Reward architecture in Deep reinforcement learning for disaster road management plan
- Focusing on the application of Envelop multi-objective optimization algorithms -

Soohyun JOO*, Yoshiki Ogawa*, and Yoshihide Sekimoto*

Abstract: Reward in Deep RL infers heuristic knowledge of the goal. To derive optimal policy for rapid disrupted mobility recovery, we set human mobility recovery rate, road connectivity, travel cost as the reward factors, and generated the reward framework. However, there exist various relative importance of reward factors, and if the preference is different from the original one, we could not be sure the optimal policy can always be identified. That is, if the framework does not properly represent optimization problem, it might fail to get optimal road recovery plan. Therefore, it is quite important to identify a set of relative preference which derive optimal policies for enhancing the robustness and generality of our model. In this paper, we would identify generalized optimal policy over the space of all possible preferences and confirm underlying preference by applying envelope multi-objective optimization algorithm.

Keywords: Deep reinforcement learning, relative importance, road restoration, optimal policy

１. Introduction

In 2018, successive heavy rain resulted in multi-locational, devasting flood and landslide. Approximately 80% of high-risk areas in Hiroshima and Okayama Prefecture had been flooded or corrupted. The restoration of roads related to real life did not begin until 14 days after the disaster. In addition, social, economic activities harmed, and rescuing victims and restoring other facility had difficulty with temporarily blocked roads.

The government's road management plan focused on the rapid recovery of disrupted human mobility. However, with increased uncertainty about post-disaster, their plan relies on the current situation and their experience. Moreover, they have difficulty predicting and evaluating the change of human mobility under recovery operation. These limitations make the current road recovery plans inefficient.

For effective road recovery strategies in post-disaster, we proposed data-driven Deep reinforcement learning (Deep RL) algorithm combining traffic flow analysis based on mobile phone GPS data. Concretely, traffic analysis allows to predict and evaluate mobility recovery with a series of reconstructions. Stochastic numerical models with unstructured input data of road-usage could determine optimal solution with consideration of human mobility.

Deep RL deals with many uncertainties that be difficult to fully consider and identifies optimal answer by exchanging environmental information and reward signal. The optimization goal is to maximize the sum of discount reward. Among four components in Deep RL (e.g., agent, action, state, reward), the reward infers heuristic knowledge of the goal, and encourages a behavior consistent with some prior information. For the optimal policy, the reward system should be able to express the effectiveness of agent's action based on the given state information.

The considerations in the reconstruction project are the quality, working duration, and cost. To derive optimal policy within specific number of steps which means the project period, we set human mobility recovery rate, road connectivity, travel cost as the reward factors, and created the reward framework. Certain factors are regarded as more important or less important, depending on the opinion of the decision maker. But, if the weighted reward does not properly represent optimization problem, it might

* Center for Spatial Information Science (the University of Tokyo)
〒**153-8505** Ce507, 4-6-1, Komaba, Meguro-ku, Tokyo E-mail：**soo1992@iis.u-tokyo.ac.jp**

fail to get optimal road recovery plan.

It is quite important to derive the optimal policies from any preference for enhancing the robustness and generality of our model. In this paper, we determined general optimal policy for all possible preference spaces and identified the difference for each preference by applying envelope multi-objective optimization algorithm.

The reminder of this paper is organized as follows. In Section 2, we explain the methodology. Section 3 describes the digital road map and mobile phone GPS data in Hiroshima Prefecture. In Section 4, we illustrate the suggested decision-making system. In Section 5, we explain our results. Section 6 presents discussion and conclusion.

２. Background

Deep RL is formulated via Markov decision process. The agent in this method interacts with an environment at discrete time steps t, and could learn the effectiveness of each action, and approximate the best action with given situation data.

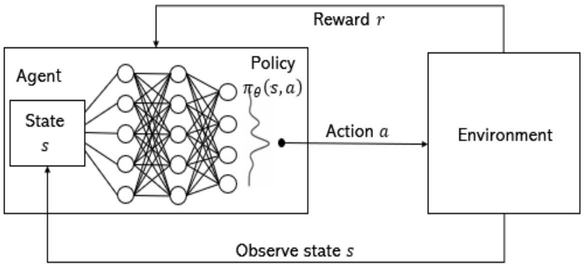

Figure 1. The framework of reinforcement learning

This method estimated action value function employing deep neural network and improving itself iteratively with the basic update method. The network extracts the contextual information from a large amount of unstructured data, and models parameterized approximate function without the curse of dimensionality.

2.1. Multi-Objective reinforcement learning

The aforementioned methods have focus on single-objective settings. On the other hands, multi-objective reinforcement learning (MORL) have been suggested to improve the performance of the Deep RL agents on real-world problem having multiple conflicting objectives.

