Flight Policy Decision during COVID-19 using Reinforcement Learning Algorithms

Qiong Hu	Mindy Saylors	Di Gao
ID: 405065032	ID: 605431628	ID: 405429117
qiong.j.hu@gmail.com	mindysaylors@ucla.edu	di0396@ucla.edu

June 11th, 2020

Contents

1 Introduction

Under the circumstance of the global pandemic caused by the COVID-19 epidemic, the airline industry has been facing a severe economic challenges. Since many authorities have posted strict travel bans for countries or certain regions, a large number of airlines have had to stop flights altogether, and the remaining are greatly reducing seat occupancy. Therefore, it is important to bring up a decision model that accounts for various influence factors and aids in developing an appropriate flight planning policy given the severity of the COVID-19 outbreak.

To simplify the problem, we only consider the revenue and disease transmission as dominant influence factors in our proposed method. The input of the decision-making model is the initial epidemic severity of the two cities where the flight would travel to and from, represented by the infectious case ratio among city population. The output is the"flight policy" which for our problem refers to the suggested ratio of flight amounts with regard to the maximum flight capacity between the two cities, and suggested average seat occupancy for these flights. The objective of the problem is to find the flight policy that makes the total revenue of the airlines as large as possible by transporting more passengers with less planes, and keeps the disease transmission as small as possible by predicting less infectious case number over time.

Using this action-oriented, time-dependent flight decision model, our goal is also to predict when the airline industry as well as how the pandemic situation might return to an new normal in this updated economic climate, as an extended application.

2 Related Work

The existing solution for making flight policies are mostly based on either experience from previous outbreaks and pandemics, such as SARS in 2003, Ebola in 2014, and MERS flu in 2015, or theoretical model-based predictions.

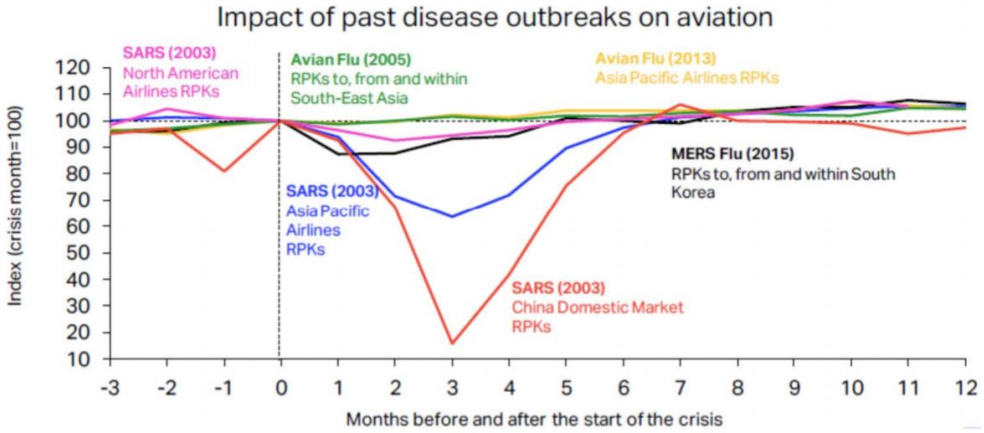

Figure 1: Impact of past disease outbreaks on aviation [1].

According to Figure 1, the average aviation recovery time is about 6-7 months, but the history model may not apply to today's situation since the impact of COVID-19 has already surpassed the 2003 SARS outbreak.

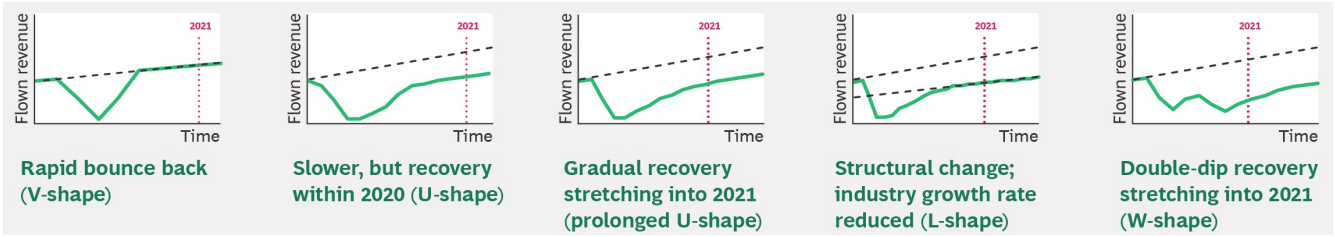

Figure 2: Five demand recovery scenarios in a highly uncertain future [2].

Another existing popular method is to use theoretical recovery scenarios to predict the proper curve trend under different flight decisions, as seen in Figure 2. There are five typical recovery scenarios, among which the prolonged U-shape is considered to be the most likely empirically. The prolonged U-shape scenario predicts that it takes about 12-18 months for the flown revenue to return to a new normal, considering the lock-down for several months, discouraged travel and slowly-reopened borders. However, a shortage of this scenario-based method is that, there are more than five types of possibly helpful flight policies, which may not fall into one of the five scenarios.

3 Problem Formalism

We propose a method using Reinforcement Learning (RL) Algorithms. The core elements needed to set up the problem environment for any RL algorithm are state, action, and reward. In the following four subsections, we will introduce how each core element is defined, and what data we will utilize for our proposed method.

3.1 State

We use the Susceptible-Exposed-Infectious-Recovered (SEIR) epidemic model to simulate the natural spread of COVID-19 in each city, which has proven reliable in tracking the Coronavirus spread [3]. Individuals are each assigned to one of the following disease states: Susceptible (S), Exposed (E), Infectious (I), and Recovered (R), which refers to individuals that are not yet infected and disease-free, experiencing incubation of disease, the confirmed and isolated cases, and recovered individuals, respectively. Using this model and three parameters (β, σ, γ) to estimate how each disease state evolves, we can describe the virus transmission by the following equations [3] as shown in Equation (1) to (4):

$$\frac{dS}{dt} = -\frac{\beta SI}{N}, \tag{1}$$

$$\frac{dE}{dt} = \frac{\beta SI}{N} - \sigma E, \tag{2}$$

$$\frac{dI}{dt} = \sigma E - \gamma I, \tag{3}$$

$$\frac{dR}{dt} = \gamma I, \tag{4}$$

where S, E, I, R represent the number of individuals in each catagory, N is the total population, $\beta = R_0 * \gamma$ controls the rate of spread, $\sigma = 1/Y$ is the incubation rate, and $\gamma = 1/D$ is the recovery rate. In our

simulation, we use reproductive number $R_0 = 2.2$, incubation duration $Y = 5.2$ days, and infectious duration $D = 7.5$ days [3]. These variables will be discussed later in Section 3.4.

In order to model the influence of flights on both cities where the planes take off and land, we define the state space as:

$$\mathcal{S} = \{S_l, E_l, I_l, R_l, S_r, E_r, I_r, R_r\}. \tag{5}$$

State space $\mathcal{S}$ has eight dimensions: the first four of them $\{S_l, E_l, I_l, R_l\}$ represent the population ratios of four disease states in the city where the outbound planes take off (termed "local city"), and the other four $\{S_r, E_r, I_r, R_r\}$ represent the population ratios of four disease states in the city where the outbound planes land (termed "remote city"). Since we are using population ratios instead of the individual numbers for each state, all the states should meet the following requirements:

$$0 \leqslant S_l, E_l, I_l, R_l, S_r, E_r, I_r, R_r \leqslant 1, \tag{6}$$

$$S_l + E_l + I_l + R_l = 1, \tag{7}$$

$$S_r + E_r + I_r + R_r = 1. \tag{8}$$

For added simplicity of calculations in later algorithms, we made two assumptions for states:

(i) The number of outbound flights and passengers from local city to remote city is the same as the number of inbound flights and passengers from remote city to local city, thus the total population N of both cities remain unchanged during the time of estimation.

(ii) The infectious ratio in the flight for passengers is the same as the infectious ratio (I) in the whole city where the flight departs.

3.2 Action

To model the flight decisions, we define our action space as:

$$\mathcal{A} = \{N_f, N_s\}, \tag{9}$$

where N_f is the relative amount of flights with regard to the maximum daily available amount of flights $N_{f,max}$ between the two cities, and N_s is the seat occupancy in each flight with regard to the maximum available seat number $N_{s,max}$ for the flight. Since both N_f and N_s are ratios, they have upper and lower bounds:

$$0 \leqslant N_f, N_s \leqslant 1. \tag{10}$$

To simplify the problem, we consider maximum number of flights $N_{f,max}$ to be time-independent, which means in real scenarios, $N_{f,max}$ is an average value of historical daily flight amounts between the two cities before the epidemic began. Temporarily, we also imagine all the planes in question have the same size with the same maximum seating amount $N_{s,max}$, which means in real scenarios, $N_{s,max}$ is an average value of the maximum amount of seats among all flights between the two cities. Based on our definitions and assumptions, we have:

$$\text{Actual number of flights: } N_{f,real} = N_f * N_{f,max}, \tag{11}$$

$$\text{Maximum number of passengers: } N_{p,max} = N_{f,max} * N_{s,max}, \tag{12}$$

$$\text{Actual number of passengers: } N_p = N_f * N_{f,max} * N_s * N_{s,max}. \tag{13}$$

For our environment setup, the effect of flight actions on states is described as perturbations of the original SEIR model due to an increased possibility for passengers to become "Exposed (E)" after staying in the same confined space with potential "Infectious (I)" passengers for a certain amount of time. More concretely, we assume a linear relationship between seat occupancy N_s, flight duration (represented by flight distance D_f), and exposure rates, and therefore the effects of flights can be expressed as shown in Algorithm 1.

