SIRV-RL: Reinforcement Learning for Optimized Policy Control during Epidemiological Outbreaks in Emerging Market and Developing Economies

Maeghal Jain[1,*] **and Ziya Uddin**[1]

[1]BML Munjal University
[*]e-mail: maeghaljain@gmail.com

ABSTRACT

The outbreak of COVID-19 has highlighted the intricate interplay between public health and economic stability on a global scale. This paper aims to develop a reinforcement learning framework to balance health and economic outcomes during infectious disease outbreaks. The framework utilizes the SIR model without vital dynamics and incorporates globally comparable government responses. The study acknowledges the limitations of deterministic models and proposes the use of deep reinforcement learning to reduce input dimensionality and normalize input. While the model offers transparency in terms of its dependency on the reward policy defined, it recognizes the need for a comprehensive consideration of decision factors beyond pure reinforcement learning results.

Introduction

In the past, global spread of infectious diseases was largely due to colonization, slavery, and war, leading to widespread illness and death from diseases like tuberculosis, polio, smallpox, and diphtheria. Medical advancements, better access to health care, and improved sanitation have worked towards improving the situation of mortality and morbidity linked to infectious diseases in the past twenty years. However, in low and lower-middle income countries the burden of infectious diseases still persists. The rapid pace of urbanization in low and middle-income countries, along with the rise in populations living in crowded, poor-quality homes, has led to new conditions that favor the emergence of infectious diseases[?,?].

Recently, the COVID-19 pandemic caused a havoc worlwide. Till date there have been 772 million cases and more than 6 million deaths[?]. The pandemic triggered the sharpest economic recession in modern history with a 3% decline, much worse than during the 2008-09 financial crisis[?]. As nations grappled with the immediate health crisis, the economic fallout disproportionately affected vulnerable populations and exacerbated existing inequalities. Lockdowns and restrictions imposed to curb the spread of the virus led to widespread unemployment, business closures, and disruptions in global supply chains[?]. The challenges faced by low and lower-middle income countries were particularly acute, highlighting the intricate interplay between public health and economic stability on a global scale[?].

The need for a nuanced understanding of how interventions impact both health outcomes and economic indicators became increasingly evident, prompting a comprehensive examination by epidemiologists to assist policy makers[?]. The outbreak of COVID-19 has prompted epidemiologists to research on various aspects, including mobility control[?,?], vaccination strategies[?,?], stringency measures/non-pharmaceutical interventions (NPIs)[?,?,?], and financial considerations[?]. Despite the numerous studies conducted, very few explore how common interventions meet multiple policy objectives or how a precise articulation of the main policy goals directs the selection of the most effective interventions in terms of health and economic results[?,?,?,?,?,?,?]. The economic impact of the COVID-19 pandemic varied between rich and poor countries.

Although COVID-19 deaths had a slightly larger negative effect on the Gross Domestic Product (GDP) in advanced economies, this difference was not statistically significant. However, lockdown restrictions were found to have a more damaging impact on economic activity in emerging and developing economies[?,?,?]. It's also suggested that an increase in COVID-19 cases was associated with the introduction of harsher NPIs and lockdown measures could be relaxed once vaccination rates increase[?,?].

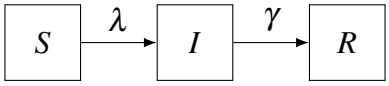
Many economists have studied the effect of COVID-19 on the economy of nations[?,?,?,?,?]. In advanced economies like Korea, where the stringency index was below the median the recession was milder than other advanced economies like the United Kingdom where the stringency was much higher[?], they achieved it mostly with very aggressive testing, contact tracing, and enforced quarantines and isolations[?,?]. In India, social distancing and containment measures have been effective in reducing the number of COVID-19 cases but have come with economic costs. Social distancing had the most adverse effect on the economy in areas with high urbanization[?].

In this paper we want to optimize the government policies regarding stringency. Therefore, alongside epidemiological data, we use the measures of globally comparable government responses[?]. We use the simple SIR model without vital dynamics[?,?,?] as it is assumed that the time scale is short enough so that can be neglected[?]. By lesioning the model, as opposed to proposing a new mathematical model with more specialized compartments to more accurately represent the actual environment[?,?], we can effectively address the real-world conditions and propose a solution that is both effective and extendable. The current model accounts for the recovery reached through vaccination[?,?,?,?,?]. However, the study has limitations; the deterministic SIR model fails to account for chance in disease spread and lacks confidence intervals on results and while stochastic models incorporate chance, they are typically more challenging to analyze than their deterministic counterparts[?].

In order to capture competing costs within the environment and achieve a balance between health and economic outcomes, we intend to employ reinforcement learning[?,?,?,?,?]. Not only does the formulation of the model deal better with competing costs, but it also offers a more transparency behind the reasoning of the decisions being made in such circumstances. When we conceptualize our problem as a reinforcement learning task, an agent is tasked with making decisions in an environment with the aim of optimizing cumulative rewards. This approach relies on trial and error, in contrast to dynamic programming, which assumes complete knowledge of the environment[?]. We make the use of deep reinforcement learning, which is an advancement to reinforcement learning as it allows to reduce the input dimensionality and normalize the input[?,?,?,?]. While our model is more transparent in terms of its dependency on the reward policy defined, it has its limitations. A universal optimal policy may not suit diverse socioeconomic contexts due to variations in healthcare resources and economic vulnerabilities across countries, regions, or cities and a comprehensive consideration of decision factors, extending beyond pure reinforcement learning results is needed[?,?,?].

Mathematical Formulation

In SIR-RL, we use a compartmental model to model infectious disease. We iteratively develop this model to fit the actual data better. In an SIR model people of the population are divided based on whether they are yet to come into contact with an infected person (Susceptible), are infectious themselves (Infectious), or have recovered from the infection (Recovered). These compartments create the SIR model which can be represented as follows:



$$\frac{dS}{dt} = -\lambda S \tag{1}$$

$$\frac{dI}{dt} = \lambda S - \gamma I \tag{2}$$

$$\frac{dR}{dt} = \gamma I \tag{3}$$

Here, λ is the force of infection, it is the rate at which susceptible individuals acquire an infectious disease[?]. It depends on other factors:

$$\lambda = pc\frac{I}{N} \tag{4}$$

Here, c is the average number of contacts a susceptible person makes per day. p is the probability of the susceptible person becomes infectious after coming into contact with an infectious person. $\frac{I}{N}$ is the proportion of the contacts that are infectious.

And, β the effective transmission rate is defined as:

$$\beta = pc \tag{5}$$

During an epidemic, the fundamental drivers of an epidemic growth is the rate of infection β i.e. the average number of infections per infected case and the infectious period $1/\gamma$ i.e. the average period for which the infected case is infected for. Epidemics can only happen if the case is infectious enough for long enough and this defined by $R_0 = \beta/\gamma$. Here, R_0 is The average number of secondary infections caused by each infected case, in an otherwise fully susceptible population.

At the peak of an epidemic R_0 declines as there are no more susceptible people left in the pool, therefore, R_{eff} (effective reproductive number) comes into play. R_{eff} is defined as the average number of secondary cases arising from an infected case, at a given point in an epidemic - therefore, it takes into account the existing immunity of the system.

$$R_{eff} = R_0\frac{S}{N} \tag{6}$$

S is the number of susceptible people, N is the total population. At the start of an epidemic when everyone is susceptible, $R_{eff} = R_0$ as, $S = N$ (i.e. the whole population is susceptible). β and γ are also used to define probability of and infectious individual infecting another individual $\beta/(\beta+\gamma)$ and the probability of recovery, $\gamma/(\beta+\gamma)$.

Most government policies look at the value of R_{eff} to come us with an effective strategy to combat the disease as the fate of the evolution of the disease depends upon it. When R_{eff} is less than one, the infected population I will steadily decline to zero. Conversely, if R_{eff} is greater than one, the infected population will increase. In other words, when $\frac{dI(t)}{dt} < 0 \Rightarrow R_{eff} < 1$ and $\frac{dI(t)}{dt} > 0 \Rightarrow R_{eff} > 1$, the

effective reproductive rate R_{eff} serves as a critical threshold that determines whether an infectious disease will rapidly extinguish or escalate into an epidemic[?].

To estimate the parameters β and γ for India, based on the data from May, 2020, to October, 2022, we simply define the cost to calibrate the model using the huber loss[?]:

$$L_\delta(y, f(x)) = \begin{cases} \frac{1}{2}(y - f(x))^2 & \text{for } |y - f(x)| \leq \delta, \\ \delta \cdot (|y - f(x)| - \frac{1}{2}\delta) & \text{otherwise.} \end{cases} \tag{7}$$

$$cost = L_{\delta=1}(S, \hat{S}) + L_{\delta=1}(I, \hat{I}) + L_{\delta=1}(R, \hat{R}) \tag{8}$$

Where, S is the number of susceptible people and $\hat{S}$ is the predicted number of susceptible people, similarly, I for infected and $\hat{I}$ for the predicted number of infected people, and R for recovered. Minimizing this cost function using the Nelder-Mead method[?], we get:

$$\beta_{optimal} = 0.04208479828695971 \tag{9}$$

$$\gamma_{optimal} = 0.02388356686032017 \tag{10}$$

$$R_{0_{optimal}} = \frac{\beta_{optimal}}{\gamma_{optimal}} = 1.7620817917644795 \tag{11}$$

$$cost = L_{\delta=1}(S, \hat{S}) + L_{\delta=1}(I, \hat{I}) + L_{\delta=1}(R, \hat{R}) = 85051490.53250012 \tag{12}$$

See figure **??** to see how the model compares with the real data.

