ISSN 2395-8618

Computing Optimized Epidemic Control Policies Using Reinforcement Learning

Jose Fernando Canto-Olvera, Uriel Corona-Bermúdez, Rolando Menchaca-Méndez, Ricardo Menchaca-Méndez

Abstract—**This paper presents a reinforcement learning-based algorithm for computing epidemic contention policies defined in terms of mobility-restriction actions. The algorithm's objective is simultaneously minimizing public health and economic affectations, which is challenging because both objectives are in conflict. We used a SEIRD (*Susceptible-Exposed-Infected-Recovered-Deceased*) epidemiological model to capture the spreading dynamics of a disease characterized by the probabilities of transitioning between the states defined in the model. To train the reinforcement learning algorithm, we implemented a discrete event simulator from scratch that considers different mobility patterns and diseases defined in terms of the SEIRD model probabilities. Extensive simulation-based results show that the proposed algorithm computes mobility restriction policies that effectively minimize the two opposite objectives and are flexible enough to allow a decision-maker to prioritize either public health or the economy.**

Index Terms—**Epidemic control, compartmental models in epidemiology, reinforcement learning.**

I. Introduction

As new infectious diseases like COVID-19 and influenza have emerged, policies have been implemented to prevent widespread contagion. These policies include wearing face masks and practicing complete isolation [9]. These policies aim to prevent large outbreaks of illness, which can negatively impact the economy and health of an entire population [1]. The effects of such outbreaks can be felt at different levels, from economic recessions and mass layoffs to overcrowded hospitals and loss of lives.

While there is a well-established epidemiological methodology that involves observation, measurement, comparison, and proposal [6]; applying these methods is sometimes challenging. This is because the spread or mutation of a disease can happen so rapidly that it's impossible to conduct a precise analysis, and consequently, taking effective actions becomes a difficult task. Fortunately, technological advancements in computer science have made it possible to simulate the behavior of populations in a short amount of time. In the context of epidemiological diseases, this allows us to estimate how the disease might spread, its potential impact, and to compute preventive policies that could be taken to minimize its effects.

Compartment models are the most popular mathematical models used to simulate infectious disease behavior. These models are typically run using ordinary differential equations, which are deterministic, but they can also be used with a stochastic framework to provide more realistic results, although the analysis is much more complicated [5]. The most well-known epidemic models are the **S**usceptible-**I**nfected–**S**usceptible (SIS) and **S**usceptible–**I**nfected–**R**ecovered (SIR) models. Another model gaining increasing interest is the **S**usceptible–**E**xposed–**I**nfected–**R**ecovered–**S**usceptible (SEIRS) [8]. Although some models include vital dynamics, like births and deaths under normal circumstances [2], recent events have shown that the spread of infection can cause many deaths to arise. A compartment model introduced in 2020 considering this is the **S**usceptible–**E**xposed–**I**nfected–**R**ecovered-**D**eceased (SEIRD) [7] model, which was used along reinforcement learning techniques to compute optimal controls for epidemic spreading.

In this article, we propose a framework using simple population behavior simulations under two common scenarios impacting infectious disease spreading through a SEIRD compartment model. From this simulation-based environment, we use reinforcement learning to compute optimized mobility restriction policies aimed at minimizing the negative impacts of an epidemic.

The remainder of this paper is structured as follows: Section II includes related work about the spreading and contention dynamics of infectious diseases. We introduce the basic compartmental models used for modeling epidemic diseases in Section III. The proposed framework and the experiments are described in Section V and VI respectively. Lastly, we include a brief discussion, our conclusions, and future work in Section VII.

II. Related Work

In 2020, following the emergence of COVID-19, there was a proposal to explore the use of reinforcement learning for computing policies of mobility restrictions without any prior knowledge [7]. The initial premise was that the impact of the pandemic could be reduced by restricting the movement of the masses, for example, an economic crisis. To simulate the masses' movements, they used a 2-dimensional grid where fixed random moves were performed daily. They used the SEIRD model to replicate the spread of the infection. Their action space contains three movement restrictions: 0%,

Manuscript received on 19/05/2023, accepted for publication on 15/09/2023.

The authors are with the Instituto Politécnico Nacional, Centro de Investigación en Computación, Mexico City, Mexico (fernando_canto@gmail.com, ucoronab@ipn.mx, rmen@cic.ipn.mx, ric@cic.ipn.mx).

https://doi.org/10.17562/PB-66(1)-8

25%, and 75%. Their states were formed by the percentage of active cases, the percentage of newly infected cases, the cumulative percentage of cured cases, the cumulative percentage of deceased cases, the reproduction rate, the daily economic contributions of the population, and the current movement restriction level. The reward function was defined in terms of three environmental parameters: the current economy ratio, cumulative death ratio, and the current percentage of active cases.

One of the study's major limitations is that several modeling assumptions were based on the early stages of the COVID-19 pandemic. Additionally, the researchers employed Deep Reinforcement Learning (DRL) techniques, but since the states and actions were discrete, tabular techniques could have been utilized for optimization purposes. Finally, the study is limited to modeling COVID-19, but the framework applies to many infectious diseases.