The MORL could be expressed the tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \boldsymbol{r}, \Omega, f_\Omega \rangle$ with state space $\mathcal{S}$, action space $\mathcal{A}$, transition matrix $\mathcal{P}$, vector rewards $\boldsymbol{r}$, one for each goal, the preference space Ω, and preference function f_Ω which typically defined with a linear scalar function.

$$f_{\boldsymbol{w}}(\boldsymbol{r}(s,a)) = \boldsymbol{w}^{\intercal}\boldsymbol{r}(s,a) \qquad (1)$$

The optimal policy in MORL depends on the relative preference among competing criteria. There are two types of policy method: single-policy methods and multi-policy methods. The agents in single-policy methods identify the optimal policy with preference they already know. Otherwise, multi-policy approaches learn a set of policies to obtain Pareto frontier of optimal solutions, so it is able to estimate optimal policies without prior knowledge of preference.

We focus on multi-policy method to make our model adapt to any government's preference and identify optimal recovery plan. Several algorithms have been devised: multiple runs of a single-policy method over several preferences, the policy-based RL learning the optimal manifold, encapsulating preferences as input. However, these methods have difficulty adapting to new preferences that have not been used in training phase.

2.2. Envelop multi-objective RL

Yang et al. suggested a MORL algorithm called envelop multi-objective reinforcement learning. Their method estimated a single policy network that is optimized over the entire space of preferences defining an optimality filter $\mathcal{H}$. This estimates the current solution frontier to produce the action value function that optimizer utility given state $\mathcal{S}$ and preference $\boldsymbol{w}$. Moreover, they increased sample efficiency detaching preference from the transitions, and accelerate the coverage of one parametric function to the optimal one with reasonable sampled trajectories.

3. Data Collection / Processing

We utilized three datasets to estimate road conditions in terms of recovery operation and define the agent's action space.

3.1. Mobile phone GPS data

The Agoop Co., Ltd. provided mobile phone GPS records for approximately 0.3% of the population in Hiroshima Prefecture (Hiroshima, Higashi-Hiroshima, Kure). This dataset contains latitude, longitude, accuracy and timestamp. As shown in Table 1, the number of users is 3,817. The GPS logs amounted to 102,821 from June 1 to June 30.

Table 1. The detail of mobile phone GPS data

Observed Period	**Average daily number of IDs in the target area**	**Average daily GPS logs in the target area**
2018/06/01 ~ 2018/06/30	3,817 (0.26% sample rate)	102,821 (avg. 27 logs/user)

The accuracy in GPS data means radius error caused by some technical obstacles. Our data has a margin of error of a 500-meter radius on average. We select the origin-destination matrix (O-D) as the representatives of human movement to alleviate the bias of inaccurate information. We confirmed that 2,970 kinds of O-Ds pass through the afflicted road by Western Japan flooding.

3.2. Local geographic information

Ministry of Land, Infrastructure, Transport and Tourism (MLIT) and Japan Statistics Bureau adjoin position factors and regional information (e.g., resident, commuter, disaster risk etc.,). The information of commuters and residents of 1 km grid are utilized to calculate the hourly travel demand of O-Ds. Besides, the road topology data is one essential factor of traffic assignment with sequential reconstruction. We get real road network's information from Japan Digital Road Map Association. Digital Road Map (DRM) is the standard national DRM supporting Japanese ITS infrastructures. The road networks are line data containing information such as road types and widths.

3.3. Damage road information

Disrupted road information provided by Municipality and MLIT includes the road name, the extent of damage, and restorative state. The road reconstruction process consists of three types: Road closed, one-way traffic, and the completion of recovery. This information help citizens and business factors identify the current state and determine the most reasonable path.

We used the location information of damage roads and the details of road in DRM (e.g., the number of lanes, width, road type, length) and defined the target road in action space and the workload (total area) of each action.

4. Decision-making system

Government, the main decision maker, devises workforce distribution plan. They want the operation crews to cooperate with others while fulfilling their responsibility. Our model is based on multi-agent Deep RL which induce to do collaboration sharing each behavior data.

4.1. Component of Decision-making system

4.1.1. Agent

One operation crew is the agent, and the number of workers is optional. The crews assumed to do "Evacuation and Embankment", which is to remove soil and rock, replace, and compact demolished road section. Hiroshima Prefecture sets the amount of available operation per a day (8 hours) with one worker. Table 2 describes the operation hour it takes to complete $100m^2$. We set that the daily workload of one worker would be $256m^2$.

Table 2. The working hour of one agent's operation

$(h/100m^2)$

Excavation Depth / Type of Machine	Under 40 cm	40cm ~ 80cm	80cm ~ 120cm
Backhoe Shovel	2.0	3.3	4.7
Large Breaker & Backhoe Shovel	2.1	2.8	3.5
Concrete Crush & Backhoe Shovel			

4.1.2. Action

The agent in Deep RL selects one action, and makes the change of environment at each time step. In this paper, the agent selects one disrupted road at each time step (one day) and had to recover within pre-defined daily workload.

