Control of Covid-19 using Agent-based modelling with Reinforcement learning.

Master Thesis
Master of Science (MSc) in Engineering (Control and Automation)

Stergios Polymenidis
Bartosz Wawrzyniak

Supervisors:
Aysegül Kivilcim
Rafał Wiśniewski

Aalborg University
Department of Electronic Systems
Fredrik Bajers Vej 7B
DK-9220 Aalborg
Spring 2021

Abstract

Humanity has been struggling with infectious diseases since the dawn of time. These diseases pose a great threat to society and the fight against them is sometimes challenging. Currently, the world is struggling with a pandemic of a virus that causes a disease called Covid-19. Disease is rarely fatal but spreads quickly and becomes easily out of control. Controlling the spread of the virus has become a challenge for governments. In this project, we propose to use a reinforcement learning algorithm to find the optimal policy to keep hospitalized and severe (requiring a respirator) cases within the imposed thresholds. We use an agent-based modelling technique to simulate society and the spread of the virus within it. Following the actions of governments regarding the pandemics, we have established a list of policies that, cause reactions similar to those in the real world. We apply the model-free value iteration reinforcement learning algorithm to the model to find a sequence of policies that will allow to control the spread of the disease and keep hospitalized and those requiring a respirator at a level that will not overload health care. We create three models with different complexities to test the operation of the algorithm. We simulate two models and the results show that the algorithm can find the desired sequence of policies.

Contents

Chapter 1

Introduction

1.1 Introduction

COVID-19 is a contagious disease that a person can get if he becomes exposed to the new severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). Initially, the virus appeared in Wuhan, China at the end of 2019. In the next months, it spreaded throughout the world and in March 2020, the World Health Organization (WHO) declared it a pandemic. After the outbreak of the Covid-19 pandemic, everyday life does not remind at all the one that it used to be during the previous years. The pandemic has directly affected public health worldwide. All over the world, healthcare systems struggled with an unknown opponent and, on most occasions at least, proved to be unprepared for such a battle. Additionally, the global economy is suffering another big crisis and at the same time, coronavirus has affected a lot of aspects of humans life such as social relationships, education and politics. Thus, the Covid-19 pandemic has become one of the most important challenges for humanity to overcome at the moment.

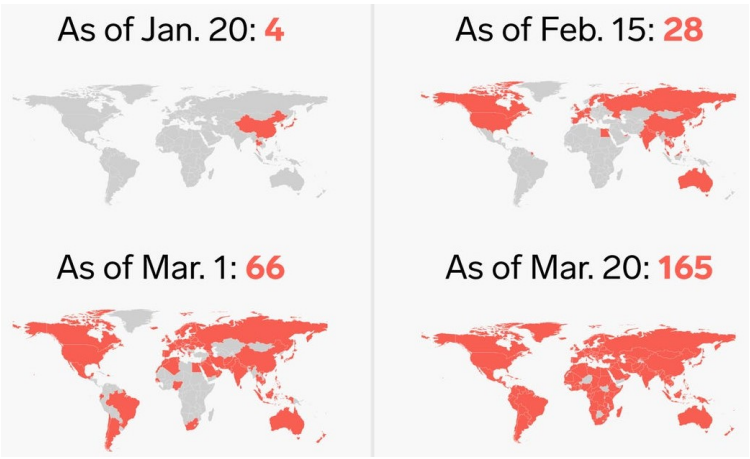

Figure 1.1: Spread of Covid-19 disease until end of March 2020. Red numbers specify number of countries with confirmed Covid-19 cases until January 20th, February 15th, March 1st and March 20th [2].

All over the world, governments enacted a variety of new laws and adopted

different policies and non-pharmaceutical interventions for controlling the spread of the disease. For instance, such policies are obligatory use of face masks in both indoor and outdoor places, international travel restrictions, social distancing, work from home, closure of schools and universities.

A significant asset to limit the coronavirus spread is the ability to forecast its epidemiological evolution and its economical effects. This can be achieved with mathematical simulation models of the pandemic dynamics, which emulate society. There is a variety of important factors to predict but in general the most vital ones are:

- Number of hospitalized people.
- Number of people in intense care units (ICUs).
- The numbers of other epidemiological compartments such as the number of infected and recovered people.
- Possible contacts/infections per day.

These mathematical simulation models of the pandemic dynamics can be easily extended to different societies by modifying the input parameters, and are able to simulate and study multiple epidemiological and lockdown scenarios. Political and health authorities can use such models as a guide to plan their actions and policies against Covid-19 pandemic. Motivated by this, this current thesis proposes an agent-based model together with reinforcement learning to understand the propagation of the Covid-19 disease and to provide policies to control it.

1.2 Thesis scope and content

The scope of this thesis is to provide a long term policy to control Covid-19 disease spread by using an Agent-based model (ABM) together with reinforcement learning (RL) methods when the dynamics of the disease model is not fully known.
This is planned to be achieved with reinforcement learning methods applied to ABM. The ABM simulates the disease dynamics using a society of agents, which emulate people, families, businesses, school and workplace environment. The epidemic model that is used to model the pandemic dynamics is a SEIHVR (Susceptible-Exposed-Infected-Hospitalized-Severe-Recovered) model, where the total population is equal to the sum of the people belonging to these health condition groups.

Six different policies of social distancing interventions are simulated:

- No restrictions.
- No restrictions with slightly lower contagion probability.

- Vertical isolation.
- Vertical isolation and obligatory use of face masks.
- Vertical and partial isolation, obligatory use of face masks and conditional lockdown.
- Partial lockdown.
- Total lockdown.

Some states must be considered constrained since the capacity of hospitals and intensive care units is limited. Therefore, a threshold is defined for the number of people in hospital and intensive care units (belonging at hospitalized and severe cases groups respectively). Objective of the reinforcement learning is to minimize the number of exposed and infected people by keeping the numbers of people in hospital and intensive care under the thresholds. Furthermore, the reinforcement learning algorithm must take under consideration the influence of the control strategies on the economy and pursue to achieve the less negative effects on economy.
In the next chapter, some related works that already have been done concerning control of epidemiological dynamics of other diseases are presented. In chapter 3, the general approach of the thesis is stated. The methods and the procedures that are followed for the design and the implementation of the system are documented in chapter 4 and 5 respectively. Chapter 6 presents the results of the testing simulations. Finally, in the last chapter, conclusions, future work and possible further steps on the project are discussed.

Chapter 2

Related Works

In this chapter, we present some existing researches conducted on mathematical modeling of pandemics in general, as well as modeling the Covid-19 pandemic. More precisely, we focus on infectious diseases models, agent-based model, agent-based simulation and policies to control the pandemic.

2.1 COVID-19 outbreak

As of today, COVID-19 is an ongoing global pandemic caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) [26]. The pandemic was publicly declared on 11 March 2020 however the first encounter with the virus took place in December 2019 in Wuhan, China. Since then, governments and scientists work tirelessly to control the rapid spread of the virus. Specialist started to come up with models of the pandemics, analyses of the dynamics and virus progression, ways of controlling and preventing the virus from the rapid spread [27]. In this thesis, we aim to focus on the control of the outbreak. Initially, we will focus on the simulation of the epidemic model with proper parameters compatible with real-world conditions and control of epidemics by using reinforcement learning.

2.2 Infectious diseases models

The purpose of modeling infectious diseases in form of mathematical models is to get a better understanding of how the disease propagates and how it reacts to different inputs. It also plays a role in the development and testing of different strategies for control.

To discuss about infectious diseases models and later about agent-based models the term model needs to be explicated. Models are abstract representations of real-world processes or objects. They are abstract since no model perfectly matches with the real-world conditions. Mathematical models can be presented in various ways such as differential equations [11]. Over the years, a wide variety of models have been developed, from simple to more complicated ones. They were developed to mimic the behaviours and propagation of the diseases. Infectious disease models can be formulated by using differential equations and proper parameters compatible with real-world data.

One of the basic infectious diseases model is SIR (Susceptible Infected Recovered) model [11] dating from the early 20th century. It consists of three compartments:

- S - Susceptible, number of susceptible individuals. These are liable to get a disease when exposed to a contact with infected individual.
- I - Infected, number of infected people. These can infect susceptible group.
- R - Recovered or removed, number of recovered/removed. This group underwent the infection and recovered from it or died.

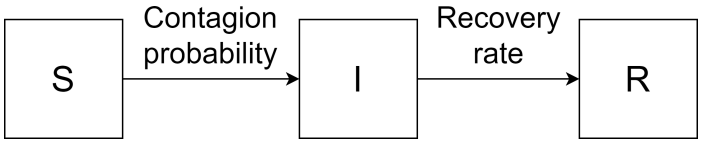

Figure 2.1: Transfer diagram for the simplest SIR model.

The number of people in each group changes in the unit of time, thus the groups can be represented as a function of time. People change states with assumed transition rates.

The classic SIR epidemic model is given by:

$$\begin{aligned} \frac{dS}{dt} &= -\frac{\beta IS}{N} \\ \frac{dI}{dt} &= \frac{\beta IS}{N} - \gamma I \\ \frac{dR}{dt} &= \gamma I \end{aligned} \tag{2.1}$$

where:

- β is an average number of contact sufficient for transmission of the infection,
- γ represents the transfer rate from infected to recovered.

Moreover, the time between contact can be stated as $T_c = \beta^{-1}$ and time to recover can be stated as $T_r = \gamma^{-1}$.

The population is given by:

$$S(t) + I(t) + R(t) = N \tag{2.2}$$

Together these two parameters divided are called basic reproduction number (also called basic reproductive ratio or basic reproductive rate):

$$R_0 = \frac{T_r}{T_c} = \frac{\beta}{\gamma} \tag{2.3}$$

This parameter represents the number of secondary infections in the population caused by one initial infection.

The model was used in simulation of Covid-19 [3] however due to the complexity of the Covid-19 disease, it is not the best choice.

According to research [1] median incubation time for Covid-19 is estimated at around 5 to 6 days. This means that after having contact with an infected individual, a susceptible person before developing an infection, need to go through an exposed state (incubation time). The SIR model can be extended to SEIR model by considering the incubation time of the disease. It is a much-preferred choice used in previous works [23]. SEIR is obtained from SIR model by using Exposed state. This exposed state can be understood as a delay in the system.

According to data [15] the transition of the coronavirus occurs through proximity in human to human interactions (1,5 meters), through droplets containing the virus, or through viral particles that float in the air which may be inhaled into the lungs. These factors can be modelled in a simulation as contagion distance and probability of contagion and will control the transition between Susceptible and Exposed states. The transitions from Infected group to Recovered group happens after a time called recovery time. The recovery time varies from 10 days onward [13]. Nevertheless SEIR model gives a good representation of Covid-19 dynamics, it is still far from complete. This model can be extended by adding more states, bringing it ever so slightly closer to to real world stages of the disease.

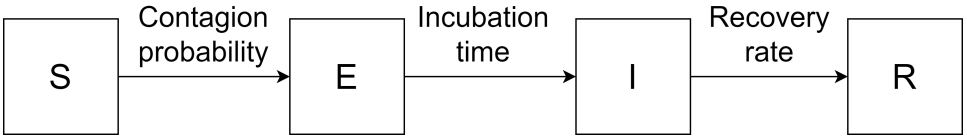

Figure 2.2: Transfer diagram for the SEIR model.

The SEIR model is given by:

$$\begin{aligned} \frac{dS}{dt} &= -\frac{\beta IS}{N} \\ \frac{dE}{dt} &= \frac{\beta IS}{N} - \alpha E \\ \frac{dI}{dt} &= \alpha E - \gamma I \\ \frac{dR}{dt} &= \gamma I \end{aligned} \tag{2.4}$$

where:

- β is an average number of contact sufficient for transmission of the infection,
- γ represents the transfer rate from infected to recovered,

- α is an incubation time.

Another research [9] uses a model which contain 8 states namely susceptible (S), infected (I), diagnosed (D), ailing (A), recognized (R), threatened (T), healed (H) and extinct (E). The model differentiates between severity of the infected cases namely non-life-threatening and potentially life threatening and also between detected and undetected cases. It also distinguishes between symptomatic and asymptomatic cases.

The simulation of the model is based on data from Italy and shows how the situation will progress and what could have happened if several different measures had been implemented earlier. It shows the benefits of the stricter lockdown and other measures such as contact tracing.

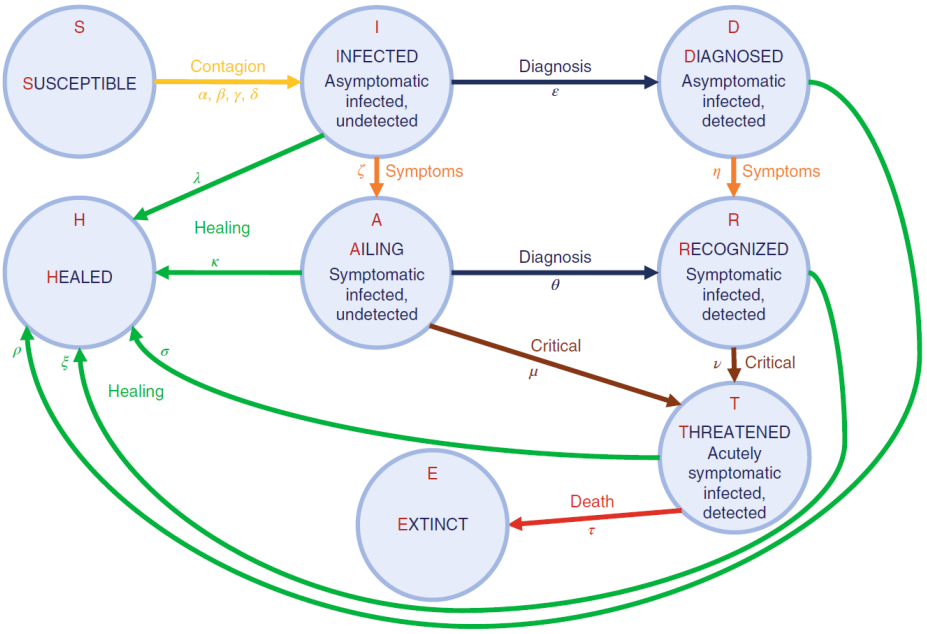

Figure 2.3: Graphical scheme representing all states in SIDARTHE mathematical model. Source [9]

There are many more details and parameters that can be added to these models. Vital dynamics (births and deaths), passive immunity's, vaccination programs and so on. All these factors will influence the simulation time and accuracy of the model.

2.3 Agent-based modeling and simulation

Agent-Based Models nowadays are widely used as a powerful and convenient modeling tool. The concept of ABM involves self-conscious (to some extent) agents,

which can perform actions on the environment. The concept development started around the 1970s. Agent-Based Simulation (ABS) can be described as micro-scale models with agents which are performing actions and interactions with each other and environment over a period of time to imitate some real world processes. From ABS one can observe the dynamics of the system. There can be several various scenarios and factors applied to the simulation to observe a result. ABMs are used in a variety of different disciplines like biology, business, technology or epidemics [4] [7].

In biomedical, the ABM is used to study how tissue patterns develop as a result of cellular interactions [5]. It proved to be useful giving the researches an insight about complex cellular activity related to cancer, immunology and others.

The ABM concept was also proven exploitable in understating the spread of diseases [22] [16]. It allows to build various models of environments with autonomous agents (humans) with set of rules such as how the agents interacts and others.

Recent studies which aim to find a way to end the Covid-19 pandemics are based on ABMs [23]. They explore and analyze several scenarios to find the strategy to dispose of the virus.

The results indicate, that to control the pandemics, 90% of the population needs to be isolated, along with travel restriction and case isolation.

Connecting the ABM with Infectious Diseases models can give an insight into how the pandemics progresses and gives a powerful tool that can be used to find a way to control the spread.

2.4 Infection spread control

The control of the disease is a great challenge. In the past, with less possibilities to travel and different people's approach, it was easier to take the pandemics under control. Nowadays with possibilities to travel around the globe it is effortless to spread the virus in no time, causing global pandemics. Also with great availability of information and disinformation it is hard to control the pandemics, making them quite hard to dispose.

The first hints of variolation (one of the method to immunize) against the contagious disease called smallpox are dated to 10th century [10]. The method was a bit different from what we know today, insufflation techniques were used. The technique involves blowing the immunological material into a body cavity. In 18th century first vaccines were developed with success, making them safer and so far the best tool against contagious diseases. Vaccines contributed to disposing a lot of diseases such as diphtheria, tetanus, pertussis, influenza and measles [21]. However, the development of the vaccines is time consuming process especially for a completely new diseases like Covid-19. Before the vaccination program can start

saving lives, there are different ways to limit the propagation of the virus.

Since Covid-19 is a disease which spread very easily through droplets and small particles floating in the air or settled on surfaces, a number of regulations were introduced to counteract. Several recommendations were proven to work against the spread namely face-masks [14], disinfection [24] and social distance [6]. Several restrictions were so far proven to have a positive impact on infection curve. Lockdown introduced all over the world become an efficient way to keep the severe cases below capacity of medical care [8]. Since most of the life-threatening symptoms occur in patients in higher age groups, isolation of these groups has a great impact on reducing the effects of the disease on population.

Nowadays, people are able to simulate the progression of the real world processes with great accuracy and based on conclusions, take appropriate actions in significantly faster time. These simulations can be also used to dispose a pandemics. Reinforcement learning techniques help people to control complicated, stochastic processes.

Researchers have used deep reinforcement learning technique to seek for a sequence of policies to control the Covid-19 spread with a great success [19].

In the present thesis, we will focus on the Q-learning technique, to find an optimal set of policies that will keep the hospitalized and severe cases below the health care limit.

Chapter 3

Project Approach

In this project, we aim to use reinforcement learning as a way to control the Covid-19 pandemics by keeping the numbers of people in hospital and intensive care below the capacity.

Initially, we develop an ABM, to mirror the real world into a smaller and simpler model. The model will contain an environment composed of houses, workplaces, school, hospital and free space. The environment is grid-based i.e., the area is fixed to a 2-dimensional x by y grid. It will also contain a fixed number of agents.

Each agent will perform fixed actions (go to work, go to school, go home) and random actions (movement in the environment). The goal is to mimic a real-life environment in which agents interact with each other to simulate the progression of the pandemics. Each agent will be assigned to one of six groups at all times namely susceptible (S), exposed (E), infected (I), hospitalized (H), severe (V) or recovered (R). Agents will change groups according to transition probabilities. Spread of the infection will be possible with contagion probability when two agents will find themselves on the same cell on the grid.

There will be several policies to follow by the agents. Each policy will introduce a layer of restrictions. Restrictions chosen for the policies are based on previous research [23].

The task for reinforcement learning controller is to learn, which policies should be applied to the simulation to keep the hospitalized and severe cases within the capacity.

Moreover, alongside the model described above which involves a lot of details and operations, two simpler, illustrative models will be programmed and simulated. The purpose of them is to evaluate the correctness of the reinforcement learning technique. Reducing complexity and having knowledge and control over processes that are taking place in the environment of illustrative models, the correctness of reinforcement learning can be easily confirmed or denied. Having two illustrative models, one with a larger environment and a larger number of agents will confirm if the reinforcement learning technique can be used, regardless of the complexity of the model.

Chapter 4

Methods

In this chapter, the methods used in this project are presented and described. These are Markov Decision Process and Reinforcement Learning technique called Q-learning.

4.1 Markov decision process

Reinforcement learning problems can be formulated with assistance of Markov decision processes (MDPs). There are two different cases of MDPs. First one have deterministic state transitions and second one have stochastic state transitions. In general, a MDP contains:

- A set of possible states S.
- A set of possible actions A.
- A real-valued reward function R(S_t,A_t).

In the MDP setup, the environment's response at time $t + 1$ depends only on the state and action at time t and it is independent of whatever happened in the past.

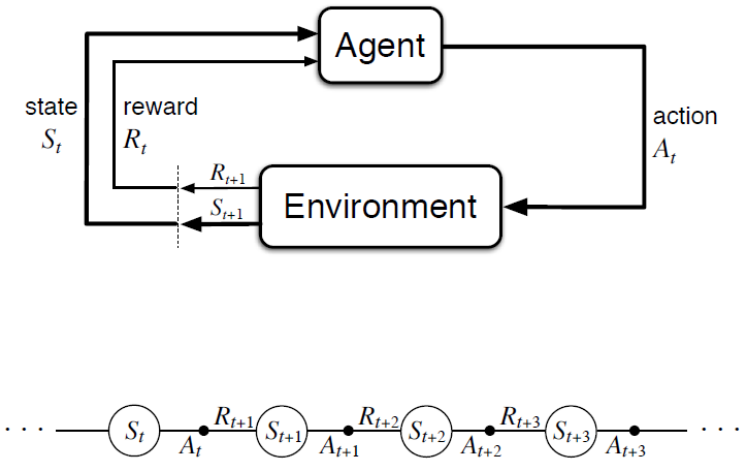

Figure 4.1: Graphical illustration of time steps of a MDP environment. Agents receive a reward R_{t+1} and end up in state S_{t+1} based on the action A_t at a particular state S_t Source: [18].