Now that, we have a model – we want to say β is a time-varying parameter controlled by the stringency index, s[?,?,?]. The stringency index is a composite measure based on nine response indicators including school closures, workplace closures, and travel bans, rescaled to a value from 0 to 100 (100 = strictest)[?]. This index simply records the strictness of government policies and does not measure or imply the appropriateness or effectiveness of a country's response i.e. a higher score does not necessarily mean that a country's response is "better" than others lower on the index.

To define the new time-varying beta that is dependent on the current stringency index, we say,

$$\frac{dS}{dt} = -\beta(1 - (s(t)/100))\frac{SI}{N} \tag{13}$$

$$\frac{dI}{dt} = \beta(1 - (s(t)/100))\frac{SI}{N} - \gamma I \tag{14}$$

$$\frac{dR}{dt} = \gamma I \tag{15}$$

Where, $s(t)$ is the stringency index at time t and is scaled down by a factor of 100 to normalize it and bring it in the range $[0,1]$. Optimizing this using the Nelder-Mead method we get:

$$\beta_{optimal} = 0.4013340889432941 \tag{16}$$

$$\gamma_{optimal} = 0.09017476605499258 \tag{17}$$

$$R_{0_{optimal}} = \frac{\beta_{optimal}}{\gamma_{optimal}} = 4.450625230328213 \tag{18}$$

$$cost = L_{\delta=1}(S,\hat{S}) + L_{\delta=1}(I,\hat{I}) + L_{\delta=1}(R,\hat{R}) = 98438821.45587364 \tag{19}$$

Finally, an additional flow from the susceptible population to the can be shown by adding a vaccination rate ν in the model.

$$\frac{dS}{dt} = -\beta(1-(s(t)/100))\frac{SI}{N} - \nu S \tag{20}$$

$$\frac{dI}{dt} = \beta(1-(s(t)/100))\frac{SI}{N} - \gamma I \tag{21}$$

$$\frac{dR}{dt} = \gamma I + \nu S \tag{22}$$

Optimizing these equations with the Nelder-Mead method we get:

$$\beta_{optimal} = 0.40897034072952304 \tag{23}$$

$$\gamma_{optimal} = 0.09196829370123338 \tag{24}$$

$$\nu_{optimal} = 2.9044029843851936e-05 \tag{25}$$

$$R_{0_{optimal}} = \frac{\beta_{optimal}}{\gamma_{optimal}} = 4.446862329077204 \tag{26}$$

$$cost = L_{\delta=1}(S,\hat{S}) + L_{\delta=1}(I,\hat{I}) + L_{\delta=1}(R,\hat{R}) = 94636860.38436058 \tag{27}$$

However, as observed by the value of $\nu_{optimal}$ from **??** which is almost negligible (close to zero) and the overestimation of infected individuals in **??** suggests that ν might be varying with time. This suggests a time-varying vaccination rate, wherein the transition from susceptibility to direct recovery fluctuates with time. Therefore, using the $\beta_{optimal}$ and $\gamma_{optimal}$ from **??** and **??**, we optimize the value of ν given a sub-interval of $[start, start + window]$ where the $window = 20^{?}$. Here, *window* is a hyperparameter that can be changed so as to reduce the cost function.

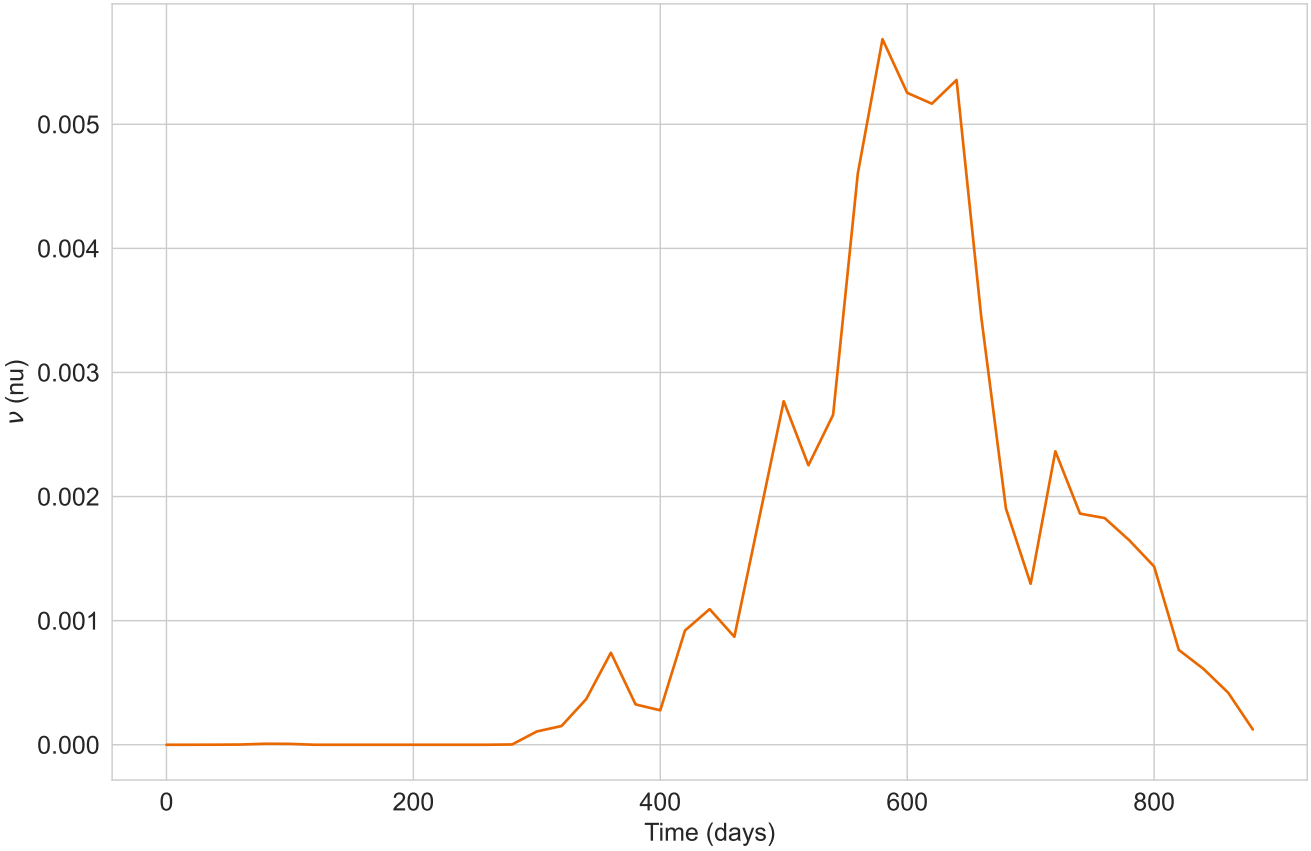
$$\frac{dS}{dt} = -\beta_{optimal}(1 - (s(t)/100))\frac{SI}{N} - \nu S \tag{28}$$

$$\frac{dI}{dt} = \beta_{optimal}(1 - (s(t)/100))\frac{SI}{N} - \gamma_{optimal} I \tag{29}$$

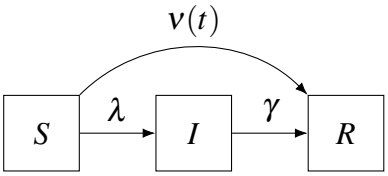
$$\frac{dR}{dt} = \gamma_{optimal} I + \nu S \tag{30}$$

This model gives us how ν varies with time.

Figure 1. ν Varying with Time



Therefore, using these values we finally, recompute $\beta_{optimal}$ and $\gamma_{optimal}$ by supplying them into the equations below:



$$\frac{dS}{dt} = -\beta(1-(s(t)/100))\frac{SI}{N} - \nu(t)S \tag{31}$$

$$\frac{dI}{dt} = \beta(1-(s(t)/100))\frac{SI}{N} - \gamma I \tag{32}$$

$$\frac{dR}{dt} = \gamma I + \nu(t)S \tag{33}$$

$$\beta_{optimal} = 0.5129414911377119 \tag{34}$$

$$\gamma_{optimal} = 0.12716245167626106 \tag{35}$$

$$R_{0_{optimal}} = \frac{\beta_{optimal}}{\gamma_{optimal}} = 4.033749620081199 \tag{36}$$

$$cost = L_{\delta=1}(S,\hat{S}) + L_{\delta=1}(I,\hat{I}) + L_{\delta=1}(R,\hat{R}) = 29742589.698910963 \tag{37}$$

See (**??**, **??**) to see how the different models compare against each other.

Now, that we have set up the following relation between β and $s(t)$, we investigate how stringency index affects the normalized Gross domestic product (GDP). To do this we fit a polynomial equation of the third degree to the points (x,y) where x is the $s(t)$ stringency at time t, and y the normalized GDP and minimize the squared error. For India after fitting a 3 degree polynomial we get the following equation:

$$\begin{aligned} GDP = -5.96640236e-05s^3 + 6.65064332e-03s^2 - 2.23109924e-01s^1 \\ +1.01357226e+02 \end{aligned} \tag{38}$$

Given that the government is an agent that takes decisions in a deterministic environment defined above, we use reinforcement learning to model the competing costs of the environment. This environment is formally as a Markov decision process, and can be described as follows:

- Set of States $\mathcal{S}$: The state of the environment are described through the descriptors like the normalized GDP ($(GDP_{predicted} - GDP_{min})/(GDP_{max} - GDP_{min})$), R_{eff}, a list of all the previous actions (in changing the stringency) and the proportion of the population that was susceptible, infected and recovered. The starting states are simply these values at the starting date and no previous actions.

