Batch deep reinforcement learning for policy responses to the COVID-19 pandemic.

Pablo Gonzalez Ginestet[1,c], Erin Gabriel[1, 2], Ziad El-Khatib[3], and Ujjwal Neogi[4]

[1]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Sweden
[2]Section of Biostatistics, Department of Public Health, University of Copenhagen, Denmark
[3]Department of Global Public Health, Karolinska Institutet, Sweden
[4]Department of Laboratory Medicine, Karolinska Institutet, Sweden
[c]Corresponding author: e-mail:pablo.gonzalez.ginestet@ki.se

May 11, 2022

Abstract

Non-pharmaceutical public health measures such as the use of face masks were used to reduce transmission of the coronavirus disease 2019 (COVID-19). The intensity of these measures over time are key to these policies being successful. Reinforcement learning (RL), an area of machine learning, studies sequential decision making processes and it has shown promising results as a framework for finding optimal treatments in healthcare. In this work, we explore the utility of RL as a decision support tool for implementing public health policies for reducing the spread of COVID-19. We illustrate this by applying two state-of-the-art deep RL algorithms to discover optimal face covering policies with the goal of minimising the risk of spreading the virus.

Keywords: Reinforcement learning; COVID-19

Supporting Information for this article is available at https://github.com/pablogonzalezginestet/batchRL_covid19

1 Introduction

The outbreak of COVID-19 and the consequent pandemic has, and continues to, challenge and test governments all over the globe. In addition to the accelerated development of a successful vaccine, political leaders have, in the meantime, been forced to adopt a variety of non-pharmaceutical public health measures in order to mitigate the spread of the virus. These measures, ranging from social distancing advice and the use of face masks, to travel restrictions, school closures and stay-at-home orders, have not only been implemented in varying degrees by different governments, but also at varying times within the course of the pandemic. The management of the pandemic has proved to be a sequential decision-making problem, with government responses developing and adapting to the current severity of the situation.

Reinforcement learning (RL), a specific area of machine learning, provides a mathematical framework to study problems that involve the task of learning to make a sequence of decisions (Sutton and Barto, 1998). RL aims at finding optimal decisions by optimizing the cumulative effect of decisions made over time. Unlike supervised learning, RL algorithm collects information by interacting with the environment through a sequence of actions. At every decision point, the RL algorithm observes relevant information that describes the current situation of the environment called states and it chooses an action. After choosing an action, it receives a new state from the environment and an immediate reward as feedback from the environment. The states can be thought as covariates measured at some point similar to adjustment in a regression problem. The reward is a function that can be positive or negative (punishment). RL has been used successfully in complex decision making tasks such as robotics and Atari games (Mnih et al., 2015). In settings such as healthcare, this *online* interaction is unfeasible. Batch or offline RL algorithms focus on learning the best possible policy from an observational data or a fixed dataset that has been previously collected and without further interaction with the environment (Lange et al., 2012; Levine et al., 2020). Batch RL algorithms receives directly a sample of transitions from the fixed dataset of the form (state, action, reward, next state), instead of the RL algorithm interacting with the environment trying an action, after observing the state, to obtain the reward and next state.

Batch RL has been applied to diverse medical problems, from tailoring therapies in patients with HIV (Ernst et al., 2006), lung cancer (Zhao et al., 2009) and type 1 diabetes (Luckett et al., 2020), to mobile health data (Liao et al., 2021). Specifically, there has been a lot of interest in applying batch RL algorithms for learning optimal treatment for sepsis patients in intensive care using a restrospective dataset MIMIC-III (Prasad et al., 2017; Weng et al., 2017; Komorowski et al., 2018; Raghu et al., 2017b,a; Peng et al., 2019). Prasad et al. (2017) used a policy fitted Q-iteration algorithm to identify when to the wean patients from mechanical ventilation in intensive care units. Weng et al. (2017) applied a policy iteration algorithm to learn a policy to find personalized optimal glycemic targets. Komorowski et al. (2018) used a policy iteration and Raghu et al. (2017b,a); Peng et al. (2019) applied a deep Q-learning approach (Mnih et al., 2015) to find suitable intravenous fluids and vasopressors treatment strategies.

In this work, we explore the utility of batch deep RL algorithms to discover non-pharmaceutical policies responses against the coronavirus disease 2019 (COVID-19) using a retrospective dataset at the national level. We examined two batch deep RL algorithms to discover optimal face covering policy response with the goal of minimising the risk of spreading the virus. To our knowledge, the paper by Kwak et al. (2021) is the only study that has applied a RL approach to learn optimal policies for controlling the COVID-19 pandemic. Similar to Kwak et al. (2021), we applied the algorithm deep Q-learning but we also apply another state-of-the-art off-policy algorithm called discrete batch-constrained (Fujimoto et al., 2019). Furthermore, we evaluated the learned policies using weighted importance sampling.

2 Methods

Assume that we observe a country trajectory composed of the sequence $\tau_i = (s_0, a_0, s_1, \ldots, a_{T_i}, s_{T_i+1})$ for $i = 1, \ldots, N$ countries where s_t is the state variable representing the country covariates, a_t is the action at time t and T_i is the length of the trajectory of country i. Let $R(s_t, a_t, s_{t+1})$ denotes the reward function. The agent, or the algorithm, selects actions with respect to a policy, π, that maps states to actions. Each policy has associated an action-value function $Q^\pi(s, a) = \mathbb{E}[R_t \mid s_t = s, a_t = a]$, which is the expected reward when following the policy after

taking action a in state s, and a state-value function $V^{\pi}(s) = \mathbb{E}[\sum_{t=0}^{T} R(s_{t+j}, a_{t+j}, s_{t+j+1}) \mid s_t = s]$ is the expected return starting from state s, and then following policy π.

The Bellman equation (Bellman, 1957) characterizes the optimal policy π^*:

$$\pi_t^*(s_t, a_{t-1}) = arg\max_{a \in \mathcal{A}} \mathbb{E}[R(s_t, a_t, s_{t+1}) + \gamma V^*(s_{t+1}) \mid s_t = s, a_t = a]$$

where $V^*(s) = \max_{a \in \mathcal{A}} Q^*(s, a)$ is optimal state value function at s_{t+1} and acting optimally accordingly the optimal policy π^* thereafter.

Q-learning approximates the following Bellman optimality equation:

$$Q^* = \mathbb{E}[R(s_t, a_t, s_{t+1}) + \gamma \max_{a \in \mathcal{A}} Q^*(s_{t+1}, a) \mid s_t = s, a_t = a].$$

In an iteratively process, Q-learning improves an approximate estimate of Q^*, denoted by Q_θ, updating θ in each step using the target $R(s_t, a_t, s_{t+1}) + \gamma \max_{a \in \mathcal{A}} Q_\theta(s_{t+1}, a)$. In deep Q-learning, approximate Q-values are obtained using neural networks, e.g. using the Deep Q-Network algorithm (DQN) (Mnih et al., 2015).

We examined two batch deep RL algorithms algorithms: DQN and discrete batch-constrained deep Q-learning (Fujimoto et al., 2019). We applied these two algorithms using raw features, the current measures observed for each country.

Deep Q-Network (DQN)

Deep Q-learning approximates $Q^*(s, a)$, the optimal action-value function, using a deep neural network model. DQN trains the network minimizing a loss function between the output of the network $Q(s, a; \theta)$ and the target $R + \gamma \max_{a \in \mathcal{A}} Q_{\theta'}(s_{t+1}, a)$ over transitions (s_t, a_t, r_t, s_{t+1}) sampled from the fixed dataset:

$$\mathcal{L}(\theta) = l(R + \gamma \max_{a \in \mathcal{A}} Q(s_{t+1}, a; \theta') - Q(s_{t+1}, a_{t+1}; \theta)) \tag{1}$$

where l is the loss function such as Huber loss (Huber, 1964) or mean-squared error. The target network parameters θ' are maintained fixed over multiple updates of θ and after a number of steps θ' is updated to θ.

We implemented DQN approximating Q-values using a three-layer feedforward neural network with ReLU activation functions after the first two layers and for the last one we used the identity activation function. The network was trained using Adam (Kingma and Ba, 2015) and smooth L1 loss function which is closely related to Huber loss.

Discrete batch-constrained deep Q-learning (dBCQ)

The framework batch-constrained RL was proposed by Fujimoto et al. (2019) to avoid extrapolation error in the Q-value. This error is due to the fact that the action selected by the target network is not contained in the dataset and thus the estimate of $Q(s_{t+1}, a_{t+1}; \theta')$ may be bad without data around (s_{t+1}, a_{t+1}). The algorithm dBCQ eliminates actions that are likely to fall outside the support of the dataset. The elimination of unseen actions is done through a sampling procedure based on a generative model of the dataset G_w, a probability distribution that maps states to actions. The generative model is trained using a supervised learning algorithm with a cross-entropy loss function. The action used to evaluate the target network used in Eq (1) is

$$a' = \underset{a' \sim G_w / \max_{a'} G_w > \tau}{\operatorname{argmax}} Q(s_{t+1}, a'; \theta)$$

where τ is a threshold used to eliminate unlikely actions.

We implemented dBCQ adapting the code developed by Fujimoto et al. (2019) at `https://github.com/sfujim/BCQ`. The Q function was approximated as in DQN using a three-layer feedforward neural network. The generative model was implemented using three-layer feedforward neural network too and a cross-entropy loss function. The network was trained using Adam (Kingma and Ba, 2015).

Data and preprocessing

We used daily COVID-19 data from *Our World in Data* (Ritchie et al., 2020) that includes data from Oxford COVID-19 Government Response Tracker and Johns Hopkins University CSSE COVID-19. This data is freely available online. The complete database includes historical data on the pandemic up to the date of publication.

We restricted the data to the period where the non-pharmaceutical interventions were the main actions to limit the spread of the virus. We used all historical data available for each country until 31 of March 2021, when vaccines may have been available, but the roll-out of the vaccine was still very slow in most countries. The earliest date in the final dataset is 23 of January 2020. Furthermore, we only considered countries with complete state variable information and we also removed remote islands. The list of countries included in the analysis as well as their date range can be found in Table 2 in the Appendix. We performed 5-fold cross-validation where in each fold the RL model was run on the training set based on 80% of the countries and its performance is evaluated on held-out test set (20%). The split was carried out at the country-level. All models were implemented in Pytorch and the codes containing details of all procedures is available at `https://github.com/pablogonzalezginestet/batchRL_covid19`.

Actions

We only considered the policy response face coverings as the action in the RL algorithm. It was downloaded from the same source. Face covering has five categories: 0) no policy; 1) recommended; 2) required in some specified shared/public spaces outside home; 3) required in all shared/public spaces outside the home, and 4) required outside the home at all times regardless of location or presence of other people. We used the following labels for the index action space: 0="No pol", 1="Recom", 2="Req some", 3="Req most" and 4="Req all".

States

We extracted from the database ten variables to be used as state variables: new confirmed cases of COVID-19; stringency index; population density; GDP per capita; human development index; life expectancy; diabetes prevalence; cardiovascular death rate; share of the population that is 65 years and older and hospital beds.

The variables new confirmed cases of COVID-19 and stringency index are the only ones recorded daily. The stringency index records the strictness of government policies on nine response indicators such as school closures, workplace closures, and travel bans except face coverings (our action in the RL algorithm). The daily stringency index was used to reflect

the level of policy response of the country other than face covering's policy. The stringency index ranges from a score of 0 to 100 where a higher score represents a higher level of policy response. The variables population density; GDP per capita, human development index and life expectancy were used to represent socio-economic and demographic factors of each country. The variables diabetes prevalence; cardiovascular death rate and share of the population that is 65 years and older as health vulnerability of the population. Lastly, the variable hospital beds, per 1,000 people, represented health system resources of the country. These variables play the role of covariates in the Q-function similar to their role in a regression problem.

Rewards

To construct the reward function we use the variable reproduction rate of COVID-19 which is recorded daily. The reward function penalizes an increase in this variable issuing a reward of -1 as the country's state deteriorates. Otherwise, a positive reward of $+1$ is issued for decreases in this variable. Given the fact that the incubation period for COVID-19 is thought to extend to 14 days, with a median time of 4-5 days from exposure to symptoms onset (Lauer et al., 2020; Li et al., 2020), we consider changes in the variable within 5 and 14 days: $R_t^k = -sign(x_{t+k} - x_t)$ with $k = 5, 14$. In this work we only consider this simple reward function but other reward schemes can be considered.