The goal of the agents is to maximise the cumulative reward that they receive in total. Total reward at any time instant t is given by the next equation:

$$Total\ Reward = R_{t+1} + R_{t+2} + R_{t+3} + ... + R_T \tag{4.1}$$

Where T is the final time step. In Equation 4.1, it can be observed that all future rewards have equal weights. This fact might not be desirable, therefore an additional concept of discounting arises. A discount factor γ is defined and each reward after the immediate reward is discounted by this factor as follows:

$$Total\ Reward = R_{t+1} + \gamma R_{t+2} + ... + \gamma^2 R_{t+3} + ... = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \tag{4.2}$$

Deterministic Markov Decision Processes

A deterministic MDP is defined by the state space of the process, the action space of the controller, the transition function, which describes how the state changes as a result of control actions, and the reward function, which evaluates the immediate control performance. As a result of an action A_t applied in the state S_t at the discrete time step t, the state changes to S_{t+1}, according to the transition function $f : SxA \rightarrow S$:

$$S_{t+1} = f(S_t, A_t) \tag{4.3}$$

At the same time, the controller obtains a reward R_{t+1}, according to the reward function $R : SxA \rightarrow R$:

$$R_{t+1} = R(S_t, A_t) \tag{4.4}$$

The reward evaluates the immediate effect of an action, or more specifically the transition from a state S_t to another state S_{t+1}. However it does not say anything about long-term effects of this action. Actions are selected according to a policy $h : S \to A$, using:

$$A_t = h(S_t) \tag{4.5}$$

If the transition and reward function are available, the current state S_t and the current action A_t are sufficient to determine both the next state S_{t+1} and the reward R_{t+1}. This is called the Markov property, which is essential in providing theoretical guarantees about reinforcement learning algorithms [17].

Stochastic Markov Decision Processes

In stochastic MDP, the next state is not deterministically given by the current state and action. In contrast, the next state is a random variable, and the current state and action give the probability density of this random variable.

Rewards are associated with transitions, and in stochastic MDP transitions are not fully determined by the current state and action. Thus, the reward function also has to depend on the next state. After a transition to a random state S_{t+1}, a reward R_{t+1} is derived according to:

$$R_{t+1} = R(S_t, A_t, S_{t+1}) \tag{4.6}$$

Note that Equation 4.6 is a deterministic equation, which means that, once S_{t+1} has been generated, the reward corresponding to the transition reward R_{t+1} is fully determined. In the stochastic case, the Markov property requires that the state S_t and the action A_t fully determine the probability density of the next state S_{t+1}.

4.2 Dynamic programming and reinforcement learning

To control the pandemics, dynamic programming (DP) and reinforcement learning (RL) techniques are considered. DP methods require a model of a system. They work offline to produce the optimal policy for a given process. RL alongside supervised and unsupervised machine learning form three basic machine learning (ML) paradigms. In comparison to ML methods, RL does not require to be fed with optimal inputs and outputs. It also does not require a model in contrast to DP. Instead, it discovers and learns by taking actions and receiving rewards. The RL can be divided into two categories - online and offline. Offline RL uses data given in advance. Online RL, seeks an optimal set of actions relying on data collected as the process run. For DP and RL, the goal is to find a set of actions that

maximize the reward. The environment is usually presented in form of the MDP. The advantage of RL over DP is that RL methods can work on huge (a great number of states), stochastic and even not fully known processes, whereas DP methods always require process model.

Online RL requires a balance between exploration (trying random or not optimal actions) and exploitation (using the current knowledge, taking currently optimal actions). This balance can be controlled by the ϵ-greedy algorithm [17] where with probability ϵ, the RL algorithm chooses to explore and with probability $1 - \epsilon$ - exploit.

4.2.1 Model-free value iteration

The method of RL which fits this project is a model-free value iteration method namely Q-learning [17]. The method was introduced by Chris Watkins in 1989 and later proven by Watkins and Dayan in 1992. Q-learning can help to find the best policy (set of actions) for a system that does not change in finite time. Q-learning depends on Q-table which stores the quality values and an updated formula. The update formula is stated as:

$$Q^{NEW}(s_t, a_t) = (1 - \alpha)Q(s_t, a_t) + \alpha(r_t + \gamma max_a Q(s_{t+1}, a)) \tag{4.7}$$

where:

- s_t is a state at time t,
- a_t is an action taken at time t,
- $Q^{NEW}(s_t, a_t)$ is the new Q-value,
- $Q(s_t, a_t)$ is a current Q-value for state s_t at time t and action a_t a time t,
- $max_a Q(s_{t+1}, a)$ is the maximum Q-value over all possible actions in next state $(s_t + 1)$,
- α is the learning rate,
- r_t is the reward,
- γ is the discount factor.

The equation can be divided into two parts. First part $(1 - \alpha)Q(s_t, a_t)$ is the old information, the Q-value value obtained so far, scaled by the learning rate. Second part $\alpha(r_t + \gamma max_a Q(s_{t+1}, a))$ is the new information. The learning rate defines how much the new information should override the old one. Choosing it as 0 will mean that the algorithm will only take the old information. This is not the desired situation as the algorithm will not learn anything from taken actions. Choosing

the learning rate as 1 will mean that the old information will be completely lost at every iteration. The choice of this value is arbitrary and depends on the system to which the reinforcement learning is implemented.

The discount factor determines how the algorithm will consider future rewards. The stopping criteria for Q-learning is:

$$Q^{NEW}(s_t, a_t) = Q(s_t, a_t) \tag{4.8}$$

Meaning that the Q^{NEW} is the same as Q-value in the Q-table, the Q value will not change. If the discount factor is chosen to 0, the Q-values will naturally converge to rewards since:

$$Q^{NEW}(s_t, a_t) = (1 - \alpha)Q(s_t, a_t) + \alpha r_t \tag{4.9}$$

$$Q^{NEW}(s_t, a_t) = Q(s_t, a_t) - \alpha Q(s_t, a_t) + \alpha r_t \tag{4.10}$$

$$Q^{NEW}(s_t, a_t) - Q(s_t, a_t) = \alpha(r_t - Q(s_t, a_t)) \tag{4.11}$$

$$Q^{NEW}(s_t, a_t) = r_t \tag{4.12}$$

If the discount factor is chosen to be other than 0, then the algorithm will take into account future rewards and converge to:

$$Q^{NEW}(s_t, a_t) = r_t + \gamma max_a Q(s_t, a) \tag{4.13}$$

To fill out the Q-table, the algorithm needs to perform the learning part. During this part, the algorithm explores the available actions, calculates and overwrites the Q-values. After learning part, the optimal policy can be found by searching for the maximum values in the Q-table.

$$\pi^*(s_t) = argmax Q(s_t, a_t) \tag{4.14}$$

where $\pi^*(s_t)$ denote the optimal policy for state s at time t

Chapter 5

System Design

This chapter presents comprehensive descriptions of all parts which assemble the whole simulation. First, the Agent-based model with all its elements is characterized namely agents with their properties and parameters, grid environment with delineated "places" (houses, workplaces, school, hospital) and processes that are taking place in the simulation such as economical and epidemiological processes. Next, the design of the control part is derived. It contains a description of the reinforcement learning technique used in the project with extensive elaboration on the principle of operation. Afterwards, a smaller model is described. The purpose and design of it are discussed. Lastly, the implementation of the simulation is described in details, such as development of environment and composition of the program.

5.1 Agent-based model

An ABM belongs to a class of models called computational models. It implies the use of autonomous agents, which interact with the environment and each other. The outcomes of such action can be observed and conclusions can be obtained.

To simulate a spread of infectious disease like Covid-19, ABM is a preferred choice due to its flexibility of implementation. The simulation presented in this thesis shows not only pandemic effects on society, but also the financial effects caused from the measures used to control the outbreak. There are two types of agents. One of them are people living in a simulated society. They interact with each other on epidemiological and economical levels. The disease can spread as the agents perform stochastic actions. Second type of agents are agents which interact only on economical level. These agents are:

- Houses
- Workplaces
- Hospital
- Government

As mentioned above, these agents interact with others (also humans) but also as they are literally structures in the simulated environment (excluding government).

There is one more structure namely school, which does not interact on any level with other agents. Another important property is that houses, workplaces, school and hospital accommodate humans inside. More elaboration on that can be found in 5.1.2.

The simulation is performed in a normal day cycle meaning there are 24 hours in a day, each hour is divided into number of timestamps in which agents perform their tasks. Number of timestamps are dependent on the tasks performed in a given hour. The distribution of the tasks in the day depends on the policy applied to the simulation. Detailed descriptions of such policies can be found in 5.5.1.

5.1.1 Humans

The total number of people in the simulated society is 400. Parameters are assigned to them when they are created. These parameters are:

Parameter	Description
Epidemiological compartment	Each human can be assigned to one of the SEIHVR compartments (more details can be found in section 5.2)
Incubation time	Time needed for the disease to incubate
Recovery time	Time needed for an infected person to recover
Age	The age of the individual
Social stratum	The economical condition of an agent
Personal wealth	Money that belong to a human agent
Personal income	Salary that a human agent earns from his workplace
Personal expenses	Money that a human agent spends during a month
Homelessness	Homeless or not
Employment status	Employed or not
Educational status	Student or not
Essential employee	Means that this worker is crucial to maintain proper life course
House	House which agent is assigned to
Workplace	Workplace which agent is assigned to

Table 5.1: Parameters of human agents.

Furthermore, each agent has x and y coordinates assigned, which represent the current position of the agent in the environment. Moreover, each agent has parameters dx and dy which represent the "distance" (in cells) that the people can cover in each timestamp iteration. These parameters defines the location of the individuals in the environment and enables them to move inside it.

The age of the people is a random variable distributed with normal distribution $N(42, 28)$. Notation $N(\mu, \sigma^2)$ represents normal distribution of a random variable X is with mean μ and standard deviation σ. Mean of this distribution is 42 which is the average population age in Denmark [25]. Standard deviation is set equal to 28 after tuning it, so that the percentages of age population groups of people under 17 years old, people between 17 and 60 years old and people over 60 years old in the ABM, correspond to the real percentages in Denmark [25]. Each person has a probability 0.002 of being homeless [12]. People that are younger than 17 years old are considered to be students, while people older than 66 years old are considered to be retired from working. Every person between 17 and 66 years old has a probability to be employed in one of the workplaces. This probability is equal to 0.9515 (Unemployment in Denmark in 2021 = 4,85% [20]). Moreover, the probability that a person is an essential employee is equal to 0.2. This means that in one of the lockdown policies described in Table 5.8 this person will go to his work while non-essential workers will stay at home. The initial distribution of humans into epidemiological groups as well as initialization of incubation and recovery time are described in the Section 5.2. The economical aspects are described in the Section 5.1.3. Humans have a set of actions that they perform during the simulation. They are described in Table 5.2.

Action	Description
Go home	The person moves one cell towards his home. If his home is 200 cells or above from him, he goes there immediately (in real life by car)
Stay home	The person stays at his home and move inside it with probability 0.4
Go to work	The person moves one cell towards his workplace. If his workplace is 200 cells or above from him, he goes there immediately
Stay at work	The person stays at the workplace and move between the cells inside it with probability 0.4
Stay at work 2	The person stays at the workplace and move between the cells inside it with probability 0.05
Walk free	The person walks free in the environment. There is a 0.25 probability to move either left, right, up or down.
Go to school	The person with status student moves one cell towards school. If school is 200 cells or above from him, he goes there immediately.
Stay at school	The person with status student stays at school and move inside it with probability 0.4

Table 5.2: Actions performed by the human agents.

5.1.2 Grid environment

The ABS takes place in a grid-like environment. The environment size is fixed and its shape is defined as a rectangle. The cells in the grid relate to 1,5m by 1,5m squares in the real world within which, agents can be infected or get infected by other agents.

There are 5 different structures in the environment namely houses, workplaces, school, hospital and outside.

Houses

Each agent is assigned to a house. The distribution of agents into houses depends on their age and house capacities. House capacities are uniformly distributed $\mathcal{U}(1,3)$. Notation $U(a,b)$ represents uniform distribution of a random variable X between a and b which are the minimum and maximum values. Each agent with an age above 45 is assigned to a house together with other agents above 45. Other houses are occupied by agents with age below and equal to 45. The total number

of houses is equal to the quotient of the total population and the average family size which is equal to 3 members.

The variables that are defined for each house are:

- Coordinates x and y indicate the centre cell of each house
- Variables dx and dy indicate for how many cells, each house agent expands to the left, right, up and down in x and y axes of the environment. The values are uniformly distributed $\mathcal{U}(1,3)$.
- Number of residents of the house (family members). The value is uniformly distributed $\mathcal{U}(1,3)$.
- Social stratum of the house
- House wealth

The people that are assigned to a house are not homeless, are assigned to only one house and belong to the same social stratum as the house.

Workplaces

A total number of businesses is defined as 160 but not all of them have assigned workers from simulated society. People are assigned to workplaces with uniform distribution $\mathcal{U}(3,9)$. This means, that around 30 to 100 workplaces are occupied (this range depends on how many employees are in the simulated society and on the distribution of them to workplaces). The other workplaces remain empty and serve a purpose for the economy which is described in Section 5.1.3 in "Economical Interactions".

The variables that are defined for each workplace are:

- Coordinates x and y which indicate the centre cell of each workplace
- Variables dx and dy which indicate for how many cells, each workplace agent expands to the left, right, up and down in x and y axes of the environment. The value is uniformly distributed $\mathcal{U}(1,7)$.
- Number of employees of the workplace, which is uniformly distributed $\mathcal{U}(3,9)$.
- Social stratum of the workplace
- Workplace wealth

The people that are assigned to a business, are not unemployed, are not assigned to any other business and belong to the same social stratum as the business.

School

A single school is defined outside of the environment. The school also has fixed grid-like dimensions defined by the centre x and y coordinates and dx and dy which indicates for how many cells, each school expands to the left, right, up and down in x and y axes. Apart from its size, the school also has the number of students as a parameter.

Hospital

Hospital is defined outside of the environment and it is simplified compared to houses and workplaces. It has only position variables x and y. When a person needs to be hospitalized, he is transferred to a pixel outside from the main simulation environment. Critical limits for the hospitalized people are 2% of the total population and for people in intensive care units is 1% of the total population. This means that in a population of 400 people, the constraint for hospitalized people is 8 and for people in intensive care is 4.

Government

Government is an economical agent. Its purpose is to manage the money from taxes and spend it on people and hospital. The government is not characterized by position variables. More about it can be read in Section 5.1.3.

5.1.3 Economy

As mentioned earlier, the ABM takes under consideration also the effects of the pandemic on the economy. This is another important effect of the pandemic which should have a place in the analysis since implementing strict lockdown policies, may lead to lower infected cases and better epidemiological results, however, affect the economy with the most negative way. Therefore, it is vital that the simulation cover the effects of the applied policies on the economy. Then, the objective should be to implement policies which result to the minimum spread of the disease in parallel with minimum financial loss.
In order to accomplish this challenge, economical parameters are assigned to the agents. During the simulation the agents interact each other financially as well. More details on these interactions and the general economical approach of the model are explained below.

Initialization Process

As a first step, the total wealth is defined. 90% of the total wealth corresponds to the government and the rest 10% corresponds to the rest agents. This amount is

shared to the agents as follows:

- 50% to businesses.
- 40% to people.
- 10% to houses.

The wealth that corresponds to each agent depends on its social stratum and on how many agents have the same social stratum. Social stratum values vary from one to five and represent the following financial statuses:

- 1 corresponds to not wealthy.
- 2 corresponds to less wealthy.
- 3 corresponds to average class.
- 4 corresponds to wealthy.
- 5 corresponds to very wealthy.

To make the above information more clear, the initial wealth of a random business will be given by:

$$Initial\ Wealth = \frac{(Total\ Businesses\ Wealth * Social\ Stratum\ Ratio)}{Number\ of\ Businesses\ with\ same\ social\ stratum} \tag{5.1}$$

Social stratum ratios are given in [23], and namely they are:

- 3.62 % of the total wealth of the respective agents group corresponds to not wealthy agents.
- 7.88 % of the total wealth of the respective agents group corresponds to less wealthy agents.
- 12.62 % of the total wealth of the respective agents group corresponds to average class agents.
- 19.71 % of the total wealth of the respective agents group corresponds to wealthy agents.
- 56.12 % of the total wealth of the respective agents group corresponds to very wealthy agents.

Furthermore, some assumptions are made so that the model becomes more realistic. In a house with a specific social stratum, only people with the same social stratum are assigned. Thus, in a wealthy house, reside only wealthy people. The

same assumption is considered for businesses. In a wealthy business, work only wealthy people. Also, all homeless people have social stratum of not wealthy.

Moreover, each person has personal income and personal expenses. Their values depend on their social stratum. Bigger social stratum means bigger income and more expenses. The income of the people is considered to be their salary. Concerning unemployed people, who have no salary, they receive a monthly aid from the government, while homeless people have zero income and zero expenses. People's income and expenses are summed up at the Table 5.3:

People	Income (money units)	Expenses (money units)
Homeless	0	0
Unemployed	400	100
Not Wealthy	900	600
Less Wealthy	950	650
Average Class	1200	900
Wealthy	1500	1200
Very Wealthy	2000	1700

Table 5.3: People income and expenses.

To summarize, at the end of the initialization process, every agent (person, house, business) should have an initial wealth. Moreover, personal income and expenses for each person are defined. During the simulation, there are economical interactions between the agents, which depend, at some point, on the applied policy.

Economical Interactions

Economical interactions between agents are divided in daily and monthly interactions.

Daily Interactions

Every day people are going for shopping during their free time to random businesses. They spend there an amount of their personal expenses (this amount is assumed equal to $\frac{1}{60}th$ of their personal expenses for each time they go shopping). The money is also deducted from the person's house wealth. This is the main income for businesses. Apparently, the more people that are moving freely outside, the more money, businesses are earning. In scenarios where people move freely, market presents more intensive movement, while in strict lockdown scenarios, people don't move freely outside and businesses are considered closed for

clients, therefore market movement is decreased significantly.

Monthly Interactions

At the end of each month there is also a series of economical interactions between the agents.

- Businesses:
  They pay salaries to their employees and bills and taxes to the government.
- Houses:
  At the end of each month, houses pay bills and taxes to the government. Furthermore, each house pays some constant amount of money to a random business for monthly supplies. To the house wealth are added the salaries of its residents. This amount is the houses' income.
- People:
  People are earning their salary (aid from the government if they are unemployed) at the end of each month.

In Figure 5.1 the economical interactions between agents are presented graphically. The continuous red lines indicate direct transfer of wealth from one agent to another. The discontinuous red lines represent indirect gain or loss of wealth. Particularly, human agents spend some money every day for shopping. This wealth is transferred directly from people's wealth to businesses' wealth. Same wealth is also removed from each person's house wealth. Therefore houses' wealth is reduced, but there is no direct wealth transfer between house and business agents in this case. Similarly, the salaries of the residents of each house are added also to the house's wealth.

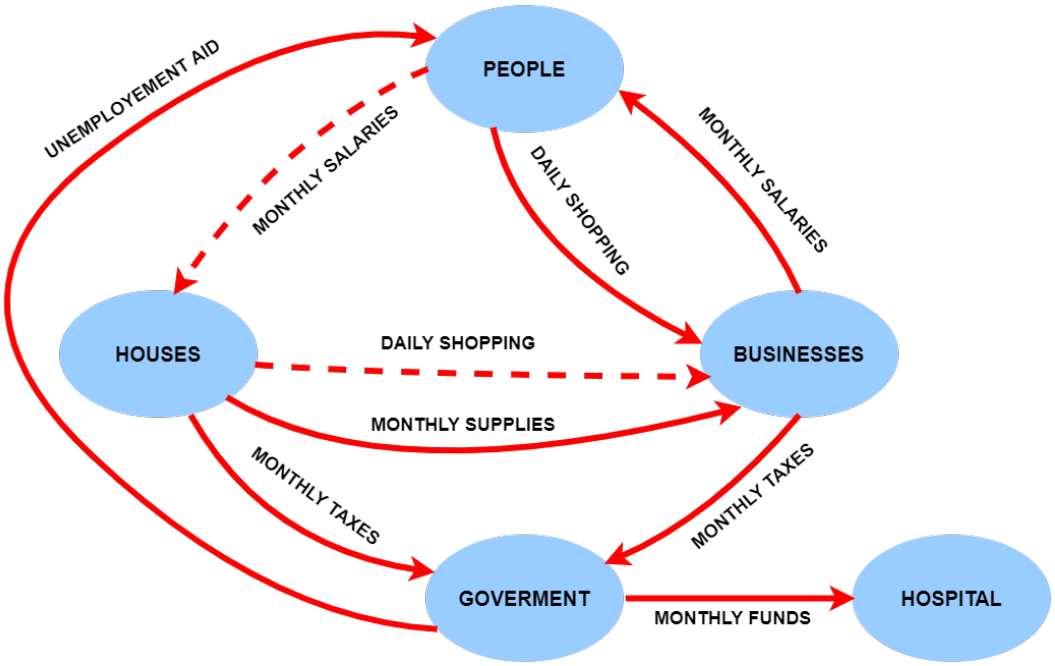

Figure 5.1: Illustration of the daily and monthly economical interactions between the agents.