- TODO: FIX BELOW ITEM AFTER MODEL RUNS

- Actions $\mathcal{A}$: The stringency index variable was analyzed with a sample size of 1034. The mean value was approximately 59.62, with a standard deviation of 22.82. The minimum value was 0, while the maximum value reached 100. And the differences between two consecutive stringencies had a mean of 0.030474, and standard deviation of 2.387331, with the minimum being -14.360000 and maximum 55.560000. Based on this we define the discrete action space. There are 7 actions for the agent, it can keep the stringency index same, reduce/increase by 2.5, reduce/increase by 5, and reduce/increase by 10 given that the stringency index doesn't exceed 100 or go below 0.

- Transition dynamics $\mathcal{T}(\mathbf{s}_{t+1} \mid \mathbf{s}_t, \mathbf{a}_t)$ that map a state-action pair at time t onto a distribution of states at time $t+1$. This state transition is defined by the SIR model with lockdown and the model of how stringency index affects the GDP.

- Immediate reward $\mathcal{R}(\mathbf{s}_t, \mathbf{a}_t, \mathbf{s}_{t+1})$. Here we define a reward strategy however it should be noted that this strategy can be easily changed to prioritize different needs.

- Discount Factor $\gamma \in [0,1]$, where lower values place more emphasis on immediate rewards. Here, we choose the default discount factor of 0.99.

In general, the policy π is a mapping from states to a probability distribution over actions: $\pi : \mathcal{S} \to p(\mathcal{A} = \mathbf{a} \mid \mathcal{S})$. If the MDP is episodic, i.e., the state is reset after each episode of length T, then the sequence of states, actions and rewards in an episode constitutes a trajectory or rollout of the policy. Every rollout of a policy accumulates rewards from the environment, resulting in the return $R = \sum_{t=0}^{T-1} \gamma^t r_{t+1}$. The goal of RL is to find an optimal policy, π^*, which achieves the maximum expected return from all states.

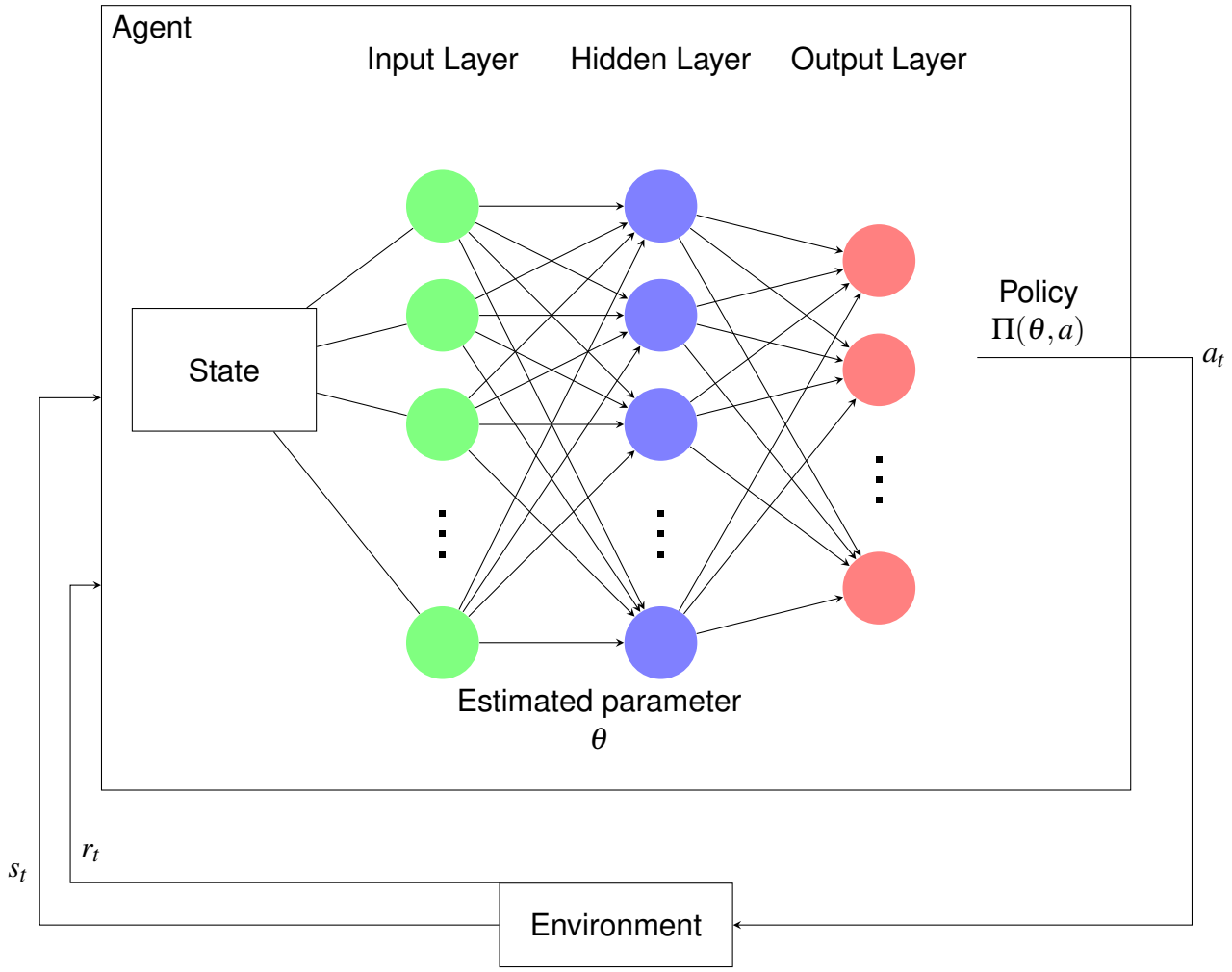
The stringency index emerges as a critical factor influencing both the Gross Domestic Product (GDP) and the rate of infection spread. The decision to escalate or de-escalate the stringency index is a strategic one, with significant implications. Increasing the stringency decreases the spread of the infection. Conversely, it must be noted that herd immunity can only be achieved when the epidemic reaches its peak i.e. when the effective reproductive number is equal to one ($R_{eff} = 1$). This can only happen by lowering the stringency index which would allow the natural dynamics of the epidemic to transpire such that the population of susceptible individuals has depleted enough such that it is insufficient to propagate the disease further. Therefore, stringency is used to control the number of infected people and slow down the rate at which the epidemic reaches its peak, so that hospitals could house the number of infected people.

With this basis we define the reward function**?**. In Deep Reinforcement Learning (DRL), positive rewards promote and negative rewards demote actions. The agent tries to generate such a policy/knowledge to avoid the discouraging situation by following the policy. By designing a proper reward function, it is possible to generate such an agent that may follow the human desired situation.

The reward function is parameterized to account for key factors influencing decision-making. To incentivize reduction of $Reff$ (effective reproductive number) a negative reward is imposed of $-20 * R_{eff}$, but as the R_{eff} is between $[1.9, 1, 5]$ we start to positively reward it with a multiple of the GDP ($100 * \text{min\_max\_normalized\_GDP}$) and if the R_{eff} is below 1.5 then $200 * \text{min\_max\_normalized\_GDP}$. Furthermore, there's a simple positive reward of 10 if R_{eff} is below the threshold value (1.9) and negative

reward of -10 otherwise. To reward not changing the stringencies frequently, we reward the absolute different between the previous stringency and the current stringency negatively ($|s(t)-s(t-1)|*-1*5$) [TODO: this factor should be reduced back to 2 from 5]. If the proportion of the infected population were to rise above 0.003 the model is punished (-5000) and otherwise rewarded ($+20$). It should be realized there can be an infinite number of ways to design the reward function to be more human and upgrade the way a decision is taken given the situation[?]. Therefore, this research act as a framework for promoting the development of more efficient reward strategies for the same.

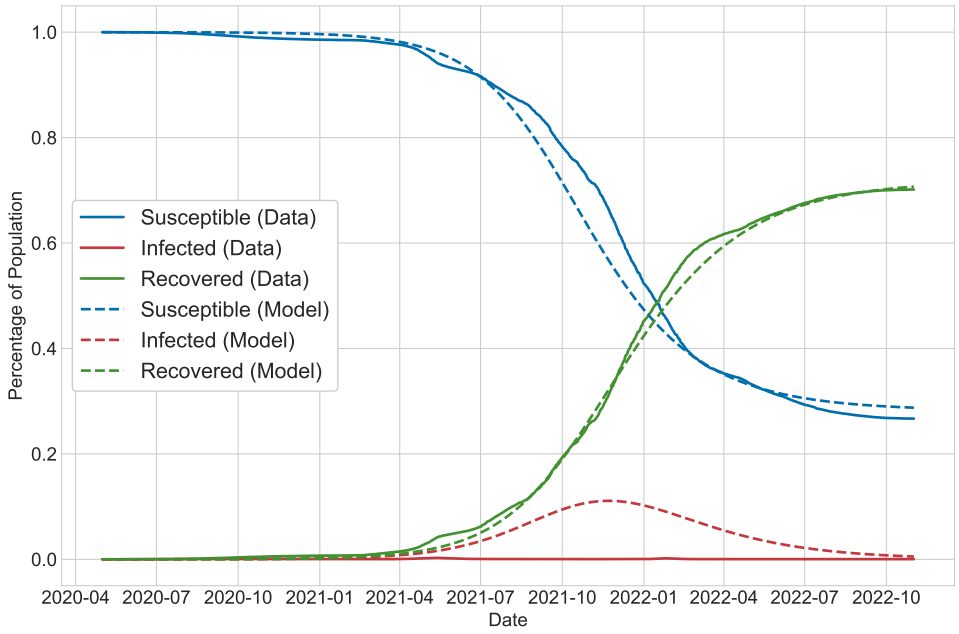
The agent observes the percentage of the population that is susceptible, infected, recovered, and the GDP which extrapolated from the stringency index and all the past moves played, i.e., all the past stringency indexes decided by the agent. The following values are fed into a simple network. Stable Baselines3 supports multiple inputs by using Dict Gym space. This can be done using MultiInputPolicy, which by default uses the CombinedExtractor features extractor to turn multiple inputs into a single vector, handled by the network arch network. For data that varies with time we use a simple LSTM architecture.



