based data (bottom), and on the the left α updated every day and right updated every 15.

Deceased population: First we compare to the results to the no reinforcement learning results (thus no containment policy). For both types of migration, the last value reached was $0.051 \pm 2E06$. For the case of changing containment policies, for fixed migration the last value is on average $0.048\% \pm 0.0018\%$ for α updated every day, and $0.048\% \pm 0.0015\%$ for α updated every 15 days. For aviation based migration, with containment policies changed every day the last value is on average $0.049\% \pm 0.008\%$, and if updated every 15 days the value is $0.049 \pm 0.001\%$. These values represent the average per country and are approximated.

From these values, because they are averaged across country and approximated it is hard to conclude if any significance difference is present. However this compartment is very significant as the aim of the experiment is to reduce the mortality in each country. Therefore, we added a red background showing if results are significantly different (following a confidence interval of 95%) for all countries, with confidence being computed per time-step. The significance is calculated comparing the type of

migration. Thus, bottom left plot of Figure 21 shows the significance when compared to the top left plot. In other words, comparing fixed migration and aviation migration for α updated every day. For plot bottom right and top right the same is applied just looking at α updated every 15 days. We can see that results seem to only be significant during the first 250 days, while during the remainder of the experiment the values are not significantly different. This shows that, whether a country maintains full closure or full freedom, during the latest stages of the virus spread, this has no effect on the mortality rate.

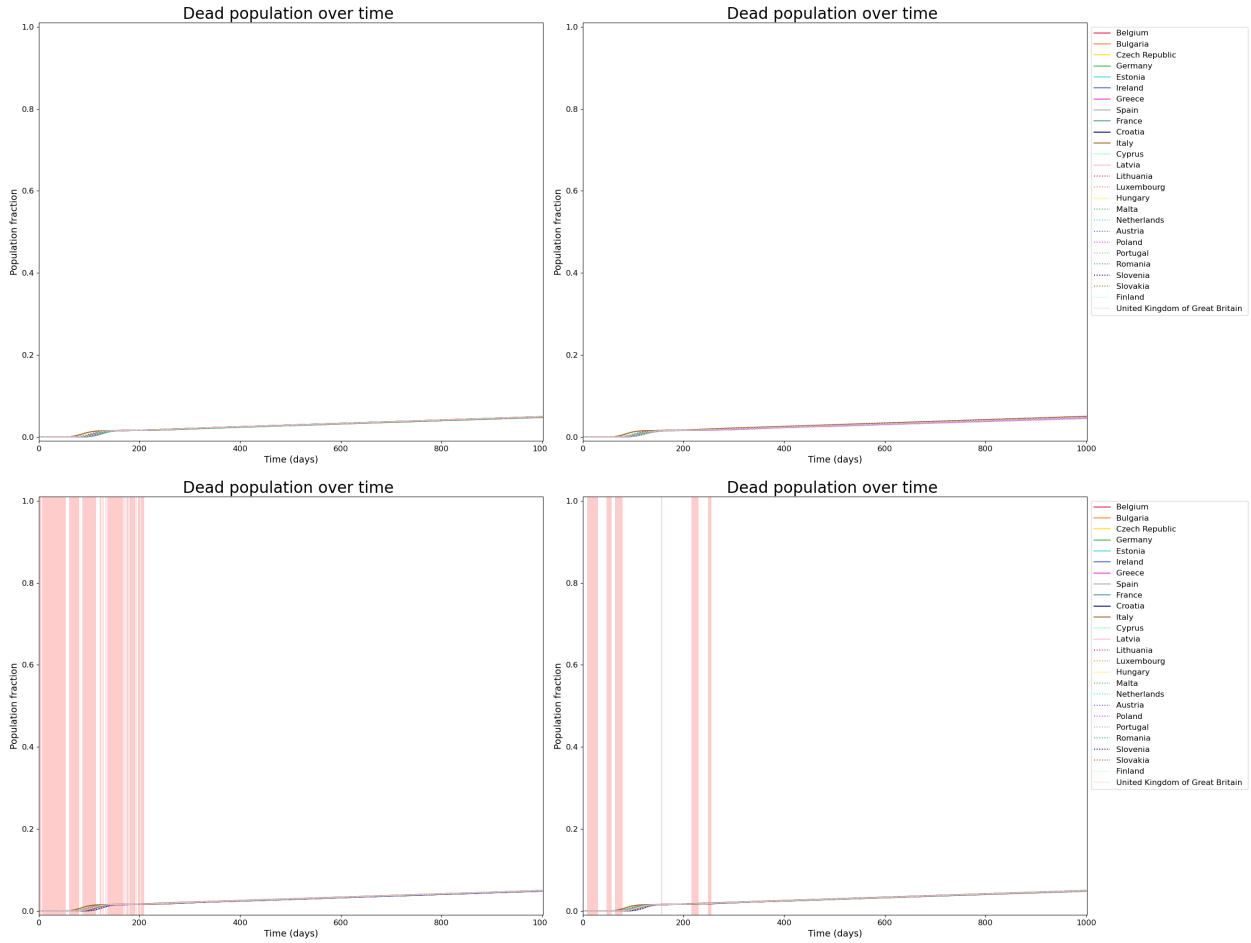

Figure 21: Deceased population fraction (no age group differentiation) over time for fixed migration (top) and aviation based data (bottom), and on the the left α updated every day and right updated every 15.

Furthermore, we can compare plots bottom left and bottom right to see how significance varies according to how often α is updated. If containment policies are updated every day there are significant difference during the peak period of virus spread when comparing the two types of migration. On the other hand, if containment policies is updated every 15 days there is some significant difference right before the first increase in deaths, and right after day 200 due to the second wave of spread of the virus. This shows that the type of migration has less effect if containment policies are changed every 15 days.