In 2022, a framework for controlling infectious diseases was proposed to help make data-driven decisions and reduce long-term costs [11]. The authors didn't propose a population interaction model because all information was obtained from official sources. The framework uses a generalized SIR model as the spreading disease model and applies a model-based, multi-objective planning algorithm to identify a set of Pareto-optimal policies, which cannot be improved for one objective without sacrificing another at each decision point. By combining this framework with prediction bands for each policy, policymakers have a real-time decision-support tool. The framework was applied to the spread of COVID-19 in China. The experiments conducted by the authors focused on six regions in China. The environment state was defined by its annual gross domestic product (GDP), population, and the number of confirmed COVID-19 cases. The state was defined per region. They used three levels of movement restriction as their actions: level 1 indicated no or few official policies, level 2 meant a public health emergency response, and Level 3 was stringent closed-off management required by the government. The cost function was sampled mobility ratio for each restriction level, scaled by the daily GDP.

The framework is highly robust and can be applied to various infectious diseases. However, since the main proposal was for a particular case study, and the population interaction was obtained from official sources instead of a simulation, we do not know the decisions that could be made under certain conditions. For instance, we do not know what decisions would be made if the cost function included the mortality rate.

III. Analytical Models for Epidemic Processes

In this section, we introduce the most common compartment models used to study the dynamics of infectious diseases. They range from basic deterministic models expressed as ordinary differential equations to more advanced stochastic models defined as Markovian processes.

A. Deterministic Compartmental SIS Model

The simplest compartment model for infectious diseases is the SIS model, first introduced by Kermack and McKendrick [4]. The model assumes two transitions between the compartments: infection (from S to I) and recovery (from I to S). The infection rate is assumed to be proportional to the sizes of the S and I compartments, while the rate of recovery is assumed to be proportional to the size of the infected compartment. This model is given by Eqs. 1:

$$\begin{aligned}\frac{dS(t)}{dt} &= -\theta I(t)\frac{S(t)}{N} + \delta I(t),\\ \frac{dI(t)}{dt} &= \theta I(t)\frac{S(t)}{N} - \delta I(t),\end{aligned} \tag{1}$$

where $S(t)$ and $I(t)$ denote the size of the susceptible and infected compartments at time t, $N = S(t) + I(t)$ is the population size, and θ and δ are positive constants called infection and recovery rates, respectively. It may be helpful to think of θ as the rate at which infected individuals make infection-transmitting contacts. Then, the total rate of infectious contacts is θI, but only a fraction S/N of these are susceptible individuals and thus lead to a new infection [5].

B. Stochastic Compartmental SIS Model

The starting point of many epidemic models is a stochastic formulation, which involves an implicit connectivity network [5]. For the SIS model, $S(t)$ and $I(t) = N - S(t)$ are random variables taking values from the set $\{0, 1, \cdots, N\}$. Some assumptions are required:

1) the network of contacts is fully connected,
2) infection is transmitted across a link between a susceptible and an infected individual at rate α, so αN corresponds to θ;
3) each infected individual recovers at rate δ independently of all others and of the network; and
4) both processes are Markovian.

Treating this process as a continuous-time Markov chain and given the state of the system at time t, $(S, I)(t)$, the following two transitions are possible:

$$\begin{aligned}(S, I) &\xrightarrow{\alpha SI} (S-1, I+1),\\ (S, I) &\xrightarrow{\delta I} (S+1, I-1),\end{aligned} \tag{2}$$

where the rates encode the transmission and recovery processes. Whether the next event is an infection or recovery is determined at random but relative to the magnitude of the two rates.

C. Stochastic Compartmental SEIR and SEIRS Models

The SEIR [3] model is another variation of the disease spread model. It includes an additional *exposed* compartment

ISSN 2395-8618

between the susceptible and infected compartments. The model is described as follows:

$$
\begin{aligned}
(S,E,I,R) &\xrightarrow{\alpha SI} (S-1,E+1,I,R),\\
(S,E,I,R) &\xrightarrow{\beta E} (S,E-1,I+1,R),\\
(S,E,I,R) &\xrightarrow{\delta I} (S,E,I-1,R+1).
\end{aligned} \tag{3}
$$

Here, β represents the infection latency rate. Similar to the SIRS model, the SEIR model also includes a transition from the recovered compartment back to the susceptible compartment, which can be modeled as:

$$(S,E,I,R) \xrightarrow{\epsilon R} (S+1,E,I,R-1). \tag{4}$$

D. Stochastic Compartmental SEIRD Model

The SEIRD model [7] is the latest addition to the models used to study the spread of disease. This model includes a new compartment, denoted by D, representing the deceased individuals. The stochastic nature of the model can be described using the Equation 5:

$$
\begin{aligned}
(S,E,I,R,D) &\xrightarrow{\alpha SI} (S-1,E+1,I,R,D),\\
(S,E,I,R,D) &\xrightarrow{\beta E} (S,E-1,I+1,R,D),\\
(S,E,I,R,D) &\xrightarrow{\gamma I} (S,E,I-1,R,D+1),\\
(S,E,I,R,D) &\xrightarrow{\delta I} (S,E,I-1,R+1,D),\\
(S,E,I,R,D) &\xrightarrow{\epsilon R} (S+1,E,I,R-1,D).
\end{aligned} \tag{5}
$$

This model assumes that each infected individual dies at a rate γ.