3 Results

After processing, the resulting dataset contained 140 countries where 40% has less than a year of trajectory, being the shortest trajectory 237 days. The total number of observations is 50760 (see Table 1). Figure 1 shows the dynamics of the four daily variables considered in the analysis. These variables exhibit a similar pattern: a peak in mid-February 2020 followed by a drastic fall and then converged to values between the two peaks. The first peak is explained by China since it was the only country in the dataset at that time. Most of the countries are already in the dataset by mid-March 2020. The pattern in the policy face covering reflects the fact that at the early of the pandemic not many countries favored

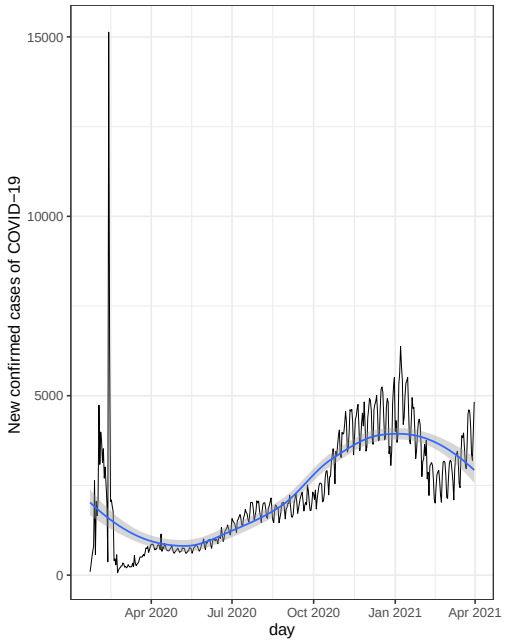
the requirement of mask wearing. Asian countries adopted widespread public mask usage early in the outbreak as opposed to Western countries (Leffler et al., 2020). On the other hand, the stringency index's figure shows that countries mostly preferred other polices such as lockdowns and travel restrictions at the beginning of the pandemic.

Figure 2 displays the reward function using both changes 5 (R_t^5) and 14 (R_t^{14}) days. Both reward functions decrease along time to converge around zero. The reward functions end up oscillating around zero, which means that there are as many rewards as punishments on average, something that is good for the learning process of the RL algorithm. The reward function with changes within 14 days is smoother than the one that considers 5 days.

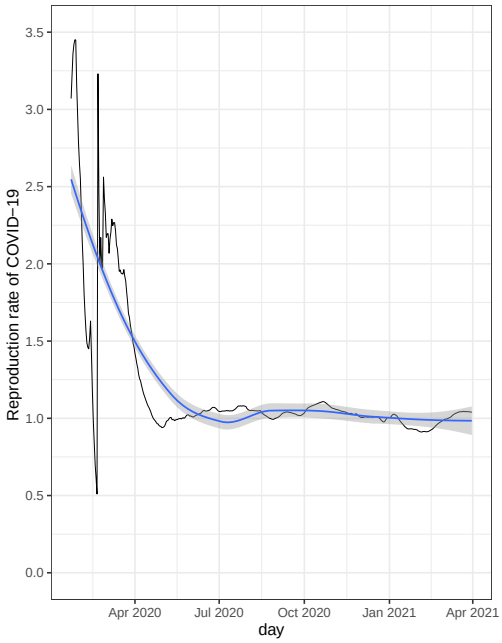
Table 1: Descriptive statistics of the state variable that are not recorded daily in the dataset. All values are expressed in means and interquartile range (Q1;Q3) unless specified.

	Overall
Total number of observations	50760
Total number of countries	140
population density	211.30 (35.61 ; 147.67)
GDP per capita	21797.3 (6222.6 ; 32605.9)
diabetes prevalence	7.649 (5.310 ; 9.590)
cardiovasc death rate	249.11 (151.69 ; 311.11)
aged 65 older	9.737 (4.213 ; 15.322)
human development index	0.7526 (0.6650 ; 0.8800)
life expectancy	74.29 (70.78 ; 78.93)
hospital beds per thousand	3.01 (1.20 ; 4.34)

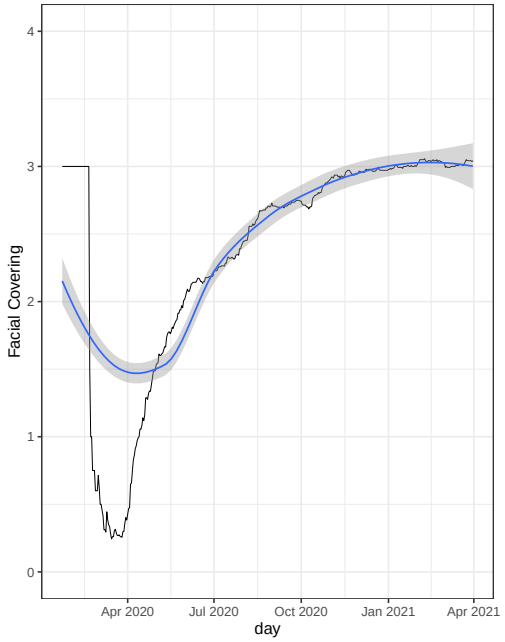
aged 65 older: share of the population that is 65 years and older; **diabetes prevalence**: diabetes prevalence (% of population aged 20 to 79) in 2017; **human development index**: a composite index measuring average achievement in a long and healthy life, knowledge and a decent standard of living; **cardiovasc death rate**: death rate from cardiovascular disease in 2017 (annual number of deaths per 100,000 people); **hospital beds per thousand**: hospital beds per 1,000 people; **gdp per capita**: gross domestic product at purchasing power parity; **population density**: number of people divided by land area, measured in square kilometers



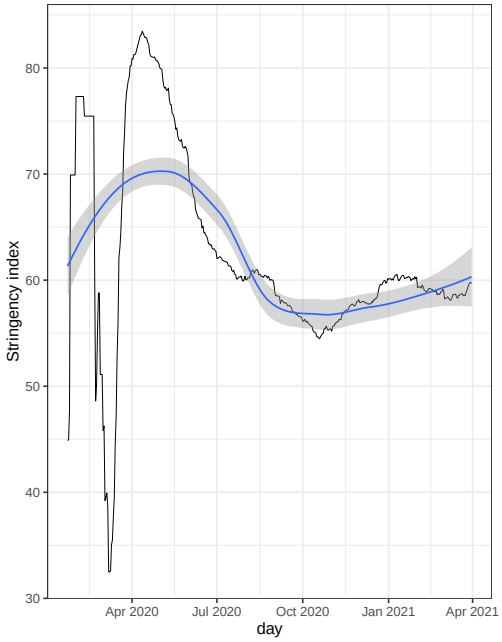
The peak is on 13 February 2020 in China.



The peak is on 27 and 28 of January 2020 in China.



The face covering policy index is: 0: "No pol", 1: "Recom", 2: "Req some", 3: "Req most" and 4: "Req all".



Stringency index goes from 0 to 100 and it does not take account for the policy face covering.

Figure 1: Average across countries of the daily recorded variables: new cases (top left), reproduction rate (top right), face covering policy (bottom left) and stringency index (bottom right panel).

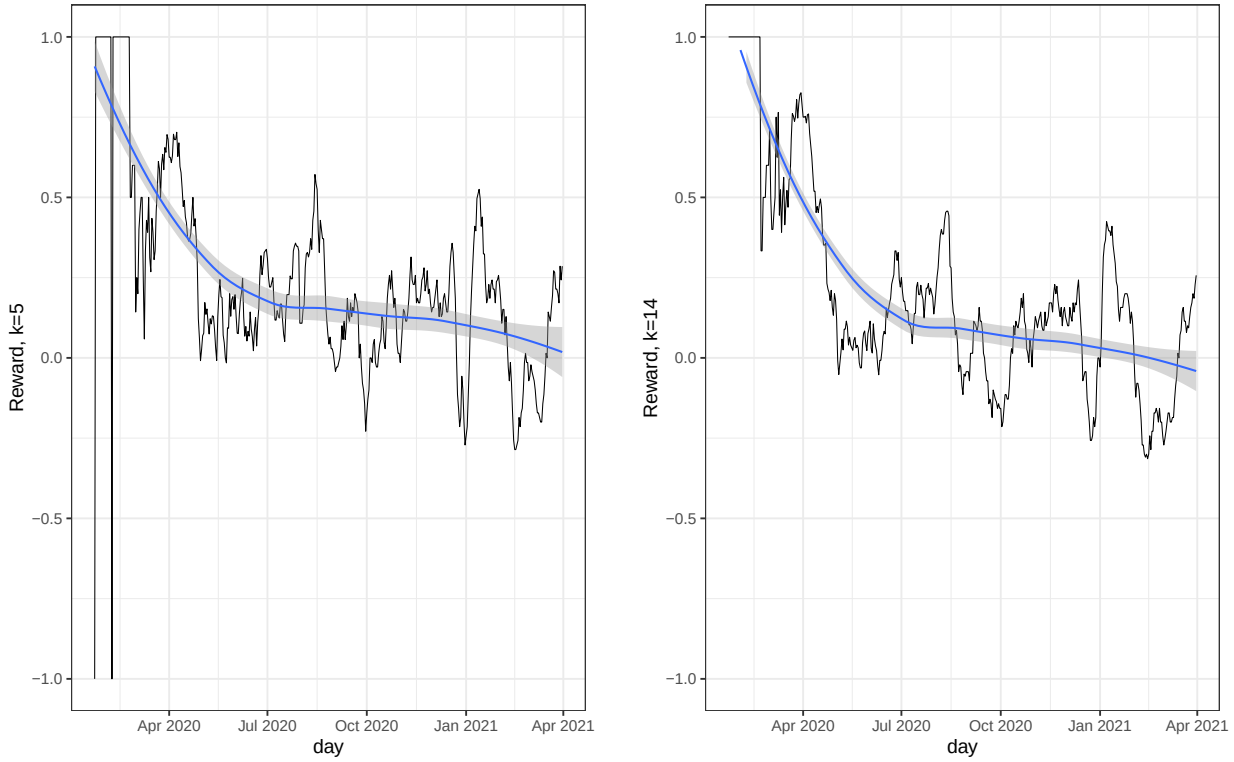


Figure 2: Visualization of the dynamic of the reward functions.

Analysis of learned policies

Figure 3 compares the frequency distributions of actions chosen by governments and the actions suggested by the RL algorithms (DQN and dBCQ) using both reward functions (R_t^5 and R_t^{14}) over the 50760 decision time instances. It shows that the RL algorithms tend to recommend less strict actions compared to the government. DQN and dBCQ performed similarly though dBCQ suggested slightly stricter measures. The only major difference in performance in terms of using the rewards resulted in DQN in the strictest measure "Req all". The RL algorithms more frequently chose the policies "No pol" or "Recom" compared to government. Governments chose much more often to require the use of face masks in all shared/public spaces outside the home with other people present, then would be suggested by either of RL algorithms would suggest. Governments and RL algorithms recommended the action of face coverings that is required in some specified shared/public spaces outside

the home at a similar frequency. However, even when the RL algorithms and governments agree on the frequency of policy use it is not clear if the RL algorithms and the governments recommended these policies for the same time periods. To investigate this we look at a few countries by continent: Australia and New Zealand in Oceania (Figure 4); Argentina, Brazil, Canada and Mexico in America (Fig 5 and Figure 6); China and Japan in Asia (Figure 7); Italy, Spain and Sweden in Europe (Figure 8 and Figure 9); Israel in Middle East (Figure 9), and Niger and South Africa in Africa (Figure 10). The RL recommendations were performed using the trained model with the countries in the training set where the target country belong to the held-out test set in that fold. In each figure, the panel to the left shows the suggested actions by the RL algorithms (dbcq5 = dBCQ using R_t^5, dbcq14 = dBCQ using R_t^{14}; dqn5 = DQN using R_t^5 and dqn14 = DQN using R_t^{14}) compared to the government (gov) throughout the pandemic course considered in the analysis and the panel to the right shows jointly the dynamics of the reproduction rate and the reward function. This latter figure visualizes the built feedback for the algorithm: when the reproduction rate decreases the RL algorithm receives a reward of 1, otherwise it gets a penalization of -1.