According to Section 13 of Road Act, MLIT and municipal government are the main authorities. MLIT mainly takes change of expressway and highway. On the other hand, municipal administrations take the responsibility of other parts in their region. In Western Japan, the number of disrupted roads under the jurisdiction of one local government is at least fifteen.

The potential users in our framework might be the municipals, that is, one working group might be placed in one damage road among the government's jurisdiction. Accordingly, we think that the agent in our model should cover fifteen disrupted roads at least. So, the action space consists of fifteen target roads.

4.1.3. State

The state space describes the input data for determining action value function. The data included the information about corresponding reward, environment and implied the agent's objective and the effect of each action.

We wanted the agents to cooperate with each other, so utilized the method suggested by Foerster et al. In detail, we add the numerical signals to the input layer, and make the agent each identify the cooperation with other agents via learning process. At first, we define the meaning of cooperation and communication protocol representing the teamwork.

Restoring inter-connected damage roads concurrently helps human mobility recover rapidly. In case of the restoration of D21 and D36 at the same time, there are three types of recovered traffic: 1) O-Ds passing only D21, 2) passing only D36, and 3) passing both D21 and D36. We define the cooperation between two agents using O-Ds' road usage. Let denoted by $\mathcal{T}_c^e$ traffic volume driving down both damage road c and damage road e. $\mathcal{R}$ refers to the set of damage roads covered in multi-agent RL system. The effect of cooperation is calculated with Equation 2:

$$\mathcal{CP}_t^{\mathcal{A}_o} = \frac{\mathcal{T}_c^e}{\sum_{d,g\in\mathcal{R}} \mathcal{T}_d^g}, \qquad c,e\in\mathcal{R} \tag{2}$$

where $\mathcal{CP}_t^{\mathcal{A}_o}$ means the cooperation effect at step t assuming that the agent A selects damage road c and the agent $\mathcal{A}_o$ selects damage road e.

The state space includes four factors: the operation progress rate of each damage road $(\mathcal{P}_t^{r_n})$ in action space, human mobility recovery rate of each disrupted road section $(\mathcal{TR}_t^{r_n})$, travel time $(\mathcal{T}_t)$, and the overall human mobility recovery rate $(\mathcal{TR}_t^{O})$ at each time step t. With these components, we add the effects of cooperation with other agents$(\mathcal{CP}_t^{\mathcal{A}_m})$and the impact of selected damage road(α_{r_t}). Using the protocol, each agent predicts the effect of teamwork of each action and determines its policy with the consideration of collaboration.

$$s_t = \left\{ \begin{matrix} \mathcal{P}_t^{r_1},\cdots,\mathcal{P}_t^{r_n},\mathcal{TR}_t^{r_1},\cdots,\mathcal{TR}_t^{r_n},\mathcal{T}_t,\mathcal{TR}_t^{O},\mathcal{CP}_t^{\mathcal{A}_1} \\ ,\cdots,\mathcal{CP}_t^{\mathcal{A}_m},\alpha_{r_t} \end{matrix} \right\} \tag{3}$$

4.1.4. Reward

The road reconstruction is related to multi-objective optimization (MOO). The goals in MOO are in complementary, conflicting or independent relationship. Brys et al. suggested the combination of basic reward and

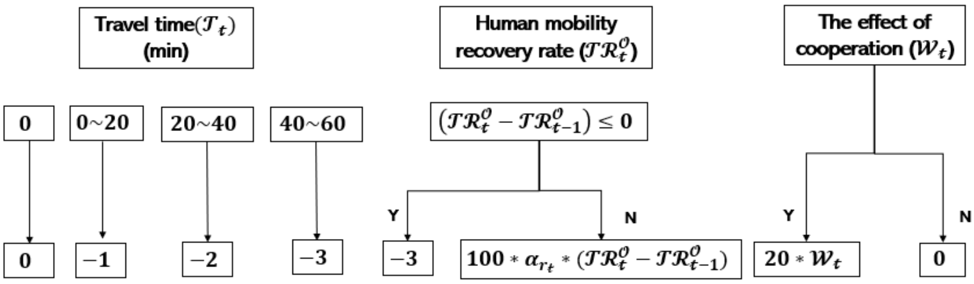
Travel time($\mathcal{T}_t$) (min)
Human mobility recovery rate ($\mathcal{TR}_t^O$)
The effect of cooperation ($\mathcal{W}_t$)
0
0~20
20~40
40~60
$(\mathcal{TR}_t^O - \mathcal{TR}_{t-1}^O) \leq 0$
Y
N
Y
N
0
−1
−2
−3
−3
$100 * \alpha_{r_t} * (\mathcal{TR}_t^O - \mathcal{TR}_{t-1}^O)$
$20 * \mathcal{W}_t$
0