IV. Simulation-based Model of the Epidemic Process

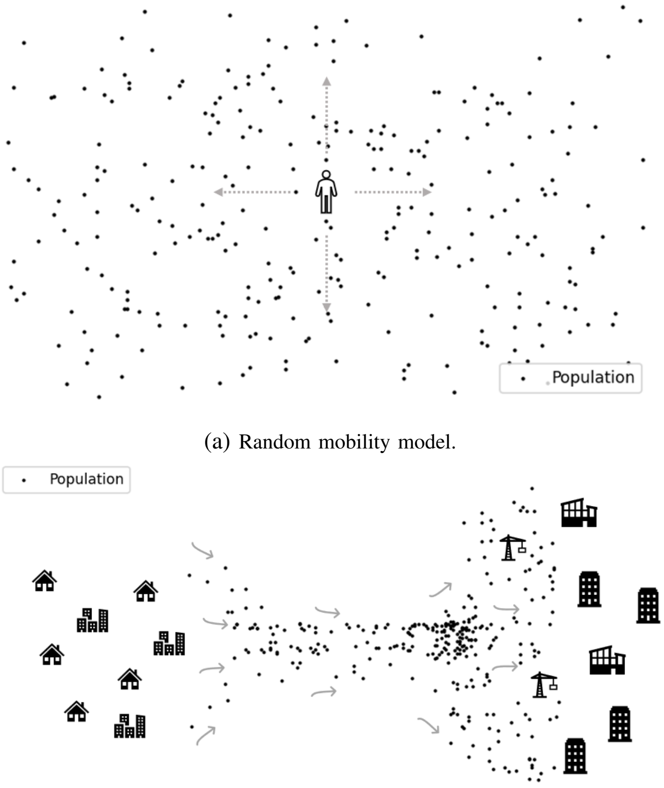
In this section, we present the design of a discrete-event simulator that models the dynamics of a population under two mobility scenarios and its impact on the dynamics of the spreading of an infectious disease. This model is then used as part of the reinforcement learning framework to simulate episodes of the epidemic process that are used to compute optimized control policies.

A. Population Dynamics

To simulate the population behavior, we designed a virtual environment where the population can move according to a pair of models. The first one considers a random free movement simulating movements in open spaces. The second scenario includes a bottleneck to simulate the dynamics of crowded urban areas such as a public transportation system.

In the first scenario, each individual moves horizontally and vertically inside a grid with a random speed restricted by a maximum value d_{max}. This scenario is illustrated in Fig. 1a.

In the second scenario, individuals also move horizontally and vertically but always from their respective housing to their workplace. The scenario includes a narrow passage that all individuals must traverse to reach their destination. This situation creates a bottleneck where individuals closely interact. As in the previous scenario, the micro-movement is random with a maximum speed. This scenario is illustrated in Fig. 1b.



(a) Random mobility model.

(b) Bottleneck mobility model.

Fig. 1. Simulated population behaviors

The environment was modeled as a 2-dimensional grid of size $N \times M$ and a set of P individuals with an initial number of susceptible persons NS and an initial number of infectious persons NI. Given a level of mobility restriction, for example, 75% from the P initial persons, just 25% of them are placed randomly in the grid, simulating those people who could face infectious interactions. The people in the grid move L times per day in the area, causing close interaction between susceptible and infected people. A close interaction is determined by a spread ratio r. These interactions trigger the dynamic process of the disease, which consequently induces health and economic effects. The same process is repeated for T days.

B. Infectious Disease Model

To model the evolution of an infectious disease triggered by infectious interactions, we used a SEIRD model. This model is applied to every person in the population. The interpretation

ISSN 2395-8618

for any infectious disease spreading cycle under this model is as follows:

1) Initially, the population is susceptible to catching the virus (Susceptible state).
2) After being near an infected person, with probability p_{SE}, a person is exposed (Exposed state) to the virus, meaning that one is infected but can not spread the virus yet. The contact may have occurred or not, but with probability $p_{SS} = 1 - p_{SE}$, the person was not exposed.
3) Then, the exposed people, after a certain period, can transmit the disease to other people (Infectious state). Here, the probability p_{EE} determines the virus' incubation time.
4) Infected people can recover from the illness, meaning they have created temporal immunity (Recovered state), or die (Deceased state) with probabilities p_{IR} and p_{ID} respectively.
5) After recovering, one can be susceptible again (return to a Susceptible state) to the virus with probability p_{RS}, which means that the person's immunity has expired. Here, p_{RR} determines the immunity time.

The Markov chain of this model is shown in Fig. 2. Each person's initial state of the Markov chain will depend on whether it is initially susceptible or infected, as described in Section IV-A. The only absorbing state in this model is the Deceased state.

The interaction between the population simulator and the infectious disease model determines the dynamics of disease-spreading process. This interaction is described in Algorithm 1.

V. Computing Mobility-Restriction-based Policies for Epidemic Contention

In this section, we present the design of a discrete-event simulator that combines a simplified SEIRD model with the dynamics of a population composed of individuals who interact with each other by moving in a predefined region. Its main purpose is to provide a tractable model that a reinforcement learning algorithm can use to compute optimized and meaningful control policies.

A. State Space

The *environment state* is composed of the percent (or equivalently, the fraction) of people in the Susceptible (S), Infected (I), Recovered (R), and Deceased (D) compartments. Additionally, we included the fraction of days that have passed before completing an episode. This is, for a number of days T and a certain amount of people P, at each time step t the environment will have a state given by Equation 6:

$$s_t = (S/P, I/P, R/P, D/P, t/T) \in \mathcal{S}. \tag{6}$$

B. Action Space

The spread of the disease can be mitigated by reducing the population's mobility. Hence, the *agent's actions* are defined in terms of mobility restrictions. We established five levels of mobility restrictions: level 0, no restriction; level 1, where 25% of the individuals are static; level 2, 50%; level 3, 75%; and level 4, where 100% of the individuals do not move. Then, at each time step t the agent will choose an action in the action space given by Equation 7:

$$a_t \in \{0, 0.25, 0.5, 0.75, 1\} = \mathcal{A}. \tag{7}$$

C. Cost Function

Our proposed cost function considers a combination of the economic and health negative impacts. The following subsections describe how these two aspects are defined and combined into the global cost function.