5.2 SEIHVR Epidemic model

5.2.1 SEIHVR model

The model used in this project adds another two states namely Hospitalized (H) and Severe (V) to SEIR model. H indicates the number of individuals who are infected and hospitalized. V indicates the number of individuals who need the help of the ventilator. The model is given by:

$$\begin{aligned}
\frac{dS}{dt} &= -\frac{\beta IS}{N} \\
\frac{dE}{dt} &= \frac{\beta IS}{N} - \alpha E \\
\frac{dI}{dt} &= \alpha E - (\gamma I + \eta I) \\
\frac{dH}{dt} &= \eta I - (\sigma H + \zeta H) \\
\frac{dV}{dt} &= \zeta H - \mu V \\
\frac{dR}{dt} &= \gamma I + \sigma H + \mu V
\end{aligned} \tag{5.2}$$

where:

β - is an average number of contact sufficient for transmission of the infection.
γ - is a recovery time for infected group
α - is a normally distributed incubation time $\mathcal{N}(7,3)$.
η - represents the transfer of infected to hospitalized group
σ - is a recovery time for hospitalized group
ζ - represents the transfer of hospitalized infected to severe group
μ - is a recovery time for severe group
The population is given by:

$$S(t) + E(t) + I(t) + H(t) + V(t) + R(t) = N \tag{5.3}$$

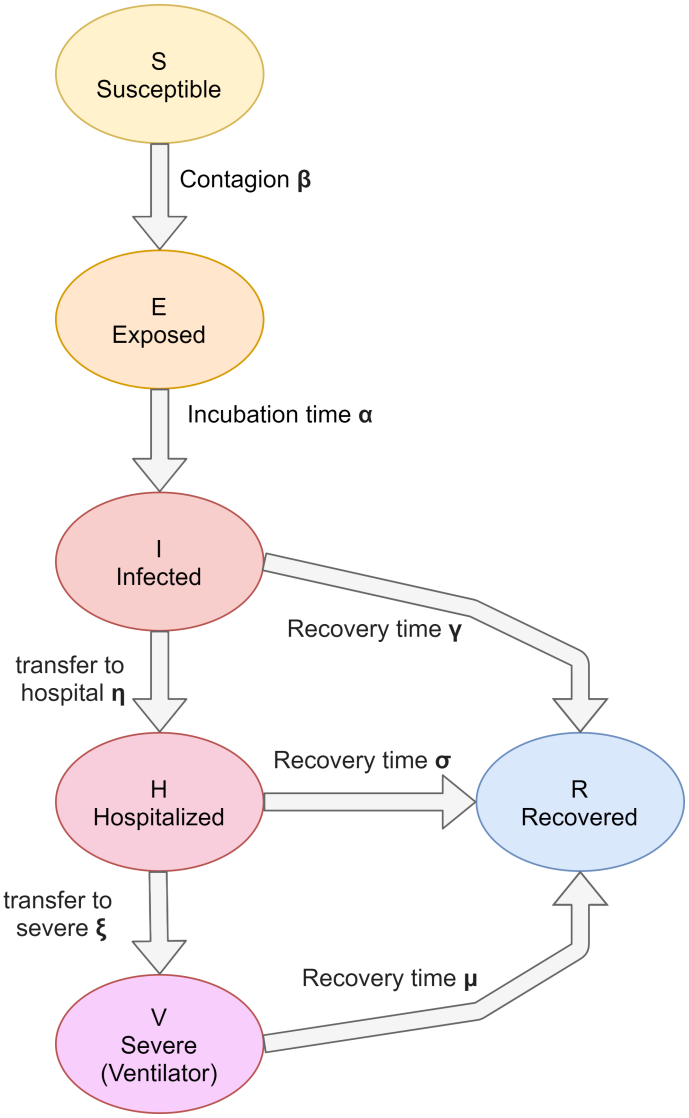

Figure 5.2: Diagram representing the transitions between the states of SEIHVR epidemic model.

Initially, 1% of the simulated population (4 people for the total population of 400) is set to belong to epidemiological group Infected. The rest of the agents are initialized in the Susceptible group.

Incubation time is chosen with normal distribution $\mathcal{N}(7,3)$ for each human. The incubation time cannot be less than 3 days.

Recovery time is dependent on the age of the agent. If the agent's age is less than 40, the recovery time is set to 14 days. If the agent's age is above or equal 60, the recovery time set for this agent is 20 days. Otherwise (between 40 and 60) the recovery time is set to 17 days.

The vital dynamics has been omitted due to short time scale of the simulation. It is assumed that births and deaths will have a very limited influence on the pandemics progression.

Contagion probabilities are set to 0.9 and 0.3 in policies with face masks [23].

At initialization of the environment and agents, every susceptible person has a probability to be immune. This probability is equal to 0.01. The initially immune individuals are automatically transferred to Recovered group. The agents from recovered group cannot get infection again.

The probability that a person needs to go to the hospital or to receive a ventilator in the hospital depends on age. The table 5.4 shows the estimates of the severity of the cases.

Age group (years)	Probability of infected requiring hospitalization	Probability of infected requiring ventilator
0 - 9	0.001	0.05
10 - 19	0.003	0.05
20 - 29	0.012	0.05
30 - 39	0.032	0.05
40 - 49	0.049	0.063
50 - 59	0.102	0.122
60 - 69	0.166	0.274
70 - 79	0.243	0.432
80 - 89	0.273	0.709

Table 5.4: Probabilities of the hospitalization and severity of cases. Source: [15]

5.3 Policies

Policies created for ABS are based on approaches applied all over the globe in confrontation with the Covid-19 pandemic. In this project, a few rules to prevent the virus from spreading are used to create a set of policies. Applying a specific

policy to the simulation, results in the difference between actions taken by the agents. Through these actions, it can be observed that dynamic of the pandemics changes. The actions also influence the economy. Table 5.5 describes a set of rules used to formulate policies. Table 5.6 presents 6 policies formulated from the set of rules.

Rule	Description
No restrictions	People, Houses, Business, School function as normal meaning there is no prevention against the spread applied.
Lockdown	All agents stay at homes. School, Workplaces remains empty.
Vertical isolation	Risk groups are isolated (stay at home)(Risk groups are people with age above and equal 60)(The school is also closed, students stay home). Business remains functioning as normal.
Face masks	Contagion probability reduced.
Conditional lockdown	Lockdown under certain condition.
Partial isolation	X% of the population are under lockdown, the rest of population are not (essential workers don't follow lockdown).
Quarantine	This option can be turned off and on. It works for every policy. Each infected individual goes to their home after 1 day of being infected.

Table 5.5: Rules used to formulate policies.

Policy	Description
Policy 0	No restrictions
Policy 1	Vertical isolation
Policy 2	Vertical isolation, Face masks
Policy 3	Conditional lockdown, Partial isolation
Policy 4	Lockdown, Partial isolation
Policy 5	Lockdown

Table 5.6: Policies used to control the pandemics in ABM simulation.

The implementation of the policies is described in section 5.8.

5.4 Reinforcement learning

The controlling strategy used in this project is online Reinforcement Learning, model-free value iteration. The method used is called Q-learning. The advantage which makes this method efficient for this project is that it can work without a model in a huge, stochastic environment.

5.4.1 Exploration - exploitation strategy

The balance between exploration and exploitation is set using ϵ-greedy strategy. The pseudo-code for the algorithm:

Algorithm 1 ϵ-greedy strategy

```
if x < e then
  take random action (uniform distribution U(0,1))
else
  take best possible action
end if
```

The probability to explore ϵ decrease exponentially.

5.4.2 Reward function

The reward function is an important part of the Q-learning algorithm. It dictates how the algorithm will converge, which behaviours and actions should be rewarded positively and which should be punished. Rewards assess the performance and efficiency of the algorithm.

The reward in the simulation is split into two main parts. Then, these main parts are split into few more parts. Two main parts are rewards for epidemic status and rewards for the economy. Each factor which makes the reward is weighted. Factors are chosen as linear or quadratic functions. This is done to keep a balance between each part of the reward and to attach a significance to each part. For example, the rewards for H and V states have a higher significance, as the goal for the algorithm is to keep the hospitalized and severe cases below the threshold.

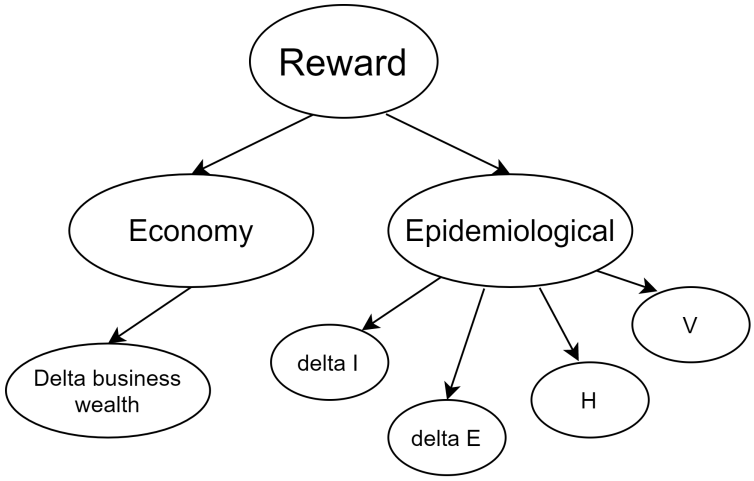

Figure 5.3: Construction of the reward.

The reward is computed at the end of the day (t). After that, Q-value for this day is computed. For the first day of the simulation, $(t-1)$ symbolizes the initial conditions. The reward function is defined as follows:

$$R(t) = FRR(t)\frac{DBW}{a} + (-b \cdot dI + (-c \cdot dE + (-d \cdot H(t)^2) + (-f \cdot V(t)^2) \quad (5.4)$$

where

- H(t) - hospitalized individuals at day t
- V(t) - severe individuals at day t
- I(t) - infected individuals at day t
- FRR(t) - financial reward rate at day t

$$FRR(t) = \begin{cases} 1 & \text{when } H(t) < 8, V(t) < 4 \text{ and } I(t) < 200 \\ 0.5 & \text{when } H(t) < 8, V(t) < 4 \text{ and } I(t) \geq 200 \\ 0 & \text{when } H(t) \geq 8 \text{ and } V(t) \geq 4 \end{cases}$$

- DBW - delta business wealth - difference in the wealth of the workplaces between current and previous day
- $dI = I(t) - I(t-1)$ - delta I - difference in number of infected between current and previous day
- $dE = E(t) - E(t-1)$- delta E - Difference in number of exposed between current and previous day

- a, b, c, d, f - tuning parameters

All parameters a, b, c, d, f need to be tuned to find a proper balance in the reward function. Situation when one part of the reward significantly outranks the others is not desirable. To tune it properly, a few test runs are needed.

The reason for choosing such components for the reward is to lower the infection peak, keep the hospitalized and severe cases below the threshold and, at the same time, keep the economy unchanged. Without adding a reward for the change of Exposed (contact tracking), the algorithm cannot react properly and learn that choosing a certain policy, leads to consequences of higher infection peak in the future.

To find the sequence of policies to follow each day, the algorithm needs to go through the learning phase. In this phase, the algorithm is encouraged by high probability to explore to find the optimal policies for each day. To do this, the program picks random policies to evaluate the performance by computing the Q-value. The probability to explore decreases exponentially in order for the algorithm to start exploiting the optimal sequence of policies. As mentioned before, the Q-value is computed from the formula 4.7. The learning rate is chosen 0.9, making the algorithm favour the new information over the old one. The discount factor is chosen to 0.9 to take into account future rewards.

5.5 Single workplace model

In this section the design of a single workplace environment simulation and the proposed Markov decision process model for this, are described. Then the reinforcement learning algorithm is applied to the simulation model. Objective is to prove the efficiency of the reinforcement learning algorithm for the training of the general agent-based model simulation and validate its results. If the algorithm gives an optimal sequence of policies for this single workplace environment model, it validates that the same algorithm will finally converge also to an optimal sequence of policies for the general ABM.

5.5.1 Simulation and learning process

The simulation environment includes a workplace and three human agents that are supposed to work at this workplace. The workplace is a 3x3 workplace (covers space of 9 cells in total). The simulation runs for 20 days and three different policies can be followed. The different groups of health condition, each agent can belong, are different comparing to the normal agent-based model simulation. The hospitalized and severe cases groups are excluded, consequently the possible groups of health condition are namely:

- Susceptible.
- Exposed.
- Infected.
- Recovered.

After the initialization, there are two susceptible agents and one infected. The position of the agents at the first day is a random cell in the workplace grid environment.

The agents are supposed to work eight hours per day. During these eight hours, the agents move inside the workplace environment with some probability, which depends on the policy being followed. Every hour is divided in 3 equal timestamps. Thus, every agent "acts" (moves or stays at the same position) 3 times per hour, 24 times per day. The rest sixteen hours of the day, the agents stay stable without moving. Movement of an agent is defined as the change of its position from its current position (cell) to a neighbor position (cell). The neighbor cell where the agents move is selected randomly. At every timestamp each agent can move to eight directions as it is presented in Figure 5.5. Thus the probability of an agent to move to a specific neighbor cell is equal to $\frac{1}{8}$. If a cell outside the grid environment is selected for an agent to move, the random neighbor cell choice process is repeated until one cell inside the grid is selected. For instance, if the current position of an agent is at the cell (3,1) as in Figure 5.5c, then at the next timestamp, there is probability equal to $\frac{1}{8}$, the agent to move to each one of the neighbor cells, which are (2,1), (2,2) and (3,2).

(1,1)	(1,2)	(1,3)
(2,1)	(2,2)	(2,3)
(3,1)	(3,2)	(3,3)

Figure 5.4: Illustration of the single workplace environment. It is composed of 9 cells (3x3). Each cell has a name, (1,1) for example, according to its position at the environment grid.

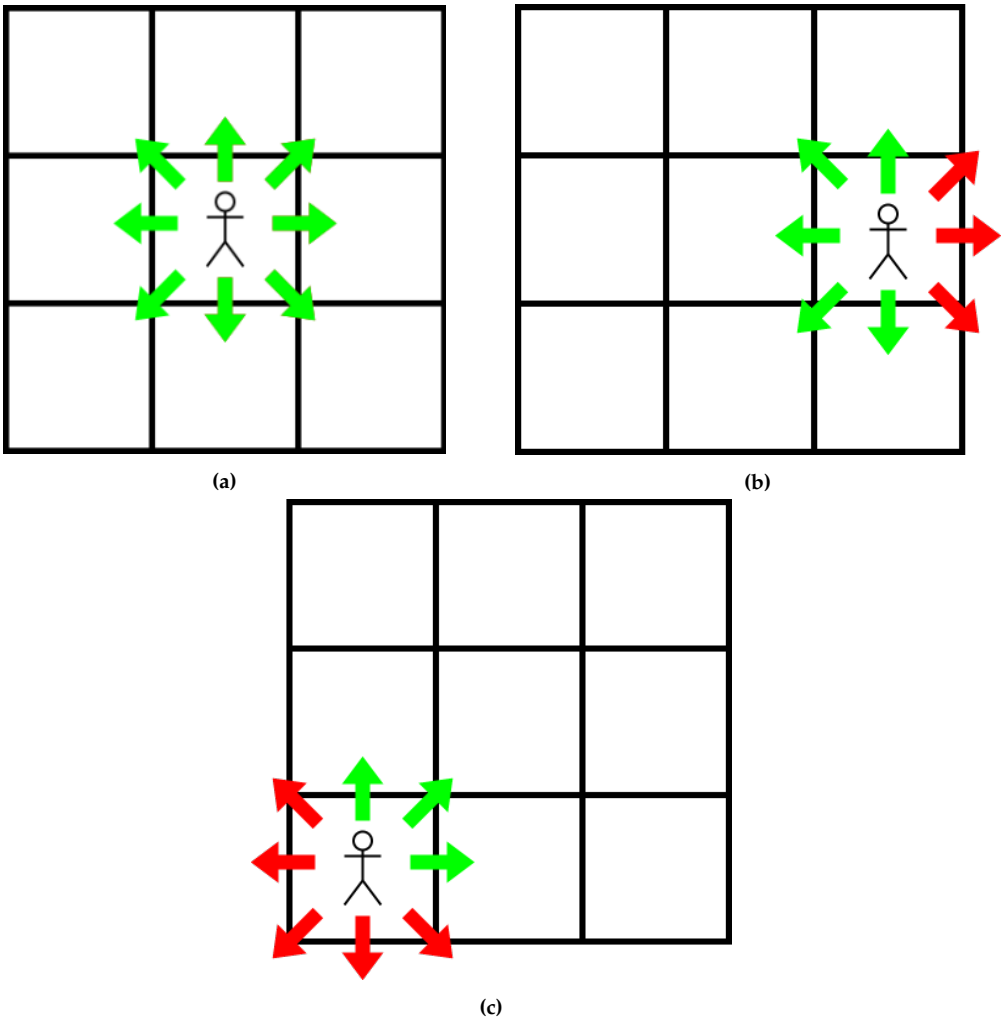

Figure 5.5: Possible movements of the agents depending on their position. (a) Possible movements when an agent is at the central position. (b) Possible movements when an agent is at an edge position. (c) Possible movements when an agent is at a corner position.

In case that an infected and a susceptible agent meet at the same position cell, there is a probability of a contagion to happen (contagion probability, depends on the policy). If the contagion indeed happens, the susceptible agent alters to exposed. In addition, after the incubation time passes, the exposed agent becomes finally infected. Afterwards, the recovery time of the agent follows. At the end of the recovery time, the agent recovers (his group of health condition changes to recovered). 20 days are simulated and policy is selected every day according the exploration and exploitation balance. The ϵ-greedy value is computed by the next equation:

$$\epsilon = e^{-0.0002x} \tag{5.5}$$

Where x is the number of run iteration (1,2,3,...,K iterations). As long the learning process proceeds, the less it explores between policies (picks a random policy to apply) and the more it take advantage of what it has learnt so far.

Iteration number	1	1000	5000
Explore Probability	0.9998	0.8187	0.3678

Table 5.7: Exploration probability values for different stages of reinforcement learning process.

It must be proved that the learning process will converge to an optimal sequence of policies to be followed every day, so that the maximum number of infected people should be minimized (in this single workplace scenario, it should stay equal to 1). Therefore, the expected outcome would be a "lockdown policy" until the infected agent recovers, and then the agents will continue work under normal conditions (during last days of the simulation the "normality policy" should be established). The recovery time is set to 14 or 17 days depending on the age of the employees, as in the normal simulation.

Policies

As referred above, there are three different policies which can be followed in the single workplace simulation, policy 0, policy 1 and policy 2. Below these three policies are explained in more detail.

Policy 0

Policy 0 is supposed to be the "normality" policy. At every timestamp during the eight working hours, each agent has a probability equal to 0.6 to stay at the same position and a probability equal to 0.4 to move to a neighbor position. Therefore the probability of an agent to move to a specific neighbor cell is evaluated as $0.4 * \frac{1}{8} = 0.05$. Face masks are not used and contagion probability is equal to 0.9. Economical reward is equal to 3.

Policy 1

In policy 1, the use of face masks is established. Thus, the contagion probability decreases to 0.3. Furthermore, at every timestamp during the eight working hours, each agent has a probability equal to 0.95 to stay at the same position and a probability equal to 0.05 to move to a neighbor position. Therefore the probability of an agent to move to a specific neighbor cell is evaluated for policy 1 as $0.05 * \frac{1}{8} = 0.00625$. Economical reward decreases now to 1.

Policy 2

In policy 2, all the agents do not move at all. They stay always at the same position. Since the agents don't move, they don't meet each other. Consequently, there are no contagions and no agents get infected. Only case that two agents meet each other at the same cell is at first day where the position of the agents is random. In this case, contagion probability is equal to 0.3. This policy is considered as the "strict lockdown" policy. The economical reward in this scenario is considered equal to -1.

5.5.2 Markov decision process

The single workplace simulation should be described by a Markov decision process so that reinforcement learning algorithm is applicable for it. The MDP is composed of N different states. State to state transitions are characterized by transition probabilities, which depend on the followed policy scenarios. A reward corresponds to each state to state transition. The reward depends on the difference of exposed and infected people between the two states, the number of exposed and infected agents in the next state and the economical reward. The economical reward is determined by the policy being followed.

Each one of the states is described by the next properties:

- Position of each agent (cell in the grid environment).
- Group of health condition of each agent.

For each state, the positions that agents 1, 2 and 3 are located and the group of health condition of each agent must be defined. Example of a random single state of the model can be seen in figure 5.6. State transition can take place once per every timestamp.

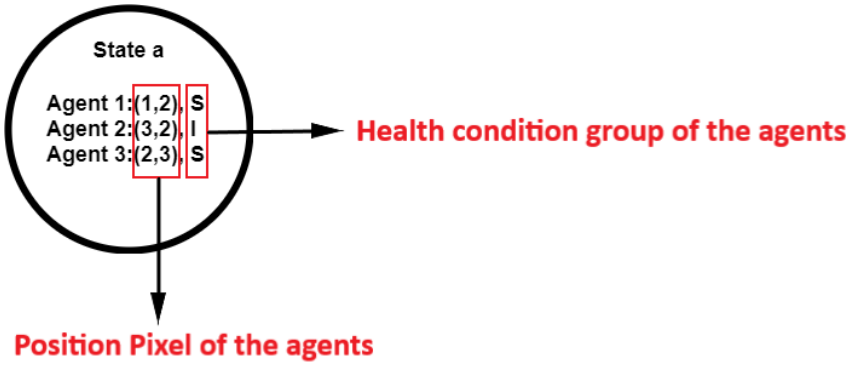

Figure 5.6: Example of a single state of the Markov decision process designed to describe the single workplace simulation.

As stated above, every transition from one state to another is characterized by two parameters, the transition probability and the reward (cost) of the transition. The parameters that affect the reward in each state to state transition are the difference of the exposed and the infected agents between the two states, the number of the exposed and the infected agents at the next state and the economical part of the reward. The reward function is defined as:

$$R_{ij} = (-2) \cdot (E_j - E_i) + (-2) \cdot (I_j - I_i) + (-1) \cdot I_j + (-1) \cdot E_j + Economical\ Reward \tag{5.6}$$

where:

- i is the current state.
- j is the next state.
- R_{ij} is the reward derived by moving from state i to state j.
- E_j and E_i are the number of the exposed agents at the respective states.
- I_j and I_i are the number of the infected agents at the respective states.
- Economical Reward is the economical reward, which depends on the followed policy.

Concerning, the transition probabilities from one state to another, in general they can be defined as:

$$P_{ij} = P((A1_{\,i \to j}) \cap (A2_{\,i \to j}) \cap (A3_{\,i \to j})) \tag{5.7}$$

where:

- P_{ij} is the transition probability from state i to state j.

- A1, A2 and A3 are the three agents of the simulation.
- $(A1_{\ i \to j})$ is the transition of the agent 1 from state i to state j (similar for A2, A3).