Algorithm 1: Effects of inbound flight actions on local disease spread

$N_p = N_f * N_{f,max} * N_s * N_{s,max}$
$N_{I,p} = I_r * N_p$
$N_{E,p} = N_{I,p} * R_0 * N_s * D_f * C$
if $N_{I,p} + N_{E,p} > N_p$ **then**
    $N_{E,p} = N_p - N_{I,p}$
    $N_{S,p} = 0$
else
    $N_{S,p} = N_p - N_{I,p} - N_{E,p}$
end
$S_{l,new} = [S_l * (N_l - N_p) + N_{S,p}]/N_l$
$E_{l,new} = [E_l * (N_l - N_p) + N_{E,p}]/N_l$
$I_{l,new} = [I_l * (N_l - N_p) + N_{I,p}]/N_l$
$R_{l,new} = R_l * (N_l - N_p)/N_l$

In Algorithm 1, N_p is the total number of passengers, $N_{S,p}, N_{E,p}, N_{I,p}$ represent the number of Susceptible (S), Exposed (E), Infectious (I) passengers, $S_{l,new}, E_{l,new}, I_{l,new}, R_{r,new}$ are the four updated ratio values for the local disease states as similarly defined in Equation 5. N_l is the local population, I_r is the remote infectious ratio before the flight has been added, R_0 is the disease reproductive number, C is an adjustable constant to more accurately describe the influence of the flight duration when using flight distance D_f. This same algorithm can also be easily reversed to calculate the effects of outbound flight actions on the remote city's disease spread by swapping all subscripts l's with r's.

The practical reason for assuming the number of exposed passengers to have a linear relationship to the number of infectious passengers, disease reproductive number, seat occupancy, and flight duration is inspired by a simulated visualization of the resulting transport of expiratory droplets in an aircraft cabin [4]. As one may expect, longer flights result in greater exposure to infected passenger which causes more exposed and infectious passengers.

A complete algorithm pipeline of updating the state s_t to a new state s_{t+1}, after taking flight action a_t and time dt is shown in Algorithm 2.

Algorithm 2: State evolution algorithm after flight action a_t and time dt

Function `state_evolve`(s_t, a_t)**:**
    $s_{t'}$ = Updated s_t with natural disease spread after dt using SEIR Equation (1)-(4)
    $s_{t'} = (s_{local}, s_{remote})$
    $s_{local'}$ = Updated s_{local} with effects of flight action a_t on local city using Algorithm 1
    $s_{remote'}$ = Updated s_{remote} with effects of flight action a_t on remote city using Algorithm 1
    $s_{t+1} = (s_{local'}, s_{remote'})$
    return s_{t+1}

3.3 Reward

To compare flight revenue, flight cost and disease transmission into balanced consideration, we define our reward function at any step of state evolution as follows:

$$\begin{aligned} R(s,a) &= c_p * N_p - c_f * N_f - c_I * (N_{I,p,out} + N_{I,p,in}) \\ &= c_p * N_p - c_f * \mathcal{A}(N_f) - c_I * (\mathcal{S}(I_l) + \mathcal{S}(I_r))N_p \\ &= c_p * \mathcal{A}(N_f)\mathcal{A}(N_s)N_{p,max} - c_f * \mathcal{A}(N_f) - c_I * (\mathcal{S}(I_l) + \mathcal{S}(I_r))\mathcal{A}(N_f)\mathcal{A}(N_s)N_{p,max}, \end{aligned} \tag{14}$$

where c_p, c_f, and c_I are three adjustable positive coefficients to balance the relative importance of these three factors, N_p is the passenger number from Equation 13, N_f is the relative flight number according to the action, $N_{I,p,in}$ and $N_{I,p,out}$ represent the infectious passengers from inbound and outbound flights, respectively to calculate the total number of infectious passengers.

We consider the reward to have a positive relationship with passenger amount because for airline companies, the more passengers per flight, the more revenue gained. The penalty for the number of flights is our attempt to encapsulate operating costs and enforce a penalty for allowing low-occupancy or near empty flights. To maintain a cautious disease transmission rate, we compute a penalty linearly dependent on the total number of infectious passengers from both inbound and outbound flights, and take a high relative weight of c_I.

The reward function is formalized this way so to not only balance between economics and the public health, but also the balance between the short-term benefits and long-term benefits of each. For example, in an extreme scenario where the immediate benefits are gained by taking a large number of flights and passengers, the disease feature of incubation ensures that the infectious passenger number will not rise too sharply over time. Therefore, under the current definition of the reward function, our policy should not target gaining maximum immediate rewards, and instead, consider upcoming developments and future rewards as well.

In the default model-training simulations, we use passenger reward $c_p = 1$, flight cost $c_f = 1$, and infectious penalty $c_I = 10$. The evaluation of policies trained from different comparative weights of these factors will be discussed in Section 6.1.

3.4 Data

Our simulation includes both synthetic experiments and real-data experiments, both of which require a set of data to describe the conditions of cities, flights and the COVID-19 virus.

The data required for our simulation was the population and the confirmed number of cases for both local and remote cities, and we acquire this data from the JHU dataset[5]. In synthetic experiments, we only implement the initial conditions, while in real-data experiments, we use the real confirmed case number from January 22 to June 6, 2020 to replace the estimated infectious number in the state space, and then follow the same algorithm procedure to generate predicted actions for a year.

As we introduced in Section 3.2 about flight conditions, we use relative ratios to define the action space, but we also need to calculate the real passenger number to update the flight effects. Therefore, we needed concrete values for maximum flight number $N_{f,max}$ and maximum seat capacity $N_{s,max}$. Using the Chinese civil aviation annual report [6], we employ an average number $N_{f,max} = 300$ flights per day and $N_{s,max} = 200$ seats per flights in all of our experiments.

As for virus-related parameters, we use reproductive number $R_0 = 2.2$, incubation duration $Y = 5.2$ days, and infectious duration $D = 7.5$ days [3] for all the cities. In future work, these parameters should be customized to each individual city to better model the disease transmission.

4 Proposed Solution and Algorithm

To solve this flight decision-making problem and find a policy that returns an optimal total reward over a long time period, we implement two algorithms: Q-learning, and Proximal Policy Optimization (PPO) algorithm.

4.1 Q-learning

Q-learning is an off-policy model-free reinforcement learning algorithm that looks for an optimal policy by maximizing the expected Q value over a certain number of successive steps, starting from the current state.

The core of the algorithm is to update state and action -related Q value using the weighted average of old value and new value in future states with optimal future actions:

$$Q^{new}(s_t, a_t) \leftarrow (1-\alpha)\underbrace{Q(s_t, a_t)}_{\text{old value}} + \underbrace{\alpha}_{\text{learning rate}} \overbrace{[\underbrace{R(s_t, a_t)}_{\text{reward}} + \underbrace{\gamma}_{\text{discount factor}} \cdot \underbrace{\max_a(Q(s_{t+1}, a))}_{\text{estimate of optimal future value}}]}^{\text{learned value}}, \quad (15)$$

where $R(s_t, a_t)$ is the reward received when moving from state s_t to state s_{t+1} with action a_t using Equation 14, α is the learning rate ($0 < \alpha \leqslant 1$), and γ is the discount factor ($0 < \gamma \leqslant 1$). We use $\alpha = 0.1$ and $\gamma = 0.95$ to update Q values.

From Equation 15, we can see that the algorithm has a mapping function $Q : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ to map state space and action space to a set of real numbers. During implementation, since our state space and action space are both continuous space, there will be an infinite number of (s_t, a_t) combinations, making it difficult to assign Q value for each of them, thus we implement a discretized Q-table as an approximation, with a finite set of states and actions.

In most of our training and testing experiments, we equally divide all the state variables to 20 bins. Since the state space from Equation 5 has 8 dimensions, the resulting Q-table would have a size of more than $20^8 = 2.56 * 10^{10}$, which is too large for most of common computers to calculate and update. Therefore, to avoid the curse of dimensionality, we use a reduced state space with only 2 dimensions: $\mathcal{S}_{reduced} = \{I_l, I_r\}$ including a local infectious ratio and a remote infectious ratio to calculate the Q-table while still using the complete state space to update the estimated disease spread.

Unlike the state space, the action state space is naturally discrete because there are a discrete number of flights and number of passengers. In Section 3.4, we show that the max number of flights $N_{f,max} = 300$ and the average maximum seat capacity $N_{s,max} = 200$; from Equation (11, 13), we can see the real flight and passenger number is the multiplication of ratios in action space and the maximum values. In reality, the decision of flight amounts and seat occupancy will be integers, therefore we discretize the action space so that the divided units would be integers. In most of our implementations, we use a

flight unit = 5 and seat unit = 5, which means our discretized decision-making policy would decide on a precision of every 5 flights and every 5 seats. Therefore, the discretized action space has a size of $(300/5) * (200/5) = 60 * 40 = 2,400$.

With above considerations, we implement a Q-table with 4 dimensions, two of which represent state space, and the other two represent action space, and the size of Q-table is $20 * 20 * 60 * 40$. We also tested and evaluated on policies generated by Q-table with different sizes, and the results are demonstrated in Section 5.2.