1) Economic Cost.: We consider two factors that determine the negative economic impact of a pandemic. The first factor is the mobility restrictions, namely, the percentage γ of people who stay at home. The second factor is the percentage of cumulative deaths σ because deceased people no longer contribute to the economy. We introduce weights ρ_1 and ρ_2, that define the relative importance of mobility restrictions and cumulative deaths respectively. This way, the economic cost can be computed by Equation 8, where ω_e is a weight that a decision maker can use to determine the importance of the economy:

$$\alpha = \omega_e[\rho_1\gamma + \rho_2\sigma]. \tag{8}$$

2) Public Health Cost.: The disease directly affects the population's health when they are infected, causing death in the worst case. That is why we assume an infection cost related to the percentage of new infections per day δ and the percentage of the cumulative number of deaths σ. Then, with ρ_3 and ρ_4 as the importance factors for the infections per day and the cumulative deaths percent, respectively, the health cost is given by Equation 9, where ω_i is a weight that a decision maker can use to determine the importance of the infections:

$$\beta = \omega_i[\rho_3\delta + \rho_4\sigma]. \tag{9}$$

Once we have defined the economic and public health costs, we define our cost function as a linear combination of them (Equation 10). We can minimize this function or maximize its negative value (reward interpretation). Situations or states with higher economic or public health costs will be highly discouraged by getting a significantly high cost (low reward):

$$R(s_t) = -C(s_t) = -(\alpha + \beta). \tag{10}$$

D. Policies Computation

We use the environment described in Section IV to compute optimized policies. We start an episode at day $t = 0$. This episode has an initial state s_t. Using a policy π, we determine

ISSN 2395-8618

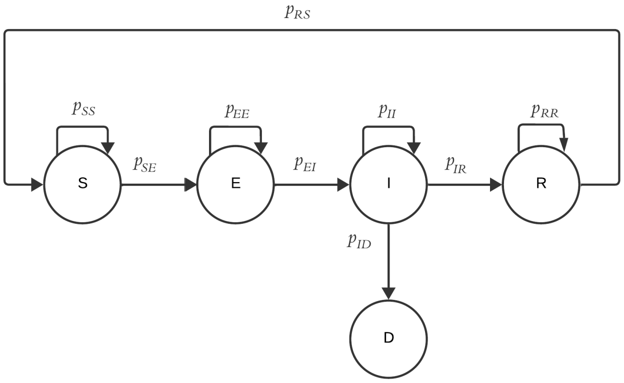


Fig. 2. Markov chain that models the dynamics of the disease within an individual

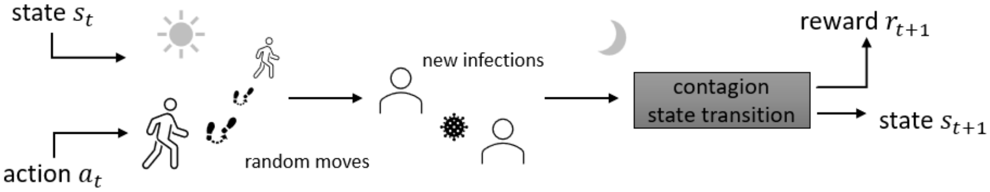


Fig. 3. The simulation-based epidemic model considers the interaction between individuals and their impact on the evolution of the virus spreading process

the episode action (mobility restriction) by sampling it from the policy $a_t \sim \pi(s_t)$. During the day, people perform random moves that generate new infections due to the close contact between susceptible and infectious persons. After the episode completion, we get a new state describing the environment s_{t+1} and determining the initial state for the next episode $t+1$. Using the new state, we can calculate the reward and then use it to improve the policy. This process is illustrated in Fig. 3 and described in Algorithm 1.

Algorithm 1 receives as input the set of probabilities $p_{SE}, p_{EI}, p_{ID}, p_{IR}, p_{RS}$ that define the properties of the SEIRD model that determine the dynamics of the epidemic process, and the set of weights $\omega_e, \omega_i, \rho_1, \rho_2, \rho_3, \rho_4$ that are defined by a decision maker to prioritize the different components of the objective function, namely, the economy or the public health. Algorithm 1 also receives the control policy π that is used to determine the proportion of the population with restricted mobility (Line 4). The algorithm's main for-loop (Lines 8-34) performs a time step (or a day) of the epidemic process which is composed of T days. Every day, the system state is updated according to the interactions of the individuals and to the values of the probabilities of the SEIRD model. As a result, the algorithm returns a sequence of the system state composed of the level of mobility restriction and the cost of applying the mobility restriction for each time step in the simulation.

For the policy improvement process, we use an every-visit Monte-Carlo algorithm (see Algorithm 2 [10]). We started by defining a uniform policy π by assigning the same probability > 0 to every action in the action space $\mathcal{A}$, random Q-values for each pair $(s, a) \in \mathcal{S} \times \mathcal{A}$ and an empty list of episode's returns for each pair state-action. For each improvement iteration (episode) e in a determined number of iteration E we:

1) Generate a simulation of the epidemic dynamics of T days following the policy π (Line 5). This simulation returns a sequence of state-action-reward s_0, a_0, R_1, $\cdots$, s_{T-1}, a_{T-1}, R_t.
2) By iterating backward in the sequence, we compute the quality $Q(s_t, a_t)$ of the taken action a_t in a state s_t. This quality is calculated by averaging the gain obtained at every visit to the state s_t and taking the action a_t (Lines 8-10).
3) The policy is updated at every backward iteration for the state s_t, considering the quality values for every $a \in \mathcal{A}$, $Q(s_t, a)$. The best action that can be taken a^* is the one with the highest quality value at the moment. The policy is updated by increasing the probability of taking the action a^* in state s_t and reducing the other actions probabilities (Lines 11-14).