GDPloss: GDPloss is negatively affected by increase in Infected group, although we did not look at the infected group results in particular, we have seen that they follow the same pattern of the

exposed group (since exposed population progresses to either infected or asymptomatic with a 75%-25% ratio). In fact, if we compare result top-right of Figure 22 to top-right of Figure 19 we can see that the country of Romania had the highest value for Exposed population fraction (compared to other countries) and it has the lowest GDPloss minimum. Furthermore, the negative peak of the GDPloss (shown in Figure 22) occurs at the same time as the Exposed group, and thus, also the Infected group. The lowest average GDPloss derivative is obtained by updating α every 15 days (for fixed migration) with a minimum value of -6.4 ± 1.2. This result is in line with the Exposed group, which had the smallest maximum value for the same settings. Top-left and bottom right plots show higher variance when comparing the minimum value reached by each country. It is noticeable, that for all experiments the GDP does not fully recover. This is due to the fact that the virus is still present within the population (this is also confirmed by the value of the Dead group steadily increasing).

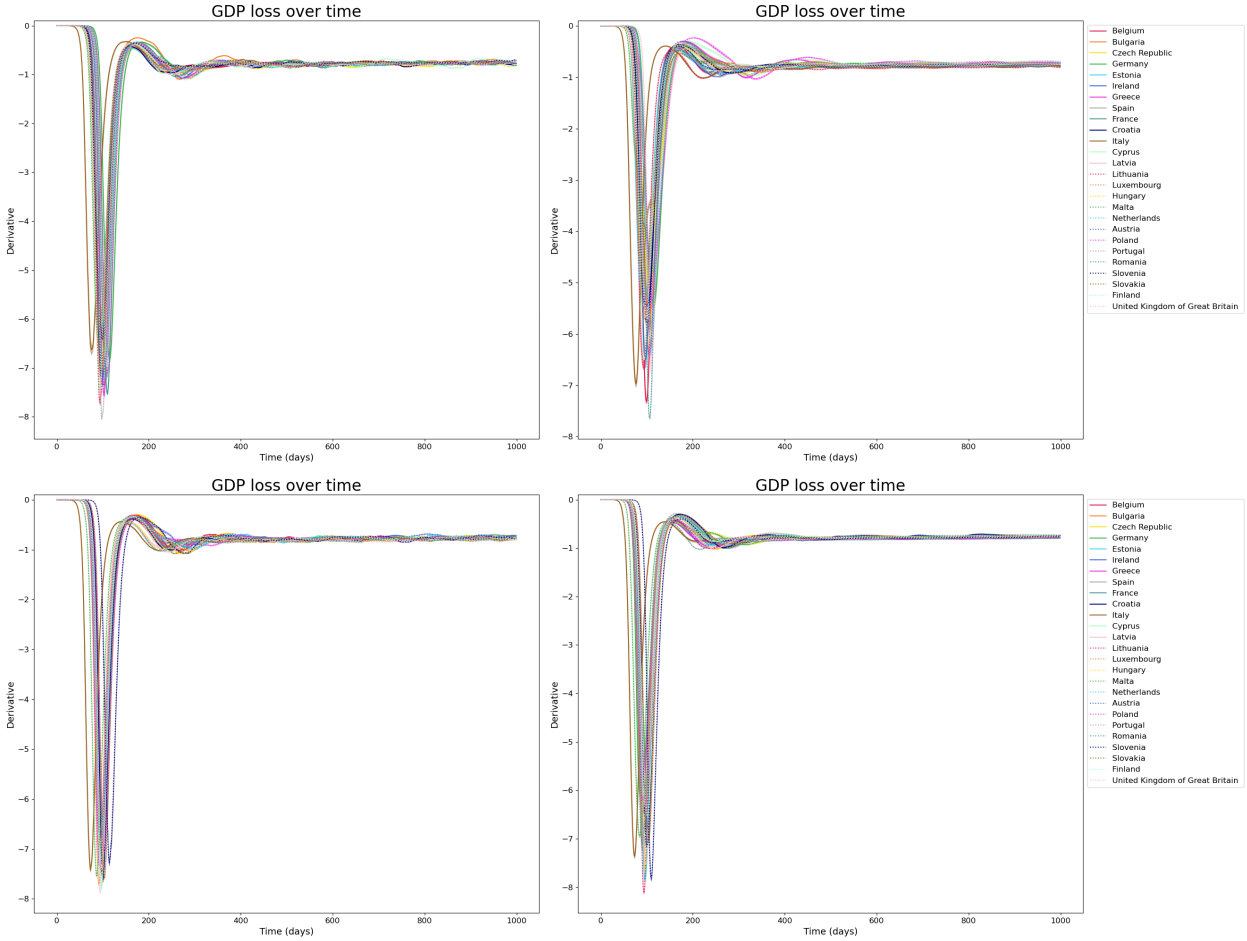

Figure 22: GDPloss (no age group differentiation) over time for fixed migration (top) and aviation based data (bottom), and on the the left α updated every day and right updated every 15.

When discussing the results for the containment policy, SEAIRDS-V groups, and GDPloss, using aviation data migration (compared to fixed migration) did result in a significant difference (mainly visible for some countries like Malta and Slovenia), however, this added data does not result in a better visible policy, in fact, trends in results are more easily visible for the case of fixed migration. This is mainly due to the fact the values for aviation based migration are higher (4E-05 on average but with some countries reaching almost 0.003)(see Figure 2) than the value chosen for fixed migration. This causes the virus to spread faster (compared to fixed migration) between the countries. Thus, also

explaining why there is less variance between countries in the results, making it hard to distinguish any specific pattern.
Furthermore, when looking at the results with no reinforcement learning, it was concluded that there is no interaction effect between migration and age-group distinction. For this reason, for the next sets of results, looking at the effects of age group we are only going to compare results with fixed migration.