Algorithm 3: Q-learning algorithm

```
Randomly initialize Q-table
repeat
    repeat
        if N_random < ε then
            a_t = random action
        else
            a_t = arg max_a(Q(s_{t+1}, a))
        end
        s_{t+1} = state_evolve(s_t, a)
        Update Q(s_t, a_t) using max_a Q(s_{t+1}, a)
    until Reach number of max-steps
    ε decays
until Reach number of episodes
Output: Q-table
```

A complete pipeline of using a Q-learning algorithm to train the Q-table and find optimal decision-making model is shown in Algorithm 3, where we utilize an episode-dependent variable, ϵ, to find the trade-off between exploration and exploitation: in each update, the training model has a probability of ϵ to take a random action and a probability of $1-\epsilon$ to take the action with maximum Q-value under current state. We tested on $\epsilon = 0.7, 0.9$, and ϵ decay rate $= 0.9999, 0.99999$, with different number of total episodes. By looking at the converge trend of episode rewards, we decide to use $\epsilon = 0.9$ and decay rate $= 0.9999$ with episode number $= 30000$ in our training and testing experiments.

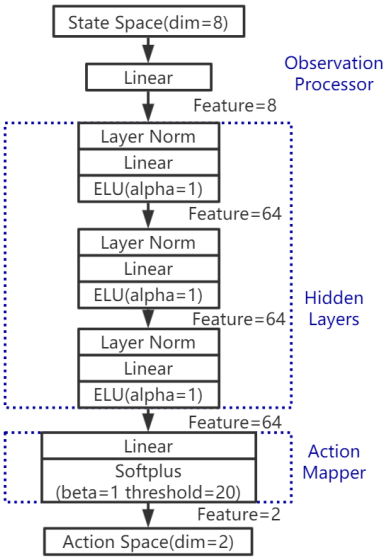

Figure 3: Pipeline diagram of PPO algorithm with a neural network architecture.

4.2 Proximal Policy Optimization (PPO)

Proximal Policy Optimization (PPO) algorithm is an online policy gradient method for reinforcement learning, which is commonly used at OpenAI due to its ease of use and good performance.

PPO algorithm is applicable for environments with either discrete or continuous state and action space. The key contribution of PPO is a loss function that combines large policy update penalty, value function

error from neural network and an entropy bonus, ensuring that a new update of the policy does not change too much from the previous policy. This definition of loss function leads to less variance in training, and makes sure the agent does not go down an unrecoverable path of taking senseless actions, at the cost of some bias [7].

The structure of PPO with a neural network architecture that we implemented is shown in Figure 3, where there are three hidden layers with 64 hidden features in each layer. We also utilize an automatic observation processor and an action mapper in PyTorch to interact our full continuous state space (dimension = 8) and action space (dimension = 2) with hidden layers, instead of using reduced discretized state and action spaces as in Q-learning algorithm. We train using 30000 episodes to obtain a resulting decision-making agent and evaluate its performance in Section 5.2.

5 Simulation Set-Up

In this section, we will introduce visualization and evaluation methods of a policy, and a brief description of our simulation scenarios, for which the results and discussions are shown in Section 6.

5.1 Result Visualization

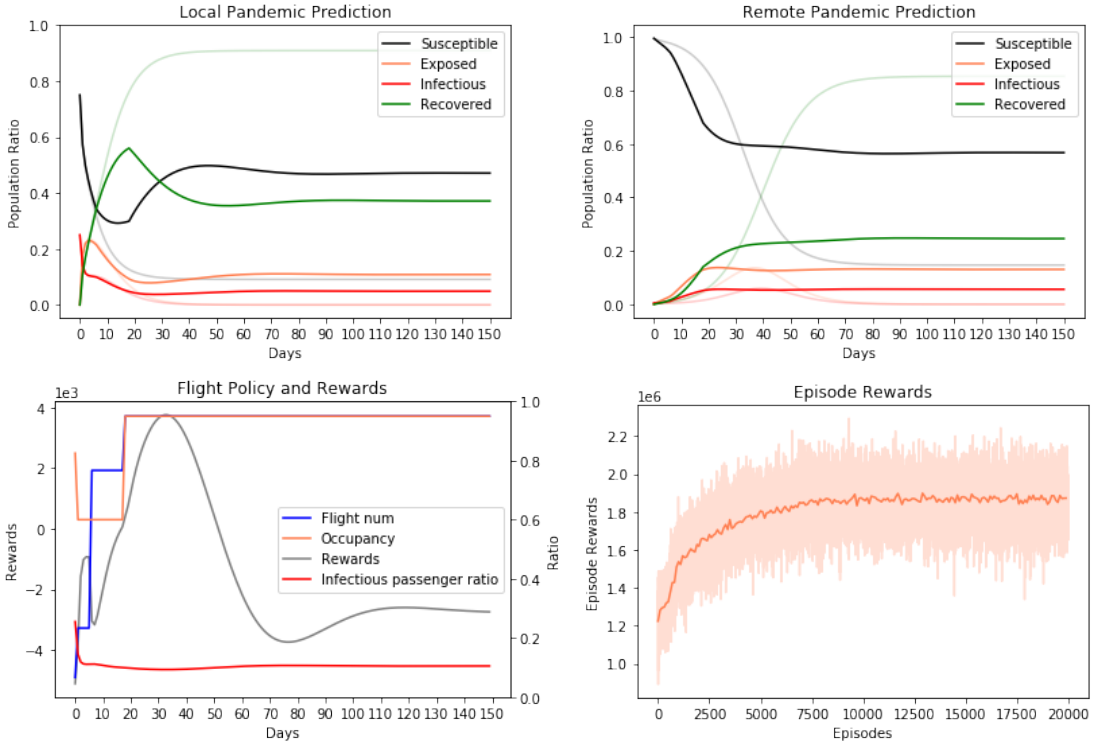

Figure 4: Flight policy visualization example.

Figure 4 is an example showing the training process with the episode rewards, and plot the policy in

three images: pandemic prediction of each disease state in the local and remote city, and flight actions along with rewards and infectious passenger ratio in 150 days.

In the two pandemic state figures, we also plot the baseline curves in a lighter color, where there are no flights. Generally, we could find an explainable difference between the predicted curve with or without flight actions. For instance, in this example, we define the local city to have 40% more population than the remote city, and have a severe pandemic stage for the local city with an initial infectious ratio of 25% among the local city population, but a 5% infectious ratio for the remote city. We can see from the figures that, with flight actions between the two cities, more infectious passengers are transported to remote city, causing an earlier rising trend in exposed and infectious curves compared to no-flight situation.

By looking at the action figure along with state figures, we find out that this resulting policy suggests a new normal where both the flight number and seat occupancy returns to 95% of their maximum availability after 18 days. However, the cost of gradually allowing flights between the two cities is that the pandemic situation might never vanish. After about 70 simulation days, the disease spread in both cities reaches a stable state, where there is about 10% of the population that remains exposed and 5% remains infectious. Considering the trade-off between the economics and public health benefits, this visualization result seriously encourages policy-making authorities to think whether this really is a cost that the general public should and would take.

5.2 Evaluation Method

To evaluate the performance of a policy, we need to use variables other than the absolute reward value. This is because some policies are trained from reward function with different weighted coefficients and we still want to compare their performances. Therefore, here we define six variables to describe and evaluate how realistic and how good a policy is: action unrealism, passenger number, flight number, infectious number, recovered number, and infectious passenger number.

- Action unrealism: To describe the level of realistic of a policy, we look at the changing rate of the actions. From action visualization figure in Figure 4, we notice that the flight number and seat occupancy are often a set of step functions, suggesting that a realistic flight decision would usually remain unchanged for a short period of time before updating to a new policy due to changed pandemic situation. The more rapid and violent the change is, the more unrealistic the actions are, therefore, we define the sum of relative changing rate as "Action unrealism":

$$\text{Action unrealism} = \sum_{\text{periods}} \left(\frac{|\Delta N_f|}{\Delta t} + \frac{|\Delta N_s|}{\Delta t} \right), \tag{16}$$

  where Δt represent the duration of time when both N_f and N_s are unchanged, $\sum(\Delta t)$ is the total simulated time, which in most of our experiments, is 150 days.

- Passenger number:

$$N_{p,total} = \sum_{\text{time}} N_p, \tag{17}$$

  where daily passenger number N_p has the same definition as in Equation 13, and the total passenger number $N_{p,total}$ is the sum of N_p over the whole simulated time period. Larger $N_{p,total}$ means more passengers travel between the two cities, bringing more revenue for the airline industry, and therefore it is considered to be a better policy.

- Flight number:

$$N_{f,total} = \sum_{\text{time}} N_f, \tag{18}$$

where daily flight number N_f has the same definition as in the action space from Equation 9. We consider the operating cost as the dominant influence on flight number, so a better policy should have a lower total flight number $N_{f,total}$.

- Infectious number:

$$N_{I,total} = \sum_{\text{time}} (I_l * N_l + I_r * N_r) = N_l * \sum_{\text{time}} I_l + N_r * \sum_{\text{time}} I_r, \tag{19}$$

where I_l, I_r represent the daily infectious population ratio in the local and remote cities, the same as in Equation 5, and N_l, N_r are time-independent population number in two cities. A good policy should have a lower total infectious individuals in both cities over the simulated time.

- Recovered number:

$$N_{R,total} = \sum_{\text{time}} (R_l * N_l + R_r * N_r) = N_l * \sum_{\text{time}} R_l + N_r * \sum_{\text{time}} R_r, \tag{20}$$

where R_l, R_r represent the daily recovered population ratio in the local and remote city, the same as Equation 5. A good policy should have a higher total recovered individuals in both cities over the simulated time.