Figure 2. Reward setting

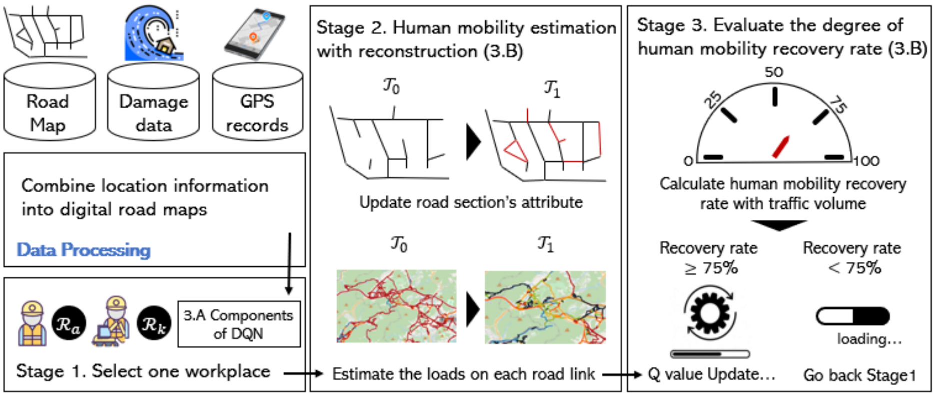
Road Map
Damage data
GPS records
Combine location information into digital road maps
Data Processing
$\mathcal{R}_a$
$\mathcal{R}_k$
3.A Components of DQN
Stage 1. Select one workplace
Stage 2. Human mobility estimation with reconstruction (3.B)
$\mathcal{T}_0$
$\mathcal{T}_1$
Update road section's attribute
$\mathcal{T}_0$
$\mathcal{T}_1$
Estimate the loads on each road link
Stage 3. Evaluate the degree of human mobility recovery rate (3.B)
0
25
50
75
100
Calculate human mobility recovery rate with traffic volume
Recovery rate ≥ 75%
Recovery rate < 75%
loading...
Q value Update...
Go back Stage1

Figure 3. The framework of decision-making system

extra reward to help the agent explore behaviors with heuristic information of system.

The fundamental goal is to restore disrupted human mobility to normal state rapidly. We define the basic reward as the change of human mobility recovery rate. This is because we want the agent to learn which roads have the high impact on human mobility recovery while choosing its action, not giving prior knowledge. Further, the agents are required to have a cooperative altitude and consider the travel time, the operation costs. We set the sum of numerical protocol ($\sum_{k=1}^{m} \mathcal{CP}_t^k$) and travel cost as the extra rewards.

Human mobility recovery rate, the main reward, is the composite effect of all agents' action. The usage of original value does not properly demonstrate the recovery effect of each action, and then the common reward interferes with achieving their goal making the agents lazy. As shown Figure 2, we utilize the traffic weight α_{r_t} indicating how much disrupted road r_t they select at time step t has affected the overall human mobility recovery rate. This weight is the proportion of traffic volume of each damage road for the total traffic volume.

4.1.5. Learning process

The framework is illustrated in Figure 3. The agent selects one disrupted road based on Q value function at each time step. With target roads' accumulative progress rate at each time step, the attribute of each damage road (e.g., travel time, basic capacity etc.,) and travel demand of each O-D are recovered and updated. And then, traffic flow on each link is estimated and human mobility recovery rate is calculated. At last, if mobility recovery rate is over target value, the simulation is over and current policy is updated using accumulated experiences that are acquired through the interaction with environments. Otherwise, the agent does the whole processes repeatedly up to the maximum number of steps.

4.2. Human mobility estimation model

We estimate loads on each road section applying stochastic traffic assignment.

4.2.1. The restoration of road capacity

Soon after the flooding, the operation crew did reconstruction to secure practicable lanes. The details of the meaning of reconstruction, the initial state of damage road are as follows:

- **Definition 1.** Reconstruction is to secure available lands up to the original level.
- **Assumption 1.** Road's capacity represents the maximum traffic using available lanes. The initial capacity of all damage roads is zero.

The recovery of decreased capacity is followed with the work progress rate. Sigmoid function is used to estimate working rate. This function describes productive efficiency in construction project. Fourteen days after Western Japan flooding, vehicles could drive on most highways and main roads. So, we assumed that 14 days are needed to complete the restoration of road with the maximum amount of operation. Road's cumulative progress rate is approximated with Equation 4:

$$\mathcal{P}_t^m = \frac{1}{1+e^{-0.8x_t}}, \tag{4}$$

$$\because x_t = -7 + 13\left(\frac{\sum_{k=0}^{t} \mathcal{W}_k}{\mathcal{S}_m}\right)$$

where $\mathcal{P}_t^m$ is the cumulative progress rate of damage road m at step t. $\sum_{k=0}^{t} \mathcal{W}_k$ means the cumulative workload up to the step t. $\mathcal{S}_m$ is the total workload of corresponding road.