5.3.2 Age groups

Containment policies: In Figure 23, we show the value of the state imposed containment policies (updated every day) for every country, in the case were the population is divided in age groups. All countries show an initial period of time up to day 200 were there is a high variation in values. At around day 200, 19 out of 26 countries increase their containment policies values (and show less variance). Other countries, however, seem to maintain the same average value and variance such as Czech Republic, Croatia, Italy, Latvia, Malta, Netherlands and Portugal (7 out of 26). Out of the 26 countries 24 maintain an average value above or equal to 0.5 , especially after day 200. However two countries maintain a value below 0.5, this countries being Malta and the Netherlands. The Netherlands in particular maintaining complete opening after day 200.

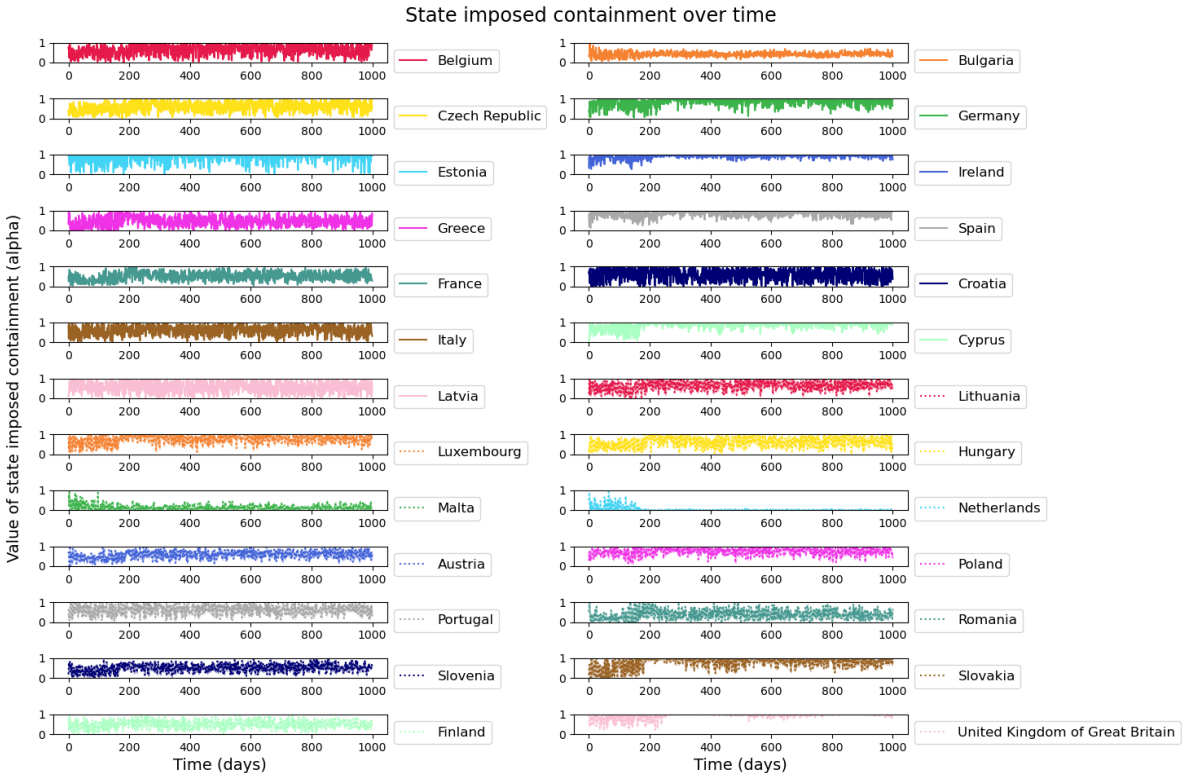

Figure 23: Containment policy over time for each country (updated every day), fixed migration and age group distinction.

On the other hand, if containment policies are updated every 15 days 8 out of 26 countries have an average value of the last 100 days over 0.7. The same number of countries have an average between 0.7 and 0.3, resulting in 10 countries having an average below 0.3.
In the results for containment policies updated every day we saw a pattern of rising the value of containment policies across 19 countries, however in this case (update every 15 days) such pattern is not

visible. One noticeable aspect is that the variance between consecutive policies is on average rather low. The countries that show the highest variance being, Belgium, Croatia , Cyprus, Luxembourg and the Netherlands.