- Infectious passenger number:

$$N_{I,p,total} = \sum_{\text{time}} (N_{I,p,in} + N_{I,p,out}) = \sum_{\text{time}} (I_l + I_r) * N_p, \tag{21}$$

where $N_{I,p,in}$ and $N_{I,p,out}$ represent daily infectious passengers for inbound and outbound flights respectively. The reason why both the total infectious passenger number $N_{I,p,total}$ and infectious individuals in the cities $N_{I,total}$ are included in evaluation is because $N_{I,p,total}$ does not explicitly involve city populations. Therefore, it would be more clear to show the direct influence of flight actions. A good policy should ensure both values to be low.

In summary, using these six variables as our evaluation method, we consider a policy to be good and realistic, if it has a comparatively low value in action unrealism, flight number, infectious number, and infectious passenger number, and a comparatively high value in passenger number and recovered number.

By applying this evaluation method, we can compare and evaluate the performance of policies obtained using different algorithms or algorithm parameters, as shown in Figure 5.

Figure 5 shows the evaluation result of three different sizes of Q-tables in comparison to PPO algorithm. For ease of understanding, we normalized the evaluation variables to a range between 0 and 1, by dividing all the values to the maximum value in each category[1].

[1]The same normalization method will be applied in other bar diagrams in this paper when showing evaluation results. Thus, the ratio value across different diagrams cannot be directly used for comparison, because the original maximum range might be different.

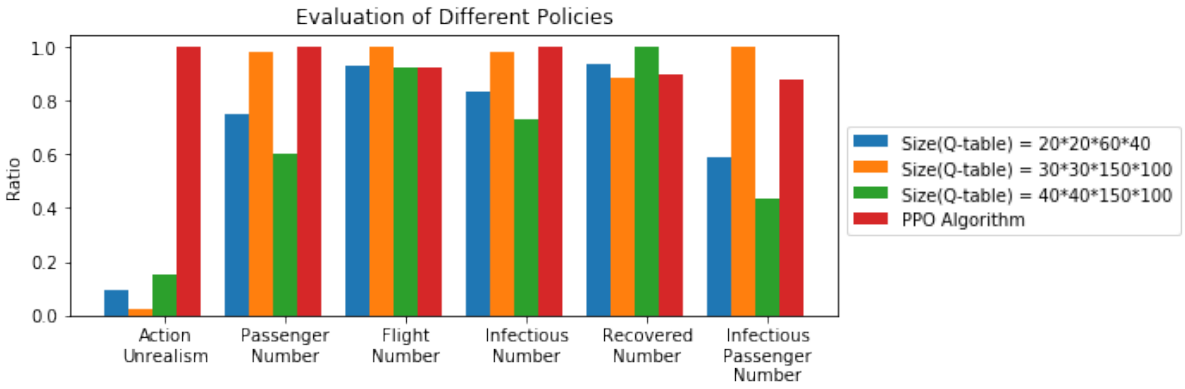

Figure 5: Evaluation of different policies.

We can see that the resulting policy trained by PPO Algorithm has the highest value of action unrealism, passenger number, infectious number, and second highest infectious passenger number, reflecting a poor overall performance, which is why we are not going to use it in the following simulation experiments.

For different sizes of Q-tables, we tested on equally dividing state space to 20, 30 and 40 bins, and having flight units and seat units to be 5 or 2. We expect the performance of the Q-table should be better with smaller increments in size, since the resolution for action decision would be higher. However, the evaluation results suggest that when the Q-table size is $30 * 30 * 150 * 100$, the infectious number and infectious passenger number are the highest among the three tests. We suspect the reason might be that, the number of episodes for training is not sufficient enough to traverse and update all the Q-values in the Q-table, resulting in some Q-values to be initially randomized values. Since the Q-table with a size of $40 * 40 * 150 * 100$ shows an overall better performance with a smaller infectious number and infectious passenger number, and a larger recovered number, we will use this pre-trained policy to experiment on most of the simulation scenarios in Section 6. We call this policy $\pi^{\mathbf{0}}$ for future reference, and since many of the simulation scenarios are different from the one this policy is trained from, we consider these experiments to be also "transfer learning" experiments.

5.3 Simulation Experiments

Based on the environment set-up in Section 3, we have a list of parameters to manipulate with, and we can divide them roughly into two classes for experiments: different city situations (including different populations, initial pandemic level and flight distance), and different coefficients for the reward function, as shown in Equation 14.

From Algorithm 2, the effects of flight actions on the disease states are updated symmetrically for local city and remote city in each step of the state evolution, which means the policy should remain the same when reversing local city situation with remote city. Thus, we only need to experiment on one of two scenarios when two cities have different population. Therefore, as for our synthetic experiments with different city situations, we use policy $\pi^{\mathbf{0}}$ to investigate in the following scenarios:

1: Both cities $\rightarrow$ same population and equally severe pandemic

2: Both cities $\rightarrow$ same population; Local city $\rightarrow$ more severe pandemic

3: Local city → larger population, and more severe pandemic

4: Local city → larger population; Both cities → equally severe pandemic

5: Local city → larger population; Both cities → equally slight (mild) pandemic

6: Local city → larger population, and slighter pandemic

In Algorithm 1, the exposed passenger number $N_{E,p}$ has a linear relationship with the flight distance D_f with adjustable constant C to model the influence of flight duration on disease transmission. We define $D_f = 1, C = 10$ for the longest direct flight in the world, the flight from Singapore to New Jersey lasting around 18.5 hours and traveling 9,534 miles [8]. Therefore, the flight distance D_f is actually a distance ratio, which can be more than 1, because the flight with transitions could take a longer time and travel larger distances. In Section 6.1.2, we present the evaluation results when applying policy $\pi^{\mathbf{0}}$ to scenarios with varying flight distance, ranging from 0 to 10, to show the influence of flight distance on flight actions and disease transmission.

In Equation 14, we defined three positive coefficients c_p, c_f, c_I for reward function. For each coefficient, we will choose a set of values to train a new Q-table while keeping the other two as default values ($c_p = 1$, $c_f = 1$, $c_I = 10$), then evaluate and compare these new resulting policies. The set of values we choose are as follows:

- passenger reward $c_p \in \{0, 0.5, 1, 2, 5, 10, 20, 50, 100, 200, 500, 1000\}$
- flight cost $c_f \in \{0, 0.5, 1, 2, 5, 10, 20, 50, 100, 200, 500, 1000\}$
- infectious penalty $c_I \in \{0, 1, 2, 5, 10, 20, 50, 100, 200, 500, 1000\}$

6 Results and Discussions

6.1 Synthetic Experiments

6.1.1 Population and epidemic situation

We test policy $\pi^{\mathbf{0}}$ under six synthetic scenarios as described in Section 5.3, and the evaluation results are shown in Figure 6.

In Figure 6, the evaluation results for Scenario 3 are most noticeable with the highest number of total passengers and infectious passengers, and second highest number of flight and total infectious cases in both cities. These visualized results have been shown as a visualization example in Figure 4. A possible reason for this result is that, since the local city has more population and a more severe pandemic situation, and therefore has larger influence on the disease spread between the two cities. The policy suggests a higher amount of flights and passengers, so that more infectious passengers would be transported to the remote city, and thus the infectious number in local city would reduce more quickly. However, this "selfish" decision would cause disastrous disease spread to remote city and in return also do harm to the local city.

We also notice that Scenarios 1 and 2 also have a comparatively high value for the flight number, infectious number and recovered number. The reason might be that, in Scenario 1 & 2, both cities have the same population, even though their initial pandemic severity might be different, after a certain amount

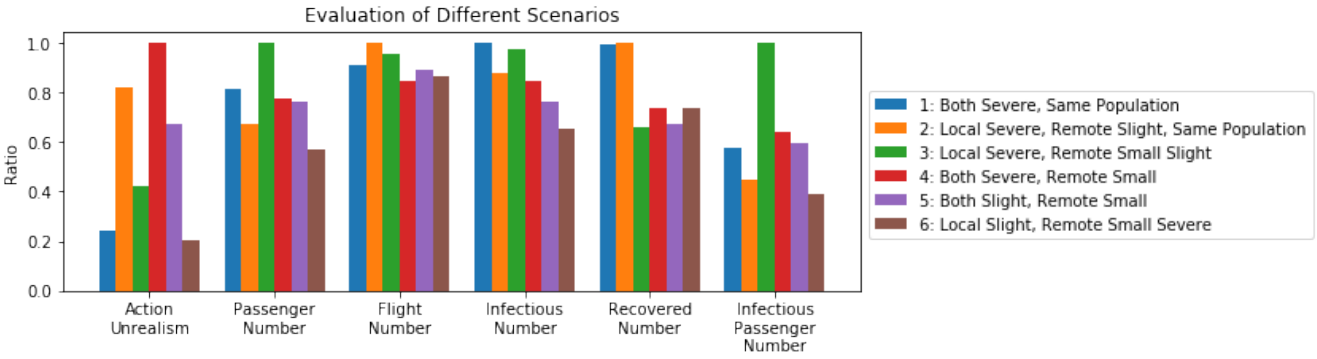

Figure 6: Policy evaluation for different city scenarios.

of time (about 50 days in our simulation), all disease states in the two cities are roughly synchronized, therefore the policy would suggest gradually increasing the flight number while still being cautious with the seat occupancy. However, the cost of this decision would be similar to that of Scenario 3, where the disease transmission might maintain a stable level without vanishing even after a long time.

To see the influence of population and pandemic severity of the remote city for the overall flight decision, we plot the evaluation result of Scenario 1 & 2 with the same population and a more continuous change of remote pandemic severity in Figure 7, and evaluation result of Scenario 4 & 5 with the same severity and a more continuous change of the remote population in Figure 8. Similar to the bar diagrams previous, the y-axises in both figures are also normalized according to the maximum value in each catalog, therefore the comparison of the absolute value across different diagrams is emphasized to be meaningless.