4.2.2. The travel demand with reconstruction

People are unable to move or isolated if the road system is seriously damaged and there exist no detour. As damaged roads have been restored, traffic in the network has been increased to pre-disaster level. Accordingly, we set one assumption with trip generation under reconstruction.

- **Assumption 2.** The travel demand is influenced by the presence of alternative routes, the minimum value of the cumulative progress of disrupted road that O-D passes through on normal days.

4.2.3. Traffic allocation assignment

Stochastic traffic assignment model is utilized to estimate vehicular flow under recovery operation. O-Ds might select the trajectories having sufficient capacity, because disrupted road has the unstable capacity and the risk of the surge of travel time.

- **Assumption 3.** The amount of traffic on one of trajectories depends on the minimum capacity of the link that constitutes this route.

The traffic allocation algorithm is described using Meng et al. [44] notations. $G = (\mathcal{N}, \mathcal{A})$ refers to transportation network, where $\mathcal{N}, \mathcal{A}$ are the sets of nodes and link, respectively. O-D with origin node r destination node s is defined by $\mathcal{H}_{(r,s)}$. $\mathcal{R}$ and $\mathcal{S}$ mean the set of origin, destination respectively. Denoted by $\mathcal{K}_{rs}$ the set of paths connecting $\mathcal{H}_{(r,s)}$ by $q_{(rs),t}$ travel demand of $\mathcal{H}_{(r,s)}$ at each time step t.

With Assumption 3, denoted by $\mathcal{C}_t^k$ the minimum capacity, by $\mathcal{U}_t^k$ road usage rate on path $k \in \mathcal{K}_{rs}$ at each time step t. $\mathcal{F}_t^k$ means the traffic flow on path $k \in \mathcal{K}_{rs}$ at each time step t and is calculated with Equation 5:

$$\mathcal{F}_t^k = q_{(rs),t} * \mathcal{U}_t^k, \qquad (5)$$

$$\because \mathcal{U}_t^k = \frac{\mathcal{C}_t^k}{\sum_{k \in \mathcal{K}_{rs}} \mathcal{C}_t^k}$$

The traffic flow of each link $\mathcal{V}_{a,t}$ at each time step t would be calculated with the fundamental flow equations

Figure 4. The damage road covered in our model

$$\mathcal{V}_{a,t} = \sum_{r \in \mathcal{R}} \sum_{s \in \mathcal{S}} \sum_{k \in \mathcal{K}_{rs}} \mathcal{F}_t^k \delta_a^k, a \in \mathcal{A}$$

$$\sum_{k \in \mathcal{K}_{rs}} \mathcal{F}_t^k = q_{(rs),t}, r \in \mathcal{R}, s \in \mathcal{S} \qquad (6)$$

where $\delta_a^k = 1$ if path $k \in \mathcal{K}_{rs}$ between $\mathcal{H}_{(r,s)}$ traverse link $a \in \mathcal{A}$, and 0 otherwise.

4.2.4. Human mobility recovery rate

We selected traffic volume as the recovery evaluation index, because our model focus on securing the stable condition. The human mobility recovery rate in each damage road and the overall road network are calculated for input layer and reward. The human mobility recovery rate of target road $\mathcal{K}$ at each time step t is the total amount of loads with respect to the total traffic volume normal days. Equation 7 is as follows:

$$\mathcal{TR}_t^{\mathcal{K}} = \frac{\sum_{a \in \mathcal{A}} \mathcal{F}_{a,t}}{\sum_{a \in \mathcal{A}} \mathcal{F}_{a,n}} \qquad (7)$$

where $\mathcal{A}$ is the set of links that make up the road $\mathcal{K}$ and each link denoted by a. $\mathcal{F}_{a,n}$ is the estimated traffic flow on link a on normal days, $\mathcal{F}_{a,t}$ is the estimated traffic flow in link a at each time step t.

5. Experiment

We utilized envelop MOO algorithm suggested by Yang et al. for identifying the generalized optimal policy over any preference.

Figure 4. The damage road covered in multi-agent system

5.1. Outline of model setting

There are three types of operation crew which consists of four, eight, and seventeen workers respectively. They have specific fifteen disrupted roads as the work target. Our system could consider forty-five damage roads simultaneously. Figure 4 describes the target roads subjected to each groups' action space. The common objective of all agents is to recover human mobility up to

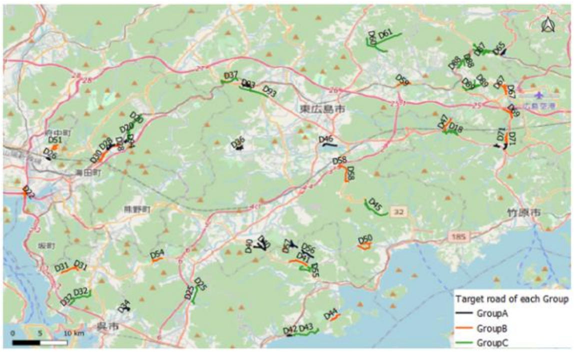

Figure 4. The damage road covered in our system

75% of pre-disaster level within 30 steps.