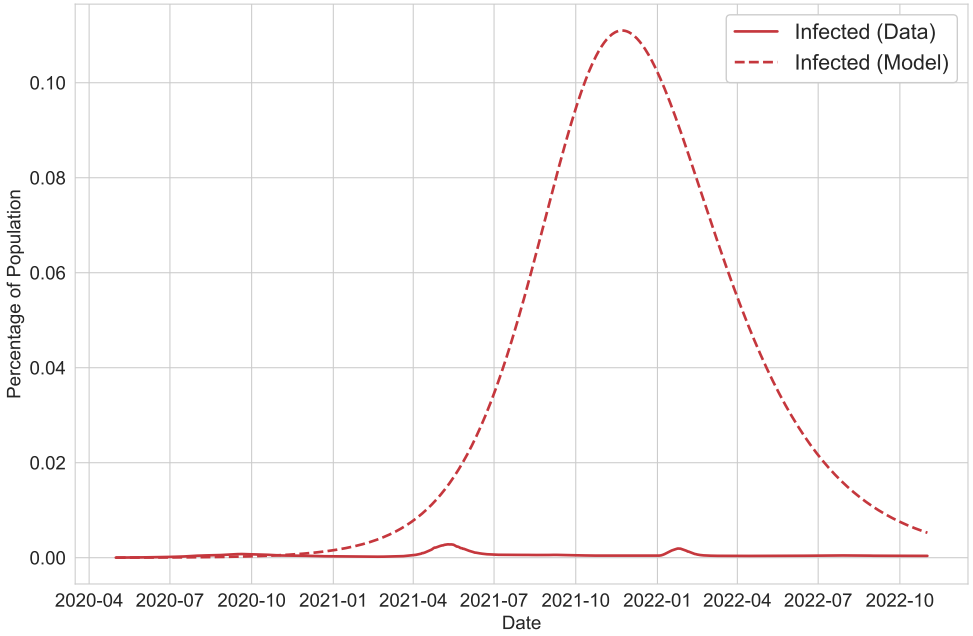
Results

Using the simple SIR model (**??**, **??**) to model the disease dynamics we get:

Figure 2. SIR Model with lockdown for India



(a) SIR model with lockdown

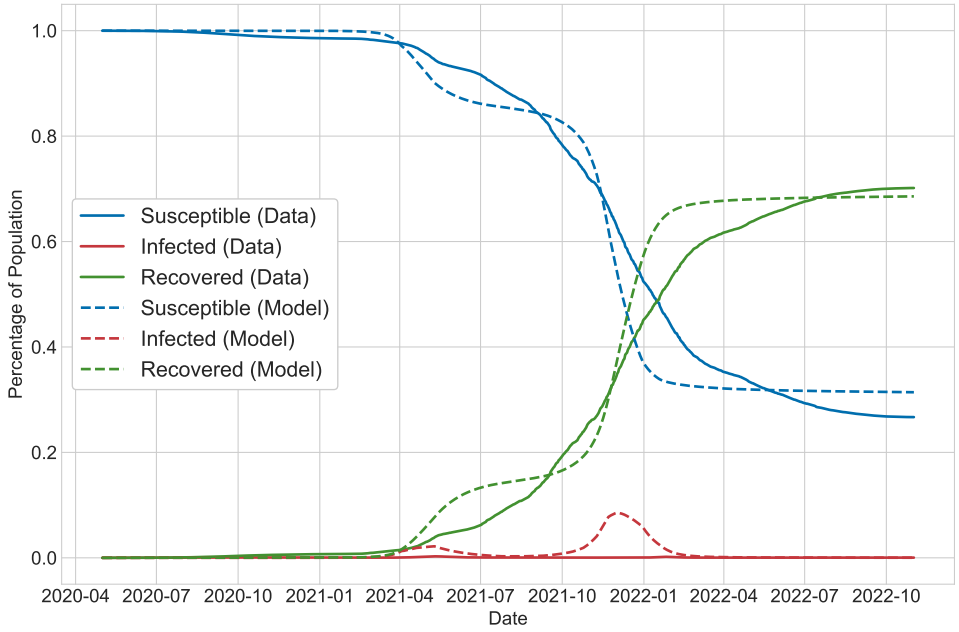


(b) Infections modelled with SIR model

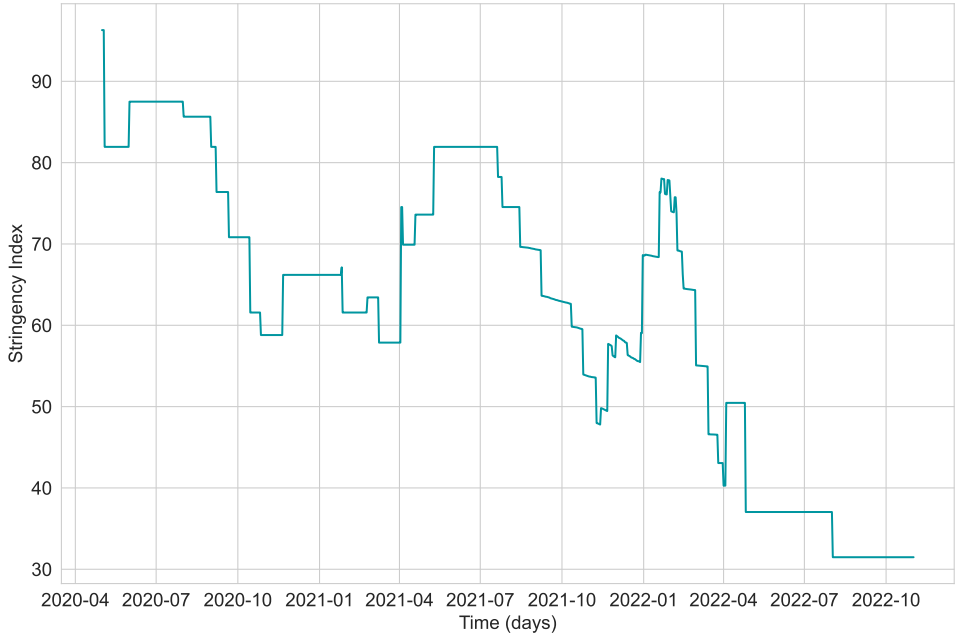
Here, it can be observed that the SIR model accurately fits the susceptible population and recovered population but overestimates the infected population by a significant margin. Although overestimating the infected population may not always be problematic, in our specific case, it can create complications. This is because our research involves rewarding the agent when the proportion of infected individuals falls below a predetermined threshold. Consequently, an overestimation of the infected population could lead to incorrect decision-making and undesirable outcomes.

Combining the lockdown dynamics in the SIR model (**??**, **??**) we get the following:

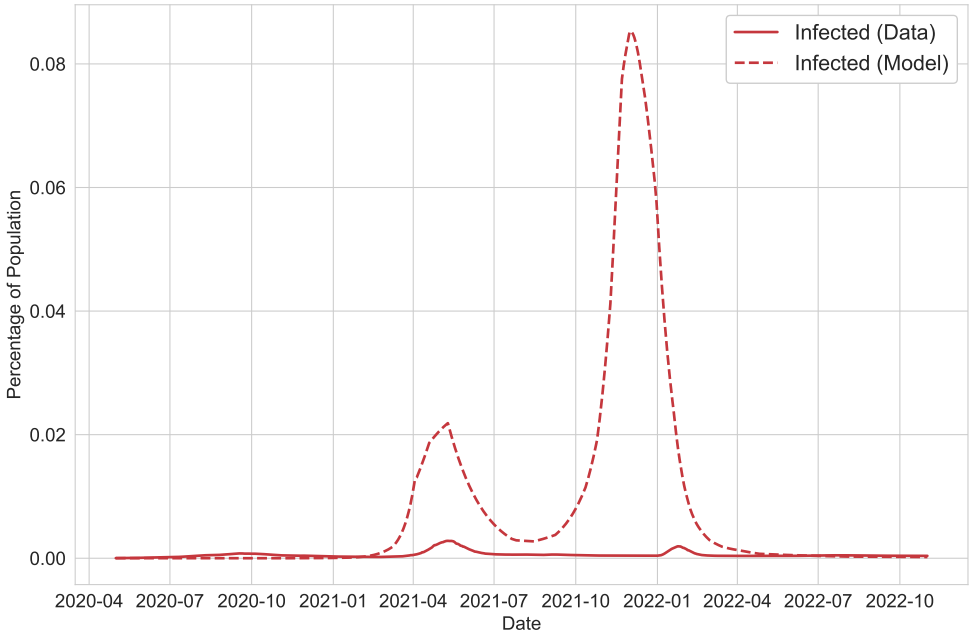
Figure 3. SIR Model with lockdown for India