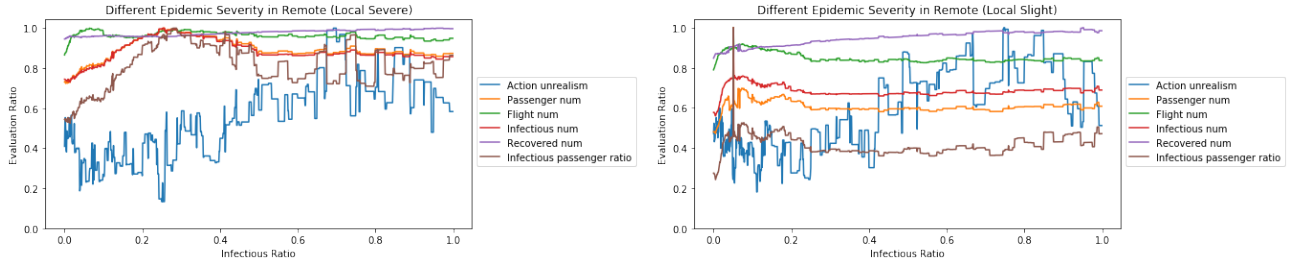

Figure 7: Flight policy evaluation for scenarios of different epidemic severity in the remote city.

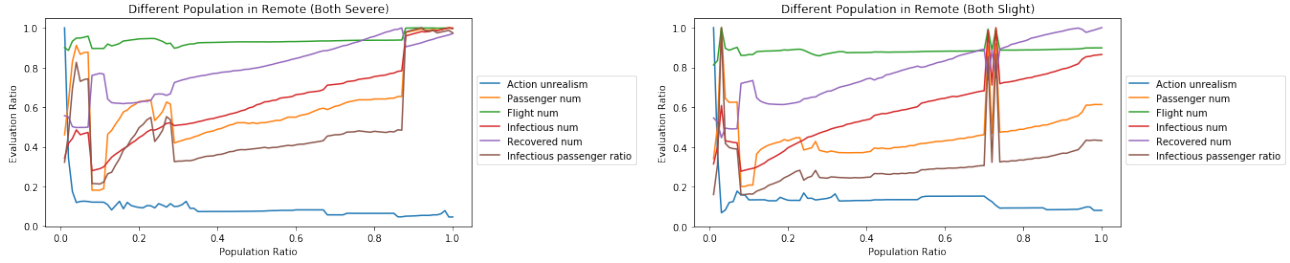

Figure 8: Flight policy evaluation for scenarios of different population in the remote city.

From Figure 7, we notice that when the local pandemic is severe (initial infectious ratio = 0.25), the evaluation curves also have a peak at around 0.25; and when local pandemic is slight (initial infectious ratio = 0.05), the peak is also around 0.05. This means that when the two cities are of same population, our policy $\pi^{\mathbf{0}}$ suggests the highest total passenger number which leads to the highest total number of infectious cases and infectious passengers when the two cities are of the same severity. When the remote infectious ratio becomes unbelievably high, the evaluation curve does not change much because the flight actions are already reduced to the lowest level, making the evaluation reflecting the natural disease spread.

From Figure 8, the evaluation result is a little surprising. It is interesting to see a sharp rise at a population ratio = 0.9 for almost every curve when both cities have a severe pandemic, and a sharp peak at a population ratio = 0.7 when they are both slight pandemics. The reason could be related to our algorithm of how the flight actions effect both cities, but this still needs time investment. We notice that the action unrealism curve reduces as remote the population rises towards the local population. This reason might be that when the population difference between the two cities reduces, the flight actions tend to remain more stable since the disease states in both cities are easier to be synchronized and stabilizable.

6.1.2 Flight distance

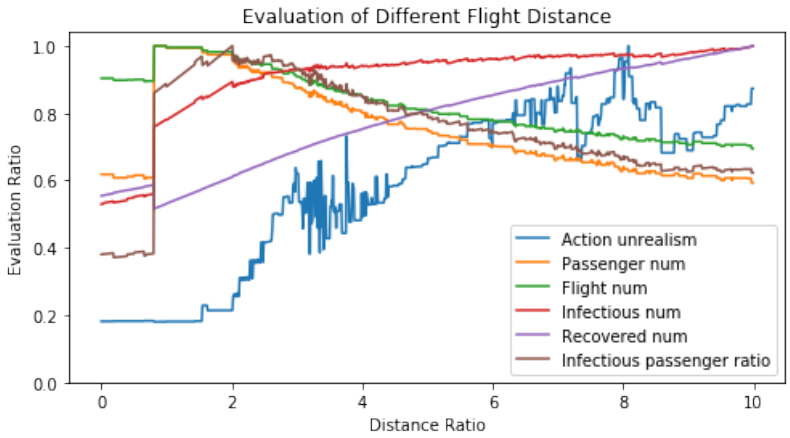

Figure 9: Flight policy evaluation with different flight distance.

As introduced in Section 5.3, we show the evaluation results with regard to the change of flight distance in Figure 9. We notice a sharp rise when $D_f = 0.8$, and then the passenger number, flight number, and infectious flight number decrease steadily, while total infectious number and total recovered number increase steadily. Based on our defined effects of flight distance (or flight duration), the results suggest that the maximum safe traveling duration to be $0.8 *$ maximum direct flight duration, which is equal to about a 14.8 hour flight.

6.1.3 Infectious penalty

We change the infectious penalty coefficient c_I from 0 to 500, train a policy for each value as shown in Section 5.3, and evaluate these policies. The evaluation results are shown in Figure 10. Since each policy is separately trained which is time-consuming, the discrete x-axis is rough but can still reveal the trend. When the infectious penalty c_I is large enough (in our case is about 200 times larger than c_f and c_p), the

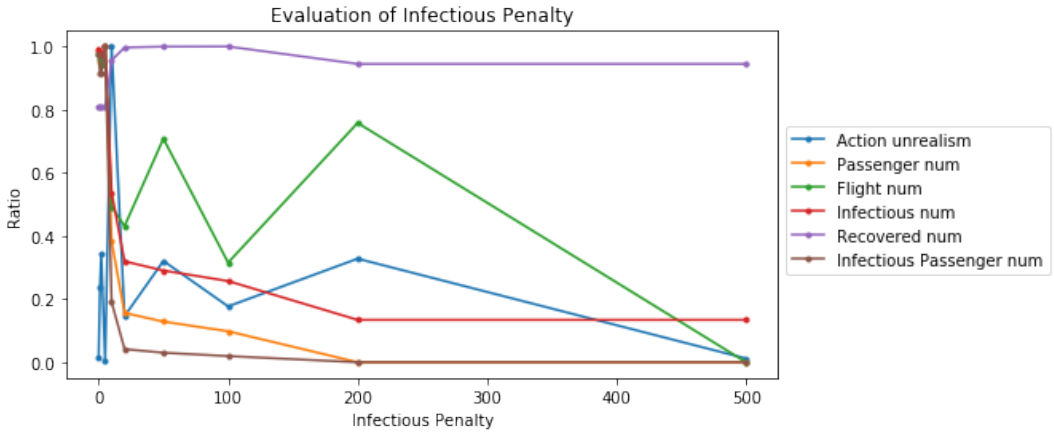

Figure 10: Evaluation of flight policies trained with different infectious penalty parameters.

passenger number reduced to 0, suggesting that the public health is the first priority when making the flight decision, leading to a complete no-flight decision.

If zoom in the plot where c_I is smaller than 5, almost all evaluation results (except for the recovered number and action realism) are near 1, meaning that the importance of disease transmission is overlooked, and the policy tends to suggest a near-maximum action. A more balanced result would be somewhere between these two extremes so our $c_I = 10$ when we trained policy $\pi^{\mathbf{0}}$ was reasonable.

6.1.4 Flight cost

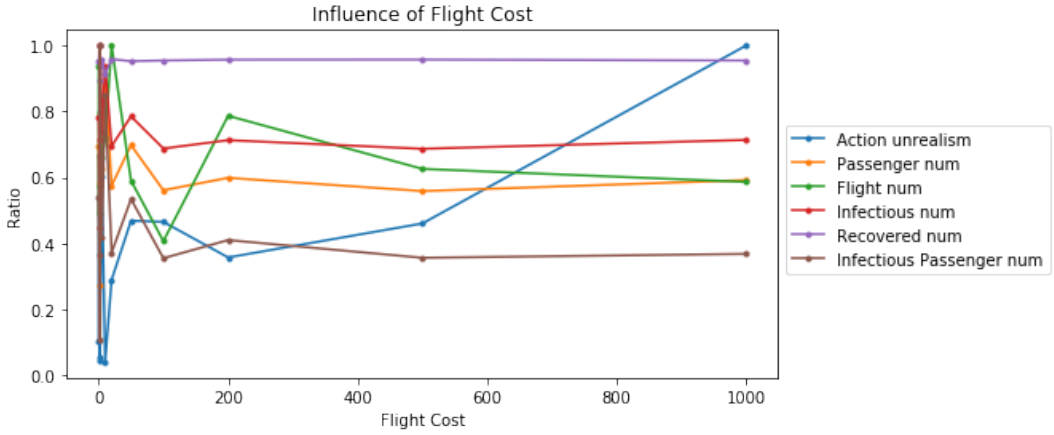

Figure 11: Evaluation of flight policies trained with different flight cost parameters.

We analyzed the flight cost coefficient c_f from 0 to 1000, train a policy for each cost coefficient as shown in Section 5.3, and evaluate these policies. From evaluation results in Figure 11, we notice that when $c_f > 200$, the curves (except for action unrealism) remain stable, suggesting that the influence of flight cost already reaches a saturation level.