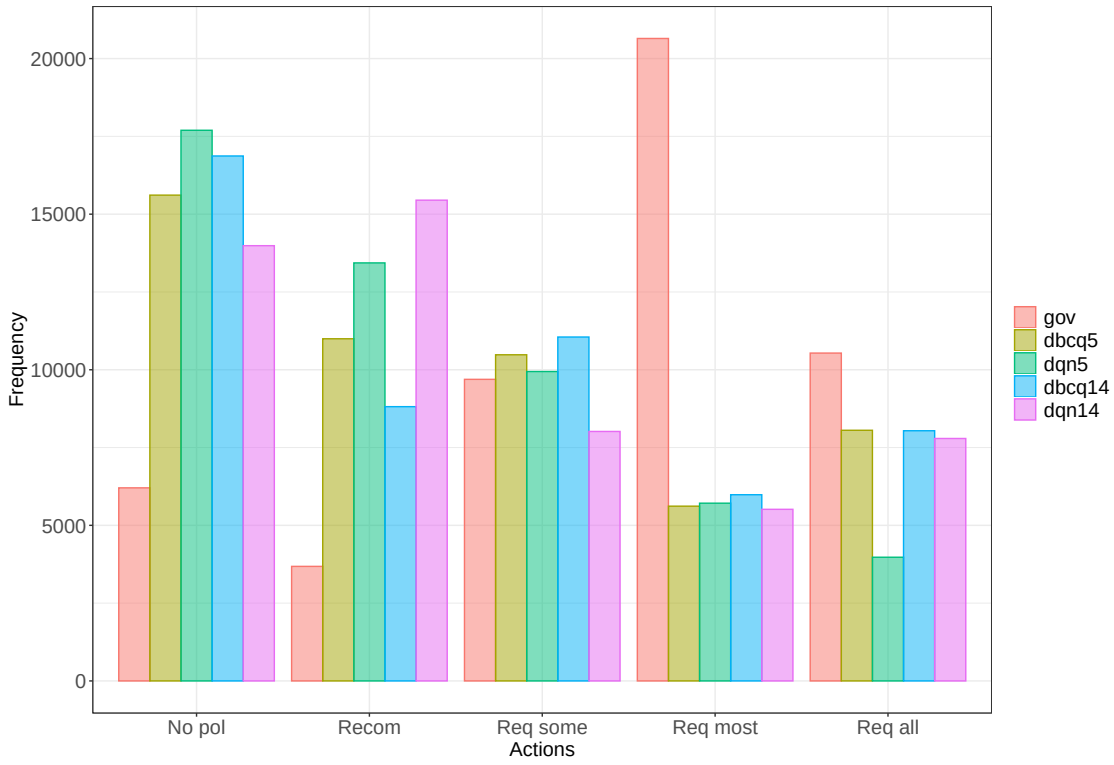


Figure 3: Actions taken by the government and those discovered by the models over the cross-validations. Notation: gov = government; dbcq5 = dBCQ using R_t^5, dbcq14 = dBCQ using R_t^{14}; dqn5 = DQN using R_t^5 and dqn14 = DQN using R_t^{14}.

In Figure 4, Australia and New Zealand gradually implemented stricter face covering policy. Although in Australia the face covering policy reached the strictest level of "Req all", the RL algorithms tended to recommend "No pol". But, in the case of New Zealand the RL algorithm tended to recommend other than "No pol" (stricter measure with R_t^5). DQN using R_t^5 closely matched the government policy in New Zealand. The stricter recommendation seen in New Zealand could be explained by the fact that stringency index was lower than in Australia (Appendix Figure 11).

In Figure 5, Argentina and Brazil's applied a strict face covering policy of "Req most" and "Req all", respectively. The RL algorithms based on R_t^{14} suggested a combination of "No pol" and "Req some" by DQN and "No pol" and "Req all" by dBCQ. Whereas, the RL algorithm suggested other than "No pol" using R_t^5. The algorithm dBCQ recommended less times "No pol" than DQN in the case of Brazil. The lower variability in the suggested

actions recommended by the algorithm in Argentina compared to Brazil could be explained again by the fact that the stringency index in Argentina was very high (above 80) at almost all time (Appendix Figure 12). In the case of Argentina, the RL algorithm DQN using R_t^5 is the closest match to the government and dBCQ using R_t^5 in Brazil.

In Figure 6, Canada implemented a stepwise approach as opposed to Mexico that has applied a fixed policy of "Req all" over the full time period except for a short period at the beginning of the pandemic. The stringency index for both countries were similar (between a range of 60 and 80, Appendix Figure 13). The actions recommended by the RL algorithms followed a similar pattern to the actions observed but different levels: Mexico is characterized by actions kept fixed for a long time and Canada several levels of action for shorter times. dBCQ and DQN using R_t^5 is the closest match to the actions observed in Mexico and Canada, respectively.

In Figure 7, Japan implemented the level "Recom" for face covering in an uniform fashion. The RL algorithms followed this pattern but suggesting the weakest level of "No pol". China applied three levels: "Req some", "Req most" and "Req all" over the period. The RL algorithm dBCQ suggested "Req some" most of the time (with R_t^5), which is the closest to the government's policy, or "Req all" (with R_t^{14}). DQN suggested "Recom" with very brief periods of stricter policies.

In Figure 8, Italy applied the strictest face covering policy of "Req all" most of the time. Instead, the RL algorithms suggested a combination of "Req all" and other less stricter levels. The RL algorithm based on R_t^5 resulted in a similar pattern to the government though a lighter level policy as opposed to variable pattern of the actions recommended by dBCQ with R_t^{14}. On the other hand, Spain gradually increased the level of the policy to reach the strictest level of "Req all". The RL algorithms showed frequent and rapid adjustment among all levels including "No pol". Finally, Sweden implemented the least strict policy of "No pol" all the time except the last months and its stringency index was lower than Italy and Spain (Appendix Figure 15). The RL algorithms based on R_t^{14} suggested a similar policy to what was implemented by the government: "No pol" and "Req some". Though the RL algorithms using R_t^5 did not closely follow the pattern of the observed, they recommended alternating among the weakest levels: "No pol", "Recom" and "Req some".

Israel, Figure 9, implemented a strict policy from the beginning of the pandemic: "Req most" and "Req all". The actions recommended by dBCQ using R_t^5 was the closest to the government despite of being slightly stricter. However, the same algorithm using rewards R_t^{14} recommended most of the times "Recom" combined with some brief period of "No pol" and "Req all". DQN using R_t^5 was closer to the observed actions than using R_t^{14}.

In Figure 10, Niger and South Africa applied mostly an uniform policy of level "Req most" but South Africa was stricter than Niger in the application of other policies as it is shown in Appendix Figure 17. dBCQ algorithm recommended "No pol" for Niger and mostly "Req some" with some brief period of "No pol" for South Africa. Whereas, DQN recommended periods of "No pol" and stricter levels.

Finally, despite the fact that off-policy evaluation methods in the context of discrete actions are not always reliable (Gottesman et al., 2019, 2018), DQN policy had the best performance using weighted importance sampling as it is shown in the Appendix Figure 18.

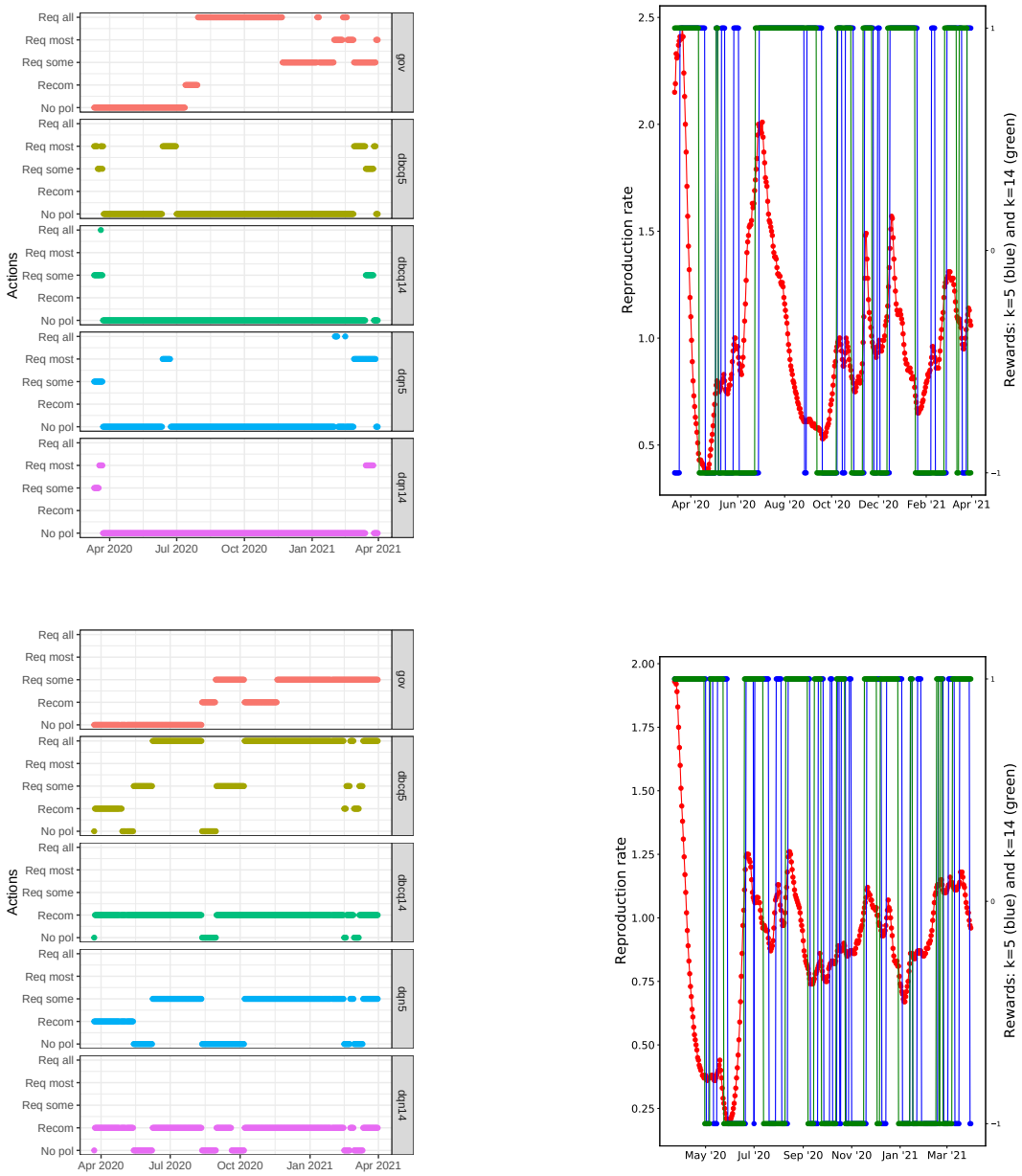


Figure 4: Australia (above) and New Zealand (below). gov = government; dbcq5 = dBCQ using R_t^5, dbcq14 = dBCQ using R_t^{14}; dqn5 = DQN using R_t^5 and dqn14 = DQN using R_t^{14}.

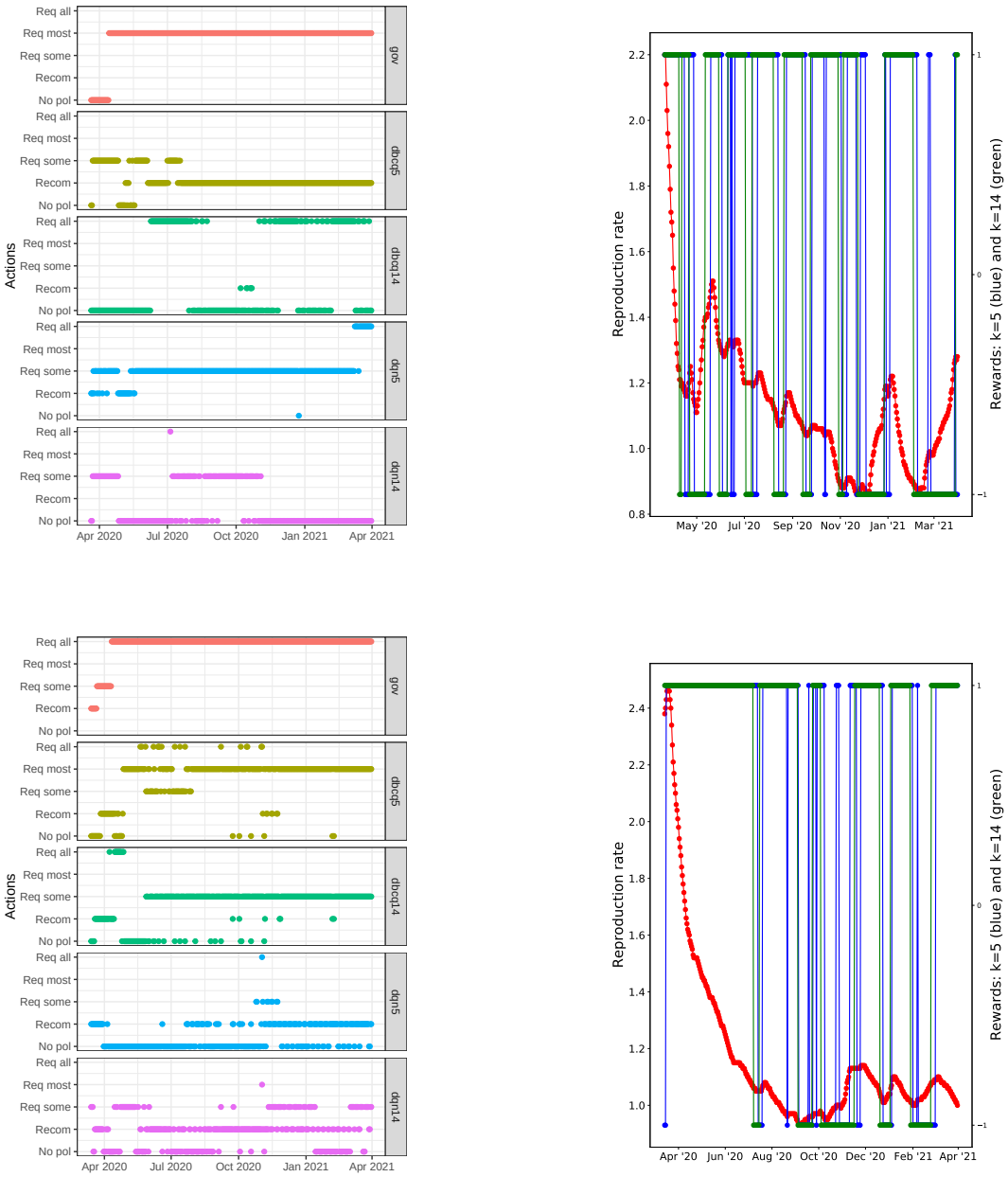


Figure 5: Argentina (above) and Brazil (below). gov = government; dbcq5 = dBCQ using R_t^5, dbcq14 = dBCQ using R_t^{14}; dqn5 = DQN using R_t^5 and dqn14 = DQN using R_t^{14}.