(a) SIR model with lockdown



(b) Stringency varying with Time

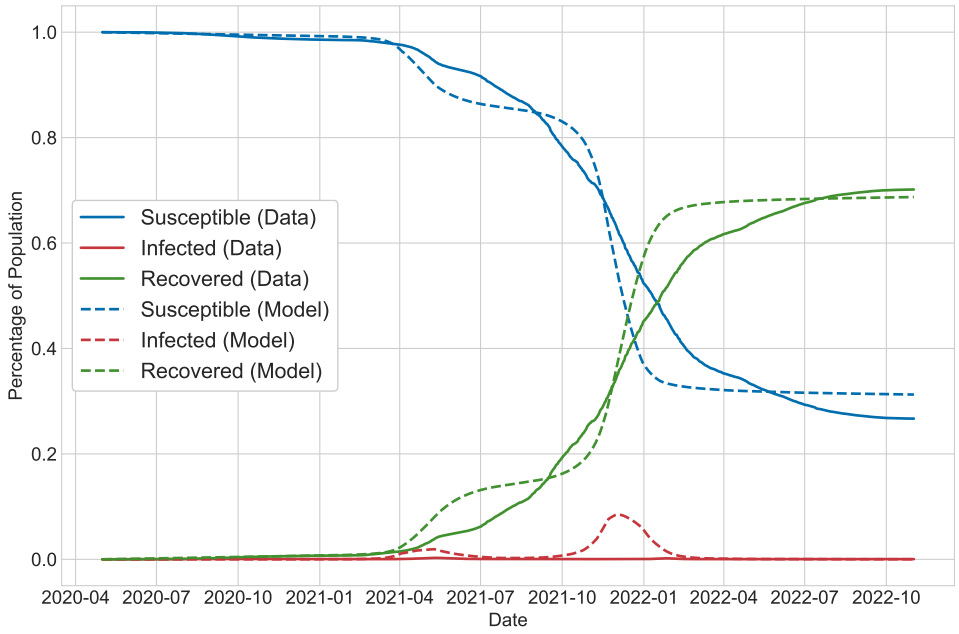


(c) Infections modelled with SIR model with Lockdown

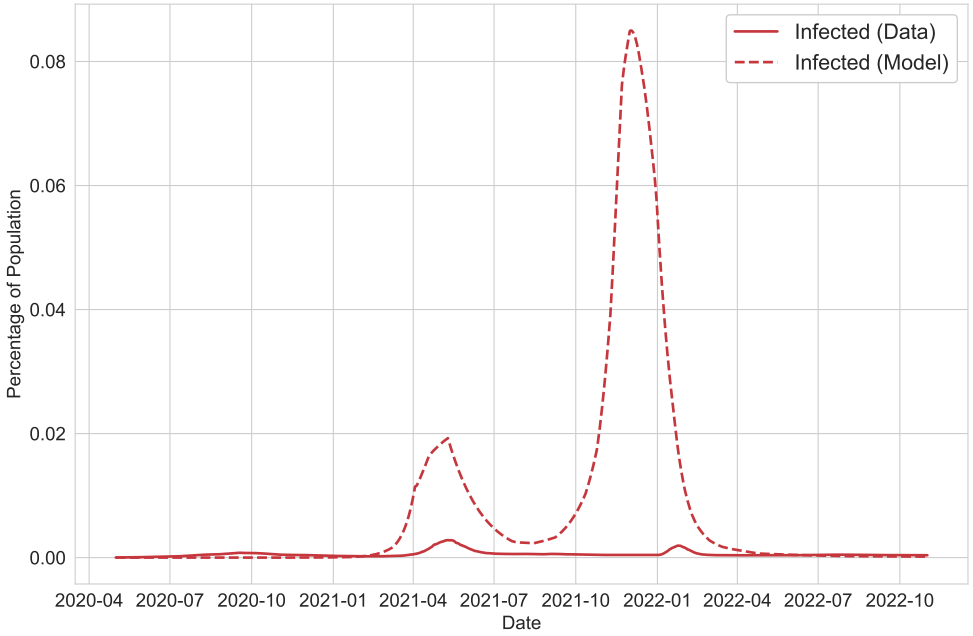
Here, it can be observed there's an overestimation of infected individuals, but, the two stages of the epidemic are being accounted for. This is what suggests that there might be depletion of infected individuals through vaccination.

Combining vaccination dynamics with SIR with lockdown model (**??**, **??**) we get the following:

Figure 4. SIRV Model with lockdown for India



(a) SIRV model with lockdown

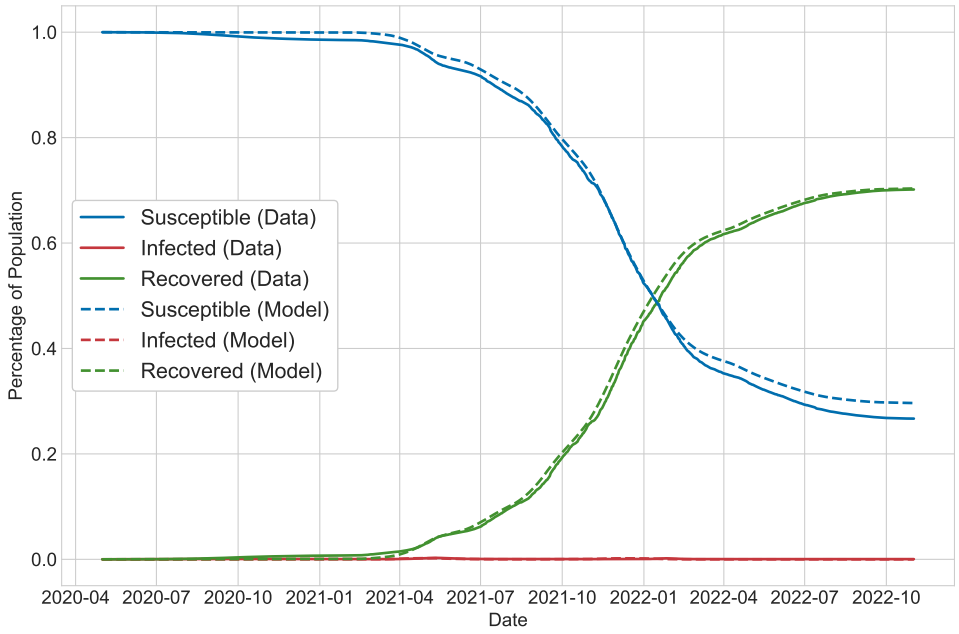


(b) Infections modelled with SIRV model with Lockdown

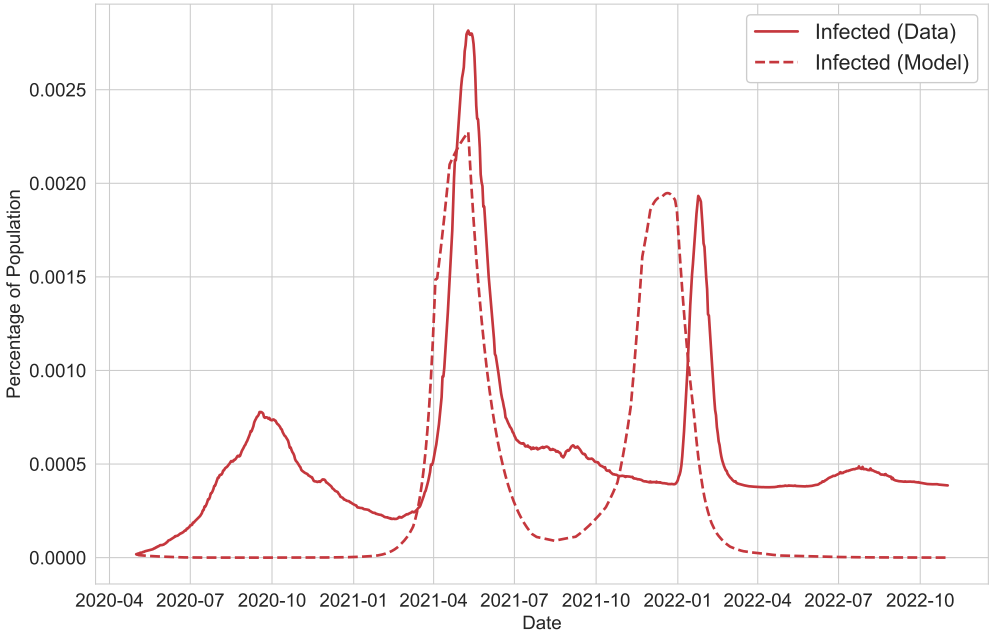
Here, because the value of ν (**??**) is negligible, it makes no change. However, time-varying ν shall be able to better account for the these dynamics.

For time-varying ν and SIRV model with lockdown (**??**, **??**), we get the folllowing:

Figure 5. SIRV Model with lockdown and time-varying ν for India

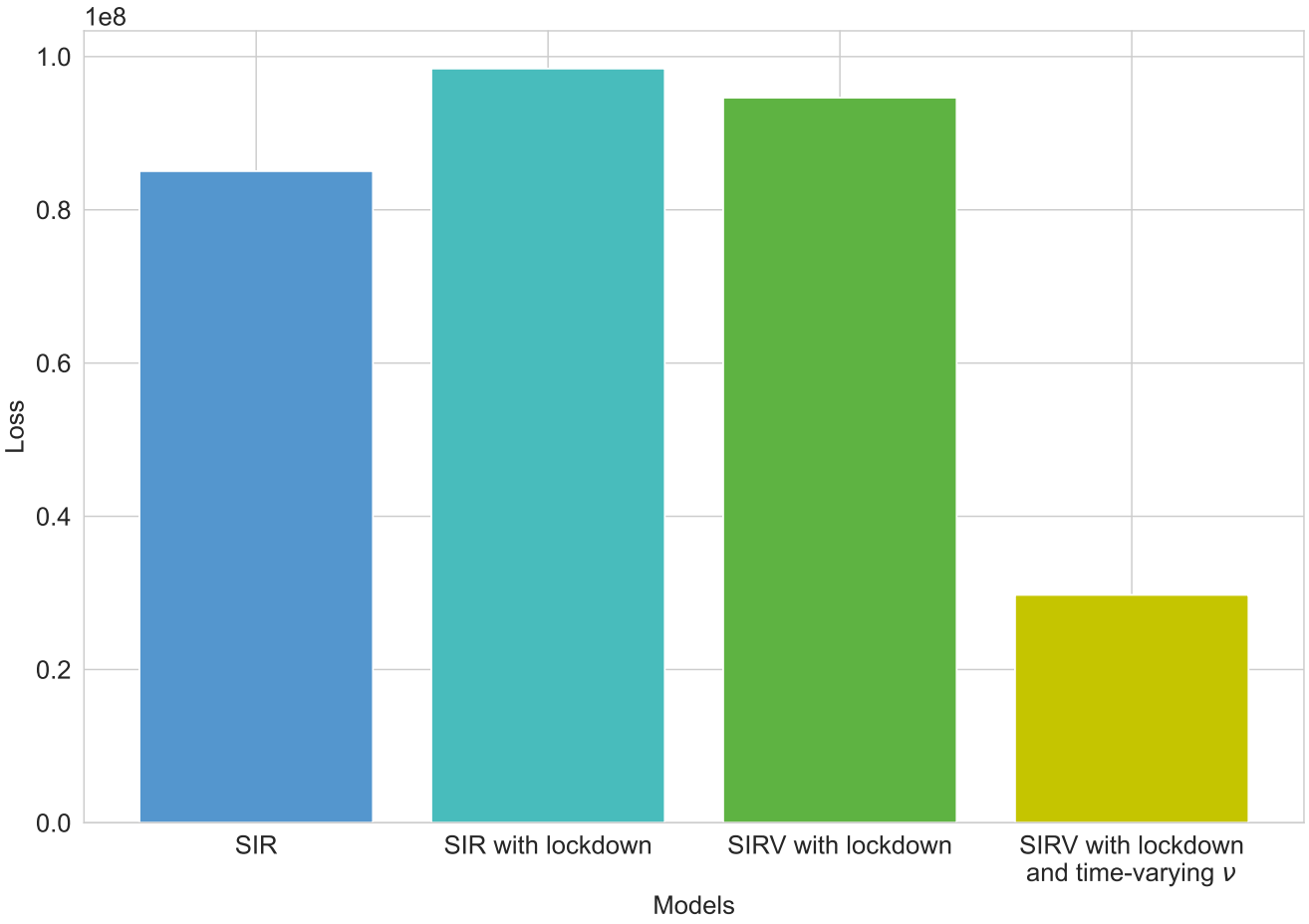


(a) SIRV model with lockdown and time-varying ν



(b) Infections modelled with SIRV model with Lockdown and time-varying ν

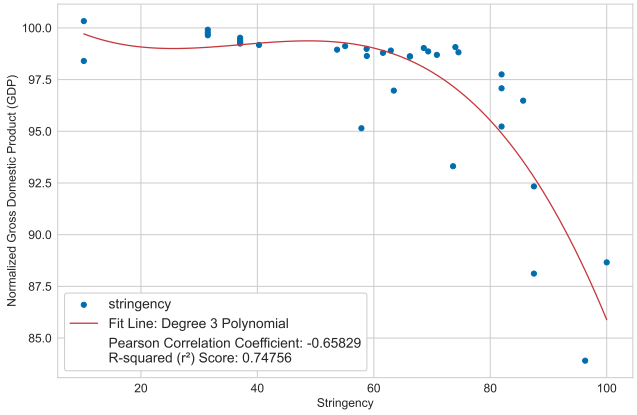
Figure 6. Loss for Different Models



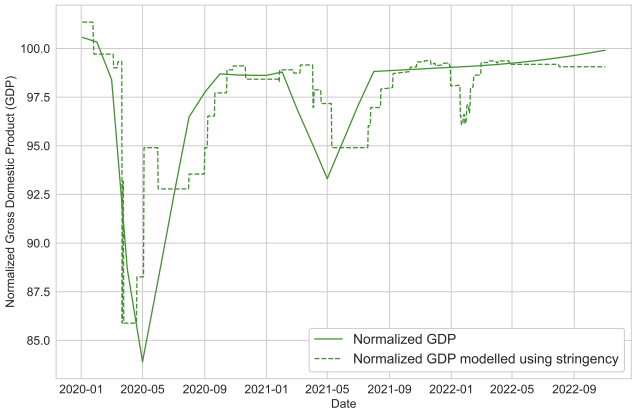
With a time-varying ν (vaccination rate) and the effect of lockdown, our model is able to account for the infected individuals and reduce the cost in comparison to all the previously formalized models for the data. This shows how interventions and changes in the way people behave in response of a epidemic[?] play a major role in the way the epidemic unfolds.

However, Non-pharmaceutical Interventions (NPIs) come with costs for developing nations. Below are scatter plots with pearson correlation, for three countries (India, Mexico, Brazil) which are Emerging Market and Developing Economies[?] from May, 2020 to October, 2022.

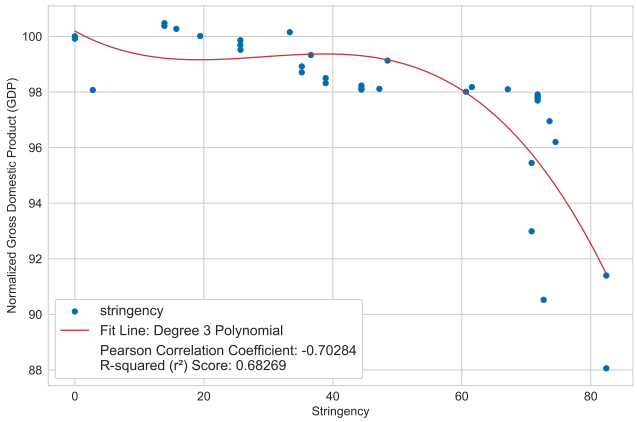
Figure 7. Stringency and GDP for Developing Economies



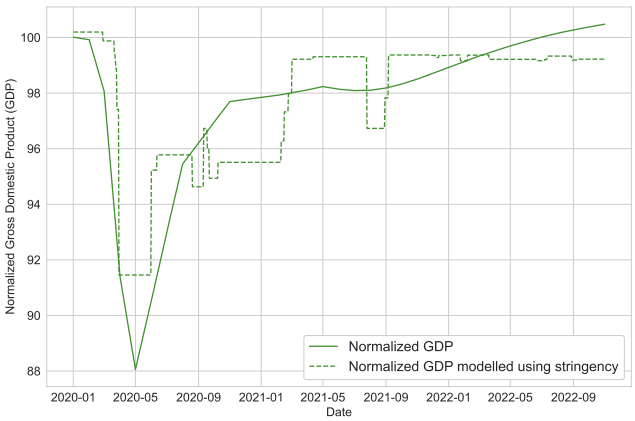
(a) Stringency and Normalized GDP for India



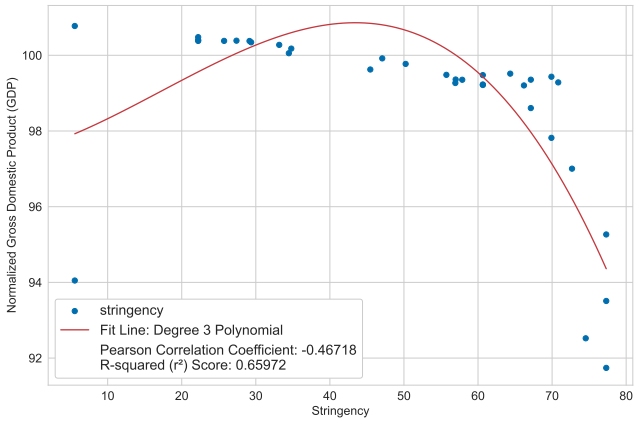
(b) Normalized GDP modelled with Stringency for India



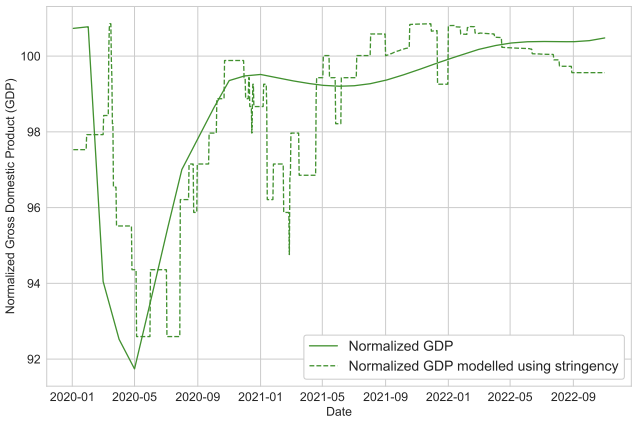
(c) Stringency and Normalized GDP for Mexico



(d) Normalized GDP modelled with Stringency for Mexico



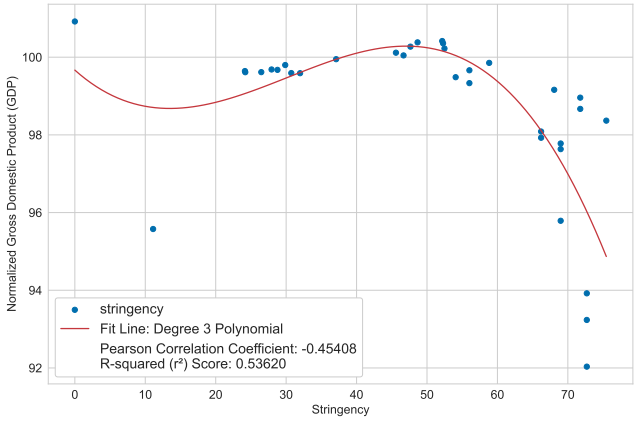
(e) Stringency and Normalized GDP for Brazil



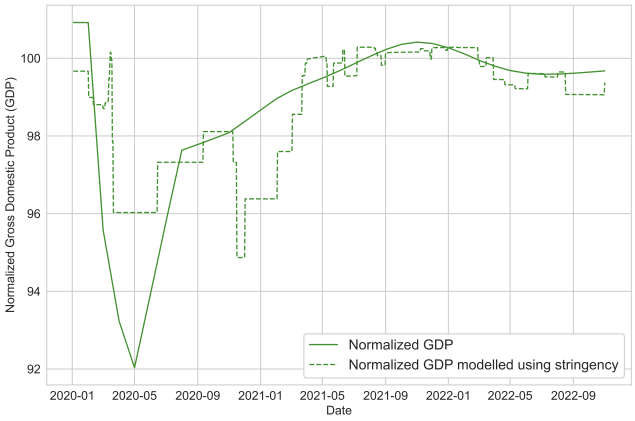
(f) Normalized GDP modelled with Stringency for Brazil

It can be observed that stringency has a negative impact on the normalized Gross domestic product (GDP). Therefore, in some countries policies made during an epidemic have competing costs. This is not the case for Advanced Economies like (USA, Japan, Canada).