Transition of an agent from one state i to another state j consists of two different events. Whether or not agent will move and his position cell will change and whether or not agent's group of health condition will change. These two events are independent from each other and also independent from the respective events of the rest human agents. Equation 5.7 takes the following form:

$$\begin{aligned} P_{ij} &= P(A1_{\ i \to j}) \cdot P(A2_{\ i \to j}) \cdot P(A3_{\ i \to j}) \\ &= P(A1_{at\ position}) \cdot P(A1_{contagion})^{\nu_1} \cdot P(A2_{at\ position}) \cdot P(A2_{contagion})^{\nu_2} \cdot \\ & P(A3_{at\ position}) \cdot P(A3_{contagion})^{\nu_3} \end{aligned} \tag{5.8}$$

where:

- $P(A1_{at\ position})$, $P(A2_{at\ position})$, $P(A3_{at\ position})$ are the probabilities of each agent to move from his position cell at state i to his position cell at state j.
- $P(A1_{contagion})$, $P(A2_{contagion})$, and $P(A3_{contagion})$ are the contagion probabilities for each agent.
- ν_1, ν_2 and ν_3 are set equal to 1, if agent 1, agent 2 and agent 3 respectively, are susceptible and at state j, they meet at the same position cell an infected human agent. In every other occasion these three parameters are set equal to 0.

In Figure 5.7 an example of how the MDP works, how the transitions from state to state take place, and how the transition probabilities and rewards are obtained, is presented. In state "b" no agents meet each other, thus there are no contagions. The transition probability from state "a" to state "b" will be the product of the probability of agent 1 to move from cell (1,2) to cell (1,1) with the probability of agent 2 to move from cell (3,2) to cell (3,1) and the probability of agent 3 to move from cell (2,2) to cell (1,3). Parameters ν_1, ν_2 and ν_3 are equal to 0. If it is assumed that policy 0 is followed this transition probability according to the equation 5.8 will be:

$$\begin{aligned} P_{ab} &= P(A1_{\ a \to b}) \cdot P(A2_{\ a \to b}) \cdot P(A3_{\ a \to b}) \\ &= (0.4 \cdot \frac{1}{8}) \cdot (0.4 \cdot \frac{1}{8}) \cdot (0.4 \cdot \frac{1}{8}) \\ &= 0.000125 \end{aligned} \tag{5.9}$$

If it is assumed that policy 1 is followed this transition probability will be:

$$\begin{aligned} P_{ab} &= P(A1_{\,a\to b}) \cdot P(A2_{\,a\to b}) \cdot P(A3_{\,a\to b}) \\ &= (0.05 \cdot \frac{1}{8}) \cdot (0.05 \cdot \frac{1}{8}) \cdot (0.05 \cdot \frac{1}{8}) \\ &= 0.000000244140625 \end{aligned} \tag{5.10}$$

If it is assumed that policy 2 is followed this transition probability will be:

$$\begin{aligned} P_{ab} &= P(A1_{\,a\to b}) \cdot P(A2_{\,a\to b}) \cdot P(A3_{\,a\to b}) \\ &= 0 \cdot 0 \cdot 0 = 0 \end{aligned} \tag{5.11}$$

In contrast, in state "e" of Figure 5.7 one susceptible and one infected agent meet each other, thus there is a contagion. Parameter ν_1 is set equal to 1, since agent 1 is susceptible and meets the infected agent 2 at the position cell (1,1). Parameters ν_2 and ν_3 are equal to 0. The transition probability from state "d" to state "e" will be the product of the probability of agent 1 to move from cell (2,1) to cell (1,1) with the probability of agent 2 to move from cell (2,1) to cell (1,1) with the probability of agent 3 to move from cell (2,3) to cell (3,3) and with the contagion probability of agent 1. If it is assumed that policy 0 is followed this transition probability according to the equation 5.8 will be:

$$\begin{aligned} P_{de} &= P(A1_{\,d\to e}) \cdot P(A2_{\,d\to e}) \cdot P(A3_{\,d\to e}) \\ &= (0.4 \cdot \frac{1}{8} \cdot 0.9) \cdot (0.4 \cdot \frac{1}{8}) \cdot (0.4 \cdot \frac{1}{8}) \\ &= 0.0001125 \end{aligned} \tag{5.12}$$

If it is assumed that policy 1 is followed this transition probability will be:

$$\begin{aligned} P_{de} &= P(A1_{\,d\to e}) \cdot P(A2_{\,d\to e}) \cdot P(A3_{\,d\to e}) \\ &= (0.05 \cdot \frac{1}{8} \cdot 0.3) \cdot (0.05 \cdot \frac{1}{8}) \cdot (0.05 \cdot \frac{1}{8}) \\ &= 0.0000000732421875 \end{aligned} \tag{5.13}$$

If it is assumed that policy 2 is followed this transition probability will be:

$$\begin{aligned} P_{de} &= P(A1_{\,d\to e}) \cdot P(A2_{\,d\to e}) \cdot P(A3_{\,d\to e}) \\ &= 0 \cdot 0.3 \cdot 0 \cdot 0 = 0 \end{aligned} \tag{5.14}$$

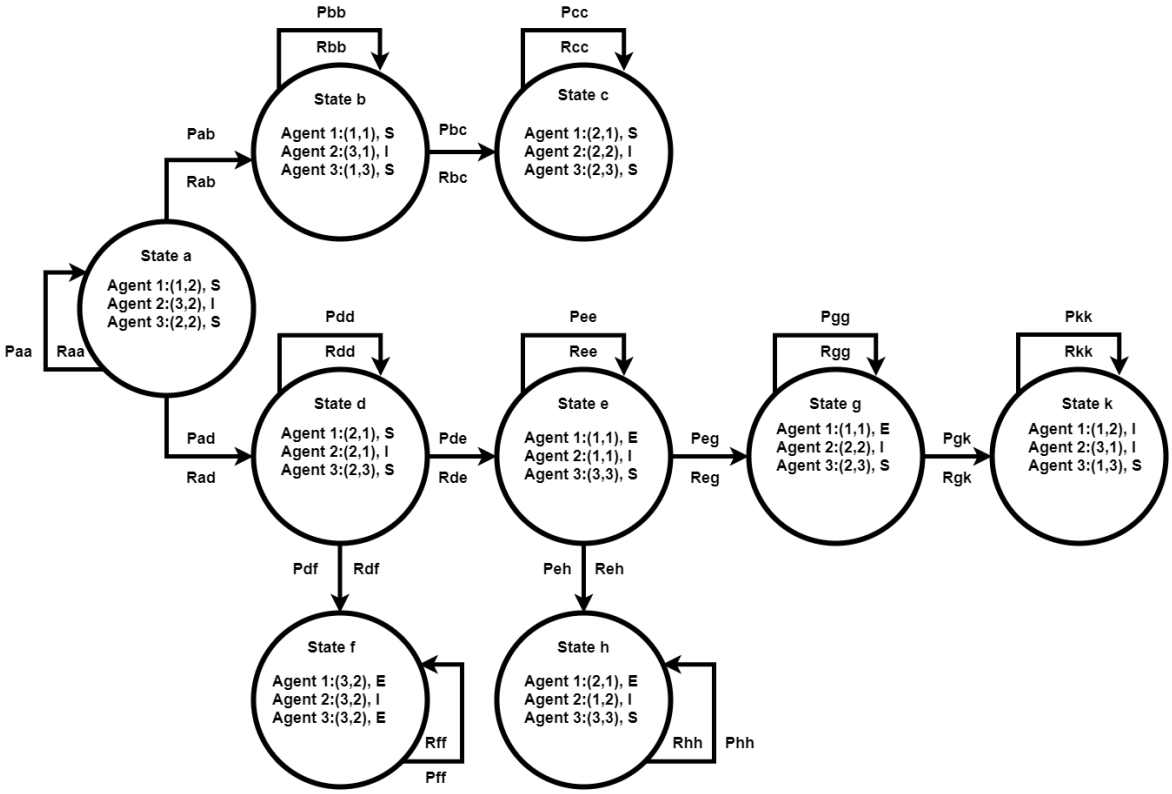

Figure 5.7: Markov Decision Process model example of the single workplace simulation.

It is assumed that at some point of the simulation the system is at the state "a" of the Figure 5.7. There is one susceptible agent in each cells (1,2) and (2,2) and one infected agent at cell (3,2). A possible transition is to state "b" where all agents move but don't meet each other, so there is no contagion and parameters ν_1, ν_2 and ν_3 are equal to 0. More particularly, agent 1 moves to cell (1,1), agent 2 moves to cell (3,1) and agent 3 moves to cell (1,3). From this motion, it is concluded that either policy 0 or policy 1 was followed (in policy 2 agents stay always at the same position). At state "b" there are two susceptible agents and one infected, consequently the difference between exposed and infected agents between states "a" and "b" is zero. The corresponding reward for this transition is evaluated as:

$$R_{ab} = (-2) \cdot (0) + (-2) \cdot (0) + (-1) \cdot 1 + (-1) \cdot 0 + \textit{Economical Reward} \quad (5.15)$$

As mentioned either policy 0 or policy 1 was followed so economical reward will be equal to either 3 or 1. Assuming that policy 0 was followed, the derived reward is evaluated as:

$$R_{ab} = (-2) \cdot (0) + (-2) \cdot (0) + (-1) \cdot 1 + (-1) \cdot 0 + 3 = (-1) + 3 = 2 \quad (5.16)$$

The transition probability P_{ab} is equal to 0.000125 as computed in Equation 5.9. At the next timestamp, another possible transition is the system to stay at state "b". There is a probability that all agents stay at the same cell, no matter which policy is followed, and since there are no contagions, groups of health condition of the agents remain the same. Parameters ν_1, ν_2 and ν_3 are again set equal to 0. The reward for this transition is evaluated as:

$$\begin{aligned} R_{bb} &= (-2)\cdot(0) + (-2)\cdot(0) + (-1)\cdot 1 + (-1)\cdot 0 + \textit{Economical Reward} \\ &= (-1) + \textit{Economical Reward} \end{aligned} \tag{5.17}$$

The transition probability P_{bb} is estimated as:

$$\begin{aligned} P_{bb} &= P(A1_{\,b\to b}) \cdot P(A2_{\,b\to b}) \cdot P(A3_{\,b\to b}) \\ &= \begin{cases} 0.6\cdot 0.6\cdot 0.6 = 0.216 & \text{for policy 0} \\ 0.95\cdot 0.95\cdot 0.95 = 0.857375 & \text{for policy 1} \\ 1\cdot 1\cdot 1 = 1 & \text{for policy 2} \end{cases} \end{aligned} \tag{5.18}$$

Moreover, a possible transition from state "a" in Figure 5.7 is to state "d". At state "d" one susceptible agent, the agent 1, and one infected agent, the agent 2 meet each other at position cell (2,1). Thus, there is probability a contagion to take place and ν_1 is set equal to 1. However, it is observed that in state "d" the agent 1 is still susceptible. This means that the contagion never happened. Contagion probability depends on the policy as stated above and can be equal either to 0.9 or 0.3. Therefore the probability of a contagion not to happen when one susceptible and one infected agent meet each other is 0.1 and 0.7 for policy 0 and policy 1 respectively. The reward for the transition from state "a" to state "d" is:

$$\begin{aligned} R_{ad} &= (-2)\cdot(0) + (-2)\cdot(0) + (-1)\cdot 1 + (-1)\cdot 0 + \textit{Economical Reward} \\ &= (-1) + \textit{Economical Reward} \end{aligned} \tag{5.19}$$

The transition probability from state "a" to state "d" is:

$$\begin{aligned} P_{ad} &= P(A1_{\,a\to d}) \cdot P(A2_{\,a\to d}) \cdot P(A3_{\,a\to d}) \\ &= \begin{cases} 0.4\cdot\frac{1}{8}\cdot 0.1\cdot 0.4\cdot\frac{1}{8}\cdot 0.4\cdot\frac{1}{8} = 0.0000125 & \text{for policy 0} \\ 0.05\cdot\frac{1}{8}\cdot 0.7\cdot 0.05\cdot\frac{1}{8}\cdot 0.05\cdot\frac{1}{8} = 0.0000001708984375 & \text{for policy 1} \\ 0\cdot 0.7\cdot 0\cdot 0 = 0 & \text{for policy 2} \end{cases} \end{aligned} \tag{5.20}$$

A possible next state is the state "e" where all agents move, agent 1 meets agent 2 at the same position cell again and he is now exposed. Thus, between states "e"

and "d" there is a difference of the number of exposed agents. The reward for the transition from state "d" to state "e" is:

$$\begin{aligned} R_{de} &= (-2)\cdot(1-0)+(-2)\cdot(0)+(-1)\cdot 1+(-1)\cdot 1+\textit{Economical Reward} \\ &= (-4)+\textit{Economical Reward} \end{aligned} \tag{5.21}$$

The transition probability P_{de} is calculated at Equations 5.12, 5.13 and 5.14. Another possible next state from state "d" is state "f". At state "f" all three agents move at the position cell (3,2). Susceptible agents 1 and 3 meet the infected agent 2. Thus parameters ν_1 and ν_3 are set equal to 1. It is observed that both agent 1 and agent 3 are exposed at state "f", so two contagions took place and there are two more exposed agents at state "f" comparing with state "d". The reward for the transition from state "d" to state "f" is:

$$\begin{aligned} R_{df} &= (-2)\cdot(2-0)+(-2)\cdot(0)+(-1)\cdot 1+(-1)\cdot 2+\textit{Economical Reward} \\ &= (-7)+\textit{Economical Reward} \end{aligned} \tag{5.22}$$

The transition probability from state "d" to state "f" is:

$$\begin{aligned} P_{df} &= P(A1_{\,d\to f})\cdot P(A2_{\,d\to f})\cdot P(A3_{\,d\to f}) \\ &= \begin{cases} 0.4\cdot\frac{1}{8}\cdot 0.9\cdot 0.4\cdot\frac{1}{8}\cdot 0.4\cdot\frac{1}{8}\cdot 0.9 = 0.00010125 & \text{for policy 0} \\ 0.05\cdot\frac{1}{8}\cdot 0.3\cdot 0.05\cdot\frac{1}{8}\cdot 0.05\cdot\frac{1}{8}\cdot 0.3 = 0.00000002197265625 & \text{for policy 1} \\ 0\cdot 0.3\cdot 0\cdot 0\cdot 0.3 = 0 & \text{for policy 2} \end{cases} \end{aligned} \tag{5.23}$$

Finally, if agent 1 gets exposed, he will become infected after incubation time passes. Assuming that the system is at state "g" of Figure 5.7 exactly one timestamp before agent 1 group of health condition turns to infected, then a possible transition from state "g" is to state "k". At state "k" there are two infected agents, one more comparing to state "g", no exposed agents and one susceptible. The three agents are at three different position cells. The reward for this transition is:

$$R_{gk} = (-2)\cdot(0-1)+(-2)\cdot(2-1)+(-1)\cdot 2+(-1)*0+\textit{Economical Reward} \tag{5.24}$$

It is observed that the difference of the exposed agents between states "g" and "k" $(E_k - E_g)$ is equal to -1. In this case this difference term is set to zero because only difference caused from agents becoming exposed from susceptible is taken under consideration and not difference caused from exposed turning to infected. Same happens with the difference of infected agents between two states, when an infected agent recovers. In other words, terms $(E_j - E_i)$ and $(I_j - I_i)$ of the Equation 5.6 should be positive. If they are negative, they are set equal to zero. Therefore,

the corresponding reward for the transition from state g to state k will ultimately be equal to:

$$R_{gk} = -4 + Economical\ Reward \tag{5.25}$$

5.6 Dynamic programming for single workplace simulation

In the next section, a dynamic programming algorithm is implemented for the single workplace simulation. The objective is to find an optimal sequence of policies for this simulation and show that the reinforcement learning algorithm converges to the same sequence of policies. This will validate the results obtained from our reinforcement learning algorithm and prove its efficiency also for the general agent-based model simulation.
In the single workplace simulation, the goal is to minimize the number of the infected people. Three human agents are generated as two of them being susceptible and one of them being infected. Optimally, none of the other two human agents should get infected during the simulation. This can be achieved if the human agents don't meet at all until the infected agent recovers. As a result, no contagion will take place and the virus will not be transmitted to any susceptible person. In simulation, such a sequence state can be the following:

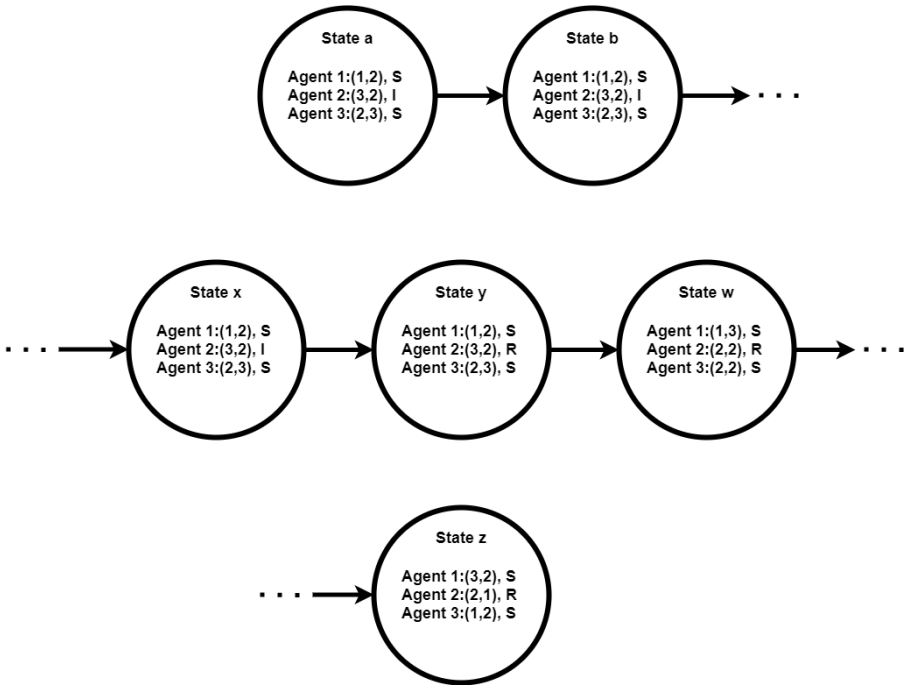

Figure 5.8: Example sequence of states at the single workplace model simulation. There is no virus transmission.

In general, three kinds of state to state transitions are performed for the scenario presented in Figure 5.8. One of them is when the system goes from a state where two agents are susceptible and one is infected to another state where still the same two agents are susceptible and the other one is still infected. This transition is depicted in Figure 5.9:

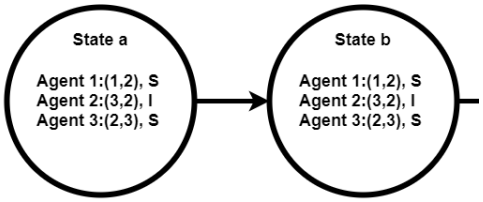

Figure 5.9: Example of transition from a state with two susceptible agents and one infected to another state where both susceptible agents remain susceptible and the infected agent remains infected.

Another transition is the transition to the state that the infected agent recovers and it is depicted in Figure 5.10.

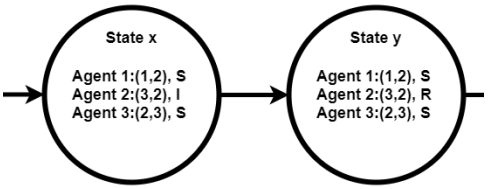

Figure 5.10: Example of transition from a state with two susceptible agents and one infected to another state where both susceptible agents remain susceptible and the infected agent recovers.

After the infected agent recovers, the rest transitions will be from a state where two agents are susceptible and the other one is recovered to another state where the same two agents are still susceptible and the previously recovered agent remains recovered. This kind of transition is depicted in Figure 5.11:

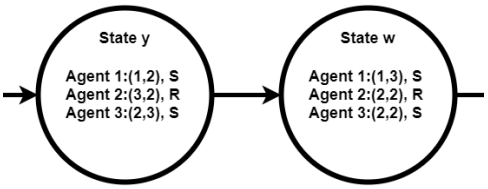

Figure 5.11: Example of transition from a state with two susceptible agents and one recovered to another state where both susceptible agents remain susceptible and the recovered agent remains recovered.

Considering that the recovery time for every person is either 14 or 17 days, the expected sequence of states should have the form of the sequence in Figure 5.12:

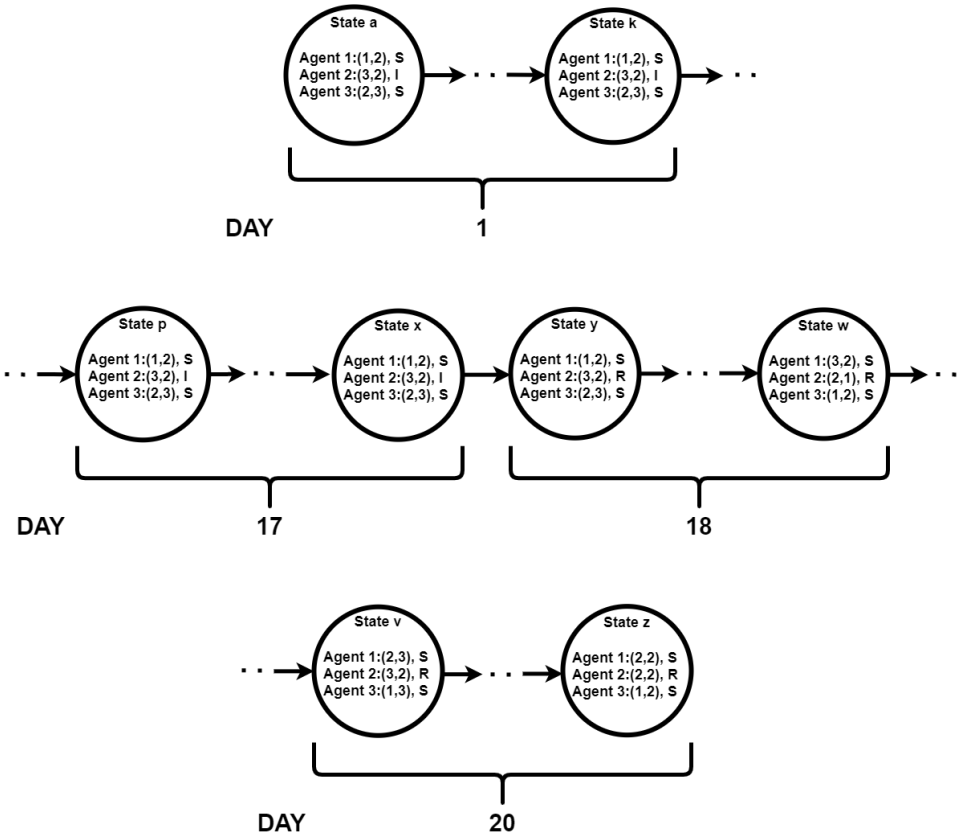

Figure 5.12: Example sequence of states at the single workplace model simulation time period. There is no virus transmission.