When zoomed in to analyze the smaller cost coefficents, we find that when $c_f = 2$, passenger number, infectious number and infectious passenger number reach a high peak, and when $c_f = 1$, they stay in low level. Therefore, it is reasonable for us to set $c_f = 1$ as default value when training policies.

6.1.5 Passenger reward

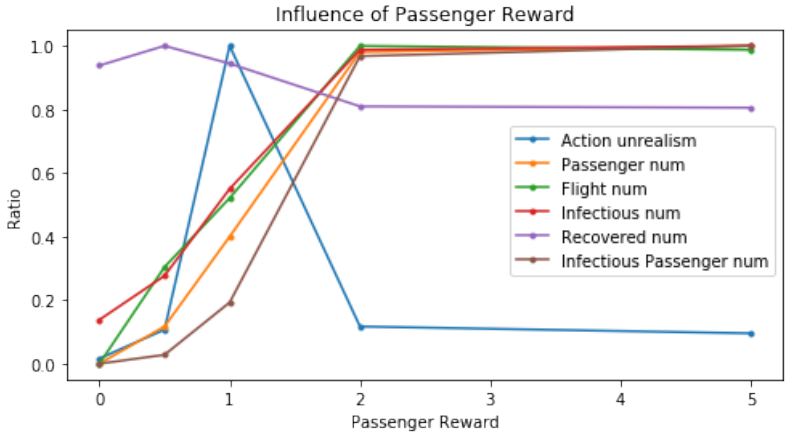

Figure 12: Evaluation of flight policies trained with different passenger reward parameters.

We analyzed the passenger reward coefficient from 0 to 1000, train a policy for each value as shown in Section 5.3, and evaluate these policies. We find that when $c_p > 2$, all the curves remain nearly constant, so we only plot the evaluation results when $c_p \leqslant 5$ in Figure 12 for demonstration. When $0 \leqslant c_p \leqslant 2$, the flight number, passenger number, infectious passenger number and total infectious number all increase along with c_p, showing that the policy simply suggest more flights and passengers with higher reward value for each passenger. After $c_p = 2$, the policy rarely changes because all the flights and seat occupancy have already reach the maximum capacity. Therefore, to properly model the airline revenue from each passenger, it is reasonable that we set $c_p = 1$ as a default value when training policies.

6.2 Real Data Experiments

6.2.1 Population and epidemic situation

We collect data from the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University[5] and test the policy $\pi^{\mathbf{0}}$ by making flight decisions among Shanghai, Los Angeles, New York, San Francisco, and Houston. The reason why choosing these five cities as a representative is based on the consideration of evaluating cities with important and busy airports, different range of population and pandemic phase, to better simulate different scenarios mentioned in Section 5.3 and also as a real-life correspondence to synthetic scenarios in Section 6.1.

Figure 13 shows the real epidemic data in five cities including time-dependent infectious ratio and the amount of population. As we can see from this figure, New York, Los Angeles and Shanghai have relatively large population (more than one million) and Houston, San Francisco have relatively small population (less than one million). New York, Houston have relatively severe pandemic with infectious ratios larger

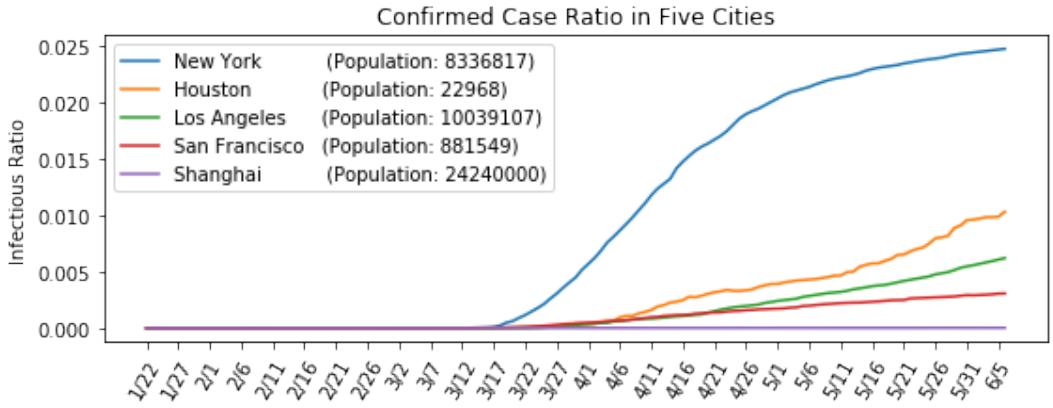

Figure 13: Real epidemic development data in five cities.

than 0.01 and Los Angeles, San Francisco and Shanghai have relatively slight pandemic with infectious ration less than 0.01 after June 5th.

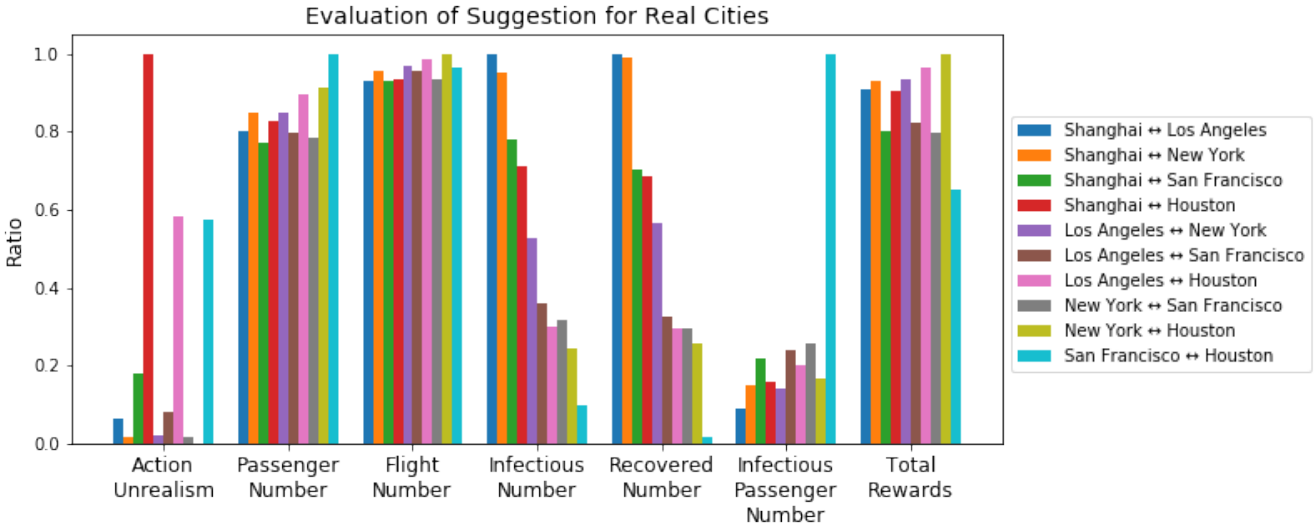

Figure 14: Policy evaluation for different city scenarios using real data.

We apply policy $\pi^{\mathbf{0}}$ on all ten scenarios among these five cities, and evaluate the suggested flight decisions over 365 days since the beginning of the dataset (January 22, 2020). The complete evaluation results are shown in Figure 14.

We observed a high level of action unrealism for flight decision between Shanghai and Houston, Los Angeles and Houston, San Francisco and Houston, suggesting that the action decision might be unrealistic and unreliable with too many action fluctuations. On the other hand, we notice that the flight decision between New York and Los Angeles, New York and San Francisco, New York and Houston could possibly be realistic and worth taking a closer look at. Also, through the visualization of time-dependent state development, we think these three scenarios might reflect some interesting disease spread features so we will to briefly present them in following Sections 6.2.2 to 6.2.4.