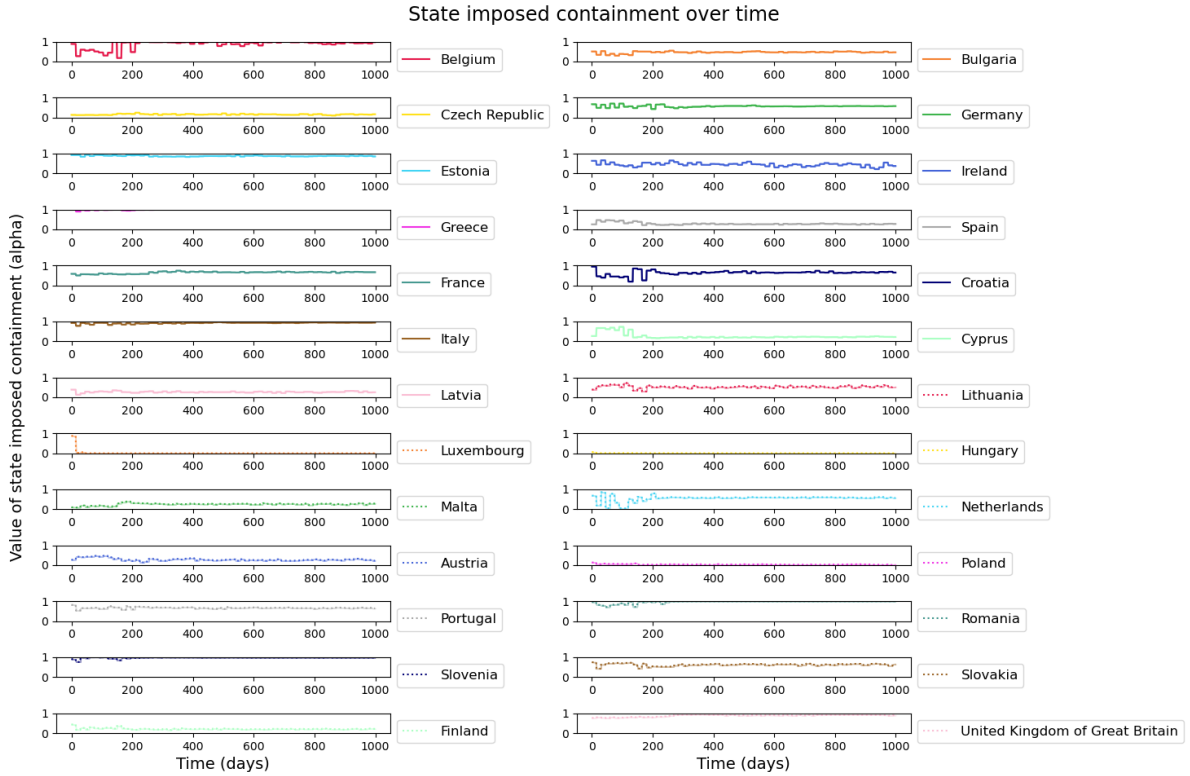

Figure 24: Containment policy used over time for each country (updated every 15 days), fixed migration and age group distinction.

Susceptible population: For the case in which the population is divided in age categories, we are also interested in the results for the Susceptible group as this gives information on how population is distributed among the countries, which might be insightful when looking at other groups. We can see from the results that most counties have a averaged ratio of 0.65% of the population belonging to the adult category. The child category ranges between 0.13% to 0.18% for most countries with Ireland having the most children with around 0.21%. For the senior group the percentages range from 0.15% to almost 0.25%.

For the case where containment policies (α) are updated every day (shown in Figure 25-right) the overall the minimum value reached by the Susceptible population ranges between Italy reaching the lowest value of 0.27% and the highest minimum being 0.46% for Ireland, with the minimum taking place at day 200 for most countries.

For the children group the Susceptible population diminishes of around 0.06%, the adult population of around 0.55% and the senior population of around 0.08%.

If, on the contrary containment policies are updated every 15 days the pattern vary greatly. In this case, the shape at the lower point that was visible earlier is now less clear, and there is a higher distinction between the countries. The lowest value is reached by Slovenia and Luxembourg with a value of 0.39%. The highest minimum values where Greece and United Kingdom, of respectively 0.46% and 0.48%. Therefore we can see that the lowest minimum value is around 0.12% higher if the

containment policies are updated every 15 days, however the highest minimum value does not have the same change.

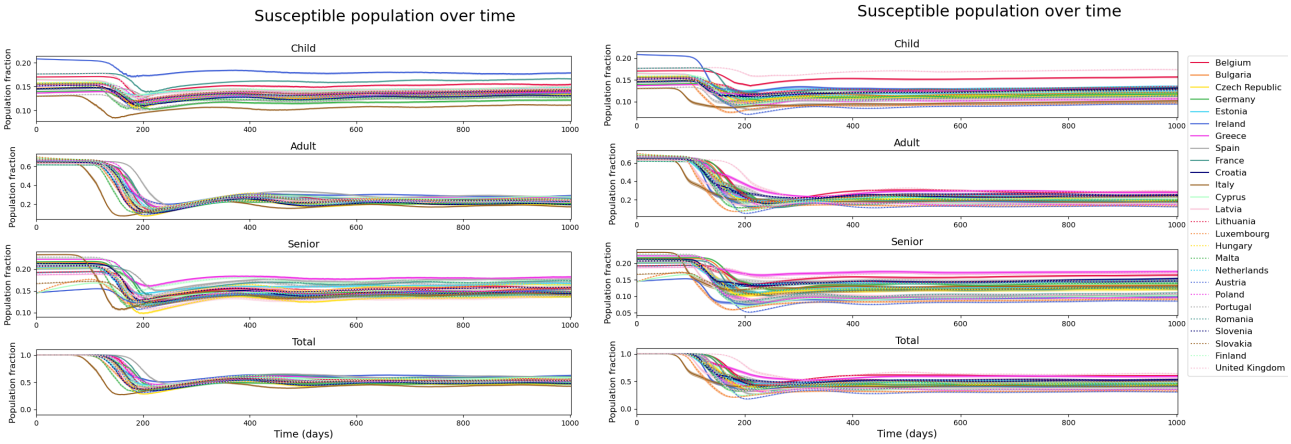

Figure 25: Susceptible population fraction over time for each country updated every day (left) and every 15 (right), fixed migration and age group distinction.

Exposed population: First we are going to explore the results under conditions where α is updated every day (shown in Figure 26-left). First we look at the results for the child group which show the values ranging from value is 0.0021% for Greece to 0.0050% for Czech Republic (roughly 0.0015% to 0.004% in the plot). the adult group ranges from 0.068% for Austria and 0.047% for Ireland. The total Exposed population ranges between 0.054% an 0.085%.

For the case where containment policies are updated every 15 days, the results show that the children group ranges between 0.0013% for the United Kingdom and 0.0074% for Ireland. For the adult group the values are between 0.038% for the United Kingdom and 0.084% for Luxembourg. In total the Exposed population ranges from 0.042% to 0.1%.

One important thing to notice is that the values described do not exactly match those represented in the plot. This can be see in both plots. The most striking example is the country of Luxembourg reaching a value of 0.084% (for the adult group with α updated every 15 days), however in the plot (see Figure 26-right) Luxembourg only reaches roughly 0.04%.