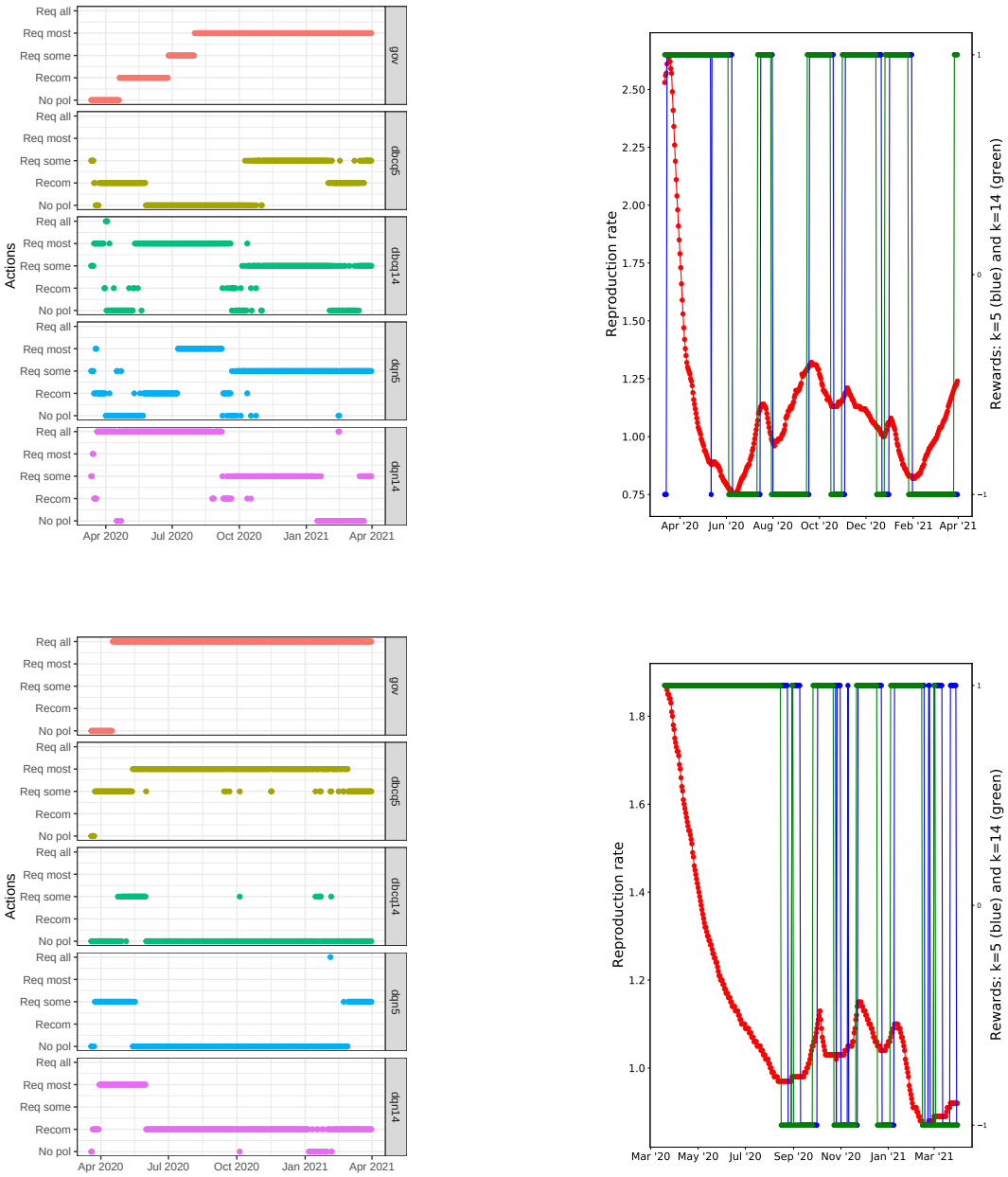


Figure 6: Canada (above) and Mexico (below). gov = government; dbcq5 = dBCQ using R_t^5, dbcq14 = dBCQ using R_t^{14}; dqn5 = DQN using R_t^5 and dqn14 = DQN using R_t^{14}.

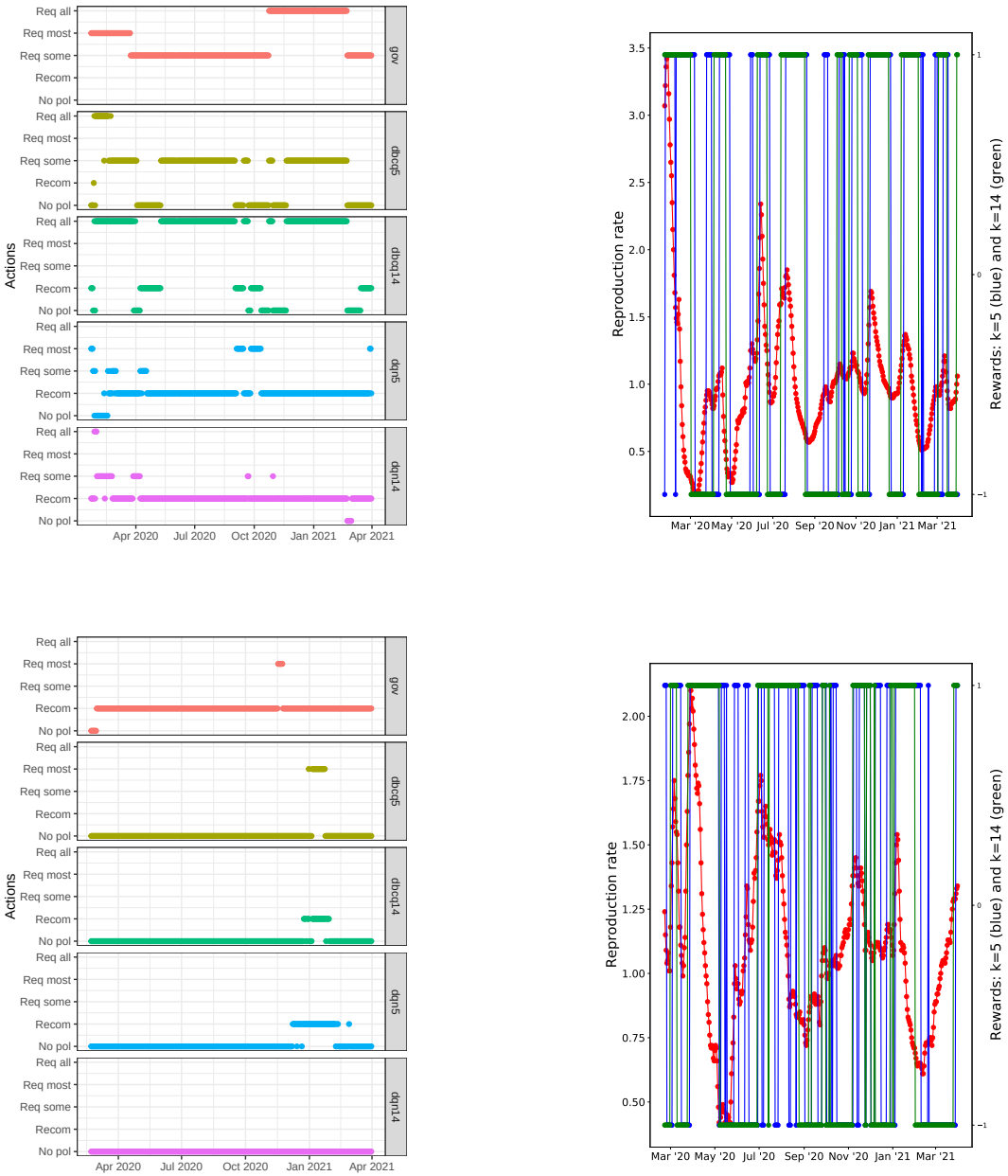


Figure 7: China (above) and Japan (below). gov = government; dbcq5 = dBCQ using R_t^5, dbcq14 = dBCQ using R_t^{14}; dqn5 = DQN using R_t^5 and dqn14 = DQN using R_t^{14}.

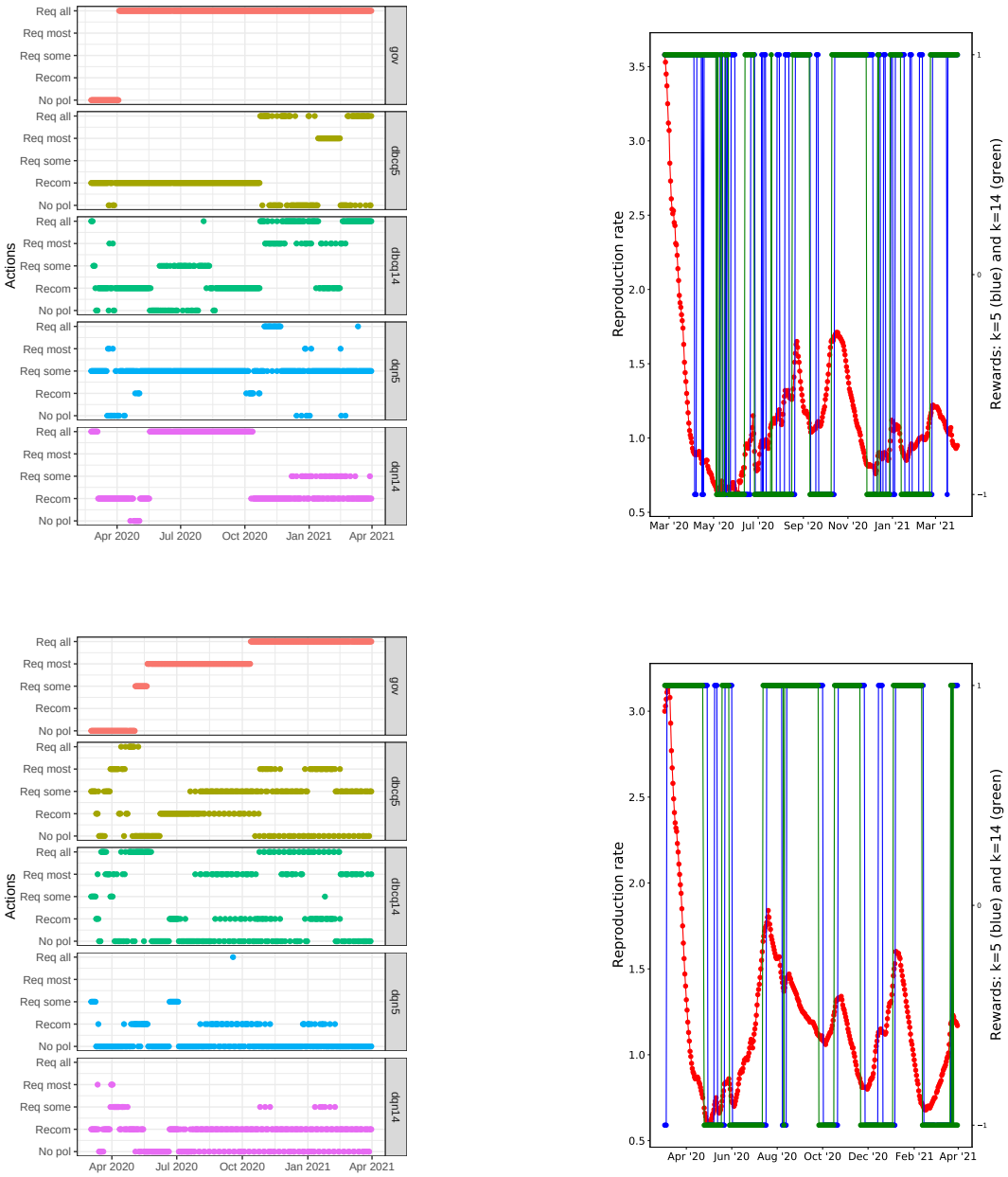


Figure 8: Italy (above) and Spain (below). gov = government; dbcq5 = dBCQ using R_t^5, dbcq14 = dBCQ using R_t^{14}; dqn5 = DQN using R_t^5 and dqn14 = DQN using R_t^{14}.