Dynamic programming algorithm is applied to compute the optimal gain of each one of the above transitions, and to find out which policy should be followed, in order to obtain an optimal gain. The optimal gain of each transition is calculated by the following equation:

$$J_i = \max_{h_i}\{R_{i\to i+1}\} + J_{i+1} \tag{5.26}$$

where:

- J_i is the gain of the current state.
- J_{i+1} is the gain of the state $i+1$.
- h_i represents the followed policy.
- $R_{i\to i+1}$ is the reward of the transition $i \to i+1$.

Equation 5.26 is called Bellman equation. The reward for each transition is given by Equation 5.6. State transition takes place at each timestamp. In the single

workplace simulation, each simulated hour is divided to three timestamps, thus during the whole simulation there are 20 days x 24 hours x 3 timestamps = 1440 timestamps. As a result, there are 1440 state transitions and 1441 different states in total during the simulation. The gain J_{1441} of the last state is set equal to 0.

During day 20 at the transition from state 1440 to the last state 1441, two susceptible agents remain susceptible and the recovered one remains recovered. Considering the reward function, it can be concluded that there will be no difference of exposed or infected agents between the two states and also there will be no exposed or infected cases. The transition between these two days is the same as the one in Figure 5.11. The reward of this transition depends only on the economic reward. In this case, the gain of the state 1440 is evaluated as:

$$J_{1440} = \max_{h_{1440}} \{R_{1440 \to 1441}\} + J_{1441} \tag{5.27}$$

If policy 0 is followed, Equation 5.27 takes the form:

$$J_{1440} = (-2) \cdot (0) + (-2) \cdot (0) + (-1) \cdot (0) + (-1) \cdot (0) + 3 + 0 = 3 \tag{5.28}$$

If policy 1 is followed, Equation 5.27 takes the form:

$$J_{1440} = (-2) \cdot (0) + (-2) \cdot (0) + (-1) \cdot (0) + (-1) \cdot (0) + 1 + 0 = 1 \tag{5.29}$$

If policy 2 is followed, Equation 5.27 takes the form:

$$J_{1440} = (-2) \cdot (0) + (-2) \cdot (0) + (-1) \cdot (0) + (-1) \cdot (0) - 1 + 0 = -1 \tag{5.30}$$

The gain of state 1440 is larger when policy 0 is followed, thus policy 0 is the optimal one for the transition $1440 \to 1441$ and for all similar transitions with the one, which is illustrated in Figure 5.11. Same applies for all the transitions during the days 19 and 18.

On day 17, one agent is still infected and recovers at day 18. The other two agents remain susceptible. This transition takes place between the last state of day 17 and the first state of day 18. These two are states 1225 and 1226 respectively. The term $I_{1226} - I_{1225}$ is equal to -1, but it is taken under consideration only when it is positive, so it is set to 0. The transition is similar with the one in Figure 5.10. The gain for the state 1225 is evaluated as:

$$J_{1225} = \max_{h_{1225}} \{R_{1225 \to 1226}\} + J_{1226} \tag{5.31}$$

If policy 0 is followed, Equation 5.31 takes the form:

$$J_{1225} = (-2) \cdot (0) + (-2) \cdot (0) + (-1) \cdot (0) + (-1) \cdot (0) + 3 + J_{1226} = 3 + J_{1226} \tag{5.32}$$

If policy 1 is followed, Equation 5.31 takes the form:

$$J_{1225} = (-2)\cdot(0) + (-2)\cdot(0) + (-1)\cdot(0) + (-1)\cdot(0) + 1 + J_{1226} = 1 + J_{1226} \tag{5.33}$$

If policy 2 is followed, Equation 5.31 takes the form:

$$J_{1225} = (-2)\cdot(0) + (-2)\cdot(0) + (-1)\cdot(0) + (-1)\cdot(0) - 1 + J_{1226} = -1 + J_{1226} \tag{5.34}$$

From previous calculations, we can conclude that policy 0 is the optimal one since maximum gain is obtained under the application of policy 0.

All the state to state transitions from state 1 to state 1225 are transitions between states where the two agents remain susceptible and the other one remains infected. These transitions are similar to the transition in Figure 5.9. When a transition from state 1224 to state 1225 is considered and policy 0 or policy 1 is applied, there is a possibility that the infected agent meets either one or both of the susceptible agents. Therefore, there is a possibility that one or two contagions will take place and at state 1225 there will be one or two exposed agents. However, this is not the case, if policy 2 is followed, since agents don't move. The gain for this transition is evaluated as:

$$J_{1224} = \max_{h_{1224}}\{R_{1224\to1225}\} + J_{1225} \tag{5.35}$$

If policy 0 is followed, Equation 5.35 takes the form:

$$J_{1224} = (-2)\cdot(1)\cdot P_{01} + (-2)\cdot(2)\cdot P_{02} + (-2)\cdot(0) + (-1)\cdot(1) + (-1)\cdot(1)\cdot P_{01} + (-1)\cdot(2)\cdot P_{02} + 3 + J_{1225} = -3P_{01} - 6P_{02} + 2 + J_{1225} \tag{5.36}$$

If policy 1 is followed, Equation 5.35 takes the form:

$$J_{1224} = (-2)\cdot(1)\cdot P_{11} + (-2)\cdot(2)\cdot P_{12} + (-2)\cdot(0) + (-1)\cdot(1) + (-1)\cdot(1)\cdot P_{11} + (-1)\cdot(2)\cdot P_{12} + 1 + J_{1225} = -3P_{11} - 6P_{12} + J_{1225} \tag{5.37}$$

If policy 2 is followed, Equation 5.35 takes the form:

$$J_{1224} = (-2)\cdot(0) + (-2)\cdot(0) + (-1)\cdot(1) + (-1)\cdot(0) - 1 + 0 = -2 + J_{1225} \tag{5.38}$$

where:

- P_{01} is the probability that one susceptible agent meets the infected agent at the same position cell when policy 0 is followed.
- P_{02} is the probability that both susceptible agents meet the infected agent at the same position cell when policy 0 is followed.

- P_{11} is the probability that one susceptible agent meets the infected agent at the same position cell when policy 1 is followed.
- P_{12} is the probability that both susceptible agents meet the infected agent at the same position cell when policy 1 is followed.

It is expected that the gain obtained from policy 2 will be higher than the gains obtained from the other two policies. However, the computation of the the probabilities P_{01}, P_{02}, P_{11} and P_{12} is complicated, since there are multiple different combinations of the motion of these agents in the $3x3$ grid environment. The main difficulty is that agents move only to neighbor cells. Policy 2 is expected to be the optimal one for the transition $1224 \rightarrow 1225$. Moreover, it is expected to also be the optimal one for all the transitions from state 1 to state 1225. These transitions are all of the same type, which is depicted in Figure 5.9.

In conclusion, policy 2 is the expected policy to be followed during the simulation, until the infected agent recovers. After that policy 0 will be the optimal policy until the end of the simulation.

5.7 Implementation

Both initial models with a population of 400 and the illustrative model are implemented using a Visual Studio integrated development environment from Microsoft. Both applications were developed using a C++ programming language. The Python programming language was considered in the development phase, but the much slower simulation time influenced the choice of C++.

5.7.1 Structure of the program, initial model

The image below presents a structure of the C++ program for the initial model.

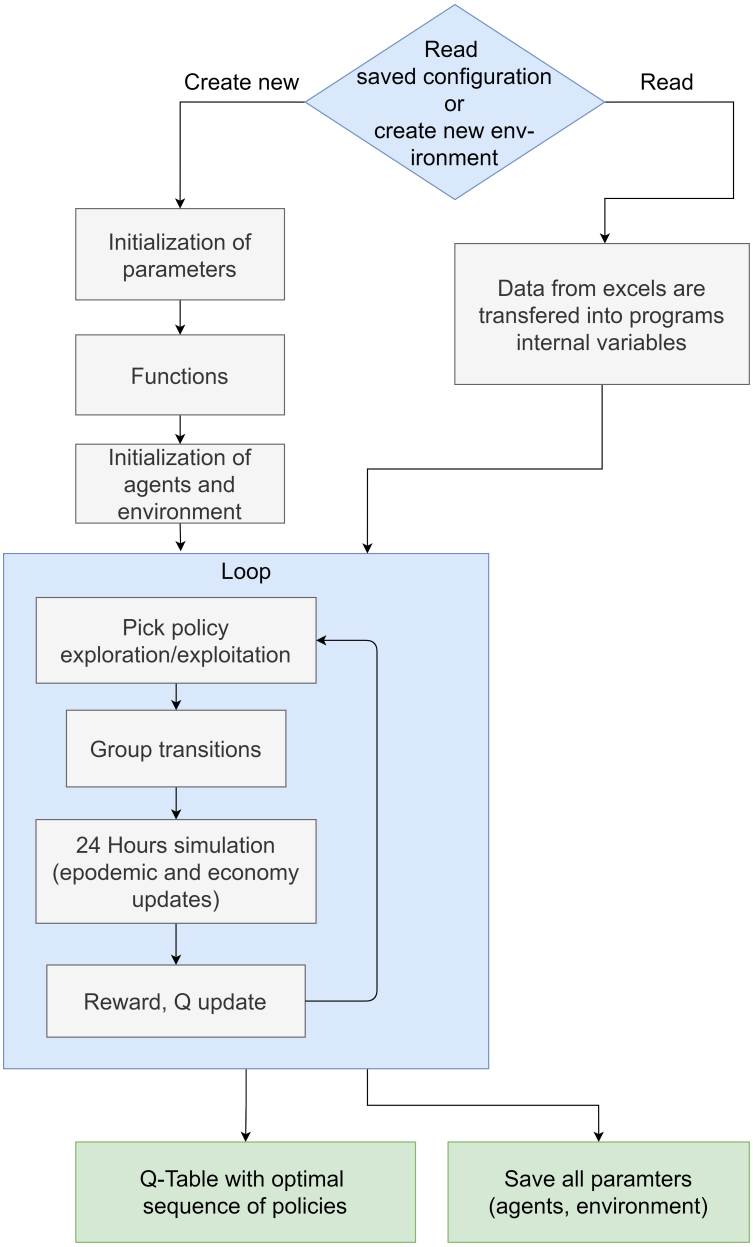

Figure 5.13: Flowchart of the program.

First, all the parameters are initialized. These parameters are environmental parameters like population, unemployment rate, simulation time and epidemiological parameters, for example, the number of initial immunes, economical parameters like wealth as well as all parameters needed to simulate for example various counters etc. Moreover, classes for all agents are defined with all parameters and functions (actions) of agents.

Then all functions are defined. There are functions that control the flow of money in the environment. There is a function that controls the transfer between epidemiological groups. Another function checks all contacts between humans and if the contact is contagious and the relevant conditions are met, the certain individual is transferred to the appropriate group. Another function grants proper reward according to the performance of the simulation. The remaining function updates the Q-table.

All 6 policies are also defined as functions. Each of them is divided into certain time blocks in which people perform different actions. Implementation of each policy is described in Table 5.8.

Next, the environment with structures and humans with proper parameters is created. There is 1% of infected individuals at the start.

When all preparations are finished, the program enters into a loop which length is specified by the simulation time in days. During this loop, the program goes through various steps to update the environment as well as find the optimal control.

First, the policy for the upcoming day is chosen. The choice depends on the probability of exploration versus exploitation. If the exploration is chosen, the policy is picked randomly with uniform distribution. If the exploitation is chosen, the program checks in the Q-table which policy is optimal for the upcoming day and applies it to the environment.

Secondly, the group transitions of humans are performed. The transition path between the groups can be seen on Figure 5.2.

Next, the simulation of 24 hours of the environment is executed. During this part, the proper function which describes the chosen policy is executed. Each hour is divided into 200 timestamps. During each timestamp, humans perform actions and their positions are monitored in order to track the contagious contacts. Overall, each human perform 4800 actions each day, whether it is to stay or move in the grid environment.

The last step in the loop is to grant a reward and update the Q-value in Q-table for chosen policy.

After the loop, the program saves the Q-table into an excel file, to be able to read it and perform operations on it in the future. The program can also save all the details about the environment and agents, so when the program is executed again, it can run with the same parameters as before.

5.7.2 Structure of the program, single workplace model

The structure of the simulation of the workplace is similar to the bigger model with differences in the environment and the policies. The model was developed

to validate the correctness of the reinforcement learning algorithm. The reason to reduce the complexity of the model substantially is that it is easier to control since fewer actions are occurring making the model dynamics predictable.

A the beginning, only 3 individuals are initialized in the 3 by 3 calls workplace environment. The actions of the individuals are to stay in the cell or move to one of the surrounding cells. The simulations go through an 8-hour loop. This time, each hour is divided into 3 timestamps. There are 3 policies prepared for this simulation.

The program has an option to save and read the initialized environment as well as Q-table.

5.7.3 Learning procedure

The learning procedure consists of running the program many times to fill out the Q-table with converged Q-values. The learning procedure was conducted on the small model. The result - Q-table can be seen in Chapter 7. The learning procedure was not conducted on the initial model.

5.7.4 Implementation of policies

The policies are implemented as functions in the C++ program. In each function for each policy, certain actions are specified for each hour of the simulated environment.

Hour \Policy	Policy 0	Policy 1
0-8	stay at home	stay at home
8-12	go to work go to school (students) stay at home (quarantined)	go to work stay at home (quarantined, risk group)
12 - 14	walk free stay at home (quarantined)	walk free stay at home (quarantined, risk group)
14 - 18	go to work stay at home (quarantined, students)	go to work stay at home (quarantined, risk group)
18 - 24	walk free stay at home (quarantined)	walk free stay at home (quarantined, risk group)
Hour \Policy	Policy 2	Policy 3
0-8	stay at home	stay at home
8-12	go to work 2 stay at home (quarantined, risk group)	go to work 2 stay at home (quarantined, risk group)
12 - 14	walk free stay at home (quarantined, risk group)	walk free stay at home (quarantined, risk group)
14 - 18	go to work 2 stay at home (quarantined, risk group)	go to work 2 stay at home (quarantined, risk group)
18 - 24	walk free stay at home (quarantined, risk group)	walk free stay at home (quarantined, risk group)
Hour \Policy	Policy 3 Lockdown	Policy 4
0-8	stay at home	stay at home
8-12	stay at home go to work (essential workers)	stay at home go to work (essential workers)
12 - 14	stay at home go to work (essential workers)	stay at home go to work (essential workers)
14 - 18	stay at home go to work (essential workers)	stay at home go to work (essential workers)
18 - 24	stay at home	stay at home
Hour \Policy	Policy 5	
0-8	stay at home	
8-12	stay at home	
12 - 14	stay at home	
14 - 18	stay at home	
18 - 24	stay at home	

Table 5.8: The implementation of the policies in the program. Each day is divided into 5 periods of time in which agents perform an actions. Some actions are performed only by specific groups denoted in parenthesis next to action.

Chapter 6

Tests and Results

In the following chapter, the results of various tests that were implemented are presented. The chapter is divided into two sections. Initially, in the first part of the chapter, epidemiological and financial results of every single policy of the general ABM are described. Each policy is tested separately, in a separate simulation. Each simulation takes a time period of 60 days. The agents are generated at the first run (test for policy 0). The agents and their various characteristics are then saved to an excel file and they are imported to the simulation at the next runs so that the tests for all policies take place in the same environment. Afterwards, the results of the different policies are compared to each other.

Moreover, epidemiological results of the implemented policies in the single workplace environment are presented. Once more, each policy is tested separately in a separate simulation. Each simulation takes a time period of 20 days.

Afterwards, in the second section of the chapter, the derived results from the application of the reinforcement learning algorithm are presented, for the single workplace environment, the second illustrative model and the general ABM.

6.1 Agent-based model: policies tests

As mentioned above, each policy of the general ABM is tested separately. In addition, the obtained results are compared to each other, so that the "performance" and the "efficiency" of each policy, regarding both epidemiological (results of SEIHVR epidemiological model) and economical level can be concluded.

6.1.1 Policy 0

First, policy 0 is tested. In this first simulation the initialization process takes place and the agents are generated. Some of the most important characteristics of the generated environment are:

- Initially the infected people are four, the immune people (they are considered already recovered) are 3, and the rest 393 people are susceptible.
- Number of students is 75.

- Number of people over 60 years old which are considered to belong to risk group are 93.
- Number of employees, which are essential and are supposed to go to their workplaces even under lockdown policies are 51. One of them is also initially infected.

As stated, simulated time period is 92 days.
In policy 0 there is no restrictions. People move freely, shops are open, employees and students go to their workplaces and school respectively. It is considered as the "normality" policy. The resulting epidemiological graphs of the SEIHVR model for policy 0 are:

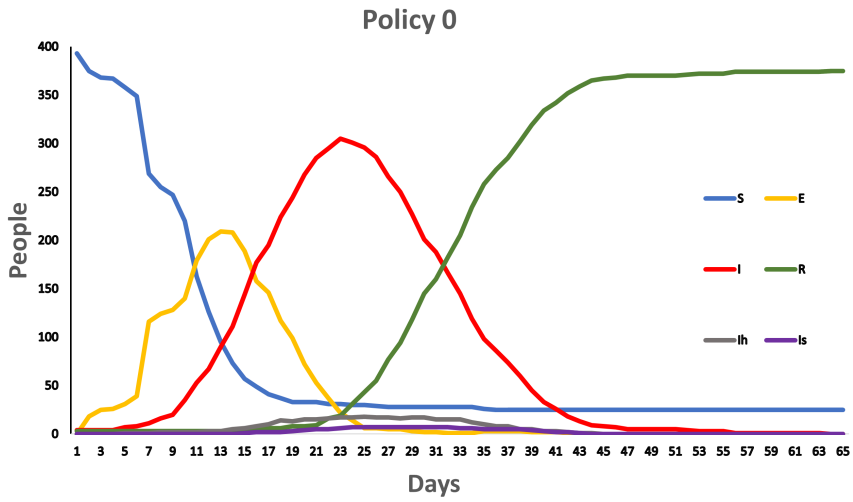

Figure 6.1: Results of epidemiological model when policy 0 is applied to the ABM.

A total of 372 people get infected from the virus, while the peak of the infected curve is 305 people. The peak number of hospitalized patients is 18, exceeding the desired threshold. The peak number of patients in intensive care units (severe cases) is 7. After 64 days all the infected patients have been recovered and 25 people remain susceptible.

Considering the economical effects, it can be observed from Figure 6.2, that during each month the wealth of the businesses is increasing, since they are open for the clients and people move freely. The more people move, the more "shopping" actions take place, which means a bigger profit for the businesses. The decrease of the businesses' wealth at the end of each month is due to salaries and taxes payments. At the same time, houses' wealth is decreasing through the month due to the everyday expenses of their residents, while at the end of the month people are getting paid, thus houses' wealth increases.

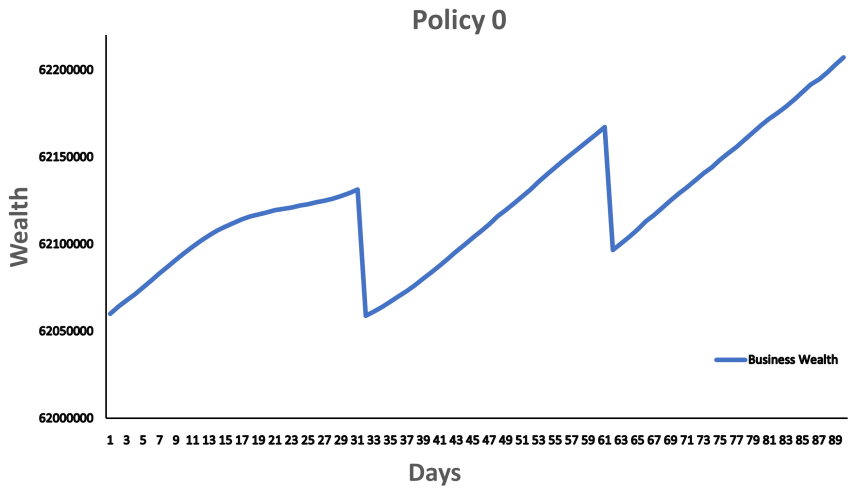

Figure 6.2: Economical effects of policy 0 to businesses.

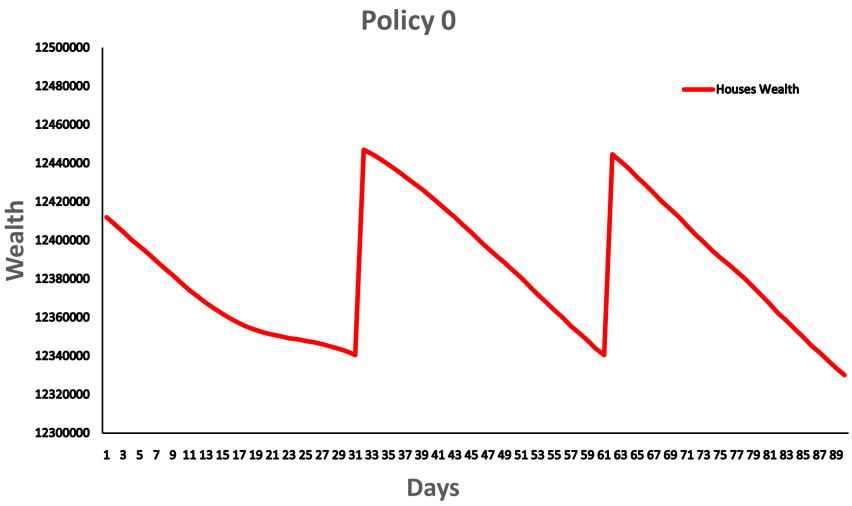

Figure 6.3: Economical effects of policy 0 to houses.