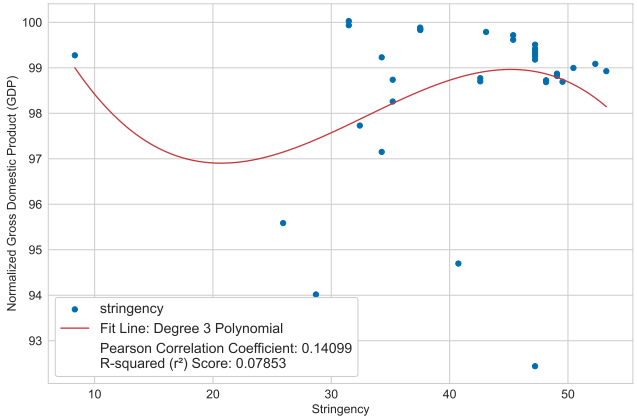
Figure 8. Stringency and GDP for Advanced Economies



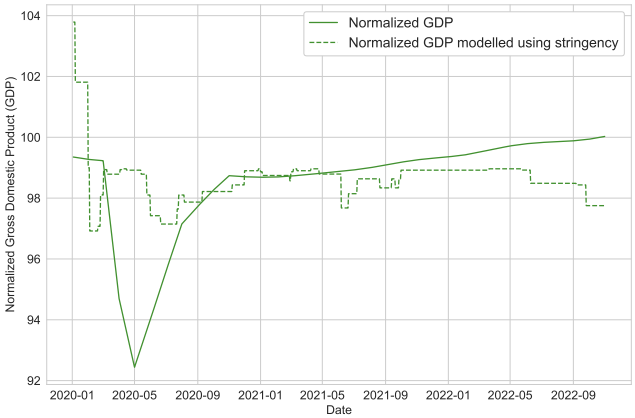
(a) Stringency and Normalized GDP for United States



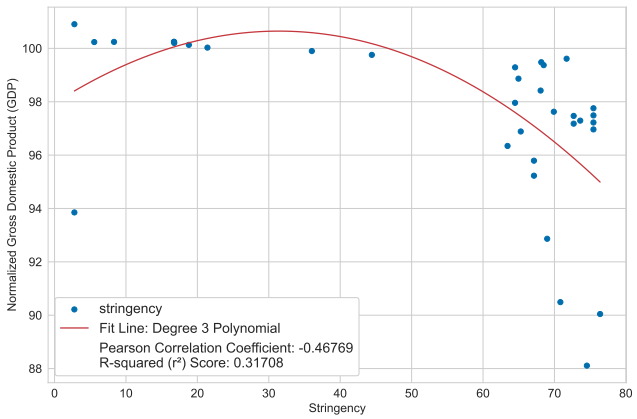
(b) Normalized GDP modelled with Stringency for United States



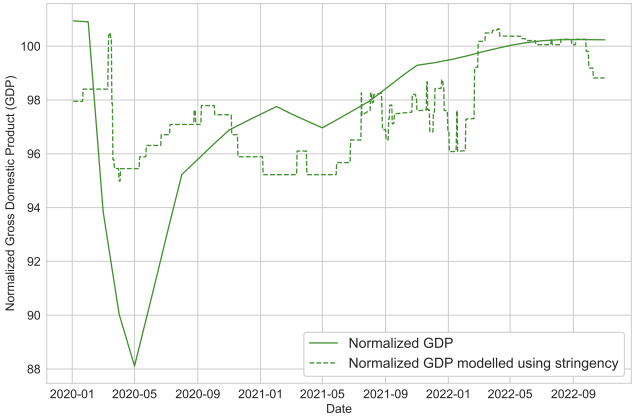
(c) Stringency and Normalized GDP for Japan



(d) Normalized GDP modelled with Stringency for Japan



(e) Stringency and Normalized GDP for Canada



(f) Normalized GDP modelled with Stringency for Canada

From this we conclude, that maybe there are other factors that explain the dip in the normalized GDP than just an increase in the stringency.

Discussion

The paper seeks to inspire epidemiologists by highlighting the advancements achieved through the application of reinforcement learning in policy making during the pandemic. We introduce a virtual environment that closely simulates a pandemic scenario and thoroughly explore innovative strategies for disease mitigation using reinforcement learning. Our proposed approach demonstrates compelling efficacy in achieving optimal decision-making, effectively balancing the formidable challenges posed by the pandemic and economic considerations. We are confident that this research contribution will forge a connection between epidemic studies and reinforcement learning, offering valuable insights to help humanity better defend against the ongoing pandemic crisis.

Experiment Settings

0.1 Dataset

The population-level epidemiological data can be obtained from the 'Our World In Data COVID-19' dataset: https://ourworldindata.org/coronavirus or more specifically: https://github.com/owid/covid-19-data/blob/master/public/data/owid-covid-data.csv[?]. Quaterly GDP data can be obtained from the 'Organisation for Economic Co-operation and Development': https://www.oecd-ilibrary.org/economics/data/main-economic-indicators/main-economic-indicators-complete-database_data-00052-en[?].

0.2 Code

We used stable-baseline3[?], Pytorch[?], Scipy[?], Pandas, Matplotlib, Python[?]. Code: https://github.com/psymbio/sir_rl

References

1. Baker, R. E. *et al.* Infectious disease in an era of global change. *Nat. Rev. Microbiol.* **20**, 193–205, DOI: 10.1038/s41579-021-00639-z (2022).

2. Tan, M. K. I. COVID-19 in an inequitable world: the last, the lost and the least. *Int. Heal.* **13**, 493–496, DOI: 10.1093/inthealth/ihab057 (2021). https://academic.oup.com/inthealth/article-pdf/13/6/493/41430650/ihab057.pdf.

3. Who coronavirus (covid-19) dashboard. https://covid19.who.int/. Accessed: 2024-01-12.

4. World economic outlook, april 2020: The great lockdown. Accessed: 2024-01-12.

5. Nicola, M. *et al.* The socio-economic implications of the coronavirus pandemic (covid-19): A review. *Int. J. Surg.* **78**, 185–193, DOI: 10.1016/j.ijsu.2020.04.018 (2020).

6. Gagnon, J. E., Kamin, S. B. & Kearns, J. The impact of the covid-19 pandemic on global gdp growth. *J. Jpn. Int. Econ.* **68**, 101258, DOI: 10.1016/j.jjie.2023.101258 (2023).

7. Anderson, R. M., Heesterbeek, H., Klinkenberg, D. & Hollingsworth, T. D. How will country-based mitigation measures influence the course of the covid-19 epidemic? *The Lancet* **395**, 931–934, DOI: 10.1016/s0140-6736(20)30567-5 (2020).

8. Song, S., Liu, X., Li, Y. & Yu, Y. Pandemic policy assessment by artificial intelligence. *Sci. Reports* **12**, 13843, DOI: 10.1038/s41598-022-17892-8 (2022).

9. Chinazzi, M. *et al.* The effect of travel restrictions on the spread of the 2019 novel coronavirus (covid-19) outbreak. *Science* **368**, 395–400, DOI: 10.1126/science.aba9757 (2020).

10. Nguyen, T. *et al.* Covid-19 vaccine strategies for aotearoa new zealand: a mathematical modelling study. *The Lancet Regional Heal. - West. Pac.* **15**, 100256, DOI: 10.1016/j.lanwpc.2021.100256 (2021).

11. Kim, D., Keskinocak, P., Pekgün, P. & Yildirim, The balancing role of distribution speed against varying efficacy levels of covid-19 vaccines under variants. *Sci. Reports* **12**, DOI: 10.1038/s41598-022-11060-8 (2022).

12. Jalloh, M. F. *et al.* Drivers of covid-19 policy stringency in 175 countries and territories: Covid-19 cases and deaths, gross domestic products per capita, and health expenditures. *J. Global Heal.* **12**, DOI: 10.7189/jogh.12.05049 (2022).

13. Caldwell, J. M. *et al.* Understanding covid-19 dynamics and the effects of interventions in the philippines: A mathematical modelling study. *The Lancet Regional Heal. - West. Pac.* **14**, 100211, DOI: 10.1016/j.lanwpc.2021.100211 (2021).

14. Ferguson, N. *et al.* Report 9: Impact of non-pharmaceutical interventions (npis) to reduce covid19 mortality and healthcare demand. DOI: 10.25561/77482 (2020).

15. De Foo, C. *et al.* Health financing policies during the covid-19 pandemic and implications for universal health care: a case study of 15 countries. *The Lancet Global Heal.* **11**, e1964–e1977, DOI: 10.1016/s2214-109x(23)00448-5 (2023).