6.2.2 New York - Los Angeles

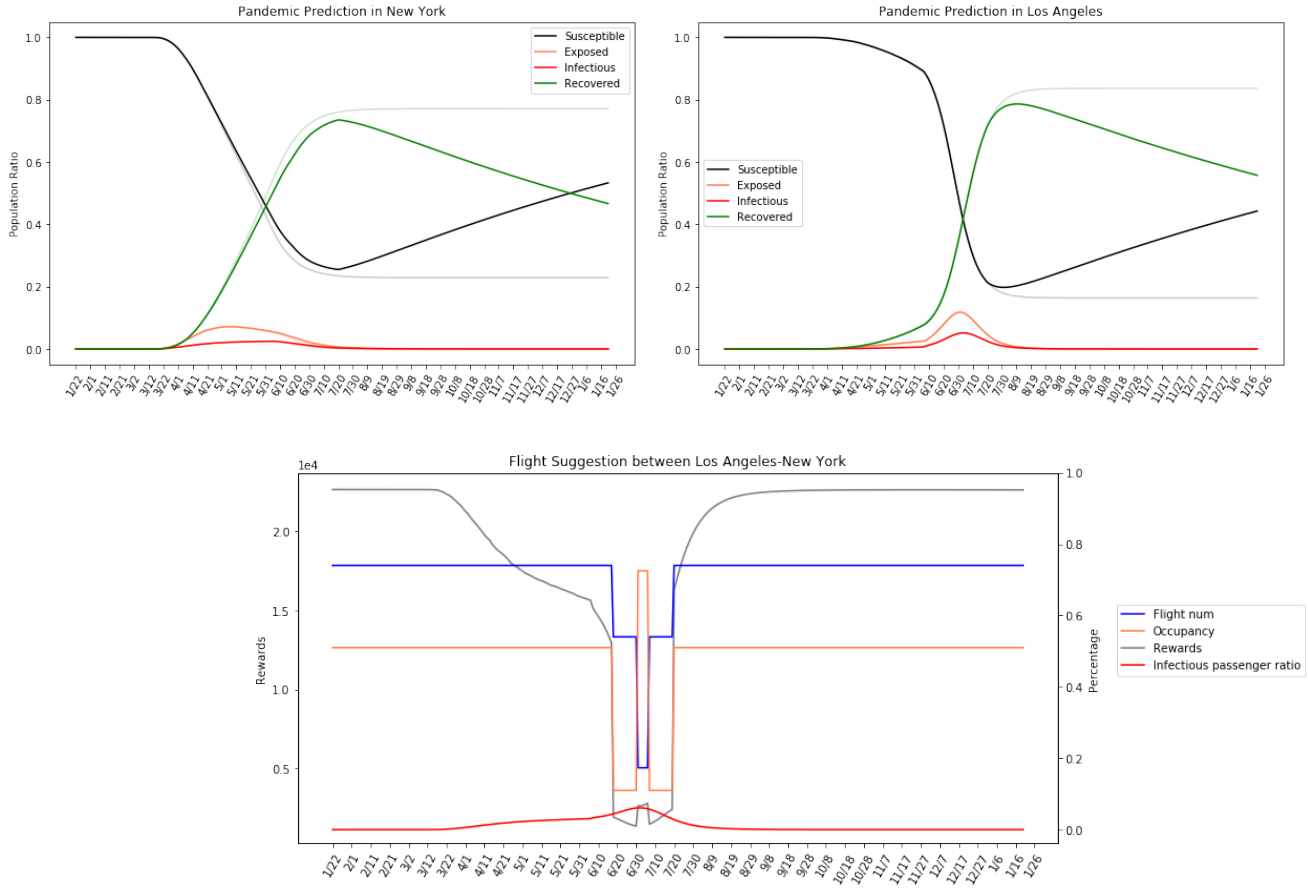

Figure 15: Flight decision between New York and Los Angeles

Figure 15 shows the flight suggestion between New York and Los Angeles from January 22, 2020 to January 23, 2021, along with pandemic prediction in two cities due to the flight decision.

We notice that in both cities, there is not too much difference in Exposed and Infectious number between the no-flight baseline and with flight suggestion. The reason might be a similarity of the population and epidemic circumstance in these two major cities.

We also find that the flight suggestion is only influenced by the pandemic during June 18 and July 20, 2020. During this period, the infectious passenger ratio have a small peak and then gradually return to zero, along with the exposed and infectious ratio in both cities. After July 20, 2020, the policy suggests the flight number returns to 75% of the maximum number of flights, and seat occupancy returns to 50%.

The policy gives a promising flight suggestion and pandemic prediction that the airline industry and disease situation would soon returns to a new normal.

6.2.3 New York - San Francisco

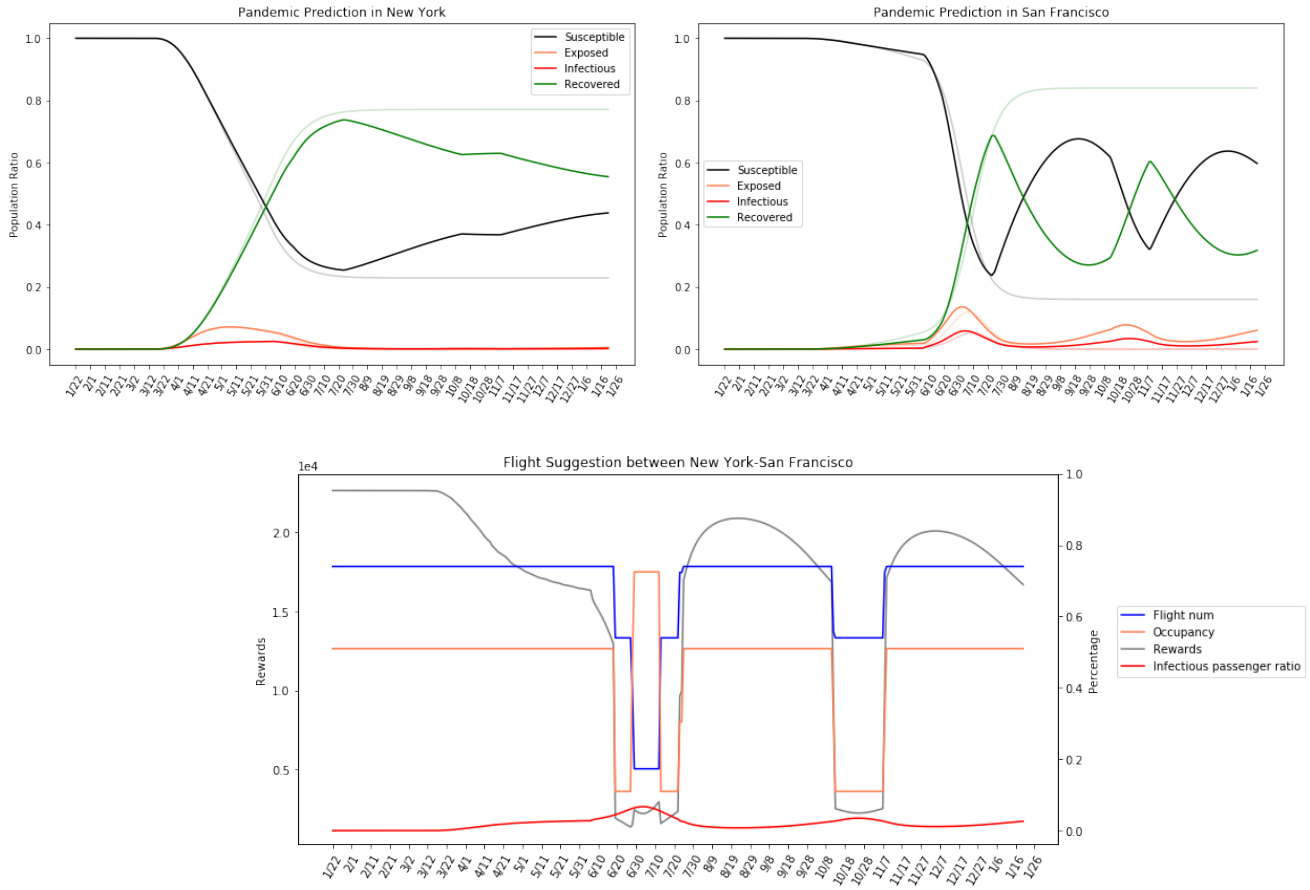

Figure 16: Flight decision between New York and San Francisco

Figure 16 shows the flight suggestion between New York and San Francisco from January 12, 2020 to January 23, 2021, along with pandemic prediction in two cities due to the flight decision.

Unlike no-flight situation and any synthetic scenarios, the pandemic curves show a regular periodicity in San Francisco. We obtain a similar trend when simulating the flight decision between San Francisco and other cities, such as Los Angeles. The estimated recurring time is slightly different depending on which other city we are making the flight decision for, but roughly it is always at the end of October, 2020 or at the beginning of November, 2020. None of the other cities except San Francisco show a recurring trend of the disease, of which the reason is not yet been understood.

The policy $\pi^{\mathbf{0}}$ suggest a flight decision that reduces (flight number, seat occupancy) to (50%, 10%) about a week before the recurring outbreak in the city, and returns to (75%, 50%) when it passes. The period between lasts for about three months.

The flight suggestion and pandemic prediction by the policy could possibly cause some panic, but we might need to take some precautionary measures in case this prediction comes to reality.