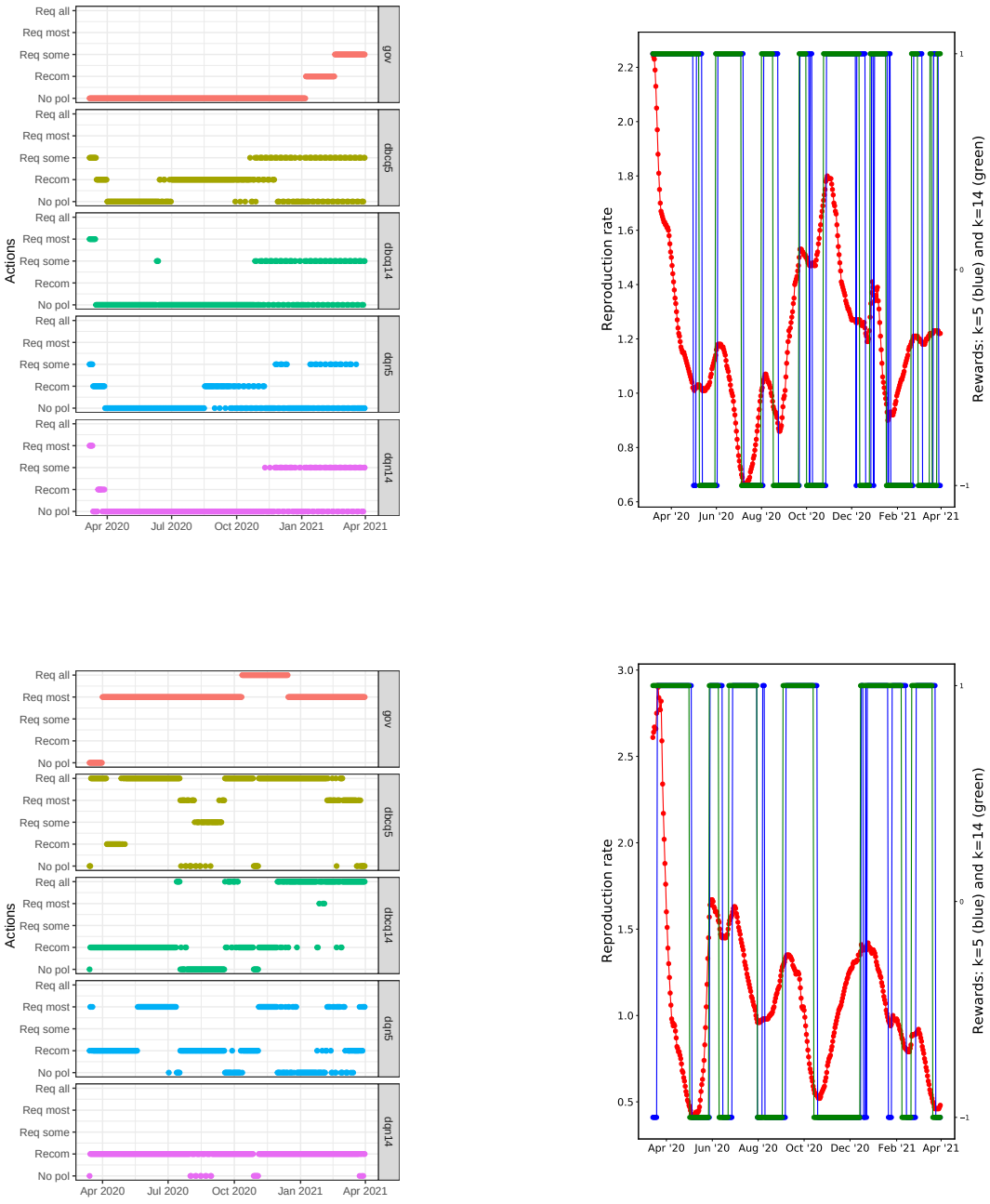


Figure 9: Sweden (above) and Israel (below). gov = government; dbcq5 = dBCQ using R_t^5, dbcq14 = dBCQ using R_t^{14}; dqn5 = DQN using R_t^5 and dqn14 = DQN using R_t^{14}.

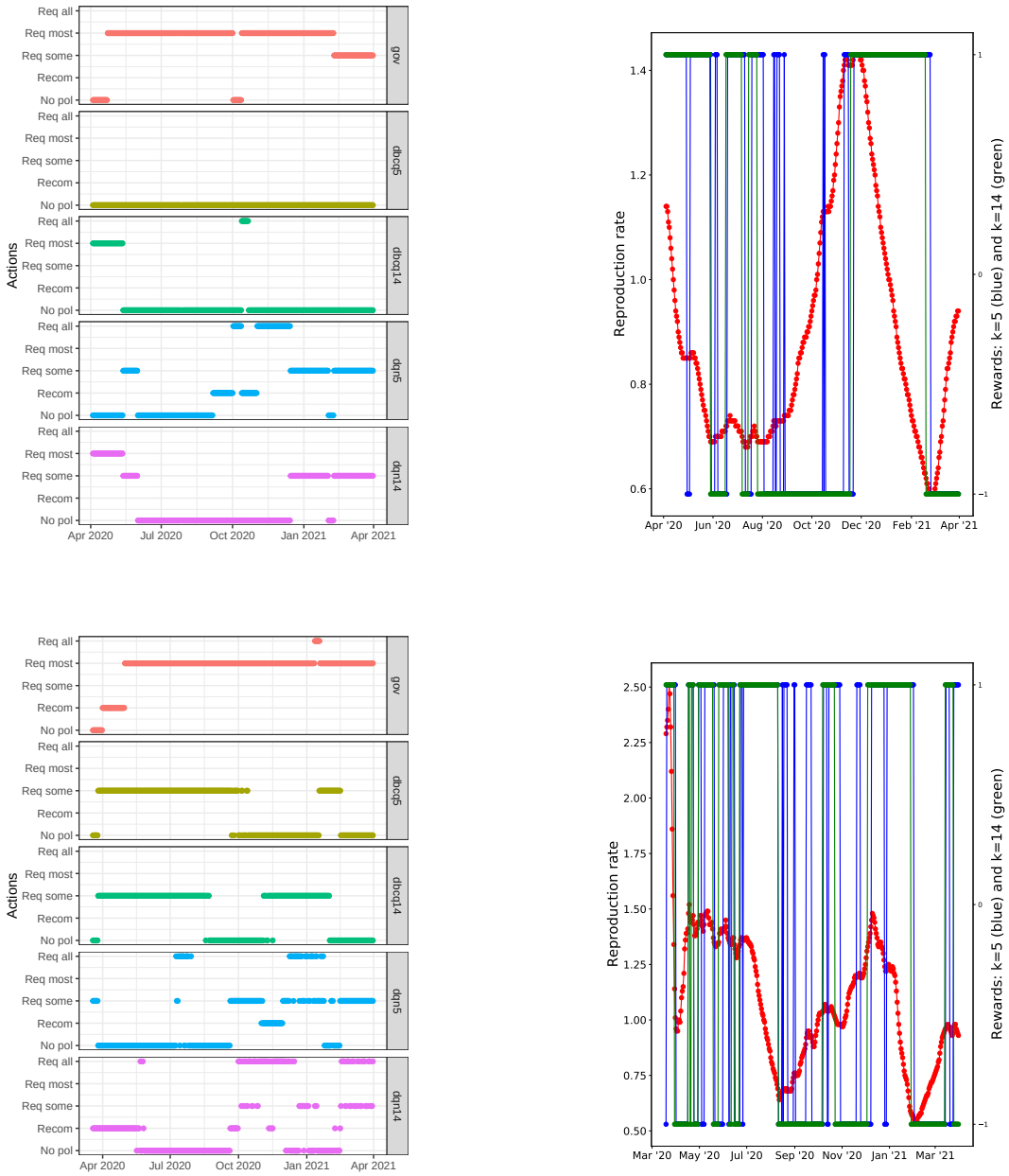


Figure 10: Niger (above) and South Africa (below). gov = government; dbcq5 = dBCQ using R_t^5, dbcq14 = dBCQ using R_t^{14}; dqn5 = DQN using R_t^5 and dqn14 = DQN using R_t^{14}.

Conclusion

In this work, we explored a data-driven approach to discover non-pharmaceutical policy responses against COVID-19. Using publicly available COVID-19 epidemiological data, we examined two state-of-the-art RL algorithms with continuous state-space and discretized action space to find optimal face covering policy response with the goal of minimising the risk of spreading the virus. Though batch RL is in too early a stage of development to deploy the learned policies from these methods in the real world currently, we think further development of batch RL will make it a useful decision support tool for implementing public health policies to the COVID-19 and future pandemics.

With this caveat, our results suggest that in most cases countries followed a similar policy to what was selected by the RL algorithms though that some recommendations shown frequent and rapid adjustment which is unfeasible in the real world. We found examples where the RL algorithms suggested a similar dynamic but a different level of the policy like Argentina or Japan where the algorithm recommended for most of the time slightly less strict than the policy applied by the country. We also found cases where the RL algorithms recommended the same level of the policy implemented by the country for a long, for example Sweden, or short period of time, for example Italy.

This work is limited not only by the current state of the art of the batch RL algorithms, but also by the available data. The ideal setting for RL algorithms is when units, here countries, can be considered the same conditional on state and time and a number of units with the same state take on all possible actions. It is unlikely that countries are interchangeable even at the same time in the same state, and as countries followed similar policy trajectories over time few, if any, took on all actions and not all actions were taken on at all time points. In practice, this limits the action space the RL algorithms can consider.

Appendix

List of countries

Table 2: List of countries included in the analysis sorted chronologically by their starting point in the dataset given by the column first date.

Country	First date	Last date
China	2020-01-23	2021-03-31
South Korea	2020-02-21	2021-03-31
Japan	2020-02-22	2021-03-31
Italy	2020-02-24	2021-03-31
Iran	2020-02-27	2021-03-31
France	2020-03-01	2021-03-31
Singapore	2020-03-01	2021-03-31
Germany	2020-03-02	2021-03-31
Spain	2020-03-03	2021-03-31
United Kingdom	2020-03-03	2021-03-31
United States	2020-03-05	2021-03-31
Switzerland	2020-03-06	2021-03-31
Belgium	2020-03-07	2021-03-31
Netherlands	2020-03-07	2021-03-31
Norway	2020-03-07	2021-03-31
Sweden	2020-03-07	2021-03-31
Austria	2020-03-09	2021-03-31
Malaysia	2020-03-10	2021-03-31
Australia	2020-03-11	2021-03-31
Bahrain	2020-03-11	2021-03-31
Denmark	2020-03-11	2021-03-31
Canada	2020-03-12	2021-03-31
Qatar	2020-03-12	2021-03-31
Iceland	2020-03-13	2021-03-31

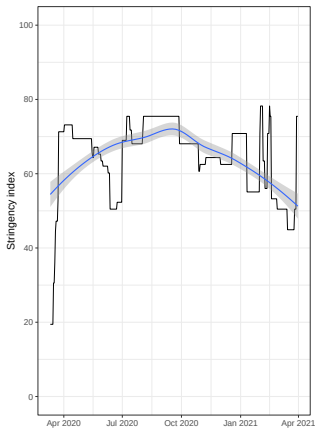
Country	First date	Last date
Brazil	2020-03-14	2021-03-31
Czechia	2020-03-14	2021-03-31
Finland	2020-03-14	2021-03-31
Greece	2020-03-14	2021-03-31
Iraq	2020-03-14	2021-03-31
Israel	2020-03-14	2021-03-31
Portugal	2020-03-14	2021-03-31
Slovenia	2020-03-14	2021-03-31
Egypt	2020-03-15	2021-03-31
Estonia	2020-03-15	2021-03-31
India	2020-03-15	2021-03-31
Ireland	2020-03-15	2021-03-31
Kuwait	2020-03-15	2021-03-31
Philippines	2020-03-15	2021-03-31
Poland	2020-03-15	2021-03-31
Romania	2020-03-15	2021-03-31
Saudi Arabia	2020-03-15	2021-03-31
Chile	2020-03-16	2021-03-31
Indonesia	2020-03-16	2021-03-31
Lebanon	2020-03-16	2021-03-31
Pakistan	2020-03-16	2021-03-31
Thailand	2020-03-16	2021-03-31
Luxembourg	2020-03-18	2021-03-31
Peru	2020-03-18	2021-03-31
Russia	2020-03-18	2021-03-31
Colombia	2020-03-19	2021-03-31
Ecuador	2020-03-19	2021-03-31
Mexico	2020-03-19	2021-03-31
South Africa	2020-03-19	2021-03-31