16. Hollingsworth, T. D., Klinkenberg, D., Heesterbeek, H. & Anderson, R. M. Mitigation strategies for pandemic influenza a: Balancing conflicting policy objectives. *PLoS Comput. Biology* **7**, e1001076, DOI: 10.1371/journal.pcbi.1001076 (2011).

17. Pangallo, M. *et al.* The unequal effects of the health–economy trade-off during the covid-19 pandemic. *Nat. Hum. Behav.* DOI: 10.1038/s41562-023-01747-x (2023).

18. Ash, T., Bento, A. M., Kaffine, D., Rao, A. & Bento, A. I. Disease-economy trade-offs under alternative epidemic control strategies. *Nat. Commun.* **13**, DOI: 10.1038/s41467-022-30642-8 (2022).

19. Ohi, A. Q., Mridha, M. F., Monowar, M. M. & Hamid, M. A. Exploring optimal control of epidemic spread using reinforcement learning. *Sci. Reports* **10**, DOI: 10.1038/s41598-020-79147-8 (2020).

20. Padmanabhan, R., Meskin, N., Khattab, T., Shraim, M. & Al-Hitmi, M. Reinforcement learning-based decision support system for covid-19. *Biomed. Signal Process. Control.* **68**, 102676, DOI: https://doi.org/10.1016/j.bspc.2021.102676 (2021).

21. Redlin, M. Differences in npi strategies against covid-19. *J. Regul. Econ.* **62**, 1–23, DOI: 10.1007/s11149-022-09452-9 (2022).

22. Liang, L.-L., Kao, C.-T., Ho, H. J. & Wu, C.-Y. Covid-19 case doubling time associated with non-pharmaceutical interventions and vaccination: A global experience. *J. Global Heal.* **11**, DOI: 10.7189/jogh.11.05021 (2021).

23. Patel, M. D. *et al.* The joint impact of covid-19 vaccination and non-pharmaceutical interventions on infections, hospitalizations, and mortality: An agent-based simulation. DOI: 10.1101/2020.12.30.20248888 (2021).

24. Gagnon, J. E. & Rose, A. 23-8 how did korea's fiscal accounts fare during the covid-19 pandemic? Tech. Rep., Peterson Institute for International Economics (2023).

25. Deb, P., Furceri, D., Ostry, J. & Tawk, N. The economic effects of covid-19 containment measures. *IMF Work. Pap.* **20**, DOI: 10.5089/9781513550251.001 (2020).

26. Eichenbaum, M. S., Rebelo, S. & Trabandt, M. The macroeconomics of epidemics. *The Rev. Financial Stud.* **34**, 5149–5187, DOI: 10.1093/rfs/hhab040 (2021).

27. Lim, S. & Sohn, M. How to cope with emerging viral diseases: lessons from south korea's strategy for covid-19, and collateral damage to cardiometabolic health. *The Lancet Regional Heal. - West. Pac.* **30**, 100581, DOI: 10.1016/j.lanwpc.2022.100581 (2023).

28. Coronavirus: South korea seeing a 'stabilising trend'. https://www.bbc.com/news/av/world-asia-51897979. Accessed: 2024-01-12.

29. Hale, T. *et al.* A global panel database of pandemic policies (oxford covid-19 government response tracker). *Nat. Hum. Behav.* **5**, 529–538, DOI: 10.1038/s41562-021-01079-8 (2021).

30. Hethcote, H. W. *Three Basic Epidemiological Models*, 119–144 (Springer Berlin Heidelberg, 1989).

31. Hethcote, H. W. *The Basic Epidemiology Models: Models, Expressions for R0, Parameter Estimation, and Applications*, 1–61 (WORLD SCIENTIFIC, 2008).

32. Allen, L. J. A primer on stochastic epidemic models: Formulation, numerical simulation, and analysis. *Infect. Dis. Model.* **2**, 128–142, DOI: https://doi.org/10.1016/j.idm.2017.03.001 (2017).

33. Cooper, I., Mondal, A. & Antonopoulos, C. G. A sir model assumption for the spread of covid-19 in different communities. *Chaos, Solitons amp; Fractals* **139**, 110057, DOI: 10.1016/j.chaos.2020.110057 (2020).

34. Bjørnstad, O. N., Shea, K., Krzywinski, M. & Altman, N. The seirs model for infectious disease dynamics. *Nat. Methods* **17**, 557–558, DOI: 10.1038/s41592-020-0856-2 (2020).

35. Mwalili, S., Kimathi, M., Ojiambo, V., Gathungu, D. & Mbogo, R. Seir model for covid-19 dynamics incorporating the environment and social distancing. *BMC Res. Notes* **13**, DOI: 10.1186/s13104-020-05192-1 (2020).

36. Marinov, T. T. & Marinova, R. S. Adaptive sir model with vaccination: simultaneous identification of rates and functions illustrated with covid-19. *Sci. Reports* **12**, DOI: 10.1038/s41598-022-20276-7 (2022).

37. Maurício de Carvalho, J. P. S. & Rodrigues, A. A. Sir model with vaccination: Bifurcation analysis. *Qual. Theory Dyn. Syst.* **22**, DOI: 10.1007/s12346-023-00802-2 (2023).

38. Thäter, M., Chudej, K. & Pesch, H. J. Optimal vaccination strategies for an seir model of infectious diseases with logistic growth. *Math. Biosci. Eng.* **15**, 485–505, DOI: 10.3934/mbe.2018022 (2018).

39. Turkyilmazoglu, M. An extended epidemic model with vaccination: Weak-immune sirvi. *Phys. A: Statistical Mech. its Appl.* **598**, 127429, DOI: 10.1016/j.physa.2022.127429 (2022).

40. Yaladanda, N., Mopuri, R., Vavilala, H. P. & Mutheneni, S. R. Modelling the impact of perfect and imperfect vaccination strategy against sars cov-2 by assuming varied vaccine efficacy over india. *Clin. Epidemiology Global Heal.* **15**, 101052, DOI: https://doi.org/10.1016/j.cegh.2022.101052 (2022).

41. Nguyen, Q. D. & Prokopenko, M. A general framework for optimising cost-effectiveness of pandemic response under partial intervention measures. *Sci. Reports* **12**, DOI: 10.1038/s41598-022-23668-x (2022).

42. Bastani, H. *et al.* Efficient and targeted covid-19 border testing via reinforcement learning. *Nature* **599**, 108–113, DOI: 10.1038/s41586-021-04014-z (2021).

43. Francois-Lavet, V., Henderson, P., Islam, R., Bellemare, M. G. & Pineau, J. An introduction to deep reinforcement learning. DOI: 10.48550/ARXIV.1811.12560 (2018).

44. Arulkumaran, K., Deisenroth, M. P., Brundage, M. & Bharath, A. A. Deep reinforcement learning: A brief survey. *IEEE Signal Process. Mag.* **34**, 26–38, DOI: 10.1109/msp.2017.2743240 (2017).

45. Henderson, P. *et al.* Deep reinforcement learning that matters. *Proc. AAAI Conf. on Artif. Intell.* **32**, DOI: 10.1609/aaai.v32i1.11694 (2018).

46. Mnih, V. *et al.* Human-level control through deep reinforcement learning. *Nature* **518**, 529–533, DOI: 10.1038/nature14236 (2015).

47. Dunn, W. N. *Public policy analysis* (Routledge, London, England, 2017), 6 edn.

48. Demir, T. & Miller, H. Policy communities. In *Handbook of Public Policy Analysis*, 137–147 (CRC Press, 2006).

49. HENS, N. *et al.* Seventy-five years of estimating the force of infection from current status data. *Epidemiology Infect.* **138**, 802–812, DOI: 10.1017/S0950268809990781 (2010).

50. Huber, P. J. Robust Estimation of a Location Parameter. *The Annals Math. Statistics* **35**, 73 – 101, DOI: 10.1214/aoms/1177703732 (1964).

51. Gao, F. & Han, L. Implementing the nelder-mead simplex algorithm with adaptive parameters. *Comput. Optim. Appl.* **51**, 259–277, DOI: 10.1007/s10589-010-9329-3 (2010).

52. Lockdowns in sir models. https://benjaminmoll.com/wp-content/uploads/2020/05/SIR_notes.pdf. Accessed: 2023-12-26.

53. Lockdowns in sir models (code). https://benjaminmoll.com/wp-content/uploads/2020/05/SIR_lockdown.m. Accessed: 2023-12-26.

54. Łukasz Rachel. An analytical model of covid-19 lockdowns. (2020). Https://www.lse.ac.uk/CFM/assets/pdf/CFM-Discussion-Papers-2020/CFMDP2020-29-Paper.pdf.

55. Mathieu, E. *et al.* Coronavirus pandemic (covid-19). *Our World Data* (2020). Https://ourworldindata.org/coronavirus.

56. Aws deepracer. https://aws.amazon.com/deepracer/league/. Accessed: 2024-01-12.

57. OECD. Main economic indicators - complete database. DOI: https://doi.org/https://doi.org/10.1787/data-00052-en (2015).

58. Raffin, A. *et al.* Stable-baselines3: Reliable reinforcement learning implementations. *J. Mach. Learn. Res.* **22**, 1–8 (2021).

59. Paszke, A. *et al.* Pytorch: An imperative style, high-performance deep learning library (2019). 1912.01703.

60. Virtanen, P. *et al.* Scipy 1.0: fundamental algorithms for scientific computing in python. *Nat. Methods* **17**, 261–272, DOI: 10.1038/s41592-019-0686-2 (2020).

61. Oliphant, T. E. Python for scientific computing. *Comput. Sci. Eng.* **9**, 10–20, DOI: 10.1109/MCSE.2007.58 (2007).