In Yang et al.'s paper, they estimate an action value function through over 3000 training episodes, and verify the robustness of their algorithm. Our model includes the /stochastic traffic assignment. It is practically unreasonable to do thousands of training process. Accordingly, the agent's preference space is defined as follows: Each weight is $w_i \in [0, 0.1, \cdots, 0.9, 1]$, and the sum of three weights is 1.

In training phase, the operation crews learned about each preference three times. And then, the agents act greedily following to the optimal policy for each preference.

5.2. Learning Result

In sub-section, we identify a set of actions selected in accordance with policy. Figure 5 describes the ternary contours plots of the human mobility recovery rate and the number of steps of each episode with different weight set in test phase. The generalized policies were able to determine optimal policies for 36 cases of 62 preferences in the set of weights.

The human mobility recovery rate in final varies depending on the given preferences. However, we could confirm that the lower the weight on the mobility recovery rate, the lower final recovery rate. It is important to give the agent the evaluation signal related to the agent's goal, even a small value.

5.3. Verification

Deep RL used a parameterized function approximator, and adjust the weights incrementally during learning. Before approximation phase, the optimization objective function of action value function, the loss function, is needed to define. Envelop MOO algorithm utilized *homotopy optimization* method which trades off between main loss $\mathcal{L}^{\mathcal{A}}$ and auxiliary loss function $\mathcal{L}^{\mathcal{B}}$. This method updates the policies in a direction that obtain better utility rather than approaching the optimization result from the previous step.

The gradient descent algorithm is utilized to solve the weight parameter which could find the minimum loss, the goal of loss function. In other words, if the loss value of a set of sample data is close to zero, it might lead to a local / global optimal solution.

Figure 6. The loss value and weighted recovery effect

$$\mathcal{L}(\theta) = (1-\lambda)\mathcal{L}^{\mathcal{A}}(\theta) + \lambda\mathcal{L}^{\mathcal{B}}(\theta)$$

$$\because \mathcal{L}^{\mathcal{A}}(\theta) = \mathbb{E}_{s,a,w}[\|y - Q(s,a,w;\theta)\|_2^2] \qquad (8)$$

$$\because \mathcal{L}^{\mathcal{B}}(\theta) = \mathbb{E}_{s,a,w}[|w^{\top}y - w^{\top}Q(s,a,w;\theta)|]$$

5.3.1. Case analysis

Based on the characteristics of the loss function, we estimated the loss value of the experience data from each

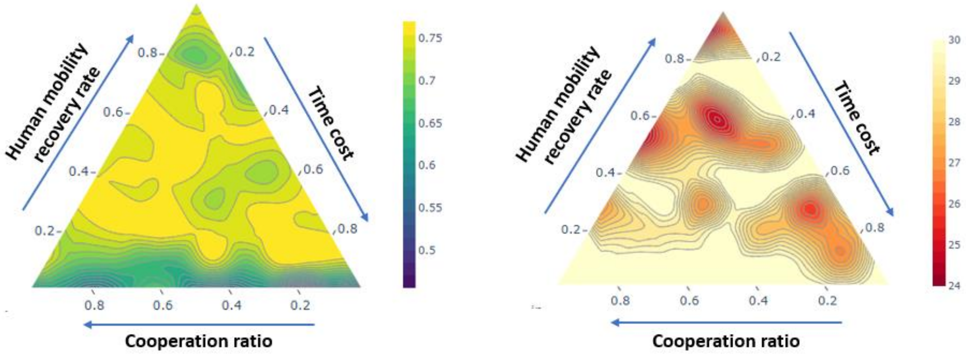

Figure 5. The ternary contours plots of human mobility recovery rate (left) and the number of steps (right)