Algorithm 1 Episode simulation of disease spreading dynamics.

```
Input Total population $P$, radio spread $r$, number of days $T$, steps per day $L$, $p_{SE}, p_{EI}, p_{ID}, p_{IR}, p_{RS}$, area height $N$, area width $M$, $\omega_e, \omega_i, \rho_1, \rho_2, \rho_3, \rho_4$, maximum displacement $d_{max}$, policy $\pi$.
Output Episode sequence $s_0, a_0, R_0, \cdots, s_{T-1}, a_{T-1}, R_T$
Initialize the population grid $G$ of size $N \times M$
$S \sim U(0, 1)$ (Initial susceptible population fraction)
$s_0 \leftarrow (S, I = 1 - S, R = 0, D = 0, t = 0)$
$a_0 \sim \pi(s_0)$ (Mobility restriction under $\pi$)
Initialize $M_i$ as $< mathxmlns = ”http : //www.w3.org/1998/Math/MathML” display = ”block” >< msub >< mi > M < /mi >< mi > i < /mi >< /msub >< /math >$ for $i = [1, \cdots, P]$
Initialize $M_i$ on state $S$ for $i = [1, \cdots, S]$
Initialize $M_i$ on state $I$ for $i = [S + 1, \cdots, P]$
for $t$ in $[0, 1, \cdots, T]$ do
    $S_r, E_r, I_r, R_r, \leftarrow S \cdot a_t,\ E \cdot a_t,\ I \cdot a_t,\ R \cdot a_t$ (Restricted fraction per compartment)
    $N_S \leftarrow (S - S_r) \cdot P$ (Susceptible people)
    $N_I \leftarrow (I - I_r) \cdot P$ (Infectious people)
    $M_S \leftarrow M_S \subseteq \{M_i | M_i \text{ is on state } S\}, |M_S| = N_S$
    $P_i \leftarrow (X_i, Y_i)$ for $i = [1, \cdots, N_S + N_I]$, $X_i \sim U(0, M)$, and $Y_i \sim U(0, N)$
    $\delta_0 \leftarrow 0$ (Infections counter)
    for $l$ in $[1, \cdots, L]$ do
        $P_i \leftarrow P_i + (X_i, Y_i)$ for $i = [0, \cdots, N_S + N_I]$, and $X_i, Y_i \sim U(-d_{max}, d_{max})$
        $A \leftarrow [dist(P_i, P_j)]_{ij}$ for $i = [1, \cdots, N_S]$, $j = [N_S + 1, \cdots, N_S + N_I]$
        for $A_{ij} < d$ do
            if $M_{S,i}$ is on state $S$ then
                Trigger $M_{S,i}$ one step
                if $M_{S,i}$ is on state $E$ then
                    $\delta_0 \leftarrow \delta_0 + 1$
                end if
            end if
        end for
    end for
    Trigger each $M_i \in \{M_i | M_i \text{ is not on state } S\}$
    $\gamma \leftarrow S_r + E_r + I_r + R_r$
    $\sigma \leftarrow 100 \cdot D$
    $\delta \leftarrow \delta_0 / P \cdot 100$
    $C_{t+1} \leftarrow \omega_e[\rho_1 \gamma + \rho_2 \sigma] + \omega_i[\rho_3 \delta + \rho_4 \sigma]$
    $s_{t+1} \leftarrow [S, I, R, D, (t + 1)/T]$
    $a_{t+1} \sim \pi(s_{t+1})$
end for
return $s_0, a_0, -C_1, \cdots, s_{t-1}, a_{t-1}, -C_T$
```

Following the previous steps, we get an optimized policy π^* at the end of the improvement iterations. The Algorithm 2 describes this policy improvement process.

Algorithm 2 Monte Carlo-based reinforcement learning algorithm.

```
Input Training episodes E, days per episode T, states space $\mathcal{S}$, actions space $\mathcal{A}$
Output Optimized policy $\pi^*$
$\pi(a|s) \leftarrow 1/|\mathcal{A}|$ for $a \in \mathcal{A}, s \in \mathcal{S}$ (Random policy)
$Q(s, a) \leftarrow U(-\infty, \infty)$ for $a \in \mathcal{A}, s \in \mathcal{S}$ (Q function)
$R(s, a) \leftarrow \{\}$ for $a \in \mathcal{A}, s \in \mathcal{S}$ (Set of returns)
for $e = [1, \cdots, E]$ do
    Generate an episode following $\pi$: $s_0, a_0, R_1, \cdots, s_{T-1}, a_{T-1}, R_T$ (Algorithm 1)
    $G \leftarrow 0$
    for $t = [T - 1, \cdots, 0]$ do
        $G \leftarrow G + R_{t+1}$
        $R(s_t, a_t) \leftarrow R(s_t, a_t) \cup \{G\}$
        $Q(s_t, a_t) \leftarrow 1/|R(s_t, a_t)| \sum_{r \in R(s_t, a_t)} r$
        $a^* \leftarrow argmax_a Q(s_t, a)$
        for $a \in \mathcal{A}$ do
            $\pi(a|s_t) \leftarrow \begin{cases} 1 - \epsilon + \epsilon/|\mathcal{A}| & \text{if } a = a^* \\ \epsilon/|\mathcal{A}| & \text{if } a \neq a^* \end{cases}$
        end for
    end for
end for
return $\pi$
```