Country	First date	Last date
United Arab Emirates	2020-03-19	2021-03-31
Croatia	2020-03-20	2021-03-31
Panama	2020-03-20	2021-03-31
Serbia	2020-03-20	2021-03-31
Slovakia	2020-03-20	2021-03-31
Turkey	2020-03-20	2021-03-31
Argentina	2020-03-21	2021-03-31
Bulgaria	2020-03-21	2021-03-31
Latvia	2020-03-21	2021-03-31
Uruguay	2020-03-21	2021-03-31
Algeria	2020-03-22	2021-03-31
Costa Rica	2020-03-22	2021-03-31
Dominican Republic	2020-03-22	2021-03-31
Hungary	2020-03-22	2021-03-31
Bosnia and Herzegovina	2020-03-23	2021-03-31
Jordan	2020-03-23	2021-03-31
Lithuania	2020-03-23	2021-03-31
Morocco	2020-03-23	2021-03-31
New Zealand	2020-03-23	2021-03-31
Vietnam	2020-03-23	2021-03-31
Albania	2020-03-24	2021-03-31
Malta	2020-03-24	2021-03-31
Moldova	2020-03-24	2021-03-31
Brunei	2020-03-25	2021-03-31
Burkina Faso	2020-03-25	2021-03-31
Cyprus	2020-03-25	2021-03-31
Sri Lanka	2020-03-25	2021-03-31
Tunisia	2020-03-25	2021-03-31
Ukraine	2020-03-26	2021-03-31

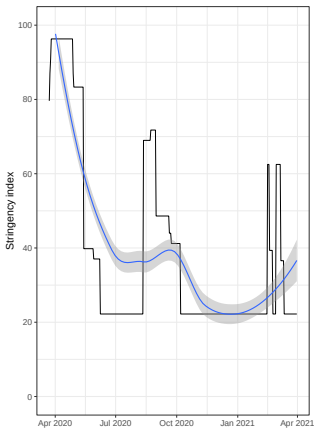
Country	First date	Last date
Azerbaijan	2020-03-27	2021-03-31
Ghana	2020-03-27	2021-03-31
Kazakhstan	2020-03-27	2021-03-31
Oman	2020-03-27	2021-03-31
Venezuela	2020-03-27	2021-03-31
Afghanistan	2020-03-29	2021-03-31
Uzbekistan	2020-03-29	2021-03-31
Cambodia	2020-03-30	2021-03-31
Cameroon	2020-03-30	2021-03-31
Honduras	2020-03-30	2021-03-31
Belarus	2020-03-31	2021-03-31
Georgia	2020-03-31	2021-03-31
Bolivia	2020-04-01	2021-03-31
Kyrgyzstan	2020-04-01	2021-03-31
Kenya	2020-04-03	2021-03-31
Niger	2020-04-04	2021-03-31
Paraguay	2020-04-05	2021-03-31
Trinidad and Tobago	2020-04-05	2021-03-31
Bangladesh	2020-04-07	2021-03-31
Djibouti	2020-04-09	2021-03-31
El Salvador	2020-04-10	2021-03-31
Guatemala	2020-04-11	2021-03-31
Madagascar	2020-04-12	2021-03-31
Mali	2020-04-13	2021-03-31
Jamaica	2020-04-16	2021-03-31
Gabon	2020-04-18	2021-03-31
Tanzania	2020-04-18	2021-03-31
Ethiopia	2020-04-19	2021-03-31
Myanmar	2020-04-20	2021-03-31

Country	First date	Last date
Sudan	2020-04-22	2021-03-31
Liberia	2020-04-23	2021-03-31
Cape Verde	2020-04-27	2021-03-31
Togo	2020-04-30	2021-03-31
Eswatini	2020-05-01	2021-03-31
Zambia	2020-05-01	2021-03-31
Tajikistan	2020-05-05	2021-03-31
Haiti	2020-05-07	2021-03-31
Uganda	2020-05-07	2021-03-31
Benin	2020-05-08	2021-03-31
Nepal	2020-05-08	2021-03-31
Central African Republic	2020-05-09	2021-03-31
Guyana	2020-05-11	2021-03-31
Mozambique	2020-05-12	2021-03-31
Yemen	2020-05-16	2021-03-31
Mongolia	2020-05-17	2021-03-31
Nicaragua	2020-05-20	2021-03-31
Bahamas	2020-05-24	2021-03-31
Malawi	2020-05-26	2021-03-31
Zimbabwe	2020-05-28	2021-03-31
Libya	2020-05-29	2021-03-31
Suriname	2020-06-07	2021-03-31
Eritrea	2020-06-16	2021-03-31
Burundi	2020-06-17	2021-03-31
Botswana	2020-06-30	2021-03-31
Barbados	2020-07-12	2021-03-31
Gambia	2020-07-21	2021-03-31
Bhutan	2020-07-31	2021-03-31
Belize	2020-08-07	2021-03-31

Stringency index

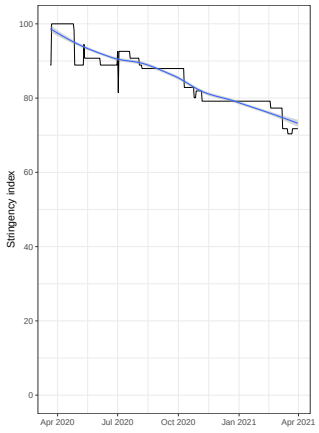


(a) Australia

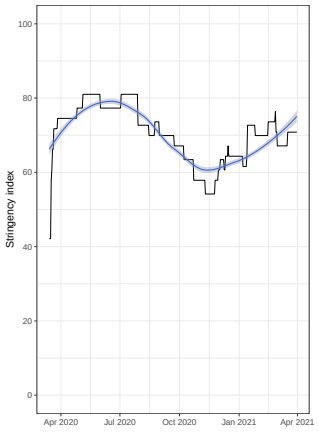


(b) New Zealand

Figure 11: Dynamic of the stringency index in two representative countries from Oceania.

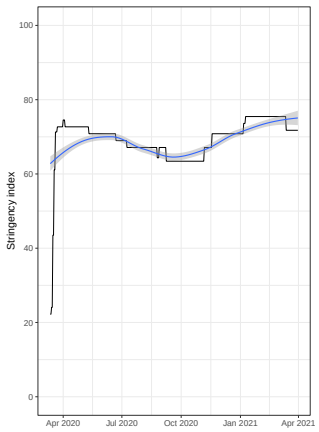


(a) Argentina

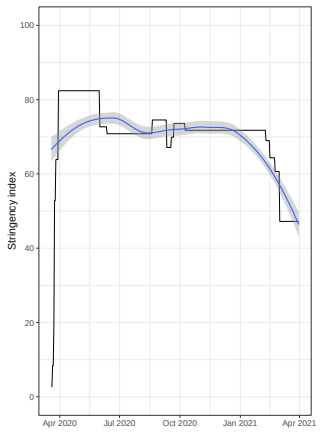


(b) Brazil

Figure 12: Dynamic of the stringency index in the two representative countries from South America.

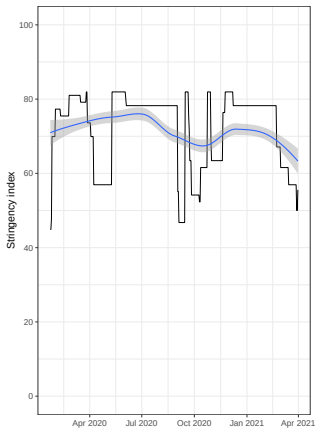


(a) Canada

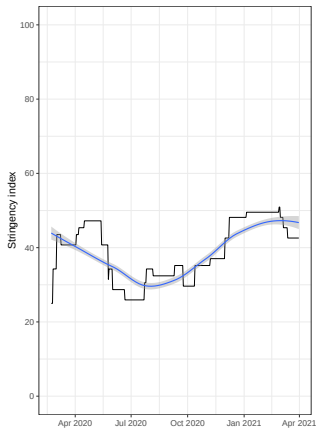


(b) Mexico

Figure 13: Dynamic of the stringency index in the two representative countries from North America.

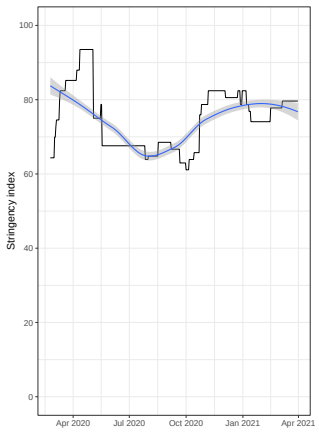


(a) China

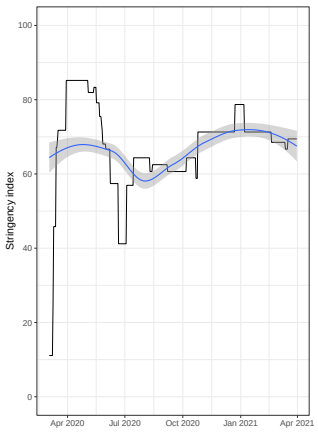


(b) Japan

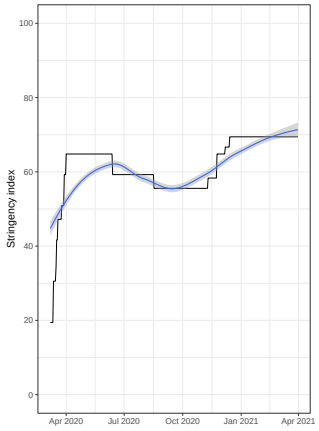
Figure 14: Dynamic of the stringency index in the two representative countries from Asia.



(a) Italy

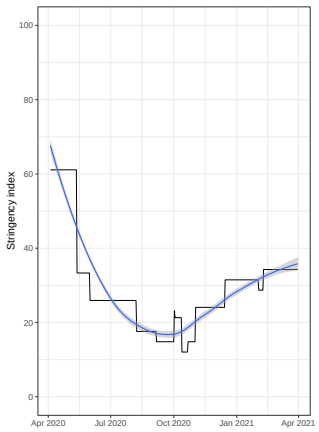


(b) Spain

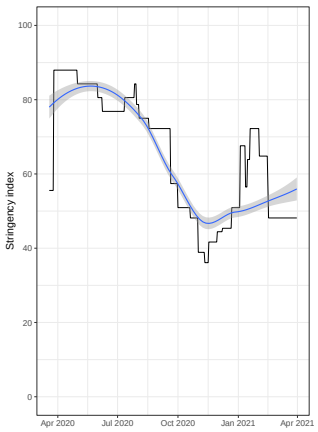


(c) Sweden

Figure 15: Dynamic of the stringency index in the three representative countries from Europe.

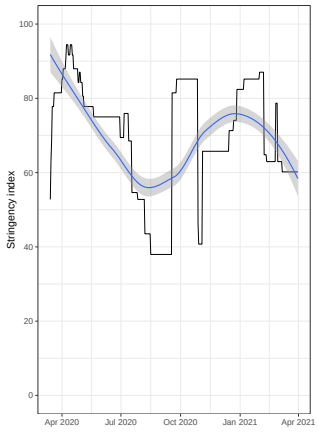


(a) Niger



(b) South Africa

Figure 16: Dynamic of the stringency index in the two representative countries from Africa.



(a) Israel

Figure 17: Dynamic of the stringency index in the representative country from Middle East.

Evaluation

We used weighted importance sampling (WIS) to evaluate the policy Precup et al. (2000); Jiang and Li (2015). WIS is weigthed variant of IS that weights the reward at time t by the cumulative importance sampling ratio given by $\rho_{1:t} = \prod_{k=1}^{t} \frac{\pi^e(a_k|s_k)}{\pi^b(a_k|s_k)}$, where $\pi^e(a_k|s_k)$ is the proposed policy that one want to evaluate and $\pi^b(a_k|s_k)$, known as the behaviour policy, is the probability used to generate actions in the retrospective dataset. Then, the WIS estimator is given as follows:

$$V_{step-WIS} = \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T_i} \gamma^{t-1} \frac{\rho_{1:t}^i}{w_t} R_t^i \tag{2}$$

where $w_t = \sum_{i=1}^{N} \frac{\rho_{1:t}^i}{N}$ as the average cumulative importance sample ratio at time t.

The behaviour policy is unknown and it must be estimated from the data. We estimated the behaviour policy $\pi^b(a_t|s_t)$ using a deep neural network model trained with a cross entropy loss function.