6.1.2 Policy 1

Next, policy 1 is simulated. The same environment of agents as in policy 0 is used. In policy, 1 vertical isolation is applied. This means all people that are over 60 years old and belong to the risk group stay isolated at their home. Furthermore, school is closed, therefore students also stay at home. The resulting epidemiological graphs of the SEIHVR model for policy 1 are:

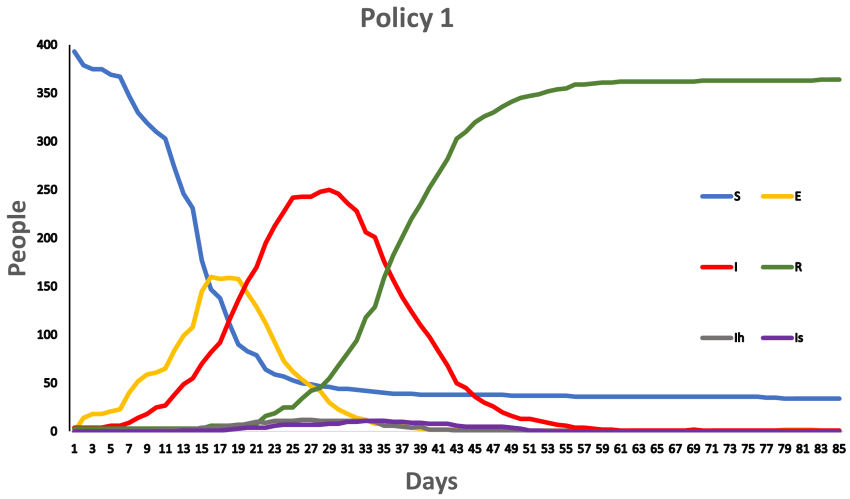

Figure 6.4: Results of epidemiological model when policy 1 is applied to the ABM.

The total number of people that get infected falls at 363 people and the peak of the infected curve is 250 infected cases. The peak number of the hospitalized patients is 12 and of the severe cases is 11. After 60 days, there are still 2 infected people, and 34 people remain susceptible.

Regarding economical effects, businesses' wealth still increases during the months, since there are a lot of people who move freely. However, their profits are smaller comparing to the respective profits in policy 0. Houses' wealth decreases during the month as in policy 0, but with a lower rate.

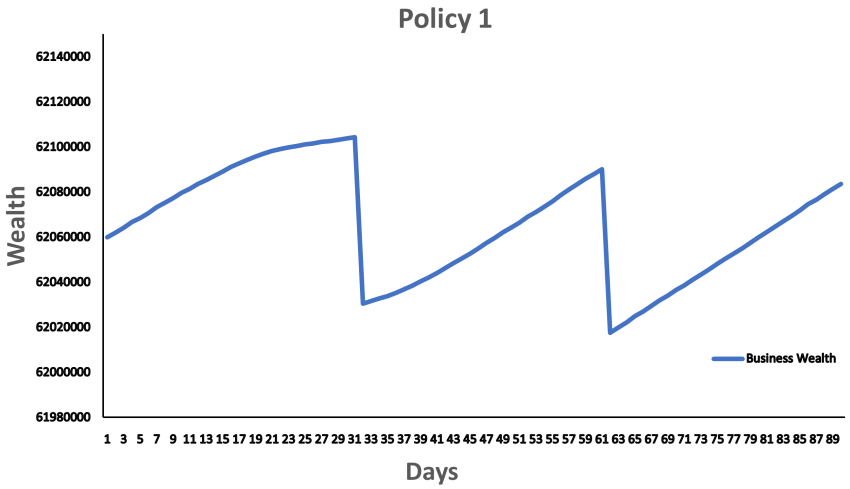

Figure 6.5: Economical effects of policy 1 to businesses.

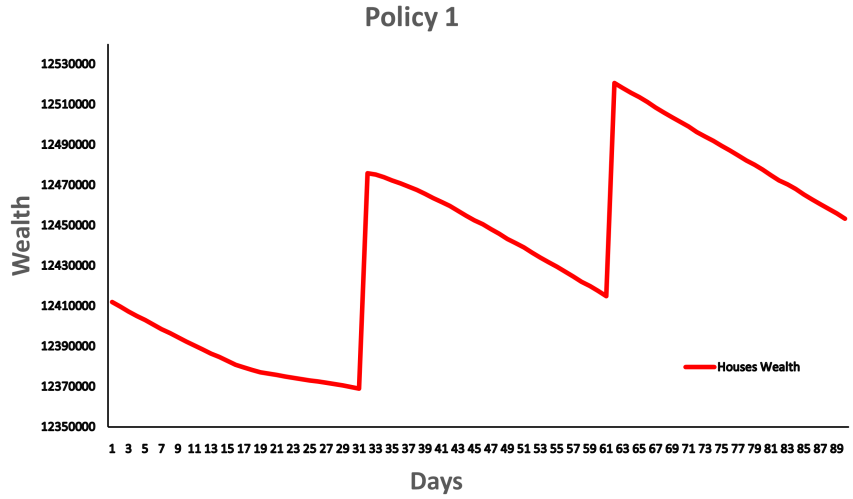

Figure 6.6: Economical effects of policy 1 to houses.

6.1.3 Policy 2

In policy 2 the same restrictions as policy 1 are implemented. The only difference is that the use of face masks become obligatory. Practically this means that the contagion probability reduces from 0.9 to 0.3. The derived epidemiological graphs for policy 2 are:

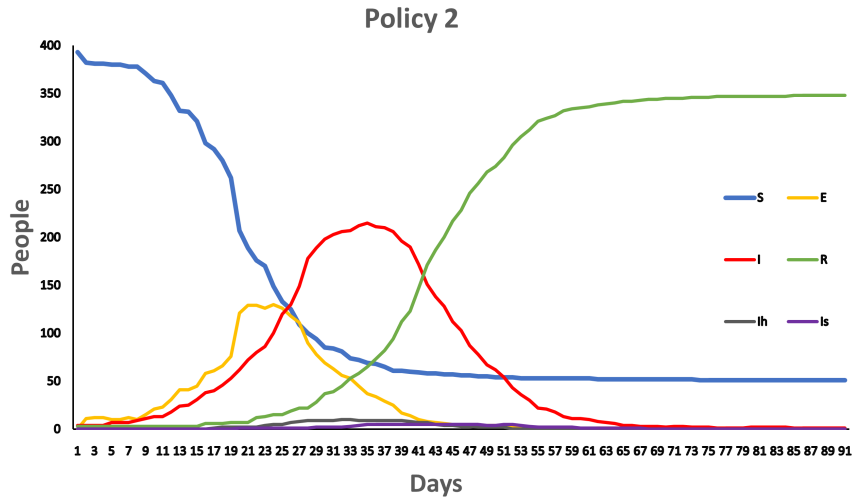

Figure 6.7: Results of epidemiological model when policy 2 is applied to the ABM.

This time 346 people get infected and the peak value of the infected people at the same time is 211. The peak number of hospitalized patients is 10 and the peak number of people in intense care units is 5, which is decreased compared

to the previous policies but still, it overpasses the desired threshold. There is still one infected person at the end of the simulation and 51 people remain susceptible. Financially, policy 2 has similar effects with policy 1, since nothing changes regarding the movement of human agents. As a result, "shopping" profits for businesses are similar to the respective in policy 1. Similarly, houses' expenses follow the same trend as policy 1.

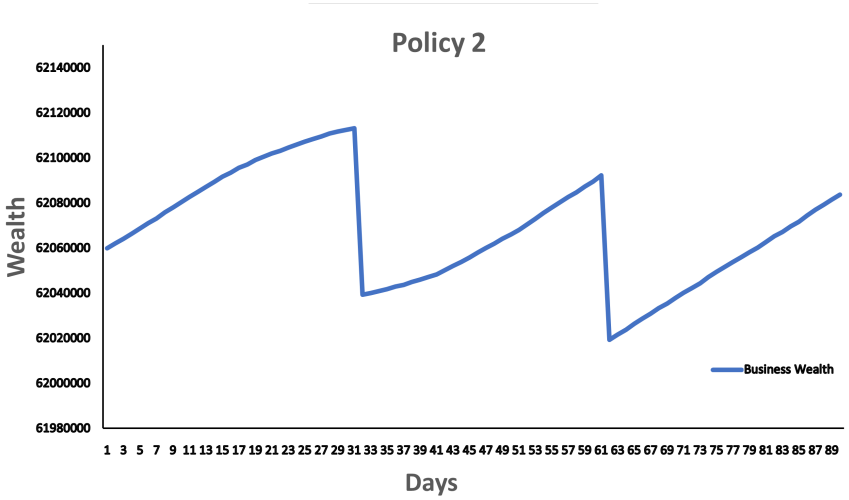

Figure 6.8: Economical effects of policy 2 to businesses.

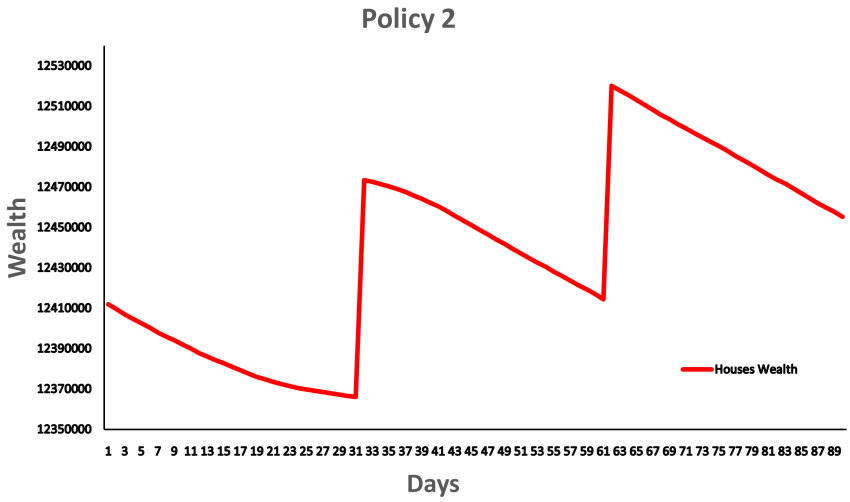

Figure 6.9: Economical effects of policy 2 to houses.

6.1.4 Policy 3

In policy 3, if more than 10 new infected cases are spotted in one day, lockdown and partial isolation is applied for 15 days. All people stay isolated at home for 15 days except the essential workers, who go to their workplaces as normal. The epidemiological results of this policy are depicted in Figure 6.10.

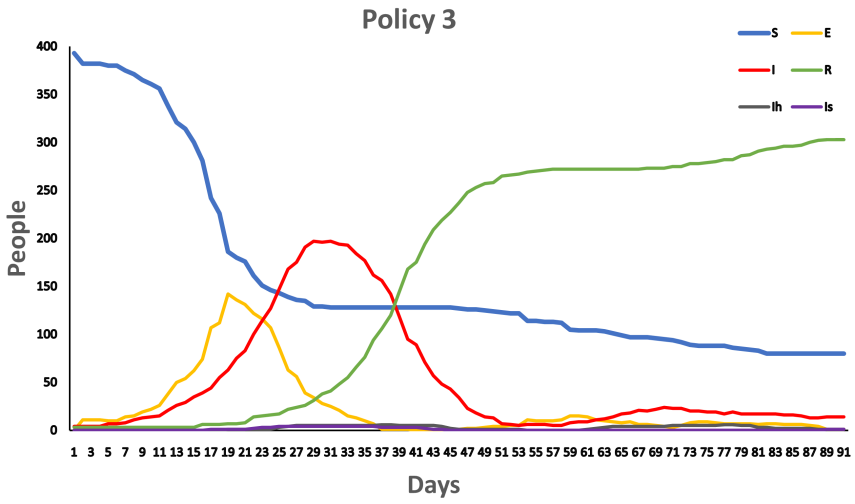

Figure 6.10: Results of epidemiological model when policy 3 is applied to the ABM.

In this scenario, 318 people get infected, which is less compared to the previous policies. The peak value of the infected cases is also decreased compared to the previous policies and equals to 197. Maximum hospitalized and severe cases are 6 and 4 respectively. At the end of the simulation, there are still 13 infected agents, while 79 people remain susceptible. On day 19, 11 more infected cases are observed compared to day 18. Thus the lockdown and the partial isolation is applied. As it can be observed also in Figure 6.10, after approximately 15 days the infected cases start to decrease while at the same time exposed people are less than 15. Some time after the lockdown period passes, a small increase of exposed and infected agents is observed, but soon the cases start to decrease again.

As for economical effects, businesses' wealth appears to increase, except during the lockdown period. During the lockdown period, people stay isolated at home, thus business agents have no "everyday" financial interaction with human agents, resulting in no profits. At the end of the simulation, the wealth of the businesses is approximately the same as it was at the beginning of the simulation. Regarding houses' wealth, it is expected to increase, since the people spend much less during the lockdown, considering that they don't move freely. Economical effects of policy 3 are depicted in Figures 6.11 and 6.12.

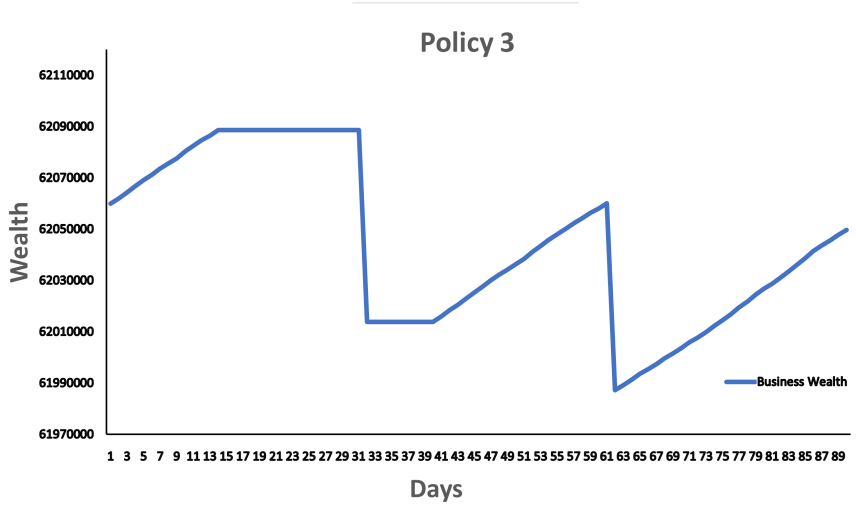

Figure 6.11: Economical effects of policy 3 to businesses.

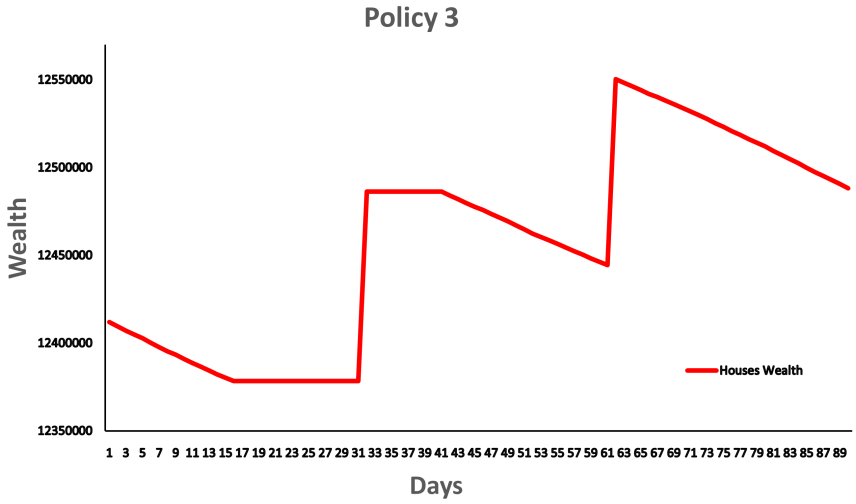

Figure 6.12: Economical effects of policy 3 to houses.

6.1.5 Policy 4

In policy 4, restrictions become more strict. Throughout the simulation, there is lockdown and partial isolation of the agents. All people stay at home, except the essential workers who go to their workplaces as normal. This policy is expected to reduce significantly the infected cases of the epidemiological model.

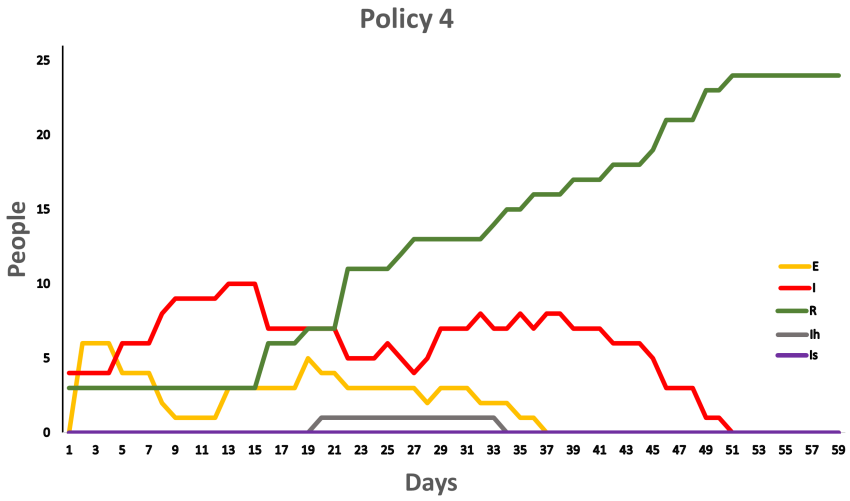

Figure 6.13: Results of epidemiological model when policy 4 is applied to the ABM.

Only 21 people get infected in this simulation and the peak of the infected curve is 10 cases. Only one agent needs to go to the hospital and none needs intensive care. Therefore the respective thresholds of hospitalized and severe cases are not exceeded. After 52 simulation days, there are no more infected people and 376 people remain susceptible until the end of the simulation. In figure 6.13 susceptible curve is skipped because it just shows a small change from 400 people to 376.

Although the epidemiological results of policy 4 are encouraging, the applied policy does not have positive effects on the economy. As mentioned, the biggest part of the population stays isolated at home. Furthermore, businesses are closed and shopping is not available neither for essential workers. This results in zero profit for businesses and causes a drop in the businesses wealth. Houses' wealth increases because people don't spend money on shopping.

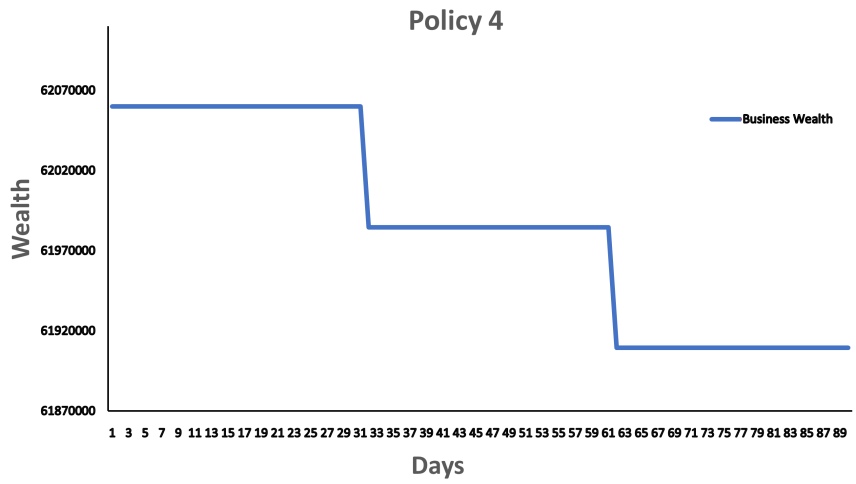

Figure 6.14: Economical effects of policy 4 to businesses.

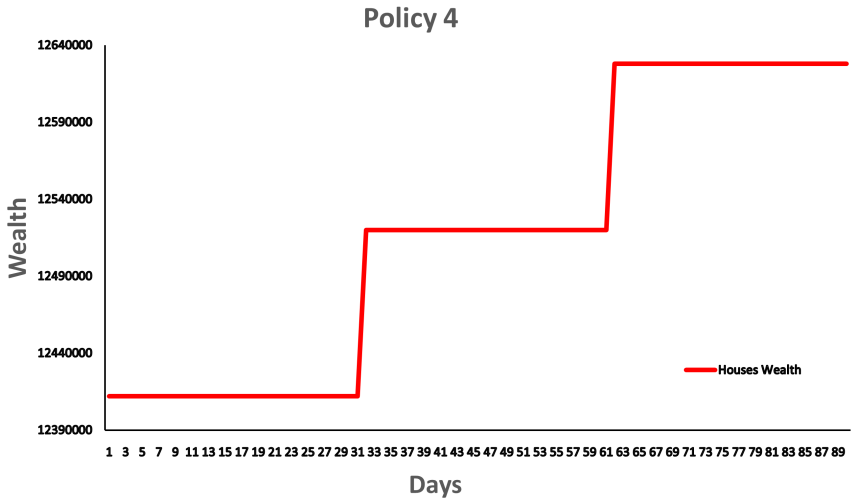

Figure 6.15: Economical effects of policy 4 to houses.

6.1.6 Policy 5

In policy, 5 total lockdown is applied. All the human agents are isolated at their homes The resulting graphs of the SEIHVR epidemiological model for this scenario can be seen in Figure 6.16.