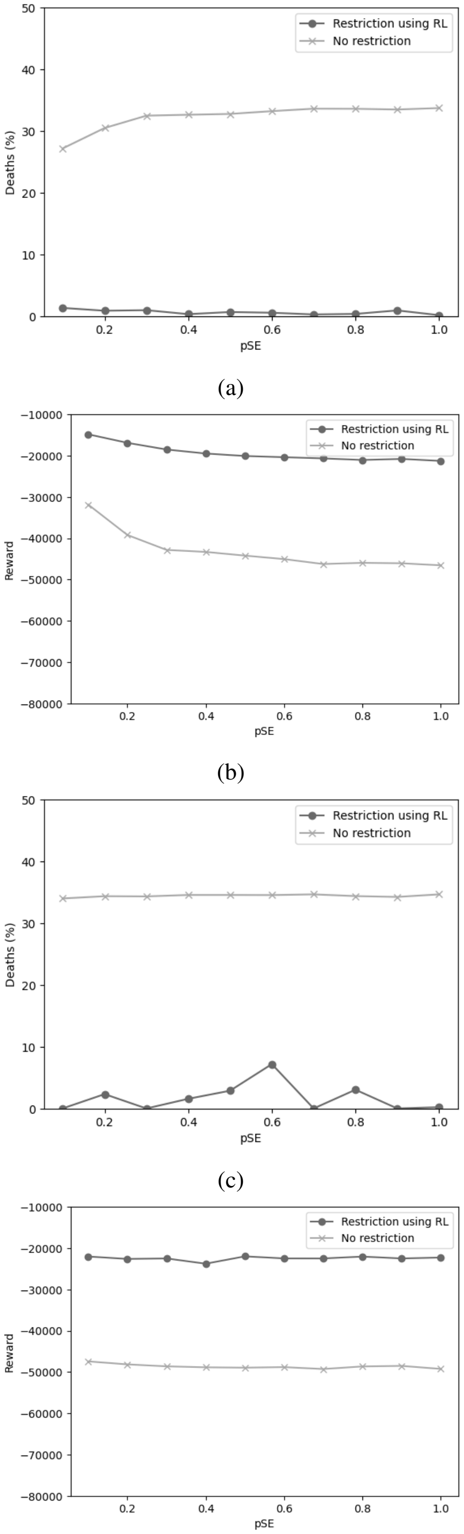
VI. Experimental Results

In this section, we present the results of a series of experiments where we evaluate the performance of the contention control policies computed by the reinforcement learning algorithms. We consider different values of the transition probabilities defined in the SEIRD model that define the particular properties of the virus causing the epidemic. We also consider different values of weights that assign priorities to the two components of the reward function, namely, the economic and public health. We compare the performance of the optimized agents against the case where no mobility restriction policies are implemented.

A. Experimental Settings

The system is initialized with $S = 99\%$ of susceptible people and $I = 1\%$ of infected people. The virus spread radius is set to $r = 1$ (meter), considering the healthy distance proposed by the Mexican government during the COVID-19 pandemic. The random daily movements are $L = 1000$, and we are studying the epidemic process for $T = 30$ days. The weight of restricting mobility and cumulative deaths in the economic cost are set to $\rho_1 = 0.25$ and $\rho_2 = 0.75$. The weight of daily infections and cumulative deaths in the public health cost are set to $\rho_3 = 0.25$ and $\rho_4 = 0.75$. The values of the probabilities of the Markov chain of the SEIRD model are set to $p_{EI} = 0.25$, $p_{IR} = 0.20$, $p_{ID} = 0.05$, $p_{RS} = 0.10$ with the

ISSN 2395-8618



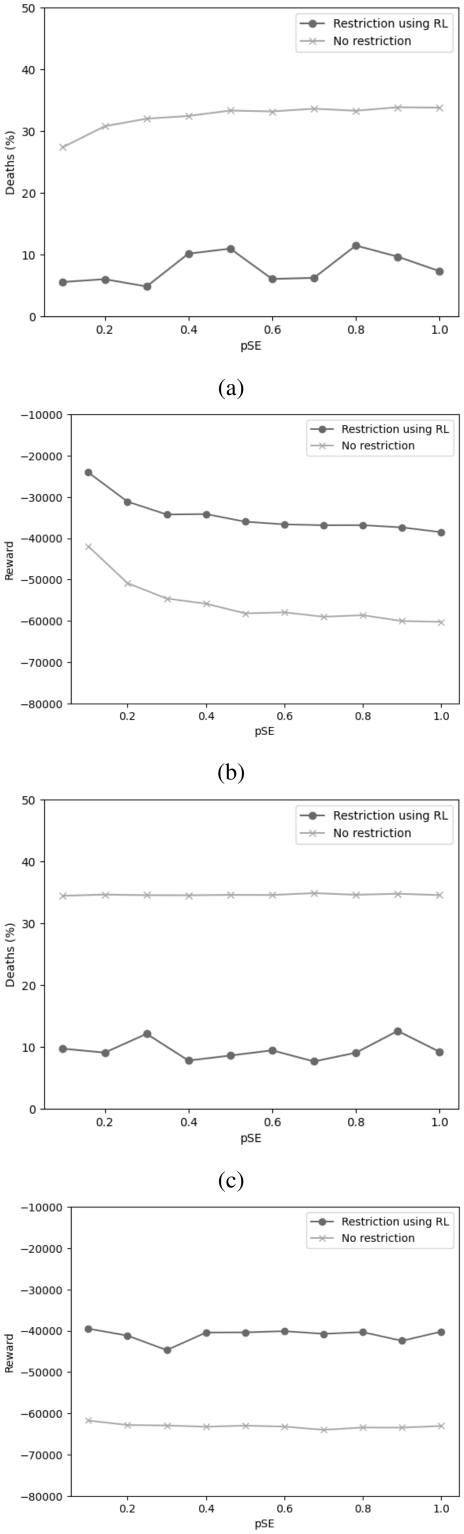
(a)