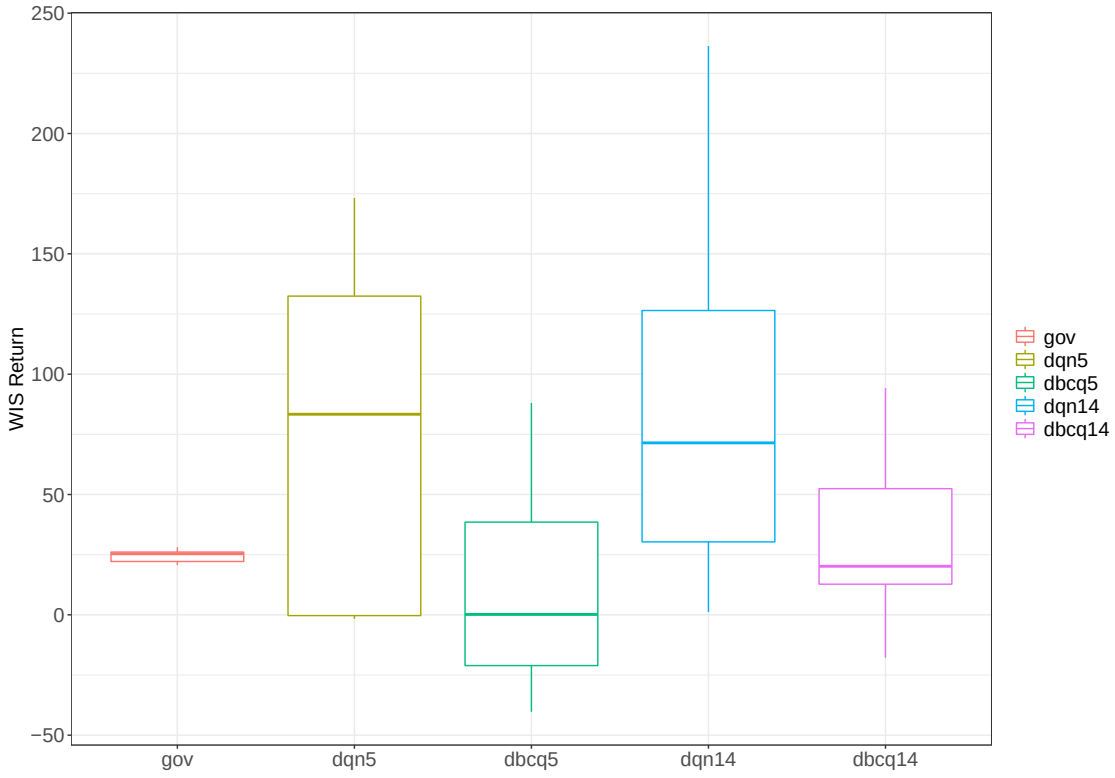


Figure 18: Estimated policy value for the observed government policy (gov) and the learned RL policy on the five held-out test set. dbcq5 = dBCQ using R_t^5, dbcq14 = dBCQ using R_t^{14}; dqn5 = DQN using R_t^5 and dqn14 = DQN using R_t^{14}

Training Loss, Reward R_t^5

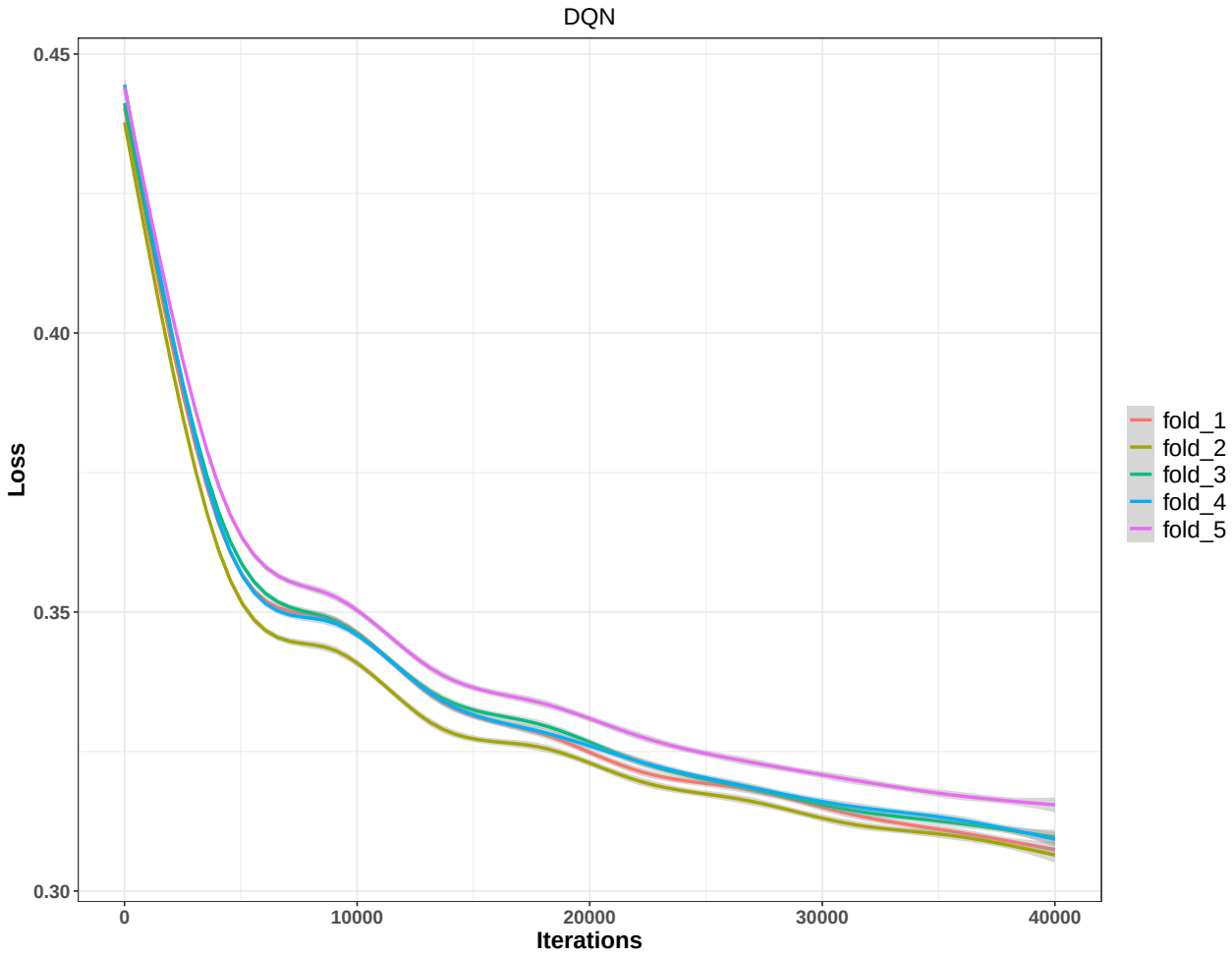


Figure 19: Train loss of the model DQN on the training set for 40000 iterations.

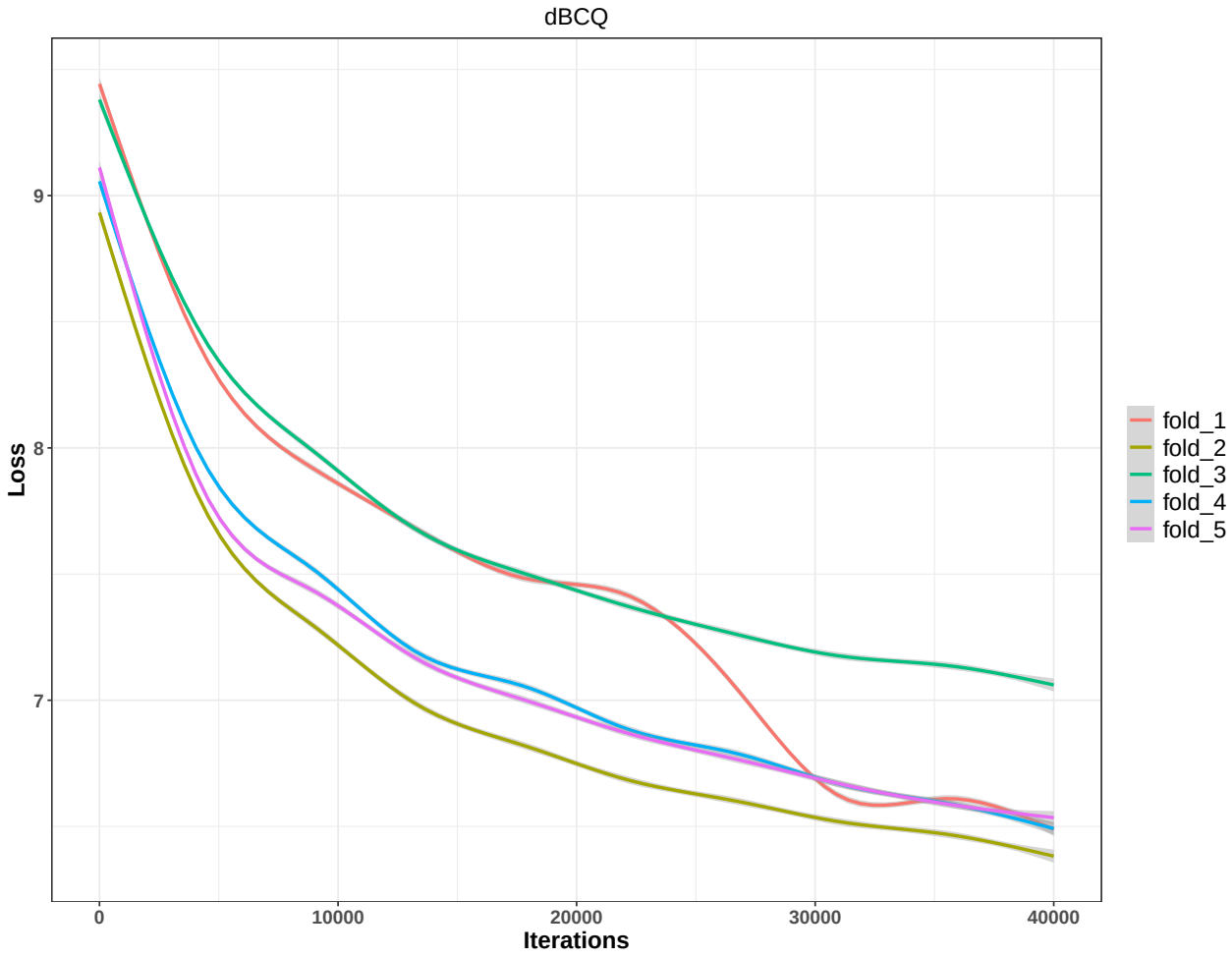


Figure 20: Train loss of the model discrete BCQ on the training set for 40000 iterations.

Training Loss, Reward R_t^{14}

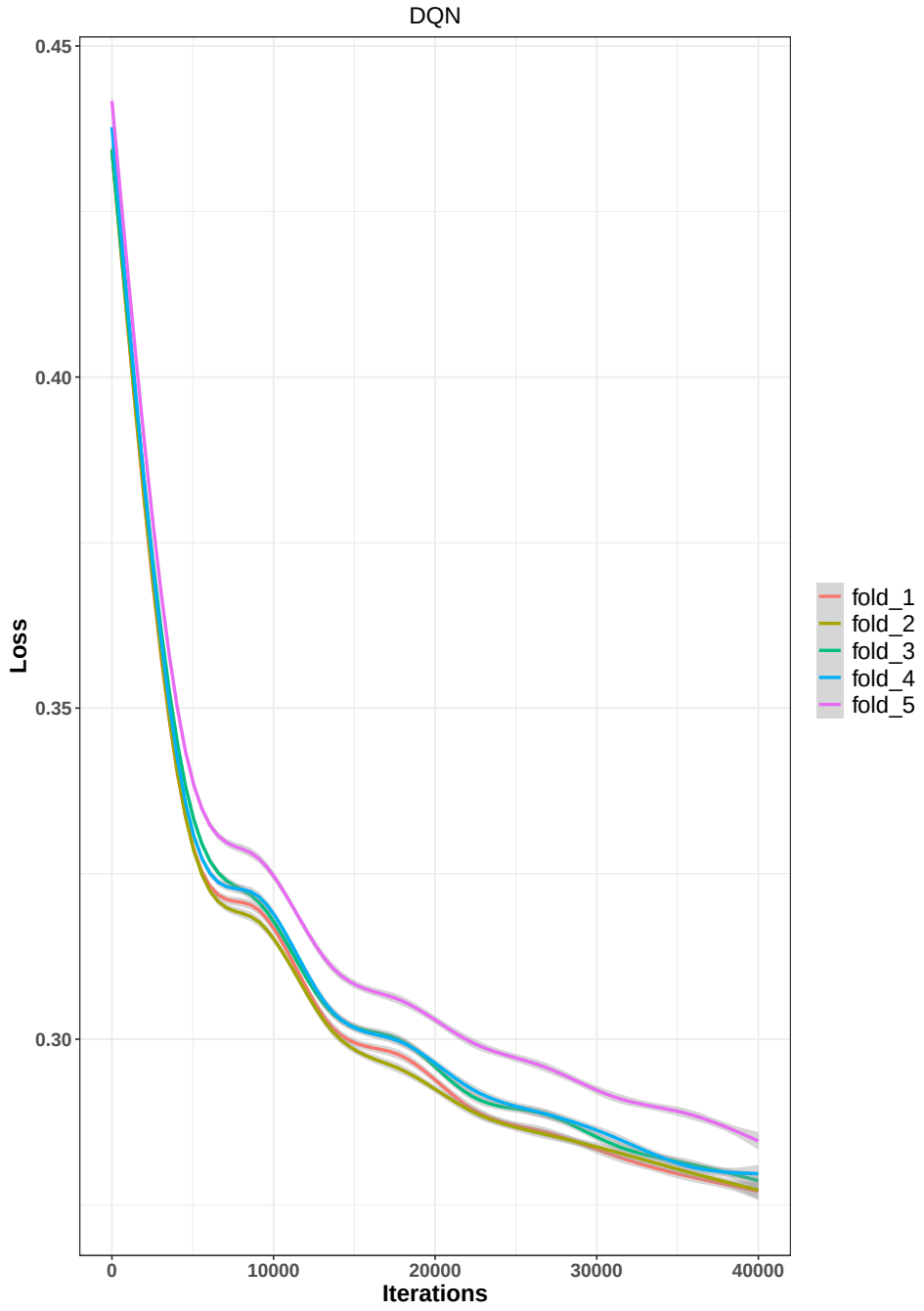


Figure 21: Train loss of the model DQN on the training set for 40000 iterations.