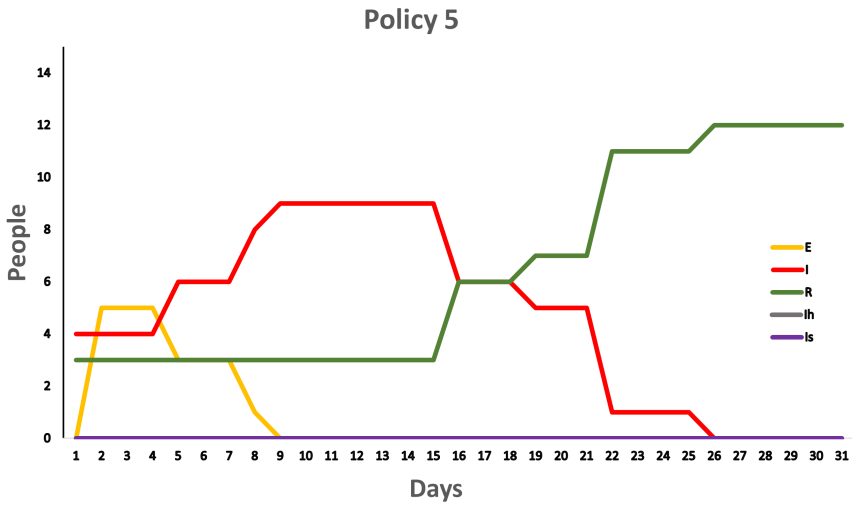

Figure 6.16: Results of epidemiological model when policy 5 is applied to the ABM.

In this scenario, 9 people get infected in total and the maximum number of infected cases is 9. None of the infected cases needs neither to get hospitalized nor to receive intensive care. After 26 days of simulation, the last infected person recovers, and 388 people remain susceptible until the end of the simulation.

From the financial point of view, policy 5 has the same negative effects on the economy as policy 4, since people do not move at all outside from their homes and businesses are closed for shopping.

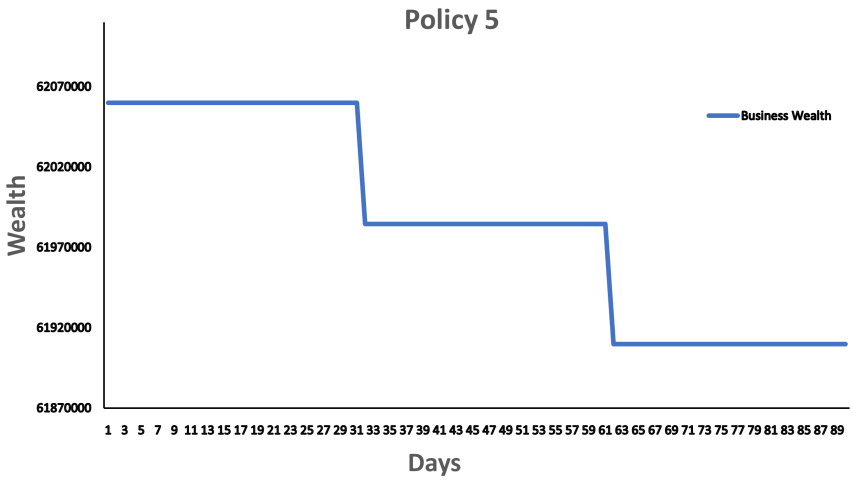

Figure 6.17: Economical effects of policy 5 to businesses.

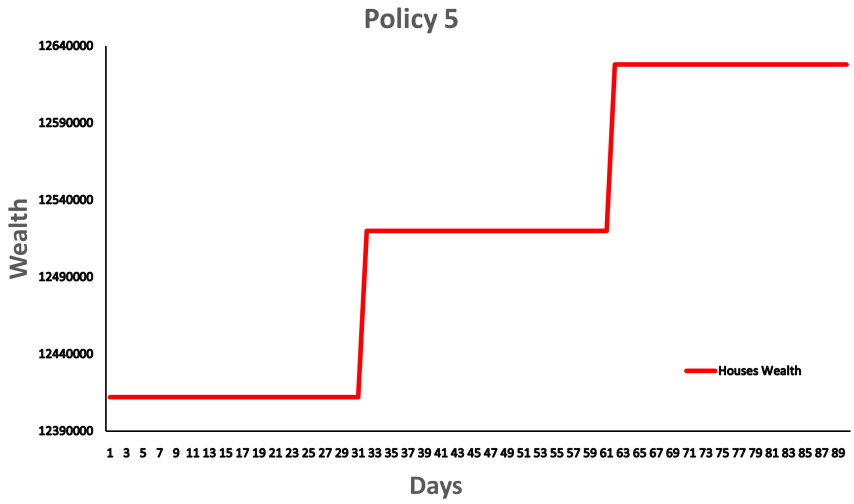

Figure 6.18: Economical effects of policy 5 to houses.

6.1.7 Policies Comparison

In this section, important attributes of the tested policies are compared each other, so that performance and efficiency of each policy can be evaluated.

Figure 6.19 demonstrates the infected curves from policies. An obvious conclusion is that the more strict the restrictions that are applied, the fewer people get infected. While in policy 0 the virus spreads quickly, in policies 4 and 5 the virus spread is limited and much fewer people get infected. Moreover, in policy 2 a lower infected curve peak is observed compared to the one of policy 1. In policy 2, the spread of the virus is also slower compared to policy 1 since infected people increase at a lower rate. Considering that in policies 1 and 2 the same restrictions apply, but in policy 2, face masks are obligatory, shows that the use of face masks can be an important asset to limit the coronavirus spread. The same happens at policy 3, where also the virus spreads slower during non-lockdown periods compared to policies 0 and 1. Also, the application of conditional lockdown decreases significantly both the people that get infected and the peak of the infected curve compared to policies 0, 1 and even 2.

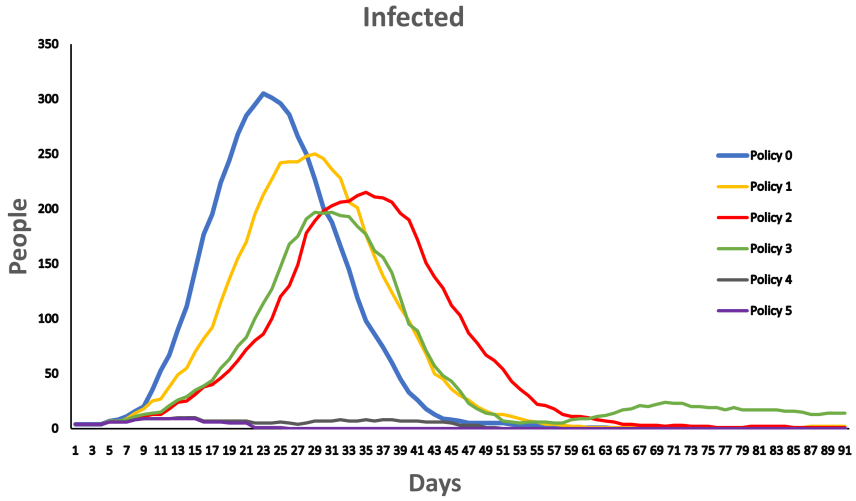

Figure 6.19: Comparison of infected curve between all policies.

In addition, hospitalized and severe cases must stay below the thresholds so that healthcare systems are capable to take care of them. In the current system, these thresholds are 8 and 4 patients respectively. In policies 0 and 1, both hospitalized and severe cases are higher enough compared to the respective thresholds. However, the cases present a significant drop as the restrictions become more strict. In policies 4 and 5 there are no patients that need to receive intensive care, and only one person needs to get hospitalized in policy 4. Also in policies 2 and 3, these numbers are close to the thresholds, which means that with a little more strict isolation rules for some people groups, these policies can also keep hospitalized and severe cases under the thresholds.

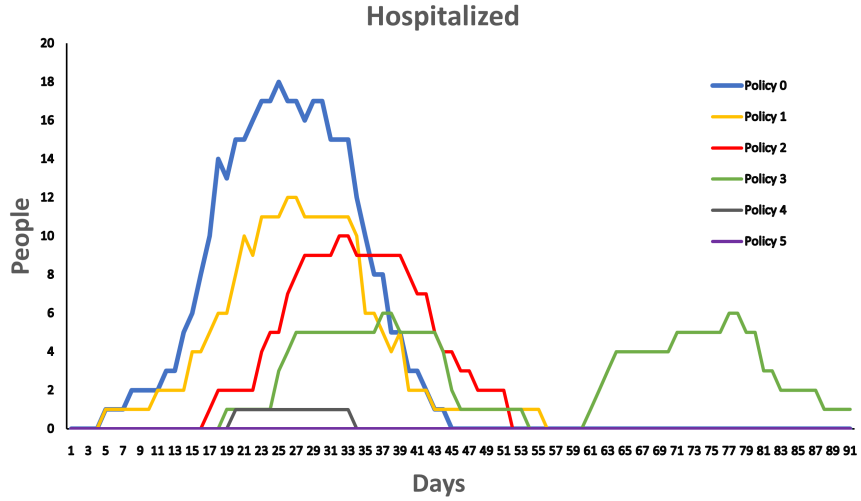

Figure 6.20: Comparison of hospitalized curve between all policies.

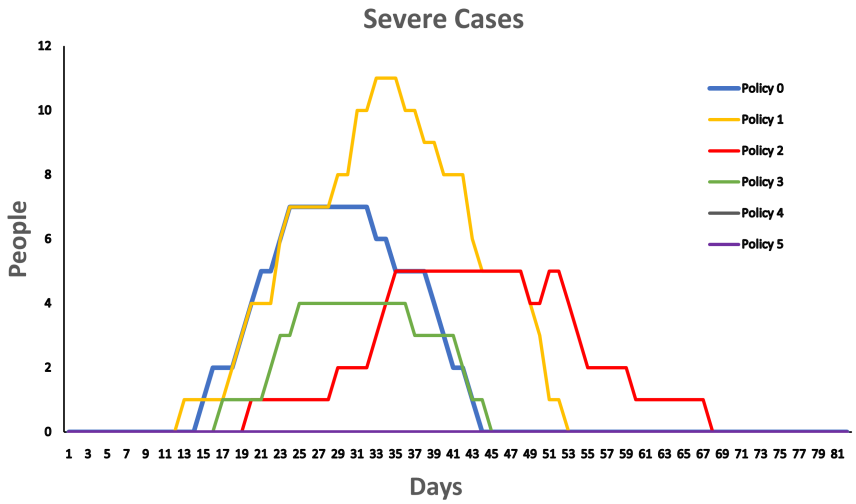

Figure 6.21: Comparison of severe cases curve between all policies.

Lastly, the efficiency of the policies regarding the economical sector must be evaluated. As expected, in policy 0 the profits of the businesses are larger compared to the respective ones of the remaining policies. Policies 4 and 5 may decrease the spread of the Covid-19 disease significantly, however strict lockdown causes big losses to the businesses and the economy. On the other hand, if policies with less strict lockdown rules are followed, like policies 1, 2 and 3, people can move freely at some point, interact with businesses, which continue to be functional and the economy can keep a balance between profits and expenses. Thus, an optimal sequence of policies to reduce the coronavirus spread should combine some strict lockdown policies to limit the virus at its peak, which are replaced in time by policies with more relaxed lockdown rules so that society economy receives the less possible financial "damage".

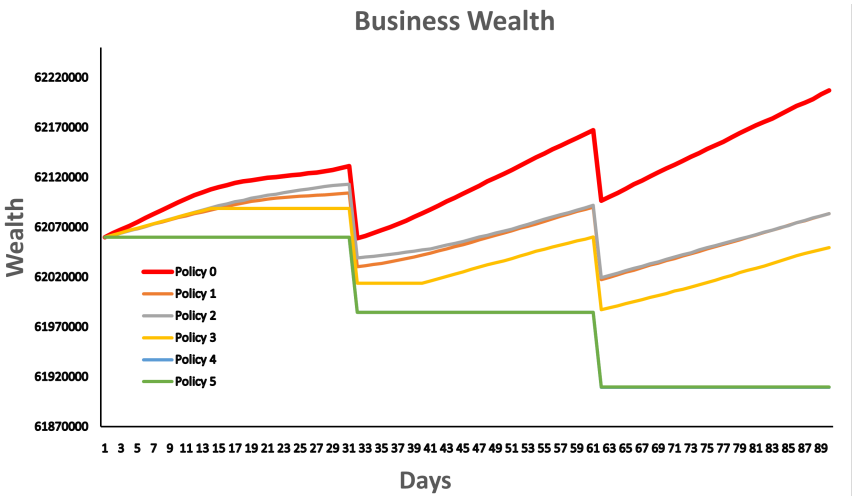

Figure 6.22: Comparison of economical effects of all policies to businesses.

6.2 Single workplace model: policies tests

The policies that are implemented in the single workplace model are also tested separately. The objective of these tests is to validate the expected results of each policy. Three different policies are implemented. In policies 0 and 1, it is expected that both susceptible agents will get infected. In contrast, in policy 2, it is expected that none of the susceptible agents will get infected due to the strict restriction of the mobility.

6.2.1 Policy 0

First, we will present the characteristics of the generated environment:

- Three human agents: two susceptible and one infected.
- A workplace which covers a grid of 3x3 cells.

The initial position of the human agents is a random cell in the workplace grid environment. The simulation includes a time period of 20 days.

In policy 0, at every timestamp, the agents move either to a neighbor cell or stay at the same cell. The probability to move is equal to 0.4 and the probability to stay at the same cell is 0.6. There is no use of face masks and contagion probability is 0.9. The resulting epidemiological graph for policy 0 is:

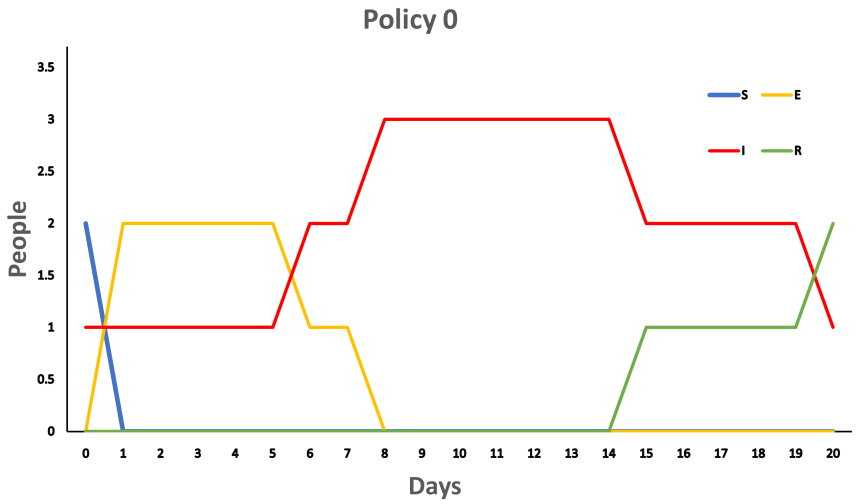

Figure 6.23: Results of epidemiological model when policy 0 is applied to the single workplace model.

As expected, both susceptible agents become exposed during the first day. Then incubation time of each exposed agent follows and they get infected after 5 and 6 days respectively. After 14 days, the initially infected agent recovers. Also one more agent recovers at the 20th day of the simulation while the other one is still infected when simulation time ends.

6.2.2 Policy 1

Similarly with policy 0, in policy 1, the agents move either to a neighbor cell or stay at the same cell. However the probability to move is equal to 0.05 and the probability to stay at the same cell is 0.95. The use of face masks is obligatory and contagion probability decreases to 0.3. The resulting epidemiological graph for policy 1 is:

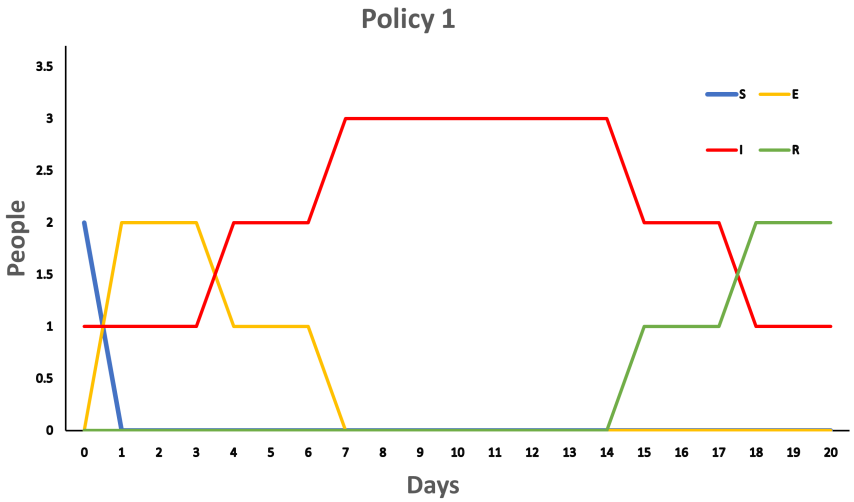

Figure 6.24: Results of epidemiological model when policy 1 is applied to the single workplace model.

Even though the probability of the agents to move and the contagion probability are decreased, the two susceptible agents get exposed again during the first day of the simulation. One of them has incubation time equal to 3 days and gets infected at 4th day and the other one has incubation time equal to 6 days and gets infected at the 7th day of the simulation. After 14 days of the simulation the initial infected agent recovers. Another recovers recovers 18th day, while the last one is still infected at the end of the simulation time.

6.2.3 Policy 2

In policy 2, the agents don't move at all. There is a possibility that either one susceptible agent or both of them, meet the infected agent after the initialization process, since their initial positions are random. In this situation, contagion probability is 0.3 because face masks use is obligatory. The resulting epidemiological graph corresponding to policy 2 is:

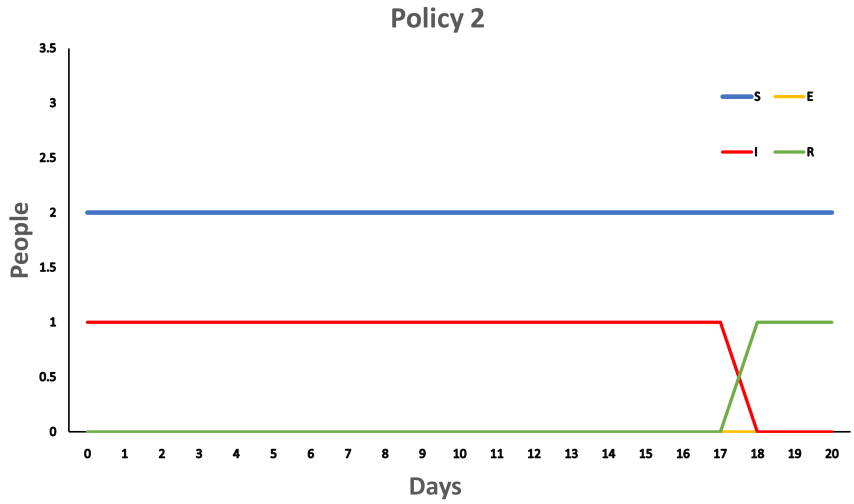

Figure 6.25: Results of epidemiological model when policy 2 is applied to the single workplace model.

As expected none of the susceptible agents get infected, during the simulation and the initially infected agent recovers after 17 days.

6.3 Reinforcement learning

In the second part of the tests, the reinforcement learning algorithm is applied to the ABM. The objective is to derive an optimal sequence of policies that will be followed during the 60 days simulation, so that the peak of the infected curve is minimized and the hospitalized and severe cases stay under control and do not exceed the respective thresholds. At the same time, the economy should not be negatively affected. The reward function in Equation 5.4 is responsible for this by giving the appropriate rewards to the respective transitions.

Also, a single workplace environment is designed and modelled with MDP. Then reinforcement learning is applied to it. This aims to prove the efficiency of the reinforcement learning algorithm.

6.3.1 Single workplace environment

In a single workplace environment, the objective is to minimize the number of infected people. Since three human agents are generated and one of them is initially infected, optimally none of the other two human agents will get infected. This can be achieved if the human agents do not move at all until the infected one recovers. As a result, a meeting between the infected and a susceptible agent will be avoided, no contagion will take place and the virus will not be transmitted to any suscepti-

ble person. A sequence of states that achieves the above scenario is illustrated by Figure 5.8.

As mentioned three possible policies can be followed in the single workplace environment simulation. In policy 0, the human agents move freely in the workplace and the economical reward is the maximum possible. In policy 1, the agents move with a much smaller probability in the workplace and the economical reward is decreased. In policy 2, people do not move at all and the economical reward becomes negative. To guarantee that the states sequence will be similar as the desired one of Figure 5.8, policy 2 should be followed until the infected person recovers and any virus transmission is avoided. Then, policy 0 should be followed so that the economy takes some boost. This is the policy sequence that is expected to be derived from the application of the reinforcement learning algorithm.