6.2.4 New York - Houston

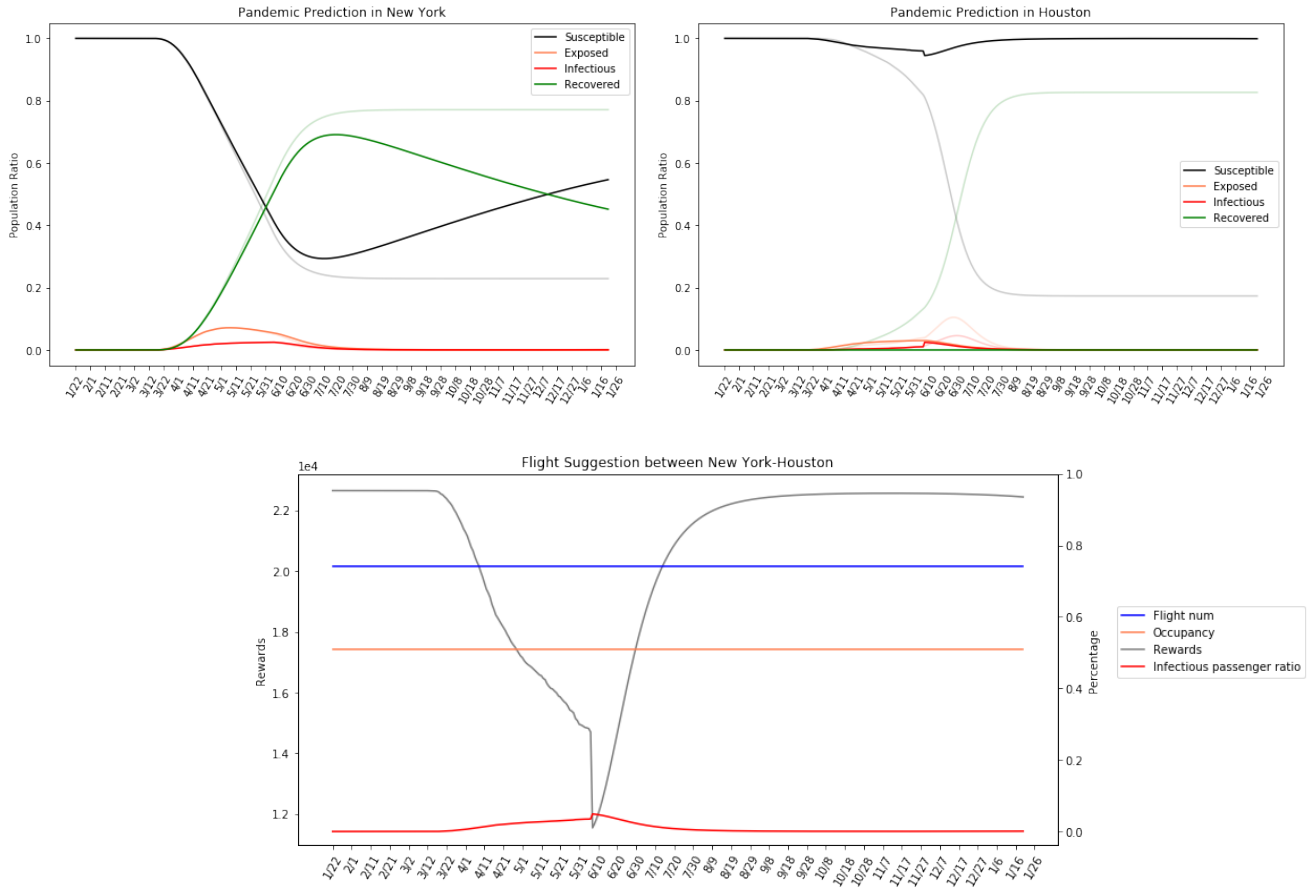

Figure 17: Flight decision between New York and Houston

Figure 17 shows the flight suggestion between New York and Houston from January 12, 2020 to January 23, 2021, along with pandemic prediction in two cities due the flight decision.

The pandemic evolution in Houston is yet another unexpected simulated result, with the number of Exposed and Infectious in Houston growing even lower than those of no-flight baseline. In other words, flights between New York and Houston do not deteriorate the pandemic situation in Houston but cure it instead. The pandemic peak in infectious state even disappears at the start of the rising trend.

One of the reasons for this counter-intuitive result could possibly be simply a technical mistake. Since the Q-table in policy $\pi^{\mathbf{0}}$ has a size of $40 * 40 * 150 * 100$, the ratio resolution in state space is $1/40 = 0.025$. Therefore when the change of all the disease state ratios are smaller than the resolution, the flight decision could go wrong.

On the other hand, another possible explanation is herd immunization. Given that the population in Houston is no more than 0.3% compared to New York, and also considering the high infectious ratio in New York, the infectious passenger number could possibly turn to a high number of susceptible individuals in Houston city after flights, meaning that the majority of Houston population would become immune to the disease before the real pandemic peak arrives. Therefore, Houston does not appear to be influenced

too much by the disease in our simulation.

The herd immunization could also explain why policy $\pi^{\mathbf{0}}$ suggests a constant flight number (remains at 75%) and seat occupancy (stays at 50%) over the whole period of simulation time.

In summary, the three scenarios with real data in Section 6.2.2 to 6.2.4 show three possibilities of post-COVID19 world development:

- New York - Los Angeles: The pandemic situation in both cities are resolved after several months, the airline industry returns to a new stable state which is lower than the maximum availability before pandemic.
- New York - San Francisco: The pandemic situation reoccurs in a regular periodicity, the airline industry repeatedly drops and rises to deal with the recurring.
- New York - Houston: The pandemic situation resolves due to herd immunization, the airline industry remains stable over time.

7 Limitations and Future Work

Based on our decision-making results in both the synthetic experiments and real-data experiments, most of the disease states and flight decisions are influenced by the pandemic situation for no more than 30 days, as shown in Figure 4, 15-17, which is not true because the lock-down has already lasted for more than two months, and the influence of pandemic on disease states and flights still lasts. Some possible reasons for the pandemic prediction and flight policy to be not accurate or realistic are as follows.

Firstly, the Q-learning algorithm that we used has built-in limitations. As we explained in Section 4.1, Q-table utilizes a discretization of the continuous state and action space, so it is inevitable to involve some approximation. The performance of the policy is limited by the resolution of discretization, which is related to the size and dimension of Q-table, constrained by the computation ability and training episodes and time. In order to improve the performance, we can use algorithms such as Deep Q-Network (DQN), Double Deep Q-Network (DDQN), and et al.

Secondly, the dual-city model in our method is a rough simulation of the reality. We make the assumption that the two cities are only affected by flights between each other, while in reality, each city has multiple remote cities to commute with. For simplicity, we consider all the remote cities as a whole, and use the flight action to update local city and this simplified remote city. This will cause certain problems because different remote cities have much more complicated pandemic situations, including infectious ratios, airline conditions and also government policies. To make the city and flight environment more accurate and multi-city model should implemented.

Also, in our state evolution algorithm (Algorithm 2), our model assumes both cities have the same disease evolution rate and remains in the epidemic phase only with different initial conditions. However, in reality many cities may be in different phases, so further investigation should be carried out for different cities to be in different epidemic environments.

Another direction to improve our decision-making model is to modify the SEIR-model, including finding more accurate parameters to describe the spread of disease and the effect of flights, and adding more disease states to the model. For the first one, we are currently using a same set of (β, σ, γ) for the rate of

the spread, incubation and recovery, which are a set of global average values. However, it is proved that Coronavirus is slowly mutating [10], indicating that the virus may not remain the same in all the cities over time, therefore the virus-related parameters should be customized to each city. A possible approach is to use machine learning algorithms to learn these variables for each city by using past confirmed case numbers. Furthermore, we describe the effects of flight duration with a ratio of flight distance and some adjusting constant, which could be a quite rough estimation. It might also be possible to use machine learning method to train and obtain these variables using real laboratory data.

For modification of the SEIR model, we notice that the current model does not include vital dynamics, especially the death rate of the disease, which is not a sufficient simulation of the reality, therefore we can edit the SEIR model to SEIRD model, which involves death cases and takes change of population into consideration[11]. SEIRS model is another modification which considers the temporary immunity and virus mutation, so that the recovered individuals still have a possibility to become susceptible [12]. SuEIR model pays extra attention on the asymptomatic and unreported cases of COVID-19, which is trained by machine learning algorithms based on the reported historical data. [13].

Finally, there are certain countries that implement a 14-day quarantine rule for all inbound passengers, taking China as an example, which we do not take into account. If this effective prevention policy becomes a routine for more cities and countries, it is possible to modify our model to allow more and earlier flight recovery than current prediction.

8 Conclusions

In order to make flight decisions for airline industry while considering both the economics and public health in the long term and short term, we implemented a Q-learning algorithm to look for an optimal flight policy. We also defined six variables as our evaluation method to evaluate the performance of resulting policies. We tested the policies on synthetic experiments with different city scenarios and reward coefficients, as well as real-data experiments. From these simulations, we observed three possibilities of post-COVID19 world development: returning to a new normal which is slightly less than the maximum level before pandemic, facing a reoccurring pandemic situation in a regular periodicity, and remaining stable over time due to immunization. For each of these scenarios and possibilities, our model provides a flight policy decision. In the future, we wish to further investigate a more accurate and realistic model to construct flight policies across the globe, making the world a more informed, healthier place to live.

References

[1] Effects of Novel Coronavirus (COVID-19) on Civil Aviation: Economic Impact Analysis. 1 May 2020. http://www.capsca.org.

[2] The Post-COVID-19 Flight Plan for Airlines. 31 March 2020. https://www.bcg.com.

[3] Binti Hamzah FA, Lau C, Nazri H, Ligot DV, Lee G, Tan CL, et al. CoronaTracker: Worldwide COVID-19 Outbreak Data Analysis and Prediction. [Preprint]. *Bull World Health Organ*. E-pub: 19 March 2020. doi: http://dx.doi.org/10.2471/BLT.20.255695

[4] Gupta, Jitendra K., Chao-Hsin Lin, and Qingyan Chen. Transport of expiratory droplets in an aircraft cabin. Indoor Air 21.1 (2011): 3-11.

[5] COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University. https://github.com.

[6] 2014 National Civil Aviation Flight Operation Efficiency Report. May 2015. http://www.cata.org.cn.

[7] Schulman, John, et al. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347* (2017).

[8] Lakritz, Talia. 12 of the longest flight routes in the world. *Insider*. Mar 10, 2020. https://www.insider.com/longest-flight-routes-in-the-world

[9] Poole, David L., and Mackworth, Alan K. Artificial Intelligence: foundations of computational agents (2nd Edition). *Cambridge University Press*, 2010.

[10] Draghi, Jeremy, and Ogbunu C. Brandon. The Coronavirus Is Mutating. That's Not Necessarily Good or Bad. *Undark*. May 14, 2020. https://undark.org/2020/05/14/covid-19-evolution-mutation/

[11] Piccolomini, Elena Loli, and Fabiana Zama. Preliminary analysis of COVID-19 spread in Italy with an adaptive SEIRD model. *arXiv preprint arXiv:2003.09909* (2020).

[12] Trawicki, M. B. (2017). Deterministic Seirs Epidemic Model for Modeling Vital Dynamics, Vaccinations, and Temporary Immunity. Mathematics, 5(1), 7. https://www.mdpi.com/2227-7390/5/1/7

[13] Zou, Difan, et al. Epidemic Model Guided Machine Learning for COVID-19 Forecasts in the United States. *medRxiv* (2020).