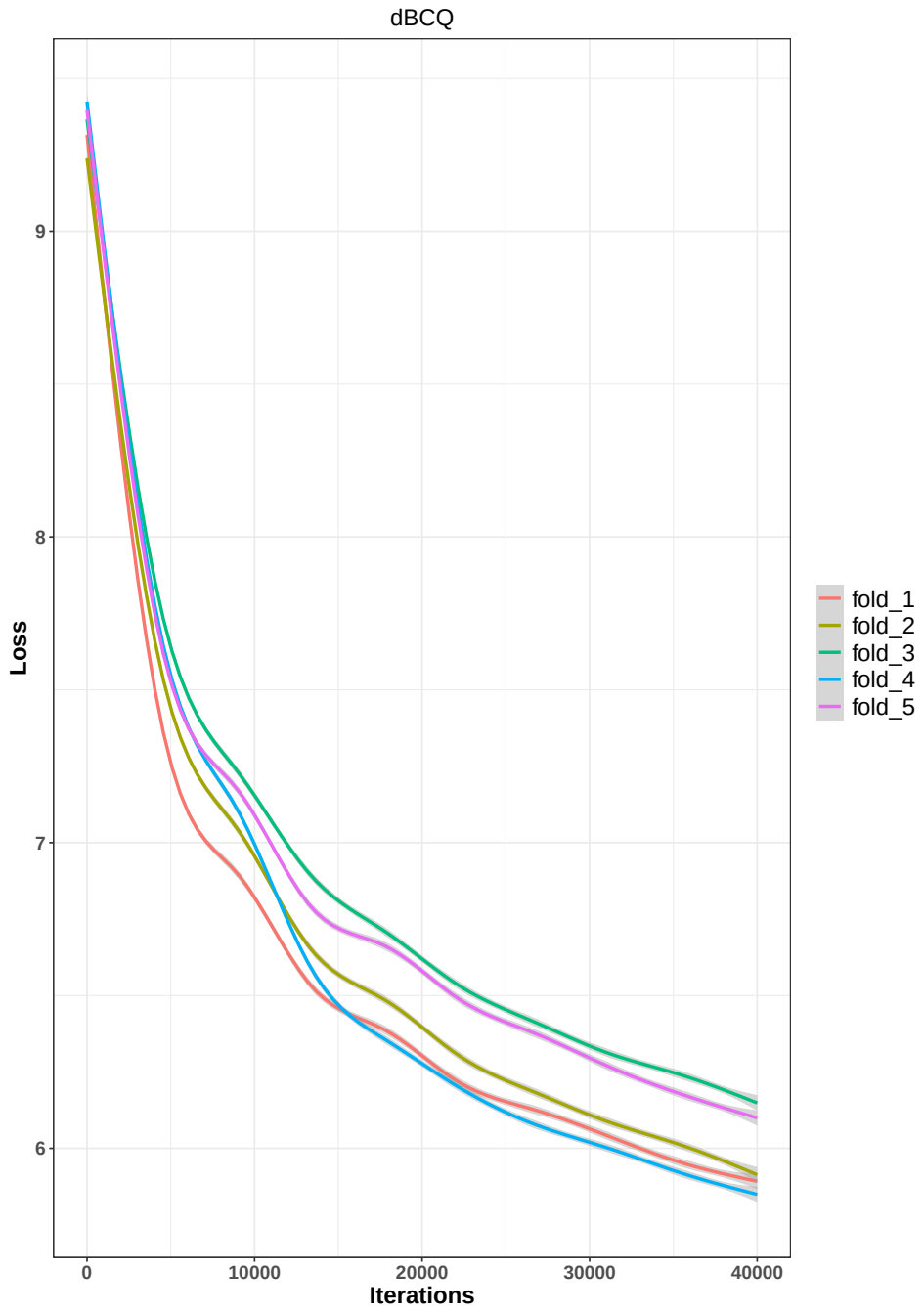


Figure 22: Train loss of the model discrete BCQ on the training set for 40000 iterations.

References

Richard Bellman. *Dynamic programming*. Princeton University Press, 1957.

Damien Ernst, Guy-Bart Stan, Jorge Goncalves, and Louis Wehenkel. Clinical data based optimal sti strategies for hiv: a reinforcement learning approach. In *Proceedings of the 45th IEEE Conference on Decision and Control*, pages 667–672, 2006. doi: 10.1109/CDC.2006.377527.

Scott Fujimoto, Edoardo Conti, Mohammad Ghavamzadeh, and Joelle Pineau. Benchmarking batch deep reinforcement learning algorithms, 2019.

Omer Gottesman, Fredrik D. Johansson, Joshua Meier, Jack Dent, Donghun Lee, Srivatsan Srinivasan, Linying Zhang, Yi Ding, David Wihl, Xuefeng Peng, Jiayu Yao, Isaac Lage, Christopher Mosch, Li wei H. Lehman, Matthieu Komorowski, A. Aldo Faisal, Leo Anthony Celi, David A. Sontag, and Finale Doshi-Velez. Evaluating reinforcement learning algorithms in observational health settings. *ArXiv*, abs/1805.12298, 2018.

Omer Gottesman, Fredrik D. Johansson, M. Komorowski, Aldo A. Faisal, D. Sontag, Finale Doshi-Velez, and L. Celi. Guidelines for reinforcement learning in healthcare. *Nature Medicine*, 25:16–18, 2019.

Peter Huber. Robust estimation of a location parameter. *Annals of Statistics*, 53(1):73–101, 1964.

Nan Jiang and Lihong Li. Doubly robust off-policy evaluation for reinforcement learning. *CoRR*, abs/1511.03722, 2015. URL `http://arxiv.org/abs/1511.03722`.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2015.

Matthieu Komorowski, Leo Anthony Celi, Omar Badawi, Anthony C. Gordon, and Aldo A. Faisal. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature Medicine*, 24:1716–1720, 2018.

Gloria Hyunjung Kwak, Lowell Ling, and Pan Hui. Deep reinforcement learning approaches for global public health strategies for covid-19 pandemic. *PLOS ONE*, 16(5):1–15, 05 2021. doi: 10.1371/journal.pone.0251550. URL https://doi.org/10.1371/journal.pone.0251550.

Sascha Lange, Thomas Gabel, and Martin Riedmiller. *Batch Reinforcement Learning*, pages 45–73. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012. ISBN 978-3-642-27645-3. doi: 10.1007/978-3-642-27645-3_2. URL https://doi.org/10.1007/978-3-642-27645-3_2.

Stephen A Lauer, Kyra H Grantz, Qifang Bi, Forrest K Jones, Qulu Zheng, Hannah R Meredith, Andrew S Azman, Nicholas G Reich, and Justin Lessler. The incubation period of coronavirus disease 2019 (covid-19) from publicly reported confirmed cases: Estimation and application. *Annals of Internal Medicine*, 172(9):577–582, 2020. doi: 10.7326/M20-0504. URL https://doi.org/10.7326/M20-0504. PMID: 32150748.

Christopher T. Leffler, Edsel Ing, Joseph D. Lykins, Matthew C. Hogan, Craig A. McKeown, and Andrzej Grzybowski. Association of country-wide coronavirus mortality with demographics, testing, lockdowns, and public wearing of masks. *The American Journal of Tropical Medicine and Hygiene*, 103(6):2400 – 2411, 2020. doi: 10.4269/ajtmh.20-1015. URL https://www.ajtmh.org/view/journals/tpmd/103/6/article-p2400.xml.

Sergey Levine, Aviral Kumar, G. Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *ArXiv*, abs/2005.01643, 2020.

Qun Li, Xuhua Guan, Peng Wu, Xiaoye Wang, Lei Zhou, Yeqing Tong, Ruiqi Ren, Kathy S.M. Leung, Eric H.Y. Lau, Jessica Y. Wong, Xuesen Xing, Nijuan Xiang, Yang Wu, Chao Li, Qi Chen, Dan Li, Tian Liu, Jing Zhao, Man Liu, Wenxiao Tu, Chuding Chen, Lianmei Jin, Rui Yang, Qi Wang, Suhua Zhou, Rui Wang, Hui Liu, Yinbo Luo, Yuan Liu, Ge Shao, Huan Li, Zhongfa Tao, Yang Yang, Zhiqiang Deng, Boxi Liu, Zhitao Ma, Yanping Zhang, Guoqing Shi, Tommy T.Y. Lam, Joseph T. Wu, George F. Gao, Benjamin J. Cowling, Bo Yang, Gabriel M. Leung, and Zijian Feng. Early transmission dynamics in wuhan, china, of novel coronavirus–infected pneumonia. *New Eng-*

land Journal of Medicine, 382(13):1199–1207, 2020. doi: 10.1056/NEJMoa2001316. URL https://doi.org/10.1056/NEJMoa2001316. PMID: 31995857.

Peng Liao, Predrag Klasnja, and Susan Murphy. Off-policy estimation of long-term average outcomes with applications to mobile health. *Journal of the American Statistical Association*, 116(533):382–391, 2021. doi: 10.1080/01621459.2020.1807993. URL https://doi.org/10.1080/01621459.2020.1807993. PMID: 33814653.

Daniel J. Luckett, Eric B. Laber, Anna R. Kahkoska, David M. Maahs, Elizabeth Mayer-Davis, and Michael R. Kosorok. Estimating dynamic treatment regimes in mobile health using v-learning. *Journal of the American Statistical Association*, 115(530):692–706, 2020. doi: 10.1080/01621459.2018.1537919. URL https://doi.org/10.1080/01621459.2018.1537919. PMID: 32952236.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin A. Riedmiller, Andreas Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015. URL http://dblp.uni-trier.de/db/journals/nature/nature518.html#MnihKSRVBGRFOPB15.

Xuefeng Peng, Yi Ding, David Wihl, Omer Gottesman, Matthieu Komorowski, Li-wei H. Lehman, Andrew Ross, Aldo Faisal, and Finale Doshi-Velez. Improving sepsis treatment strategies by combining deep and kernel-based reinforcement learning, 2019. URL https://arxiv.org/abs/1901.04670.

Niranjani Prasad, Li-Fang Cheng, Corey Chivers, Michael Draugelis, and Barbara E. Engelhardt. A reinforcement learning approach to weaning of mechanical ventilation in intensive care units. *ArXiv*, abs/1704.06300, 2017.

Doina Precup, Richard S. Sutton, and Satinder Singh. Eligibility traces for off-policy policy evaluation. In *ICML*, 2000.

Aniruddh Raghu, Matthieu Komorowski, Imran Ahmed, Leo Anthony Celi, Peter Szolovits, and Marzyeh Ghassemi. Deep reinforcement learning for sepsis treatment. *ArXiv*, abs/1711.09602, 2017a.

Aniruddh Raghu, Matthieu Komorowski, Leo Anthony Celi, Peter Szolovits, and Marzyeh Ghassemi. Continuous state-space models for optimal sepsis treatment: a deep reinforcement learning approach. In *MLHC*, 2017b.

Hannah Ritchie, Edouard Mathieu, Lucas Rodés-Guirao, Cameron Appel, Charlie Giattino, Esteban Ortiz Ospina, Joe Hasell, Bobbie Macdonald, Diana Beltekian, and Max Roser. Coronavirus pandemic (covid-19). `https://ourworldindata.org/coronavirus`, 2020.

Richard S. Sutton and Andrew G. Barto. *Introduction to Reinforcement Learning*. MIT Press, Cambridge, MA, USA, 1st edition, 1998. ISBN 0262193981.

Wei-Hung Weng, Mingwu Gao, Ze He, Susu Yan, and Peter Szolovits. Representation and reinforcement learning for personalized glycemic control in septic patients. *ArXiv*, abs/1712.00654, 2017.

Yufan Zhao, Michael Kosorok, and Donglin Zeng. Reinforcement learning design for cancer clinical trials. *Statistics in medicine*, 28:3294–315, 11 2009. doi: 10.1002/sim.3720.