This difference is because the plot shows for each country, an average of the five values obtained from five experiments, for each time step. Instead the value discussed, considers the highest and lowest maximum value disregarding at which time step it occurred. The difference is therefore caused by the peaks of Ireland not occurring on the same day (or close range). The same also holds for United Kingdom however the value is less visible from the plot.

Looking at the span of the peaks (for the adult group we can also see that for α updated every day the peaks seemed to occur around the same time-period, more in detail, starting to increase at day 100 reaching the peak at day 175, and finally reaching the end of the peak around day 285. On the other hand. for α updated every 15 days the values start to increase at day 85 reaching the peak at day 175, and finally reaching the end of the peak around day 300, however the peaks range between day 175 and day 200.

Recovered population: Let's now analyze the results for the Recovered group shown in Figure 27.

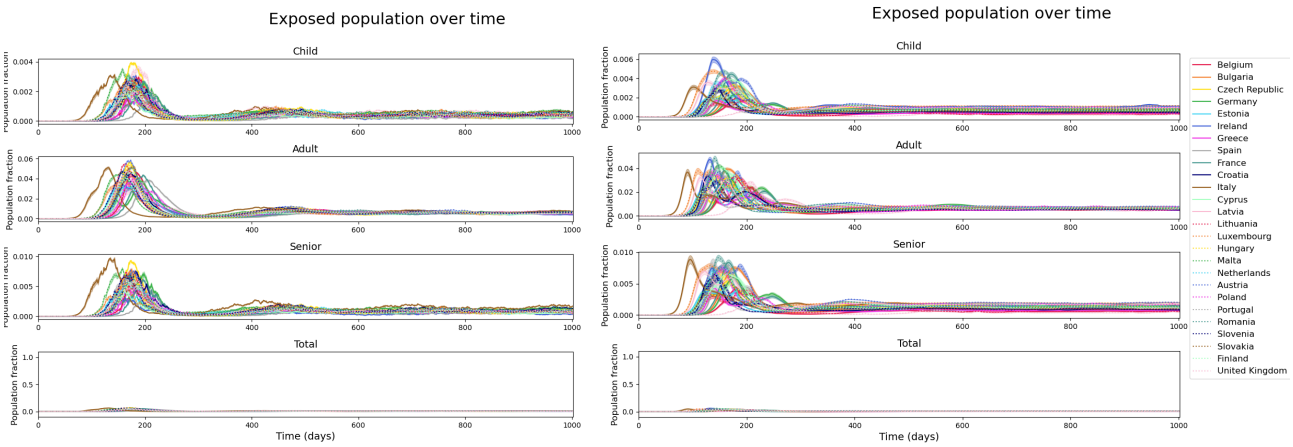

Figure 26: Exposed population fraction over time for each country updated every day (left) and every 15 (right), fixed migration and age group distinction.

First we take a look at the results if containment policies are updated every day. For all population groups the recovered population reaches the highest value at day 230. The maximum value for children ranges from 0.014% for Spain to 0.041% for the United Kingdom. For the adult group the values range from 0.42% for Ireland to 0.51% for Austria. In total the values range from 0.47% to 0.64%, and the results are inline with the plots showing low variance between countries trends.
Looking at the results for containment policies updated every 15 days, the highest values are found around the same day (230). In this case the children group has values ranging from 0.014% for the United Kingdom to 0.071% for Ireland. The adult group has values ranging from 0.39% for Greece to 0.54% for Austria. In total the values range between 0.45% to 0.74%.

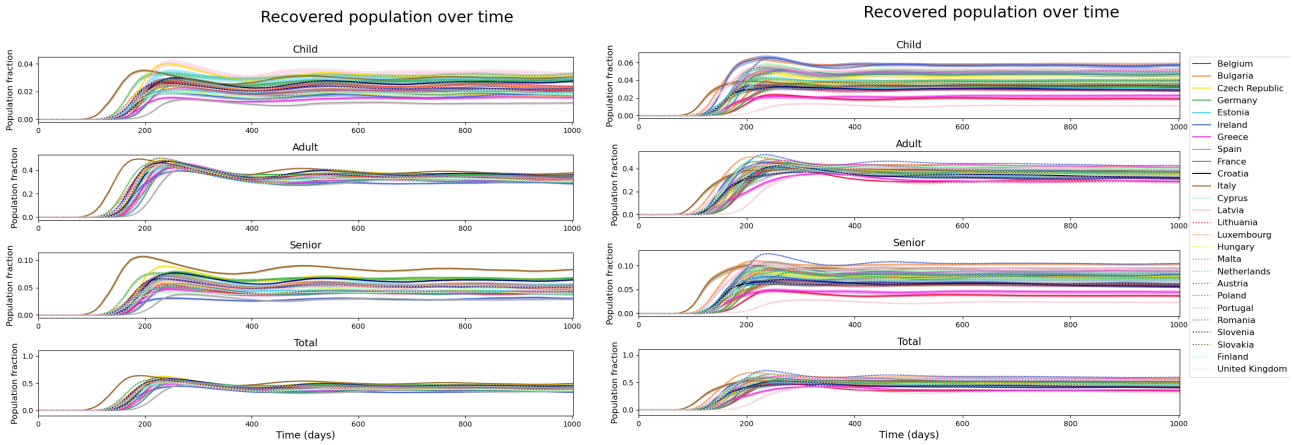

Figure 27: Recovered population fraction over time for each country updated every day (left) and every 15 (right), fixed migration and age group distinction.

Deceased population: The last population group to be explored is the Dead group.
First, as before, we are going to explore the values for containment policies being update every day (see Figure 28-left). In this case, the range for the children is from 0.00068% to 0.0018%, for adults the range is 0.013% to 0.017%. Overall, for all population the range is between 0.022% to 0.027%.