Reinforcement learning is applied to the single workplace environment simulation. During the learning process, the simulation is executed 16622 times. At every iteration, the program creates a 3x3 environment with randomly placed human agents inside it, as described before. A table with Q-values is constructed, where rows represent days of simulation (20 days in total) and columns represent policies. The goal is to find an optimal policy for each day. The policy with the highest quality value is the optimal policy. The algorithm converges after 16622 runs and the following Q-table is obtained:

Simulation Time (Days)	Policy 0	Policy 1	Policy 2	Max Q Value	Selected Policy
1	-23.9995	-21.9982	-19.9995	-19.9995	**Policy 2**
2	-24.0465	-21.3868	-20.0005	-20.0005	**Policy 2**
3	-24.6436	-27.351	-20.0005	-20.0005	**Policy 2**
4	-24.1782	-26.2334	-20.0005	-20.0005	**Policy 2**
5	-22.8806	-28.2268	-20.0005	-20.0005	**Policy 2**
6	-24.7911	-27.7731	-20.0005	-20.0005	**Policy 2**
7	-22.6855	-23.2074	-20.0005	-20.0005	**Policy 2**
8	-31.94	-35.4216	-20.0005	-20.0005	**Policy 2**
9	-27.004	-27.9307	-20.0005	-20.0005	**Policy 2**
10	-20.0488	-21.3511	-19.0005	-19.0005	**Policy 2**
11	-22.0403	-23.4169	-20.0005	-20.0005	**Policy 2**
12	-20.2864	-22.052	-19.9995	-19.9995	**Policy 2**
13	-21.5591	-24.4666	-19.9995	-19.9995	**Policy 2**
14	-22.6265	-23.5764	-19.9995	-19.9995	**Policy 2**
15	-17.3417	-17.3331	-12.6909	-12.6909	**Policy 2**
16	-20.2959	-18.0952	-12.6909	-12.6909	**Policy 2**
17	-20.6858	-17.6828	-12.6909	-12.6909	**Policy 2**
18	24.9557	1.33282	-0.445877	24.9557	**Policy 0**
19	28.3186	-0.954001	-0.900122	28.3186	**Policy 0**
20	28.3186	6.59329	4.50645	28.3186	**Policy 0**

Table 6.1: Q-table derived from reinforcement learning algorithm for the single workplace environment simulation.

The policies sequence that reinforcement learning "suggests" is to use policy 2 until the 17th day and then, for the rest days, policy 0. Following policy 2 for 17 days means that the agents don't move for this time period. Therefore a possible meeting between agents is avoided, no contagions take place and no one gets infected. Furthermore, considering that the recovery time of all infected people is either 14 or 17 days, it is concluded that the initial infected person will be recovered by day 17. Therefore, from day 18 there is no infected agent and the system can go back to policy 0, which also gives a boost to the economy, since policy 0 gives the maximum economical reward.

In other words, the obtained policy sequence from the reinforcement learning corresponds with the expected one. If this sequence of policies is followed, the states sequence is similar to the one in Figure 5.8. This validates the results of the reinforcement learning for the general ABM and is considered as proof that the algorithm will ultimately result in a sequence of policies that will indeed limit the spread of coronavirus and affect the economy in a less harmful way.

6.3.2 Second illustrative model

To ensure that the reinforcement learning algorithm can be applied to other, more complicated models, another illustrative model was prepared. It consists of 10 by 10 gird-like with 20 people, 4 houses and 2 workplaces. Initially, 1 person is infected and the rest is susceptible. The policies are the same as in the previous illustrative model. In learning part, the algorithm was executed 18143 times.

The result in form of Q-table is:

Simulation Time (Days)	Policy 0	Policy 1	Policy 2	Max Q Value	Selected Policy
1	-51.9995	-53.9995	-19.9995	-19.9995	**Policy 2**
2	-28.1514	-30.7237	-20.0005	-20.0005	**Policy 2**
3	-50.9098	-59.8723	-20.0005	-20.0005	**Policy 2**
4	-39.8503	-42.5356	-20.0005	-20.0005	**Policy 2**
5	-73.1306	-36.4568	-20.0005	-20.0005	**Policy 2**
6	-65.1482	-59.4451	-20.0005	-20.0005	**Policy 2**
7	-53.0485	-40.2816	-20.0005	-20.0005	**Policy 2**
8	-72.1525	-84.404	-20.0005	-20.0005	**Policy 2**
9	-61.7591	-40.657	-20.0005	-20.0005	**Policy 2**
10	-86.2928	-45.9561	-20.0005	-20.0005	**Policy 2**
11	-37.7446	-65.6725	-20.0005	-20.0005	**Policy 2**
12	-89.7422	-91.68	-20.0005	-20.0005	**Policy 2**
13	-92.1216	-93.9579	-20.0005	-20.0005	**Policy 2**
14	-95.0651	-90.0961	-20.0005	-20.0005	**Policy 2**
15	-85.538	-51.3401	-14.4827	-14.4827	**Policy 2**
16	-63.9354	-83.6107	-14.4827	-14.4827	**Policy 2**
17	-92.9957	-81.9205	-14.4827	-14.4827	**Policy 2**
18	29.9995	12.471	10.4634	29.9995	**Policy 0**
19	29.9995	6.72061	2.96392	29.9995	**Policy 0**
20	29.9995	14.1925	12.1882	29.9995	**Policy 0**
21	29.9995	10.5549	8.75447	29.9995	**Policy 0**
22	29.9995	3.88991	4.47755	29.9995	**Policy 0**
23	29.9995	10.3956	7.35431	29.9995	**Policy 0**
24	29.9995	16.2028	14.408	29.9995	**Policy 0**
25	29.9995	6.4759	2.61678	29.9995	**Policy 0**
26	29.9995	17.432	16.3706	29.9995	**Policy 0**
27	29.9995	16.4803	13.4631	29.9995	**Policy 0**
28	29.9995	16.2892	11.473	29.9995	**Policy 0**
29	29.9995	19.3394	18.2862	29.9995	**Policy 0**
30	29.9995	20.2181	17.787	29.9995	**Policy 0**

Table 6.2: Q-table derived from reinforcement learning algorithm for the second illustrative simulation.

The results indicate that until day 17, policy 2 should be applied. After that, policy 0 should be applied. This makes sense since the infected person recovers after 14 or 17 days. There could be a situation when the infected is initialized at the same cell with susceptible. In that case, it will take more time to dispose of the virus. But this situation would occur rarely so it did not influence the result.

6.3.3 Simulation with optimal policies

The following section contains the simulation of the illustrative models with optimal policies. The results are presented at Figures 6.26, 6.27, 6.28 and 6.29.

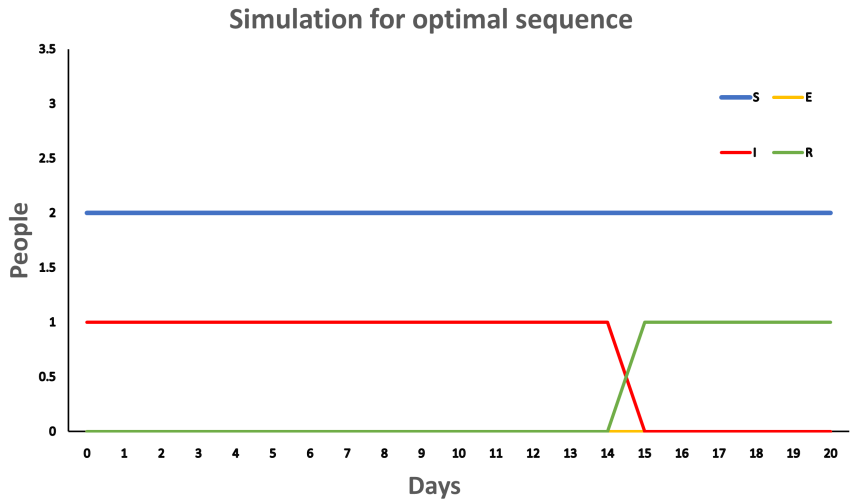

Figure 6.26: Results of epidemiological model when the optimal sequence of polices is applied to the single workplace model. The infected agent recovers after 14 days.

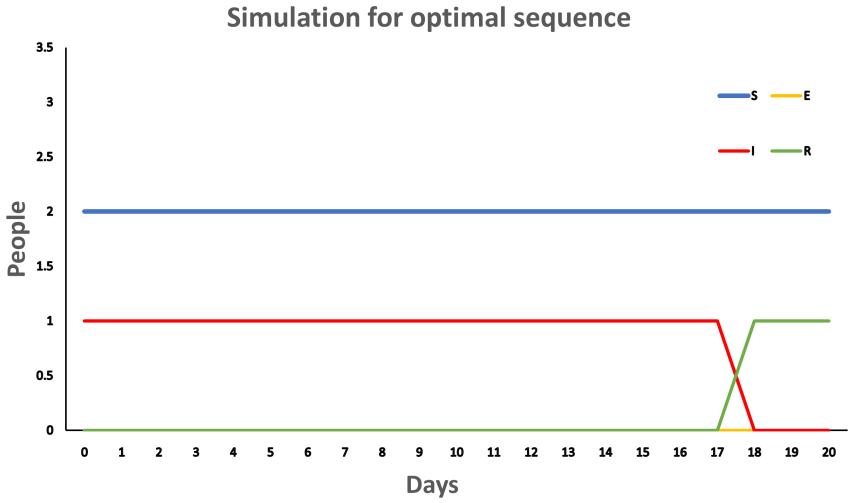

Figure 6.27: Results of epidemiological model when the optimal sequence of polices is applied to the single workplace model. The infected agent recovers after 17 days.

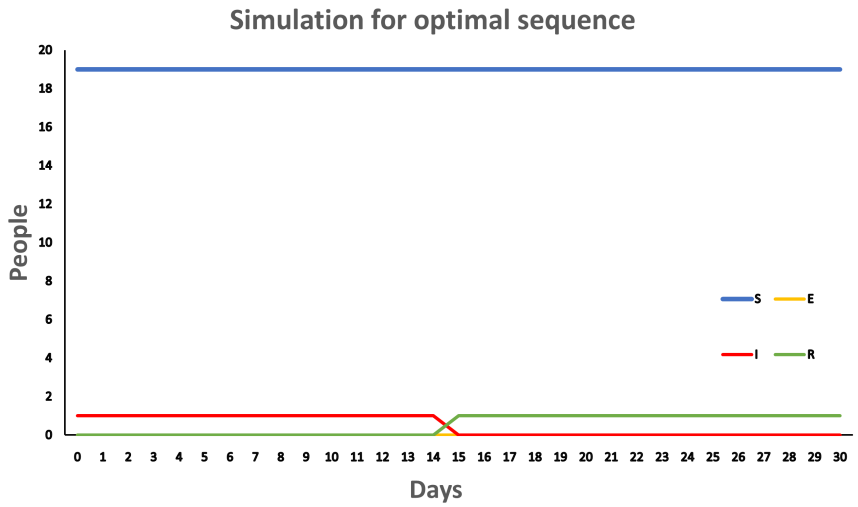

Figure 6.28: Results of epidemiological model when the optimal sequence of polices is applied to the second illustrative model. The infected agent recovers after 14 days.

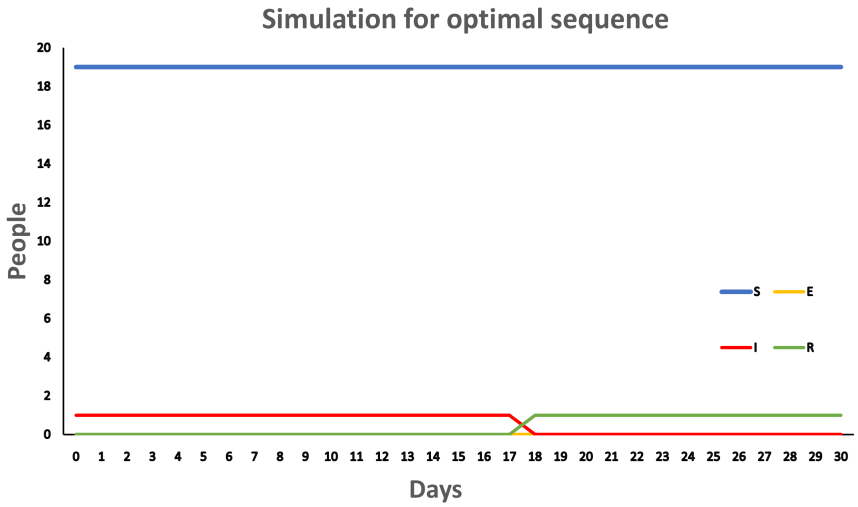

Figure 6.29: Results of epidemiological model when the optimal sequence of polices is applied to the second illustrative model. The infected agent recovers after 17 days.

From the graphs, it can be concluded, that the virus is disposed after either 14 or 17 days. These are recovery times and the value depend on the age of the infected individual. During the first 17 days, the algorithm suggested applying policy 2. This policy restricts the movement to none. Thus, the infected cannot spread the infection to other agents. After 17 days, when the virus is disposed, the algorithm suggested applying policy 0. This is desired, as policy 0 has the most economical benefits. Thus, it can be concluded that the RL algorithm works correctly and as expected.

Chapter 7

Conclusions and Further Development

The present thesis aimed to develop an Agent-Based Model and use Reinforcement Learning to control it. The ABM was supposed to simulate the society and the spread of Covid-19. It included a grid-like environment with 24h day cycle in which agents (humans) move around and perform actions. The pandemic was simulated using the SEIHVR model. On top of that, economic processes were added to influence the reinforcement learning algorithm's operation.

To illustrate our results more clearly, we presented a simulation model with a fewer number of agents and workplaces. The objective of it was to test and validate the reinforcement learning algorithm. This simulation consisted of a grid-like, 3 by 3 environment with 3 agents (people). Initially, one of them was infected and the remaining ones were susceptible. The agents were able to move in the environment with probabilities dictated by the applied policy. In this simulation, the economy did not play a significant role. The epidemiological states were emphasized to check if the reinforcement learning algorithm will be able to stop the infection by applying certain policies.

Indeed, we expected the algorithm to apply the policy in which the agent's moves are disabled. This would allow the infected person to go through the infection without spreading it to other agents. Moreover, when the infected agent has recovered, the algorithm changed the optimal policy to policy with movement and the biggest economic benefits which is also expected and desirable. The expected optimal set of policies for an illustrative model was extracted by the algorithm after around 16000 iterations. During 1 iteration the simulation period was 20 days in which 120 actions per day were performed.

We designed an MDP for the illustrative model. The states described agents positions together with their epidemiological groups. MDP also described sets of possible actions. The reward functions for state transitions were also specified.

We believe, that the states designed for MDP for the illustrative model can be scaled to an initial model. To prove this, we designed another illustrative model with 10 by 10 environment and 20 agents. We applied reinforcement learning with the same reward as presented in the single workplace model. The algorithm converged to the expected result after around 20000 iterations.

For the initial model, we expect the reinforcement learning, like in illustrative models, to learn the set of optimal policies as well. It has to be noted that for this model, the reward function needs to be adjusted to take into account the economy and the medical care thresholds. The MDP for the model needs an additional description of economic agents to fully describe the dynamics of the system.

An agent-based model like every other model does not fully reflect real-life conditions. Thus, the model can be improved in many ways. The simulation already takes a good amount of time to run therefore the functionality needs to be limited. To properly mirror the real-life conditions, a wider variety of factors needs to be taken into account such as vital dynamics, vaccinations, immunity or lack of it in the recovered compartment. Because of performance reasons, the epidemiological model was kept simple but efficient to reflect Covid-19 dynamics. The main focus of the project was to keep the hospitalized and severe cases below the medical care threshold. Thus, the SEIHVR model used in the project includes basic epidemiological compartments needed to simulate the Covid-19 spread as well as mentioned hospitalized and severe compartments. Random movements of people do not fully reflect real-life conditions. However, they are efficient to represent the spread of the virus in a similar way as in real life.

For future work, we propose to simulate the initial model and adjust the reward function according to the results of performed simulations. We believe that the reinforcement learning controller should work with the initial model, and the simulation time and the number of iterations will be larger than in the illustrative models. Some optimizations and updates can be made to the model as well to make it more accurate and to reduce simulation time. Additionally, few tests could be conducted to check how the reinforcement learning reacts to fluctuations in the model. One of them could include the change of transition probabilities between groups. Another one could include changes in the model, for example adding vaccinations or interference from outside the society like travels (people coming and leaving with or without the virus).

Bibliography

[1] Yismaw L. Assemie M.A. Alene M. "Serial interval and incubation period of COVID-19: a systematic review and meta-analysis". In: *BMC Infect Dis* 21 (2021). DOI: https://doi.org/10.1186/s12879-021-05950-x.

[2] *An animated map tracks the spread of the coronavirus as cases were reported in more than 180 countries*. https://www.businessinsider.com/map-tracks-novel-coronavirus-spread-in-countries-around-the-world-2020-3?r=US&IR=T. Accessed: 2021-5-28.

[3] Tsakris A Siettos C Anastassopoulou C Russo L. *Data-based analysis, modelling and forecasting of the COVID-19 outbreak*. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0230405.

[4] Eric Bonabeau. "Agent-based modeling: Methods and techniques for simulating human systems". In: *Proceedings of the National Academy of Sciences* 99.suppl 3 (2002), pp. 7280–7287. ISSN: 0027-8424. DOI: 10.1073/pnas.082080899. eprint: https://www.pnas.org/content/99/suppl_3/7280.full.pdf. URL: https://www.pnas.org/content/99/suppl_3/7280.

[5] Shayn M. Peirce Bryan C. Thorne Alexander M. Bailey. "Combining experiments with multi-cell agent-based modeling to study biological tissue patterning". In: *Briefings in Bioinformatics* 8 (2007), 245–257. DOI: https://doi.org/10.1093/bib/bbm024.

[6] Francesco Sannino Corentin Cot Giacomo Cacciapaglia. *Mining Google and Apple mobility data: Temporal Anatomy for COVID-19 Social Distancing*. https://arxiv.org/pdf/2008.02117.pdf.

[7] Bernardo Furtado and Isaque Eberhardt. "A simple agent-based spatial model of the economy: tools for policy". In: (Oct. 2015). URL: https://www.researchgate.net/publication/282975962_A_simple_agent-based_spatial_model_of_the_economy_tools_for_policy.

[8] Marino Gatto et al. "Spread and dynamics of the COVID-19 epidemic in Italy: Effects of emergency containment measures". In: *Proceedings of the National Academy of Sciences* 117.19 (2020), pp. 10484–10491. ISSN: 0027-8424. DOI: 10.1073/pnas.2004978117. eprint: https://www.pnas.org/content/117/19/10484.full.pdf. URL: https://www.pnas.org/content/117/19/10484.

[9] Blanchini F. Bruno R. Giordano G. "Modelling the COVID-19 epidemic and implementation of population-wide interventions in Italy." In: *Nat Med* 26 (2020), 855–860. DOI: https://doi.org/10.1038/s41591-020-0883-7.

[10] U.S. Department of Health Human Services. *Smallpox*. https://www.cdc.gov/smallpox/.

[11] Herbert W. Hethcote. "The Mathematics of Infectious Diseases." In: *SIAM Review* 42.4 (2000), 599–653. DOI: www.jstor.org/stable/2653135.

[12] *Homelessness In A Welfare State: Perspectives From Copenhagen*. https://www.humanityinaction.org/knowledge_detail/homelessness-in-a-welfare-state-perspectives-from-copenhagen/. Accessed: 2021-5-26.

[13] Sheel M. Housen T. Parry A.E. *How long are you infectious when you have coronavirus?* https://theconversation.com/how-long-are-you-infectious-when-you-have-coronavirus-135295.

[14] Jeremy Howard et al. "An evidence review of face masks against COVID-19". In: *Proceedings of the National Academy of Sciences* 118.4 (2021). ISSN: 0027-8424. DOI: 10.1073/pnas.2014564118. eprint: https://www.pnas.org/content/118/4/e2014564118.full.pdf. URL: https://www.pnas.org/content/118/4/e2014564118.

[15] *Impact of non-pharmaceutical interventions (NPIs) to reduce COVID19 mortality and healthcare demand*. https://spiral.imperial.ac.uk/bitstream/10044/1/77482/14/2020-03-16-COVID19-Report-9.pdf.

[16] Dunham JB. "An Agent-Based Spatially Explicit Epidemiological Model in MASON." In: *Journal of Artificial Societies and Social Simulation.* 9 (2005). DOI: http://jasss.soc.surrey.ac.uk/9/1/3.html.

[17] Bart De Schutter Damien Ernst Lucian Busoniu Robert Babuska. *Reinforcement Learning and Dynamic Programming Using Function Approximators*. CRC Press, 2010. ISBN: 978-1-4398-2109-1.

[18] *Nuts Bolts of Reinforcement Learning: Model Based Planning using Dynamic Programming*. https://www.analyticsvidhya.com/blog/2018/09/reinforcement-learning-model-based-planning-dynamic-programming/.

[19] Mridha M.F. Monowar M.M. et al. Ohi A.Q. "Exploring optimal control of epidemic spread using reinforcement learning." In: *Sci Rep 10* 22106 (2020). URL: https://doi.org/10.1038/s41598-020-79147-8.

[20] Aaron O'Neill. *Unemployment rate in Denmark 2020*. https://www.statista.com/statistics/318316/unemployment-rate-in-denmark/.

[21] World Health Organization. *Vaccines and immunization: What is vaccination?* https://www.shorturl.at/uwxBG.

[22] Dragicevic S. Perez L. "An agent-based approach for modeling dynamics of contagious disease spread." In: *Int J Health Geogr.* 8 (2009), 245–257. DOI: 10.1186/1476-072X-8-50.

[23] Hélder S. Lima Marcos A. Alves Frederico G. Guimarães Rodrigo C.P. Silva Petrônio C.L. Silva Paulo V.C. Batista. "COVID-ABS: An agent-based model of COVID-19 epidemic to simulate health and economic effects of social distancing interventions". In: *Elsevier* 139 (2020). DOI: https://www.sciencedirect.com/science/article/abs/pii/S0960077920304859?via%3Dihub.

[24] ECDC TECHNICAL REPORT. *Disinfection of environments in healthcare and nonhealthcare settings potentially contaminated with SARS-CoV-2*. https://www.ecdc.europa.eu/sites/default/files/documents/Environmental-persistence-of-SARS_CoV_2-virus-Options-for-cleaning2020-03-26_0.pdf.

[25] *STATISTICS DENMARK*. https://www.dst.dk/en. Accessed: 2021-5-26.

[26] *WHO Coronavirus disease information.* https://www.who.int/health-topics/coronavirus#tab=tab_1.

[27] *WHO global literature on coronavirus disease.* https://search.bvsalud.org/global-literature-on-novel-coronavirus-2019-ncov/.

List of Figures

List of Tables