(b)

(c)

(d)

Fig. 4. Percentage of deaths and reward function when the population moves according to a random mobility model(4a,4b), bottleneck mobility model (4c,4d) and $\omega_e = 30$ to give priority to the public health over the economy



(a)

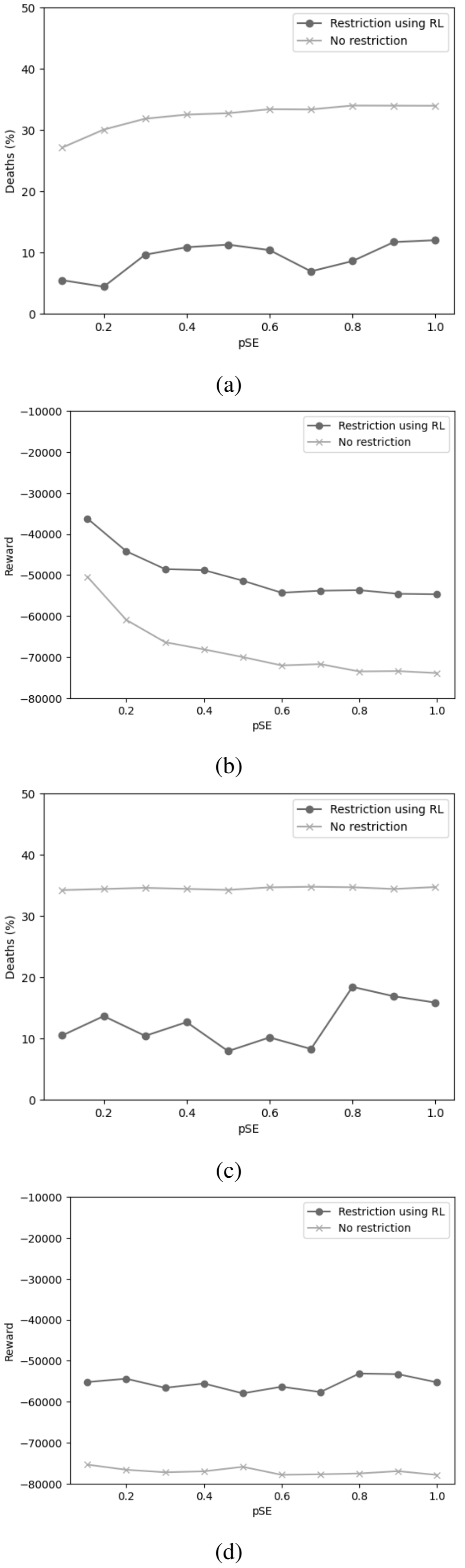
(b)

(c)

(d)

Fig. 5. Percentage of deaths and reward function when the population moves according to a random mobility model (5a,5b), bottleneck mobility model (5c,5d) and $\omega_e = 60$ for a balanced impact of economy and health

ISSN 2395-8618



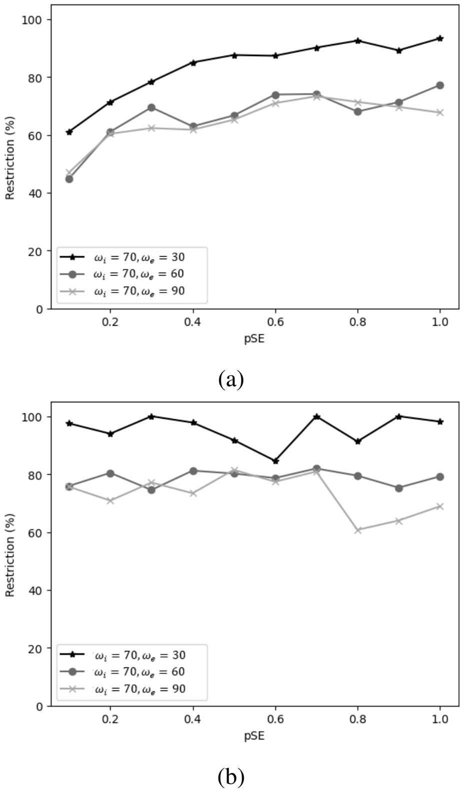
(a)

(b)

(c)

(d)

Fig. 6. Percentage of deaths and reward function when the population moves according to a random mobility model (6a,6b), bottleneck mobility model (6c,6d) and $\omega_e = 90$ to give priority to the economy over the public health



(a)

(b)

Fig. 7. Mobility restrictions for the random mobility model (7a) and the bottleneck mobility model (7b) under different values of ω_e

exception of p_{SE} which is increased to evaluate the impact of having different types of viruses. The weight that determines the infection impact is also fixed to $\omega_i = 70$ and we evaluate different values for the weight ω_e of the economic impact. Results are the average of 100 simulations.

In all the scenarios, we consider a population density similar to that of Mexico City, namely, we have $P = 300$ persons moving inside a simulation area of 216×216 meters.