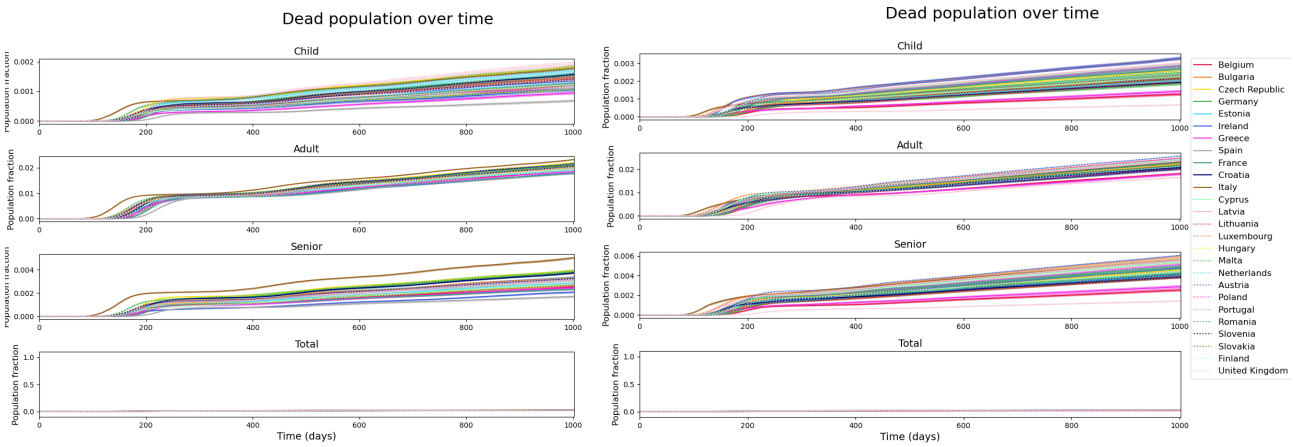

Figure 28: Dead population fraction over time for each country updated every day (left) and every 15 (right), fixed migration and age group distinction.

For the other case if containment policies are updated every 15 days (see Figure 28-right), we have a range from 0.0067% to 0.0032% for children, 0.016% to 0.025%, with an overall range from 0.018% to 0.035%. All the values discussed are also reflected on the respective plots.

GDP loss: For the GDP loss plotted in Figure 29, we only consider the GDP for the adult group. This, is due to a modelling choice made in our methods. For the case of age-groups the GDP is calculated on the basis of the adult population only as it constitutes the working force of a country.

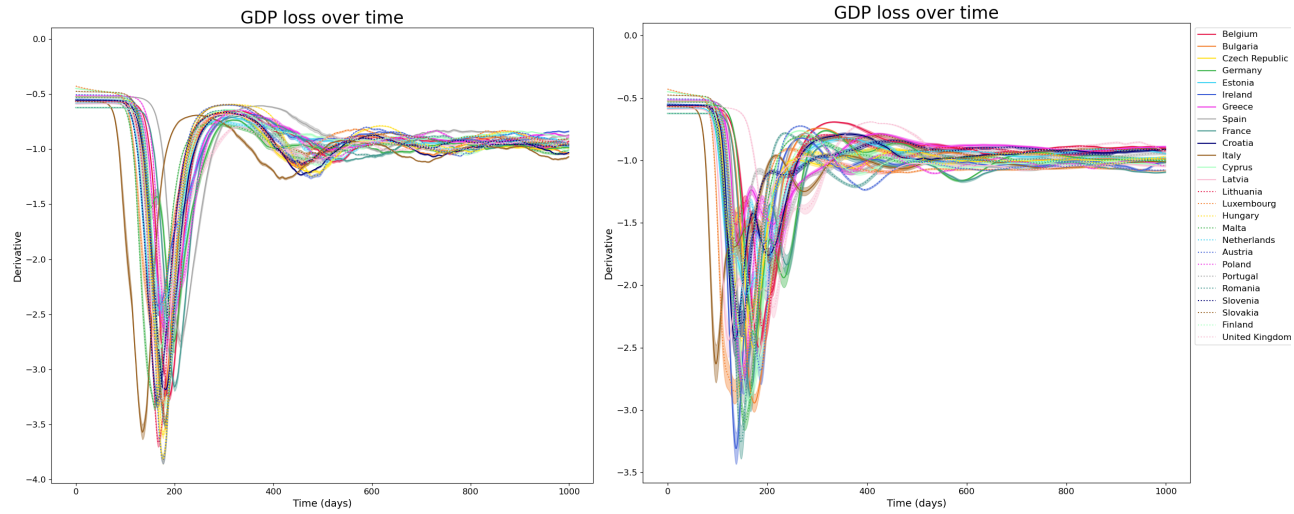

Figure 29: GDP loss over time for fixed migration, on the the left α is updated every day and on the right is updated every 15.

For the case in which containment policies are updated every day the lowest minimum GDP loss derivative has a value of -3.8 and the highest minimum has value -2.4 (see Figure 29-left). If containment policies are updated every 15 days the lowest value is -3.3 and the highest is -1.7 (see Figure 29-right).
The last values reached are higher for the former case, between -0.7 and -1.0, while for the latter case the values are between -0.9 and -1.1.

Overall comparing the dip we can see that if α is updated every day, there is less variance between countries, while in the other case it is possible to see a few countries showing two dips (Greece and Croatia) in the same period of time where countries only have one (Luxembourg and Malta). This results are in line with the results shown for the Exposed adult group (see Figure 26).

6 Conclusion

In this section we are going to first discuss and summarize the results found, secondly we are going discuss the advantages and limitation of our work, and finally we are going to mention possible future work.