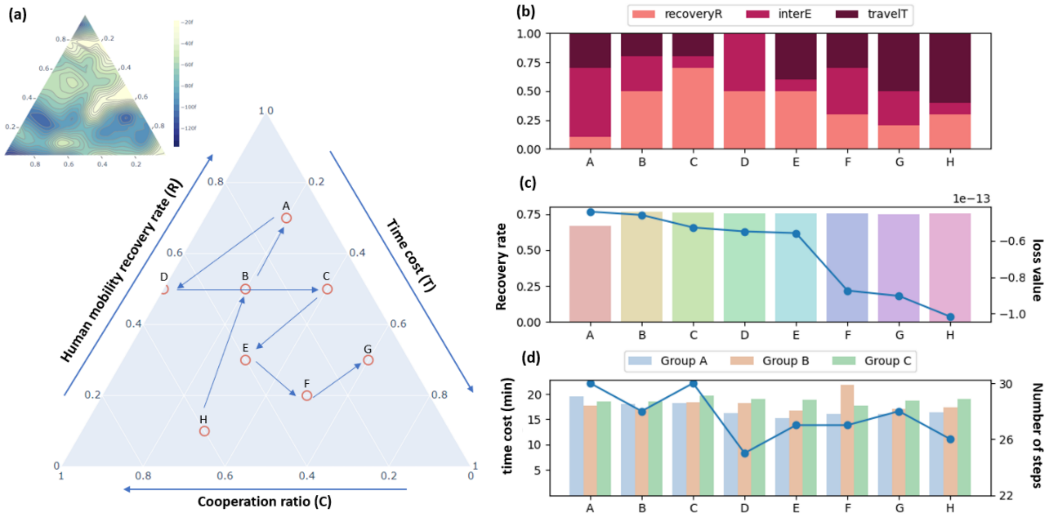

Figure 6. The detail information of each episode having specific preference; (a) the ternary plot of loss value and the selected eight cases, (b) the weight values, (c) the loss value (line graph) and the human mobility recovery rate (bar graph), (d) The average travel time of each workgroup (bar graph) and the number of steps (line graph)

episode of the test phase (with different preference). And then, we tried to understand how similar a set of agents' actions with each weight is to the generalized optimal policy network.

We selected eight cases based on the sum of all agents' loss value. As Figure 6.b, 6.c shown, except for A with the largest loss value, all episodes achieved the goal within 30 steps. The smaller the loss value, the shorter the number of steps it takes to reach human mobility recovery rate of over than 75%. While the relative preferences of each weight (Figure 6.a) vary, the average travel cost of agents' group (Figure 6.c) in cases with large weight of travel cost is generally shorter than that of other episodes.

The agent in our framework regards other agents as the environment, and has different action space depending on the group it belongs to. With these characteristics, they have their own policies, but the policies of workers in the same group might have the similarities. So, we compared the policies of each agent group with the eight cases in Figure 6.a.

Before comparing, we extracted general priority order of each episode using the operation order and the frequency. Plus, we estimated the recovery effect of each action, using the average of human mobility recovery rate variation for work progress variation. Figure 7 describe the loss value and the weighted average of recovery effects using operation priority as the weighting value.

We think that the policy for the rapid human mobility recovery is to prioritize roads with high recovery effect. That is, the larger the weighted average of recovery effects, the better policy for the goal. However, as Figure 8 shown, there is a slight gap between the estimated optimal policy and our thoughts.

This is because the difference between the preference used when updating policies and when selecting actions.

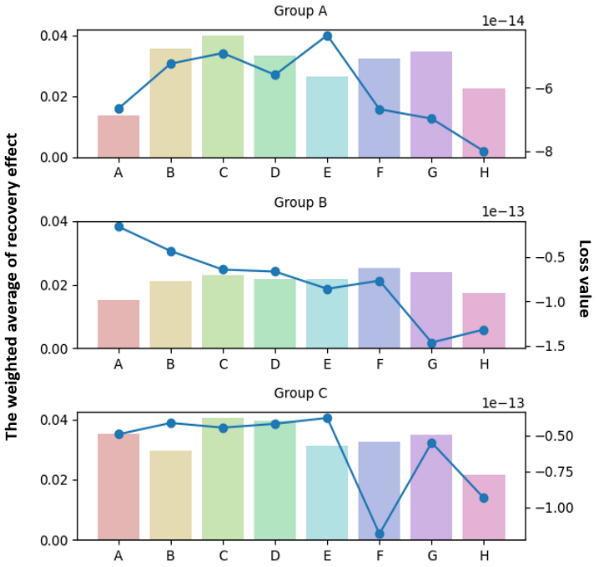

Figure 7. The loss value and weighted recovery effect

For the former, the agent estimated the approximate function based on the utility of sample data for all preferences. On the other hand, the agent selects the actions greedily using function value derived by entering specific preference into the generalized policy networks. The actions might make the weighted reward maximize.

Figure 8 shows the relationship between the operation order and recovery effect of selected roads for each episode. We identified that the smaller the loss value (the more similar the optimal policy), the agents tend to choose the actions with high recovery effect preferentially.

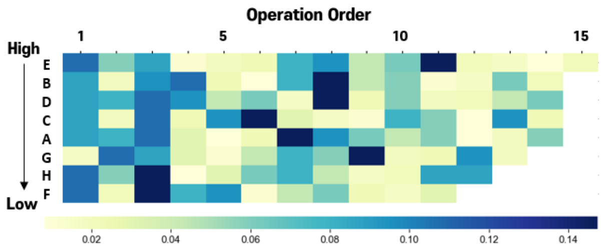

Figure 8. The relationship between the priority order and recovery effect for each episode (Group C)

6. Conclusion

This study combined envelope MOO and human mobility data to identify generalized restoration plan over any preference. In terms of practical run-time, we defined the preference space with 62 sets of weighted values. The sum of three factors in one set of preference is one.