B. Exploration Scenarios

To evaluate the impact of adopting optimized policies under different disease contagion dynamics, we considered a set of different transmission probabilities:

$$p_{SE} \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1\}$$

Additionally, we considered three different situations for economic impact; when $\omega_e = 30$, considering the health impact more important than the economic one, when $\omega_e = 60$ meaning both impacts are balanced, and $\omega_e = 90$ indicating that the economic impact is more relevant.

Fig. 4 shows the average percentage of deaths and the average value of the reward function (Equation 10) when the population moves according to random and bottleneck mobility models and $\omega_e = 30$ for each value of p_{SE} for 100 simulations.

ISSN 2395-8618

In both mobility scenarios, the behavior is similar. Compared with a fixed policy of no restrictions, when the optimized policies control the mobility restrictions, the number of deaths is lower, and the reward is higher. As expected, the bottleneck mobility induces more infections and consequently, deaths. An additional observation is that in the bottleneck mobility model the probability of being exposed p_{SE} looks to be insignificant, meaning that due to these bottlenecks, infectious interactions will occur in the whole population.

Fig. 5 shows the average percentage of deaths and the average value of the reward function, for 100 simulations, when the population moves according to random and bottleneck mobility models and $\omega_e = 60$ for each p_{SE}. In this case, we observe the same phenomenon as in Fig. 4. Nevertheless, the reward values are lower than in the case of $\omega_e = 30$. As the economic importance is higher, the mobility restrictions are lower, indirectly causing more deaths and, consequently, lower rewards.

In the last scenario, we set $\omega_e = 90$ to give higher priority to the economy. The results are shown in Fig. 6. Comparing these results with that of $\omega_e = 60$, the number of deaths did not increase drastically, but the cost did. Despite the importance given to the economy, having soft restrictions will indirectly cause more deaths, which will also affect the economy. Then, it is preferable to keep more stringent mobility restrictions even in a situation where the economy is more important than public health.

Lastly, Fig. 7 shows the restrictions applied by the reinforcement learning algorithm for all the scenarios. These results confirm two of the previous observations: when the population moves along a bottleneck space, p_{SE} becomes less relevant, and after passing a threshold, even when the economy is prioritized by making the value of ω_e large, implementing stringent mobility restrictions is the best course of action.

VII. Conclusions and Future Work

In this work, we presented a simulation-based model of the spreading dynamics of an airborne virus that is characterized by a set of probabilities that determine how the corresponding disease evolves within the individuals. The simulation model incorporates different mobility models that govern the way people interact with each other.

From the simulation-based model, we trained a reinforcement learning agent to learn the mobility restriction policies that simultaneously minimize the negative impacts of the epidemic on public health and the economy. Our results revealed that the policies derived by our agent effectively reduce these negative impacts and that can be fine-tuned to prioritize either protecting the economy or the public health.

Future work includes scaling up the simulation experiments, incorporating more realistic mobility models, and more detailed characterization of the individuals, for instance, to consider relevant attributes such as age, comorbidities, and the use of facemasks. There is also room to investigate using different reinforcement learning algorithms and new formulations of the related multi-objective optimization problem.

References

[1] R. M. Anderson, H. Heesterbeek, D. Klinkenberg, and T. D. Hollingsworth, “How will country-based mitigation measures influence the course of the COVID-19 epidemic?” *Lancet*, vol. 395, no. 10228, pp. 931–934, Mar. 2020.

[2] G. Bhatt, *Modeling epidemics with differential equations*, 02 2023.

[3] D. Faranda, I. P. Castillo, O. Hulme, A. Jezequel, J. S. W. Lamb, Y. Sato, and E. L. Thompson, “Asymptotic estimates of sars-cov-2 infection counts and their sensitivity to stochastic perturbation,” *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 30, no. 5, p. 051107, 05 2020. [Online]. Available: https://doi.org/10.1063/5.0008834

[4] W. O. Kermack and À. G. Mckendrick, “A contribution to the mathematical theory of epidemics,” *Proceedings of The Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 115, pp. 700–721, 1927.

[5] I. Z. Kiss, J. C. Miller, and P. L. Simon, “Introduction to networks and diseases,” in *Mathematics of Epidemics on Networks: From Exact to Approximate Models*. Springer International Publishing, 2017, pp. 1–26.

[6] S. López-Moreno, “Salud pública y medicina curativa: objetos de estudio y fronteras disciplinarias,” *Salud Pública de México*, vol. 42, 04 2000.

[7] A. Q. Ohi, M. F. Mridha, M. M. Monowar, and M. A. Hamid, “Exploring optimal control of epidemic spread using reinforcement learning,” *Scientific Reports*, vol. 10, no. 1, p. 22106, Dec 2020. [Online]. Available: https://doi.org/10.1038/s41598-020-79147-8

[8] P. E. Paré, C. L. Beck, and T. Başar, “Modeling, estimation, and analysis of epidemics over networks: An overview,” *Annual Reviews in Control*, vol. 50, pp. 345–360, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1367578820300614

[9] N. Qualls, A. Levitt, N. Kanade, N. Wright-Jegede, S. Dopson, M. Biggerstaff, C. Reed, A. Uzicanin, and CDC Community Mitigation Guidelines Work Group, “Community mitigation guidelines to prevent pandemic influenza - united states, 2017,” *MMWR Recomm Rep*, vol. 66, no. 1, pp. 1–34, Apr. 2017.

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

[11] R. Wan, X. Zhang, and R. Song, “Multi-objective model-based reinforcement learning for infectious disease control,” 2020. [Online]. Available: https://arxiv.org/abs/2009.04607