6.1 Discussion of results

When comparing the effects of fixed migration and aviation based migration, it was found that aviation based migration had higher values compared to fixed migration. This resulted in the virus spreading more quickly between countries, and thus having less variation in the timing of the starting and peak of infection. This reduce in variability resulted in less noticeable patters.
We also compared the effects of using age group distinction (with contact matrices) in comparison to the no age group condition. If age groups are used, more population follows the containment measures imposed by the state (50% of adult and 100% of children and senior) compared to the no age group condition (50% of population follows containment policies). An increase in population following state imposed closure reduces the contact rates between demographics, therefore explaining the decrease in infection and mortality. Furthermore reduce in contact also explains the delay in peak of infection.
When comparing the results, for both the no age group and age group condition, between our baseline (no reinforcement learning and no containment measure) to the reinforcement learning approach, it was found that results for the RL approach (both daily updated and 15 days update) showed reduced mortality (or in some cases a similar result) and overall increase delay in peak of infection.
In comparing the results daily update and 15 days update in RL, the results showed that freezing policies for 15 days led to increased variance between countries, and therefore easier to identify patterns. The results showed lower minimum but also higher maximum compared to the daily update condition. Another important pattern that emerged form the results, in the 15 day update was peaks being broken down in two smaller peaks but in a longer period of time, This does not decrease the cumulative value of infection overall, but it does spread the infection in time.
To put the results more explicitly in terms of our research questions, a consistent best policy according to the reinforcement learning approach is not found. In the results, 19 out of 26 countries simulated do increase containment policies at the time of peak of infection for one specific condition (fixed migration, age group and daily update), and 24 out of 26 countries raise containment policies after peak of infection toward complete closure. This pattern was the only pattern visible across multiple countries. An opposite pattern that was also noticed, a smaller group of countries, maintain relaxed containment policy (toward complete freedom) also experience reduced mortality. This suggests that countries may take advantage of other countries closure to lower mortality while not reducing GDP. Other more in depth patterns where not observable in the results.
Overall the containment policy chosen was successful in containing the simulated epidemic. However, it must be acknowledged that our baseline did not implement any containment policy. Therefore, much of the success is inherent from the use of containment policies them-self. Some positive aspects of the policy, are as previously mentioned lower peaks of infection, or high peaks being broken in two smaller peaks of infection, effectively softening the infection trough time.
With regards to the different additions made in this model, the results show that using country demographics was useful in being able to model different countries by using one generic approach. The contact matrices reduced the contact in the SEAIRD-V model, therefore reducing the spread of the virus. Furthermore, our method was able to find policies both in the unrealistic case of changing

policy every day, but also to deal with policies being frozen for a certain period of time, with result showing comparable results between the two models. Migration is also a very important part of our model, as it allows countries to interact together and infection to spread. However, using realistic data did not seem to help in our aim of having more distinct patterns, if not the opposite. Furthermore the difference in results could have been obtained by adapting the parameters of the migration itself.

6.2 Methodology advantages and limitations

Although our reinforcement learning approach is very simple, the environment has been created by combining several elements. The parameters used were chosen for our particular experiments. Some important parameters are: rate of transmission, and percentage of people following state imposed containment policies. However, parameters used in the economic model are also important, as our aim (based on the reward function) is to balance the increase (or maintain) in GDP loss value while reducing mortality. In our model, the starting GDP value is the same for each country and not real data is used. For example if real GDP data would be included then the parameters would have to be changed to account for this change.
The main difficulty in combining a reinforcement learning approach in this type of environment stems from combining all different elements (migration, contact matrices and age group, and economic model) in a way to maintain a coherent and well defined model the agent would be able to learn in, by adapting formulas from previous research. Formulas had to be rewritten to allow for continuous actions and to account for contradictory values of the containment measure across papers. Another issue that was faced in regard to combining the economic model with reinforcement learning was the value of the GDP not being bound. To solve this problem, instead of considering the full GDP value we considered the GDP loss between time steps, which allowed the value to be contained between a certain interval.
Furthermore, it has to be mentioned that in our case the baseline results were considered for the case of no containment policies applied. This choice was made to be able to isolate the effect of reinforcement learning alone. However, our approach could be also compared to a baseline consisting of a heuristic policy, where, for instance, a certain percentage of infected population triggers a partial or complete lockdown. Although this approach was not investigated, form the insight gather in this study, we expect that such policy would result in oscillating SEAIRDS-V values. Mainly because containment policies should be implemented before a peak is reached. If this is not done in a timely fashion, the model may then spike a second time. Another possibility is that imposing containing measures too late may not spread peaks through time, something that was observed when employing RL methods to determine optimal containment strategies. Finally, binary (or discrete) values for containment policies may not give the agents enough freedom to allow a more organic recovery of the countries' GDPs.

6.3 Future Work

To investigate more in depth our research questions one possibility would be to allow countries to receive their own current state but also the other countries. This approach could lead to insight on cooperation between countries.
Another interesting expansion would be to group countries in sub-regions (e.g. northern Europe, southern Europe, eastern Europe) and have migration in between sub-regions only, this could lead to more easily interpreted results. Furthermore, this grouping allows a possible extension consisting in giving as input to our Actor-critic approach not only the current state of the country itself, but also the state of the countries in the same sub-region.

Another possibility would be to expand on the epidemic compartmental models since they are versatile models and can be expanded easily. For instance, a possibility would be to implement more compartments, simulate virus testing on infected and asymptomatic population, or add vaccination simulation.

7 Code Availability

Code, processed data and results for this project are available on GitHub:
https://github.com/FrancescaPerin/Covid-first-trial.
For any questions please contact: perinfrancesca9@gmail.com

Bibliography

[1] N. W. Ruktanonchai, J. R. Floyd, S. Lai, C. W. Ruktanonchai, A. Sadilek, P. Rente-Lourenco, X. Ben, A. Carioli, J. Gwinn, J. E. Steele, O. Prosper, A. Schneider, A. Oplinger, P. Eastham, and A. J. Tatem, "Assessing the impact of coordinated COVID-19 exit strategies across Europe," *Science*, vol. 369, pp. 1465–1470, Sept. 2020. Publisher: American Association for the Advancement of Science Section: Research Article.