We tested the adaptability of the agent's preference space with generalized approximation policy network. The agents could achieve human mobility recovery rate over 75% in 30 steps, depending on the optimal policies for 36 types of weights.

We confirmed which the sequence of actions with each preference is similar to the generalized policy networks using the loss value. 8 cases are selected based on the loss value, and the preference, the human mobility recovery rate, and the average of travel time of each episode were graphically plotted. As a result, the greater the loss (the greater the distance from the optimal choice), the lower the recovery rate. For achieving the agents' objective, the rewards related to their targets should be provided noticeably, even at low values.

For rapid human mobility recovery, it is natural to restore road with high recovery effect first. After learning, the agents tend to choose this type of roads preferentially. However, there is a slight gap between estimated policies and conventional thought. This is because the differences arise between generalized policies for overall preferences and those adjusted for individual preferences.

Reference

K. Shahabi, and J. P. Wilson (2018) Scalable evacuation routing in a dynamic environment, Comput Environ Urban Syst, 67, 29-40

N. Y. Aydin, H.S. Duzgun, H.R. Heinimann, F. Wenzel, and K.R. Gnyawali (2018) Framework for improving the resilience and recovery of transportation networks under geohazard risks, Int. J. Disaster Risk Reuct, 31, 832-843

R.S. Sutton, and A.G. Barto (2018) Reinforcement learning: An introduction, MIT Press

T. Brys et al. (2014) Multi-objectivization of reinforcement learning problems by reward shaping, IJCNN, 2315-2322

Hiroshima Prefecture (2015) Standard specification for civil construction management

D.L. Adam (2004) Theory and application of reward shaping in reinforcement learning, Ph.D. dissertation Dept. Computer Science., Univ of Illinois, MA, USA

N.O. Sjomlj, and M. Radujkovic (2012) S-curve modeling in early phases of construction projects, Gradevinar, 64(8), 647-654

H. Mao, Z. Gong, and Z. Xian (2020) Reward design in cooperative multi-agent reinforcement learning for packet routing, arXiv preprint arXiv:1605.06676

J.N. Foerster, Y.M. Assael, N. De Freitas, and S. Whiteson (2016) Learning to communicate with deep multi-agent reinforcement learning, arXiv preprint

Q. Meng, W.H. Lam, and L. Yang (2010) General stochastic user equilibrium traffic assignment problem with link capacity constraints, J. Adv. Transp,. 42(4), 429-465

R. Yang, X. Sun, and K. Narashimhan (2019) A generalized algorithm for multi-objective reinforcement learning and policy adaptation, arXiv preprint arXiv:1908.08342

I. Ekowicaksono, F. Bukhari, and A. Aman (2016) Estimating origin-destination matrix of bogor city using gravity model, IOP Conf. Ser. Earth Environ. Sci., 31(1), 012021

L. Baird, and A.W. Moore (1999) Gradient descent for general reinforcement learning, Adv. Neural Inf. Process. Syst, 968-974

M. Hausknecht, and P. Stone (2015) Deep reinforcement learning in parameterized action space, arXiv preprint arXiv:1511.04143

H. Mossalam, Y.M. Assael, D.M. Roijers, and S. Whiteson (2016) Multi-objective deep reinforcement learning, arXiv preprint arXiv:1610.02707

B. Rouhanzadeh, S. Kermanshachi, and T.J. Nipa (2019) Identification, categorization, and weighting of barriers to timely post-disaster recovery process, Comput. Civ.Eng.2019: Smart Cities Sustain. Resil. – Sel. Pap. ASCE In. Conf. Comput. Civ. Eng., 41-49

A. Rajabifard, R.G. Thompson, and Y. Chen (2015) An intelligent disaster decision support system for increasing the sustainability of transportation networks, Nat Resour Forum, 39(2), 83-96

L. Nguyen, Z. Yang, J. Zhu, J. Li, and F. Jin (2018) Coordinating disaster emergency response with heuristic reinforcement learning, arXiv preprint arXiv:1811.05010

S. Yang, Y. Ogawa, K. Ikeuchi, Y. Akiyama, and R. Shibasaki (2019) Firm-level behavior control after large-scale urban flooding using multi-agent deep reinforcement learning, GeoSim'19, 24-27

Japan News, "Worst damage from heavy rain and single flood in Western Japan: 1.215 trillion yen.", Mar. 24, 2020. [Online] Available:https://www.yomiuri.co.jp/national/20200324-OYT1T50237/

Japan Digital Road Map., "What is the DRM Database?", 2005. [Online] Avaliable: http://www.drm.jp/english/drm/database/structure.html

CNN, "Japan floods: Death toll rises to 200 as UN offers assistance", Jul. 12, 2018. [Online] Available: https://edition.cnn.com/2018/07/10/asia/japan-floods-intl/index.html