[2] A. Parisi, S. P. C. Brand, J. Hilton, R. Aziza, M. J. Keeling, and D. J. Nokes, "Spatially resolved simulations of the spread of COVID-19 in three European countries," *PLOS Computational Biology*, vol. 17, no. 7, p. e1009090, 2021. Publisher: Public Library of Science.

[3] L. Jonung and W. Roeger, "The Macroeconomic Effects of a Pandemic in Europe - a Model-Based Assessment," 2006.

[4] E. Carletti, T. Oliviero, M. Pagano, L. Pelizzon, and M. G. Subrahmanyam, "The COVID-19 Shock and Equity Shortfall: Firm-Level Evidence from Italy," *The Review of Corporate Finance Studies*, vol. 9, no. 3, pp. 534–568, 2020.

[5] C. Liu, "A microscopic epidemic model and pandemic prediction using multi-agent reinforcement learning," 2020.

[6] Y. Vyklyuk, M. Manylich, M. Škoda, M. M. Radovanović, and M. D. Petrović, "Modeling and analysis of different scenarios for the spread of COVID-19 by using the modified multi-agent systems – Evidence from the selected countries," *Results in Physics*, vol. 20, p. 103662, jan 2021.

[7] A. Charpentier, R. Élie, and C. Remlinger, "Reinforcement Learning in Economics and Finance," *Computational Economics*, Apr. 2021.

[8] J. Langhorst, "Finding coordinated lockdown policies using multi-agent reinforcement learning." unpublished, 2021.

[9] H. W. Hethcote, "Three Basic Epidemiological Models," in *Applied Mathematical Ecology* (S. A. Levin, T. G. Hallam, and L. J. Gross, eds.), Biomathematics, pp. 119–144, Berlin, Heidelberg: Springer, 1989.

[10] F. Brauer, "Compartmental Models in Epidemiology," in *Mathematical Epidemiology* (J. M. Morel, F. Takens, B. Teissier, F. Brauer, P. van den Driessche, and J. Wu, eds.), vol. 1945, pp. 19–79, Berlin, Heidelberg: Springer Berlin Heidelberg, 2008. Series Title: Lecture Notes in Mathematics.

[11] A. Aspri, E. Beretta, A. Gandolfi, and E. Wasmer, "Mortality containment vs. economics opening: Optimal policies in a seiard model," *Journal of Mathematical Economics*, vol. 93, p. 102490, 2021. The economics of epidemics and emerging diseases.

[12] T. Oraby, M. G. Tyshenko, J. C. Maldonado, K. Vatcheva, S. Elsaadany, W. Q. Alali, J. C. Longenecker, and M. Al-Zoughool, "Modeling the effect of lockdown timing as a COVID-19 control measure in countries with differing social contacts," *Scientific Reports*, vol. 11, no. 1, p. 3354, 2021.

[13] K. Prem, A. R. Cook, and M. Jit, "Projecting social contact matrices in 152 countries using contact surveys and demographic data," *PLOS Computational Biology*, vol. 13, no. 9, p. e1005697, 2017. Publisher: Public Library of Science.

[14] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. 2018.

[15] H. Van Seijen, A. R. Mahmood, P. M. Pilarski, M. C. Machado, and R. S. Sutton, "True online temporal-difference learning," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 5057–5096, 2016.

[16] J. Peters and S. Schaal, "Policy Gradient Methods for Robotics," in *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 2219–2225, Oct. 2006. ISSN: 2153-0866.

[17] V. Konda and J. Tsitsiklis, "Actor-Critic Algorithms," in *Advances in Neural Information Processing Systems* (S. Solla, T. Leen, and K. Müller, eds.), vol. 12, MIT Press, 1999.

[18] J. Schulman, P. Moritz, S. Levine, M. Jordan, and P. Abbeel, "High-Dimensional Continuous Control Using Generalized Advantage Estimation," Oct. 2018. arXiv:1506.02438 [cs].

[19] Z. Ding, Y. Huang, H. Yuan, and H. Dong, "Introduction to Reinforcement Learning," in *Deep Reinforcement Learning: Fundamentals, Research and Applications* (H. Dong, Z. Ding, and S. Zhang, eds.), pp. 47–123, Singapore: Springer, 2020.

[20] World Bank, "Population ages 0-14, total - population ages 15-64, total - population ages 65 and above, total - population, total," 2019. data retrieved from World Development Indicators, `http://data.worldbank.org/indicator/SP.POP.0014.TO`, and indicators SP.POP.1564.TO - SP.POP.65UP.TO - SP.POP.TOTL . Accessed 23/11/2021.

[21] Eurostat, "International intra-eu air passenger transport by reporting country and eu partner country [avia_paincc]," 1993-2022. data retrieved from Eurostat Air transport measurement - passengers, `https://ec.europa.eu/eurostat/databrowser/view/AVIA_PAINCC__custom_2292592/default/table`. Accessed 15/03/2022.

[22] United Nations, "Database on household size and composition 2019," 2019. data retrieved from United Nation - Department of Economic and Social Affairs - Population Division, `https://www.un.org/development/desa/pd/data/household-size-and-composition`. Accessed 22/02/2022.

[23] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba, "Openai gym," *arXiv preprint arXiv:1606.01540*, 2016.

[24] A. G. Barto, R. S. Sutton, and C. W. Anderson, "Neuronlike adaptive elements that can solve difficult learning control problems," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-13, no. 5, pp. 834–846, 1983.